Sparse optimal control-based features generation for physiological time-series : application to the diagnosis of human diseases & disorders

Houssem Meghnoudj

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)
Spécialité : Automatique - Productique
Unité de recherche : Grenoble Images Parole Signal Automatique

Génération de caractéristiques à partir de séries temporelles physiologiques basée sur le contrôle optimal parcimonieux application au diagnostic de maladies et de troubles humains

Sparse optimal control-based features generation for physiological time-series: application to the diagnosis of human diseases & disorders

Présentée par

Houssem MEGHNOUDJ

Direction de thèse :

Mazen ALAMIR DIRECTEUR DE RECHERCHE, Université Grenoble Alpes	Directeur de thèse
Bogdan ROBU MAITRE DE CONFERENCES, Université Grenoble Alpes	Co-ancadrant

Rapporteurs :

Antoine CHAILLET
PROFESSEUR DES UNIVERSITES, CentraleSupélec
Axel HUTT
DIRECTEUR DE RECHERCHE, INRIA Strasbourg

Thèse soutenue publiquement le **18 janvier 2024**, devant le jury composé de :

Mazen ALAMIR DIRECTEUR DE RECHERCHE, CNRS	Directeur de thèse
Antoine CHAILLET PROFESSEUR DES UNIVERSITES, CentraleSupélec	Rapporteur
Axel HUTT DIRECTEUR DE RECHERCHE, INRIA Strasbourg	Rapporteur
Isabelle QUEINNEC DIRECTRICE DE RECHERCHE, CNRS	Examinatrice
Didier GEORGES PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Président

Résumé :

Dans cette thèse, une nouvelle méthodologie a été proposée pour la génération de caractéristiques à partir de signaux physiologiques afin de contribuer au diagnostic d'une variété de maladies cérébrales et cardiaques. Basée sur le contrôle optimal parcimonieux, la génération de caractéristiques dynamiques parcimonieuses (SDF) s'inspire du fonctionnement du cerveau. Le concept fondamental de la méthode consiste à décomposer le signal de manière parcimonieuse en modes dynamiques qui peuvent être activés et/ou désactivés au moment approprié avec l'amplitude adéquate. Cette décomposition permet de changer le point de vue sur les données en donnant accès à des caractéristiques plus informatives qui sont plus fidèles au concept de production des signaux cérébraux. Néanmoins, la méthode reste générique et polyvalente puisqu'elle peut être appliquée à un large éventail de signaux. Les performances de la méthode ont été évaluées sur trois problématiques en utilisant des données réelles accessibles publiquement, en abordant des scénarios de diagnostic liés à : (1) la maladie de Parkinson, (2) la schizophrénie et (3) diverses maladies cardiaques. Pour les trois applications, les résultats sont très concluants, puisqu'ils sont comparables aux méthodes de l'état de l'art tout en n'utilisant qu'un petit nombre de caractéristiques (une ou deux pour les applications sur le cerveau) et un simple classifieur (souvent linéaire) suggérant la robustesse et le bien-fondé des résultats. Il convient de souligner qu'une attention particulière a été accordée à l'obtention de résultats cohérents et significatifs avec une explicabilité sous-jacente.

Mots clés : Systèmes dynamique; Apprentissage automatique; Caractéristiques éparses; Données physiologique; Control optimal.

Abstract :

In this thesis, a novel methodology for feature generation from physiological signals (EEG, ECG) has been proposed that is used for the diagnosis of a variety of brain and heart diseases. Based on sparse optimal control, the generation of Sparse Dynamical Features (SDFs) is inspired by the functioning of the brain. The method's fundamental concept revolves around sparsely decomposing the signal into dynamical modes that can be switched on and off at the appropriate time instants with the appropriate amplitudes. This decomposition provides a new point of view on the data which gives access to informative features that are faithful to the brain functioning. Nevertheless, the method remains generic and versatile as it can be applied to a wide range of signals. The methodology's performance was evaluated on three use cases using openly accessible real-world data: (1) Parkinson's Disease, (2) Schizophrenia, and (3) various cardiac diseases. For all three applications, the results are highly conclusive, achieving results that are comparable to the state-of-the-art methods while using only a few features (one or two for brain applications) and a simple (often linear) classifier supporting the significance and reliability of the findings. It's worth highlighting that special attention has been given to achieving significant and meaningful results with an underlying explainability.

Keywords: Dynamical systems; Machine learning; Sparse features; Physiological data; Optimal control.

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!"
but "That's funny..."

Isaac Asimov, *1920–1992*

God grant me the serenity to accept the things I cannot change,
Courage to change the things I can,
and Wisdom to know the difference.

Reinhold Niebuhr, *1892-1971*

Acknowledgements

Je tiens à exprimer ma gratitude à toutes les personnes qui ont contribué de près ou de loin à cette thèse.
Ce document marque l'achèvement d'un chapitre de ma vie, et de cette aventure, j'ai beaucoup appris. Pour cela, je suis entièrement reconnaissant.

En premier lieu, je souhaite réitérer mes remerciements à Antoine Chaillet et Axel Hutt d'avoir accepté d'être les rapporteurs de ma thèse, d'évaluer son contenu scientifique et de fournir des retours et corrections qui ont permis d'améliorer ce document ainsi que son contenu. Je tiens également à remercier les examinateurs : Isabelle Queinnec et Didier Georges pour leurs retours.

Un grand merci à Mazen Alamir et Bogdan Robu pour m'avoir donné la chance d'entamer ce doctorat à leurs côtés, pour leur encadrement tout au long de la thèse, et pour m'avoir accordé leur confiance et une totale liberté sur le sujet. Merci pour votre soutien technique et humain, vous étiez toujours présents quand j'en avais besoin. J'ai beaucoup appris grâce à vous, et je tiens à vous l'exprimer.

Je tiens à exprimer ma gratitude envers mes enseignants de l'École Nationale Polytechnique d'Alger pour le précieux savoir qu'ils m'ont transmis, dont je tire bénéfice chaque jour, ainsi que pour leur rigueur et leur sympathie. Un remerciement spécial à Messieurs Tadjine, Manaa, Kebli et Ouadjaout. Je souhaite également rendre un vibrant hommage à Abdelouel, qui malheureusement nous a quittés.

Merci à Virginie et Sonia, nos RH qui ont toujours été très efficace, à l'écoute de nos problèmes administratifs (qu'on adore tant) ainsi que pour leur sympathie et leur bonne humeur.

Nous avons souvent tendance à résumer l'expérience de ces trois années de thèse dans un manuscrit de quelques pages, où se trouvent résumés les travaux qui ont abouti. Le lecteur oublie parfois que ces trois années de thèse représentent trois années de vie humaine, remplies non seulement de travail, mais aussi de sentiments, de joies et de peines. Ce document s'efforce de relater l'histoire de nos recherches scientifiques, mais il occulte les personnes humaines qui ont été derrière nous, qui sans leur soutien, ce manuscrit ne serait pas ce qu'il est.

Aziza, toi qui m'as accompagné durant cette thèse, je tiens à exprimer ma profonde gratitude pour ton soutien indéfectible et tes encouragements. Merci de m'avoir poussé vers l'avant et pour toute ton aide et surtout pour ta précieuse compagnie durant cette thèse (et celle d'après). Merci du fond du cœur.

Un immense merci à Adlene qui m'a prodigué tant d'encouragements, me poussant sans relâche à aller de l'avant, même dans les moments les plus difficiles. Je te suis également reconnaissant pour toutes les aventures que nous avons partagées, pour tout le temps que nous avons passé ensemble, et pour ta précieuse compagnie.

Je tiens également à exprimer ma gratitude envers mes frères d'armes : Estebanc et Mariana, sans lesquels cette thèse aurait été beaucoup moins joyeuse. Nous avons partagé tant de moments ensemble au fil des années, des rires, des peines et tant d'activités. Merci du fond du cœur.

Un grand merci également à mes camarades qui ont enrichi ma vie de leur présence et pour tous les moments que nous avons partagés ensemble : Bob, Andrea, Elise, Adrien, Maria, Ana, Lucas, Louise, Ariel, Chhay, Kaouther, et Mukhtar.

Enfin, je souhaite conclure en exprimant ma profonde gratitude envers mes parents et mes frères, sans qui je ne serais pas la personne que je suis aujourd'hui et je n'aurais pas accompli autant dans ma vie (bien qu'il reste encore beaucoup à faire).
Ils ont toujours été présents à mes côtés quelles que soient les circonstances et ils sont mon modèle et mon moteur. Il y a une chose pour laquelle je serai reconnaissant toute ma vie, c'est leur éducation, et les quelques mots écrit ici ne sont suffisant pour exprimer ma gratitude et amour envers eux.

ⵜⴰⵏⴰⵎⵉⵔⵜ : *thanemirt* [merci]

Table of Contents

Notations

T_s	Sampling period
f_s	Sampling frequency
N	Number of subjects
m	Number of harmonic oscillators (pendulums — modes) used in the model
ω_i	Natural pulsation of the i-th oscillator
Y	Signal of interest
$\hat{Y}$	Predicted signal
L	Length of the signal of interest and the predicted signal
y_k, $\hat{y}_k$	k-th element of the signal of interest and predicted signal respectively
Σ	State space system describing the model
a_i, b_i , c_i	Respectively the state, input, and output matrices describing the dynamics of the i-th oscillator
A, B, C	Respectively the state, input, and output matrices describing the dynamics of the system merging the m decoupled oscillators
x_k	State vector at the time instant k
x_0	Initial state (initial position and velocity of each oscillator)
$u_i(k)$	Excitation force/input of the i-th oscillator at the time instant k
u_k	Input vector that contain the excitation forces of all the m oscillators at the time instant k
U	Control sequence vector
$\tilde{U}$	Control sequence rewritten in matrix/image form
ϕ_1, ϕ_2	Covariate matrix of the free and forced responses respectively
$\underline{\phi_1}$, $\underline{\phi_2}$	Vector-wise standardised version of ϕ_1 and ϕ_2
ϕ	Concatenation of ϕ_1 and ϕ_2
$\phi_{*,i}$	i-th vector of ϕ

β	Vector that contain the initial state x_0 and the control sequence U
β_{OLS}	Solution of β using Ordinary Least Square algorithm
β_0	Intercept term
α	Weighting factor that regularises the sparsity of β
α_{init}	Initial value of α initialised by the Lasso-LARS
α_f	Final value of α set by the expert
$\alpha_{k\%}$	k-th percent sparsity level
ω	Weighting constant favouring forced/free response
r	Residue between Y and $\hat{Y}$
l	Linear spacing of the sparsity interval
n	Number of solution produced by the Lasso-LARS
G	Configuration composed of: a number of features and samples, a classifier, an underlying data distribution
R	The a priori accuracy associated being in configuration G
K	Number of folds for K-fold cross-validation
W	ICA matrix projection (channels base $\rightarrow$ independent components base)
IC_i	i-th Independent Component
I_1, I_2	Time intervals of interest
F_1, F_2	Respectively Feature 1 and Feature 2 extracted from U
$F_1(Y)$, $F_2(Y)$	Feature F_1 and F_2 extracted on the signal of interest Y
H_0	Null hypothesis
p	Probability / p-value of observing the result given H_0
TP	True Positive rate

Introduction

To describe the evolution of a phenomenon over time, often a temporal representation of signals is employed. This representation describes the phenomenon as a sequence of well-indexed and ordered numbers that follow one another. The importance of time and the ordering of events of the phenomenon make time series an interesting and natural representation. This representation is widely used in science and engineering, and it becomes difficult to remember that it is one among many possible representations that one might use to describe and study the phenomenon.

Sometimes, this perspective or representation can hinder our understanding. Changing our point of view might allow us to uncover certain hidden information that was initially obscured by the representation. The most striking example of a change in perspective is the Fourier decomposition. It enables a shift from a classical time-based representation of a signal to a representation that is more informative and useful in some specific cases. As illustrated in Figure 1, the signal is decomposed into multiple sinusoids, the sum of which forms the original signal. This representation finds numerous applications, and to mention just one: filtering. It is by changing the perspective that the contribution of each sinusoid can be discerned. It is easier to identify those that pertain to the observed phenomenon and are desirable, as well as those that are undesirable, such as noise. This task is impossible to accomplish directly with the temporal representation and it is the change of perspective that renders this information accessible and clearer.

The aim of our research is to propose a novel decomposition that is much more suitable to excitation-based signals, and where the temporal aspect that is lost during Fourier decomposition is detrimental. The methodology draws inspiration from the functioning of the brain, yet remains generic and applicable to various signal types. It enables the signal to be decomposed into what, control experts, refer to as modes. The key aspect lies in the fact that these modes can be activated or deactivated at any given moment. The modes' excitations are computed optimally to fit the signal to be decomposed. To enhance the decomposition robustness and meaningfulness, it has been designed following the principle of parsimony.

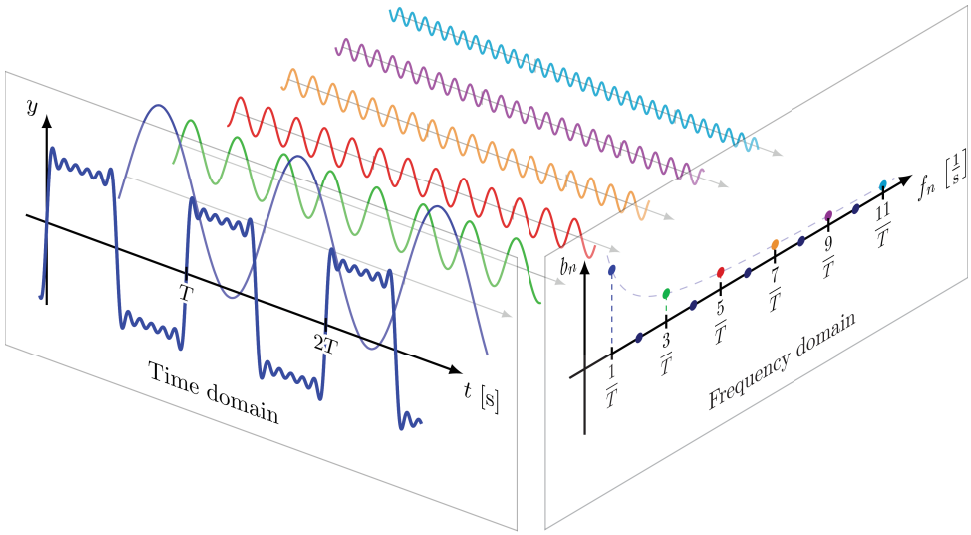


Figure 1: Fourier decomposition.

The ultimate objective is to successfully apply this methodology to diverse real-world problems, thereby demonstrating its relevance and assessing its performance. The applications considered are all healthcare-related, spanning from the analysis of electroencephalogram (EEG) signals to electrocardiogram (ECG) signals.

This manuscript focuses on the proposed decomposition that provides what we have named "Sparse Dynamical Feature" (SDF). The methodology's performance was evaluated across three problems using real data, encompassing diagnostic scenarios related to: (1) Parkinson's Disease, (2) Schizophrenia, and (3) six Cardiac Abnormalities. Special attention has been given to achieving consistent and meaningful results with an underlying explainability. The various problems and biases encountered in the literature that affect the assessment of the model's performance are also addressed.

The manuscript is structured as follows:

- Chapter 1 describes the methodology behind Sparse Dynamical Feature generation. Starting with the conceptual and inspirational foundation, followed by the implementation of the idea, encompassing model definition and resolution of the of specified optimization problem. The method's various parameters are meticulously detailed while illustrating their effect on the generation of SDFs. Automatic procedures to set these latter are proposed. Drawbacks of high-dimensional features are recalled and explained followed by the feature extraction that is performed on the SDFs. The classification & evaluation methods that are less biased and that will be used for the entire manuscript are addressed in the final part. In the conclusion of this Chapter, the main novelty and contribution of our work have been outlined.

- Chapter 2 is entirely devoted to the recording systems that produced the data used in the manuscript. In this chapter, a detailed study of the fundamental principles of electroencephalography and electrocardiography is provided.
- Chapter 3 is dedicated to the first application, which focuses on diagnosing Parkinson's disease based on EEG signals.
- Chapter 4 is devoted to the second application, which is the diagnosis of Schizophrenia and which is also based on EEG data. This part was mainly done to test the genericity of the method for another brain disease.
- Chapter 5 is dedicated to the third application, which is the detection of six abnormalities affecting the heart using ECG signals. The aim is to test the methodology in a completely different context (different signals) with a particular emphasis on the visualisation of the SDFs which show notable differences between the different categories and on which the results are based.

The manuscript is ended by a general conclusion that summarizes the contribution of the PhD work and draws some interesting tracks for further investigation.

1 Methodology

Contents

1.1 Introduction

The most important contribution of the PhD work is the development of the so-called Sparse Dynamical Features (SDF) which is a block that decomposes pre-processed signal, thus offering a new point of view on the data and providing access to new biomarkers that are much more informative and suitable for the applications considered later on. Figure 1.1 is a flowchart illustrating all the steps through which the data passes starting from raw data until classification and performance evaluation. The main contribution is coloured in blue.

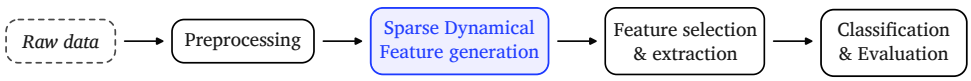


Figure 1.1: Flowchart showing the important steps that the raw data underwent until the classification.

This chapter will be exclusively dedicated to the study of the last 3 blocks of Figure 1.1. The core idea of the method and the crucial points around which the decomposition revolves are enumerated first and foremost. These will serve as guidelines to shape the methodology, which is subsequently detailed. The model that will serve as the foundation for signal decomposition, along with the optimization problem, is then described. Various approaches for solving the underlying optimization problem are presented, each with its own advantages and drawbacks.

The choice of Lasso-LARS method is discussed, followed by an in-depth description of the associated key parameters. Subsequently, the different methods of feature extraction and selection that have been applied are described, highlighting the significance and advantages that feature selection offers. Finally, the classification problem, along with the different classifiers employed, is addressed, followed by a discussion of the various data leakage issues and cross-validation procedures used to address them, at least partially.

1.2 Concept & idea

To process, encode, retrieve and transmit information, biological neuronal networks oscillate [101, 11]. This oscillation can be compared to that of a classical harmonic oscillator, which results in a sine wave of position versus time. The frequencies and timings (phase) of these network oscillations are important as they are at the basis of the mechanism underlying cognitive processes [6, 29]. The oscillation frequencies, as suggested by various studies, are task-dependent. A slow oscillation is observed for tasks that do not require attention e.g. biological needs, and a rapid oscillation is observed for tasks where the person's attention is required [101].

At the scale of one neuron, the oscillation is not important as it will not have a great impact on its own. However, it is this mechanism of individual oscillation performed with the right timing, that leads local-excitatory neurons to a global synchronisation. Furthermore, synchronization often emerges through the interplay between excitatory and inhibitory signals which leads to coordinated activity among neurons. The synchronisation of multiple neurons oscillation is how the brain achieves the large-scale integration of its many parallel, distributed information-processing activities, which leads to coherence cognition and behavior. Therefore, the timing of these oscillations has a great importance. Contrary to the synchronisation, a global inhibitory mechanism exists in the brain and can lead to rapid de-synchronization of a group of neurons [101].

It should be noted that the measured activity of the brain is a measure of a group of neurons oscillating in synchrony. Thus, the mechanism of synchronization and de-synchronization of a group of neurons suggests that the rhythms contributing to that measure occur in a **pulsatory** manner [68].

It is important to note that this rhythmic and pulsating excitation is not unique to the functioning of the brain, other systems that are familiar to the control community also show the same particularity. One can imagine, for instance, that there is a change in a single mode that contributes to the output of that system, or an outright change in the mode of operation of the dynamic system, such as a model with a switched parameter. As a consequence, the modelling one seeks must include this kind of aspects. In summary, the method is inspired by the functioning of the brain, but it can be applied to other problems that share these same properties to a sufficient extent.

The majority of signals we encounter in real life can be decomposed into two types of signal/response:

1. **A free response**: this is the system output without any excitation being perceived, which is often zero, but in some cases, it is not.
2. **A forced response**: this represents the part of the system's behaviour that is induced by an excitation signal.

We want the method to respect this philosophy too, and not only separate the two types of response, but also to be able to control the amount of information carried in one or the other. Indeed, each task has its own necessities and particularities, so we want in our framework to have the choice of putting more information on the forced regime or the self-perpetuating regime. More details will be given later for each application. This amounts to having tunable parameters as it is always the case for the models to be learned or trained.

Considering a set of m harmonic oscillators (pendulums — modes) with distinct natural pulsations $\{\omega_1, \omega_2, \ldots, \omega_m\}$, the signal of interest can be represented as a combination of the outputs of these oscillators, where each one of them contributes with its own amplitude as in the Fourier series. However, compared to the Fourier series, now the modes are allowed to be active over either on all the temporal support or only on a portion of it. Moreover, the temporal information about when an oscillator is switched on or off is now recovered. Knowing that several rhythms can coexist at the same time and that each one of them has a different activation duration, the temporal information that cannot be recovered using the Fourier transform, is of great importance.

Before moving on and describing in depth, the detailed implementation of this intuition, let's assume that we have a real signal of some kind and define the objectives and expectations for our method applied to this signal. The method will transform the signal and this transformation should:

- Disentangle and separate the contribution of each mode to the signal.
- Indicate the instants at which an oscillator is excited if any.
- Indicate which oscillators are involved as well as their amplitude contribution.
- Separate the forced response from the free response composing the signal.

Such a transformation allows for a change in the point of view on the data, this latter is no longer perceived as a traditional time series (sequence of numbers), but rather as the consequence of different modes that are switched on and off at the right moments with the right amplitude and which hopefully imitate or discover what has physiologically happened that produced the measured signals. This change of point of view hopefully gives access to new features that might be more informative for the applications considered. In our approach, we have turned to a principle that we consider important:

PARSIMONY

This principle consists of using the minimum number of elementary causes to explain a phenomenon.

Given that the studied signals are noisy, it is essential to find only the oscillations that are important, rather than finding all the oscillations in the signal (which may also be due to noise). This enhances the prediction accuracy and interpretability of the resulting statistical model. But parsimony is a rather fuzzy concept. This is why a parameter is required in order to adjust the parsimony's level. This parameter controls the number of oscillators that can be awakened/used by the method and the frequencies to be involved. A high level of parsimony leads to the use of fewer oscillators and they will be excited less frequently. This is obviously tightly lined to the concept/technique of ℓ_1-regularization that is often used to avoid the over-fitting issue in the learning literature.

In the following section, a rigorous presentation of the method briefly sketched above is proposed.

1.3 Sparse Dynamical Features

1.3.1 Model definition

In order to put the idea described in the previous subsection into practice, a model consisting of a battery of oscillators with distinct angular frequencies is used. The appropriate oscillators are switched on and off at a specific timing with the appropriate excitation amplitude, so that the output of the model $\hat{Y}$ approximately matches to the signal of interest Y to an extent that is tailored through a sparsity parameter. The excitation inputs applied to the oscillator array are named Sparse Dynamical Features (SDF). The word *sparse* refers to the fact that the appropriate excitation inputs have been generated in a parsimonious way.

Consider a signal of interest Y of length L written in a vector form along with it is predicted signal $\hat{Y}$, the output of our model, denoted as follows:

$$Y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{L-1} \end{pmatrix} \in \mathbb{R}^L, \qquad \hat{Y} = \begin{pmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_{L-1} \end{pmatrix} \in \mathbb{R}^L, \tag{1.1}$$

where y_k and $\hat{y}_k$ are k-th element of the signal Y and $\hat{Y}$ respectively.

Consider m oscillators (modes) with an increasing angular frequency $\{\omega_1, \omega_2, \ldots, \omega_m\}$. The system combining the m decoupled harmonic oscillators can be described by the following discrete-time state space representation given a sampling period $T_s = 1/f_s$:

$$\Sigma: \begin{cases} x_{k+1} = Ax_k + Bu_k \\ \hat{y}_k = Cx_k \end{cases}; \qquad x_k \in \mathbb{R}^{2m}, \quad u_k \in \mathbb{R}^m, \quad \hat{y}_k \in \mathbb{R}, \tag{1.2}$$

where x_k is the state vector, u_k is the input vector and $\hat{y}_k$ is the output of the system Σ at the instant k, and:

$$A = \mathrm{diag}(a_1, \ldots, a_m), \quad B = \begin{pmatrix} b_1 & 0_{2\times1} & \ldots & 0_{2\times1} \\ 0_{2\times1} & b_2 & \ldots & 0_{2\times1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2\times1} & 0_{2\times1} & \ldots & b_m \end{pmatrix}, \quad C = (c_1 \ \cdots \ c_m)\,.$$

Notice that $\mathrm{diag}(a_1, \ldots, a_m)$ is a block diagonal matrix with the elements $\{a_1, \ldots, a_m\}$ being on the main diagonal and zero entries elsewhere. More precisely we have:

$$a_i = expm(T_s \begin{pmatrix} 0 & 1 \\ -\omega_i^2 & 0 \end{pmatrix}) \approx \begin{pmatrix} 1 & T_s \\ -T_s\omega_i^2 & 1 \end{pmatrix}, \quad b_i = \begin{pmatrix} 0 \\ T_s \end{pmatrix}, \quad c_i = \begin{pmatrix} f_s\,\omega_i & 0 \end{pmatrix}, \tag{1.3}$$

are the matrices of a unitary gain forced single harmonic oscillator: a_i contains the pendulum dynamics, b_i is the input matrix, and c_i is the output matrix. expm denotes matrix exponential, it is important to note that the approximation is given for the reader only, and that the exact value of the matrix exponential was used in our implementation.

The m modes are merged throughout the matrices A, B and C such that the m oscillators remain decoupled but also that their contribution is summed up to form the signal $\hat{Y}$.

The input vector u_k contains the excitation forces of all the m modes at instant k :

$$u_k = \begin{pmatrix} u_1(k) & u_2(k) & \ldots & u_m(k) \end{pmatrix}^T \tag{1.4}$$

The form of the matrix b_i has been chosen such that the i^{th} pendulum is controllable by its corresponding control input $u_i(k)$.

The role of the scaling term ω_i present in the matrix c_i is to ensure that an oscillator at rest ($x_i = 0$), excited by an input $u_i(k) = z$, starts oscillating at time instant k with the same amplitude as the excitation, i.e. z. This gives a physical meaning to the values of u_k as they will be directly proportional to the oscillation amplitude present in the signal of interest Y. This also opens up the possibility of comparison between the u_i values since they are now on the same scale. We can now compare the modes in terms of their amplitude contribution, but also compare their energy, activity duration, starting time, number of times they have been excited, etc. From this representation it can be noted that the analysis of the excitation inputs u can be conducted on a single specific oscillator or a desired set of oscillators.

The system (1.2) can be then rewritten in the explicit form as:

$$x_k = A^k x_0 + \sum_{i=0}^{k-1} A^i B\, u_{k-1-i}, \quad k = 1, 2, \ldots, \tag{1.5}$$

$$\hat{y}_k = CA^k x_0 + \sum_{i=0}^{k-1} CA^i B\, u_{k-1-i}, \quad k = 1, 2, \ldots, \tag{1.6}$$

where x_0 represents the initial oscillating state of our model (initial position and velocity of each pendulum). If no external input is applied on these pendulums as in the case of the non-forced regime, i.e. $u_k = 0$ for $k = 1, 2, \ldots$, the pendulums will keep swinging in the same manner. Therefore, the information about the free response where no external stimulus is carried by x_0.

To simplify and rewrite the system in the appropriate format for the following equations, the control sequence U which embeds how the oscillators are excited over time is defined as [see equation (1.4)]:

$$U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{L-1} \end{pmatrix} \in \mathbb{R}^{m(L-1)}, \quad \begin{matrix} u_1(2) \\ u_2(2) \\ \vdots \\ u_m(2) \end{matrix} \tag{1.7}$$

By so doing, the information regarding which of the m oscillators is activated is embedded in U. The temporal information of when the excitation arrives is indicated by the active subscript k. Notice that this is relevant only because parsimonious computation of U is used later on, otherwise, all the components of U would have been non-zero.

Given (1.1) and (1.7), the equation (1.6) can be rewritten in a matrix form as:

$$\begin{pmatrix} \phi_1 & \phi_2 \end{pmatrix} \cdot \begin{pmatrix} x_0 \\ U \end{pmatrix} = \phi \cdot \beta = \hat{Y}, \tag{1.8}$$

where:

$$\phi_1 = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{L-1} \end{pmatrix} \in \mathbb{R}^{L \times 2m}, \qquad \phi_2 = \begin{pmatrix} 0 & 0 & \dots & 0 \\ CB & 0 & \dots & 0 \\ CAB & CB & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{L-2}B & CA^{L-3}B & \dots & CB \end{pmatrix} \in \mathbb{R}^{L \times m(L-1)}.$$

The ϕ matrix is known as the **covariate** matrix and a column vector of ϕ is known as a **covariate** vector.

1.3.2 Model familiarisation

In this section we will familiarise ourselves with the proposed model. It is important to note before starting that the model does not come from a modelling or a study carried out on a real signal but rather it is a model that we have imposed and that we will excite in the right manner so that it fits the designed signal of interest. For the following examples, instead of considering a battery of oscillators, a single oscillator will be considered. Since all the oscillators are decoupled, the same scenarios that will be defined in the following will apply:

1) From a resting oscillator to the first excitation (forced regime)

Initially we assume that the oscillator is at rest ($x_0 = 0$). If no force is applied to it i.e. ($U = 0$), the oscillator will remain at rest as shown in Figure 1.2.a. This oscillator is then excited by an impulse of amplitude 1 at time $k = 100$ in other words $U = (0, 0, \dots, 0, u_1(k = 100) = 1, 0, \dots, 0)^T$. It can be noted that the U vector is sparse (an important point for what follows). The profile of the excitation U and the output $\hat{Y}$ of the model Σ are shown in Figure 1.2.b and 1.2.c respectively.

Figure 1.2 shows that the output of the model $\hat{Y}$ is 0 before the impulse arrival, then as soon as the impulse arrives the model starts to oscillate in a pure sinusoidal way with the same amplitude as the given impulse, i.e. 1. This oscillation continues until the simulation stops. It is important to note that the system is stable and that the oscillation amplitude does not diverge, and that the oscillator considered in (1.2) is not damped.

Figure 1.3 shows the response of the system to an excitation with amplitude -4. We can observe that the oscillation amplitude is 4 and that the direction of the first oscillation points downwards (whereas it pointed upwards in the case of Figure 1.2.c). Moreover, the angular frequency of the oscillator is ω_1.

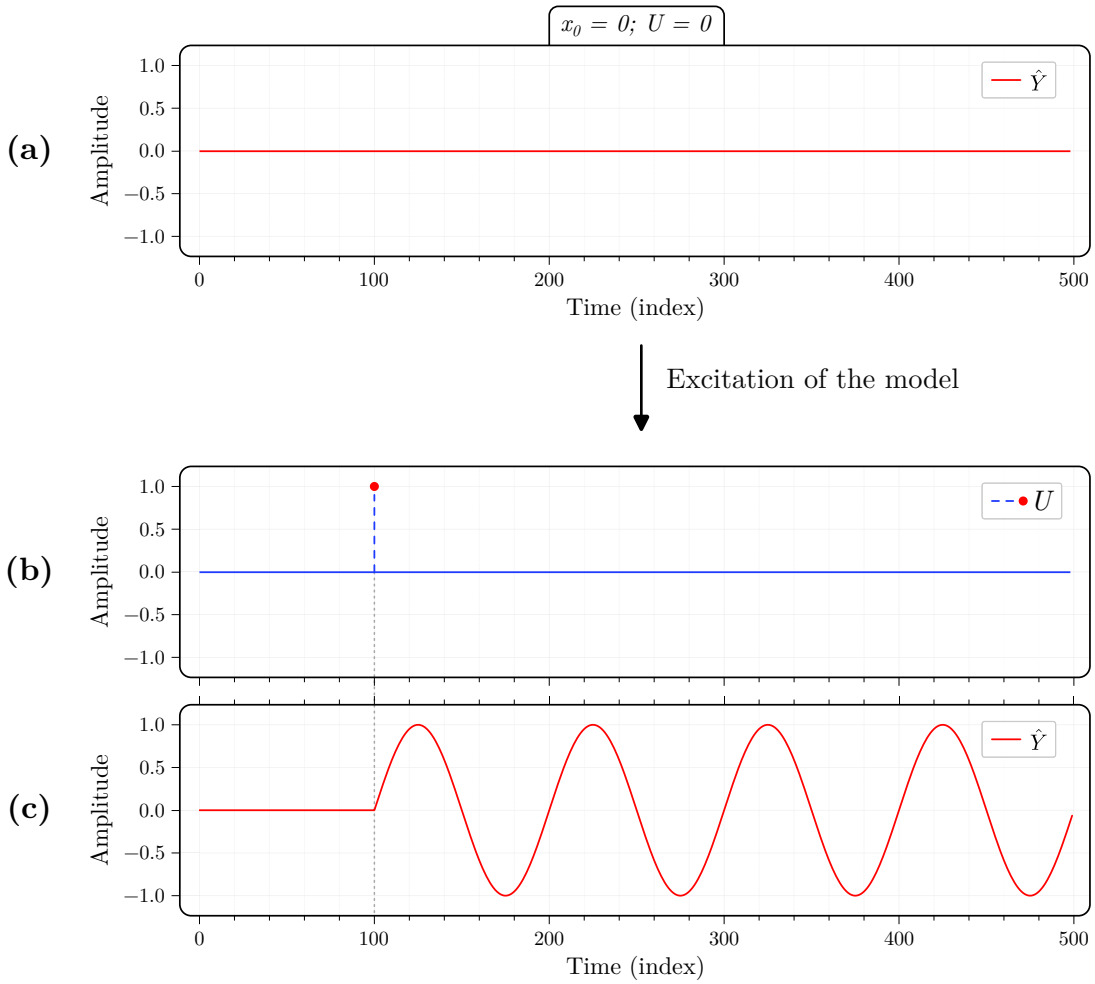


Figure 1.2: **(a)** Output $\hat{Y}$ of the system Σ at rest and where no excitation is applied: $(x_0 = 0)$ and $(U = 0)$. **(b)** Impulse excitation applied to generate the output presented in **(c)**.

2) Stopping the oscillation

An excited oscillator can be switched off with an impulse of the opposite amplitude to that which made it start[1] as shown in Figure 1.4. For this example, the oscillation frequency has been set to ω_2 with $\omega_2 > \omega_1$ for demonstrative purposes.

3) Free regime contribution

In the previous examples, the oscillator was assumed to be initially at rest $(x_0 = 0)$, however, in a practical case it can be interesting to have oscillators with non-zero initial positions (or speeds). The benefit of having the contribution of x_0 is that it allows the free response of a signal to be captured. Figure 1.5 shows the output of the model if no excitation U is applied $(U = 0)$ while the system starts from an initial point $(x_0 = [1, 0]^T)$.

[1] The same principle also works with the excitation of x_0.

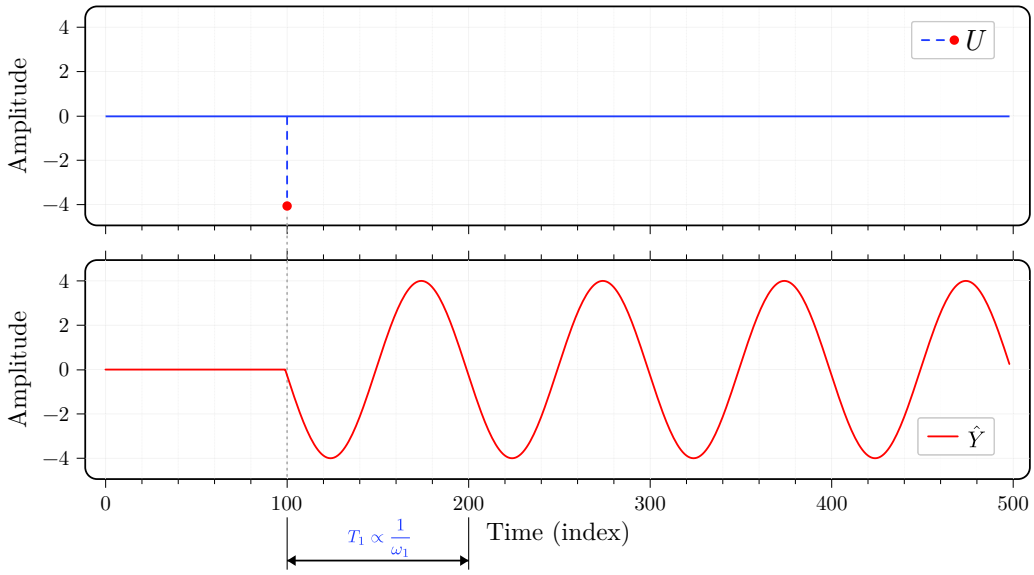


Figure 1.3: Amplitude variation.

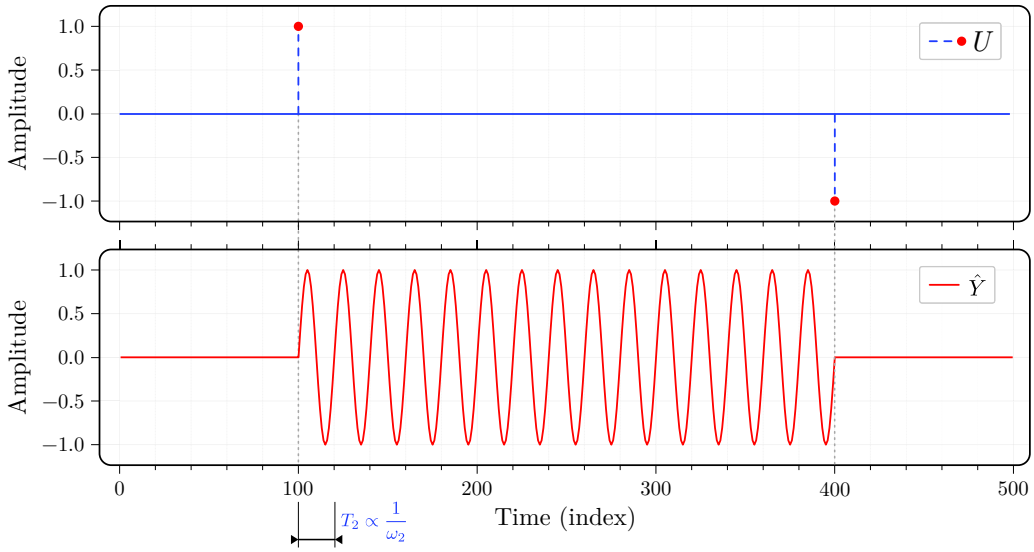


Figure 1.4: Switching off an oscillator.

4) Superposition of the forced and free regime

The free and forced regimes can coexist at the same time, giving our model more degrees of freedom to generate outputs as shown in Figure 1.6.

In Figure 1.6 it can be observed the individual contribution of the free regime between 0 and 100 and between 400 and 500. The forced response of the system can be observed between time instants 100 and 400, where the forced activity if entangled with the free regime and has a different frequency.

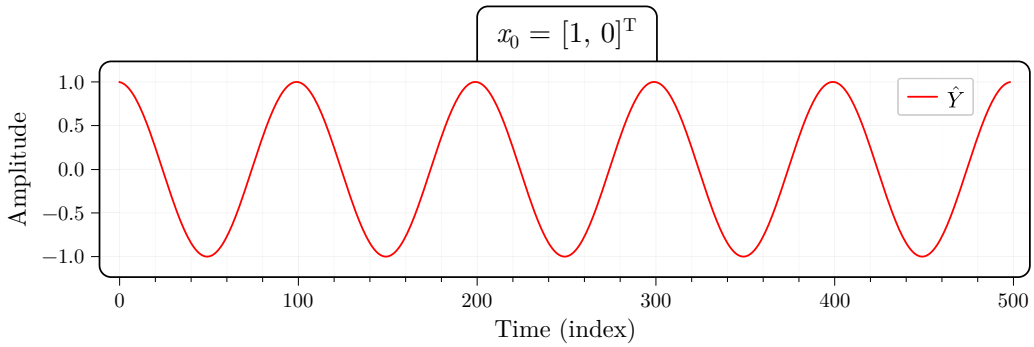


Figure 1.5: Free response of the system Σ while $x_0 = [1, 0]^T$ and $U = 0$

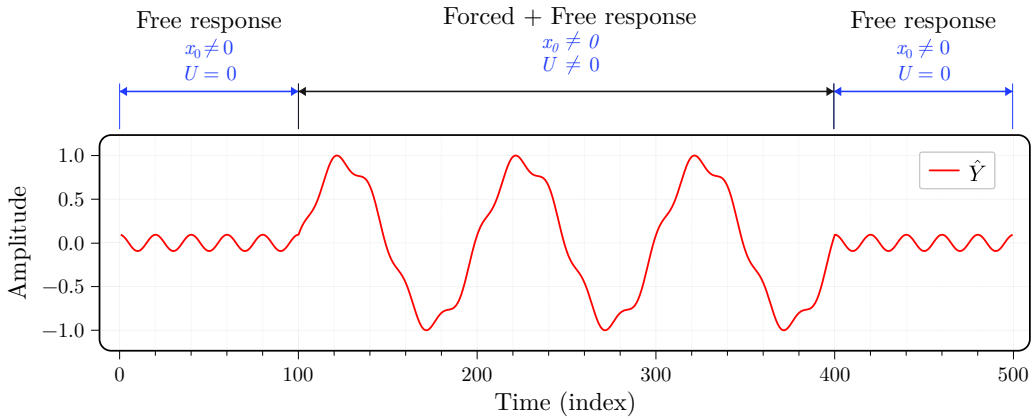


Figure 1.6: Superposition of the forced and free regime.

The examples given previously are only explanatory. The use of a battery of oscillators with an appropriate frequency span $\{\omega_1, \omega_2, \ldots, \omega_m\}$, the right initial state x_0, and excitation U are advantageous for fitting a real signal that is more complex as we will see in the following sections (to give an overview, one can see Figure 1.12).

In the previous examples, we applied U commands and set x_0 values manually. In the next section, it is shown that the computation of the vector U together with the initial state x_0 can be performed for a given measured profile Y using a dedicated parsimony-aware solver.

1.3.3 The excitation input profile as solution to parsimonious optimisation problems

A primary solution

Various approaches to find a solution for the unknown β of (1.8) exist. The first, which is the simplest and cheapest, is the Ordinary Least Squares (OLS) method [40, chap. 2].

It minimises the squared prediction error and its solution is given by:

$$\beta_{OLS} = \min_{\beta} \| Y - \phi\,\beta \|_2^2, \tag{1.9a}$$

$$\beta_{OLS} = \left[\phi^T \phi\right]^{-1} \phi^T Y, \tag{1.9b}$$

where β_{OLS} is the solution of solution of the optimization problem. Despite its simplicity, the least square method does not meet our requirements due to the fact that the resulting β will be full (not sparse), thus, not respecting the parsimony principle.

Methods that comply with the parsimony principle

Complying with the principle of parsimony translates into using as few excitation inputs as possible to fit the signal of interest to a decent level. Obviously, this statement is quite fuzzy and only a rationale tuning of the degree of sparsity viewed as a parameter of some learning strategy would enable to perform an appropriate sparsity level.

One type of method which allows this are referred as subset selection. Different methods exist such as: Sequential Forward Selection, Sequential Backward Elimination, Forward-Stagewise Regression, etc. (for more details see [40]).

Several tests using these methods were carried on during this PhD work, however, the results were not conclusive. Moreover, for some methods, the computation time is unrealistic (several hours for a single signal). On the contrary, the ones based on sparse optimization methods, via ℓ_1 regularization showed more effectiveness. This is explained in the remainder of this section.

Lasso: (Least absolute shrinkage and selection operator)

Another type of method that complies with the principle of parsimony is the shrinkage method. The most fundamental of these is the Lasso which stands for Least Absolute Shrinkage and Selection Operator. It solves the linear regression (1.8) while imposing certain coefficients to be zero, thus excluding them from impacting the prediction and thus respecting the principle of parsimony.

Its aim is to find the subset of $\leq j$ covariates that results in the smallest value of the objective function (1.10). The optimisation problem is formulated as follows:

$$\beta = \underset{\beta}{\operatorname{argmin}} \left\| Y - \phi\,\beta \right\|_2^2, \tag{1.10}$$
$$s.t. \quad \left\| \beta \right\|_0 \leq j.$$

where the l_0-norm[2] (nuclear norm) is defined as the number of non-zero elements of β. By doing so, the solution of the problem (1.10) is restricted to use only k elements of β.

[2] l_0 is not a proper norm in the mathematical and conventional way as if violates the absolute homogeneity of a norm ($\| \lambda x \|_0 \neq |\lambda| \| x \|_0$ where $\lambda \in \mathbb{R}\backslash\{0,1\}$ and $x \neq 0$), by abuse of language the term norm will be used.

The l_0-norm is non-convex, as illustrated in Figure 1.7. This non-convexity is problematic, as we are no longer able to use the various solvers and optimizers available in the field of quadratic programming. The latter are powerful and handle very well quadratic problems subject to convex constraints. The Lasso method can also be viewed as a convex relaxation of the best subset selection problem described above as it utilizes the best convex approximation to the latter by using l_1-norm which is the closest convex norm to l_0.

The lasso problem can be written as:

$$\begin{aligned} \beta = \underset{\beta}{\operatorname{argmin}} \left\| Y - \phi\,\beta \right\|_2^2, \\ s.t. \quad \left\| \beta \right\|_1 \leq j. \end{aligned} \tag{1.11}$$

The equivalent manner to write the problem is:

$$\beta = \underset{\beta}{\operatorname{argmin}} \left\| Y - \phi\,\beta \right\|_2^2 + \alpha \left\| \beta \right\|_1, \tag{1.12}$$

where α is a weighting constant that manages the trade-off between fitting the Y signal tightly and using a high number of active excitation inputs (on zero elements of β) or not fitting the Y signal tightly resulting in a sparser β.

As expected, this scheme results in obtaining sparse solutions, but the correct value of α which gives the best classification score is not known beforehand. As a consequence, the value of α has to be changed several times, thereby requiring all the calculations to be redone. In addition, the calculation time is relatively high, making the method impractical for a database containing a large number of signals.

The next section will be dedicated entirely to the Lasso-LARS method (a variant of the lasso) which was selected to solve the optimization problem (1.12). The modifications introduced to the method will be detailed, along with the different procedures used to set the different hyper-parameters.

1.4 Lasso-LARS: Lasso — Least-Angle Regression

Among all the shrinkage and subselection methods, the Lasso-LARS (Least Absolute Shrinkage and Selection Operator — Least-Angle Regression) was retained in this work. The method is very different from Lasso, but it has been shown that it produces exactly the same result as Lasso provided a variation applied to the LARS algorithm [40]. More details on the LARS and Lasso-LARS algorithms can be found in [23].

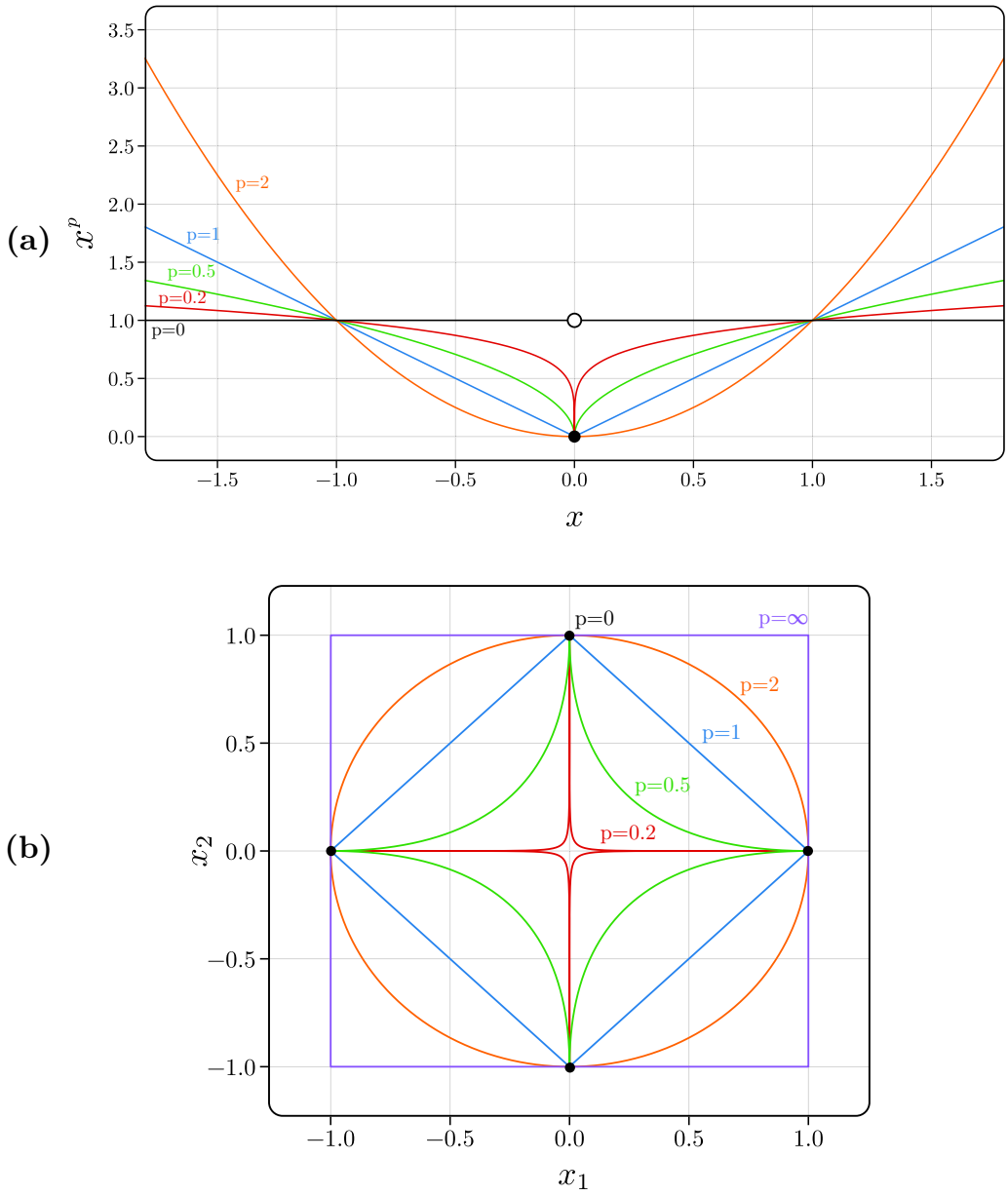


Figure 1.7: **(a)** Plot of x^p for various values of p. **(b)** Unitary circle for a range of p-norms (where $\left\| x \right\|_p = (|x_1|^p + |x_2|^p)^{\frac{1}{p}} = 1$).

The Lasso-LARS minimises the same cost function as the Lasso, i.e. a minimisation of the squared error with a weighted penalty applied to β using the l_1 norm. The corresponding cost function is as follows:

$$\beta = \underset{\beta}{\operatorname{argmin}} \left\| Y - \beta_0 - \phi\,\beta \right\|_2^2 + \alpha_f \left\| \beta \right\|_1. \tag{1.13}$$

Note that now the intercept term β_0 has been added. This term represents the value of the model's output $\hat{Y}$ when all modes are inactive i.e. $\beta = 0$. In this case, the value of β_0

is equal to the mean value of the Y signal. Note also the change of name of the weighting constant from α to α_f where the subscript f stands for final. More details will be given next.

We have opted for the Lasso-LARS method over all the existing shrinkage and subset-selection methods for several reasons. Firstly, because it is a **very efficient** algorithm for computing the Lasso solution especially when $\dim(\beta) \gg L$ which is our case[3] (as a reminder, L is the length of the signal of interest Y). Secondly, our choice was motivated more particularly for its inherent ability to compute the full solution path as α varies (when we refer to α we are referring to the value for which the optimization problem is solved during an intermediate iteration). Indeed, Lasso-LARS will yield all the solutions to the optimisation problem starting from an initial value of α_{init}:

$$\alpha_{init} = \frac{1}{L} \max\left(cov(\phi_{*,i}, Y)\right), \quad i \in \{1, \ldots, m(L+1)\}, \tag{1.14}$$

where $cov(\phi_{*,i}, Y)$ refers to the covariance between the covariate i ($i-th$ column vector of ϕ) and Y.

The value of α will decrease over the iterations, until reaching the value provided to the algorithm α_f as argument. Figure 1.8.a gives an example of how the value of α evolves over iterations. Typical evolution of the number of non-zero elements of β, namely $\|\beta\|_0$, can also be observed in Figure 1.8.b.

At each iteration, the Lasso-LARS generates one β solution to the optimization problem (1.13). This particular solution corresponds to a single α value for the iteration in question[4]. As shown in Figure 1.8, during the first iteration, the first solution produced corresponds to the value of α_{init}. As iterations progress, Lasso-LARS mobilizes more activation input and the β solution produced gradually becomes less sparse. At the final iteration, once the algorithm has mobilized sufficient activation input, the value α_f will be reached. The total number of iterations denoted n, required to reach this final value varies for each distinct signal of interest (this issue will be further elaborated in section 1.4.2). As each iteration produces one β solution, a total of n solutions to the optimisation problem are generated, characterized by gradually decreasing levels of sparsity.

To briefly explain how lasso-lars works, the active set is defined as the set of covariates that are active and contribute to the model output $\hat{Y}$, and β_i is a scalar that defines the contribution value of each covariate i. Algorithm 1 describes step by step how the algorithm works (taken from [40], where further details can be found).

[3] In our case $\dim(\beta) = m(L+1)$ and generally $m > 10$ which verifies the condition, however, even if the condition is not verified, the algorithm will have the same performance, except that the gain in calculation time will not be significant.

[4] The value of the current α depends on the correlation between the residual $r = Y - \hat{Y}$ and on the ϕ co-variates which have not been mobilized.

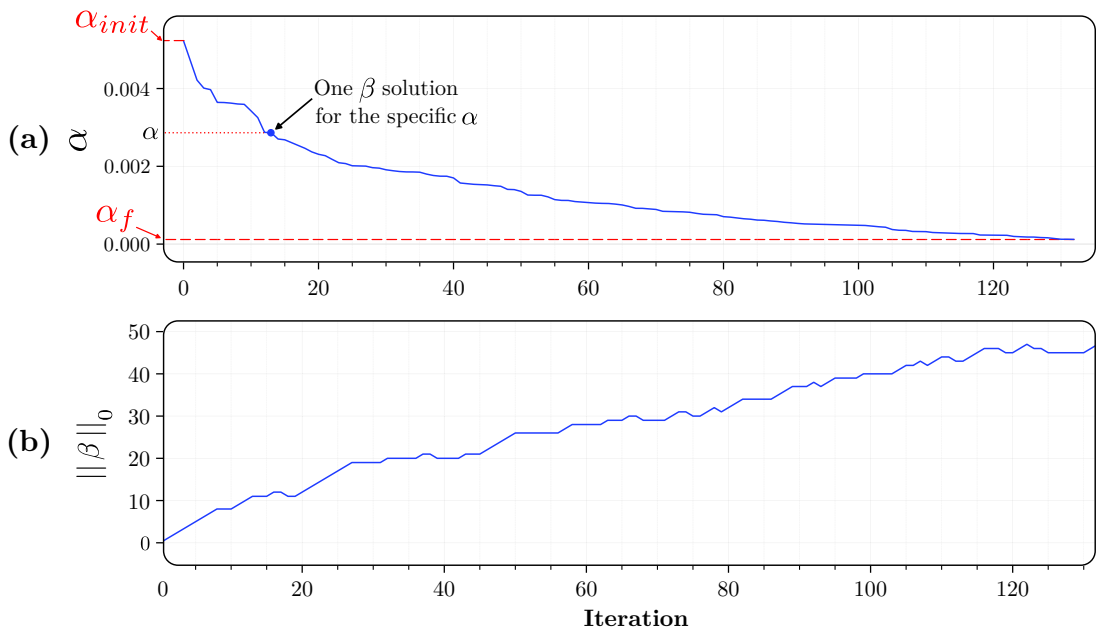


Figure 1.8: Example of the evolution of the value of α and $\|\beta\|_0$ over the iterations of the Lasso-Lars.

Algorithm 1 Least Angle Regression: Lasso variation.

1: **Initialization:** Initialize the residual $r = Y - \hat{Y}$ and $\beta_i = 0, i \in \{1, \ldots, m(L+1)\}$.
2: **Compute Correlations:** Find the predictor $\phi_{*,j}$ the most correlated with r.
3: **Update Coefficients:** Move the value of β_j from 0 to its least square coefficient $\langle \phi_{*,j}, r \rangle$, until some other feature $\phi_{*,k}$ has as much correlation with the current residual as does $\phi_{*,j}$.
4: Move $\phi_{*,j}$ and $\phi_{*,k}$ in the direction defined by their joint least squares coefficient of the current residual on $(\phi_{*,j}, \phi_{*,k})$, until some other competitor $\phi_{*,l}$ has as much correlation with the current residual r.
a: If a non-zero coefficient hits zero (a sign switch), drop its variable from the active set and recompute the current joint least squares direction.
5: **Stopping:** Continue until reaching the stopping criteria.

The algorithm will stop if one of the following stopping criteria is met:

- The number of iterations exceeds the maximum number of iterations allowed.
- All the predictors $\phi_{*,j}$ have been included in the active set, so there are no more predictors to add to the model.
- The final value of α_f specified to the algorithm has been reached: $\alpha \leq \alpha_f$ where $\alpha = max(cov(\phi_{*,i}, r)), \text{for } i \in \{1, \ldots, m(L+1)\}$.

The following section is dedicated to the visualization of the output of the method described above, namely, how to go from stacked numbers in a vector that is difficult to analyse to a display that is simpler to analyse even with the bare eye. The latter is

applicable for a single β solution chosen, but we are free to vary the level of parsimony of the solutions visualised.

1.4.1 Visualisation

As previously discussed, the generated SDFs depend on two parameters, namely, the excitation input U and the initial state of the set of oscillators x_0. The concatenation of these parameters defines the vector β that is to be computed:

$$\beta = \begin{pmatrix} x_0 \\ U \end{pmatrix},$$

> *The computation of β is performed considering both x_0 and U, but for the remainder of this manuscript, only the forced-regime response will be analyzed, i.e. U only.*

The remaining part of the SDF, i.e. x_0, is not considered for the analysis, as no conclusive results were obtained after studying it. Nevertheless, the presence of x_0 in the calculation remains essential, as it captures the free response, leaving the forced regime to U only. However, care must be taken as it may be that for some applications x_0 is more informative than U. By abuse of language, when we refer to Sparse Dynamical Features in the following, we will be referring to the U part of β.

As U contains the amplitude, frequency as well as the instant of activation of the modes involved, it is difficult to visualise it. As the approach is meant to be explainable, it is important for us to be able to plot the SDF. To do this, the U vector was rewritten in matrix form as follows:

$$\tilde{U} = \begin{bmatrix} | & | & \dots & | \\ u_1 & u_2 & \dots & u_{L-1} \\ | & | & \dots & | \end{bmatrix}, \tag{1.15}$$

$$= \begin{bmatrix} u_1(1) & u_1(2) & \dots & u_1(L-1) \\ u_2(1) & u_2(2) & \dots & u_2(L-1) \\ \vdots & \vdots & \ddots & \dots \\ u_m(1) & u_m(2) & \dots & u_m(L-1) \end{bmatrix}, \tag{1.16}$$

where $u_i(j)$ is the activation amplitude of the mode i at the time instant j.

For a better visual representation, the matrix $\tilde{U}$ can be considered as an image where each element $u_i(j)$ represents a pixel. The amplitude of modes activation will be represented by a colour. More precisely:

1. **Blue:** for negative amplitude.
2. **White:** for inactive mode.
3. **Red:** for positive amplitude.

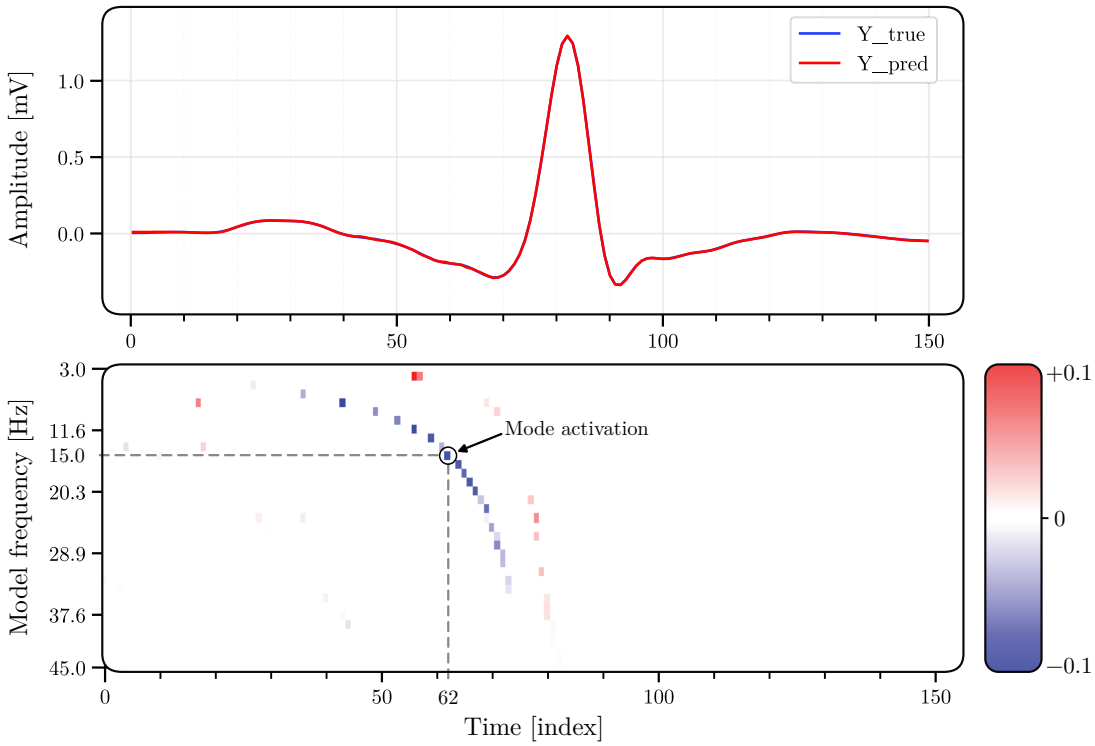


Figure 1.9: Visual representation as a bit-map of the modes' activity.

Figure 1.9 shows an example of a graphical representation applied on a heart (EEG) signal, where on the top plot, the real signal and the fit of our model are observed (almost identical for the chosen level of sparsity), while on the bottom plot, the image representation of the modes activation input inducing the predicted signal is shown.

In the lower part of the Figure, an example of how a single activation input of a single mode is illustrated. This example concerns the mode with a frequency of 15 [Hz], which was activated with an amplitude of -0.1 at the time instant indexed by 62.

We have already discussed the reasons why Lasso-LARS produces a total of n solutions to the optimisation problem with varying levels of sparsity, but in the next section, more details will be given about this point. We start with a discussion about why it is interesting to have solutions with a variable sparsity level, the practical problems encountered, and then the solutions proposed for these problems.

1.4.2 Sparsity level variation

The greater the value of α, the greater will be the product of $\alpha \| \beta \|_1$. As a result, the algorithm will tend to reduce the number of active excitation inputs as well as their amplitude (β will have to be sparse) but at the expense of a lesser accurate fit between the output of the model Σ and the target signal of interest. By contrast, the smaller the value of α, the plentiful β will be, and the model's output $\hat{Y}$ will fit perfectly the Y signal. Figure 1.10 illustrates this phenomenon very well, by showing an example where the squared error ϵ evolves for decreasing values of α can be observed.

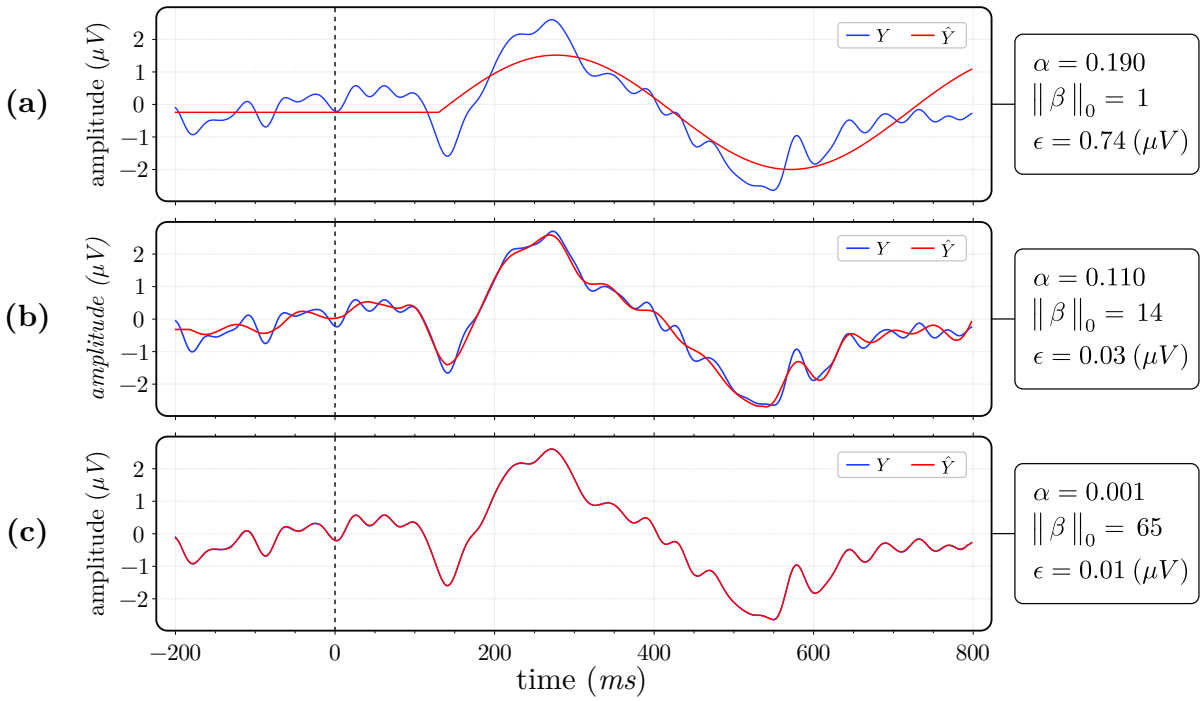


Figure 1.10: Example illustrating the effect of α values on the predicted signal $\hat{Y}$ and on the number of activation inputs mobilised by the model. **(a)** Shows a single activation of a single mode, selected with the appropriate frequency and excited at the right timing with the appropriate amplitude. **(b)** Utilisation of more excitation inputs and its effect on the $\hat{Y}$ prediction. **(c)** Shows the ability of our model to fit the Y signal by exciting more modes and more frequently. The number of distinct modes used in this example is 40.

To recall, having different solutions with varying levels of parsimony is necessary. Indeed, the best solution that yields the best result does not necessarily have to fit the signal Y perfectly as there is a trade-off between fitting the signal Y tightly and risking capturing the noise as well, and not fitting the signal Y tightly at the risk of not capturing the important information. This phenomenon is accentuated for the specific signal studied in the remainder of this manuscript, which is known to be very noisy. So a compromise has to be made, and this is made possible by the varying levels of sparsity of the many β solutions corresponding to many varying levels of parsimony.

As already mentioned, the value of α_{init} is fixed by the Lasso-LARS based on the correlation between the signal of interest Y and the co-variates of ϕ (see eq. (1.14)), thus, the only remaining choice is that of the α_f. The total number of iterations n required by Lasso-LARS to reach its stopping criteria depends on several factors, including factors related to (1) the signal of interest: Its complexity, length and frequency content. And (2) factors related to the model used: The frequency distribution of the modes, the number of modes used, and the w weighting (which we will describe later in Section 1.4.3). All these factors make it impossible to have the same number of iterations performed by Lasso-LARS to reach the same final value α_f for the different signals. Consequently, the total number of solutions n produced differs from one signal to another. This is not practical for the part that follows the generation of the SDFs, so for each signal, we would like to have the same number of solutions produced. In addition, the solutions should have an equivalent degree of fit so that they can be compared or used together[5].

To overcome this consistency problem, the sparsity interval $[\alpha_{init} — \alpha_f]$ for which the n solutions have been produced and that is specific to each signal, is therefore mapped to the new sparsity interval $[0 — 100]$ %. The most sparse solution is at the 0% level and the most full solution is at the 100% level, with the remaining solutions spread between these two values. Once all the solutions produced are brought into the same sparsity interval, the inconsistency problem of the total number of solutions n produced for each signal remains. The proposed approach to tackle this problem is to linearly slice the new sparsity interval with steps of l %, yielding thus, the same number of β solutions with decreasing level of parsimony taken at each interval. This transformation and slicing procedure is illustrated in Figure 1.11. In the case where the Lasso-LARS produces more elements than required ($n > 100/l$), the appropriate number of exceeding solutions is discarded. Contrary, in the case where the algorithms produce fewer elements than required ($n < 100/l$), the appropriate number of solutions is duplicated once, as depicted in the step 1 of Figure 1.11. For both scenarios, all solutions have an equal probability of being selected.

For the same scenario using the same signal, Lasso-LARS yields the solution of the optimisation problem 8 times faster than the classic Lasso. Furthermore, unlike the Lasso, which returns only a single solution, its Lasso-LARS variant yields several solutions, each corresponding to a given α.

We mentioned earlier that the only remaining parameter, which needs to be determined, is the value of α_f. This latter can be set by visual inspection[6] in such a way that over the entire interval $[\alpha_{init} — \alpha_f]$, a good variety of solutions is observed i.e. sparse at first and become less sparse until reaching a level of fit that is almost perfect, as shown in Figure 1.10. If the value of α_f is too large, the majority of solutions will be very sparse, and we won't obtain well-fitting solutions. Conversely, when α_f is too small, the majority of

[5] It seems cautious to avoid comparing SDFs resulting from a partial fit to SDFs resulting from another signal with a complete fit.

[6] As you might guess, this is not the option adopted in the sequel!

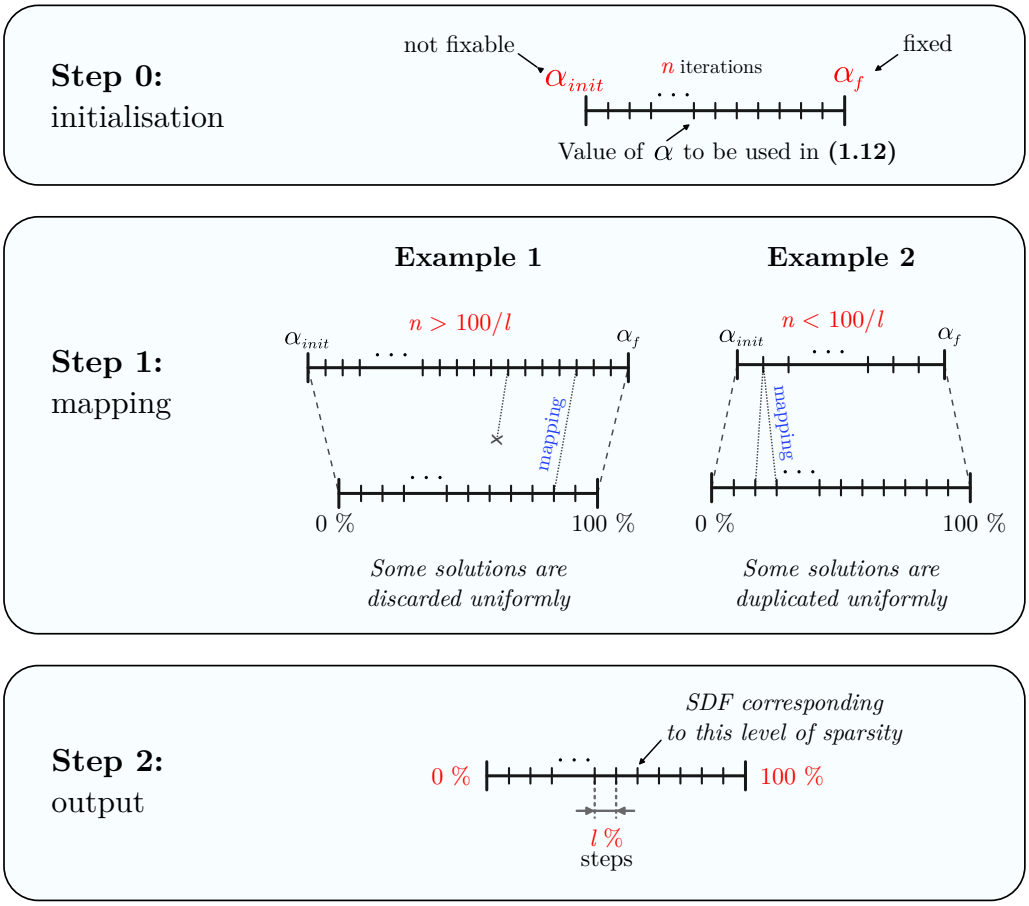


Figure 1.11: Mapping solutions from $[\alpha_{init} — \alpha_f]$ interval to $[0 — 100]$ % interval.

solutions won't be sparse at all. This behaviour is due to how the Lasso-LARS works in combination with the mapping discussed earlier in Figure 1.11.

To avoid manually selecting the value of α_f, which could introduce some bias and human subjectivity and even more obviously because it is not scalable regarding the number of profiles one needs to inspect in order to derive a rational and robust enough value of this important parameter, the selection of the α_f value has been automated. More precisely, a set of signals is selected randomly, and then, for each signal, the value of α that results in a fitting error of 1% is saved. This value is averaged across all signals to obtain an α_f value that, based on experiments, is consistent with the initial strategy. The fitting error is defined as the ratio between the energy of the residual and the energy of the initial signal:

$$error_{\%}(\alpha) = 100 \times \frac{\|Y - \hat{Y}(\alpha)\|_2^2}{\|Y\|^2}, \tag{1.17}$$

Note that this step must be repeated whenever the parameters defining the model are changed. The parameters we are referring to are: the number of modes m, their

distribution, and the weighting constant w (that will be discussed in the next section). Moreover, if for a given application there are multiple channels with a very different signal complexity over the channels, it is up to the expert to decide whether it is needed to have a different value of α_f for each channel, thus, requiring to redo this procedure for each channel. The number of times the procedure in question has to be repeated can increase rapidly, fortunately, almost all the parameters discussed above are set according to strategies and procedures we have established. Furthermore, the method is very lightweight and does not require a massive computational load.

1.4.3 Free response / forced response trade-off

The Lasso-LARS algorithm heavily involves the computation of correlations, so if the input features (columns of ϕ) are standardised, the algorithm will run faster as inner products might be used in the algorithm. The standardisation serves also to mitigate the effect of the differing magnitude of the input features on the correlations, therefore the entries of β have an equal chance of being selected a priori.

Instead of standardising the ϕ matrix in the conventional way, a slight modification is proposed in this work. More precisely, let's denote $\underline{\phi_1}$ and $\underline{\phi_2}$ as the vector-wise standardised version of the matrices ϕ_1 and ϕ_2 respectively (defined in eq. 1.8). We have introduced a weighting constant scalar $w \in (0, 1)$ so that the finally used standardised $\underline{\phi}$ matrix is defined as follows:

$$\underline{\phi} = \begin{pmatrix} w\,\underline{\phi_1} & (1-w)\,\underline{\phi_2} \end{pmatrix} \in \mathbb{R}^{L \times m(L+1)}. \tag{1.18}$$

This trick is used in order to manage the trade-off between a priori likelihood on one hand and favouring/guiding the choice of modes that are used first (either those of x_0 or U) on the other hand. Therefore, the factor w serves to adjust whether we put more emphasis on the free or the forced response. With the value of w closer to 1, the resulting SDF will be placed predominantly in x_0, contrary to a value of w closer to 0 that will result in SDFs placed predominantly in U. Therefore, this parameter serves to balance between the number of entries that will be dispatched either in x_0 or in U. This variation was intended in the needed case to favour adjusting for the x_0 features first since they affect the whole prediction unlike the U which are active only on a specific small time span (see Figure 1.12). Moreover, the modification helps on stabilising the algorithm and yields better results.

From the equation (1.8), we can observe that the predicted signal $\hat{Y}$ can be decomposed into two parts: (1) a free response $\hat{Y}_{x_0}$ and (2) a forced response $\hat{Y}_U$. In Figure 1.12 we can visualize how a signal Y is decomposed into a free response and a forced response that started at time 0 (s).

$$\hat{Y} = \underbrace{\phi_1\, x_0}_{\hat{Y}_{x_0}} + \underbrace{\phi_2\, U}_{\hat{Y}_U}. \tag{1.19}$$

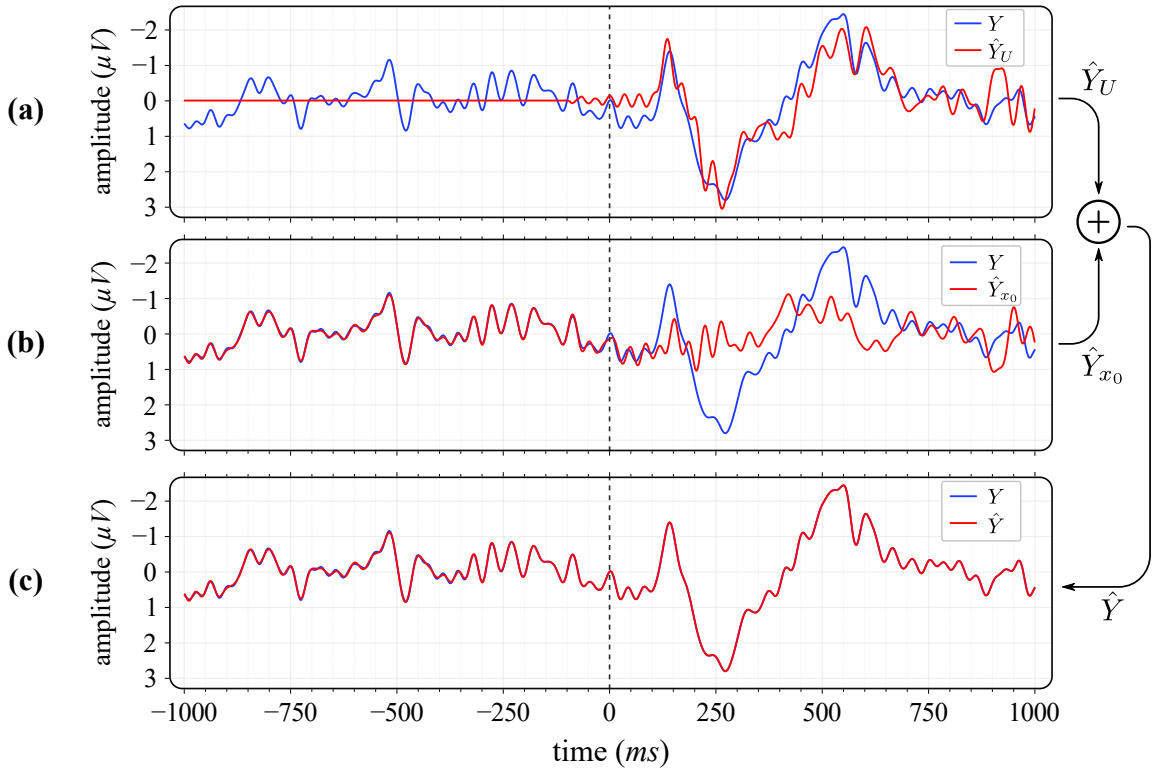


Figure 1.12: Example illustrating how a signal of interest Y (in blue) is fitted by the model (1.2) (in red). **(a)** Estimated contribution of the forced response alone $\hat{Y}_U$. As shown in the sub-figure, the modes are triggered near the forced regime onset and the shape of this response matches the shape of the signal Y. **(b)** Estimated free response $\hat{Y}_{x_0}$, the oscillations match the Y response prior to excitation arrival and keep oscillating in a similar way after the arrival of the excitation. **(c)** The contribution of (a) and (b) are summed to give the fitted signal $\hat{Y}$.

The value of the parameter w was set in such a way that the trade-off between placing the SDFs in the forced response or in the free response is consistent. More precisely, the parameter value is manually tuned so that the forced response starts to kick after the arrival of the forced regime, as shown in Figure 1.12. Experimentally, we have observed that values ranging from $w = 0.4$ to $w = 0.6$ are sufficient. For the example shown in Figure 1.12, a value of $w = 0.55$ is used. It should be noted that this tuning is performed for the lowest level of parsimony[7] and it is carried out only once for all the signals as they often have the same shape so the same value remains consistent for all the signals. Moreover, the variability of the solutions with regard to this parameter is minimal (a value of 0.52 will have almost the same results as a value of 0.5).

In cases where there is low free response activity (low background activity), as in Figure 1.13, a value of $w = 0.5$ remains reasonably consistent, as the model will know how to handle this point since it perceives that there are no modes that are active through-

[7] Nearly perfect fit, when $\alpha = \alpha_f$.

out the entire signal and will therefore activate only the appropriate modes (forced regime).

This parameter needs to be set only once for each application. If the application is changed, the relationship between the forced and free regime is no longer the same, requiring this step to be repeated.

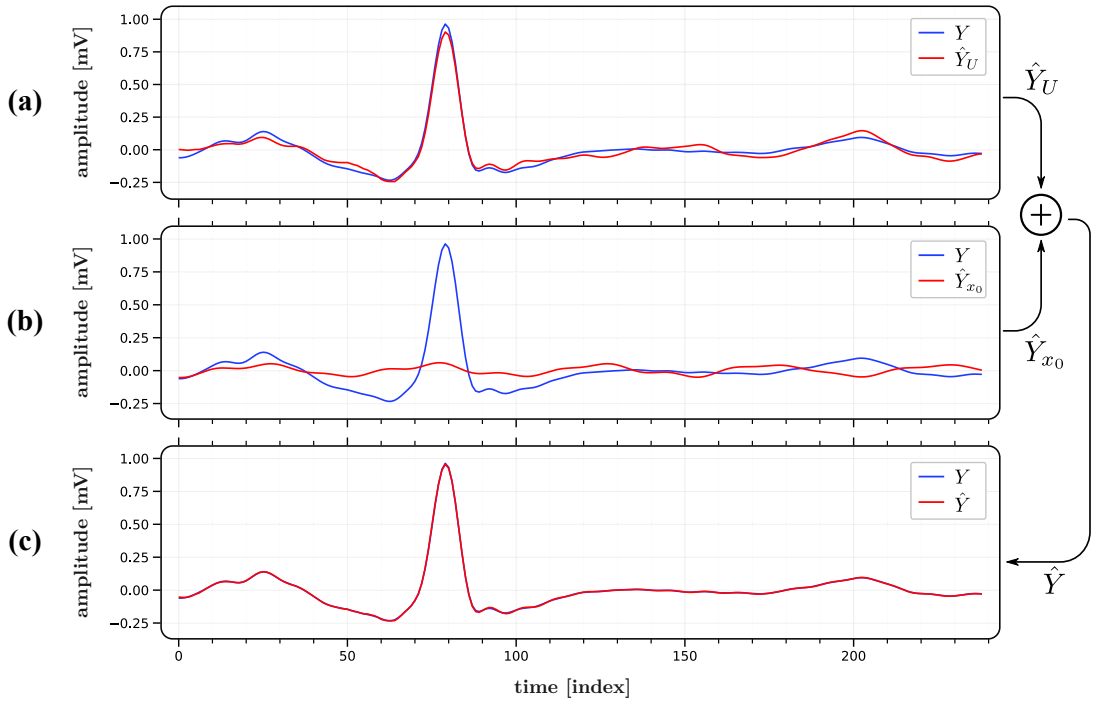


Figure 1.13: Example of the decomposition of a signal with low activity in the free regime. We can observe in **(b)** the low activity of the model in the free response contribution, which is consistent with the signal.

1.4.4 Fit completion

The Lasso-LARS algorithm is mainly used to induce sparsity in the solution of β. The main objective of the algorithm is to minimize the cost function (1.13) and to find its estimated optimal solution. As the Lasso-LARS algorithm is shaped, the final resulting β is not completely fitted, i.e. to minimise the cost function (1.13) β must be small in magnitude so that $\| \beta \|_1$ is small. As stated previously, we are only interested in finding the few most important entries of β that are used in fitting as much as possible the signal Y.
To pursue the initial design objective of the method, a new module has been added and is applied after the SDFs generation, and this for each level of parsimony. This block is used for fit completion and is divided into 3 steps:

1. **Masking:** consists of selecting only the elements that have been activated (non-zero-entries of β). Once selected, the column vectors corresponding to these elements will be kept, and the remaining regressors will be removed, thus producing the $\underline{\tilde{\phi}}$ matrix.

2. **Complete fit:** Once the regressors that have been activated are selected, a traditional least squares regression is applied using these regressors. This reduces the estimation error ϵ without affecting the number of active modes and their frequency of use.

3. **Update:** Once the new values have been computed, the previous $\hat{\beta}$ values will be overwritten by the new ones, ensuring that the new value corresponds to its previous location.

These 3 steps are depicted in Figure 1.14 for illustrative purposes.

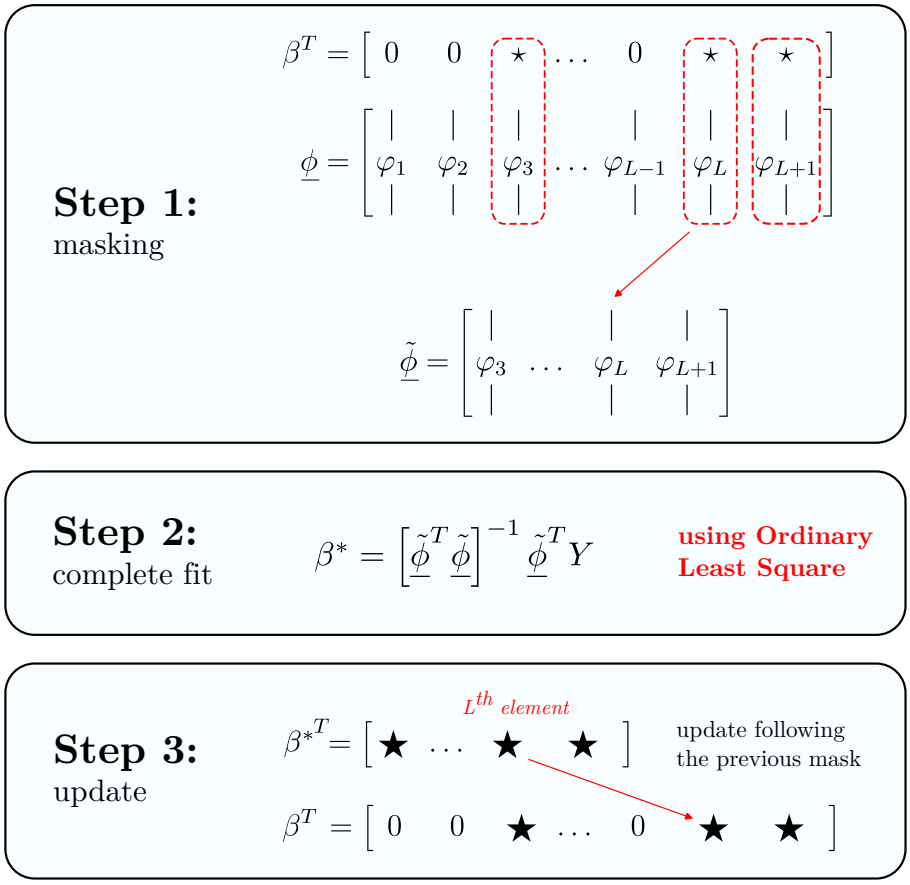


Figure 1.14: Fit completion for each SDF generated by the Lasso-LARS.

Figure 1.15 shows an example of the use of the method with and without the fit completion block. We can notice the use of exactly the same modes and excitation timing, we can obtain a better fit without affecting the sparsity of β.

1.4.5 Modes number and distribution

This section discusses how to determine the number of modes to be included in the system and their distribution over the frequency range.

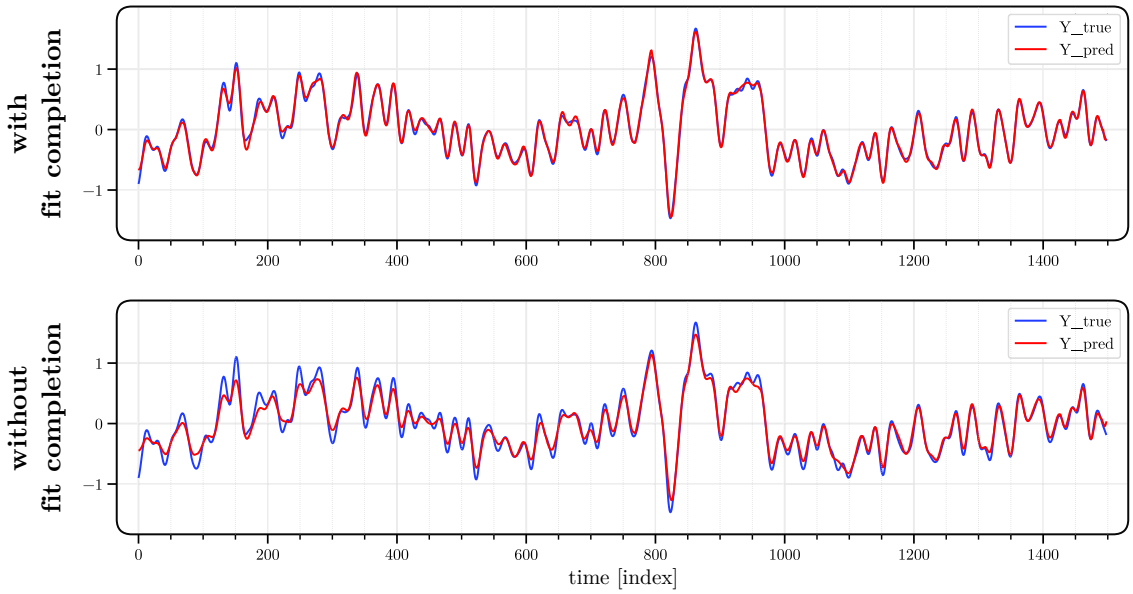


Figure 1.15: Comparison between the methodology fit with and without the completion fit block.

Number of modes used by the model

As mentioned earlier, the number m of modes involved in the model remains to be determined. This investigation can be carried out in a conventional manner, i.e., by considering the parameter m as a hyper-parameter to be optimized during the classification. However, this approach is computationally intensive and moreover, it is challenging to implement due to the significant increase in the number of generated SDF that leads to storage and memory issues. As our approach is based on explainability, we leaned more towards strategies where we can explain the reasons behind the choices we made, thus making it easier for the reader to re-implement and adjust to their particular problem.

The approach employed is to find the number of modes, denoted as m, which allows the signal to be decomposed with the same degree of fit while using the lowest number of active excitation inputs. Adhering to this criterion upholds the principle of parsimony upon which the entire methodology is founded. To accomplish this, a grid with different values of m is considered, and for each value on the grid, the following process is pursued: (1) Select a sufficiently large[8] subset of signals (identical for all grid values) to ensure significance. (2) Determine the value of α_f for the given fixed value of m that yields a fitting error of 1%. Generate the SDFs for the chosen signals and compute the average number of active excitation inputs used for each signal. The averaging is carried out for all signals and across the levels of parsimony axis; in other words, if, for a signal resulting in 50 SDF solutions with varying degrees of sparsity, the displayed average excitation count represents the mean of excitation counts across all 50 solutions. Figure 1.16 depicts the outcome of this analysis for the following grid: $\{1, 5, 10, 15, \ldots, 45\}$.

[8] The appropriate sample size depends on the application and the richness of the selected data. In general, we need to ensure that the statistics have converged, and usually 200 samples is good number.

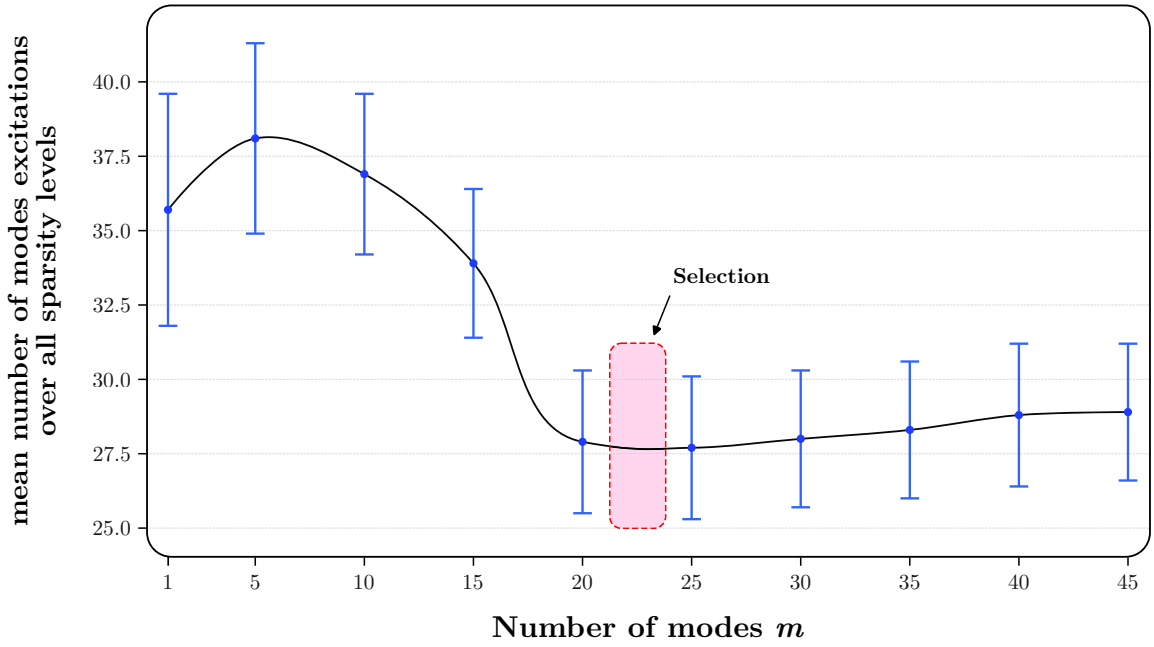


Figure 1.16: Evolution of the excitation number as a function of the number of modes m considered in the model. The same degree of fit is required for all the values.

We can observe from the figure that the appropriate number of modes denoted as m, lies within the range of 21 to 24. This range not only adheres to the principle of parsimony by minimizing the required excitation but also represents the smallest number of modes. It is more advantageous to take the smallest number of modes as this considerably reduces the complexity of the problem (see matrix size in eq.(1.8)). The exact value of the parameter m is not very critical, but its order of magnitude is.

The strategy proposed is not specific to the considered applications. Nevertheless, moving forward, it is essential to take into account the context of the application, as diverse choices can be made depending on the specific application's requirements.

Modes' frequency distribution

The only condition imposed on the m modes was that their angular frequency $\{\omega_1, \omega_2, \ldots, \omega_m\}$ should be strictly increasing. In this sub-section, two different ways of distributing these angular frequencies over the entire frequency spectrum of our signals will be considered. The first is a uniform distribution, and the second is a distribution weighted by the spectral energy, both will be discussed in the following:

1. **Linear distribution:** The simplest method, the frequencies of the m modes are linearly spaced to cover the entire frequency spectrum of the signal.

2. **Spectral Energy Weighted distribution:** To understand how this part was operated, we will first look at Figure 1.17 which shows an example of the Power

Spectral Density (PSD) average over multiple signals with a confidence interval (CI) of 95%.

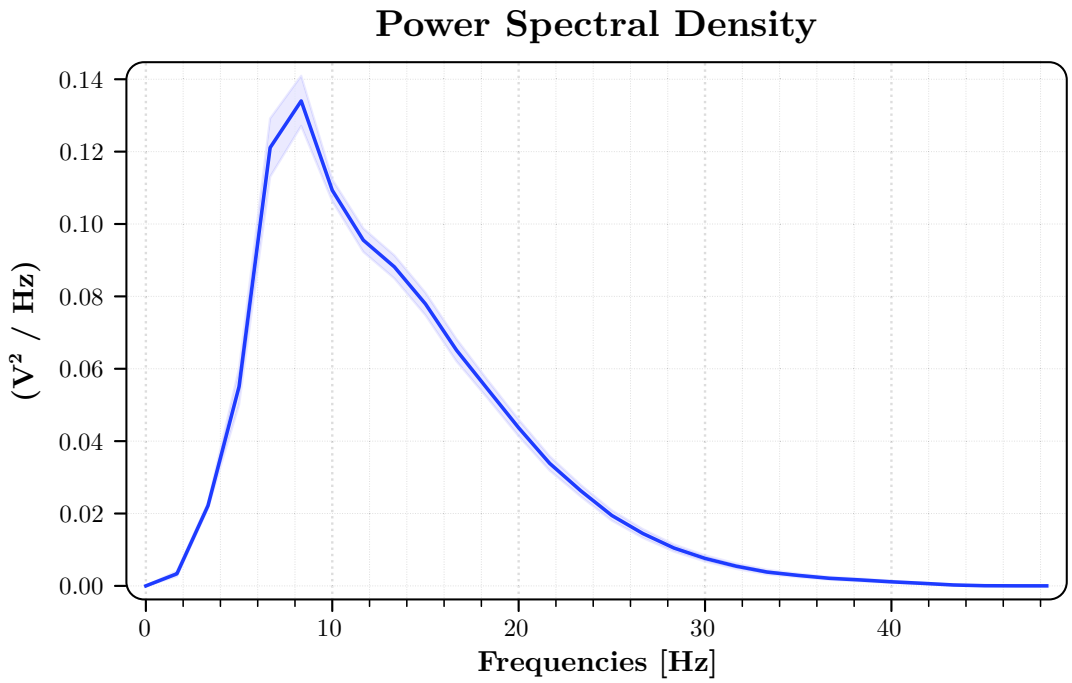


Figure 1.17: Example of average Power Spectral Density of multiple signals.

The more *energetic* an interval is, the more modes will be placed over this interval. To achieve this, the power spectral density is normalized so that it can be considered as a Probability Density Function (PDF). Next, the cumulative density function is computed, giving the result shown in Figure 1.18.

Next, the interval $[0 — 1]$ is linearly partitioned into m equal parts. We need to determine the inverse function for each ordinate value. In other words, we horizontally project the value onto the cumulative density function (CDF), and then the obtained value is vertically projected onto the x-axis. The resulting frequency represents the frequency of that mode (as indicated in red in Figure 1.18). By repeating this process for the m ordinate values, we establish the distribution of frequencies for these modes, thus allocating them according to the spectral energy.

To compare the outcome of the two distributions, the same procedure as for fixing the number of modes m has been replicated. This involves testing both approaches and determining which one complies more with the principle of parsimony. The outcomes of this analysis are illustrated in Figure 1.19. An example can be observed[9] for $m = 40$, showcasing a comparison between the number of excitations generated for the same configuration. It can be observed that the linearly spaced frequency distribution produces,

[9] We conducted similar tests for other values of m, and the same result was observed.

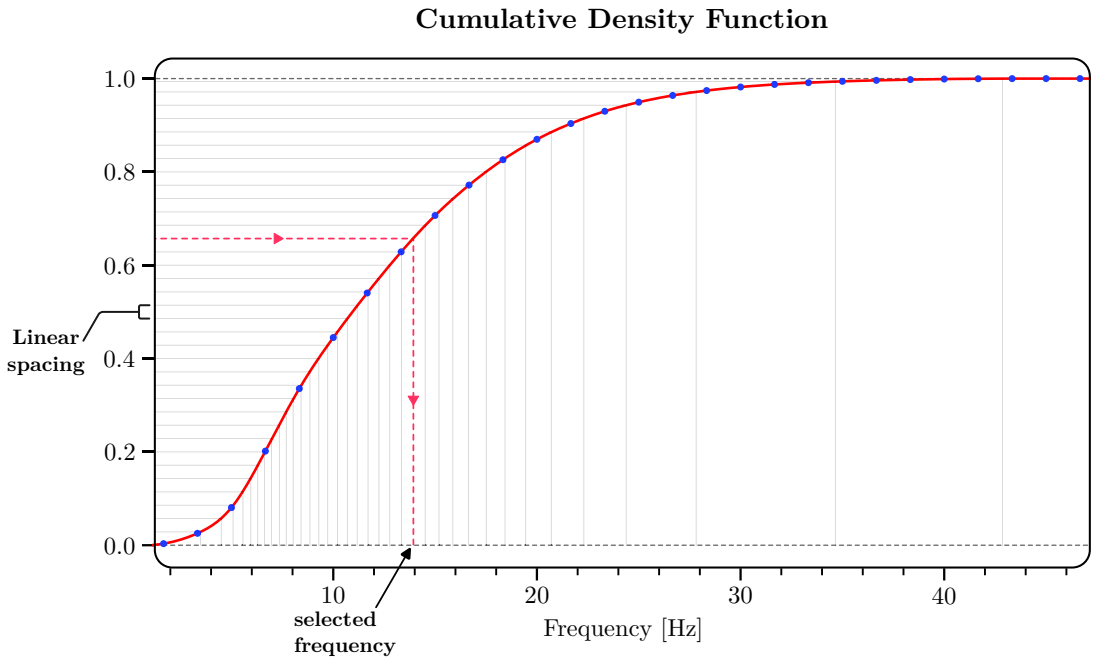


Figure 1.18: Cumulative Density Function of the PSD illustrated above. Illustrates how the frequency distribution of the modes is related to the PSD.

for the same configuration, the same outcome while utilizing only half the number of excitations. The latter is the one that best complies with the principle of parsimony.

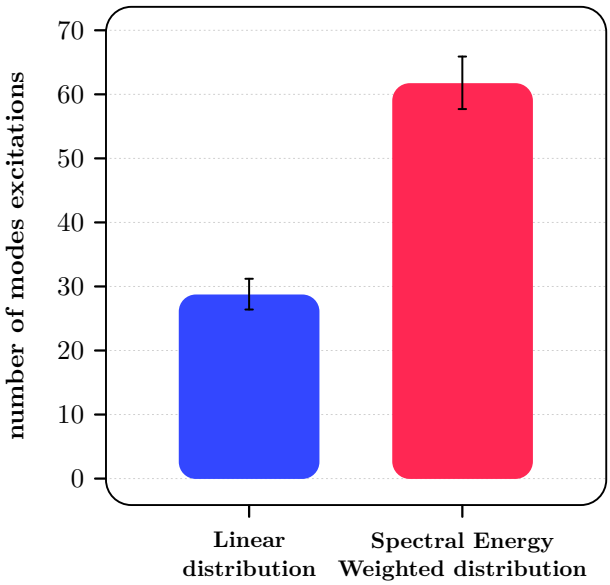


Figure 1.19: Comparison of the number of modes excitations between a linear distribution and the spectral energy weighted distribution of the model frequencies.

From the previous analysis, we concluded that the most important factor is that the

frequency distribution of the model should cover the entire frequency spectrum as best as possible. If an interval is more energetic, the modes will be activated much more often or have a greater amplitude. To illustrate this point, Figure 1.20 shows an example with a linear frequency distribution, and we can observe the correlation between the PSD and the calculation of the energy of U carried by each frequency band[10]. The distribution of the two PSDs the signal and the model's output are quite similar. This clearly shows that the model's philosophy respects both the frequency aspect and the physical essence of the signal as intended. It is essential to notice the slight discrepancies between the PSDs of the signals (depicted in red) and those of our model (depicted in blue). We have no explanation for this, but we believe it could be explained by the lack of consideration of the power captured by x_0.

1.5 Features extraction

Feature extraction is a fundamental process in machine learning and signal processing, where raw data is transformed into a more compact and meaningful representation, known as features. These features are selected to capture essential patterns, characteristics, or relevant information from the original data without being redundant. Once the features have been extracted, the model relies on them to make its predictions.

1.5.1 High dimensionality drawbacks

Before discussing feature extraction, it is very important to note that the more features we have, the more data will be needed to converge to the true distribution underlying the data. When the dimension of the features increases, the volume of the space containing the latter grows exponentially, so that the available data will be scattered in that space. In order to have reliable results, the data must cover the space well enough, so the number of data must also grow exponentially [8]. This phenomenon is known as the *curse of dimensionality*.

Suppose we have a set of discriminant features and we want to study, for example, the difference between Parkinson's disease and healthy individuals for this particular set of features. For practical reasons, in our study, we cannot take the entire worldwide population suffering from the disease and compare it with the entire worldwide healthy population. It is necessary to take a representative sample of both populations. For the global population, and for the features chosen, we have one data distribution for each condition (healthy or not), these distributions are called the true distribution underlying the data.

Taking a representative sample from the entire population is equivalent to drawing a sample from this distribution. In the case where a large number of features are used and

[10] Which is also a PSD but not as we know it classically.

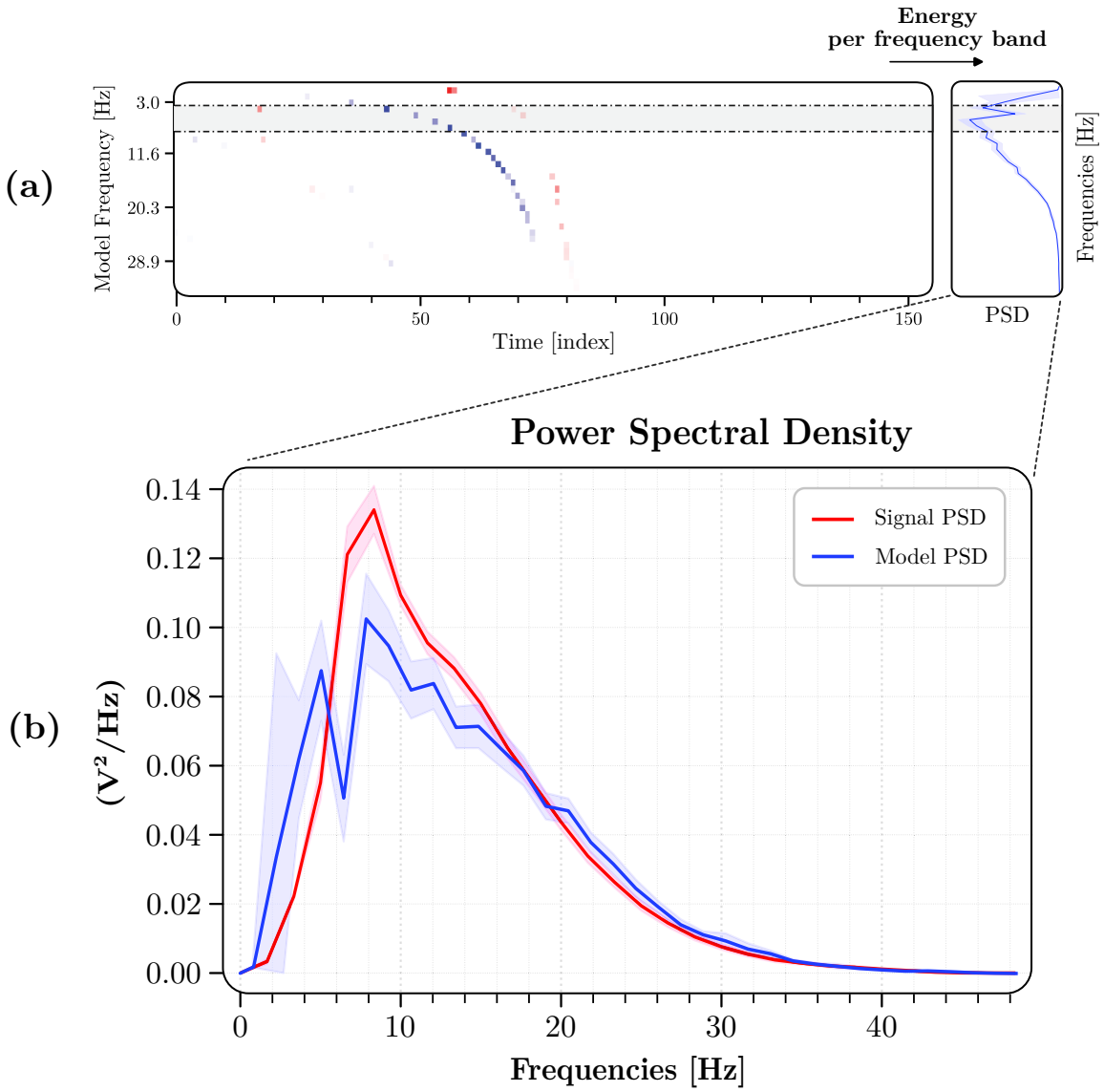


Figure 1.20: **(a)** Shows how the PSD of our model is calculated. **(b)** Comparison between the Power Spectral Density of the signals used and that produced by our model using (a).

a small sample is drawn, the space in which the data resides is very large (high dimension) and the samples drawn end up scattered in that space. Given that most classifiers tend to maximise the separation between the two classes, it is easier to obtain very good results when there are many features and a few number of data samples. These good results can also be obtained if we simply had a good draw[11] during our sampling [30]. More alarming is the fact that these good results can also be obtained if the true distribution (worldwide) of the two classes is exactly the same (overlapped) and therefore theoretically

[11] By pure chance, it is possible to have the majority of the samples from the first class in one area, and those from the other class in another far away area, the likelihood of this happening increases with the number of features used.

not separable. Non-separability is equivalent to having a separation accuracy of 50% for the case of 2 classes.

This shows the importance of studying the statistical significance of the results obtained. These considerations should be taken into account prior to the experiment in what is so-called "experiment design" where statistical methods and calculations are used to determine an appropriate sample size to ensure the validity and reliability of the experimental results. We are aware that it is not always straightforward to put this into practice, throughout this discussion, we want to raise the alarm about the problem where the number of features used is not adequate given the number of subjects in the experiment. In what follows, the problem we have just mentioned will be illustrated using a simplified example.

In the following example, the goal is to separate and categorize data from two classes: "x" and "o". It is assumed that these two classes share the same underlying distribution (overlapped), therefore, if somehow we have access to this information, we know that the two classes are not separable. Figure 1.21.a is a discrete example of the same underlying distribution where if we draw an item in class 1, the probability of drawing it from the left or the right is the same, similarly for class 2. For simplicity reasons, at first, only 4 elements will be drawn (2 for each class). Two possible drawing are illustrated in Figure 1.21.b.

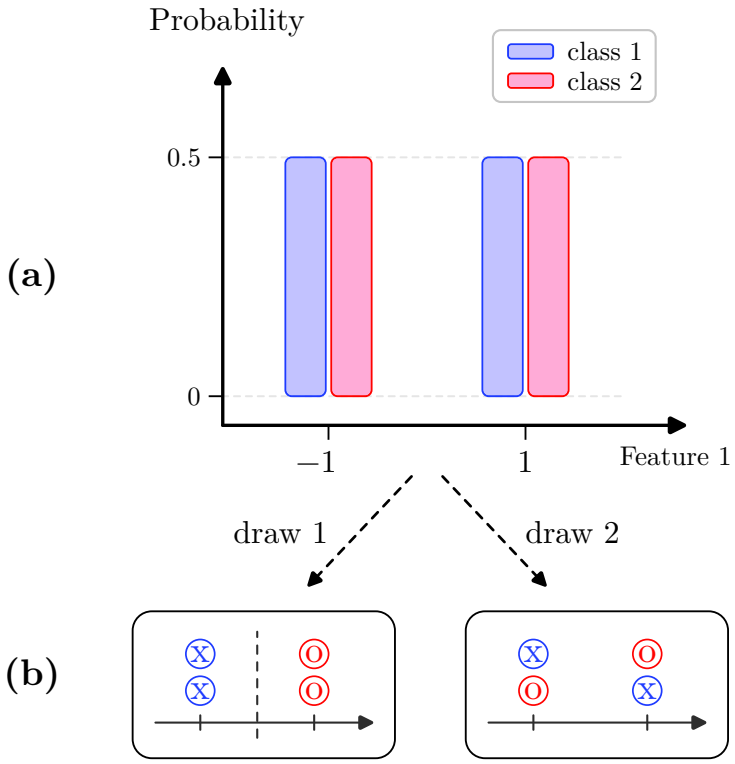


Figure 1.21: **(a)** Example of a discrete overlapping true distribution. **(b)** Two possible drawings taken from the distribution (a).

It can be observed that for the left-hand draw, the data samples are separable even though we know that the underlying distributions that generated this draw are not separable. It is simply by chance that we have obtained perfect separability of the two classes. The likelihood of this happening decreases with the number of samples taken, and the more points we sample, the more we converge on the true distribution.

By drawing only 4 elements, there are 16 possible combinations, for lack of space, we have illustrated the most important draws[12] in Figure 1.22. At the bottom, we have noted the number of occurrences of a similar draw as well as the classification accuracy obtained using a Linear Support Vector Machine (SVM). Note that the lowest precision is 50%, so in the worst scenario, we would say that the data are not separable. But it is possible to have a good draw by chance and have a precision of 100%. It is important to note that each occurrence is equally likely to be drawn. Moreover, the decision boundary of the SVM is the same for each draw (no randomness in the classifier that will lead to different results).

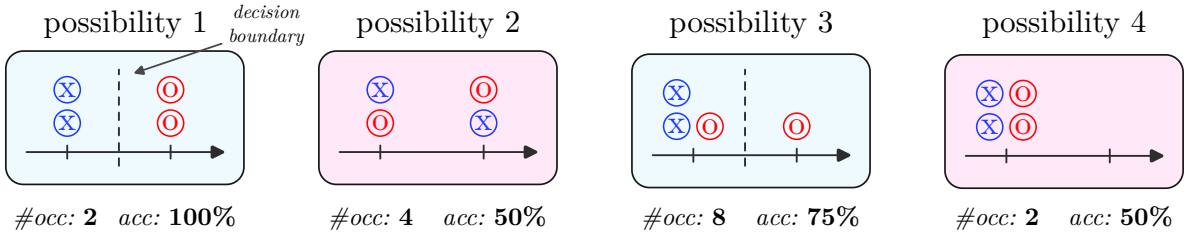


Figure 1.22: Important possible draws of 4 elements from the distribution in Figure 1.21.a with their number of occurrences and the classification accuracy.

For configuration G, which consists of 1 feature, 4 samples, a linear SVM classifier, and this specific true distribution underlying data, we will compute the a priori accuracy (R) associated with this configuration. This a priori accuracy is inherently linked to the configuration itself, and it assesses the expected accuracy that is yielded by the configuration before incorporating any evidence or making any sample draw (a prior). The expected accuracy can be calculated following:

$$\mathbb{E}(R;G) = \sum_i r_i \, p(r_i;G) = \frac{2}{16} \times 1 + \frac{8}{16} \times 0.75 + \frac{6}{16} \times 0.5 = 0.6875, \tag{1.20}$$

where r_i is the accuracy obtained for the possibility i and $p(r_i;G)$ is the probability of observing the possibility i under configuration G which in other words is the number of occurrences of possibility i.

In the following example, the same procedure will be used to calculate the expected accuracies for different configurations G (the classifier and the underlying distribution will be kept the same). The impact of adding more samples on the expected accuracy

[12] The remaining draws can be obtained by horizontal symmetry and class swapping.

will be studied first, followed by the impact of adding one feature at a time. Note that the added feature has the same marginal distribution as feature 1 shown in Figure 1.21.a. The obtained results are depicted in Figure 1.23.

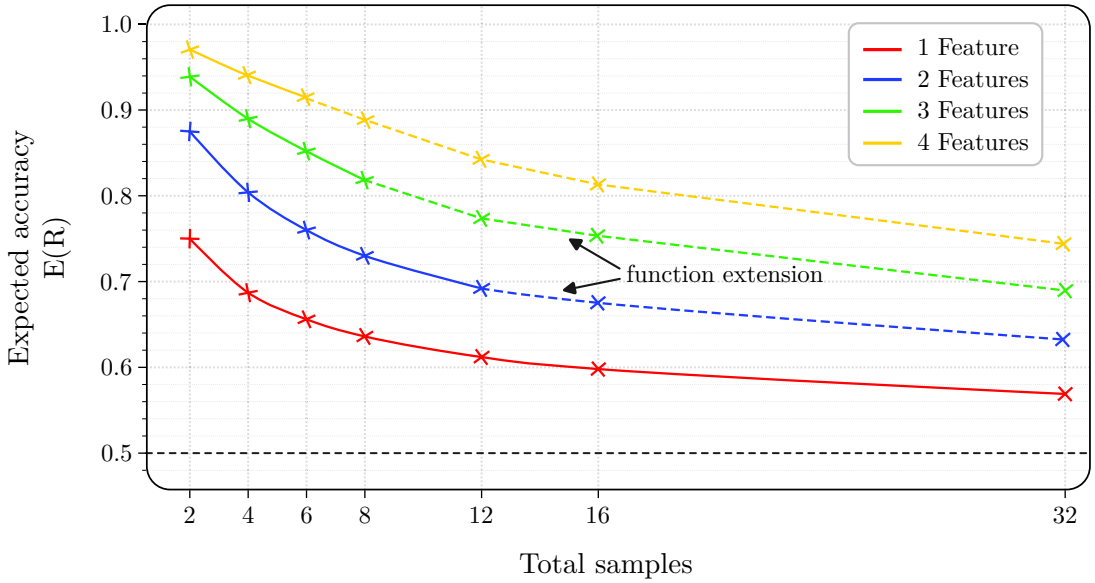


Figure 1.23: The expected accuracy as a function of the total number of samples for a single feature up to four. Computed points are in solid, their corresponding extension is dashed.

The curves were obtained by listing all the possible splits as shown in Figure 1.22 for an increasing number of samples. In the case of the red curve (1 feature) the expected accuracy can be computed according to the subsequent formula:

$$\mathbb{E}(R) = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{n} \left| k - \frac{n}{2} \right| + \frac{1}{2} \right), \tag{1.21}$$

where the even n is the number of samples taken in the data-set. We could not derive a formula for the other cases, thus, the first points were determined using a numerical computation and are represented by the solid line. The dashed curves are what we believe the extensions of the curves as we could not compute for more samples as the number of possibilities grows very fast.

Figure 1.23 gives an order of magnitude of the expected accuracy that can be obtained for a data-set which, if the true underlying distribution is known, is not separable. It can be observed that as the number of samples increases, the accuracy tends towards 50%, independently of the number of features taken. Furthermore, for the same number of samples, the expected accuracy increases as the number of features increases. If we take, for example, 16 samples for 4 features, the accuracy we can expect is of the order of 80% simply by being in this configuration, even if the underlying distribution is not separable.

This simplified experiment highlights two important points:

1. Importance of studying the significance of the results obtained: In some cases, such a study is difficult, however, we want to warn of the various biases and problems that can arise with data-based methods. Something that can help we believe is the use of explainable features, regularization, etc. as having 20 features for 50 data obviously leads to fragile results. This should be handled by either reducing the number of features used despite the loss of accuracy or by adding more quality data.

2. Importance of data-set sizes: The number of data samples used should be significantly higher than the number of features. A good rule of thumb given by Richard Bellman [8] is to have at least 5 data samples for each feature that covers well the space. The importance of data-set size is widespread, however, in the previous example, we wanted to give an idea of the expected accuracy, simply due to the randomness of the draws in the case where there is little data or a high number of features.

The considerations we have just discussed are often absent in the works we have encountered and that will be addressed in the following chapters. This aspect enhances the robustness and relevance of our results, especially considering that we will be utilizing an extremely limited set of features.

1.5.2 Advantages of feature extraction and selection

As our work is built up around the explainability and significance of the results, the problem of high dimensionality has been avoided by carrying out feature selection. This latter offers several advantages:

1. **Reducing the risk of over-fitting:** Classifiers with strong fit capability, will no longer be able to over-fit to data that is scattered in a high-dimensional space, but will be pushed to fit using a small number of features that contain relevant information. Feature selection helps to project the data into a lower-dimension feature space that should not be sparse. As will be discussed later on, over-fitting is intolerable for health applications for obvious reasons.

2. **Improving the stability** of the classifier: Upon adding new data samples, the decision boundary should remain stable and not vary a lot. Feature selection helps to stabilize the decision boundary by giving less degrees of freedom to the classifier. This effect is more relevant for applications based on small data-sets.

3. **Physical explainability:** Using explainable features, which have a direct relationship with the targeted application, not only helps to make sense of the model used but also to understand the model's decisions. In addition, the final model is more trustworthy as it is derived from prior knowledge.

4. **Extract valuable information:** Only the useful and important information that the model should utilise to perform the classification is extracted and well packed in these features. The classifier will no longer need to perform the feature extraction.

5. **Reducing computational time:** Using fewer features reduces the size of the problem to be solved, thus considerably reducing computing time. This is important even in the case of off-line training, as it gives access to methods for testing the significance of results which are often not feasible due to their heavy computational demands.

In our case, the SDF (the U vector) at the output of the algorithm has a size of $m(L-1)$, where L can be very large. So we often end up with SDFs that are certainly sparse but occupy a large dimension[13]. Given that the applications considered in this manuscript are limited in terms of data quantity, the advantages mentioned above are even more relevant.

1.5.3 Feature extraction and selection

For the sake of brevity, we have only referred to feature extraction earlier, although, feature extraction and feature selection will be discussed in greater detail. The difference between the two is that feature selection is a process that chooses a subset of features from the original features following a certain criterion. However, feature extraction is a process through which a set of new features is created based on the original signals or features.

The different types of feature extraction/selection performed on the generated SDFs can be split into two categories respectively based on temporal or frequency selection. This is described hereafter:

Temporal selection:

In some situations, it is preferable to rely on our prior knowledge and experience regarding the application. For this reason, it is important to tailor the time interval studied and select the features that have been activated over the time interval of interest only. Figure 1.24 is an example of a brain signal where it is physiologically and physically meaningful to consider only features taken over the indicated temporal interval of interest.

The temporal selection is directly performed on the vector U by discarding the elements of the vector that do not belong to the time interval of interest, as illustrated by Figure 1.25.

Frequency selection:

In other situations, some prior knowledge suggests the selection of a frequency band of interest. Similarly, this selection is made possible by the proposed method and is illustrated in Figure 1.26.

[13] Typically for the applications addressed $m(L-1) = 40000$, due to sparsity the SDF dimensions is lower and is in the order of 4000.

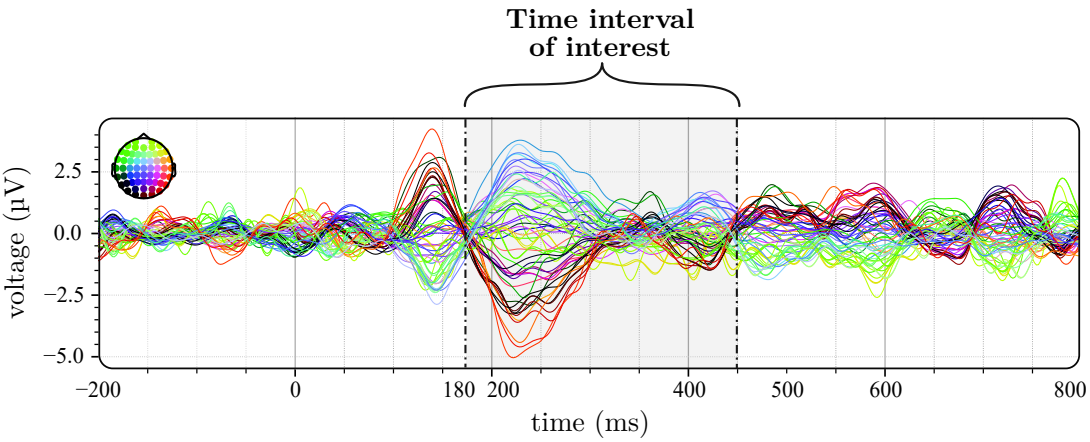


Figure 1.24: Example of brain signals with time interval of interest.

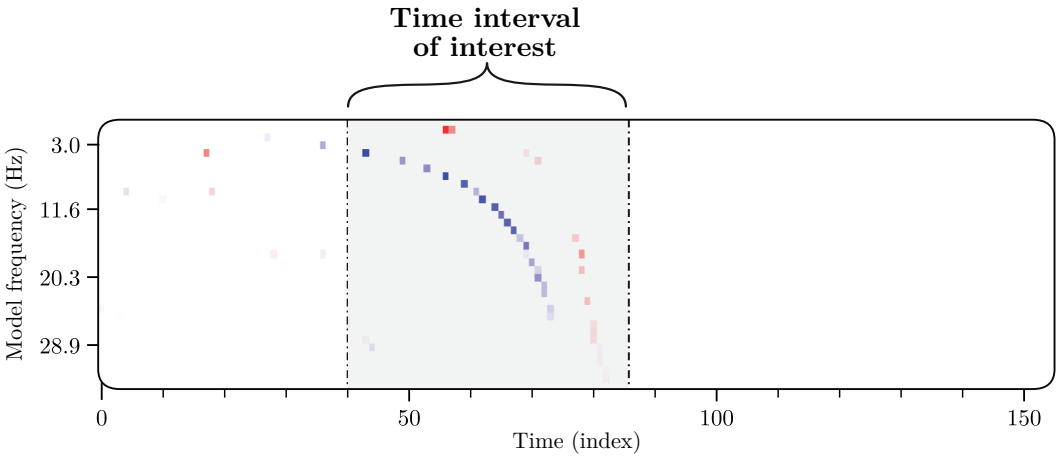


Figure 1.25: Temporal feature selection.

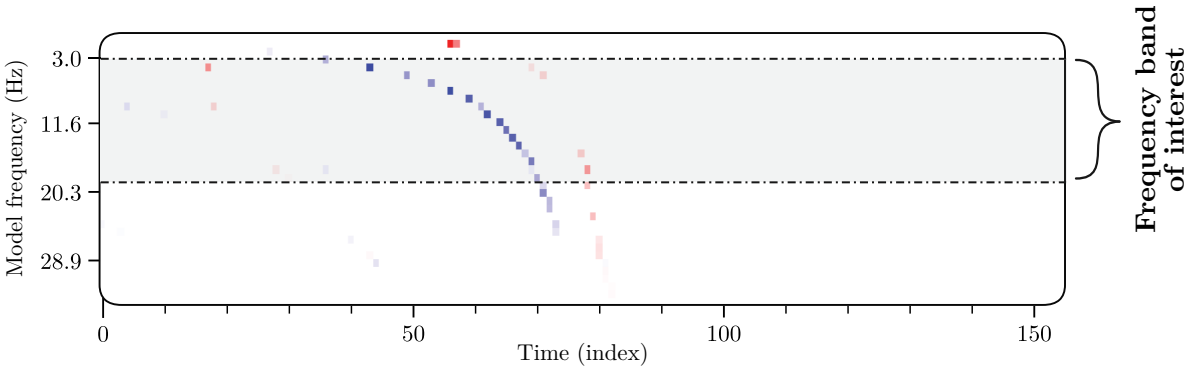


Figure 1.26: Frequency feature selection.

Feature extraction:

The methodology proposed has been used as a decomposition, a change in the point of view on the data. Instead of treating our data as a time series, we chose to consider them

as modes with their associated frequencies, which are activated and deactivated at given instants with a given amplitude. All signal information is therefore embedded in these SDF.

Just as a temporal signal can be fed directly to the classifier, so can the SDFs. Thus, the entire vector U can be fed directly to the classifier, or a selection applied to it (frequency and/or time). Although the feature selection is carried out on the SDFs, we will remain in the high-dimensional case. For the advantages presented in Section 1.5.2 and the high-dimensional issues discussed in Section 1.5.1, we are forced to perform feature extraction. The different combinations of feature selection and feature extraction that can be applied to the generated SDFs are shown in Figure 1.27. It is important to note that only one case will be used at the time and there is no aggregation of the different possibilities. Moreover, to recall, x_0 must be used for the computation, however, for the applications considered, no significant results were obtained using it, so it was discarded for the remainder of the analysis.

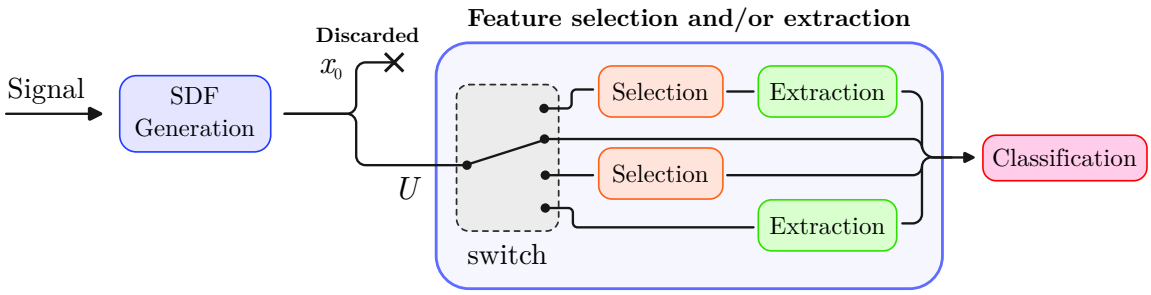


Figure 1.27: The different combinations of feature selection and/or extraction that we can apply on the SDFs.

Different features have been tried out that we thought would be interesting to consider in the applications considered in the following chapters. Table 1.1 summarizes all the features that have been used. It should be noted that, depending on the application, only a subset of these features has been evaluated, not all of them. The features that have been evaluated will be indicated for each application in its respective section.

Combination of the above:

Note that for this part, the different feature selection/extraction options can be used individually or in combination. For instance, a first selection of the frequency band of interest is performed, followed by a feature extraction applied to the resulting selected features. Thus, yielding the desired features extracted only from the frequency band of interest.

Once the desired features have been selected and extracted, they will be used by the next step, which is classification. This will be covered in greater detail in the next section.

Number	Feature	$f(U)$
1	Energy	$\|U\|_2^2$
2	Count non-zero	$\sum \mathbb{1}(U \neq 0)$
3	Mean	$\overline{U}$
4	Max	$max(U)$
5	Min	$min(U)$
6	Pk-pk	$max(U) - min(U)$
7	Argmin	$argmin(U)$
8	Argmax	$argmax(U)$
9	Argmax−Argmin	$argmax(U) - argmin(U)$
10	Variance	$\sigma^2 = \left[U - \overline{U}\right]^T \left[U - \overline{U}\right]$
11	Skewness	$skew(U)$
12	Kurtosis	$kurtosis(U)$
13	Count above mean	$\sum \mathbb{1}(U > \overline{U})$
14	Count below mean	$\sum \mathbb{1}(U < \overline{U})$
15	Norm l_1	$\|U\|_1$
16	Norm ∞	$\|U\|_\infty$
17	Positive argmax	$arg(\|U\|_\infty)$

Table 1.1: List of extracted features a subset of which is used in the forthcoming applications.

1.6 Classification & Evaluation

The aim of the classification is to use a model, also called "classifier", to categorize the input data into different classes. The classifier utilizes the data to learn patterns from the inputs in order to separate them and predict the output (the different classes). Concerning our case, the input data are either the SDF directly or the result of feature selection/extraction applied on the SDF. The output is the class to which the input data belongs to.

Different classification algorithms exist with different levels of complexity and use cases. The less complex models are often appreciated because of their simplicity of use, their low computational costs and especially for their explainability and the simple interpretation of the results. As we are aiming to use very few features, simple classifiers are sufficient for the task. Depending on the complexity of the application the following 3 classifiers were used:

1. **Linear Discriminant Analysis (LDA):** The objective is to find a linear combination of the features such that the between-class variance is maximized relative to the within-class variance. Figure 1.28.a shows that although the line with the greatest centroid spread, projected data overlap. However, the discriminant directions minimize the overlap of the projected data and maximize the between-class

variance as illustrated in Figure 1.28.b [40].

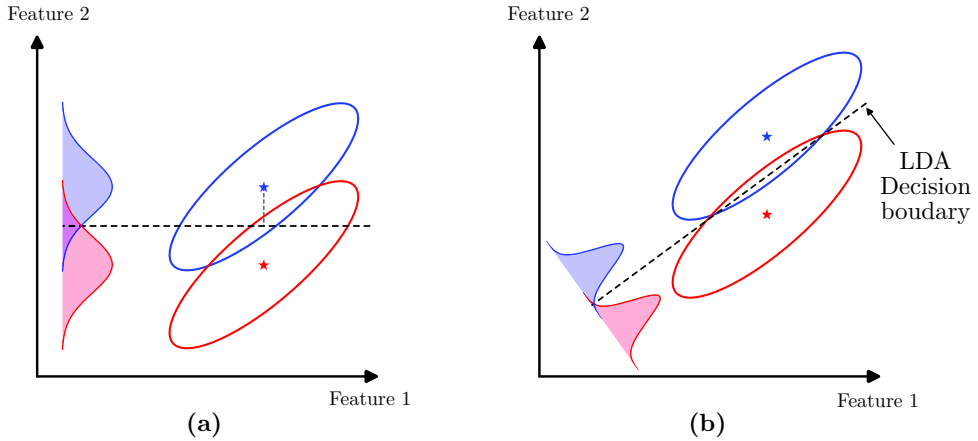


Figure 1.28: **(a)** Greatest centroid spread compared to **(b)** LDA separation.

2. **Quadratic discriminant Analysis (QDA):** This algorithm works exactly the same as the LDA, however, instead of having a hyperplane (linear combination of features), a quadratic curve separates the different classes (quadratic combination of features). QDA is often used when the data is more complex, and the linear classifier fails to separate the data. Figure 1.29 shows a comparison between LDA and QDA separation capabilities. [40].

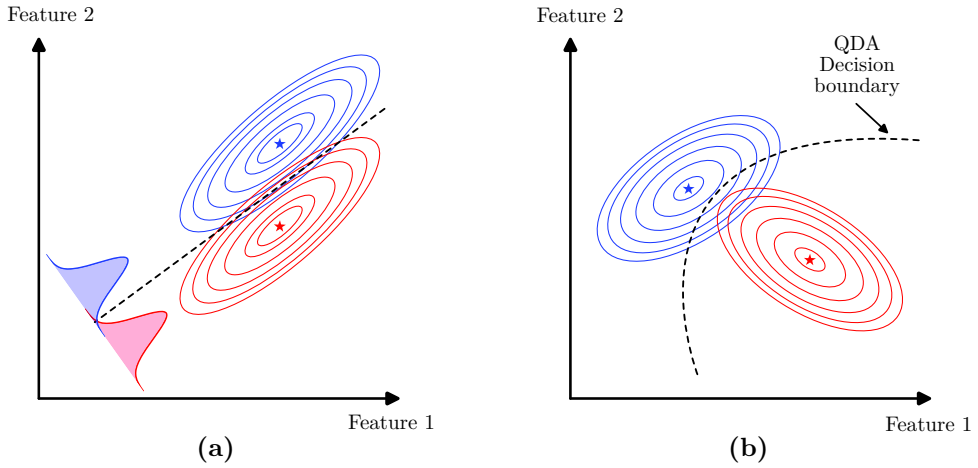


Figure 1.29: Comparison between decision boundary between: **(a)** LDA and **(b)** QDA.

3. **Adaptative Boosting (AdaBoost):** Is an ensemble learning algorithm designed to enhance the performance of weak classifiers by combining their predictions. It iteratively trains a sequence of weak models, such as simple decision trees [89], assigning higher weights to misclassified instances in each iteration as presented in Figure 1.30. The final model is formed by giving more influence to those weak models that perform well. AdaBoost adapts its training focus based on the errors of the

previous models, effectively creating a strong classifier that excels in generalization and is less prone to overfitting. More details can be found in [28].

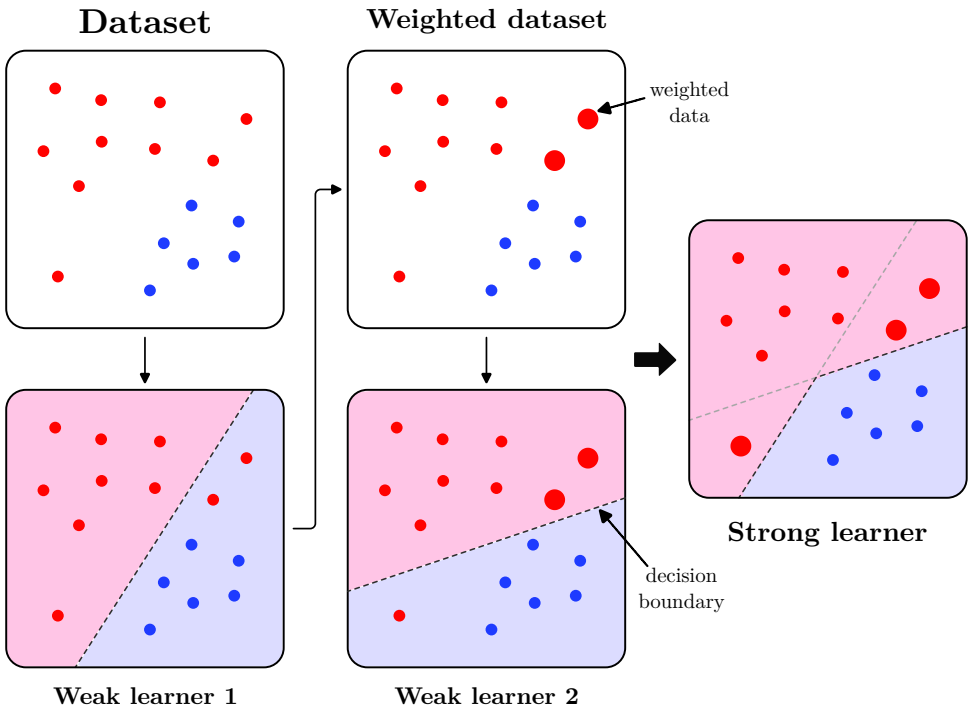


Figure 1.30: AdaBoost algorithm in few steps.

The first two classifiers have no hyper-parameters to tune, making them easy to use and straightforward to implement. However, the AdaBoost classifier has the maximal number of weak learners as well as a learning rate to tune. Details about these algorithms are out of the scope of this manuscript and can be found in the given reference [40].

Before discussing the so-called Cross-Validation technique and its relevance in the design and performance evaluation of the model, which will be addressed in the next sub-section, we should first discuss the so-called data leakage problems. Data leakage is defined as the use of information in the model training process that is not supposed to be available at the time of prediction [45]. This would not be possible in a real-life scenario, where a new sample of unlabelled data is received and needs to be categorized. The data leakage introduces a bias in the evaluation of the model and therefore the claimed performances. Indeed, the developed model will have excellent performances on the training data and even the presumably designated as test data should data leakage be present in the process of splitting the data, but it will perform poorly on new unseen data. Two types of data leakage are important to note which are explained hereafter:

1. **Group leakage:** Where correlated data/information is present in both the training and the test sets [4]. This shouldn't happen in a realistic scenario where new unseen data arrives. This problem is so serious that sometimes complex models (e.g. neural

networks) rely on the signature of the data[14] and its similarity to the training data to make their predictions. As a result, the model has poor generalization capabilities, as little relevant information has been captured. Appendix 6.3 demonstrates the effect of group leakage with an example using a Convolutional Neural Network (CNN).

2. **Optimisation over test-set:** This concerns the optimization of model hyper-parameters and/or the methods used, as well as the features that worked the best using the test set. This leakage can occur either directly, i.e. using the test set alone, or indirectly, using the entire data set (which is partly composed of the test set) [45]. The evaluation of the method becomes biased because the choice of hyper-parameters and features was made by including the test set, which is not possible in a real case. This bias is even more accentuated by the fact that today's models often have a large number of hyper-parameters and a high generalization capability, which leads to even greater over-fitting of the latter on the test-set. This problem can be alleviated by adding a validation-set (more details in the next sub-section).

1.6.1 Cross-Validation

To evaluate the generalisation performance of our method on new unseen data, we first used the simplest sampling method, namely a Nested Train-Test-Validation split (NTTV) [60]. This procedure consists first of all of splitting the data-set into a holdout test-set and a learning-set, this step takes place in the outer loop and the splitting is performed in a random and stratified manner (keeping the same proportion of the classes as they appear in the initial population if possible). The learning-set is then randomly sampled in a stratified manner into a training-set and a validation-set within the inner loop (see Figure 1.31). This sampling method allows us to have a validation set in order to select on the latter the best hyper-parameters, but at the cost of reducing the size of the training-set. It is important to recall that we cannot select the hyper-parameters that perform the best on the test-set as this is considered as data leakage. [99] is an excellent study of this bias with different Cross Validation (CV) procedures.

Firstly, the model is trained on the training-set, then the hyper-parameters[15] are fine-tuned and selected on the validation-set. The model is then retrained on the learning-set, using the best hyper-parameter obtained. Finally, the trained model performances are evaluated on the test-set. This assessment will result in only one performance estimate, which will vary considerably and will depend heavily on the splits (which subject is in which group). One solution to cope with this variance problem is to repeat the splitting procedure several times with different randomization in each repetition, giving several performance estimates, one for each given split. The overall performance will then be the average performance obtained on these different splits. However, even if we randomly split

[14] For the applications studied, models recognize the subject's signature to make the prediction.

[15] In our case they are the number of oscillators m, their angular frequency ω_i and most importantly the sparsity level α.

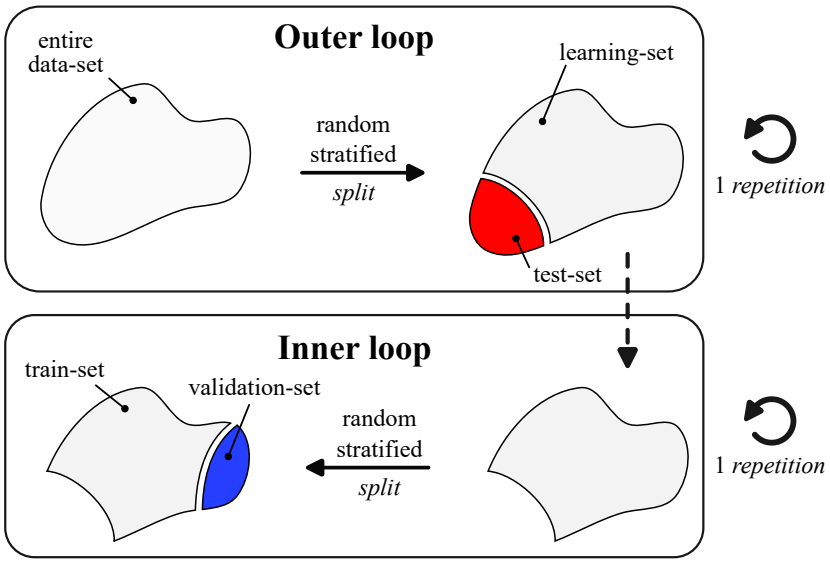


Figure 1.31: Nested train-test-validation splitting for a single trial.

the subjects into three sets, some subjects will be chosen much more frequently in one set than in the others. This creates a selection bias that significantly affects the results (the effect is much more pronounced when the dataset is small) [60]. This selection bias is reduced as the number of repeated splits is increased but at the expense of computational cost and time.

K-fold CV

K-fold CV is another solution to reduce the selection bias that works well on small data sets [92]. It consists of randomly partitioning the data set into K-folds (K partitions). For K times, we keep a fold that was not previously chosen as a holdout test-set and use the remaining $K-1$ folds as a learning-set (see Figure 1.32 for K=5 example). K may be set to the total number of data instances so that each observation is the holdout once; This procedure is called Leave-One-Out (LOO) Cross-Validation [60]. This is the best way to reduce the selection bias and improve the learning performance of the model, but at the expense of the computational load [99, 60]. In our case, this is practicable because our model is simple and computationally inexpensive, moreover, we are working on a small data set.

Nested Leave-One-Out CV

To take advantage of the benefits of the two sampling methods presented above (NTTV and LOO) and to compensate for their drawbacks, they have been combined to obtain the Nested Leave-One-Out (NLOO) Cross-Validation. This combination offers the possibility to perform hyper-parameter tuning, reduce the selection bias and compensate for the small training-set bias present in NTTV. The NLOO cross-validation procedure is composed of two nested loops: (1) An outer-loop and (2) an inner-loop, within each of which is a

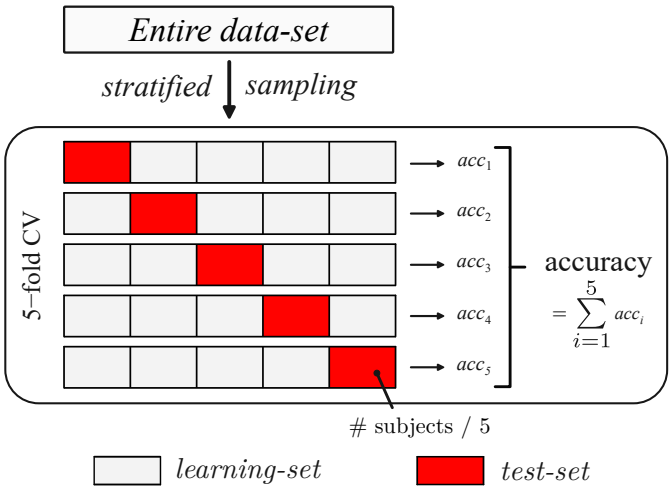


Figure 1.32: Example of 5-fold Cross Validation.

LOO procedure. The outer-loop splits the entire data-set into a learning-set and a test-set while the inner-loop splits the learning-set into a train-set and a validation-set (for more details, see Figure 1.33).

This approach is the least biased since it simulates the procedure that occurs in a real case for each repetition. In a real scenario, the entire dataset is partitioned for training and validation. Once the hyper-parameters are set and the model is trained, evaluation is performed on a new test set that has never been seen before. Therefore, the training and validation procedure is blind to the evaluation, and the choice of hyper-parameters is based only on the dataset at hand. This is exactly the same procedure replicated by the NLOO CV procedure. Each inner-loop split is independent and blind to the test set, and in each iteration, the chosen hyper-parameters will differ based on what worked best within the inner loop.

The metric used in this manuscript is accuracy, which is defined as the number of correct classifications divided by the total number of classifications made. Since we have an accuracy value for each split and for each repetition, the reported result is the overall accuracy, which is the average of all the accuracies across all repetitions and splits:

$$\text{accuracy} = \sum_{\text{repetitions}} \sum_{\text{folds}} \frac{\text{correct classifications}}{\text{all classifications}} \tag{1.22}$$

For the remainder of this manuscript, when an “N” is mentioned before the CV procedure, this indicates that a Nested CV procedure is used, either Nested Leave-One-Out (NLOO) or Nested K-fold (N K-fold). Moreover, in the nested case, the internal loop always corresponds to a Leave-One-Out (LOO) CV procedure unless it is explicitly mentioned that this is not the case.

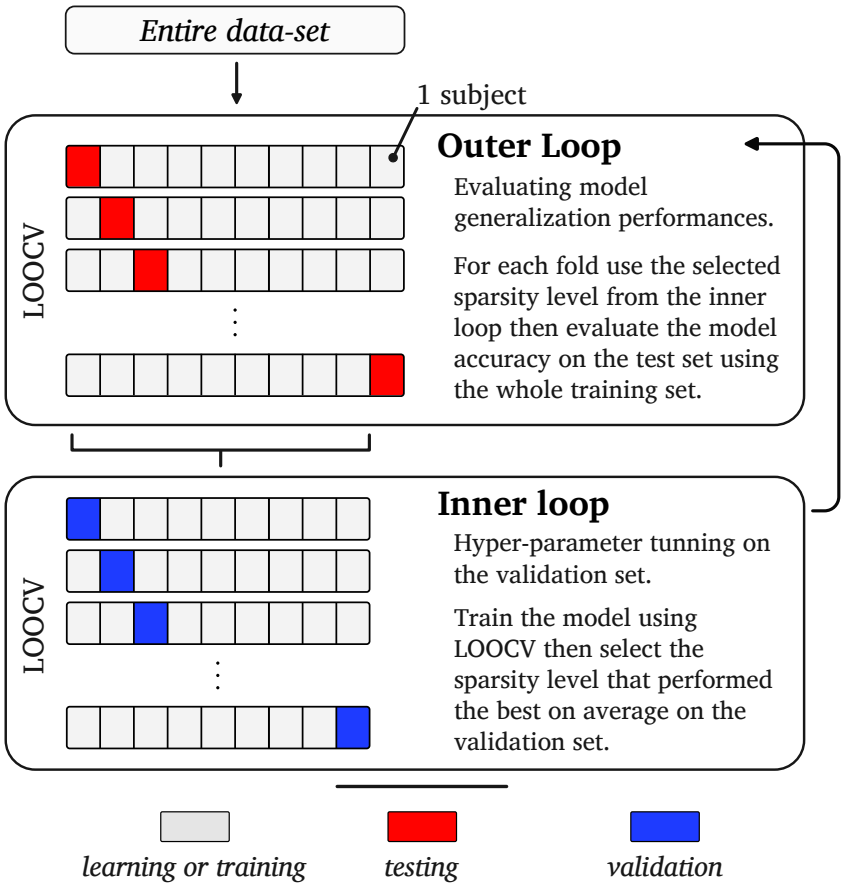


Figure 1.33: Schematic display of the nested LOO cross-validation.

1.7 Conclusion

In this chapter, we first presented the foundation of our modal decomposition. This decomposition is generic and can be employed for any physical signal. The oscillators or modes within the proposed model can be activated or deactivated at any given moment. The activation timing, along with the amplitude, is determined through the resolution of an optimal control problem. To solve the latter, the principle of parsimony constitutes an important pillar, guiding the choice of the optimiser towards the Lasso-LARS. The parameter α of the method allows for various solution profiles with varying levels of sparsity, effectively tackling the high noise problem present in the data. Therefore, only the necessary and meaningful information is captured, while mitigating the impact of noise.

The most important parameters of the method are the number of modes m, the frequency distribution of the modes, the parameter w which manages the trade-off between free response and forced response, as well as the final value of α_f which manages the degree of fit and the level of parsimony of the solutions have been discussed. The choice of the latter depends on the application, however, different procedures guiding the choice have been discussed. Selecting these parameters procedurally reduces the effect of human

subjectivity on the one hand, and on the other hand, makes the method more explicable and repeatable compared to a purely algorithmic (heuristic) selection methods.

The selection and extraction of features used in this manuscript have been discussed, highlighting their advantages and utility. The latter is even more pronounced when dealing with small datasets. Furthermore, the various employed classifiers have also been covered. The issue of data leakage is a challenge encountered in the literature, where we have presented the key points and strategies to tackle it.

Finally, the cross-validation procedure, which ensures results that are closest to reality with minimal bias, has been addressed. The ultimate solution chosen is the Nested Leave-One-Out procedure.

Main contributions

The main contributions related to the methodology presented in this chapter are the following:

- Proposal of a novel feature generation paradigm based on linear state space systems and parsimonious optimisation, offering a new point of view on the fitted signals. To our best knowledge, this is the first time we have seen this type of paradigm.
- The proposed method enables the direct disentanglement of the forced and free regimes from the signal, a topic of significant research interest in fields like Brain Computer Interfaces.
- The proposed method uses the parsimony level parameter to perform denoising which is detrimental for some applications.

Additional contributions resulting from the application of the methodology that will be presented in the remainder of the manuscript are:

- Rigorous analysis of the relevance of results in scenarios with low number of patients (data instances).
- Achieving results comparable to the state of the art that are explainable and easier to accept by the clinicians in order to be used as assisting tools. And this, while using simple classifiers and a relatively low number of features.
- Application of the proposed methodology to three distinct problems using real and publicly accessible data, without necessitating the modification of the methodology.

The next chapter will cover in depth the recording systems and their specific aspects. This is necessary to understand the context of our application, be aware of potential pitfalls, and know how to conduct the required pre-processing.

2 Recording systems

Contents

2.1 Introduction

This chapter will cover the basic aspects of the data recording systems that have been used to collect the data needed for the future applications considered. It is important to note that the experiments and data collection were not carried out by us and that the data used is publicly available. Nevertheless, this study remains important to understand the application context, and to avoid potential pit-falls and problems that may be encountered.

In Section 2.2 an introduction to the fundamental principles of electroencephalography is given, followed by the different methods used to attenuate recorded noise. Subsequently, the electrode nomenclature, montages, and positions are detailed in Section 2.2.3. The complexity of EEG recording and the various artefacts that contaminate the recorded signals are detailed in Section 2.2.4 where we can also see how these artefacts are removed using Independent Component Analysis (ICA).

In Section 2.3, a brief overview is given of the heart conduction system followed by leads position of an ECG in Section 2.3.1.

2.2 Electroencephalography

An action potential, also known as a neural impulse, is a short event during which the electrical potential of a cell rises and then falls rapidly, producing a local current [101]. This phenomenon is specific to neuronal cells as well as muscle cells [53]. Measurement of the local current reflects the underneath activity of the brain, but direct

measurement of this current is not possible. An alternative way of measuring the electrical current, thus, brain activity is to measure the voltage, which is equivalent to the current.

Electroencephalography (EEG) is a non-invasive[1] method to record the electrical activity of the brain through electrodes placed directly on the scalp. Due to the low amplitude of the electrical signal produced by the neurons, which is in the order of 10 mV; but measured at the scalp level with an amplitude generally ranging from 0.5 to 100 μV, a very high amplification gain device is required to amplify the small electrical signal measured by the electrodes. EEG has been used by several studies to assess individuals' health conditions and to study brain function in healthy individuals as well as to diagnose various diseases that alter the brain electrical activity such as: Parkinson's Disease, epilepsy, Alzheimer's, sleep disorders, schizophrenia, etc [90].

The Figure 2.1 shows an example of EEG recordings.

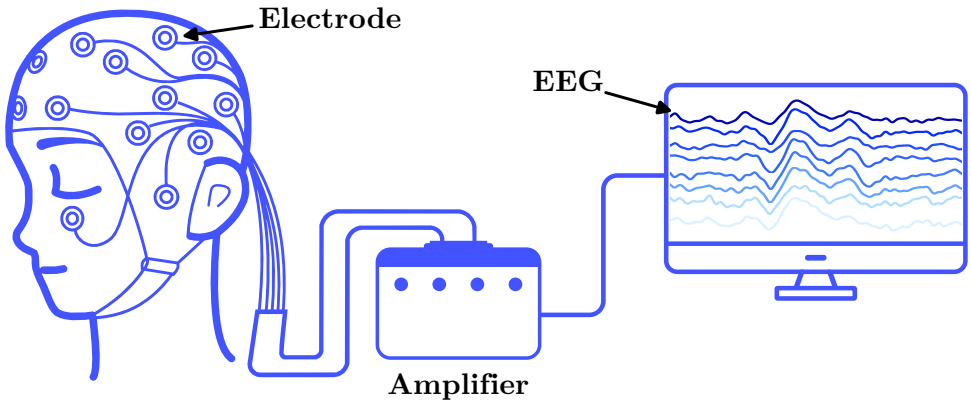


Figure 2.1: Illustration of EEG recording.

2.2.1 Electrode types

There are two types of electrodes, each with its own advantages and disadvantages:

1. **Wet electrodes** :
    - Generally made from silver, pewter or gold.
    - A conductor acting as a bridge between the electrode and the scalp. This can be either an electrolytic gel (for research purposes) or a glue (for clinical use).
    - **Low impedance**: the gel gets into the skin's pores, reducing impedance between the electrode and the scalp.
    - Movement resilience: the application of the gel creates a contact surface enabling the electrode to remain connected, so the recorded signal is not disrupted a lot by the subject's movement.

[1] Does not require passage through the skin.

- It takes around 30 minutes to install, and requires a professional.
- Risk of creating an electrolytic gel bridge between two relatively close electrodes, resulting in an identical measurement signal from both electrodes.

2. **Dry electrodes** :
   - Usually made from graphite or coal.
   - No conductive gel is required to be used.
   - **High impedance**: the contact of the electrodes is not good enough and not directly with the skin, sometimes it even touches the hair which is a poor electrical conductor.
   - Highly sensitive to movement: each movement is amplified, reducing the Signal-to-Noise Ratio (SNR).
   - Rapid set-up due to the fact that it is a slip-on helmet (non-medical use).

Figure 2.2 shows the use and contact point between a wet and a dry electrode.

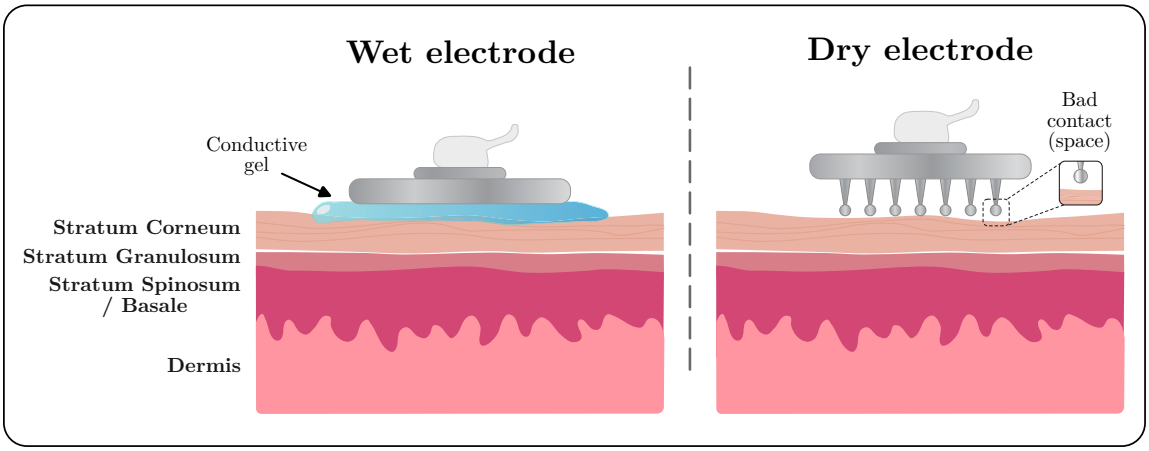


Figure 2.2: Wet electrode (left), dry electrode (right) with the different skin layers.

A high impedance impedes the passage of current from the scalp to the electrodes. As a result, fewer EEG signals of interest will be transmitted to the electrodes and therefore recorded, which has the effect of reducing the signal-to-noise ratio. Therefore, to reduce noise and improve the quality of the EEG recording, the impedance of the electrodes should be as low as possible. Figure 2.3 shows a comparison between the impedance over different frequencies between the two types of electrodes (wet and dry) (Figure has been adapted from [107] only style changes have been made). We can observe that wet electrodes have a lower impedance and therefore a better recording quality, which explains why they are widely used in the medical field.

We have discussed this point about electrode types as currently, wet electrodes, which are widely used in practice and in the medical field, are much more practical for serious

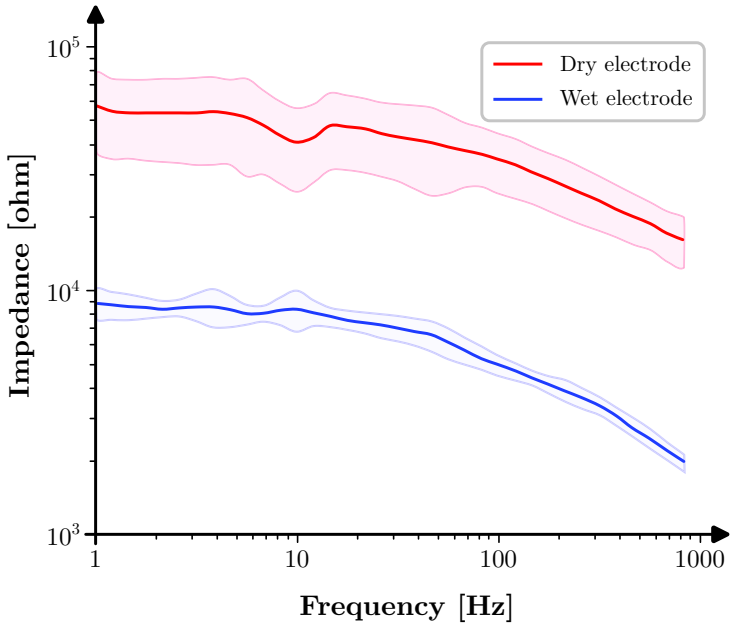


Figure 2.3: Comparison of impedance at different frequencies of wet/dry electrodes (Figure has been adapted from [107] only style changes have been made).

experiments because they allow recording with much less noise than dry electrodes ([107] studied this difference). However, dry electrodes are much easier to set up, so we can imagine helmets that can be put on for everyday use. We can imagine that we will be able to diagnose various brain diseases, but we think that this application is a long way from reality at the moment. The current aim of this work is rather to assist clinicians in diagnosis.

2.2.2 Principles of EEG measurement and noise attenuation

As presented in the previous section, the electrical signal produced by the neurons is of very low amplitudes. Therefore, an amplification device with a very high gain is necessary. Operational Amplifiers are electronic components that amplify an input voltage to produce a higher output voltage. Figure 2.4 is a simplified circuit of an Operational Amplifier.

The use of an OA requires the comparison of two electrical potentials (V_+ and V_-).

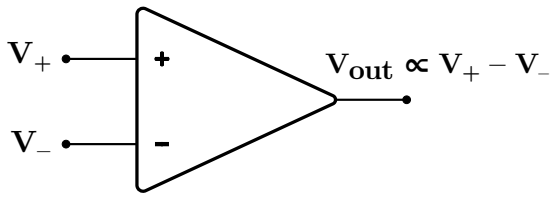


Figure 2.4: Operational Amplifier.

Usually, the input potential is V_+ and V_- is the reference potential, so the question arises:

which reference potential should we compare our electrode potential to? In the following, we will answer this question.

Referencing and grounding

Since the ground is intrinsically at a potential of 0 V, it is in most cases chosen as the reference point. Figure 2.5 shows the output using one Active electrode (A) and the ground (G) as a reference point.

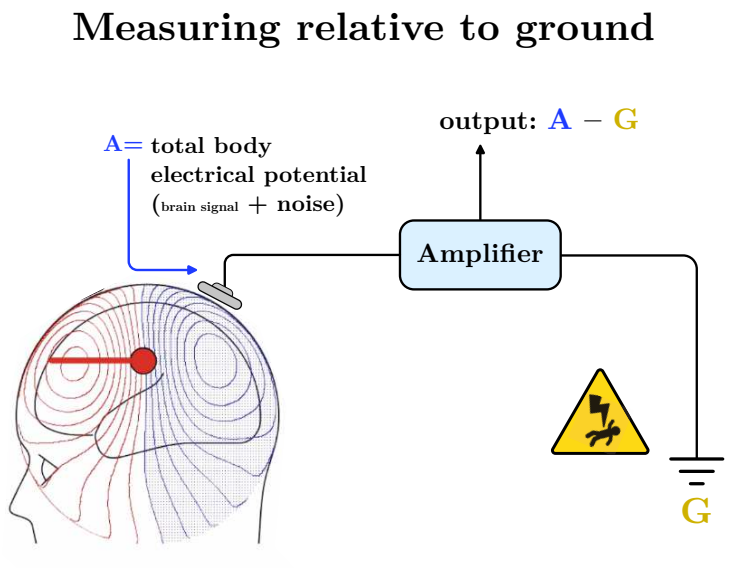


Figure 2.5: EEG recording using the ground as a reference point.

In electroencephalography, using the ground as a reference is dangerous and has several inconveniences which include:

1. The potential measured does not correspond to brain activity, but rather to the electrical voltage of the body as a whole since the human body is an electrical conductor. The measured brain signal is small and drowns in the measured electrical potential of the whole body, which in this case is considered noise.

2. Usually, the measuring instrument is affected by noise, and this noise affects all the electrodes equally. When the ground is used as a reference, the noise is not attenuated contrary to other montages[2] that we will present next.

3. Our aim to reduce impedance for better recording quality makes ground reference dangerous. Indeed, the danger comes from the fact that the current (as it is easier now with a low impedance) from the recording device can pass through the subject and electrify him or her.

[2] It refers to a specific arrangement or configuration of the electrodes placed on a person's scalp to record the electrical brain activity.

In order to overcome these problems, different montages exist, such as the one illustrated in figure 2.6. This montage is known as common mode rejection and uses another measuring electrode placed on the skull as a reference point.

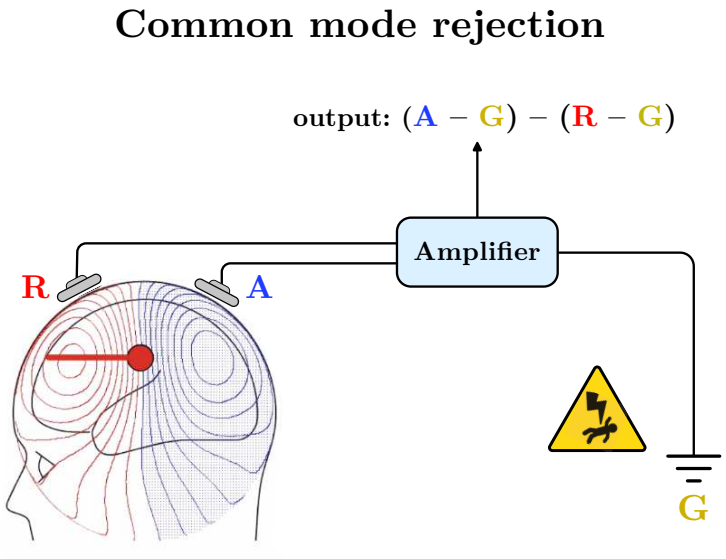


Figure 2.6: Common mode rejection montage.

As we can observe in Figure 1, the measured potential is no longer that of the whole body, but only the brain activity between points A and R (reference), this addresses problem 1 mentioned above. Since the measured potential is the difference between A and R, the same noise carried in A and R will be cancelled out, which addresses problem 2 mentioned above. The amplifier's internal noise and external noise are also attenuated by maintaining the G measurement point. The only remaining problem, yet very important, is the electrification of the subject, which is due to the fact that the subject is not galvanically isolated and is still in direct contact with the ground wire.

To overcome this, an internal point to the amplifier (floating ground) is used to isolate the subject from the power grid, resulting in the configuration shown in Figure 2.7, which addresses problem 3 mentioned above. However, information about the absolute potential (ground) is lost.

The ground electrode is often placed on the forehead (but could be placed anywhere else on the body; the location of the ground on the subject is generally irrelevant). However, Some reference electrode locations are more coveted and used in practice, and are shown in Figure 2.8.

It is important to note that the head can be considered as a dipole, and if we assume that the head is perfectly spherical and place the electrodes equidistantly and symmetrically, the sum of the signals recorded by all the electrodes will be zero 0 V. So we can

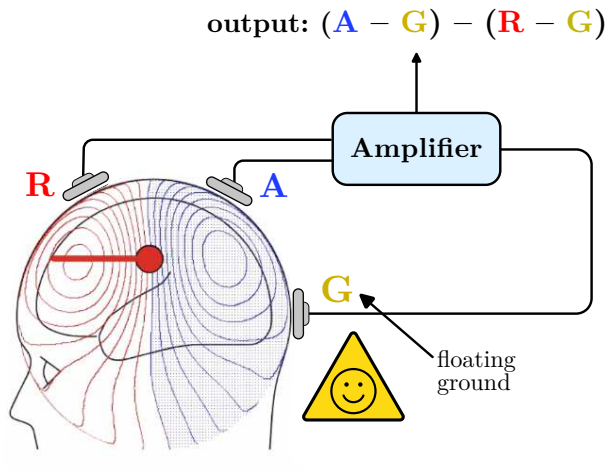


Figure 2.7: Typical montage with: an active electrode, a reference, and a floating ground.

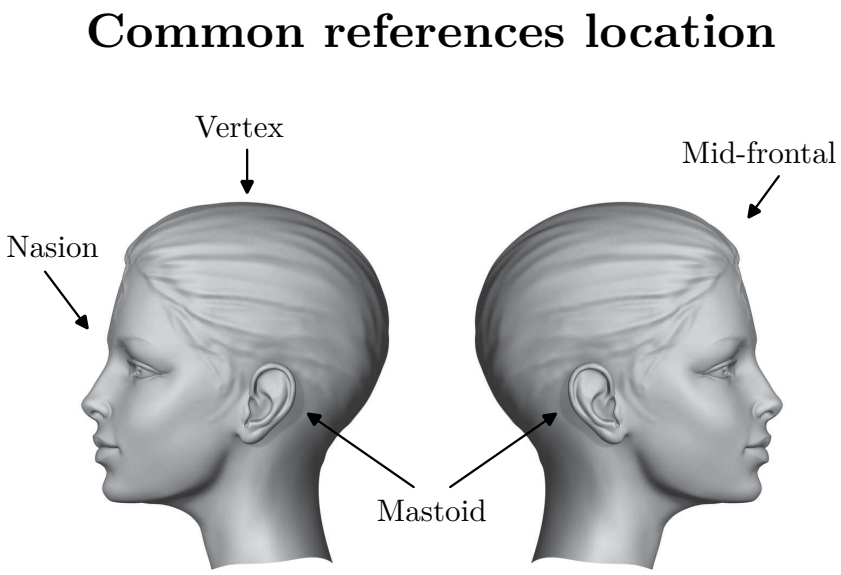


Figure 2.8: Location of commonly used references.

use this average as a virtual reference point. This makes the use of the average reference interesting, as it gives us information relative to the absolute potential. In practice, the average reference is the most widely used reference.

It should be noted that the choice of reference during recording is not important since re-referencing can be done post-recording using simple subtractions. However, even if it is simple to change a reference, it has a great impact on the EEG and on our point of view on it as indicated by Figure 2.9 where the same signals are illustrated for various reference points.

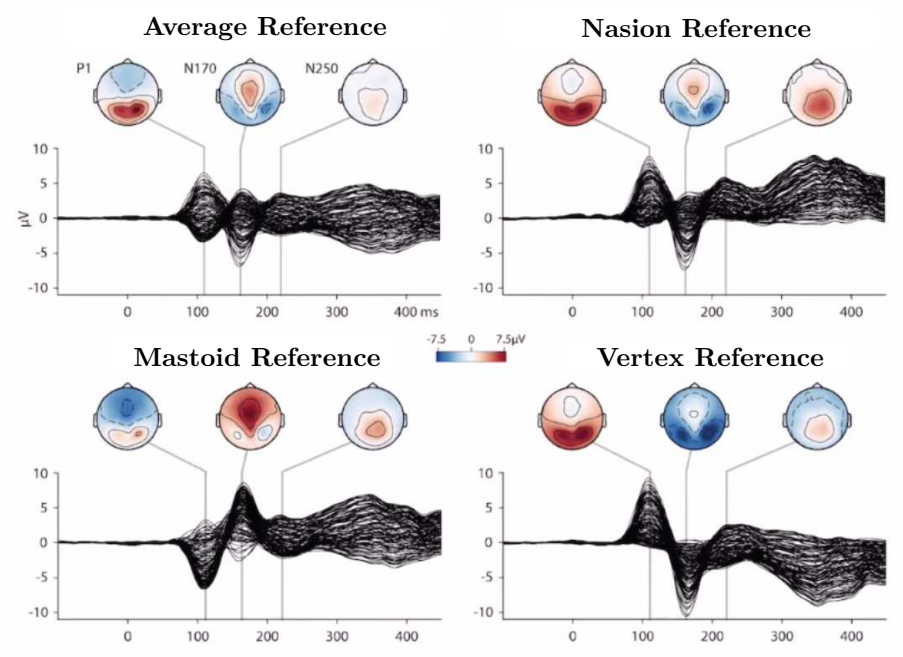


Figure 2.9: Effect of reference change on EEG.

2.2.3 Nomenclature

The electrodes must be well placed in order to maintain consistency, reproducibility of results, analysis and scientific comparison of the different available methods. Not only should the electrodes be placed in the same position on the skull, but the same electrodes should be kept in the same place. This is why there are standards for placement as well as a nomenclature of electrodes.

Montages

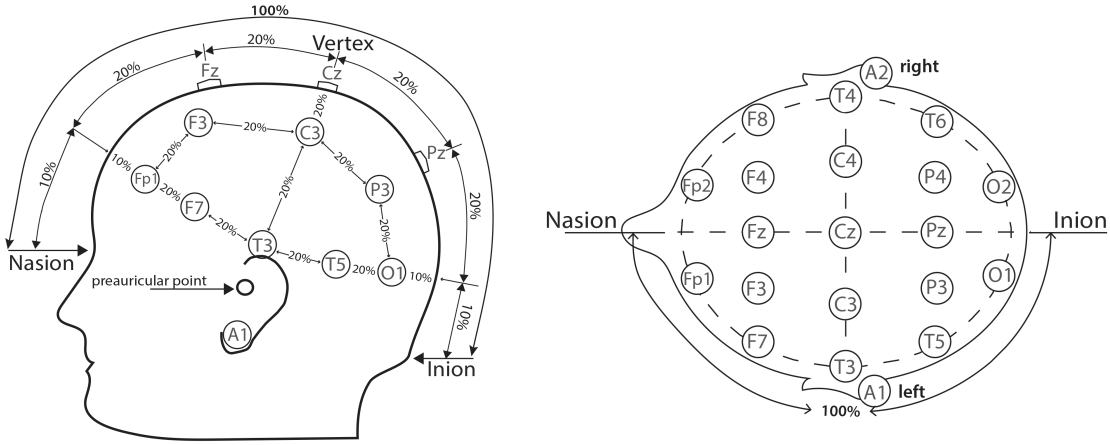
There are different EEG recording systems with different montages that differ in the number of electrodes used. There are those with 64 electrodes and those with 32 or even 16. Each has its own standard of placement, but for the example that follows the international 10-20 system is presented.

For the correct placement of the EEG electrodes, two anatomical measuring points are used and they are (see Figure 2.10):

1. The **nasion**: the point between the nose and forehead.
2. The **inion**: the lowest point of the skull on the occiput.

The line connecting the nasion to the inion passing by the vertex (highest point of the skull) is first to be considered. Two other lines connect the nasion to the inion passing by the left eat canal opening (A1) and the right one (A2). Theses lines are divided into

10—20—20—20—20—10 intervals as presented in the Figure 2.10. At each interval an electrode is placed, the latter should be symmetrical and equally spaced.



(a) **(b)**

Figure 2.10: Figure showing the international 10-20 electrodes placement system. (Taken from [95] after author's approval)

Nomenclature

Once we know where to place the electrodes on the skull, it is important to know which electrode to put where. There is a fairly intuitive nomenclature for naming the electrodes. It is based on numbers and letters to identify the electrode and therefore its placement.

- **The numbers** :
  They denote the location of the electrodes according to the hemisphere in which they are located. The letter z defines the location of electrodes positioned on the median line (the line that connects the nasion to the inion). Odd numbers are used for electrodes located in the left hemisphere of the brain, and even numbers for the right hemisphere.

- **The letters** :
  They are used to identify the location of the electrodes according to the lobes of the brain. C = central; FC = fronto-central; F = frontal; P = parietal; T = temporal; A = auricular; O = occipital.

Figure 2.11 shows an example of a montage with 64 electrodes (the one used in this work) where the position of each electrode is depicted. Depending on the colour used, it also shows which electrodes records/is closed to which brain lobe. This is important to know as each region is specialised in specific tasks (a sensory region, a motor region, etc.).

To give an idea of what an EEG signal looks like, Figure 2.12 is an example of a recorded EEG signal using 64 electrodes (channels). Each signal represents the electrical activity recorded at the corresponding electrode.

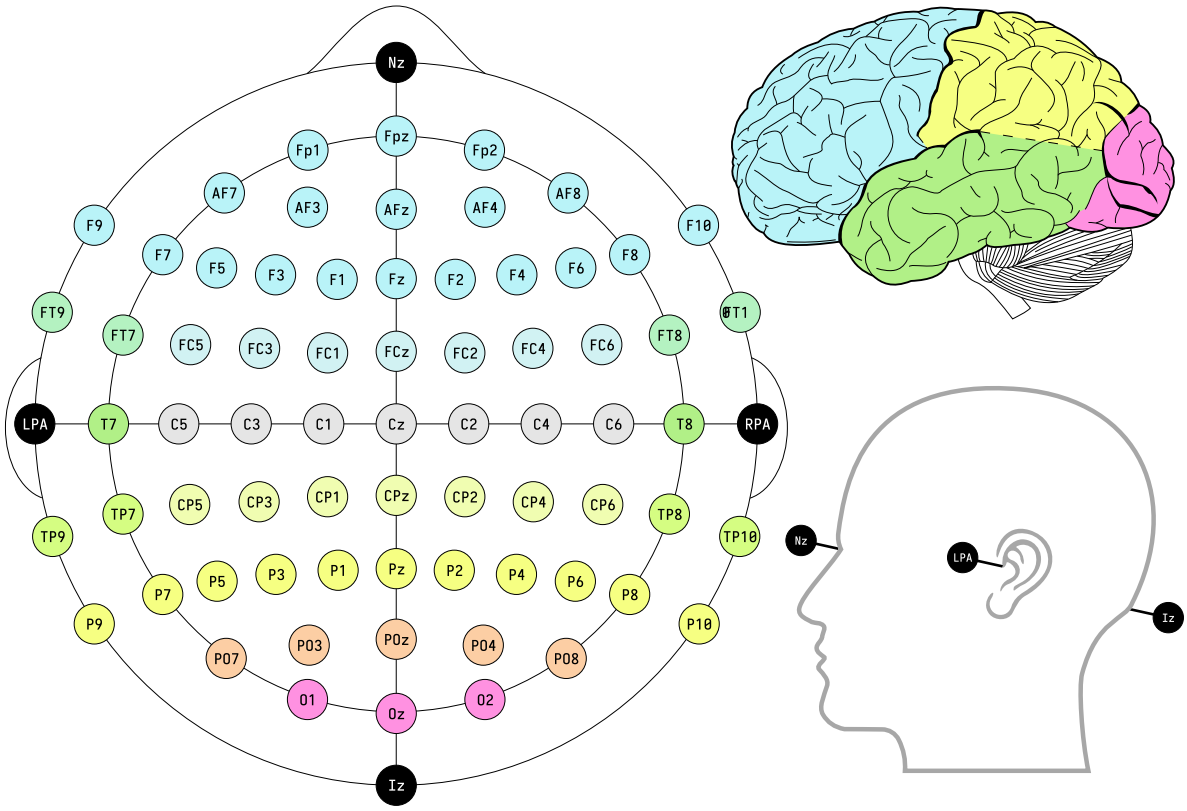


Figure 2.11: Placement and position of a montage with 64 electrodes.

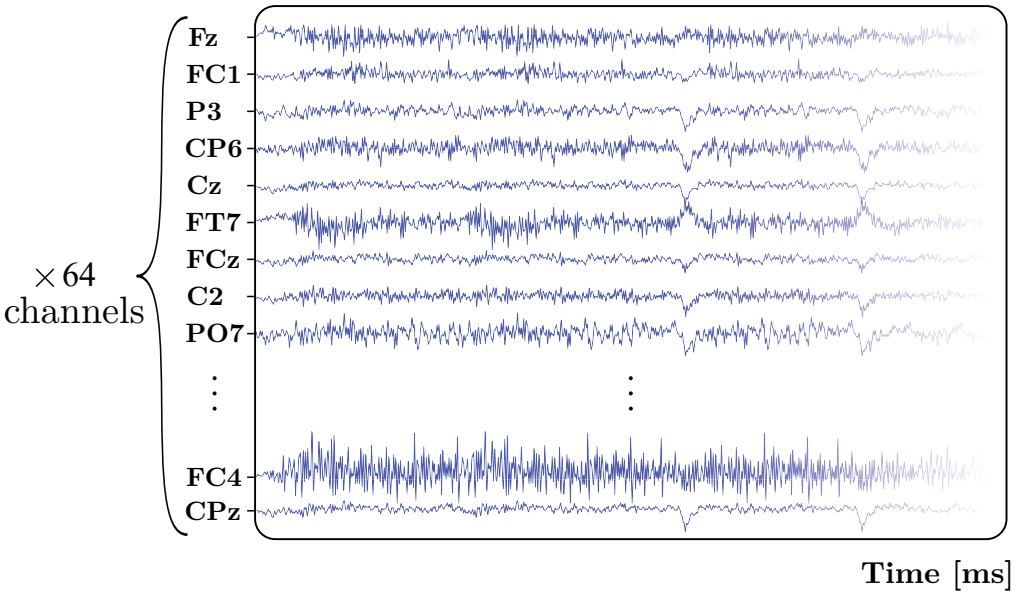


Figure 2.12: Recorded EEG example with 64 channels.

2.2.4 EEG difficulties and artefacts

EEG signals are known to have a low signal-to-noise ratio and present many difficulties. EEG noise is defined by any measured signal whose source is not the coveted brain activity [98]. Unfortunately, in most cases, the EEG signal is contaminated by various unwanted artefacts, even though we try to limit their occurrence during the recording sessions. These artefacts are entangled with the desired brain activity and can have an amplitude up to 100 times that of the brain activity. In most EEGs we encounter the following undesired artefacts: ocular, muscular, cardiac, perspiration, line noise, etc. (more details are given in the following section).

Another difficulty that we may encounter during EEG analysis is volume conduction, i.e. the transmission of electric fields from a primary current source through biological tissue towards the recording electrodes [68]. The current may pass through different conducting mediums[3] and is altered as the different tissues have different impedances. Because of volume conduction, unwanted artefacts will impact a broader region and therefore will contaminate more electrodes. In addition, we lose the ability to study a single source or brain region of interest; information is diluted and a signal recorded at one electrode is a combination of all the electrical activities present in the neighbourhood of that electrode [98].

A good EEG analogy is like placing several microphones over a football stadium. What each person is saying couldn't be recovered, but other information could be obtained, such as whether a goal has been scored. In the same analogy, the EEG records a group of neurons oscillating in synchrony but does not record the individual activity.

EEG artefacts

This section will be devoted entirely to the study of artefacts that may be present in our signals. The latter have a large amplitude and greatly affect the EEG. A method widely used in the field of EEG signal processing and which aims at isolating the various brain signals from the artefact signals is Independent Components Analysis (ICA) (more details are available in [12]). Briefly, this method decomposes a multivariate signal into additive subcomponents with the assumption that the subcomponents are Gaussian and independent, which explains the name given to them: Independent Components (IC). Once the decomposition is computed, the data are projected using a matrix W from the channel space to the independent components space. It is to be noted that the number of ICs is equal to the number of channels. After that, the components that contain unwanted artefacts are removed, so that only the good components remain. Then, the remaining good components are projected back using the inverse projection W^{-1} to their initial space (channels) as presented in Figure 2.13.

[3] Skull, meninges and cerebrospinal fluid.

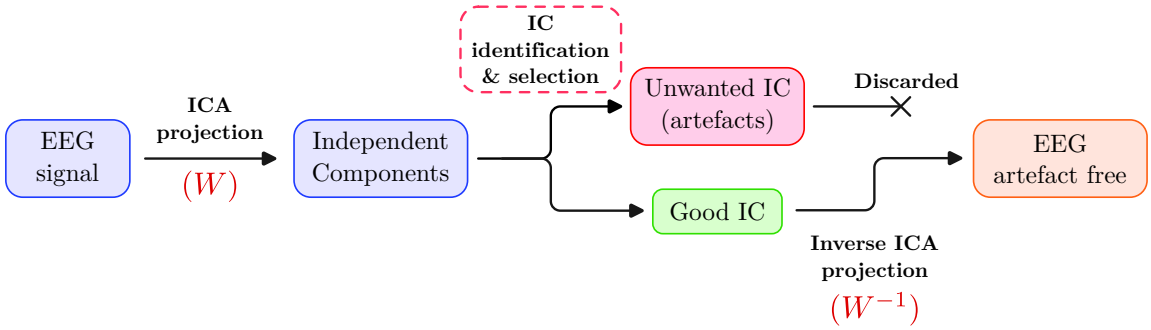


Figure 2.13: Flowchart showing how removal of unwanted Independent Components using ICA is performed.

Before removing the bad components, we need to know how to identify them and how to distinguish them from the good ones, which is the subject of this section. The ICs are computed using dedicated libraries such as EEGLAB (a Matlab library specialised in EEG processing and analysis) [18], and their content can be examined in greater detail. The analysis is based on the layout shown in Figure 2.14 where each element is defined as follows:

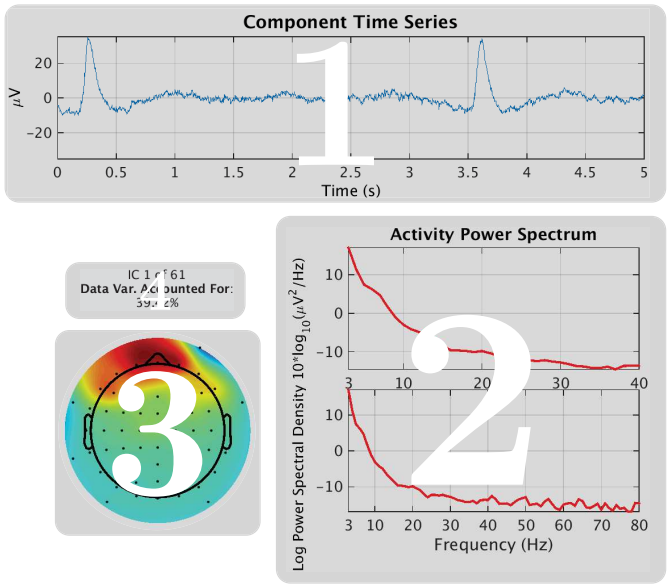


Figure 2.14: Independent component analysis template.

1. **Component Time Series:** Shows a segment of the component's activity.
2. **Activity Power Spectrum:** Shows the power spectrum of the IC activity. To help with scaling, the top plot shows the spectrum from 3 [Hz] to 40 [Hz] while the lower plot shows the spectrum to at most $f_s/2$ [Hz], where f_s is the sampling frequency.
3. **Scalp Topography:** Shows what effect the independent components process has on each electrode. Electrodes' position is represented in black dots, green colours

represent no effect, where red and blue show positive and negative contributions, respectively. It is to note that colours far from any electrodes are computed following an interpolation/extrapolation.

4. **IC Number** and **Percent Data Variance Accounted For (pvaf):** Components are classified according to their importance. The metric used to achieve this is the pvaf (details will be given at the end of this section) but what is important to know is that the higher the pvaf, the more important is the IC.

The unwanted components can come from physiological sources (eye blink, muscle movement, etc.) but also from non-physiological (external) sources such as power line noise, etc. Each of these components has its own distinctive mark and specificities. The following is a non-exhaustive list of the different electrical activities thus components usually encountered in practice:

1) Brain components:

They correspond to the brain's own activity and are quite often the desired components to be retained. Figure 2.15 is the analysis panel of a brain component.

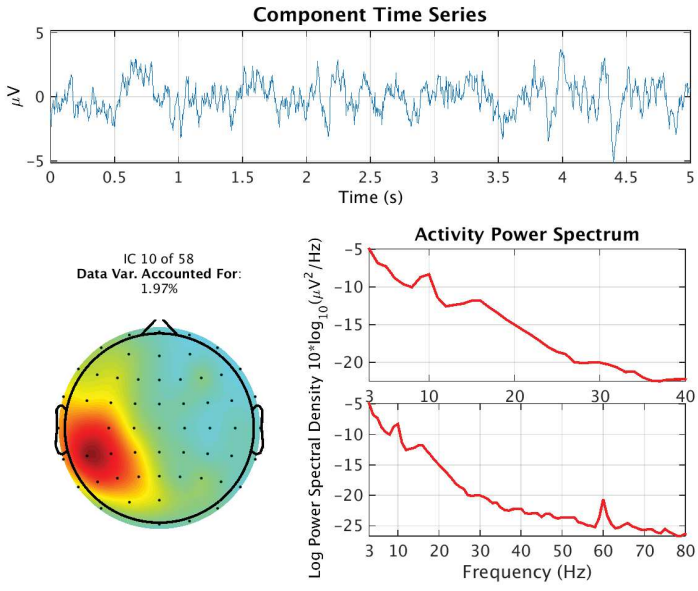


Figure 2.15: Independent component representing brain activity

Its main characteristics are:

- A scalp topography generally similar to a dipole.
- Spectral energy decreases with frequency.
- The spectral energy generally contains peaks varying between 5 and 30 Hz, with a greater concentration around 10 Hz.

2) Ocular components:

They correspond to the effect caused by the movement of the eye on the recorded EEG. Figure 2.16 is the analysis panel of an ocular component.

Ocular activity is mainly characterised by :

- Strong topographical activity at the eyes region.
- A spectral energy of less than 5 Hz.
- A visible blink on the temporal signal.

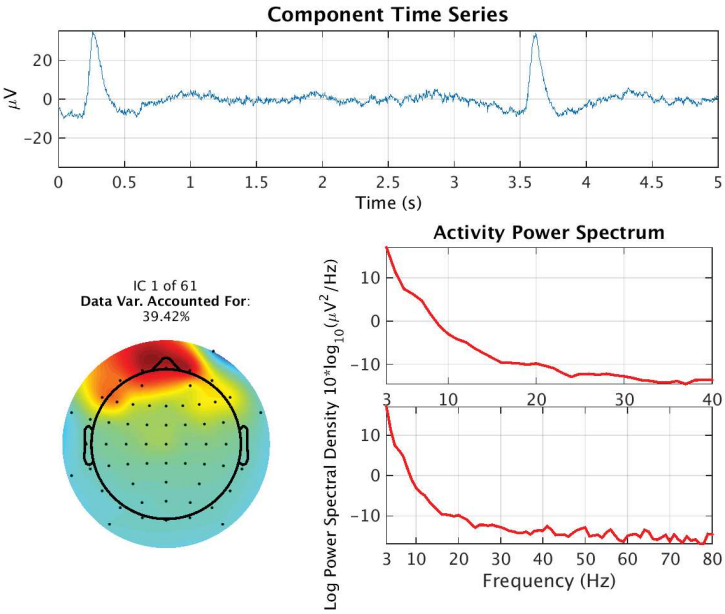
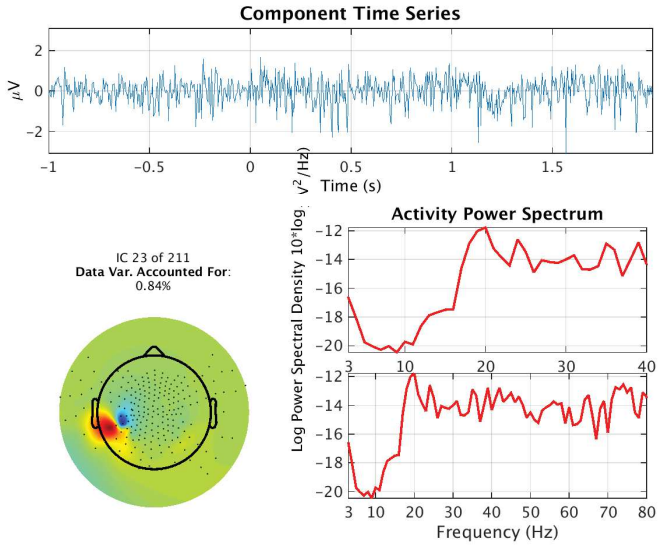


Figure 2.16: Independent component containing ocular artefact.

3) Muscle components:

They describe the electrical fields generated by muscle activity. Their activity is more powerful relative to EEG. Their unit action potential are not synchronized causing more of the power to be spread out among higher frequencies. Figure 2.17 is the analysis panel of a muscular component.



Muscle activity is mainly characterized by:

- A spectral energy above 20 Hz.
- A very close dipole.

Figure 2.17: Independent component representing muscular activity.

4) Heart components:

Heart components capture the electrical potentials generated by the heart. The pattern observed is very typical to a heart electrocardiogram (ECG). As the heart is far from the scalp, the topography map looks as a far dipole which looks like a linear gradient. Figure 2.18 shows the analysis panel of a heart component.

Heart activity is mainly characterized by:

- Clear heart shape (QRS complex) in the data (see section 2.3 Figure 2.24).
- Linear gradient scalp topography.
- No peak in the power spectrum.

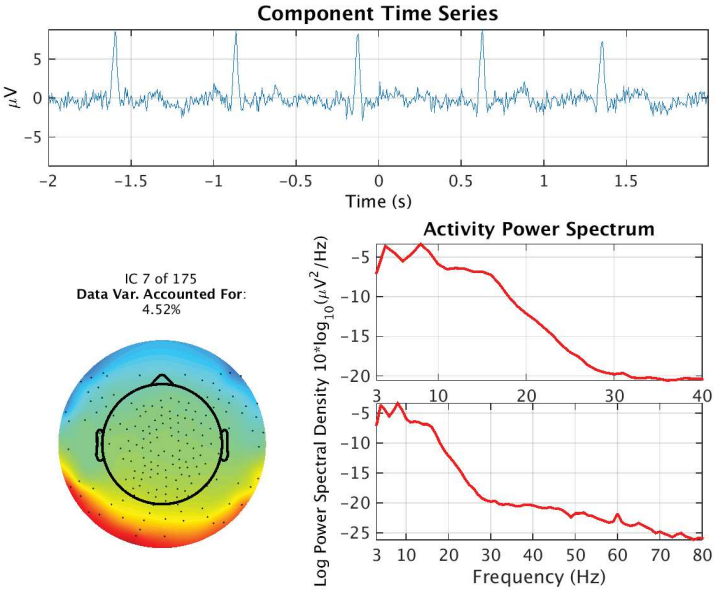
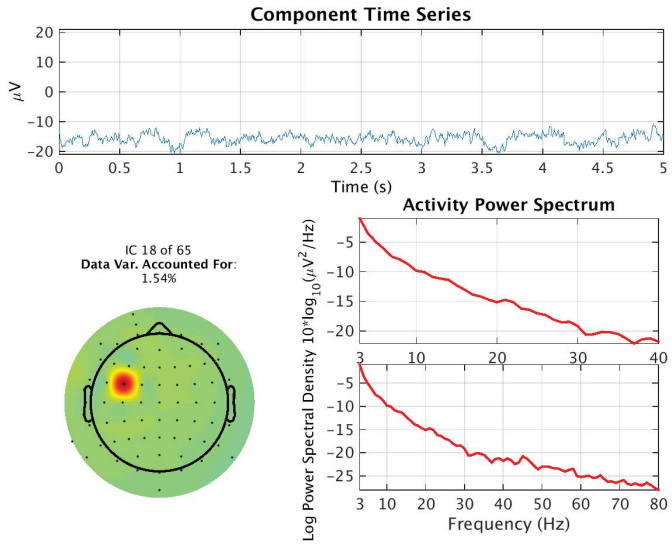


Figure 2.18: Independent component representing cardiac activity.

5) Channel noise:

If a channel has poor contact or gets bumped a lot during a recording, it will often generate large artefacts that do not affect any other channels. It can be recognized by its scalp topography which puts almost all the weighting on a single channel. Figure 2.19 is the analysis panel of a channel noise component.



This type of noise is mainly characterised by:

- A highly focal topography.
- They have similarities with muscle components, which makes them easy to confuse. However, their respective Spectral Energies are very different. Moreover, they differ in the scalp topography, which is a dipole for the muscle component and a single point for the channel noise component.

Figure 2.19: Channel noise independent component.

6) Other:

It is important to note that not all components are meaningful. ICA assumes that there are as many independent components as there are electrodes which is almost never the case. When a component does not converge onto a meaningful signal, it can either capture a mixture of signals or some noise. Anything that does not fit the above component categories can be deemed "other". However, usually, the contribution of these components to the EEG is negligible, if this is the case they can be left untouched. Figure 2.20 is one example among many analysis panels of the component other.

Mainly characterized by:

- A non-dipolar topography.
- A spectrum that may have a low 10 Hz peak, as the brain signals are probably mixed up with other noisy signals.
- Anything that does not fall into one of the above component categories.

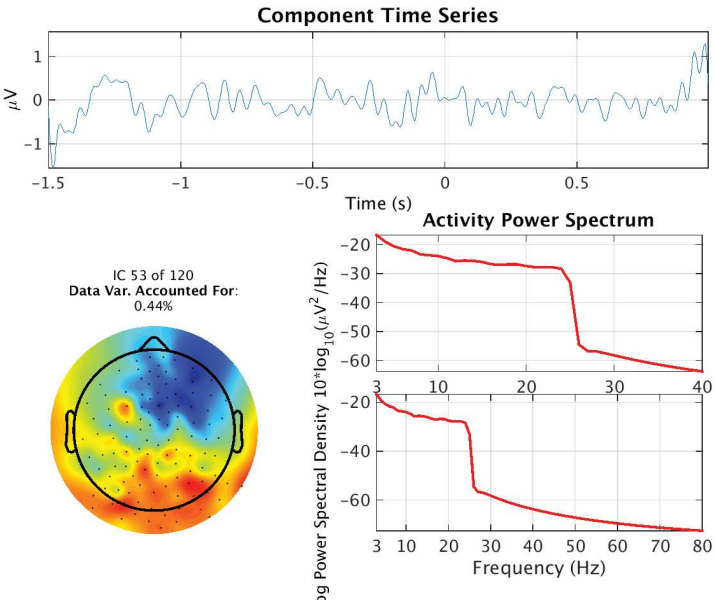


Figure 2.20: Other independent components.

Once the identification of independent components has been covered, in what follows we will illustrate the usefulness and potential of this decomposition with a practical example. The same procedure depicted in Figure 2.13 will be followed. First, a raw EEG signal over its 64 channels is presented, the latter is contaminated by ocular and cardiac artefacts that are visually distinguishable[4]. Next, an ICA decomposition is applied to this EEG and the various resulting components are shown and identified. To demonstrate that we can target and remove only a single artefact type, we will only remove the eye blinks from the EEG and leave the cardiac artefact. Finally, the result of the same EEG without the eye-blink effect will be shown.

Figure 2.21 is an example of a raw signal where, highlighted in red, the distinctive signal of a blink and a cardiac artefact can be observed. We have shown all the channels to illustrate that these artefacts can affect certain channels more than others (the amplitude of the blink is greater in the channels close to the eyes).

[4] This example has been chosen because an in-depth analysis is not strictly required; it can be done with the naked eye because the artefacts are quite visible.

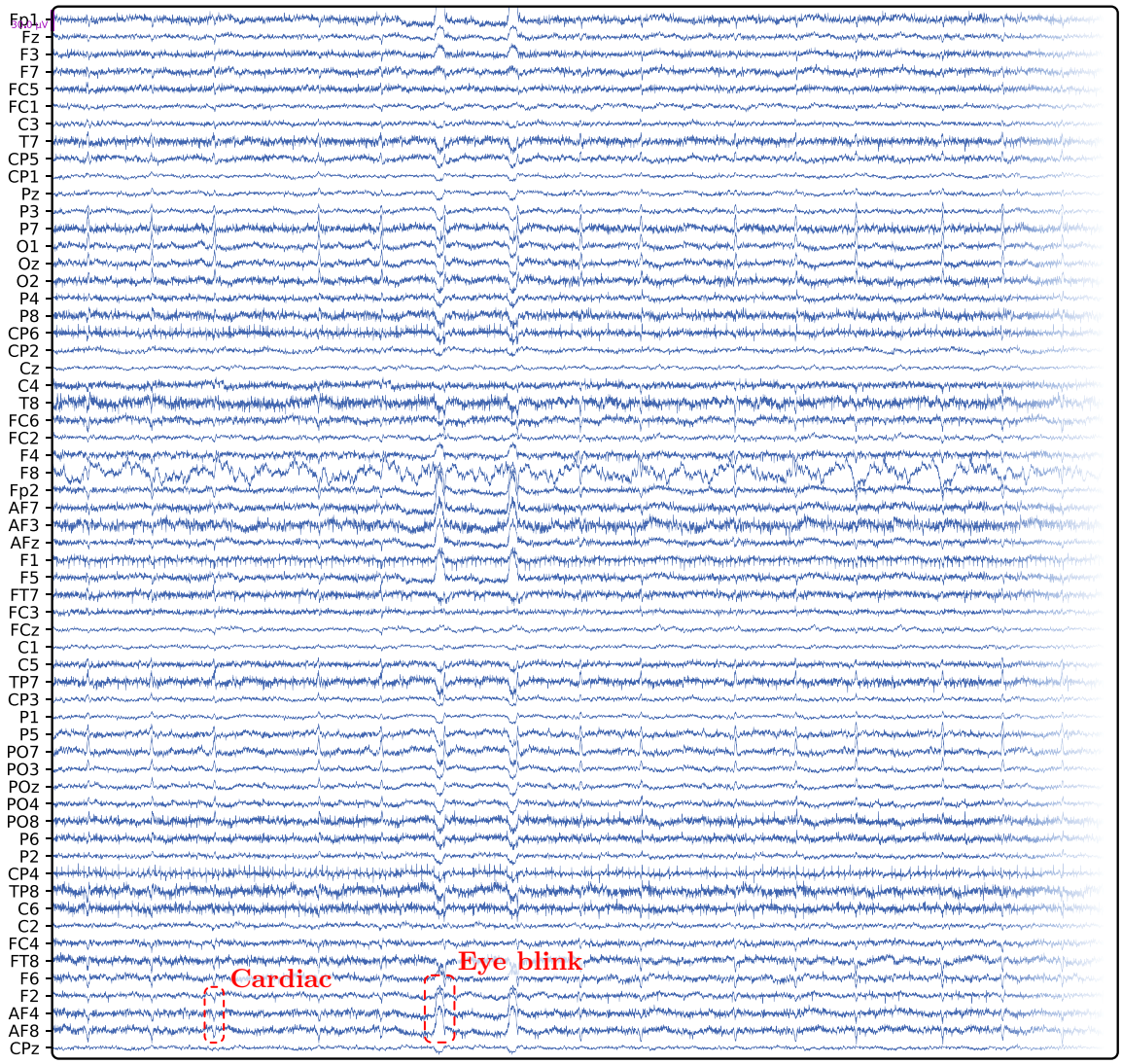


Figure 2.21: Example of a raw EEG with 64 channels. Cardiac and ocular activity can be observed on it.

In Figure 2.22 only the first 5 of the 64 available Independent Components are shown[5] (as a reminder, there are as many channels as there are components) with their corresponding pvaf between brackets. By visual inspection, eye blinks can be recognised on the first component and cardiac activity on the second and third components. For the sake of the example, only the first component (ocular only) will be removed, by setting all its values to 0, thus eliminating its contribution to the EEG.

After removing this component, the signal will be projected back from components space into channels space using the inverse projection matrix computed by the ICA (W^{-1}). The result of this projection is shown in Figure 2.23. We can observe that the ocular activity that affected all EEG channels has been removed, while the cardiac activity remains.

[5] For space-saving reasons and clarity purposes, only the first 5 are shown.

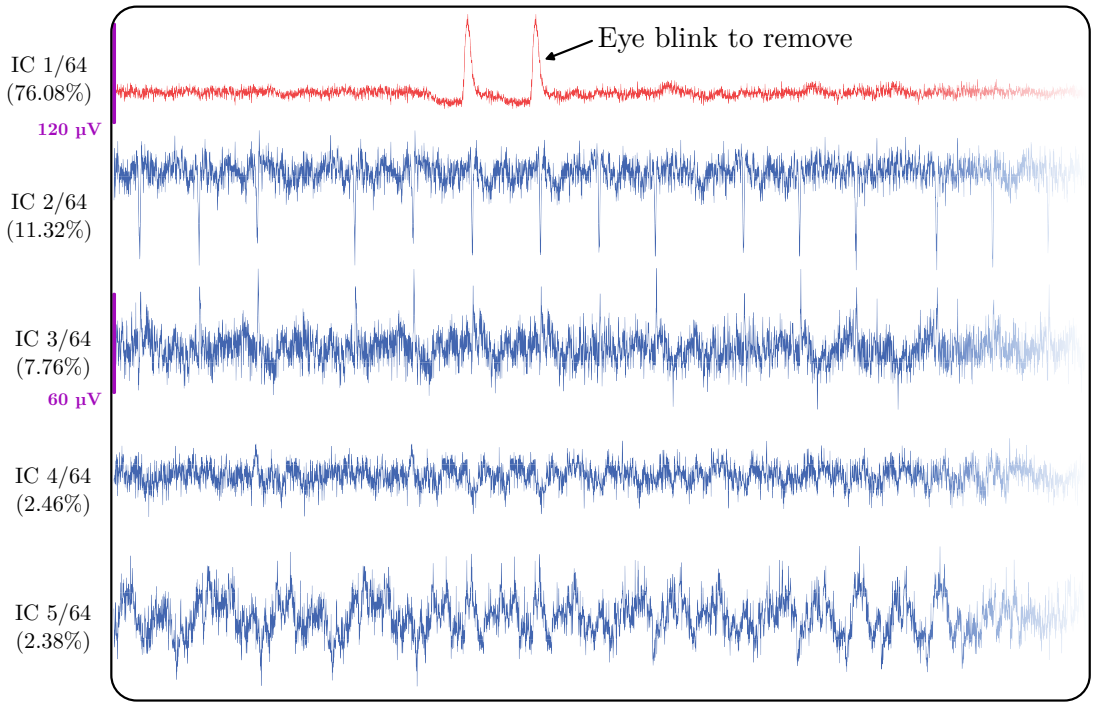


Figure 2.22: Example of first 5 Independent Components, pvaf is represented between parenthesis. It should be noted that the amplitude scale of the first component is different from the others due to its relatively high amplitude.

Percent Data Variance Accounted For (pvaf)

The percentage of representative variance "pvaf" of a component describes the extent to which the initial variance of the channel data can be attributed to that component. In other words, it indicates the percentage of variance that the component represents on the initial data. It is calculated by projecting back the component to the channel representation and then assessing its variance contribution following this formula [see EEGLAB [18], version 14, function: "eeg_pvaf"]:

$$pvaf(IC_i) = 100 - 100 \times \frac{var\left[EEG - W^{-1}IC_i\right]}{var(EEG)},$$

where IC_i is the i-th component for which we want to calculate the pvaf, $W^{-1}IC_i$ is the inverse projection from independent component space to channels space, $EEG - W^{-1}IC_i$ is the difference channel by channel between the initial EEG signals and the contribution of the IC_i component to the EEG.

This measure is useful for determining the relative importance of each component [54]. Once the independent ICA components and their corresponding pvaf have been computed, the components are ranked in descending order of pvaf. Thereafter, the number assigned to each component will represent its rank such as the first component will have the highest pvaf and so on.

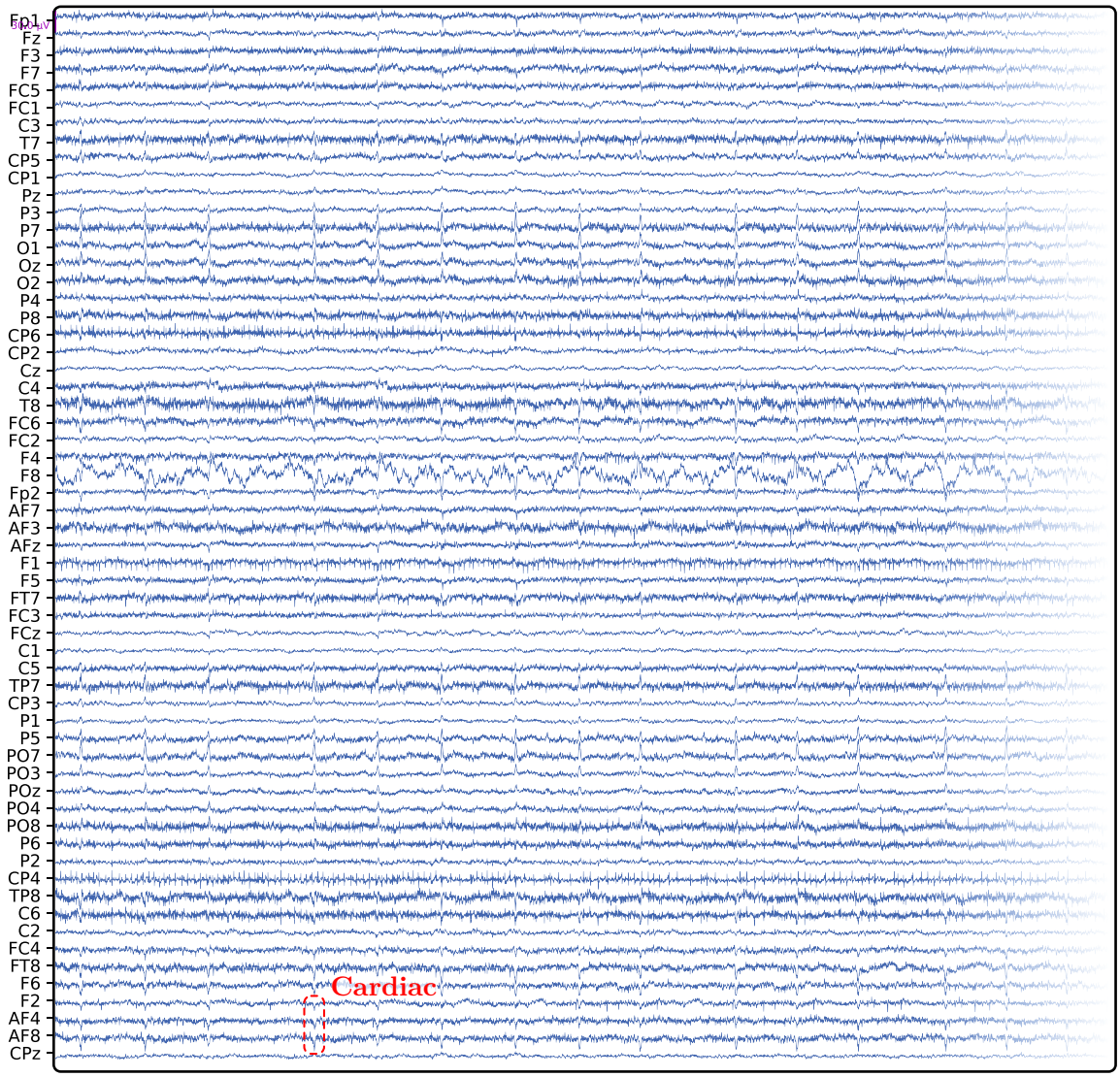


Figure 2.23: Same EEG signal as Figure 2.21, with only cardiac activity removed following ICA decomposition.

2.3 Electrocardiography

The heart has four chambers, the two upper chambers known as the atria and the two lower chambers known as the ventricles. In a healthy heart, the signal that triggers the heartbeat begins in the upper right chamber of the heart (right atrium) at the Sinoatrial node[6] which is the pacemaker of the heart. From there, the signal activates the left atrium and travels to the AtrioVentricular (AV) node. Then, it travels to the lower chambers (right and left ventricles) via the bundle branches. As the signal travels along the cardiac conduction system, it triggers nearby muscles to contract in a coordinated manner which

[6] The name sinus rhythm comes from the fact that the depolarisation of the cardiac muscle begins at the sinoatrial node.

produces the electrical wave (depolarisation wave) observed on the ECG as the one in Figure 2.24. The most important points along the electrical pathway discussed above are shown in Figure 2.25.

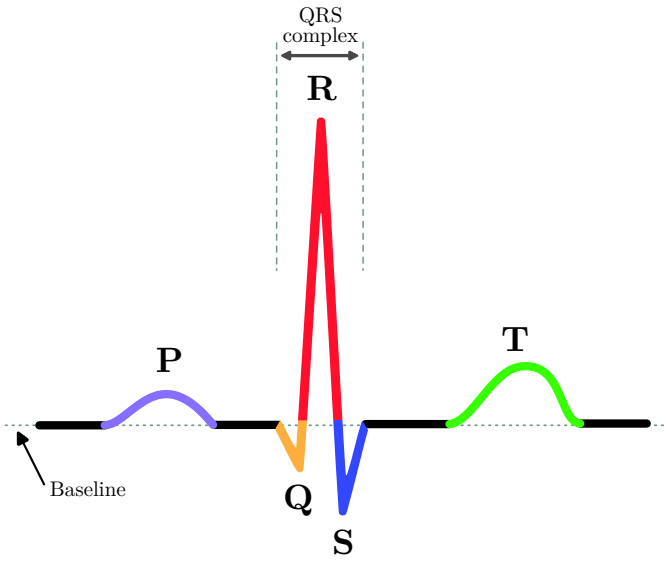


Figure 2.24: Sinus rhythms, a heartbeat perceived by the ECG.

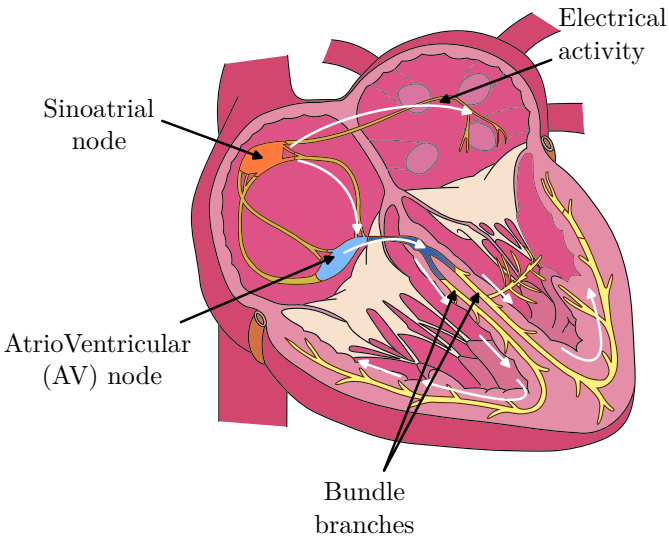


Figure 2.25: Heart conduction system.

2.3.1 Leads placement

To measure the electrical activity produced by depolarization of the cardiac muscles, electrodes (leads) are placed in different positions on the subject. In a typical ECG recording, 10 electrodes are placed, subdivided into two categories:

- 4 limb lead electrodes (RA, LA, RL, LL).
- 6 precordial electrodes (V1, V2, ..., V6).

The placement of these electrodes is shown in Figure 2.26. In Figure 2.26.a a cross-section shows the recording area of the six precordial electrodes. These six electrodes provide a horizontal view of the heart. In Figure 2.26.b six other measurements are shown, which provide a vertical view of the heart. With these 12 lead-ECG, different views of how the depolarization wave moves through the heart are recorded.

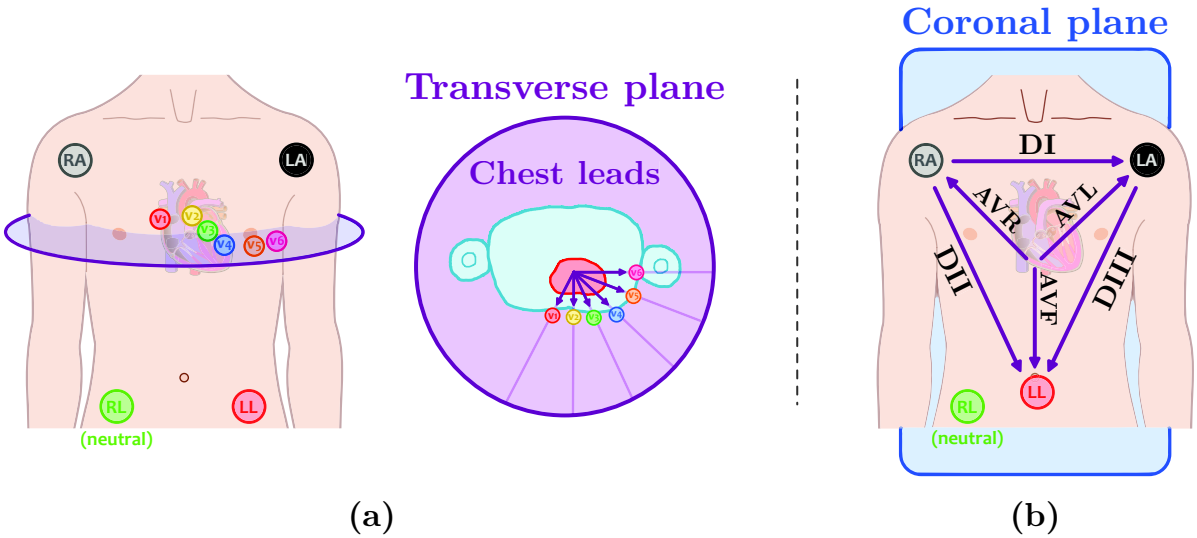


Figure 2.26: Leads placement. **(a)** Placement of precordial electrodes along with the measured sides. **(b)** Placement of limb electrodes.

The next chapter will cover the first application of the methodology, which is dedicated to the diagnosis of Parkinson's disease. In the latter, the implementation of what we have covered in the Chapter 1 will be put into practice.

3 Application 1: Parkinson's Disease diagnosis

Contents

3.1 Introduction

Parkinson's Disease (PD) is a chronic neurodegenerative disorder affecting more than 6 million persons worldwide as reported by the World Health Organization [103, 21]. PD is known by the general public for its motor symptoms such as: tremor at rest, rigidity, bradykinesia, etc. [103, 5], however, non-motor symptoms may accompany or precede the onset of motor symptoms, sometimes even arriving 20 years before the onset of the latter [16, 44].

Currently, the diagnosis of Parkinson's disease is entirely clinical. Medical professionals or clinicians assess patients' medical condition based on a thorough assessment of the patient's symptoms, medical history, and physical examination. The symptoms on which the diagnosis is often based are motor symptoms [9]. However, as we have already mentioned, these may appear after the non-motor symptoms which are not specifically associated with Parkinson's disease. Leaving the patients in suffering, raising the urge to look for new biomarkers allowing the early diagnosis of PD.

The aim of this work is not to provide a replacement for healthcare professionals, but

to offer a data-driven methodology that complements and assists these professionals. The application is entirely based on real electroencephalogram (EEG) signals, with the aim to separate Parkinson's disease patients from healthy subjects. The main objective is not to have the highest accuracy at the expense of the repeatability and significance of the results, but rather to have a valid method that is statistically significant, with the least bias possible, and that can be used in real-world scenarios.

The method presented in Chapter 1 makes it possible not only to deal with the problems of EEG signals, which are known to be very noisy, but also to generate new features that are much more faithful to the EEG generation mechanism, as well as being more informative than the classic features.

This chapter begins with a general presentation of the disease in Section 3.2, followed by an examination of the current state of diagnosis, whether clinical or non-clinical. The dataset used within the scope of this application is then detailed, followed by the implementation of the methodology outlined in the Chapter 1. The obtained results are presented in detail, accompanied by various tests of validity and robustness. Additionally, an evaluation in a more restricted context with less learning data is proposed. Finally, our work on this application is concluded by providing a synthesis of the key elements.

3.2 Generalities on the Parkinson's Disease

Dopamine deficiency is the main cause of Parkinson's disease [103, 5]. Dopamine is a neurotransmitter that enables communication[1] within the nervous system, and is particularly involved in regulating movement. A lack of dopamine has a direct impact on an individual's behaviour, reflexes and movements. The dopamine deficiency is caused by the disruption and degeneration of the dopaminergic cells located in the pars compacta of the locus niger [103, 5], the area of the brain where dopamine is produced.

The cause of the degeneration of dopaminergic neurons remains unknown, however, researchers have conjectured that the presence of Lewy bodies is the main trigger of this degeneration [34].

Epidemiology

PD is the second most common neurodegenerative disorder after Alzheimer's disease. According to the World Health Organisation's latest technical report, 5.8 million people worldwide are affected by Parkinson's disease (statistics from 2019). It is also reported that worldwide, disability and mortality attributable to Parkinson's disease are increasing more rapidly than for any other neurological disorder, up to double the rate recorded in

[1] Has also a role in regulating the feeling of motivation and reward.

the 2000s [70]. The average onset of Parkinson's disease is between the ages of 55 and 65, with the possibility of onset at any age. The number of people affected by the disease increases with age, reaching 2% over the age of 65. For unknown reasons, men are slightly more affected than women.

Causes

Parkinson's disease is considered to be idiopathic, i.e. the exact cause of the disease remains unknown. Studies show that 5% to 10% of cases are genetic. Researchers suspect that environmental factors increase the risk. These include exposure to pesticides, solvents and air pollution throughout life [70].

The next section is devoted entirely to Parkinson's disease symptoms. Both motor and non-motor symptoms will be discussed, as both are important in clinical diagnosis.

3.2.1 Symptoms

Parkinson's disease is known to the general public for its motor symptoms (tremors). However, non-motor symptoms may accompany or precede these symptoms. It is important to know that various symptoms are associated with the disease. Each patient has its own journey, and the course of the disease, the onset of symptoms and their amplitude differ among patients. This makes the clinical diagnosis of the disease even more complex.

Motor symptoms

At the onset of the disease, motor symptoms usually are unilateral, i.e. they affect only one side of the body. As the disease progresses, they become bilateral and affect the whole body. [93, 74, 103]. The following is a non-exhaustive list of the various motor symptoms:

- ***Bradykinesia*** and ***hypokinesia***: The patient will tend to make smaller and slower movements. Bradykinesia can be identified by: smaller handwriting, an expressionless face with difficulty in blinking, and a slow, hesitant gait.
- ***Resting tremor***: It can affect the hands, feet, chin, jaw and tongue. Disappears with voluntary movement.
- ***Stiffness***: The patient's muscles feel stiff and tighten involuntarily. It can occur in the arms, legs, neck, back, and even smaller facial muscles.
- *Akinesia*: Difficulty in initiating movements. Often observed when the patient is walking and then suddenly stops involuntarily. The blocking lasts about 10 seconds before the patient is able to resume walking.
- *Loss of balance.*

- *Swallowing difficulties*: as well as a difficulty in communicating verbally (speaking in a low tone).

Symptoms in bold are often those on which the movement disorder specialist relies for the clinical diagnosis of PD.

Non-motor symptoms

As already mentioned, these symptoms can appear many years before the onset of motor symptoms. The problem with these symptoms is that they are very bothersome on a daily basis, and their proper treatment often requires the correct diagnosis of the disease, which is fairly complex at this (early) stage [93, 74].

- *Mild cognitive impairment*: Memory deterioration, slowed thinking with difficulty in following a conversation.
- Mental health issues such as: *depression*, *anxiety*, *hallucination*, and *apathy*.
- *Impaired sense of smell*.
- *Pain*.
- *Sleep disorder*: Due to muscle stiffness. Animated dreams with verbal and motor agitation.
- *Constipation*.

As one might observe, most of these symptoms are fairly recurrent in the life of a healthy person. Furthermore, it is important to distinguish between the symptoms' onset and normal ageing that can lead to these symptoms.

Figure 3.1 illustrates the time course of Parkinson's disease progression starting 20 years prior to diagnosis and 20 years after (adapted from [44]).

The next section will be devoted to the clinical diagnosis of Parkinson's disease, how it is carried on in practice, and on what it relies. Finally, some statistics on this clinical diagnosis accuracy are given, as well as the difficulties encountered in diagnosing Parkinson's disease in its early stages.

3.2.2 Diagnosis

Currently, the diagnosis of Parkinson's disease is entirely clinical. Medical professionals or clinicians assess patients' medical condition based on a thorough assessment of the patient's symptoms, medical history, and physical examination. The symptoms on which the diagnosis is often based are motor symptoms (section 3.2.1) [9]. However, as we have already mentioned, these may appear after the non-motor symptoms which are not

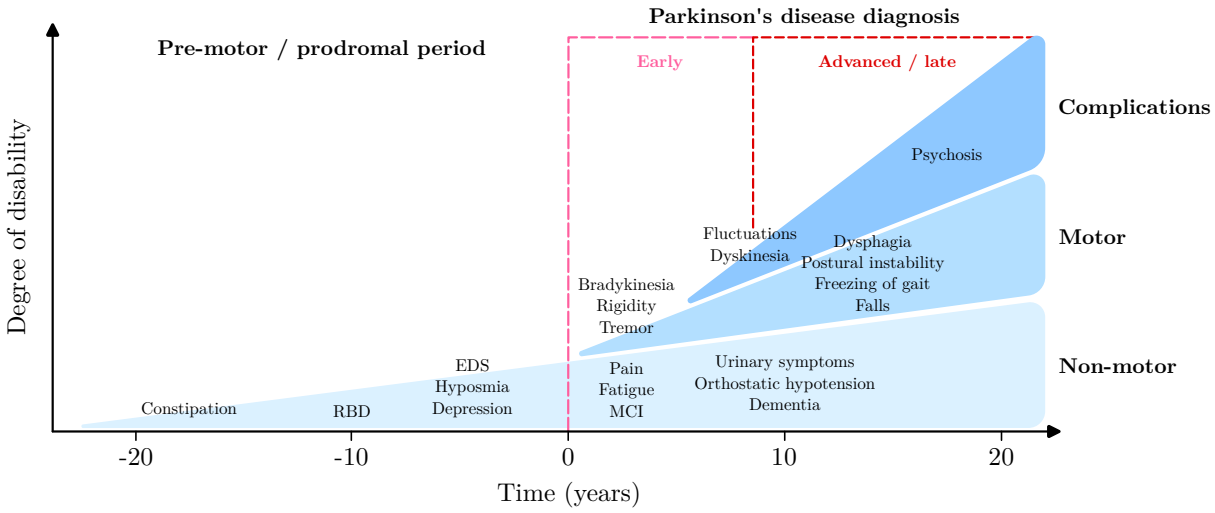


Figure 3.1: Time course of Parkinson's disease progression along with their symptoms. Abbreviations: EDS: Excessive Daytime Sleepiness; MCI: Mild Cognitive Impairment; RBD: REM sleep Behaviour Disorder (adapted from [44]).

specifically associated with Parkinson's disease. Leaving the patients in suffering, raising the urge to look for new biomarkers allowing the early diagnosis of PD.

Various tools and guidelines helping clinicians with diagnosis criteria have been developed (such as the one from the United Kingdom PD Society Brain Research Center [33]). The overall clinical diagnosis accuracy is in the order of 75 % according to the World Health Organization [103] or around 79 % following [80]. It is very important to note that the clinical diagnosis accuracy did not significantly improve during the last years particularly in the early stages of the disease where the response to dopaminergic treatment is not clear and less prominent [80]. Actually, during the early disease manifestation (<5 years of disease duration) the clinical diagnosis accuracy is around 53 % and even lower, around 26 % accuracy, for patients with <3 years disease duration [1].

Once the diagnosis has been established, treatment of the disease follows, which is the subject of the next section.

3.2.3 Treatment

The dopamine deficiency caused by the degeneration of dopaminergic cells is the main cause of Parkinson's disease. The treatment consists mainly of delivering dopamine to the brain. The dopamine can be administered in a direct form (as dopamine)[2] or indirectly (other substances that the body will convert into dopamine later on).

Taking dopamine directly orally is problematic because it has to pass through the bloodstream before reaching the brain. This has the effect of disrupting the entire func-

[2] Currently being studied on monkeys affected by PD [61]

tioning of the body, as it responds to this change in hormones. This undesirable effect on other parts of the body makes it impossible to administer dopamine by injection or orally (pills). In practice, only indirect administration of dopamine is used. The combination of Levodopa + Carbidopa remains the most widely used treatment in practice. Carbidopa inhibits the conversion of Levodopa into dopamine outside the brain, but once in the brain, Carbidopa loses its effect, allowing Levodopa to be converted into dopamine.

Despite substantial development of drugs to treat PD, none has proven to be effective in slowing progression. Levodopa remains the most effective drug for improving most motor and various non-motor symptoms of PD and for improving functioning and quality of life, although it cannot halt the neurodegenerative process as shown in Figure 3.2.

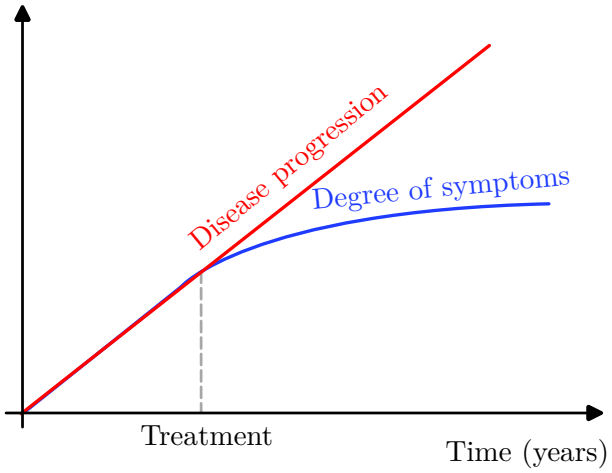


Figure 3.2: Unchanged disease course after treatment, but treatment effectiveness in reducing the severity of symptoms.

The clinical diagnosis of Parkinson's disease has been discussed, and now we will move on to data-driven diagnosis methods that are based on electroencephalogram (EEG) signals. The latter being widely used in practice as the patients' condition is directly reflected on the brain electrical activity. In the following section, the state of the art in EEG-based diagnosis of PD will be detailed.

3.3 State of the art of EEG-based PD diagnosis

Parkinson's disease diagnosis using EEG has been studied in several works. [12] uses a selection of Fourier transform coefficients to achieve a maximum accuracy of 82 %. It is to be noted that in our study we use the same data as the former. In [67], the authors propose a fully automated approach based on a 1-Dimensional Convolutional Neural Network (1-D CNN). The model directly classifies the temporal EEG epochs achieving an accuracy of 88.2 %. To perform the diagnosis, [10] relies on correlation coefficients calculated between channels as well as the coefficients of an AR model identified on the EEG to yield a presumable accuracy of 99.1 % see below for a discussion on the relevance of

such results. [108] uses high-order spectra to perform the diagnosis by extracting thirteen features from the EEG frequency spectrum, he achieved a presumable accuracy of 99.25 %. The work in [38] uses the coefficients of an AR model and the wavelet packet entropy to analyse and investigate whether there is a difference between Parkinsonians and healthy individuals with no attempt to separate the subjects. Finally, [52] utilises entropy-based features of 10 channels and a three-way decision model to obtain a classification accuracy of 92.9 %. This last study would have been more relevant if the author addressed the problem of unbalanced data-set.

We note that the majority of studies are based only on the frequency features of the EEG and that few studies focus on the temporal features while the two domains should complement each other. Only a few of the features used are explainable and we can understand their design basis to derive conclusions for future work. Most importantly, there has been no study of the significance of the results obtained, given that in the majority of studies the number of features used is very large in relation to the number of patients present in the data set used, which accentuates the problem of the curse of dimensionality discussed in Section 1.5.1, thus, accentuating the problem of the significance of the findings.

We strongly believe that some of the above-mentioned methods [12, 67, 10, 108] are subject to data leakage problems. Data leakage is defined as the use of information in the model training process that is not supposed to be available at the time of prediction [45] (presented in more detail in Section 1.6). This would not be possible in a real-life scenario, where we receive new samples of unlabelled data that we need to categorize. The first type of data leakage that some of the proposed methods suffer from is group leakage, where correlated data from the same subject are present in both the training and the test sets [4]. In this case, and using limited amounts of data, a complex model such as the 1-D CNN can even identify the subject's signature (one can see example in Appendix 6.3). The second type of data leakage comes from the fact of optimising hyper-parameters and performing feature selection directly on the test-set (therefore an absence of a validation set) [45].

3.4 Sparse Dynamical Features applied to Parkinson's Disease diagnosis

This section is entirely devoted to the application of the method presented in Chapter 1. Before moving on to the application, the dataset used is firstly described, along with the motivations that led us to this data-set choice. Next, the pre-processing (filtering, cropping, etc.) applied to the data will be described in Subsection 3.4.2, followed by an important passage addressing Event-Related Potentials (ERP). The values of the selected hyper-parameters of the proposed methodology are given, and finally, all that has been discussed previously (methodology, feature selection, evaluation, etc.) is combined in order to produce the results obtained. The obtained results are subjected to various validity and

constraint tests, which will also be discussed at the end of this section. Results omitting the Sparse Dynamical Features generation step are also presented to demonstrate the effectiveness of the method.

3.4.1 Description of the Data-set

Several EEG datasets dealing with the diagnosis of PD exist, we examined these before selecting the one on which we can evaluate and test our method. As we seek a significantly large number of patients as well as a large amount of data, the following dataset: http://predict.cs.unm.edu *(ID: d001)* [13] was chosen. One particularity of this dataset is that the data were almost untouched (raw data) in the sense that few pre-processing steps have been applied to them. An additional reason for this choice is that the data-set is publicly available, which is important for transparency and reproducibility of results purposes. Moving further in this direction, we made our implementation code as well as the execution steps to reproduce the results accessible to the public through the link: https://github.com/HoussemMEG/SDF_PD. Additionally, an explanatory animation is included in the same link.

The experimental EEG data was recorded from $N = 50$ participants, 25 of whom were suffering from PD and an equal number of sex and age-matched participants serving as a control group (CTL). This matching is important to reduce the variability in results due to the age and gender of the subjects. The PD group were subject to the same experiment twice, once on medication and the other time off medication. In this thesis, only the off-medication sessions are considered as they showed a noticeable separability from the CTL group in comparison to the on-medication sessions.

The PD group were subject to a Unified Parkinson's Disease Rating Scale (UPDRS) [78] assessing the severity of their disease which was scored by neurologists, the mean UPDRS score is (24.80 ± 8.66). All participants underwent a Mini Mental State Exam (MMSE) [97], and all obtained a score above 26 (PD: 28.68 ± 1.03, CTL: 28.76 ± 1.05) confirming their ability to comprehend the task they will be subjected to. Complete details and information regarding the subjects and the experimental procedure can be found in [12].

The experiment consisted of a 3-Oddball auditory task, during which the subjects were presented with a series of 200 repetitive auditory stimuli (trials) infrequently interrupted by a deviant stimulus. Three types of stimuli can be distinguished:

1. *Standard* (70 % of the trials).
2. *Target* (15 % of the trials).
3. *Novel* / *Distractor* (15 % of the trials).

To avoid confusing stimulus types with words in the text, each time we refer to a stimulus type it will be written in italic.

During this task, the subjects had to count the number of *target* stimuli they had heard throughout the whole experiment. The auditory stimuli were presented for a period of 200 ms and were separated by a random Inter-Trial Interval (ITI) drawn from a uniform distribution of (500 — 1000) ms preventing subjects' habituation and anticipation. Figure 3.3 draws an example of an auditory stimuli sequence.

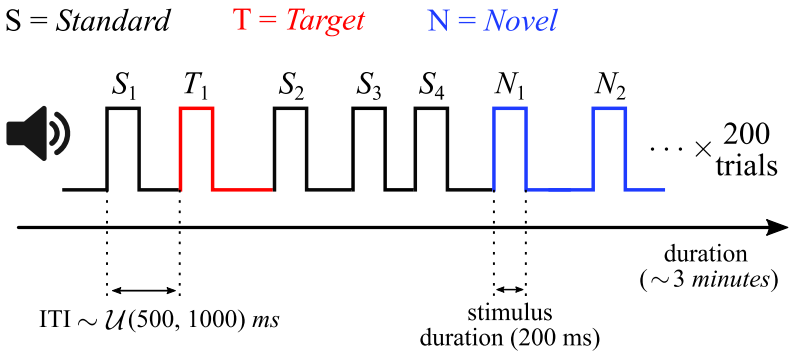


Figure 3.3: Example of a sequence of auditory stimuli.

The raw signals from a random person are shown in Figure 3.4. It can be noted that the signals are given on the entirety of the 60 channels with a high noise content.

3.4.2 Pre-processing

Throughout the experiment, the EEG signal was continuously recorded at a sampling rate of $f_s = 500$ Hz by the mean of 64 electrodes (channels). Very ventral temporal sites were removed by [12] as they tend to be unreliable, leaving at the end 60 channels. The data were then re-referenced to an average reference.

As mentioned in section 2.2.4, EEG signals are known to be very noisy and present many practical difficulties. Indeed, the coveted brain activity is of a low amplitude and is often drowned out by ambient noise, making the pre-processing stage mandatory. Despite the intrinsic complexity of EEGs and their noise content, the pre-processing steps we have applied are very mild due to the fact that our method is robust to noise. Firstly, to separate and disentangle the unwanted, high-amplitude ocular activity from the coveted cerebral activity, we conducted an Independent Component Analysis (ICA) on the data [96]. We analyzed each independent component (IC) of each subject individually, the ICs that contained eye blinking and/or very strong[3] artefacts were removed by projection following the guidelines and recommendation of [54] and [12]. Secondly, the data were then bandpass filtered using a Hamming window, attenuating the frequencies outside the (1 — 30) Hz interval. This frequency interval was selected because many studies take 20 or 30 Hz as the upper filtering limit [91]. We have taken the widest interval knowing that our method remains valid even if we widen this interval further.

[3] Which we can visually see and which affects the entire signal.

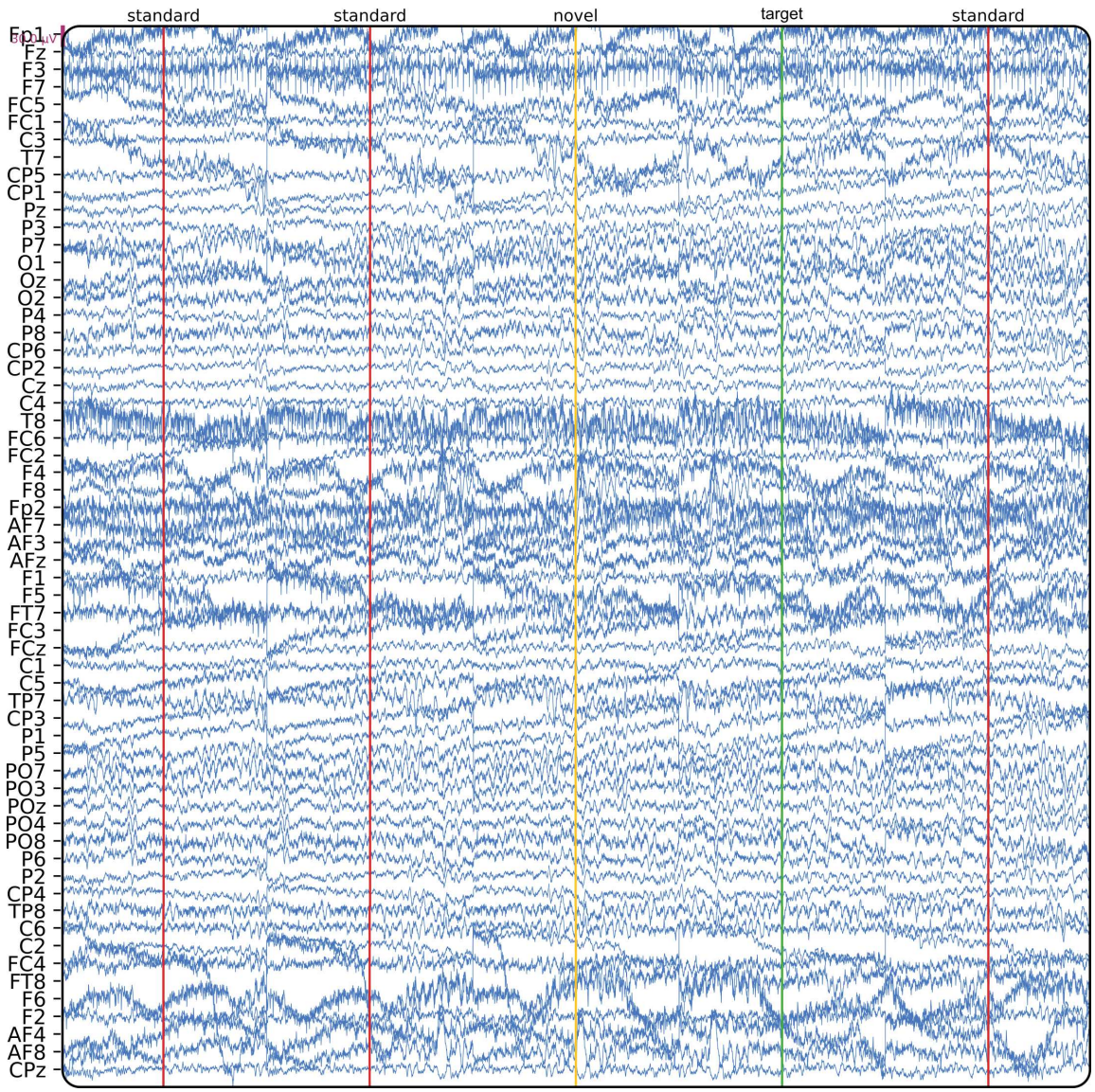


Figure 3.4: Example of a raw EEG signal composed of 5 trials and 60 channels.

Figure 3.5 shows the signals after applying the pre-processing to the same patient illustrated in Figure 3.4.

Time windows (segments) starting from stimulus onset (0 ms) up to (+500 ms) post-stimulus were formed, resulting in 200 time-locked segments, one for each stimulus (see Figure 3.6 for a more detailed graphical representation). An event-related potential (ERP) was also calculated separately for each stimulus type by vertically averaging all the signal segments corresponding to the same stimulus type and channel (see Figure 3.7) [55]. The aim of this step is to filter the signal and sum up the events occurring at the same time to make them stand out from the ambient noise. Moreover, all the pre-processing steps were performed using MNE[4] version 0.24.1 [35].

[4] An open-source Python package for exploring and analyzing human neurophysiological data.

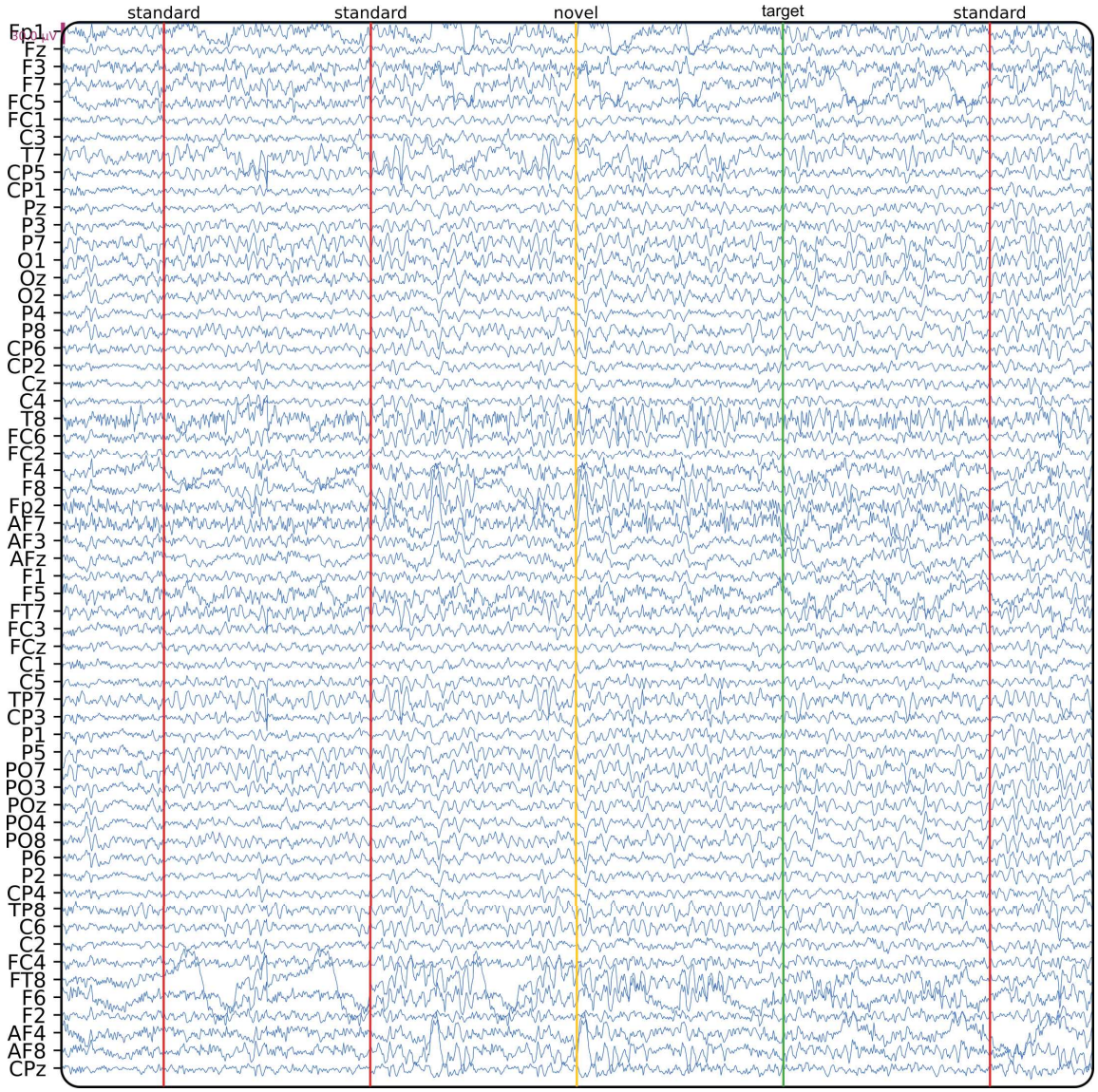


Figure 3.5: Previous EEG signal after applying the pre-processing.

3.4.3 Application of the proposed method

For each subject, we computed the ERPs corresponding to each stimulus type and channel. The window width used ranged from 0 ms to 500 ms post-stimuli[5], resulting in an ERP of length $L = 250$. The system Σ was created using $m = 40$ oscillators with an evenly spaced angular frequency taken from the interval $2\pi \times (1 \text{ to } 30)$ Hz following the procedure presented in Section 1.4.5. The weighting constant w in the equation (1.18) has been set to $w = 0.55$ following the procedure presented in Section 1.4.3. Figure 1.12 shows the trade-off of the separation between the forced and steady-state regime using

[5] The upper limit of 500 ms was chosen due to the minimal interval between stimuli min(ITI) = 500 ms.

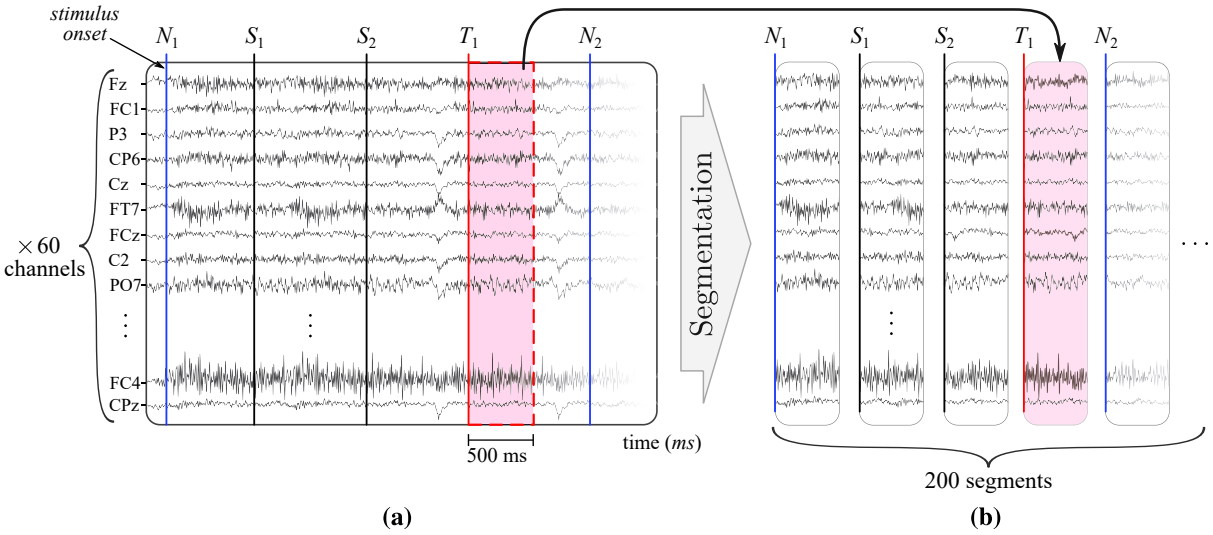


Figure 3.6: EEG segmentation. **(a)** Raw and continuous 60-channel EEG, recorded for a duration of ∼3 minutes. Stimuli are time-locked to their arrival instant. Segments of 500 ms were taken starting from the stimulus onset. **(b)** The result of the segmentation: a 60-channel signal divided into 200 time windows, each time window corresponds to one stimulus response.

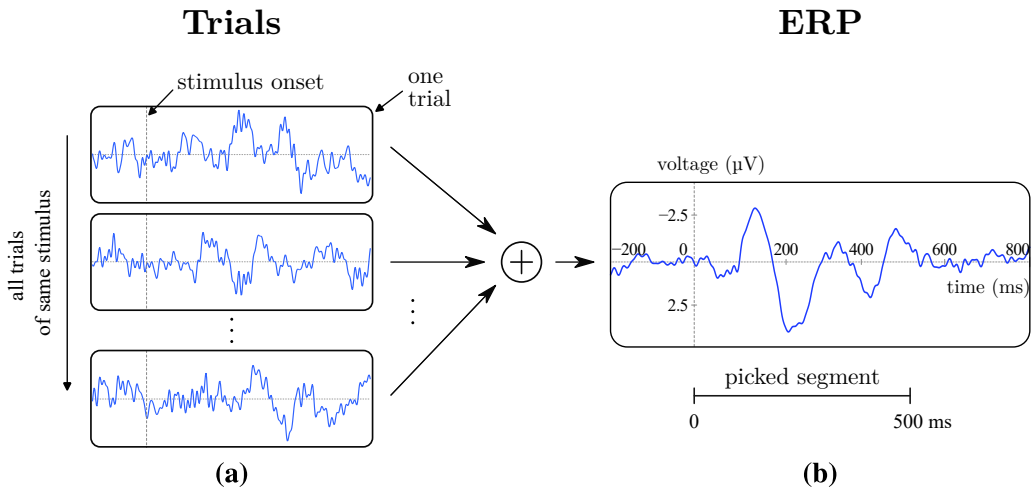


Figure 3.7: Example of ERP creation process for the *Standard* stimulus on one specified channel. **(a)** Vertical averaging of the corresponding time-locked EEG segments. **(b)** Result of the vertical averaging process. The signals length used in this example is only demonstrative, the true window length used in our method is indicated at the bottom.

the data described above.

The evaluation metric used is the accuracy, which is defined as the number of correct diagnoses divided by the number of decisions to be made. In the case of multiple repetitions, the stated accuracy is the mean value of the accuracy obtained over all folds and all repetitions.

The value of α_f was set following the procedure presented in Section 1.4.2 and the value obtained is in the order of $\alpha_f = 8 \times 10^{-4}$. The n resulting β solutions were taken and mapped as presented in Section 1.4.2 and Figure 1.11 with a regular spacing[6] $l\,\% = 2\,\%$, resulting for each ERP in 50 β variants with a decreasing sparsity levels.

To show an example of how a real EEG signal is fitted by our model, Figure 3.8 shows varying fit level for a *target* stimulus signal taken from the Cz channel. It is important to note that the contribution of x_0 is taken into account.

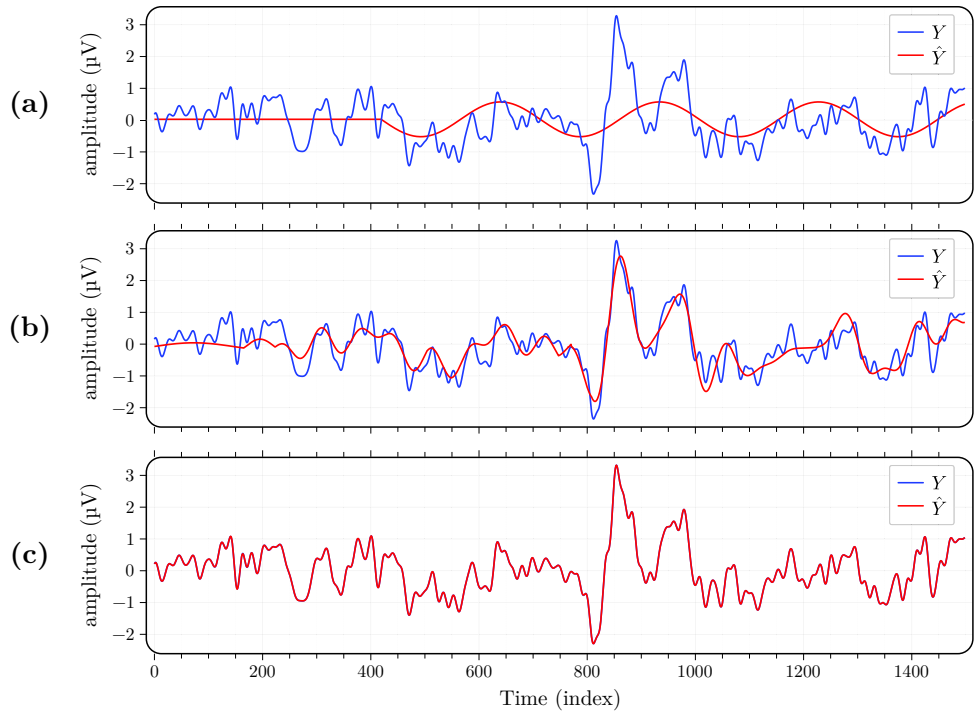


Figure 3.8: Example showing the fit of our model with different levels of sparsity. **(a)** A high level of sparsity. **(b)** A medium level of sparsity. **(c)** A low level of sparsity.

For the same signal as above, Figure 3.9 shows the separate contribution of U and x_0 to the model output. It can be seen that most of U activity takes place after the arrival of the stimulus. Nonetheless, we do observe some activity in the forced regime before the stimulus arrives, although it remains low in amplitude.

3.4.4 Features extraction

Studies have shown that patients with PD tend to be significantly slower than healthy individuals (latency observed in auditory ERP) [39, 25]. They also tend to have lower ERP amplitude over certain brain regions compared to CTL individuals [73, 25]. However, even if these differences are observed between PD and CTL groups, the last two biomarkers extracted directly from the ERP do not permit the discrimination between the two groups

[6] This is the value that induces the least solution discard or duplication (see Step 1 in Figure 1.11).

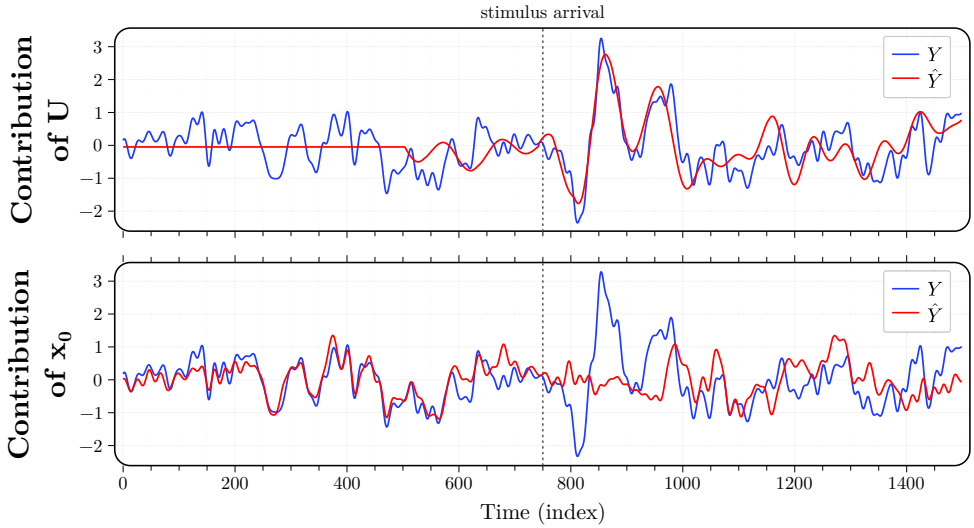


Figure 3.9: Separate contribution of U and x_0 to the output $\hat{Y}$.

with good separability and accuracy. This is mainly due to the large variance in latency and amplitude within the two groups, as well as the high sensitivity of these two indicators to the disease duration and severity [25]. We believe that the intra-group variance is due to the high noise content of the EEG, which makes the extracted measures very sensitive and variable among patients of the same group.

With the observed latency in the ERP and lower amplitude of parkinsonsians compared to healthy individuals, we moulded the information contained in the generated SDF to fit these discriminative features. The average excitations forces (mean) was used to replace the amplitude. The latency that is defined by the time elapsed from stimulus arrival to the arrival of the largest or smallest spike[7] is now replaced by its analogue: the arrival time of the lowest (in an algebraic sense) oscillator excitation (argmin).

For these two features, we have found the operating time intervals I_1 and I_2 that gave the best results and are shown in Figure 3.10:

- F_1: the instant of the lowest activation amplitude of all oscillators defined by: $\text{argmin}\Big\{U(I_1)\Big\}$.
- F_2: the average excitation forces occurring between 180 ms and 500 ms defined by: $mean\Big\{U(I_2)\Big\}$.

Only the two features presented above were tested initially. For the sake of completeness, all the features listed in Table 1.1 were also tested individually, without giving

[7] The sign of the peak depend on the channel considered, since the head can be considered as a dipole, on certain channels the value of the peak is positive whereas it is negative on other channels (see Figure 3.10 to observe the symmetry of the peaks).

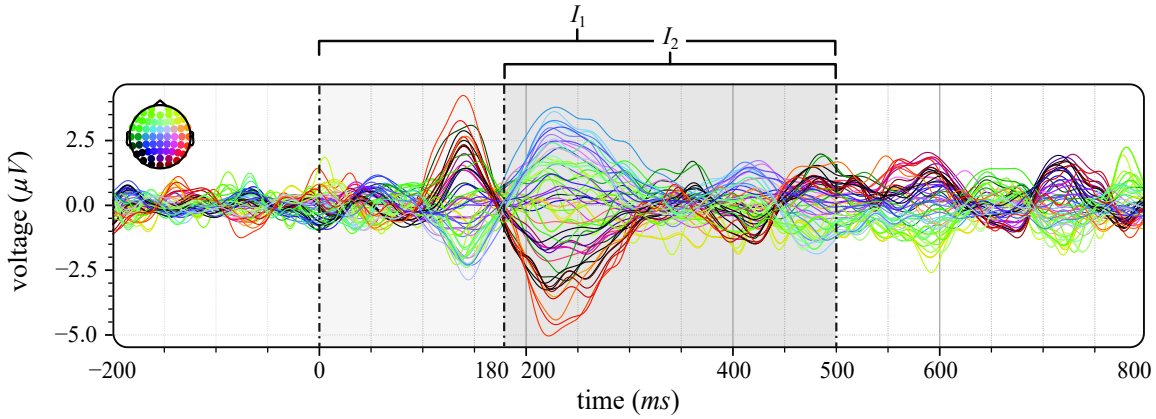


Figure 3.10: Example of one subject *standard* ERP over the 60 channels. Features operating intervals are also reported.

convincing results. Frequency selection on certain bands of interest[8] was also tested, again without any convincing results. In the remainder of this chapter, we will focus solely on the features F_1 and F_2, which were used to perform the classification using LDA classifier.

At the beginning of this Section 3.4.4, we mentioned that the two bio-markers (ERP amplitude and latency) did not give significant results due to the fact that there is a large variance even intra-group, whereas now analogous features are used? Contrary to what was previously done, the features now are extracted from the SDF instead of the pre-processed EEG. With the help of the proposed method we were able to obtain significant results for the equivalent features. We believe that this is due to the fact that the method allows us to retain the right amount of information needed for the classification[9] while avoiding capturing a lot of noise. Moreover, we believe that separating the forced regime and the free regime from the signal helped since the free regime did not bleed into the features (as a reminder, F_1 and F_2 were extracted from the U part only). More details will be given in Section 3.4.5 where we compare the same features with and without SDF generation. Before getting to that, the results will be presented first, and will be the subject of the next section.

3.4.5 Results & discussion

The form and complexity of the ERP vary according to the type of stimulus and channel position. This variation is even more pronounced when the channels are far from one another. Our model utilises fewer modes excitation to fit a less complex ERP, while it will utilise more modes excitation to fit a more complex ERP. Therefore, the best working level of sparsity that yields the best classification accuracy varies according to the channel location. Since the best working sparsity level cannot be determined in advance and since

[8] We tested beta waves (13 — 30) Hz as suggested by [51] for its functional role in PD, delta waves ($< 4Hz$), theta waves (4 — 8) Hz, and alpha waves (8 — 12) Hz.

[9] Through the nested CV, which selects the appropriate level of fit required.

we are evaluating our method on all channels and stimuli types, we have left the sparsity level as a hyper-parameter to be tuned during the validation step as presented in the Cross-Validation Section 1.6.1.

As a reminder, for each split and each repetition of the outer loop, we use the inner loop to choose the sparsity level that performs best[10] on the validation set. Then, the outer loop training is carried out with the selected level of sparsity, followed by the evaluation of the method on the test set.

As a large number of combinations (stimuli type, channels and sparsity level) are being tested while using few data instances, in some cases due to sampling noise [30], a good validation accuracy[11] is observed for a given sparsity level $\alpha_{k\%}$. While the surrounding sparsity levels $\alpha_{k-1\%}$ and $\alpha_{k+1\%}$ yield much poorer results (an accuracy jump is observed). In this case, the sparsity level $\alpha_{k\%}$ is not considered a valid choice of hyper-parameter, since it is mainly due to luck and is not consistent. We have been careful to consider only the sparsity levels that yielded good validation accuracy, while the surrounding levels of sparsity yield equally good results. To do so, we used the following triangular shape filtering window: $[1, 4, 1]/6$ and convolved it with all the validation accuracies obtained, which were ordered according to their sparsity level. Then, the sparsity level that has the highest smoothed accuracy is selected. This accuracy smoothness is expected since the variation in sparsity levels is small. This ensures us that our results are not due to luck, and that they are sufficiently consistent and contain valuable information.

Stimulus choice

We started by evaluating the performance of our method for the different stimuli. For each stimulus, we evaluated the accuracy obtained by our model, for all the channels using the NLOO CV. Table 3.1 shows the results obtained. As the EEG has 60 channels we have only indicated the channel that yielded the highest accuracy (channel*) with its corresponding accuracy.

Stimuli type	*standard*	***target***	*novel*
Channel*	AF_z	$\mathbf{CP_z}$	CP_3
Accuracy	74 %	**90** %	74 %

Table 3.1: Results of the highest accuracy obtained for each stimulus type with the corresponding channel.

As indicate the Table 3.1 the stimulus *target* showed the highest separation using the features F_1 and F_2. These results, need to be interpreted with caution as other features

[10] Exhibits the highest accuracy.
[11] Over 80%

can lead to different results. Furthermore, the results obtained for the stimuli *standard* and *novel* are not statistically significant.

Please note that for the rest of this chapter only the *target* stimulus is considered.

Luck or informative features ?

To test the significance, truthfulness and consistency of the results obtained for the CP_z channel we performed three tests:

1. *Permutation test* [65]: We define the null hypothesis H_0 as: the features F_1 and F_2 do not allow to differentiate between the PD and CTL group. Under this hypothesis, we evaluated the probability that the obtained result was simply due to chance and sampling noise as we have a small data set and we carried out a large number of runs (each per stimuli type, channel and parsimony level) [30]. After randomly permuting the labels 500 times, and assessing the best accuracy obtained, we obtain a p-value of $\left(p < 0.03\right)$ which indicates and favours the fact that the results are significant. To recall, for statistical significance, the p-value should typically be $p < 0.05$ [26, Chapter 1].

2. *Parametric consistency*: For this part, unlike the other parts, the parsimony level is fixed and it is no longer a hyper-parameter to be tuned, thus the nested CV is no longer needed and only the outer-loop is used. For each level of parsimony, the model accuracy was assessed following a LOOCV procedure. Moreover, a new spacing $l\% = 1\,\%$ was used to have a finer sparsity grid. Figure 3.11.a shows the accuracies obtained at each sparsity level. The small variation of the accuracy in the parsimony zone that yielded the highest accuracy (see the area of interest in Figure 3.11.a) may suggest the presence of information and that indeed, by using F_1 and F_2 the subjects are separable. It should be noted that this area of interest is the area selected in $100\,\%$ of the times during the model selection (in the inner-loop of the NLOO CV). To illustrate how the features F_1 and F_2 are scattered on the plane, we have plotted these in Figure 3.11.a for the sparsity level indicated by a star in Figure 3.11.b. In addition, we have also plotted the decision boundary.

3. *Spatial consistency*: We have evaluated the spatial evolution of the accuracy for the channels surrounding CP_z. If these channels also show good results, this may indeed indicate the presence of information over this region, moreover, it coincides with the suggestions of Figure 7 in [73]. Figure 3.12 shows the accuracies obtained on the test-set after a NLOO CV procedure. We observe that the channels CP_1 and CP_2 located on the same horizontal line to CP_z showed great results individually, while the other surrounding channels did not yield significant results. We can observe an increasing accuracy as we get closer to CP_z, which may suggest the presence of discriminative information over that region.

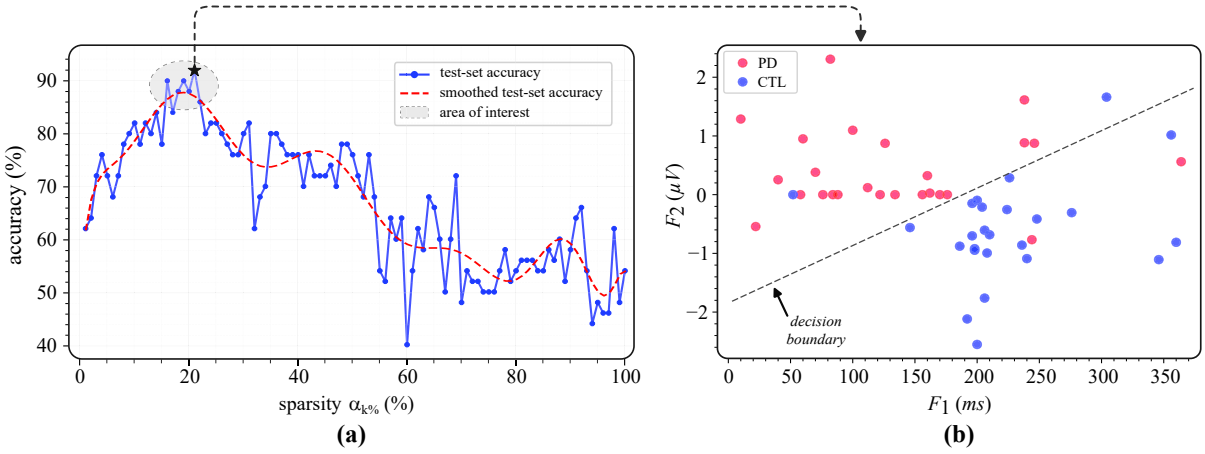


Figure 3.11: **(a)** Model's test accuracy for a varying level of sparsity using LOOCV. **(b)** Scatter plot of the features F_1 and F_2 with the corresponding decision boundary taken for the sparsity level indicated by a star in (a).

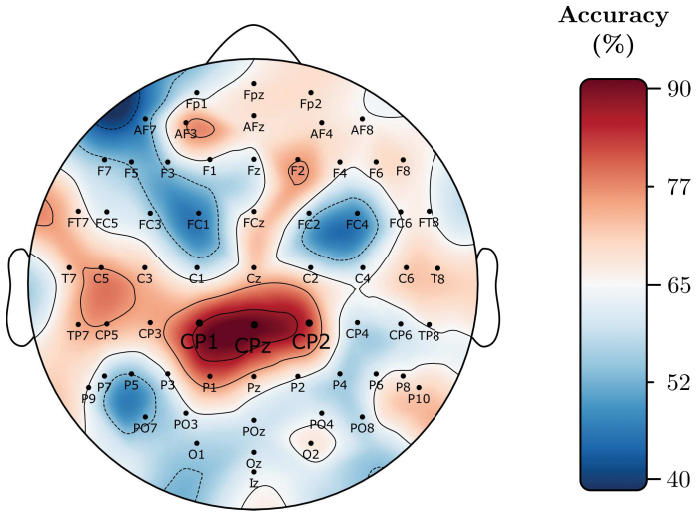


Figure 3.12: Model performances over the region surrounding CP_z.

Increasing the accuracy through voting

Observing in Figure 3.12 that the channels CP_1, CP_z and CP_2 performed well individually, we aggregated the information of these three channels by making them vote on whether the subject is healthy or non-healthy. All three channels have the same voting power, so if at least two voters agree on a subject's condition, we consider the resulting vote as the output of this new model. Table 3.2 reports the accuracies obtained after a NLOO CV procedure. The validation set is non-existent for the voting strategy as we have no parameters to tune.

Figure 3.13 represents the resulting confusion matrix of the voting strategy for the NLOO CV procedure. We can observe that the classification is balanced as the model mislabels both classes almost equally. Indeed, we can see that 1 subject out of 25 PD was wrongly categorized and that 2 subjects out of 25 CTL were wrongly categorized.

Accuracy	Individual channels			Voting strategy
	CP_1	CP_z	CP_2	
Learning	88.0 %	90.0 %	84.9 %	95.4 %
Validation	87.3 %	89.9 %	84.1 %	—
Testing	88.0 %	90.0 %	80.0 %	**94.0** %

Table 3.2: Resulting accuracies obtained for the voting strategy and the three individual channels CP_1, CP_z and CP_2 using the NLOO cross-validation.

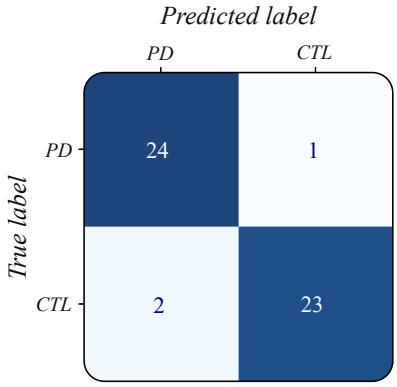


Figure 3.13: Voting strategy confusion matrix.

Regarding the early diagnosis, as presented in Section 3.2.2, in most cases patients suffer from non-motor symptoms before motor symptoms develop. The onset of the latter is slow and in some cases can take decades. The main difficulty of early clinical diagnosis is that it is mainly based on the manifestation of motor symptoms that have not yet developed sufficiently. To recall, the True Positive (TP) rate[12] of clinical diagnosis is about 26 % (9 of 34) for a duration since the first diagnosis <3 years, and of 53 % (8 of 15) for a duration <5 years [1]. Now regarding our data-set, we have 12 patients who have been diagnosed for <3 years and 2 patients for <5 years. For these 14 patients, we obtain a TP = 100 % strongly suggesting the utility of our method and its ability to work for early diagnosis cases.

Results with less learning data

For this part, we used Nested K-fold CV instead of the usual NLOO to evaluate the model performances[13] under the constraint of having fewer learning data. This step also gives an indication about the ability of our model to generalise to new, unseen data. We believe that if the model trained only by using half of the data performs almost the same as the model trained on the entire data-set it may suggest that the model trained on the

[12] Ratio of PD patients identified correctly.
[13] This only concerns the outer-loop, we kept LOO for the inner-loop.

entire data-set may obtain similar results on a data-set that is double the size of ours (under the assumptions that we have no covariate shift, for more details see [76]).

Table 2.3.b reports the accuracies obtained using a Nested 5-fold CV procedure where the learning-set is about 80 % of the entire data-set size. Table 2.3.c reports the accuracies obtained using a Nested 2-fold CV procedure where the learning-set is about 50 % of the data-set size. For both Nested 5-fold and Nested 2-fold CV the random partitioning was repeated 30 000 times. Table 2.3.a is a duplicate of Table 3.2, it was included to improve readability.

	Accuracy	Individual channels			Voting strategy
		CP_1	CP_z	CP_2	
(a) NLOO	Learning	88.0 %	90.0 %	84.9 %	95.4 %
	Validation	87.3 %	89.9 %	84.1 %	—
	Testing	88.0 %	90.0 %	80.0 %	**94.0** %
(b) 5-fold	Learning ($\pm std$ %)	88.7 % ($\pm$ 1.3)	89.7 % ($\pm$ 0.7)	86.5 % ($\pm$ 1.5)	94.1 % ($\pm$ 1.1)
	Validation ($\pm std$ %)	87.8 % ($\pm$ 1.2)	88.8 % ($\pm$ 0.7)	85.7 % ($\pm$ 1.4)	—
	Testing ($5-th$ %)	82.7 % (74.0)	85.4 % (79.7)	77.7 % (70.2)	**87.6** % (81.9)
(c) 2-fold	Learning ($\pm std$ %)	89.4 % ($\pm$ 2.5)	90.5 % ($\pm$ 2.1)	86.6 % ($\pm$ 2.1)	94.5 % ($\pm$ 2.1)
	Validation ($\pm std$ %)	88.3 % ($\pm$ 2.3)	88.9 % ($\pm$ 2.1)	85.7 % ($\pm$ 2.0)	—
	Testing ($5-th$ %)	78.0 % (67.9)	81.7 % (72.7)	73.9 % (65.7)	**83.8** % (75.5)

Table 3.3: Resulting accuracies obtained for the voting strategy and the three individual channels CP_1, CP_z and CP_2 using: **(a)** NLOO cross-validation. **(b)** Nested 5-fold CV (80 % learning data and 20 % test data). **(c)** Nested 2-fold CV (50 % learning data and 50 % test data).

From the Tables 2.3.a, 2.3.b, and 2.3.c we can observe that the accuracies obtained on the learning and validation sets have not changed significantly, but their variance increases as the learning-set gets smaller. Regarding the test accuracy, it decreased for the case of individual channels and for the voting strategy as the learning-set gets smaller which is expected as less information is present in the learning-set. However, the decrease in accuracy is not substantial, which demonstrates the robustness of our approach, which is mainly due to the utilization of only two well-separable features.

The test accuracy of the 30 000 repetitions is shown in Figure 3.14 with the $5-th$ percentile marked. This indicates that the accuracy of the model is above the indicated value in 95 % of the cases.

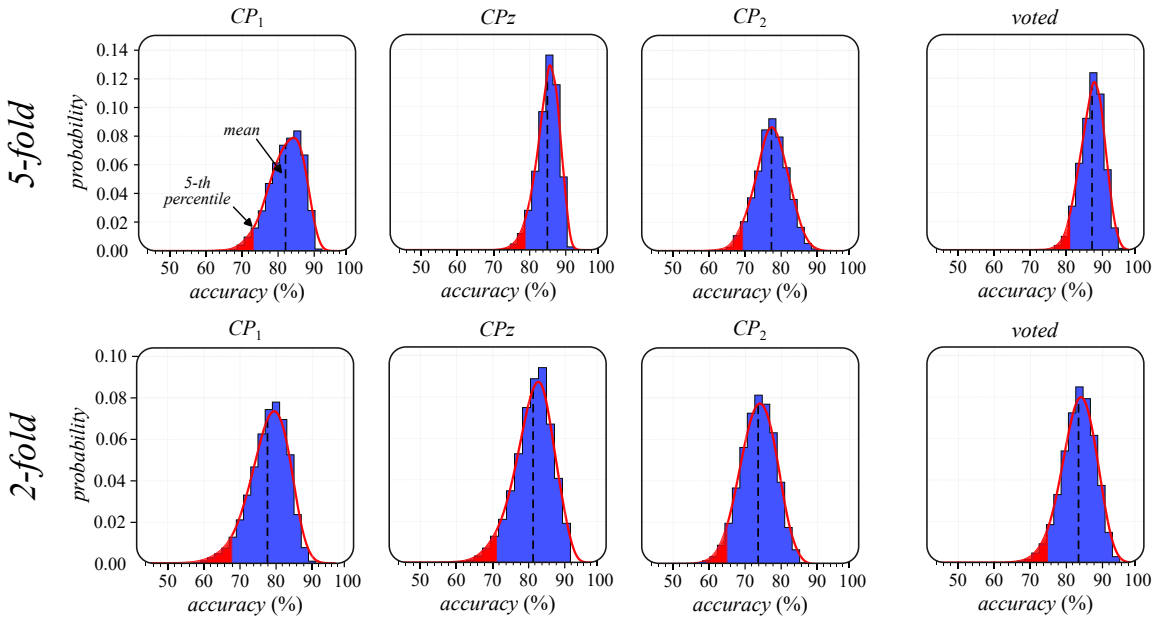


Figure 3.14: Resulting test accuracies obtained for the 30 000 random partitioning.

Omitting the feature generation step

In order to test the contribution of the Sparse Dynamical Feature generation to the results (blue block in Figure 1.1), we tried to omit it. Therefore, the features used in Section 3.4.4 will be used as they are defined and will be applied directly on the pre-processed signal Y and this for the same time intervals I_1 and I_2, as a reminder the extracted features will now be defined by:

- $F_1(Y)$: the instant of the lowest amplitude of the signal Y defined by: $\text{argmin}\Big\{Y(I_1)\Big\}$.
- $F_2(Y)$: the mean of the signal Y between 180 ms and 500 ms defined by: $mean\Big\{Y(I_2)\Big\}$.

The evaluation of the accuracy will be carried out in the same way, and the results obtained for a LOO procedure are indicated in the table 2.4.a. As a comparison, the results keeping the SDF generation block, are shown for the case of a NLOO procedure in the table 2.4.b.

Note that for the table 2.4.a, there is no validation as there are no hyper-parameters to select. We can observe that the accuracy with SDF generation performed better than without the SDF generation, both in training and in testing, for individual channels or the voting strategy.

We observe that in the case where there is no SDF generation, the accuracy during voting is quite low, and we did not benefit from the voting strategy. Indeed, the more confident voters are in their decision, the better the accuracy of the vote, as we can take advantage of everyone's good knowledge.

		Accuracy	Individual channels			Voting strategy
			CP_1	CP_z	CP_2	
(a)	***Without*** *SDF generation*	Learning	73.6 %	81.9 %	77.2 %	78.0 %
		Validation	—	—	—	—
		Testing	72.0 %	82.0 %	74.0 %	78.0 %
(b)	***With*** *SDF generation*	Learning	88.0 %	90.0 %	84.9 %	95.4 %
		Validation	87.3 %	89.9 %	84.1 %	—
		Testing	88.0 %	**90.0** %	80.0 %	**94.0** %

Table 3.4: Resulting accuracies obtained for the voting strategy and the three individual channels CP_1, CP_z and CP_2. **(a)** without the SDF generation in a LOO procedure. **(b)** with SDF generation in a NLOO procedure.

Nevertheless, we observe that the accuracy remains reasonably high. This shows first and foremost, that the selected features (F_1 and F_2) are not specific to our method and that they can be used on preprocessed signals directly without further manipulation. This demonstrates the ability of our SDF generation to preserve the physical meaning of the signal it was applied to (as it was designed for), adding an element of explainability to our method. Moreover, it can be conjectured that the proposed method can be used to discover consistent while unnoticed temporal features for different diseases and experimental protocols.

We believe that our method has yielded better results because our transformation provides different levels of parsimony, offering varying levels of fit to the noise and relevant information. The method thus acts as a filter that retains only the necessary information. The intrinsic functioning of the method, which resembles the functioning of the brain, provides a new perspective (a new point of view) on the data, which allows us to separate patients from non-patients with good accuracy.

3.5 Conclusion & perspectives

In this chapter, an overview of Parkinson's disease generalities was given first. We have seen that the main cause of PD is dopamine deficiency, which has a direct impact on the regulation of the motor system. This deficiency is caused by the degeneration of dopaminergic cells, yet the cause of this degeneration remains unclear. Then, the topic of symptoms was addressed, and we saw that even if Parkinson's disease is known to the general public for its motor symptoms (tremors), non-motor symptoms may accompany or precede by several years the motor symptoms. These non-motor symptoms are very bothersome and burdensome for patients, and unfortunately, as we have seen, they can occur repeatedly even in the life of a healthy person. This makes the clinical diagnosis of

the disease very complex.

Currently, the diagnosis of Parkinson's disease is entirely clinical, and is based mainly on the motor symptoms, which as we have already mentioned, can arrive very late. To help patients suffering on a daily basis, studies are being carried out to help and support healthcare professionals for the diagnosis of PD. These so-called data-driven studies are based on various supports, and in this chapter, we have focused on the most widespread one, which is the use of electroencephalography (EEG).

Electroencephalography is a method of recording the electrical activity of the brain using electrodes placed directly on the scalp. The principles of EEG recording and methods for reducing recording noise were covered first. Follow-up of the different possible electrode configurations and the electrodes' nomenclature. EEG signals are known to be very noisy, as they are contaminated by various artefacts that are entangled with the desired brain activity. These various artefacts were studied in order to be able to recognise and remove them using a method called Independent Component Analysis (ICA).

Most importantly, this chapter was dedicated to the application of the method presented in Chapter 1 for PD diagnosis. The method was evaluated on a publicly available data-set containing $N = 50$ subjects of which 25 have Parkinson's disease and the remaining individuals serve as a control group. The evaluation of the model was carried out with attention to induce the least bias possible so that we do not have an overly optimistic model that only works on this specific database, but rather have a simple model with a strong generalisation capability that can be used for new data samples. We were able to separate the healthy individuals from the unhealthy ones with an accuracy of 94 % using only two features extracted from the generated SDF and thus, by making the models obtained for the channels CP_1, CP_z and CP_2 vote. The different validity tests conducted suggest that the method and the results are correct and reliable, specifically the permutation test, with the resulting p-value of ($p < 0.03$) that favours the fact the results are statistically significant. The proposed new bio-markers worked also for the case where clinicians face the most problems i.e. the early diagnosis of the disease.

The simplicity and interpretability of the features extracted (F_1 and F_2) from the SDF are one of the strengths of the method. Indeed, since the SDFs are an image of the physical signal that was used to generate them, thus, the disease bio-markers can be directly captured by the SDFs. As a result, the features extracted from the SDFs are directly associated with the disease in question. So we can mould the extracted features according to our knowledge of the disease. This can be demonstrated by using the features F_1 and F_2 directly on the signals without generating SDFs, which yielded reasonably good results. These results are even better when the SDF generation block is used. This shows the interest of the method presented in Chapter 1, which allows us to go to the source of the signals that make up the EEG, thus giving access to a new point of view that we believe is more informative and more specific to the mechanism of cerebral signal generation. In addition, the hyper-parameter sparsity level, which allows the adjustment

of the degree of fit of the model to the signal of interest, acts as a filter that provides, through the nested CV, the right amount of information needed for classification without capturing noise, which for this application is very advantageous given that EEGs are very noisy. It is important to note that the method and classifiers used are very simple and computationally inexpensive.

Perspectives

We believe that a real validation with clinical tests is necessary in order to have an evaluation of the true performances of the method even if we made a lot of effort towards this point. In addition, we think it is important to have a much larger data-set with more clinical trials in order to be able to use more features and more sophisticated classifiers. It is also important to consider other similar brain diseases in the data-set, as we do not know if the other disease will have the same features as PD.

4 Application 2: Schizophrenia diagnosis

Contents

4.1 Introduction

Schizophrenia (SZ) is a chronic neuropsychiatric disorder that affects approximately 24 million individuals worldwide [32]. Schizophrenia is known by the general public for the psychosis symptoms or the so-called positive symptoms (e.g. hallucination, delusion, disorganized speech and thinking), however, other symptoms may also be connected to schizophrenia such as the negative symptoms (e.g., blunted affect), in addition to the cognitive impairments [72].

Currently, the diagnosis of SZ is entirely clinical [85]. Medical professionals assess patients' condition based primarily on the behavioural manifestation of positive symptoms, as well as on the evaluation of the patients' self-report of their own subjective experiences. The clinical diagnosis is difficult and tricky as schizophrenia shares symptomatic features with other disorders such as (bipolar disorder, severe depression, etc.) [31], furthermore, even a healthy person that does not require medical care may experience psychosis [50]. Various tools have been developed to guide and aid clinical diagnosis such as: the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [3] and the International Statistical Classification of Disease (ICD-10) [105] (published by the American Psychiatry

Association and World Health Organisation respectively). Although clinical diagnosis has been improved, it remains subject to human subjectivity [100]. Furthermore, cognitive dysfunctions may precede the onset of psychosis [71, 36], leaving patients suffering before being able to be correctly diagnosed. Hence a need for diagnostic techniques to alleviate these problems.

Data-driven methods aim to identify, directly from the data, new biomarkers of the disease. Different diagnostic directions are investigated, the main ones being: functional Magnetic Resonance Imaging (fMRI) [42, 110, 62], Electromyography (MEG) [24, 14] and Electroencephalography (EEG) [69, 46, 75, 109, 19, 86], each one with its advantages and drawbacks [87]. EEG have stood out for the diagnosis of SZ especially for its ability to record the dynamics of the brain underpinning sensory, cognitive, affective, and motor processes in response to a stimulus as well as for their high temporal resolution (in the order of milliseconds), facility of usage and price [17, 56]. These brain and motor processes are often affected by and directly correlated with the illness, and can even be observed before the onset of the latter [36].

The aim of this work is to propose a methodology designed to assist and aid healthcare professionals in diagnosing schizophrenia. The same methodology presented in Chapter 1 will be used for this new application with the aim to test its genericity in this new context. This application is also based on real electroencephalogram (EEG) signals. The objective is to separate patients suffering from Schizophrenia from healthy subjects with a desire to identify new biomarkers to perhaps shed light on the unknown biological reasons that cause the different symptoms. As with the first application, the primary goal is to have consistent and coherent results with as little bias as possible, which can be applied in real-world situations.

This chapter is structured as follows: Section 4.2 first presents the generalities on the disease of Schizophrenia, the symptoms and the way in which the diagnosis is currently carried out are the important points tackled. Section 4.2.4 briefly presents the state of the art of current studies that rely on EEG for the diagnosis of SZ and discusses the various problems that we have noticed. Section 4.3.1 details the experiment that produced the dataset used in this chapter, Section 4.3.2 describes the pre-processing that was applied before the dataset was made available online as well as the pre-processing that we applied ourselves. Section 4.3.3 initially presents the results obtained **using one feature** and without data cleaning, followed by an examination of the results' significance and trustfulness. Subsequently, additional results **using two features**, along with data cleaning, are presented in Section 4.3.5. Finally, Section 4.4 concludes this chapter and gives perspectives for future possible investigation.

4.2 Generalities on Schizophrenia

The word "schizophrenia" [from Greek 'schizo' (to split) and 'phren' (mind)] literally means "splitting of mind or personality" [22] but it rather refers to a fragmented pattern of thinking [59]. Schizophrenia is not a disease but it is a mental disorder or a syndrome with numerous symptoms that vary greatly from one individual to another. different between individuals. SZ is a complex and heterogeneous syndrome, meaning that it can manifest differently in different individuals and may involve varying combinations and severity of symptoms [43]. People with schizophrenia have a life expectancy 10 — 20 years below that of the general population [48].

As stated by the World Health Organisation [104], people with SZ often experience human rights violations by the community and even in mental health institutions. This intense and widespread stigma causes social exclusion and impacts their relationships with others, worse, including family and friends. The stigma contributes to discrimination which can limit access to general health care, education, and employment. The numbers are more alarming, as the vast majority of people with schizophrenia do not receive mental health care as only $\sim 31\%$ of people with psychosis receive specialist mental health care [102].

Epidemiology & prevalence

The most recent and largest prevalence study of Schizophrenia that we found is [15]. In this study, the prevalence of Schizophrenia is estimated by age, sex, year and for all countries. The study concluded Schizophrenia has an estimated prevalence of 0.28% with a 95% Confidence Interval (CI) of (0.24 — 0.31).

Figure 4.1 illustrates the prevalence rate of SZ as a function of age. (the figure has been adapted from [15], only aesthetic modifications have been made).

From Figure 4.1 it can be observed the onset of schizophrenia can start during adolescence and young adulthood. The prevalence peaks at around 40 years of age with a decline in the older age groups. Moreover, there is no observed difference in prevalence between males and females. In the same study, they concluded that the prevalence rates do not vary widely across countries and regions (One can see Figure 3 in [15]).

Causes

As stated by the World Health Organisation article on Schizophrenia [104]: "research did not yet identified one single cause of schizophrenia". The prevailing hypothesis suggests that the development of schizophrenia may be attributed to a complex interplay between genetic factors and a diverse range of environmental factors. Additionally, psychosocial factors may affect both the onset and progression of schizophrenia. Moreover, heavy use of cannabis is associated with an elevated risk of the disorder.

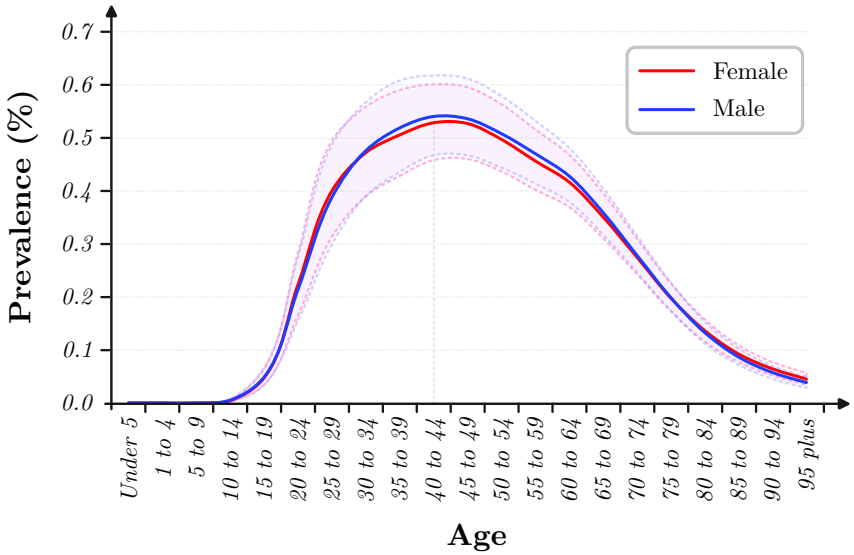


Figure 4.1: Worldwide mean prevalence rates by age and sex (with 95% CI interval).

4.2.1 Symptoms

The symptoms of Schizophrenia can be divided into 3 types [64, 104]:

Positive symptoms

Most human symptoms are an extreme version of a normal physiological process. However, the positive symptoms are new features that do not have any normal process in the counterpart. To take an example, a heart beating at a regular pace is considered a standard physiological process, in its extreme version, where the heart beats fast is tachycardia. The positive symptoms do not follow this rule of "extreme version of a normal physiological process". These positive symptoms include:

- Delusion.
- Hallucination.
- Disorganized speech.
- Disorganized behaviour.
- Catatonic behaviour.

Negative symptoms

The negative symptoms of schizophrenia can often appear several years before the patients experiences their first acute schizophrenic episode with positive symptoms. The negative symptoms often appear gradually and slowly get worse with time. Usually, the negative symptoms are a lack[1] of feelings that a healthy person have such as:

[1] hence the term "negative".

- Poor grooming or hygiene (not wanting to look after themselves).
- Changes in body language and emotions.
- Loosing interest in life and activities (avoiding interaction with people).
- Avolition (lack of motivation).
- Flat effect.

Cognitive symptoms

A cognitive deficit is also observed in patients affected by schizophrenia. As cognitive symptoms, one can find:

- SZ affects memory, learning, concentration, and understanding.
- Subtle difficulty to notice.

4.2.2 Schizophrenia phases

Schizophrenia tends to happen in episodes, during which, the patient cycles through all three phases in order [58, 20, 64]:

1. **Prodromal**: This early stage is often not recognized until after the illness has well progressed. Due to the fact that the first signs and symptoms of schizophrenia may be overlooked as they are common to many other conditions, such as depression.
2. **Active** (also called acute): During this phase, schizophrenia symptoms are usually obvious and distinctive from other conditions.
3. **Residual**: Symptoms in this phase of the illness resemble symptoms in the first phase. They are characterized by a lack of motivation and low energy, however, some elements of the active phase may remain. Moreover, patients may relapse back to the active phase after stopping their medication [58].

4.2.3 Diagnosis

As discussed previously in the introduction 4.1, the diagnosis of Schizophrenia is at present entirely clinical [85] and is based for the most part on the behavioural manifestation of positive symptoms which occurs during the active phase (when the symptoms are more prominent) [64]. This clinical diagnosis is tricky as SZ shares symptomatic features with other disorders and even a healthy person may experience psychosis [31, 50].

The important points concerning the diagnosis have already been covered in the Introduction 4.1, in what follows we would rather discuss a typical clinical diagnostic scheme and give more details on the Diagnostic and Statistical Manual of Mental Disorders

(DSM-V) for schizophrenia [3]. It is important to note that doctors must follow these guidelines when diagnosing SZ, the guidelines consist of the following elements [94]:

Criterion A. The patient has to exhibit two or more of the following, each present a significant amount of time (approximately 1-month period). At least one of these should include the first three elements (marked in bold):

(a) **Delusion**.

(b) **Hallucinations**.

(c) **Disorganized speech**.

(d) Disorganized or catatonic behaviour.

(e) Negative symptoms.

Criterion B. *Social or occupational dysfunction*: One or more areas of functioning such as work, interpersonal relations, or self-care are markedly below the level achieved prior to the onset and this, for a significant period.

Criterion C. *Duration*: The disturbance or dysfunction lasts for at least 6 months. This 6 month period must include at least 1 month of active phase (symptoms described in Criterion A).

Criterion D. *Exclusion*: The disturbance is not attributed to the direct physiological effects of a substance or another medical condition or no major depressive episode that occurred during the active phase.

Although clinical diagnosis has improved thanks to these guidelines, it remains subject to human subjectivity [100] and requires a quite heavy, expensive and meticulous observation in order to check the rules mentioned above. As previously presented in Section 4.2.1 cognitive dysfunctions may precede the onset of psychosis by several years [71, 36]. From the guidelines of (DSM-V) for schizophrenia diagnosis, it can be observed that the clinical diagnosis relies mainly on the manifestation of psychosis symptoms, leaving patients suffering before being able to be correctly diagnosed. To alleviate these problems, studies are underway to find new biomarkers for the disease, ideally at an early stage and preferably possible to check and exhibit through a rapid and repeatable process. EEG-based diagnosis falls in this category.

In the following section, more details about the current state of the art for SZ diagnosis based on EEG are given. The motivation of this choice has already been presented in the Introduction 4.1.

4.2.4 State of the art of EEG-based SZ diagnosis

The diagnosis of SZ based on EEG has been studied in several works. In [46], a robust variational mode decomposition (RVMD) is proposed in order to decompose the EEG

signal into different modes, and using six features computed on the decomposed signal they classify the SZ from the control group using an optimised extreme learning machine (OELM) to achieve a presumable classification accuracy of 92.9%. [75] extract two features corresponding to Kolmogorov Complexity and Sample Entropy from the Event-Related Potential (ERP) computed at four electrode locations. The features are separated using a Neural Network (NN) to achieve an accuracy of 91.2%. In [109], the authors proposed SZ classification based on the amplitude and latency of N1 and P2 components over the C_z channel combined with non-ERP related information (demographic) during a button-press task to generate a tone. The diagnosis accuracy was in the order of 81.1%. [19] used the mean ERP values computed between 400 — 600 ms post-stimulus during a visual task, these values were then averaged across four regions of interest (occipital, central, frontal, and parietal), and then were fed to Linear Discriminant Analysis (LDA) classifier to achieve an accuracy of 71%. [86] used 16 temporal and 4 frequency features extracted from ERP computed during a 3-oddball auditory task. Combining these features computer over 8 electrodes and by mean of a Multi-Layer Perceptron (MLP) they achieved a presumable accuracy of 93.4%. We strongly believe that the majority of the work we have seen suffers from various problems well known in the machine learning community, the same ones that have already been covered in Sections 1.6 and 3.3.

To recall, the problems raised previously concerned group leakage, optimisation over the test-set (see Sections 1.6 and 3.3 and Appendix 6.3), and being subject to the curse of dimensionality (see Section 1.5.1). The same remarks we raised previously are applicable and have been raised again for the works mentioned above. A further problem that has now been identified is the fact of not taking into account the data unbalance (numerical majority of a category). This unbalance induces a bias on the developed model as it will tend to favour the dominant category, thus biasing the evaluation and increasing the false positive or false negative rate. Concerning the curse of dimensionality, not only the statistical significance of the findings is questionable, but the associated features and choices lack explainability that might have compensated for the a priori significance issue.

4.3 The Sparse Dynamical Features generation framework applied to Schizophrenia diagnosis

This section is entirely devoted to the application of the method presented in Chapter 1 this time for the diagnosis of Schizophrenia. Before moving on to the application, the dataset used is firstly described in Section 4.3.1. Subsequently, the pre-processing (filtering, cropping, etc.) that the data had already undergone as well as those we applied will be described in Subsection 4.3.2. The values of the selected hyper-parameters of the proposed methodology are given in Section 4.3.3, and finally, all that has been discussed previously (methodology, feature selection, evaluation, etc.) is combined in order to produce the results obtained. It is important to note that the first results produced used one single feature and included all the subjects. The obtained results are subjected to

various validity and constrained tests that are discussed in Section 4.3.4. After an in-depth study, additional results are presented in Section 4.3.5 involving two features and where additional data cleaning was performed to eliminate subjects we considered to have poor signals.

4.3.1 Data-set description

The data used in the present chapter are publicly accessible (see: [81]). The study includes 81 subjects, of which 49 suffer from schizophrenia (SZ) and 32 individuals serve as a control group (CTL). Diagnoses were based on the structured clinical interview for DSM-IV[2] [7]. The subjects were age and gender-matched to reduce the effect of these two variables on the experiment. Furthermore, only right-handed subjects were included. The task that the participants were subjected to was a sensory task involving button pressing and/or an auditory tone. The stimuli are:

(a) **Stimulus 1** "Button Tone": subjects were asked to press a button at a regular pace, approximately every 1 to 2 seconds. For each button press, an immediate (without delay) delivery of a 1000 Hz tone is played.

(b) **Stimulus 2** "Play Tone": the sequence played in stimulus 1 was recorded with exact timing. The previously recorded tones were played back to the subject, who had to listen to them passively.

(c) **Stimulus 3** "Button Alone": here the subject were instructed to press the button at a similar pace, however, in this case, no sound was produced.

4.3.2 Pre-processing

The Data were recorded over 64 electrodes with a sampling rate of 1024 Hz. All the available data already underwent the following preprocessing before being made publicly available: (1) Re-reference to averaged ear lobes. (2) High-pass filtering at 0.1 Hz. (3) Outlier channels interpolation. (4) Epoching of the continuous signal and creation of 3 s time windows centred around the stimulus arrival. (5) Canonical correlation analysis to remove muscle and high-frequency white noise. (6) Outlier trial rejection. (7) Independent Component Analysis for removing unwanted artefacts. (8) Outlier channels interpolation within single trials. All the channels, single trial, and components outliers are defined as in [66]. The reader can refer to [81, 27] for the complete details about the experience. To give an idea of the state of the signals that are publicly available, Figure 4.2 shows an example of the recorded EEG signals after computing their ERP for each stimulus and channel.

[2] An old but similar version to DSM-V that has been discussed in Section 4.2.3.

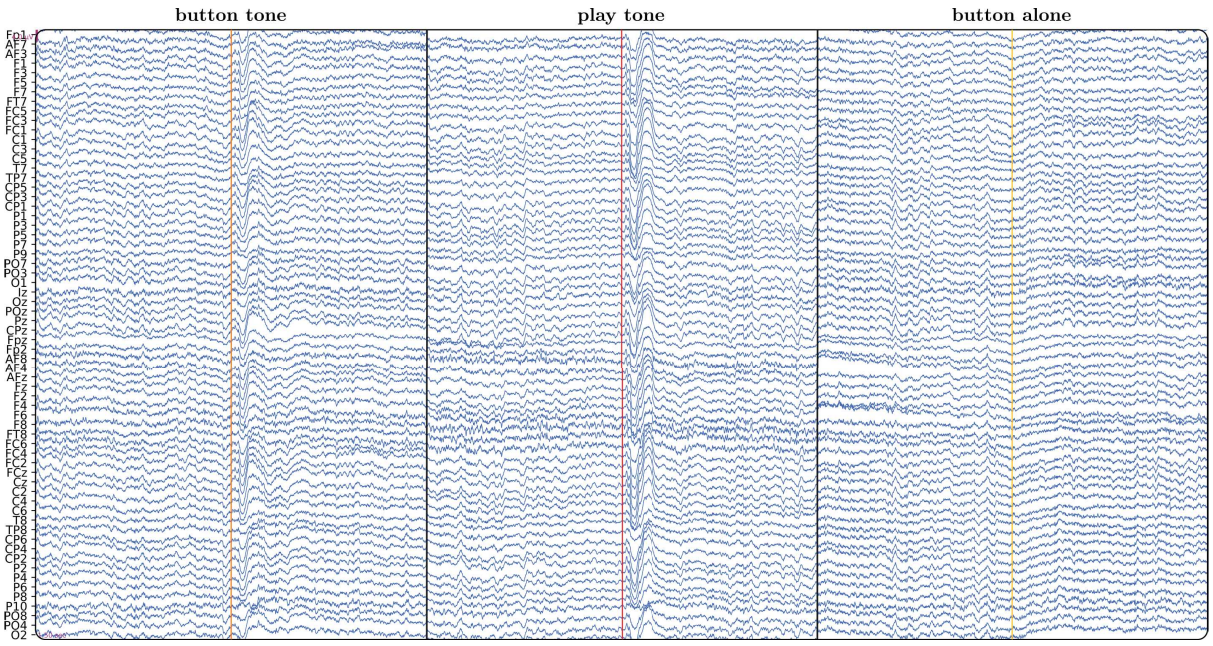


Figure 4.2: ERP of each stimulus over the 64 channels.

As for the pre-processing we applied, the data were first passed through a low-pass filter at 50 Hz[3] and then, down-sampled to 128 Hz. Finally, the data were re-sliced to form time-locked windows (trials) starting from (−300 ms) pre-stimulus to (+600 ms) post-stimulus.

An Event-Related Potential (ERP) was calculated separately for each stimulus type by vertically averaging all the signal segments corresponding to the same stimulus type and channel (see Figure 4.3) [55]. The aim of this step is to filter the signal and sum up the events occurring at the same instant to make them stand out from the ambient noise.

The result of the pre-processing and computation of the ERP for the example shown in Figure 4.2 is shown in Figure 4.4.

4.3.3 Results & discussion

The system Σ defined in Chapter 1 Equation (1.2) was created using $m = 40$ pendulums that have a linearly spaced frequency taken from (0.1—50) Hz. The frequency band is identical to that used in pre-processing, as it is the widest band found in the literature. The weighting constant α_f has been set to $\alpha_f = 0.00026$ following the procedure presented in Section 1.4.2 with a final fitting error of 0.1% in order to guarantee a progressive fit at each level of sparsity. The sparsity spacing was set to $l = 1\%$ yielding 100 β solutions with a decreasing sparsity level. The weighting constant w in the equation (1.18) has

[3] The cut-off frequency was set at 50 Hz by visual inspection of the power spectral densities of all the subjects as this was the widest frequency band that contained the most desirable activity.

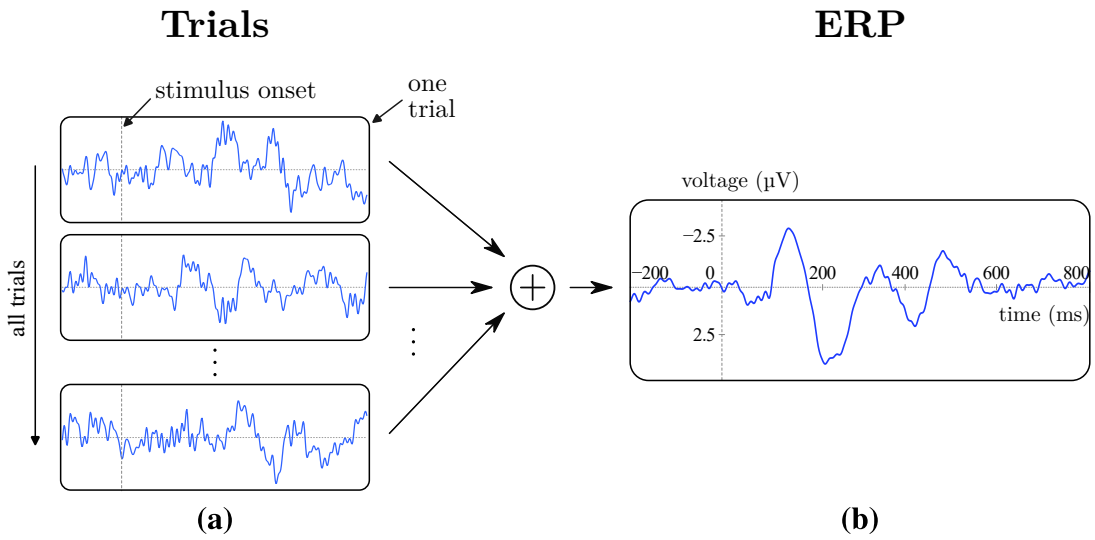


Figure 4.3: Example of ERP creation process over one channel for one stimulus type. **(a)** Vertical averaging of the corresponding time-locked EEG segments. **(b)** Result of the vertical averaging process. The signal length used in this example is only illustrative, moreover, this step was repeated for each channel and each stimulus.

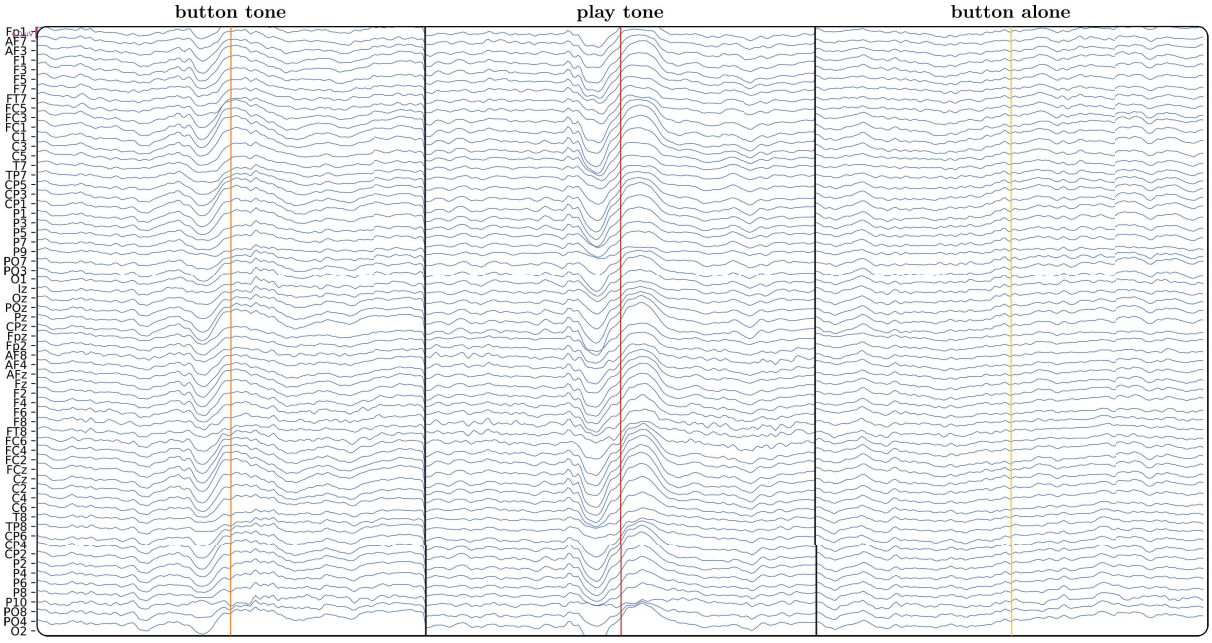


Figure 4.4: ERP of each stimulus over the 64 channels after pre-processing.

been set to $w = 0.5$ following the procedure presented in Section 1.4.3.

It is important to note that the number of modes m involved, has been set following a trial and error procedure with the objective of maximizing the validation accuracy. It is important to note that the procedure to set the number of modes m described in Section 1.4.5 was not yet developed at the time we produced the results of the current Chapter.

Out of curiosity, we wanted to test whether the value of m that has been set by trial and error is consistent with the procedure presented in Section 1.4.5.

To recall the procedure of finding m, firstly, a grid with different number of modes m is considered, and for each value in that grid, the value of α_f was determined to have a fitting error of 0.1%. The average number of excitation inputs required to fit all the signals is computed for each fixed value of m. The result is illustrated in Figure 4.5 where the evolution of the mean number of excitation inputs required for the same degree of fit is displayed versus the number of modes m present in the model.

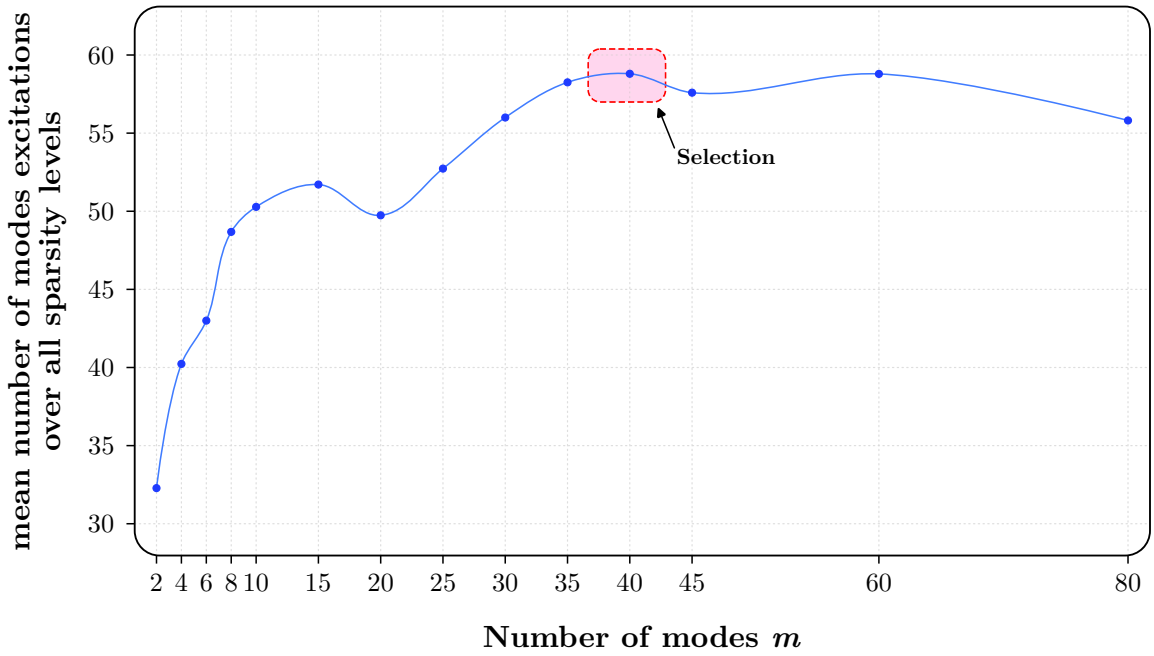


Figure 4.5: Evolution of the excitation number as a function of the number of modes m considered in the model. The same degree of fit is required for all the values.

In Figure 4.5 we can observe that the value of m found by trial and error maximizing validation accuracy does not comply with the principle of parsimony. Actually, it is the opposite, the selected value $m = 40$ is the value that required overall the largest[4] number of excitation inputs to have the same degree of fit compared to other values.

As an illustration of our model's output $\hat{Y}$ exhibiting a gradual fit with varying level of sparsity, Figure 4.6 displays the computation of the SDFs for channel C_3 and the stimulus (b) "play tone".

The SDFs for each channel and stimulus type were first generated, and then a total of 17 features presented in Table 1.1 were extracted from them. Subsequently, each feature

[4] Optimal in some sense?

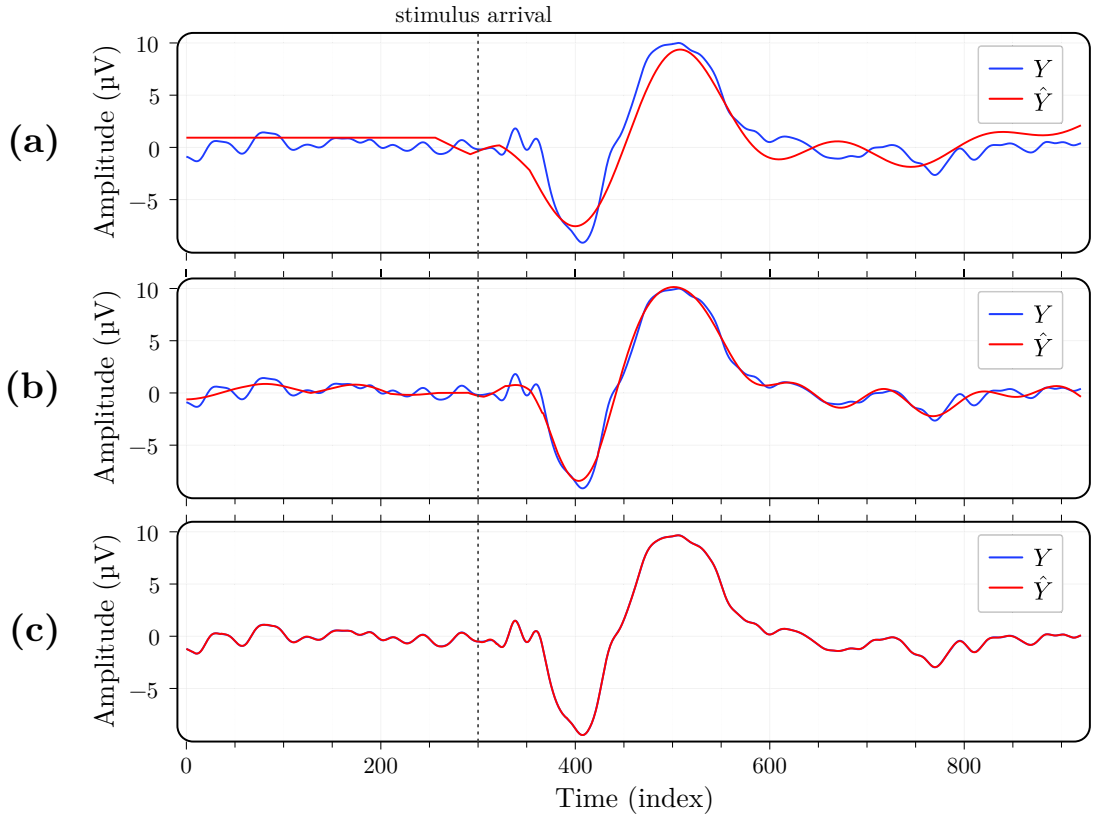


Figure 4.6: Progressive fit of the model's output $\hat{Y}$ for Channel C_3 and Stimulus (b) "Play Tone".

was tested individually. A first result was obtained using the NLOO evaluation procedure with the feature mean(U) over the EEG channel "C_3" and for the condition "(b) / Play Tone" which corresponds to passively listening to the previously recorded sequence of tones. The other channels and features did not show significant results. Using very few features, a simple linear classifier such as Linear Discriminant Analysis (LDA) was sufficient to separate the data. Table 4.1 shows the results obtained for the NLOO procedure as well as for the nested 5-fold and 2-fold CV procedures where the learning-set[5] is about 80% and respectively 50% of the entire data-set size. For both 5-fold and 2-fold CV the partitioning was repeated 10 000 times. For the test accuracy, the $5-th$ percentile is marked which indicates that the accuracy of the model is above the indicated value in 95 % of the cases. For informative purposes, the generation of the SDF takes about 2 minutes overall.

To recall, the evaluation metric used is accuracy, which is defined as the number of correct diagnoses divided by the number of decisions to be made. In the case of multiple repetitions, the stated accuracy is the mean value of the accuracy obtained overall folds and all repetitions.

From Table 4.1 we can observe that the best accuracy obtained is for the NLOO case.

[5] Composed of the training + validation set.

Evaluation scheme	NLOO	5-fold	2-fold
Learning ($\pm$ *std* %)	85.1 % (—)	84.2 % (0.6)	84.5 % (1.6)
Validation ($\pm$ *std* %)	85.0 % (—)	83.9 % (0.6)	84.3 % (1.5)
Testing ($5-th$ %)	**85.2** % (—)	80.0 %(74.1)	77.2 %(71.2)

Table 4.1: Accuracy obtained by our model using a single feature for different evaluation procedures.

This result is expected as the NLOO contains the largest learning set and the testing accuracy decreases, when the learning-set is smaller often the model loses generability which explains the accuracy loss from NLOO to 5-fold, same applies from 5-fold to 2-fold. Despite only half the data was used in the learning step, the accuracy remains reasonably high (77.2 %), this is due to the fact that a small number of features is used. We believe that if the model trained only by using half of the data performs almost the same as the model trained on the entire data-set it may suggest that the model trained on the entire data-set may obtain similar results on a data-set that is double the size of ours (under the assumptions that we have no covariate shift, for more details see [77]).

Figure 4.7 is a topomap illustrating the model's accuracy profile for each channel (regions of the brain). Its main purpose is to highlight the important regions that yielded good accuracy. On **(a)** the accuracy obtained on the test set is illustrated, while on **(b)** the accuracy obtained during the learning phase is shown. The learning accuracy is depicted to illustrate the ability of the model, using the feature presented above ($mean(U)$), to separate the data without taking into account the model's generalizability to the test set, in order to highlight important regions.

It can be observed that the regions surrounding the C_3, C_5, and FC_3 channels are those that yielded the highest accuracy for both the test and learning sets. The regions where the model was unable to separate the data are indicated in dark or light blue. Two other regions stood out and are: adjacent to the P_z channel, and the region neighbouring the C_4 channel which is symmetrical to the first region discussed. We believe that the fact that the two regions that showed the best separability (around C_3 and around C_4) are symmetrical, is synonymous with a physiological function or phenomenon occurring in these regions that can be linked to Schizophrenia.

Figure 4.8 represents the confusion matrix for the case of the NLOO evaluation procedure. The sensitivity is 90% and the specificity is 78%. We can observe that the results obtained are not perfectly balanced.

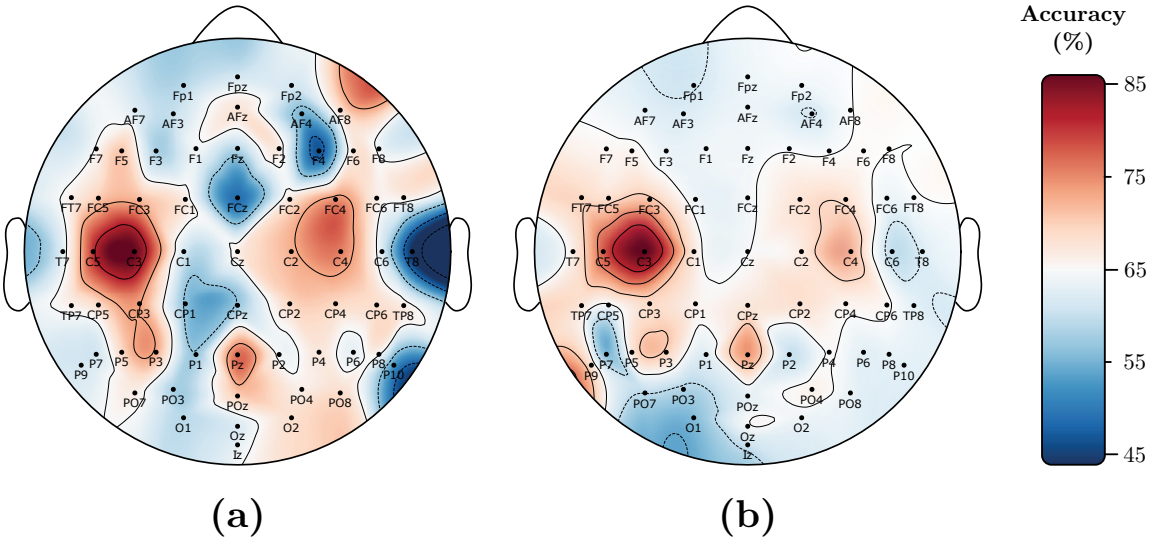


Figure 4.7: Topomap illustrating the model's accuracy profile over the brain regions. **(a)** shows the testing accuracy. **(b)** shows the learning accuracy.

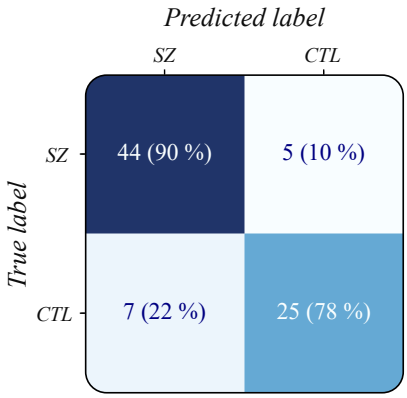


Figure 4.8: Result confusion matrix for NLOO procedure.

4.3.4 Significance and trustfulness

To assess the statistical validity and consistency of our results for the "C_3" channel we performed two tests:

1. Permutation test [65]: The probability that the results obtained are simply due to luck is discussed below as we have a small amount of data and have performed several runs (each for each type of stimuli, channel and feature). We define the following null hypothesis H_0: taking the mean of the SDF does not allow to differentiate the SZ group from the CTL group. Under this hypothesis, we randomly shuffled the labels 1000 times, and then we looked at the maximum accuracy obtained. We obtained a p-value of ($p < 0.005$) indicating that the results obtained are significant. As a reminder, for statistical significance, the p-value must be lower than 0.05 [26, Chapter 1].

2. Parametric consistency: For this part, we have evaluated the evolution of the accuracy as a function of the parsimony level. Therefore, unlike the other parts where a nested cross-validation is used, during this part only the outer loop is used in a LOO procedure for each parsimony level (in other words, the parsimony level is fixed). The obtained results are shown in Figure 4.9. It should be noted that the area of interest is the area that has been selected 100 % of the time by the inner loop when choosing the alpha hyper-parameter. This indicates that the results obtained have a parametric consistency and that it is not due to chance that one result spikes on a given parsimony level. Indeed, since it is a gradual sparsity increase, we expect the model's accuracy to be close and smooth and not have a spiking phenomenon due to luck or sampling noise.

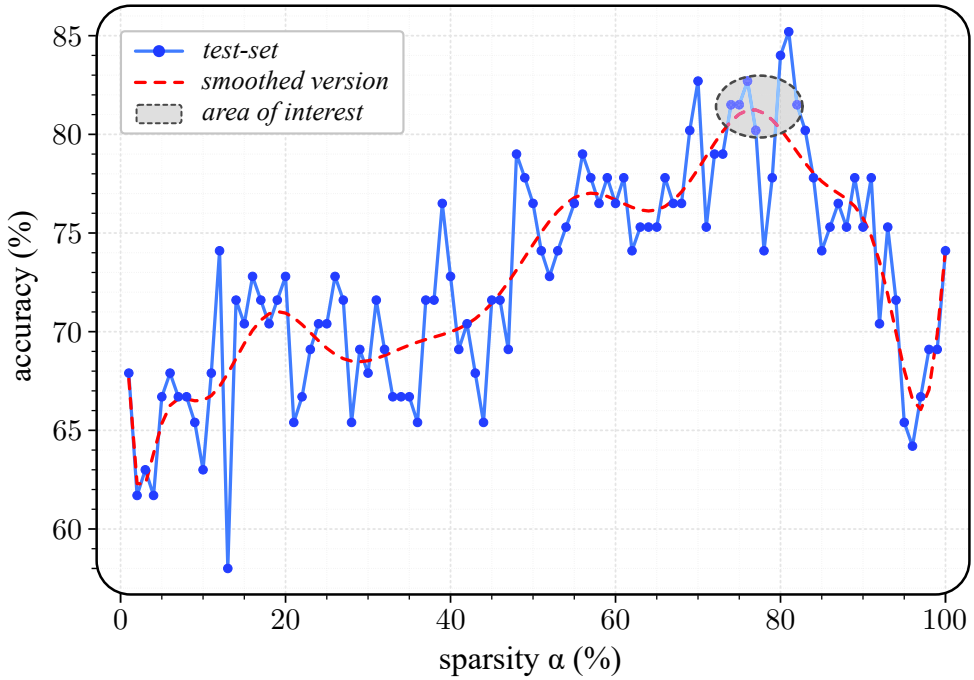


Figure 4.9: Model's test accuracy for a varying level of parsimony.

4.3.5 Data cleaning & adding of a new feature

By visually analysing the data, we have noticed that the data for some patients is very noisy. The first results in the previous section were produced using the entire data-set, we believe that it is important to provide the results without manipulating the database. However, in this section, we will be removing 7 subjects (3 CTL and 4 SZ) whose EEGs we consider to be highly noisy (see Appendix 6.1 to see their EEGs compared to a low noise EEG). It should be noted that the data from the 7 subjects that will be deleted were correctly diagnosed in the results presented in the previous section.

In addition, we have noticed that by adding another feature, the variance "$var(U)$", we can gain a few metric points. Table 4.2.a is a duplicate of Table 4.1 to improve readability. Table 4.2.b shows the results obtained using a single feature which is the mean(U) and removing the data we consider to be in poor condition. Table 4.2.c shows the results obtained using two features: the mean and the variance, while removing the noisy data.

	Features	Data cleaned?	Set	Evaluation procedure		
				NLOO	5-fold	2-fold
(a)	Mean (U)	No	Learning $(\pm\, std\,\%)$	85.1 % (—)	84.2 % (0.6)	84.5 % (1.6)
			Validation $(\pm\, std\,\%)$	85.0 % (—)	83.9 % (0.6)	84.3 % (1.5)
			Testing $(5-th\,\%)$	85.2 % (—)	80.0 % (74.1)	77.2 % (71.2)
(b)	Mean (U)	Yes	Learning $(\pm\, std\,\%)$	86.5 % (—)	86.9 % (0.5)	87.9 % (1.5)
			Validation $(\pm\, std\,\%)$	86.4 % (—)	86.7 % (0.5)	87.8 % (1.4)
			Testing $(5-th\,\%)$	86.5 % (—)	82.4 % (76.7)	**80.1** % (73.8)
(c)	Mean (U) & Var (U)	Yes	Learning $(\pm\, std\,\%)$	90.1 % (—)	88.5 % (0.6)	88.9 % (1.6)
			Validation $(\pm\, std\,\%)$	89.3 % (—)	88.0 % (0.5)	88.5 % (1.5)
			Testing $(5-th\,\%)$	**87.8** % (—)	**83.4** % (77.4)	79.4 % (72.3)

Table 4.2: Model's accuracy obtained using one or two features extracted from the generated SDF using different evaluation procedures with/without cleaning the data-set.

Firstly, by comparing Table 4.2.a with Table 4.2.b, we can observe that by removing the 7 subjects judged to have poor data, there is a slight gain in accuracy, and this, for all the evaluation procedures. By adding the second feature and keeping the same classifier (LDA), around 3 metric points were gained. As it can be observed in Table 4.2.b, it is the feature mean that has the greatest impact and carries the classification accuracy. The highest accuracy obtained is for the case presented in Table 4.2.c where two features were used and while the poor data was removed (87.8% accuracy).

Figure 4.10 shows the values of the two feautres[6] (mean and var) as well as the decision boundary of the LDA classifier.

Figure 4.11 shows the testing accuracy distribution of the $10,000$ splits of Table 4.2. The median value as well as the $5-th$ percentile are also reported. It can be observed that the values are well distributed around the median value.

[6] The level of parsimony has been fixed to produce this illustration, we have taken the one that gives the highest accuracy.

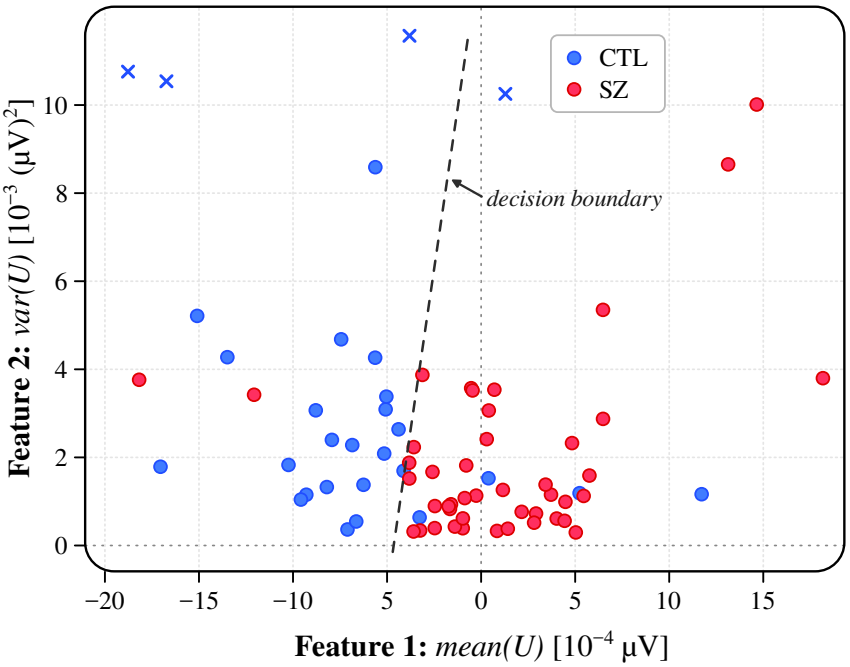


Figure 4.10: Scatter plot of the two extracted features (mean(U) and var(U)) along with the corresponding LDA decision boundary. The cross points × are distant points that have been repositioned according to their initial location for a clearer graph.

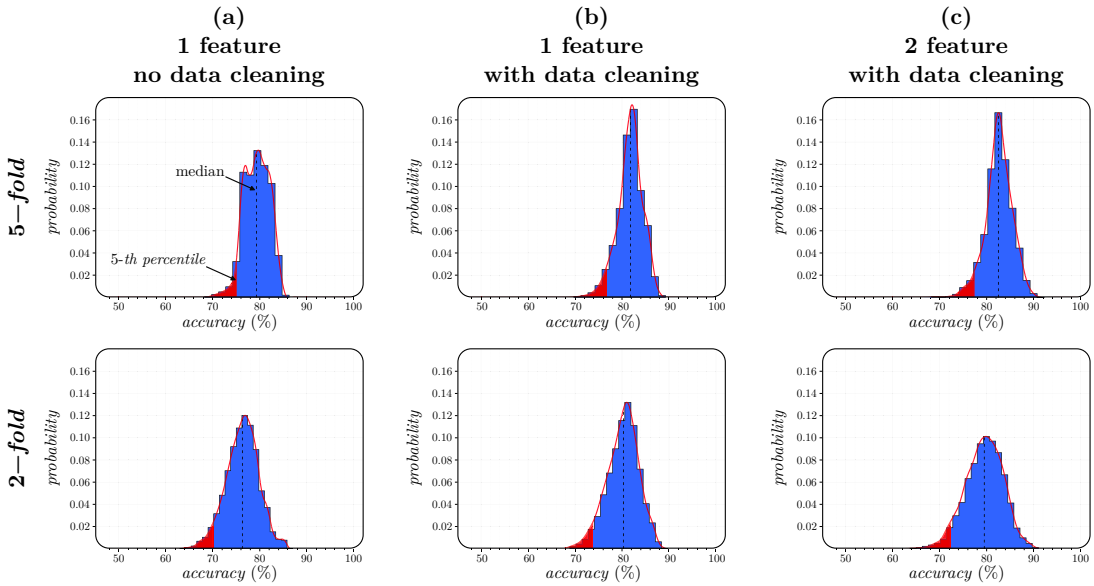


Figure 4.11: Testing accuracy distribution of the 10.000 repeated splits that produced Table 4.2.

4.3.6 SDF visualisation

This section was conducted after obtaining the results presented above. Its aim is to determine whether the SDFs of patients with SZ can be visually distinguished from those of the CTL group.

Figure 4.12 shows at the top the fit between the real signal (blue) and the model signal (red). At the bottom of the figure, we can observe which modes have been activated and when, as we have already done in Figure 1.9.

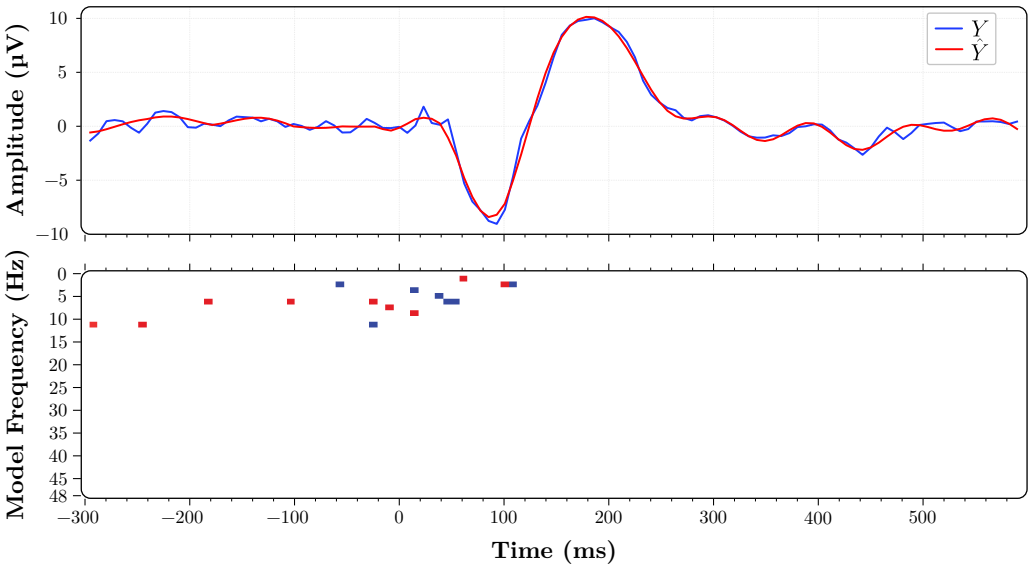


Figure 4.12: Visual representation of the model fit (top) and a bit-map of the modes' activity (bottom) of a signal taken from channel C_3.

Figure 4.12 represents the result for a single subject at a fixed sparsity level. Similarly, Figure 4.13 displays the summation of the activity of all subjects belonging to the same category. This provides us with a global view by averaging the activity of all subjects, allowing for meaningful comparisons. The selected sparsity level is 76%, which corresponds to the sparsity level determined by the inner loop (highlighted by the grey area in Figure 4.9).

Figure 4.13 shows the SDFs in their raw state. However, as it can be observed, it is difficult to draw and observe the differences directly in this figure. To overcome this problem, a spatial filter has been applied (2-D convolution) to the image[7] to smooth out the result and bring out the areas where the difference is noticeable. The spatial filter used is a standard unitary Gaussian filter $\sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$ and has the form presented in Figure 4.14.

The filtering result is shown in Figure 4.15, the areas of interest where the difference is the most noticeable are marked in dashed black. Our framework allows us to visualize the frequencies and temporal instants of the modes, which differ greatly between the two categories. We hope that this will give valuable information to practitioners on the physiological reasons causing Schizophrenia. Moreover, we think that this can make qualified persons draw new and better conclusions for the diagnosis of SZ.

[7] The image has been zero-padded to not alter the dimensions of the image.

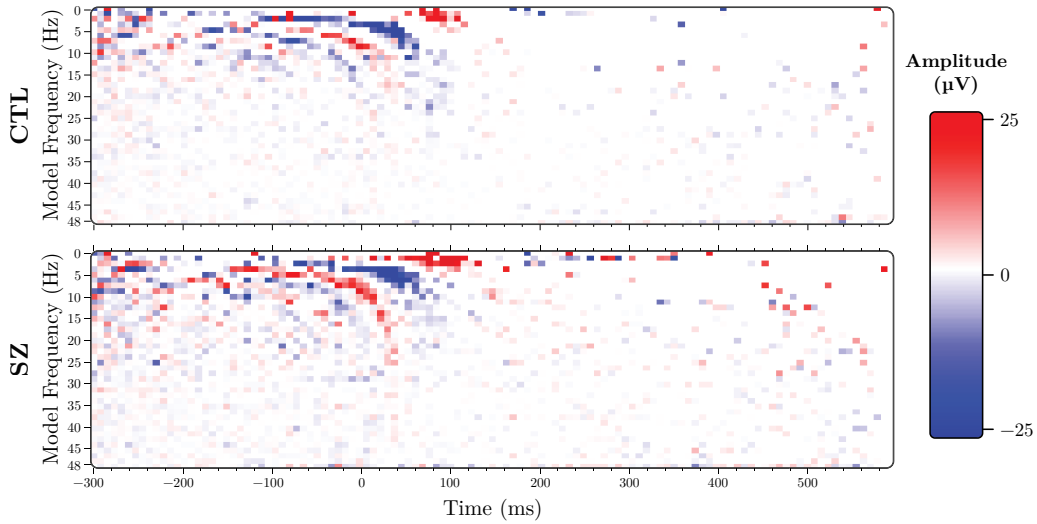


Figure 4.13: Summation of SDFs from all subjects within each category.

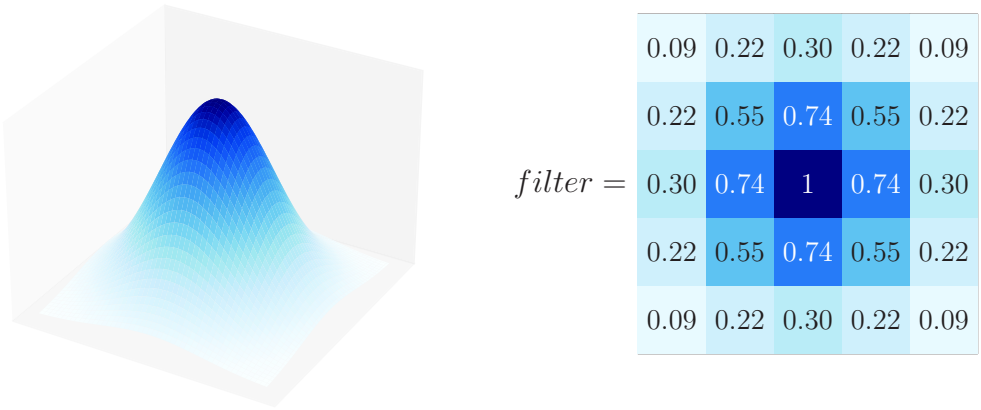


Figure 4.14: Gaussian spatial filter.

4.4 Conclusion & perspective

In this chapter, a broad overview of schizophrenia was covered first. Rather than being a disease, schizophrenia is a syndrome that encompasses several symptoms. Depending on the individual, these symptoms can manifest themselves or not with varying degrees of severity, making the illness complex and heterogeneous. The disease affects a large number of people with an estimated prevalence of 0.28% worldwide, and according to the World Health Organisation: "research has not yet identified a single cause of schizophrenia". It is important to note that schizophrenia tends to occur in episodes, where patients go through three distinct phases, each phase has its own intensity and duration. During these phases, patients suffer from a combination of symptoms with varying degrees of severity but which are separated into three categories of symptoms: (1) Positive symptoms: these are symptoms which are specific to the illness and which are known to the general public, and they are not due to a malfunction of a normal physiological process, we can find: Hallucination, delusion, etc. The other remaining symptoms are: (2) Negative symptoms,

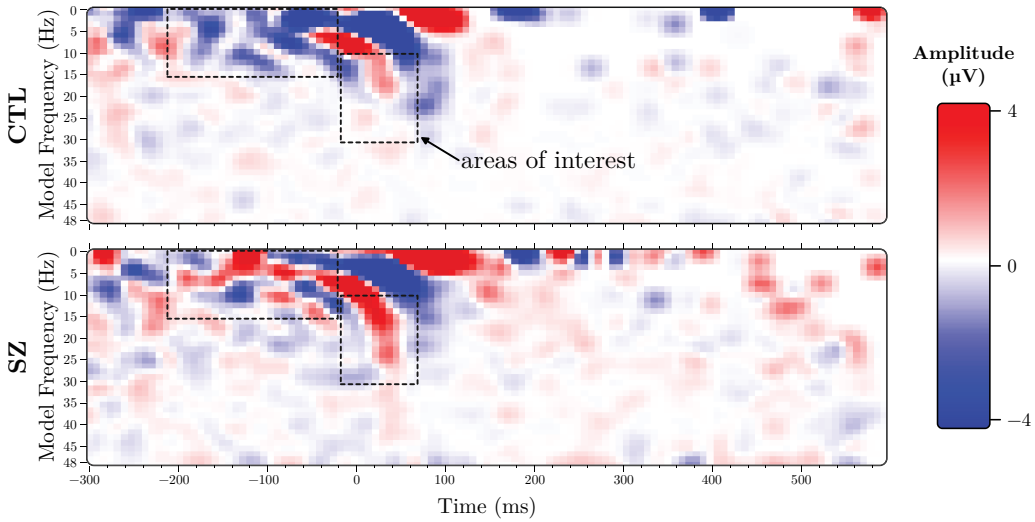


Figure 4.15: Filtering of Figure 4.13 using a Gaussian 2D filter.

and (3) cognitive impairments.

One of the guidelines that doctors must follow and that was presented in this Chapter is the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) for schizophrenia. From this it emerges that the diagnosis of SZ is entirely clinical and is based mainly and for the most part on the behavioural manifestation of positive symptoms, in addition to other considerations (duration of the illness, exclusion of disturbances, etc.). Despite the existence of these guidelines, which have improved clinical diagnosis, it is important to note that they are primarily based on positive symptoms, which, it should be remembered, can occur 1-2 years after the first symptoms, leaving patients suffering before they can be diagnosed. Various aids to diagnosis are currently being developed, and electroencephalography is one of them. EEG have stood out for the diagnosis of SZ especially for their high temporal resolution as well as its ability to record the dynamics of the brain underpinning sensory, cognitive, affective, and motor processes that are affected at the very beginning of the disease even before positive symptoms appear.

After its application in the diagnosis of Parkinson's Disease in Chapter 3, the method proposed in Chapter 1 is tested in this new setting. The method was validated on a publicly accessible data-set containing 81 subjects, 49 of whom have schizophrenia and 32 of whom serve as controls. The evaluation of our approach was carefully conducted to induce as little bias as possible so as not to have a model that is over-optimistic and that only works on the present data. The result is that the use of a single feature, which is the mean of the SDF ($mean(U)$) calculated for the C_3 channel and for the condition "(b)" which corresponds to passively listening to the same tone, was sufficient to separate the healthy from the SZ using a straight line (LDA) with an accuracy of 85.2% with a p-value of ($p < 0.005$). The model had a sensitivity of 90%, and a specificity of 78%, which indicates that the model does not classify the CTL equally well to the SZ even

though the numerical unbalance of the data-set has been taken into account. The topomap illustrating the model's accuracy profile over the brain regions for this feature highlighted 3 regions that are listed in order of importance: (1) the region around C_3, (2) the region around the C_4 channel which is symmetrical to the C_3 channel, and (3) a region around the P_z channel. The fact that the two major regions are symmetrical suggests that the results obtained are meaningful and that they are linked to a common physiological cause. The significance of the results is an important issue that we have addressed throughout this thesis, and the results suggest that the results obtained are significant. The significance of the results is an important point that has been addressed throughout this thesis, and the various tests carried out suggest that the results obtained are significant.

In order to further investigate the results, we have also visually scanned the signals, and discarded those that we judged to be poor (high noise content), those can be consulted in Appendix 6.1. It is important to note that the 7 deleted subject signals were all correctly classified by the previous model. As a result of this deletion and the addition of another feature which is the variance ($var(U)$) around 2-4 metric points were gained reaching an overall accuracy of 87.8% using only these two features and the same classifier (LDA). From the visualisation of the SDFs generated, we were able to see with the naked eye differences between the SDFs of the SZ group and the CTL group and spot the frequencies and timing of the oscillators that are the most distinctive.

Perspectives

The results obtained are still preliminary, and different perspectives can be envisaged, in particular, the use of a voting strategy between different channels allowing the aggregation of different information coming from different brain regions as it has been done for the PD in Chapter 3. We believe that having more data is also important since it allows us to use more features and more sophisticated classifiers. The possibility of using the entire area of interest spotted by SDF visual inspection will be now conceivable without affecting the statistical generability and significance of the results.

5 Application 3: Cardiac Abnormalities

Contents

5.1 Introduction

Cardiovascular diseases are the leading cause of death worldwide [82]. Electrocardiography (ECG) is a tool that offers valuable insights into cardiac and non-cardiac health and disease. The ECG is a rapid, accessible and simple test that is employed across all clinical settings, from the primary care centers to the intensive care units. Despite its accessibility and widespread uses, its interpretation requires human expertise. The reproducibility of human interpretation of the ECG varies greatly depending on the levels of knowledge and expertise. The widespread availability of ECG data has helped to develop new, automated methods for detecting various diseases and dysfunctions affecting the heart. The necessity of this automatic interpretation is helpful in developed countries, but in low and middle-income countries it is a necessity. Indeed, these countries account for more than 75% of deaths related to cardiovascular diseases (according to the World Health Organisation [106]), and often, do not have access to cardiologist with full expertise in ECG diagnosis.

In this chapter, our work is based on the study and analysis of 12-lead ECG, the most widely used ECG exam, which allows us to study the heart from different points of view. The data to be used are real data that are publicly available. These data are annotated by expert cardiologists, and our aim is to identify and classify the six types of ECG abnormalities from healthy subjects. These six types of abnormalities

are considered representative of both rhythmic disorders and conduction disturbances. Rhythmic disorders affect the electrical system that regulates the steady heartbeat, while conduction disturbances are conditions that affect the circulatory system.

The results presented in this chapter are preliminary. The main objective is not to have results comparable to the state of the art, but to show the generality and to evaluate the method in a different context, for different signals than those of Chapters 3 and 4. Moreover, we would like to investigate in more detail the visualization of the SDFs presented at the end of Chapter 4 that intrigued us. As for the state of the art, briefly, methods based on deep learning such as deep Convolutional Neural Networks (CNN) can be found. In [88], it is stated that the CNNs developed and supported by the availability of numerous digital ECGs could complete ECG interpretation similar to that of a cardiologist expert. [79] is a good example of the use of CNNs in this context. However, concerns start to raise regarding the clinical application of these CNNs in real-life scenarios, as the learned function between the input and output data is unexplainable [84].

The approach we propose addresses this problem by providing a visual explainability of the SDFs generated and used in the classification. The approach is intended to be explainable in order to properly classify the different categories. The aim is not to have the highest accuracy or to compete with the state of the art. We argue that the results obtained here are preliminary, information known to be detrimental in this context, such as the heart rate, has not been included in the model. Moreover, the entire data set has not been used.

At the end of this chapter, following on from the preliminary results obtained, we propose a scheme that could produce better results. Due to time constraints, this scheme has not been implemented.

This chapter is structured as follows: Section 5.2 presents the two groups of cardiac abnormalities studied, i.e. rhythm disorders and conduction disturbances. Section 5.2.1 defines the components of a sinus rhythm, as well as the different notions of RR and PR intervals that are widespread in the field of cardiology. Sections 5.2.2 and 5.2.3 present in detail the six abnormalities studied here. Section 5.3 is dedicated to the application of the method proposed in Chapter 1. It starts by presenting the data-set used and the pre-processing applied to the data. Results are given in Section 5.3.3 where at the end, a suggestion is given to improve the results obtained. Section 5.4 concludes this chapter and gives perspective for future work.

5.2 Generalities on the studied cardiac abnormalities

In this chapter, it is not a single condition or disease that will be studied, but rather a set of six abnormalities that affect the heart that are considered representative of both rhythmic and morphologic cardiac abnormalities. It encompasses both (1) *rhythm*

disorders and (2) *conduction disturbances*.

Rhythm disorders are caused by problems affecting the electrical system that regulates the steady heartbeat. Compared to a healthy functioning, the heart rate may be too slow or too fast, some heart rhythms are lengthened or shortened, and the heartbeat may stay steady or become irregular. The rhythm disorders that will be studied in this chapter are: Atrial Fibrillation (AF), Sinus Bradycardia (SB), and Sinus Tachycardia (ST).

Conduction disturbances are conditions that affect the circulatory system, which is responsible for maintaining the movement of blood in the body. This encompasses any condition that affects the heart or blood vessels. The conduction disturbances that will be studied include 1st degree AV block (1dAVb), Right Bundle Branch Block (RBBB), and Left Bundle Branch Block (LBBB).

5.2.1 Sinus Rhythm

Before discussing the abnormalities in more detail, it is important to define the main components of a sinus rhythm and to know how a single heartbeat is recorded on an ECG. The heartbeat is defined by the sequence (P-wave, QRS-complex, ST-segment, T-wave) which is the result of the correct synchronisation of excitation and contraction of the heart muscles. The sequence is illustrated in Figure 5.1, where each component can be visualised. Further details on the physiological aspects of sinus rhythm have been given in Section 2.3.

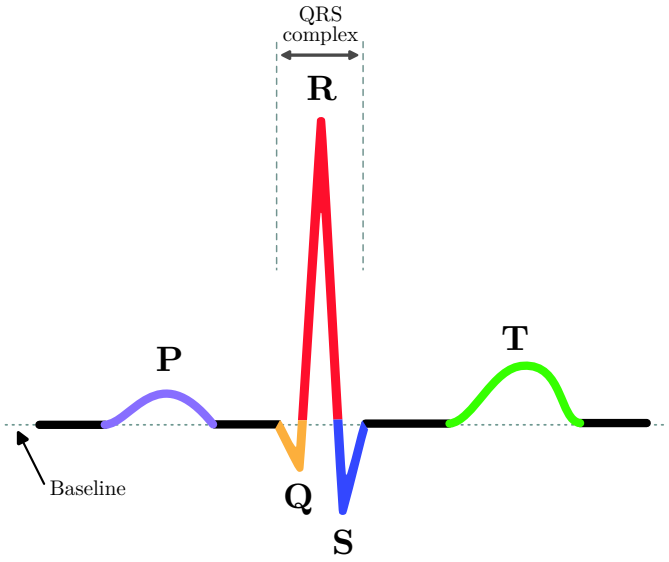


Figure 5.1: Sinus rhythms, a heartbeat perceived by the ECG.

Moreover, other concepts, such as intervals, are well-known and also important to cardiologists. Nevertheless, the most important are the RR intervals which are often

used to find the heart rate. Other existing intervals and segments that will be used are illustrated in Figure 5.2. The different segments mentioned in the figure will be referred to here and there in the description that follows.

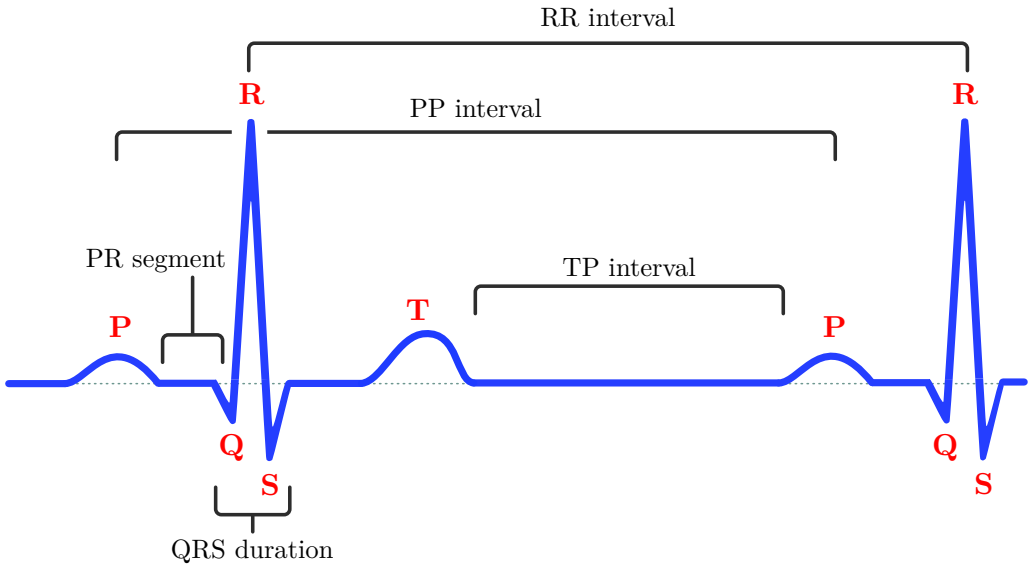


Figure 5.2: ECG important intervals.

In the following sections, each of the six abnormalities will be presented individually in what follows:

5.2.2 Rhythm disorders

Atrial Fibrillation (AF)

Atrial fibrillation is an abnormal heart rhythm (arrhythmia) characterised by rapid and irregular beating of the heart's atrial chambers. AF often begins as short periods of abnormal beating, which become longer or continuous over time [111]. In AF, the heart's upper chambers (atria) contract randomly (irregularly) and at a high pace that the heart muscle cannot relax properly between contractions. These impulses override the heart's natural pacemaker, which can no longer control the rhythm of the heart [63], which reduces the heart's efficiency and performance.

Most of the time, atrial fibrillation causes no symptoms at all, to the point where the person affected is completely unaware of the irregularity of their heartbeat [63]. The symptoms, if there are any, are:

- Palpitations: feelings of a fast, fluttering or pounding heartbeat.
- Chest pain.
- Dizziness.

- Fatigue.

Figure 5.3 shows a typical electrical activity in a healthy patient's heart (top left), and on the top right side, the unusual electrical activity present in the atrial that causes AF. At the bottom, the difference can be seen between the typical heartbeat of a healthy patient in contrast to the cardiac rhythm of a patient with AF. The spikes in the QRS complex reveal that the cardiac rhythm is uniform but irregular. Moreover, electrical activity even outside the sinus rhythms is present in AF patients.

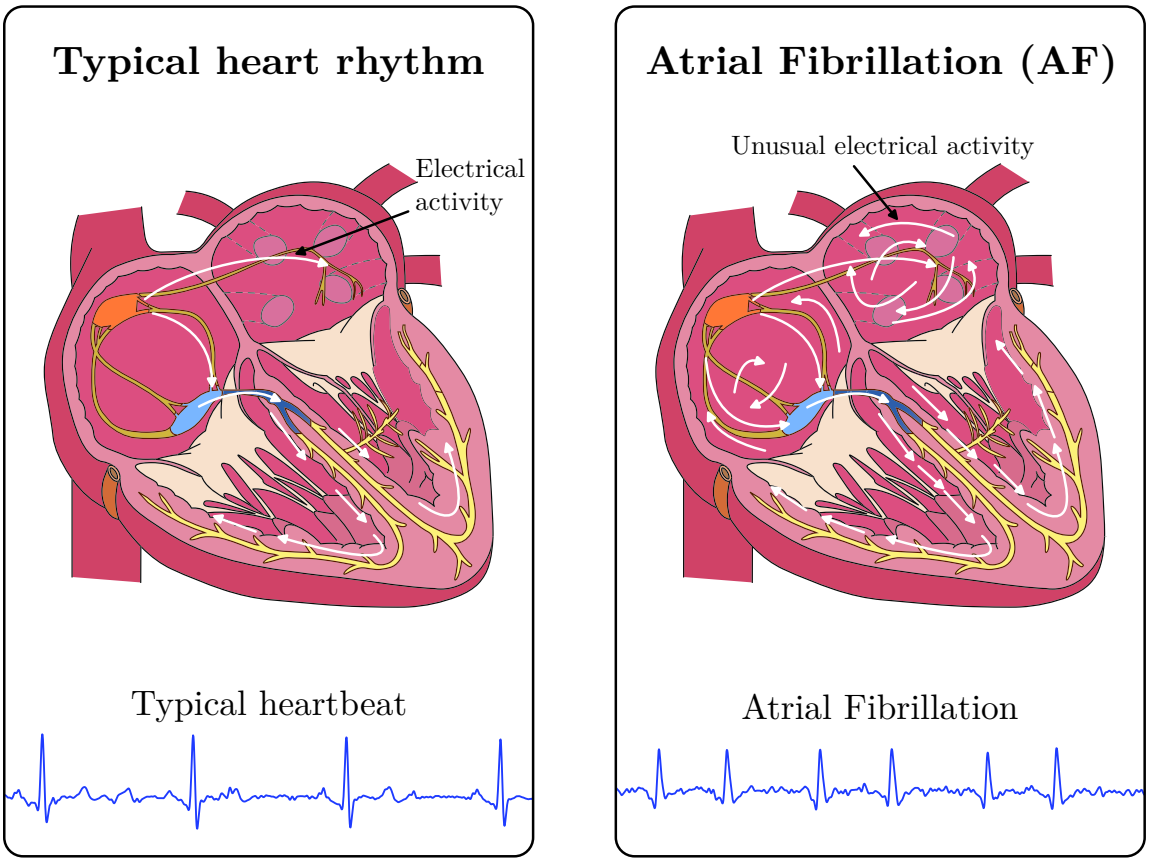


Figure 5.3: Difference in excitatory electrical activity between a healthy patient (left) and a patient with AF (right). Below, the difference between the ECG (lead *DII*) of the two categories is shown respectively.

Sinus Tachycardia (ST)

Sinus[1] tachycardia is a regular cardiac rhythm in which the heart beats faster than normal, exceeding 100 beats per minute (bpm), while for a healthy person, at rest, the heart rate should be around 60 — 100 bpm [41]. While it is common to have tachycardia as a physiological response to exercise or stress, it causes concern when it occurs at rest.

[1] The name sinus comes from the increased rate of electrical discharge from the sinus node, as presented in Section 2.3.

Most cases of ST are asymptomatic and ST itself is a symptom of a primary disease and can be an indication of the severity of that disease [49]. If symptoms are present, they often do not persist for long and may include:

- Palpitations.
- Fainting.
- Chest pain.
- Difficulty breathing.

Figure 5.4 shows the heartbeat of a person with ST (top) and a healthy individual (middle). It can be observed that the heart rate is higher in the case of ST compared to the healthy individual. This is also noticeable when comparing the RR intervals, which are much smaller in the case of ST ($RR_1 < RR_2$), almost double the heart rate of a healthy person.

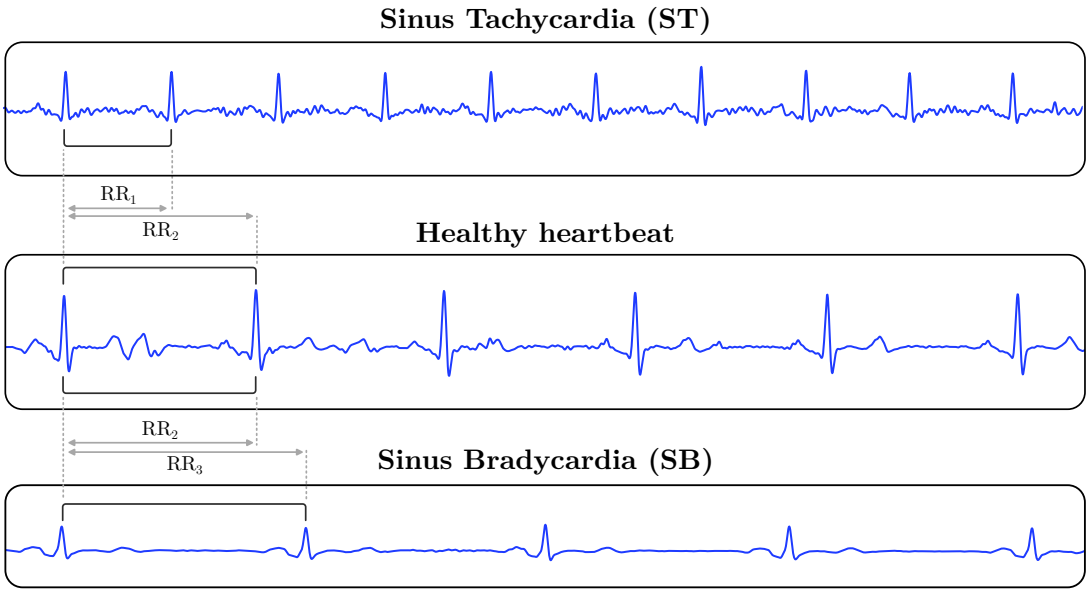


Figure 5.4: ECG (lead *DII*) of a person affected by ST (top), a healthy individual (middle), and a person affected by SB (bottom). The RR intervals are depicted to compare the heart rate.

Sinus Bradycardia (SB)

Sinus tachycardia is a regular cardiac rhythm in which the heart beats slower than expected and is inferior to 60 bpm (for an adult). It is sometimes a symptom of a certain heart condition, but it can also be a sign that a person is in a very good physical shape (professional athlete).

In most cases, SB is asymptomatic, but when symptoms do appear, they are caused by a slowing of the heart's pumping mechanism, which reduces the blood supply to the body and includes the following symptoms [47]:

- Lightheadedness.
- Dizziness.
- Hypotension.
- Vertigo.
- Syncope.

In Figure 5.4, the cardiac rhythm of a patient with SB (bottom) and a healthy patient (middle) is plotted. It can be observed that the heart rate is lower in the case of SB compared to the healthy individual. This is also noticeable when comparing the RR intervals, which are much wider in the case of SB ($RR_2 < RR_3$).

5.2.3 Conduction disturbances

1st degree AtrioVentricular block (1dAVb)

A first-degree AtrioBentricular block (1dAVb) manifests when the conduction through the AV node (presented in more detail in Section 2.3) experiences a delay, consequently slowing down the transmission of the action potential from the sinoatrial node to the ventricles. Despite its occurrence, 1dAVb typically remains asymptomatic, although there is the potential for progression to more severe forms of heart block, including second- and third-degree atrioventricular block. The diagnosis of 1dAVb is typically performed using an electrocardiogram, where it is identified by a prolonged PR interval, exceeding 200 ms. See Figure 5.5 that depicts an ECG signal of a patient affected by 1dAVb (top) and a healthy subject (bottom). The difference in the PR interval is also depicted.

Right Bundle Branch Block (RBBB)

A right bundle branch block (RBBB) is a heart block in the right bundle branch of the electrical conduction system [83]. The electrical signal cannot travel down through its normal pathway. The signal still gets to the right ventricle, but with a delay as the signal has to spread from the left bundle branch through the heart muscle and slowly activate the right ventricle. Some of the ECG criteria for the diagnosis of a RBBB include the following:

- QRS duration greater than 120 ms.
- "bunny ear" pattern in leads V1 and V3.

Figure 5.6 shows the bunny ears pattern present in the ECG of a person affected by RBBB (left top) compared to a healthy individual (left bottom). The blockage affecting the right bundle branch is depicted in the right side of the figure.

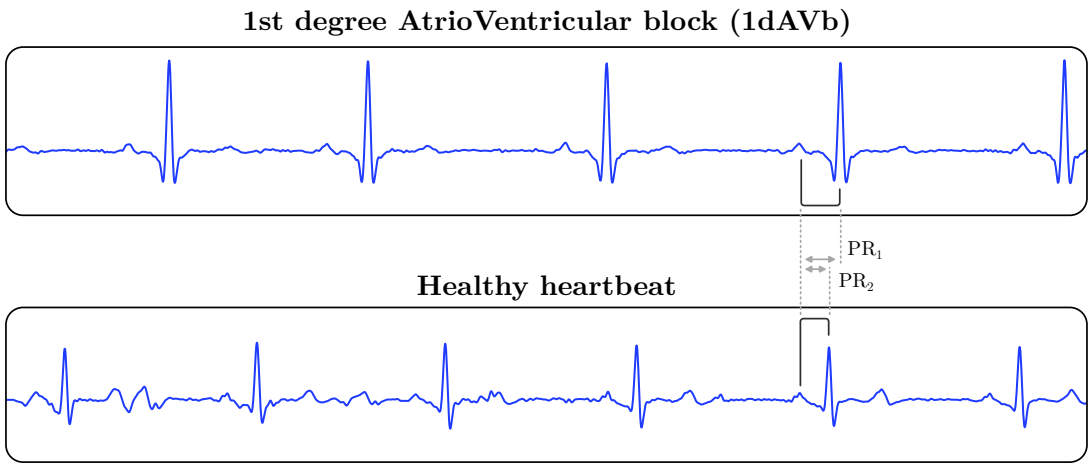


Figure 5.5: ECG (lead *DII*) of a person affected by 1dAVb (top) and a healthy individual (bottom). The PR intervals are depicted to highlight the major differences of the two categories.

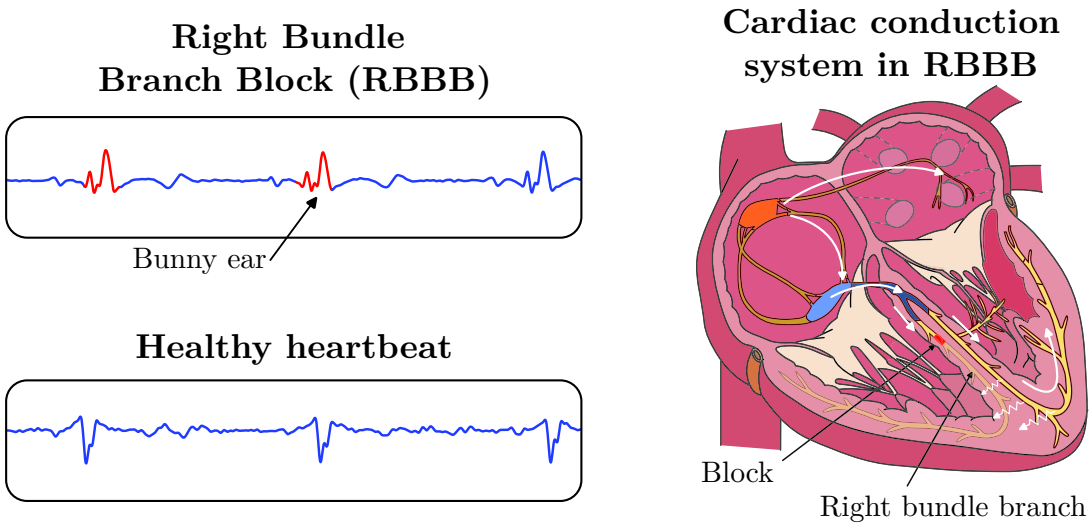


Figure 5.6: ECG (lead *V1*) of a person affected by RBBB (top) and a healthy individual (bottom).

Left Bundle Branch Block (LBBB)

Left bundle branch block is the same as RBBB, but applied to the left branch. The left ventricle contracts a little later than it normally would. This can cause an uncoordinated contraction of the heart and this can be observed directly on the ECG as:

- QRS duration greater than 120 ms.
- Dominant S wave in V1
- Prolonged R wave peak time (greater than 60 ms in leads V5 and V6)

Figure 5.7 shows the ECG signals from lead V6 of a patient with LBBB (top) and a healthy patient (bottom). Focus has been put on the V6 lead in order to show the

stretching of the R wave peak time discussed in the previous paragraph. The peak-time R1 and R2 are placed side by side to be compared.

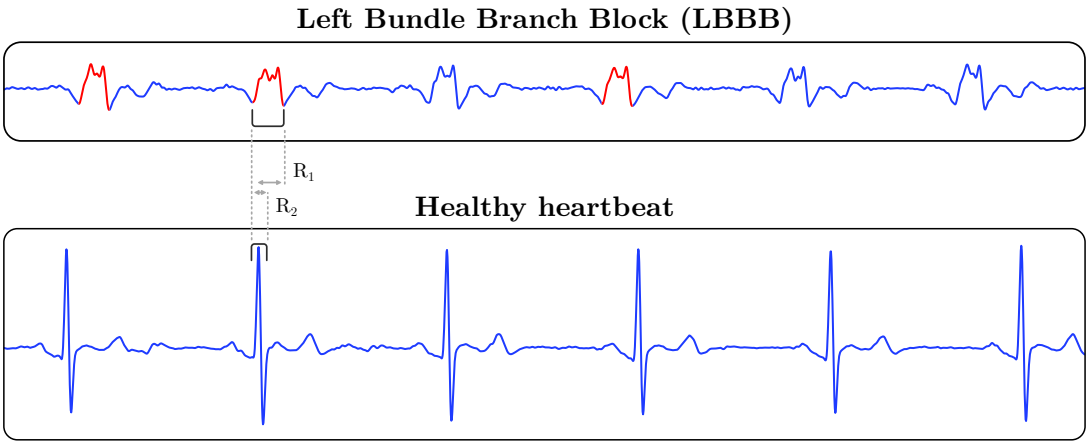


Figure 5.7: ECG (lead *V6*) of a patient with LBBB (top) and a healthy patient (bottom).

5.3 The Sparse Dynamical Features generation framework applied to cardiac abnormalities detection

This section is entirely devoted to the application of the methodology presented in Chapter 1 for the detection of the six abnormalities presented in Section 5.2. Firstly, the database used in the analysis is described. Next, the various pre-processing steps we applied to the data are detailed. In addition, preliminary results are given, which are based on direct visualization of the SDFs. Lastly, other ideas that we thought might be interesting to explore are given, unfortunately, due to time limitations, It was impossible to explore all the possible tracks.

5.3.1 Description of the Data-set

The data-set used in this Chapter is the largest we could find that is publicly available (can be downloaded from [79]). The data-set consists of 2 322 513 ECG records from 1 676 384 different patients that were recorded in the period between 2010 and 2016 by the Telehealth Network of Minas Gerais initiative[2] [2]. Six abnormalities were annotated alongside the healthy ECGs, the abnormalities are the same as those presented in Section 5.2. The annotation is based mainly on annotations provided by expert cardiologists backed up by an automatic diagnostic system. Each has limited accuracy and can be subject to error, so these two pieces of information have been combined. Briefly, if the two annotators are in agreement, the annotation is validated, otherwise other considerations have been taken into account. Complete details about the data-set and the annotation

[2] Brazilian public telehealth initiative that offers telediagnosis and teleconsultation services mainly for primary care.

can be found in [79].

The complete database is available upon request to the author, but a set containing 15% of the data is publicly accessible. For the purposes of our demonstration, this is sufficient.

Before starting the analysis and data manipulation the data set was segmented into training, validation, and test sets. Each set represents the respective proportions: 70%, 15%, and 15%. As a reminder, the learning set is a merge of the training and validation sets, so it has a proportion of 85%. We were careful to ensure that ECGs from the same subject (correlated) were not included in both the learning and test sets to avoid data leakage. Attention was also paid to keeping the same proportion of subjects from each category in each set. Moreover, the subjects that have multiple abnormalities have been removed from our analysis (a total of 3671 subjects). Table 5.1 shows the number of subjects present in each set and for each category. Moreover, in order to work with a balanced data set, and not take into account the data unbalance. For each class, the minimum number of subjects present in the condition with the fewest subjects was randomly sampled. For example, for the training set, a new training-set was created by randomly sampling 2816 subjects from each class.

Condition	Entire dataset (proportion)	Train	Validation	Test
Healthy	308004 (90.1%)	201218	53410	53376
RBBB	7685 (2.2%)	4984	1331	1370
ST	6976 (2.0%)	4491	1257	1228
AF	5609 (1.6%)	3667	968	974
LBBB	4894 (1.4%)	3107	904	883
SB	4744 (1.4%)	3103	805	836
1dAVb	4196 (1.3%)	2816	699	681

Table 5.1: Number of subjects present in each set and for each category.

The recording was performed using a 12-lead ECG system, which is the most widely adopted standard for ECG recording [57]. The sampling frequency is $f_s = 400$ Hz and the recording duration is about 10 s for a total of 4096 sample points for each signal and for each lead. This makes a signal of size $(4096, 12)$ for each subject. Some recordings have a total duration of 7 s, so they were zero-padded on both sides in order to have the same data size. The data is in its raw state and has not undergone any digital pre-processing before being made publicly available online. Figure 5.8 shows an example of an ECG over 12-leads of a healthy individual that required zero-padding.

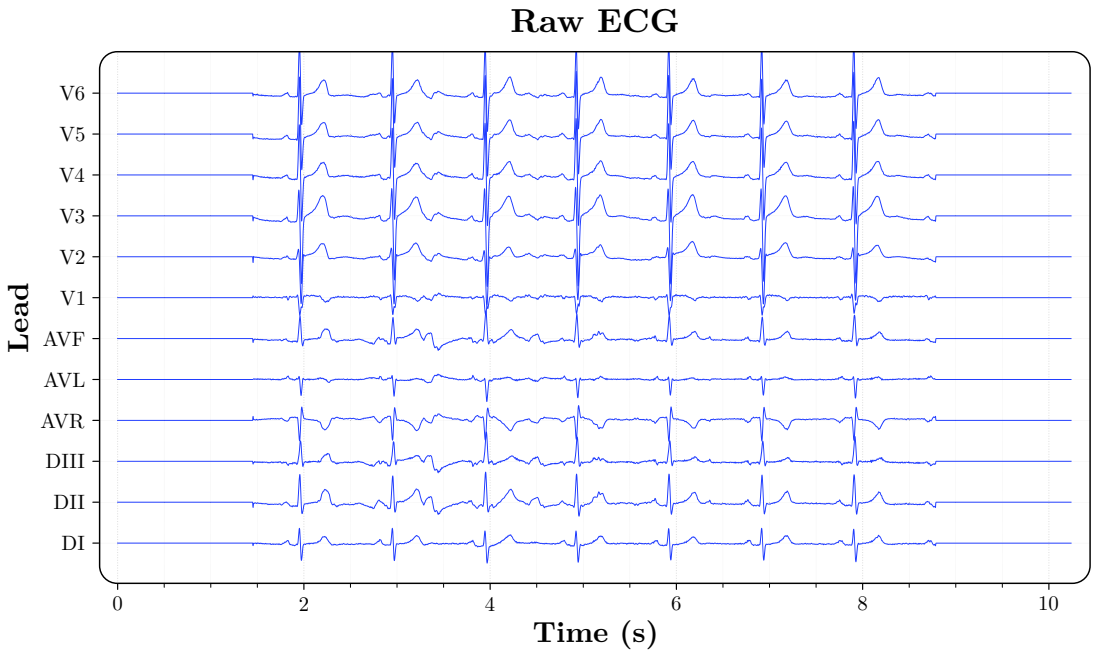


Figure 5.8: Example of a raw 12-lead ECG of a healthy patient that has been zero-padded.

5.3.2 Pre-processing

The data underwent minimal pre-processing, involving bandpass filtering between the frequencies of (3 to 45) Hz using a Finite Impulse Response (FIR) filter. The selection of this frequency band was based on a direct observation of the average Power Spectral Density (PSD) of 200 ECGs from healthy subjects in the training set, which is representative of the majority. Figure 5.9 illustrates the mean of these PSDs with a 95% confidence interval.

Once the data had been filtered, an R-peak detection algorithm was used, which is necessary for the subsequent step which is ECG segmentation. The algorithm in question is the Hamilton algorithm[3], which is widely used in the medical field, more details about its implementation can be found in [37]. The result of this process has been applied to the raw signal shown in Figure 5.8 to produce Figure 5.10. In the latter, it can be observed that the location of the R-peaks has been clearly identified (marked in red).

Once the R-peaks had been identified, the ECG was segmented into different chunks (windows). These windows start from -0.2 s to 0.4 s around the R-peak (the R-peak corresponds to 0 s). The same summation principle for creating event-related potential (ERP) which was used for Figure 3.7 Section 3.4.2 has been repeated in the present chapter. The aim of this step is no longer the same as before, i.e. to reduce noise, since the latter is very moderate in ECGs compared to EEGs. But this step served a simplification purpose

[3] The algorithm uses the lead *DI* for R-peak detection.

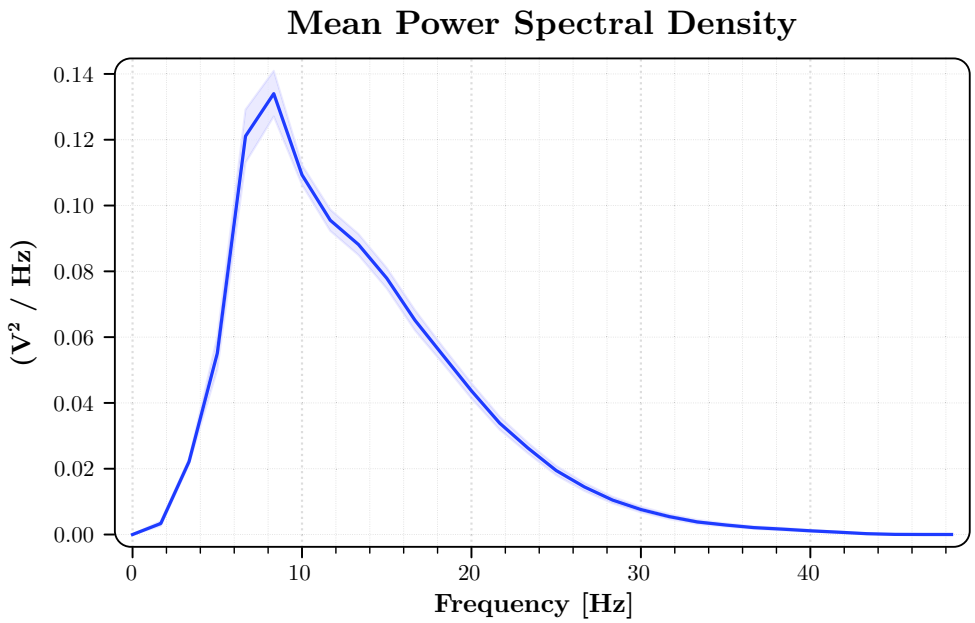


Figure 5.9: A representative mean Power Spectral Density of 200 ECGs from healthy subjects with a 95% confidence interval.

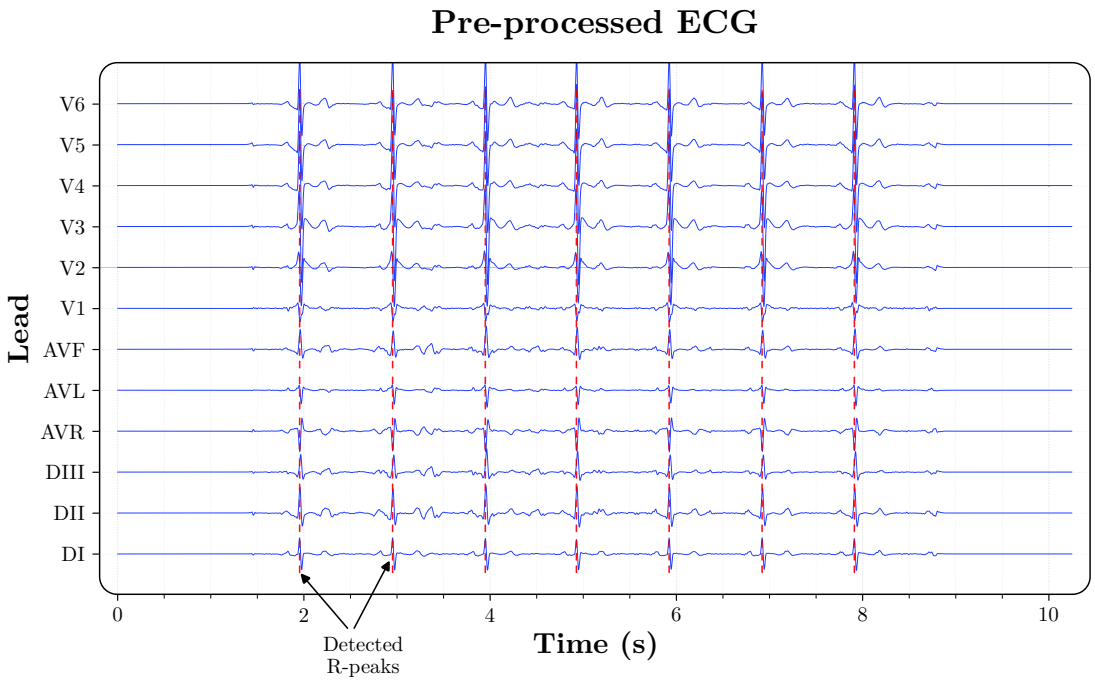


Figure 5.10: Pre-processed ECG, the detected R-peaks are marked in red.

that can be summed up in two points:

1. *Data-dimension*: As the heart rate varies among individuals, it leads to the yielding of a varying number of segments for each subject. Consequently, this results in data of varying sizes across subjects. The management of this data during classification

is complicated by the fact that it is necessary to aggregate the information from each window, for example by having them vote on the condition of the subject.

2. *Lowering algorithmic complexity*: In the end, there will be fewer signals to fit with our methodology. Although we could have kept all 10 s of ECG, however, the dimension of the optimization problem (1.8) would have grown considerably, and the model would need much more time to be fitted.

It is important to note that averaging the different windows remains representative of the whole cardiac sinus, as shown in Figure 5.11. The blue curve represents a single ECG segment, while the red curve shows the vertical averaging. It is worth noting that R-peak detection is very rarely faulty, which means that the averaging is not very representative.

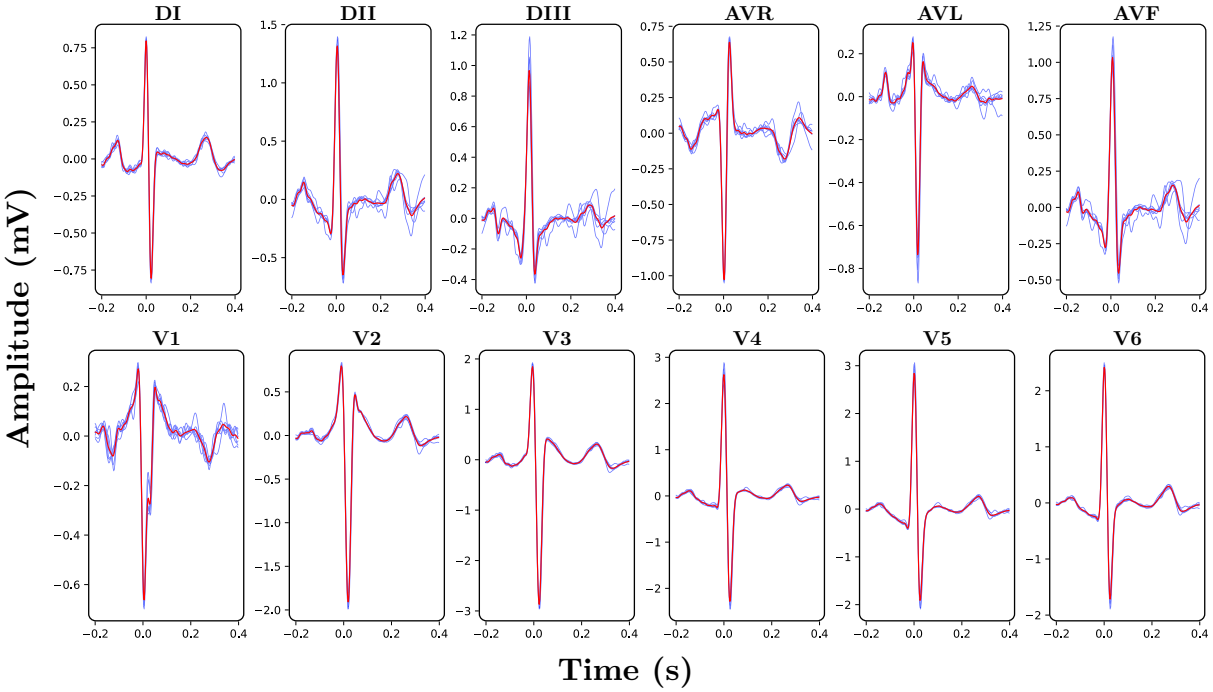


Figure 5.11: Vertical averaging (red) of the different windows each represents a sinus rhythm (blue).

Remark 1. Employing this method of averaging windows assumes that abnormality impacts all sinus rhythms, a premise supported by our observations. However, this approach harbours a significant issue that we are mindful of. Often, cardiologists rely on heart rate information, particularly for rhythmic disorders such as (AF, SB, ST), as presented in Section 5.2 and Figure 5.4, where differences in RR intervals were explored. Unfortunately, with our current method, this valuable RR interval information is lost, representing a substantial loss of valuable information. Consequently, we anticipate that the model we propose will exhibit better performance in capturing conduction disturbances (where the shape of the cardiac sinus is altered) compared to rhythmic disorders (where the primary alteration lies in the beating frequency).

Another issue we are mindful of is the fact that the PR interval exceeds 0.2 seconds for the 1st-degree atrioventricular block (1dAVb). However, in our approach, we selected a window starting from -0.2 seconds, resulting in the omission of the P-wave for this particular condition. Although we experimented with wider windows specifically for this condition with a range of $(-0.4, 0.4)$, the difference in results was not substantial. For the sake of uniformity and computational efficiency, all the segments are within the range of $(-0.2, 0.4)$ seconds centred at the R-peaks.

Once the data have been pre-processed, a signal of length $L = 240$ points is obtained. Over 12 leads, the dimension of the signal for each subject is $(240, 12)$. In the next section, the methodology presented in Chapter 1 will be applied to these signals.

5.3.3 Preliminary results

The system Σ, as defined in Chapter 1, Equation (1.2), was constructed using $m = 25$ modes with linearly spaced frequencies in the range of $(0.5 — 45)$ Hz. The upper-frequency limit is identical to that used in pre-processing; however, the lower-frequency limit is smaller. This choice was made considering that the filter is not perfect, and at the cut-off frequency the attenuation is not very effective[4]. The selection of the number of modes m in the model was determined through the procedure outlined in Section 1.4.5. The result of this procedure is depicted in Figure 5.12, where the evolution of the excitation number is plotted as a function of the number of modes m considered in the model.

The weighting constant w in Equation (1.18) has been assigned to a value of $w = 0.5$ to give equal importance to the free and forced regime. Actually, there is no free regime in the ECG at all and there is no stimulus, after which, there is a consequent change in the dynamics in the signal. Therefore, we already expect the response to be carried only by the U part of the SDF, as illustrated in Figure 5.13 where it can be observed the contribution of the free regime (carried by x_0 part of the SDF) and the forced regime (carried by the U part of the SDF).

The signal on each lead has its own complexity and therefore requires a different value of α_f, so the procedure presented in Section 1.4.2 was applied to each lead. For this procedure, the final fitting error in Equation 1.17 was set to 1%. The values of α_f found are presented in Table 5.2 for the corresponding leads.

As the ECGs are not very noisy, there was no need to have several β solutions with varying levels of sparsity, therefore, only the solution showing a near-perfect fit was taken

[4] The attenuation is 6 dB. We believe it is better to take a wider interval and observe that the model does not rely on these frequencies.

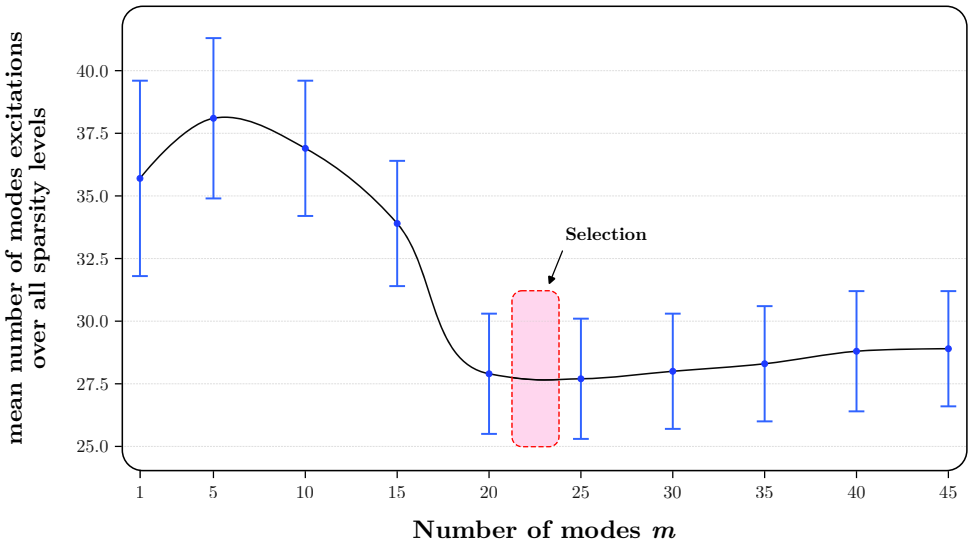


Figure 5.12: Evolution of the excitation number as a function of the number of modes m considered in the model.

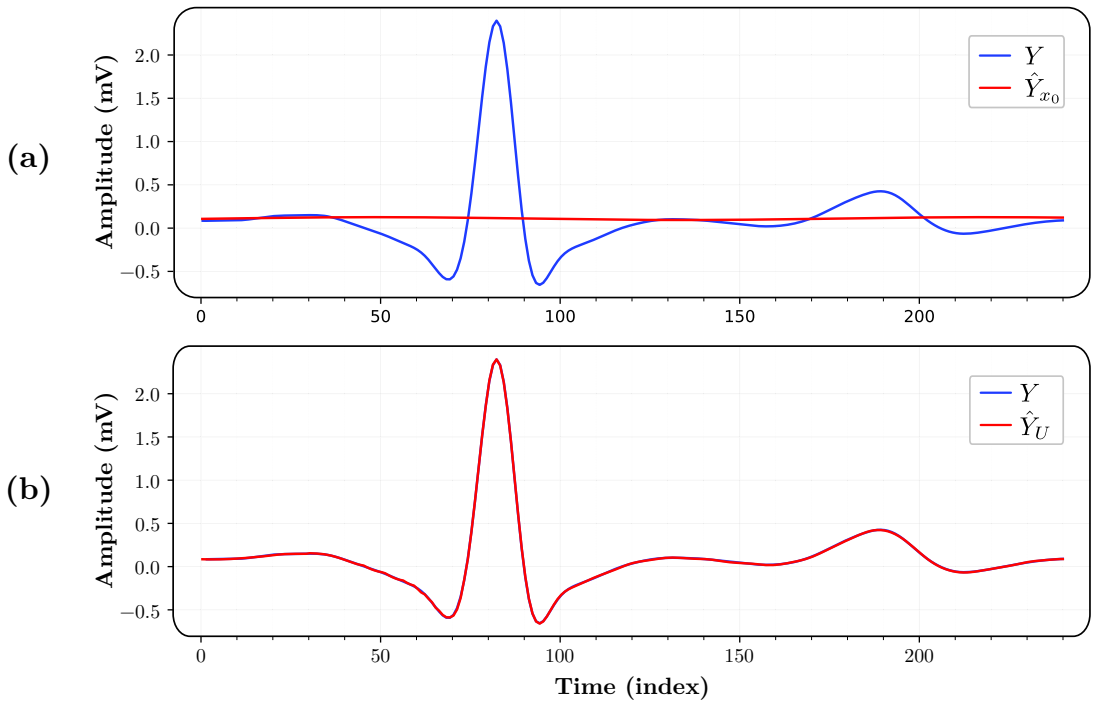


Figure 5.13: **(a)** Contribution of the free regime to the model's output. In **(b)** the contribution of the forced regime.

$(\alpha = \alpha_f)$.

Once the model parameters have been set, the SDFs were generated for all subjects and all leads. The visualization of the SDF produced for a healthy subject is shown in Figure 5.14.

Lead	DI	DII	DIII	AVR	AVL	AVF	V1	V2	V3	V4	V5	V6
α_f $(\times 10^{-5})$	15	17	10	15	9	12	16	20	23	31	33	26

Table 5.2: Resulting values of α_f for each lead.

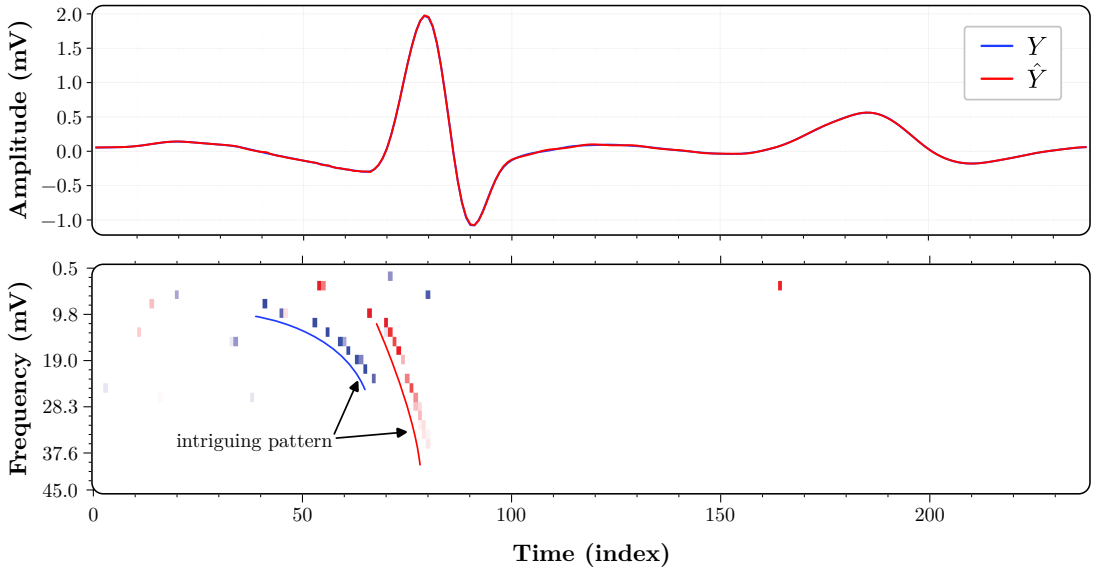


Figure 5.14: Example of the visual plot of the SDF of a healthy subject over the lead *DII*

In Figure 5.14 a pattern can be clearly seen, with modes being activated (same for deactivation) one after the other with an increasing modes' frequency. The same SDF pattern was observed for all signals. This particular shape intrigued us and motivated us to look visually at all the SDFs across all categories to determine whether there was any valuable information emerging that could separate all the categories. As a result, we summed the SDFs of patients belonging to the same category[5] as already done in Section 4.3.6 and where more details are given on this procedure. Once the SDFs had been summed, we applied the same Spatial Gaussian filter described in Figure 4.14, and the visualization result is shown in Figure 5.15. The part between 0.1 s to 0.3 s has been removed for illustrative purposes, as it did not contain any valuable information on the modes' activation.

From Figure 5.15 it can be clearly observed that the two abnormalities (RBBB, LBBB) are easily distinguishable from the other categories. Regarding the other categories, some subtle differences and shifts can be seen. More importantly, one can ask himself, are these subtle differences significant (i.e. are they representing a group pattern) or is it rather due to chance or variance? We observe that this method of visualization is not suitable to answer this question and more in-depth analysis needs to be realised.

Further on in this part, our interest is limited in determining whether an abnormality is

[5] From the training set.

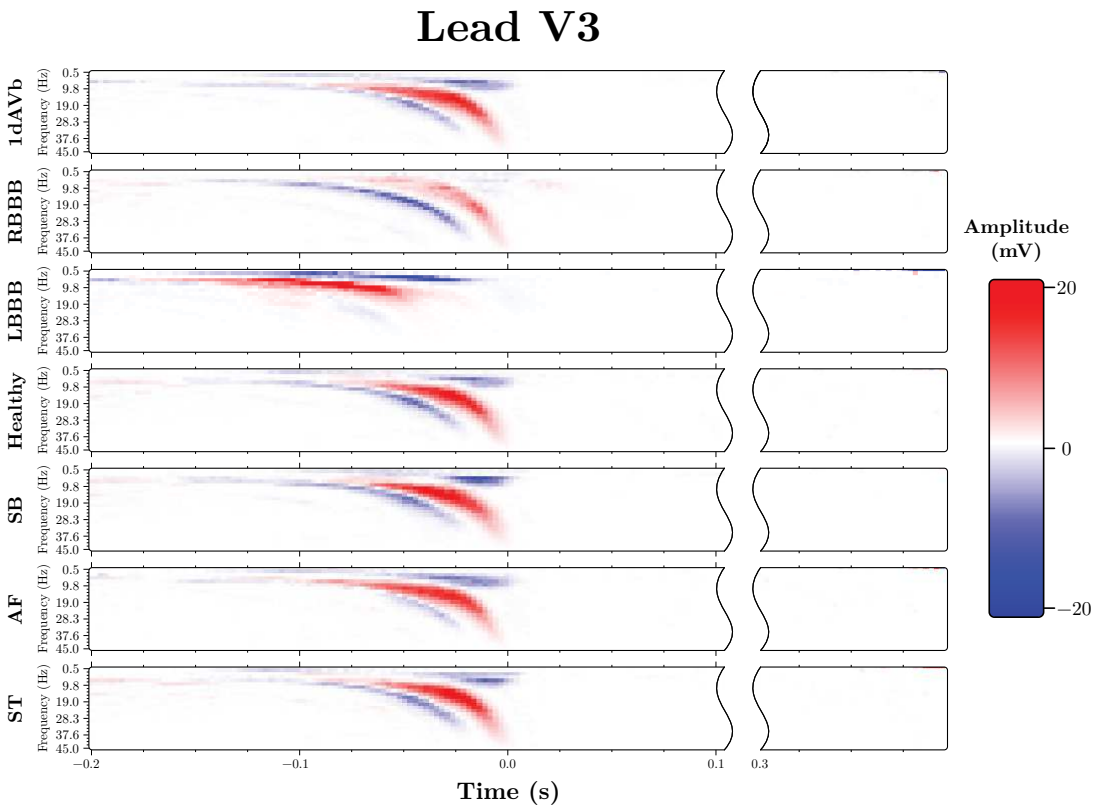


Figure 5.15: Summation of SDFs from all subjects within each category and after spatial filtering.

sufficiently distinguishable from the healthy category (please note that for the moment, we are not concerned with whether an abnormality is different from all the other categories). To better highlight the areas of difference between abnormalities and the healthy one, a subtraction has been used. For the sum of the SDFs of each category produced using the training set, the SDFs of the healthy validation subjects will be extracted. This allows us to see differential images that directly highlight the areas of interest that are different between the abnormality and healthy subjects with a dark colour. Moreover, by comparing the training and validation of healthy subjects, the effect of intra-group variance can be seen. So if this difference for a category is significantly more illuminated compared to the reference image produced by the healthy difference, this suggests that a valuable information is present in that area.

Figure 5.16 shows an example of this subtraction. It is important to note that the scale on this figure is different from that used in Figure 5.15. It is also important to note that filtering is only used for visualisation purposes (to smooth out the result), and is only used after the subtraction has been made.

In Figure 5.16, the difference between "LBBB" vs "Healthy" and "RBBB" vs "Healthy" can be easily observed. Moreover, using this visualisation, the differences in modes' activity (amplitude, frequency and timing) can be observed. It is important to remember

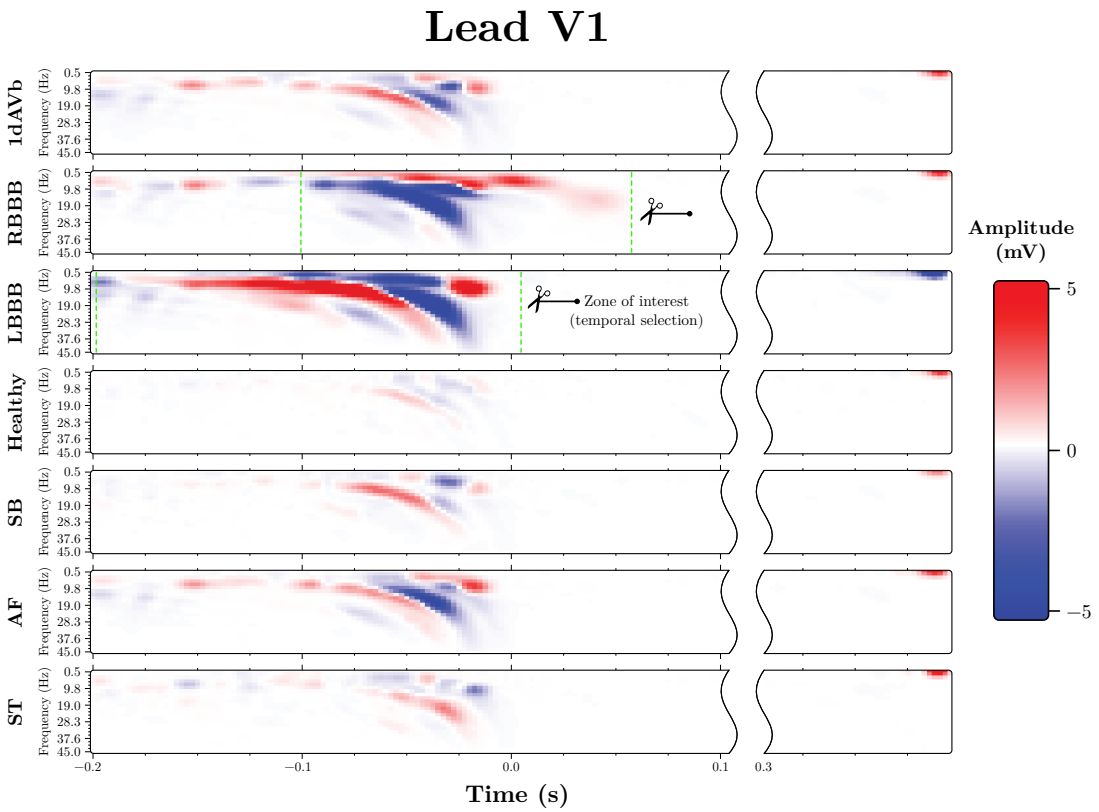


Figure 5.16: Summation of SDFs of all subjects within each category. For each category, healthy validation SDFs have been subtracted. The SDFs generated to produce this figure were made using the signals from the lead V1.

that for classification, it is not these images that are used, but rather SDFs (the U vector) or a temporal selection applied to SDFs. We have noted several zones of interest (one for each abnormality) on different leads and these are shown in Figures 5.17, and 5.18 in addition to the one shown in Figure 5.16. The temporal selection zone applied to the SDFs is delimited by green dashed lines and has a scissor mark in front of it.

For the classification, 3 possibilities were considered: (1) using SDFs directly, (2) using SDFs with temporal selection on the areas of interest marked above, (3) using SDFs with temporal selection and filtering (the same spatial filter applied above). For the classification, the AdaBoost classifier was used (detailed in Section 1.6). The weak learner is a Decision Tree with a depth of 1. Using the validation-set, the best parameters found are a learning rate of 0.5 and the maximum number of weak estimator being 100. The classification accuracy for the learning and test-set is reported in Table 5.3. It is important to note that the accuracy displayed results from the classification of each noted "abnormality" vs. the "healthy" category. Concerning the last row, all categories were included. For the abnormality marked as "1dAVb*", a wider time interval $(-0.4, 0.4)$ s was taken in order to take into account the PR interval which is specifically larger for

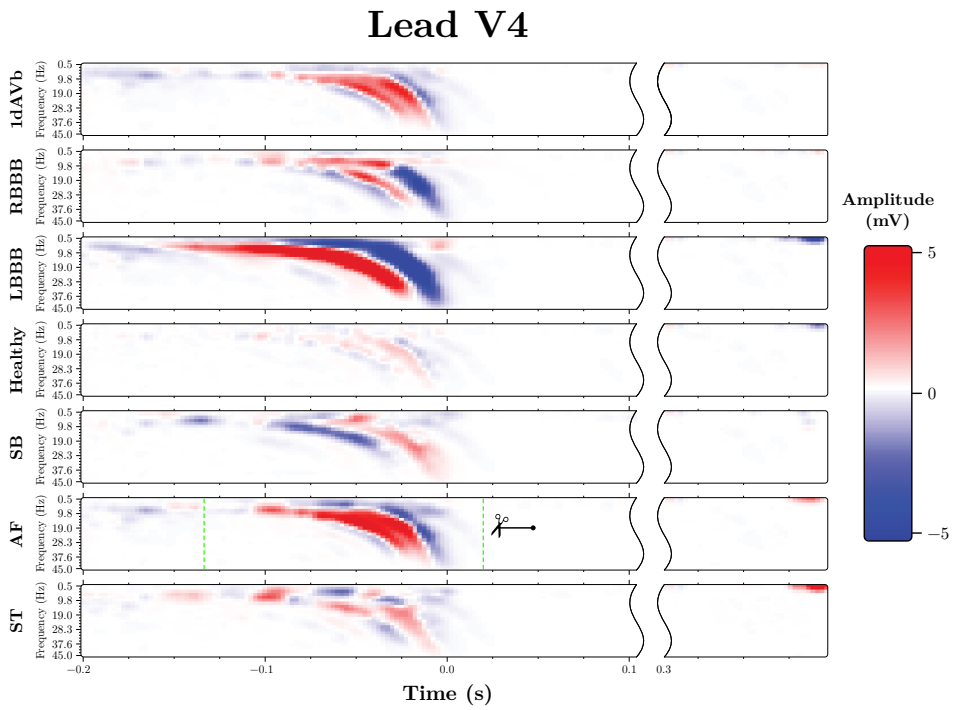


Figure 5.17: Summation of SDFs of all subjects within each category over lead *V4*. For each category, healthy validation SDFs have been subtracted.

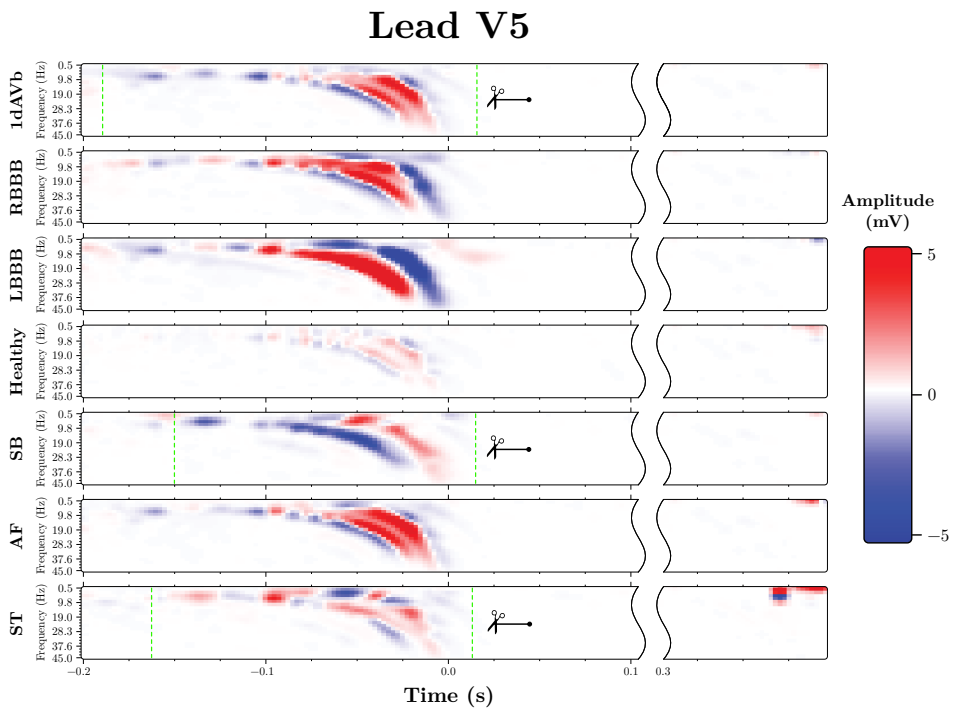


Figure 5.18: Summation of SDFs of all subjects within each category over lead *V5*. For each category, healthy validation SDFs have been subtracted.

this condition. Please remark that the SDF time selection values are written in indexes[6] where (0 corresponds to −0.2 s and 239 corresponds to 0.4 s).

Abnormality	Lead	Temporal selection	Accuracy (%)					
			no selection no filtering		selection no filtering		selection filtering	
			learning	test	learning	test	learning	test
RBBB	V1	(39 — 104)	94.6	93.1	94.3	92.5	**96.0**	**94.3**
LBBB	V1	(6 — 84)	95.5	93.4	95.2	93.4	**97.0**	**95.4**
1dAVb	V5	(8 — 90)	82.9	80.9	82.8	80.3	**85.1**	**83.0**
1dAVb*	V5	(8 — 170)	86.3	80.4	86.3	80.4	**89.8**	**89.7**
SB	V5	(20 — 86)	75.3	68.4	75.0	67.7	**78.1**	**74.3**
AF	V4	(27 — 90)	**80.6**	72.4	73.8	65.3	76.7	**72.8**
ST	V5	(14 — 90)	79.7	74.7	79.2	75.2	**82.1**	**77.8**
All	V1 & V5	(6 — 104)	—	—	—	—	53.6	52.0

Table 5.3: Result of the classification.

From Table 5.3 it can be observed that the majority of the best accuracies (noted in bold), are in the case where temporal selection was used in addition to spatial filtering. Furthermore, better test accuracy is observed for conduction disturbances (1dAVb, RBBB, and LBBB) compared to rhythm disorders (AF, SB, and ST) where the difference in accuracy is about 10% to 20%. This difference in results was expected as conduction disturbances alter the shape of the sinus rhythm, and this information could be retrieved by our model. However, as underlined above in Remark 1 since rhythm disorders affect mainly the heart rate, that important information that was lost during the creation of the windows, and therefore could not be extracted by our model, was not added. However, we are surprised to see that the model has performed quite well, with minimal information (without using the information that the cardiologist rely on, i.e. cardiac frequency). Moreover, without any changes except by taking a wider window in "1dAVb*", which includes the PR interval (to recall, an important factor for the "1dAVb" condition), an improvement in test accuracy of around 6% can be observed compared to the case with shorter windows.

The results of all leads that showed visual evidence are shown in Appendix 6.2, together with SDF visualisations of all leads. These results have not been included here due to space limitations.

[6] To avoid saying for example between −0.015 and 0.01 we say instead (6 — 84).

It is important to recall that our aim was to separate a single condition from a healthy condition, so the reference category that was subtracted for the visualisation of the SDFs was the healthy category from the validation set. If our aim is to build a classifier that works across all categories, we can and should incorporate specific zones (also identified visually) that not only separate each disease from healthy but also distinguish between different diseases. For instance, in Figure 5.18 (lead V5), it is challenging to differentiate between the SDFs of the abnormalities AF and 1dAVb. The information that needs to be added in this case is an information that allows for a clear distinction between these two categories. To address this, as we previously did, we can focus on the AF category and see where it is most distinguishable from 1dAVb (see Figure 5.19) and then use this information in a more complex classifier.

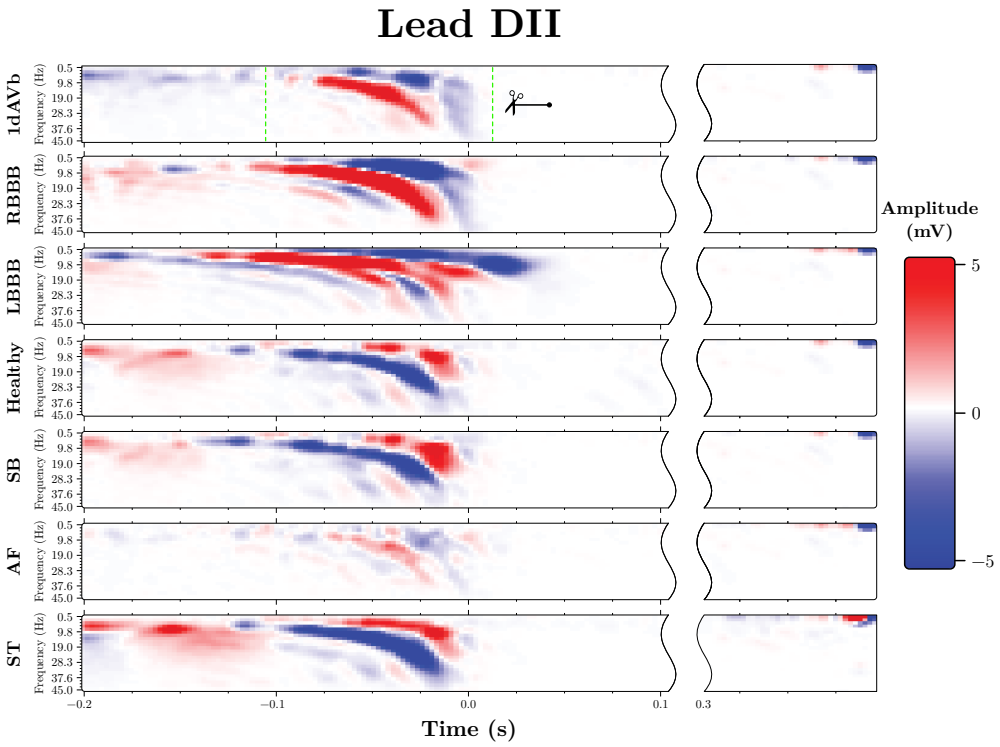


Figure 5.19: Visualisation of the SDF while the "AF" category is the focus category.

In the case where all conditions are considered at once, we believe that it is possible to improve the results of the method by: (1) adding cardiac frequency information to improve the separability of the rhythm disorders individually, and (2) using a cascade of classifiers that target specific diseases and separate them from the other conditions. This way is more appropriate and allows the right features to be used by the right classifier, so that the classification features, for example classification of "RBBB" and "LBBB" conditions, are not contaminated with other information that could lower the accuracy of that classifier. In addition, the appropriate information can be added for the appropriate classifier (for example, if we want to separate "AF" and "1dAVb", we can add the SDFs present in lead *DII* illustrated in Figure 5.19 with the appropriate time selection). A

schematic and possible proposal for this approach is shown in Figure 5.20. The values used are only a suggestion. Each classifier will have the task of classifying certain classes only, based on a specific lead and on the most appropriate temporal selection for the classes concerned.

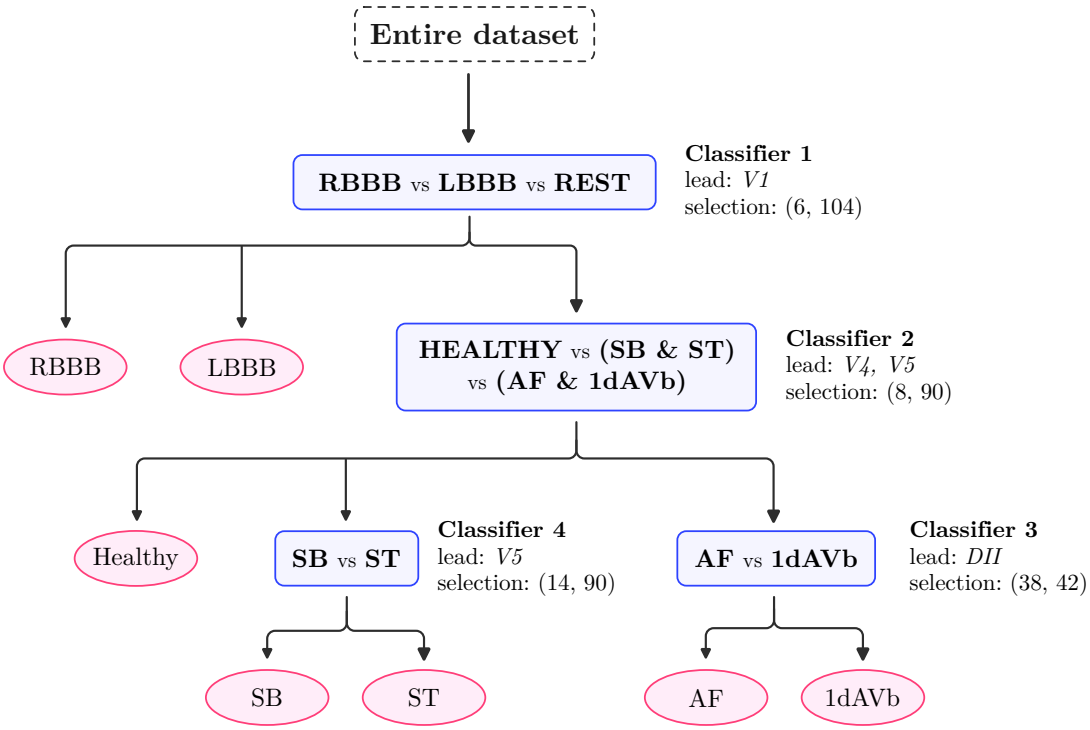


Figure 5.20: Cascade of classifiers.

5.4 Conclusion & perspectives

In this Chapter, we have seen that cardiovascular diseases are the leading cause of death worldwide and that the ECG is a quick, easy-to-access, inexpensive tool that provides valuable information on the cardiac condition of patients. The interpretation of these ECGs is complex and requires a high level of expertise (cardiologist). This level of expertise is not easily accessible in low and middle-income countries, where 75% of cardiovascular-related deaths occur (according to the World Health Organization). In addition to data availability, this helped and guided the emergence of automatic diagnosis methods for ECG classification.

Two families of cardiac abnormalities have been covered: rhythm disorders and conduction disturbances for a total of six abnormalities. Each abnormality has been detailed, with its possible causes, and how to identify it on an ECG. Important notions in the world of cardiology have been addressed, such as sinus rhythm and intervals, in addition to the heart conduction system that produces sinus rhythm. The principle of ECG recording was briefly discussed, with emphasis on the positioning of the 12-leads used to provide different angles of view on the heart (horizontally and vertically).

The methodology presented in Chapter 1 was tested on a publicly available dataset containing 348000 12-lead ECG recordings. The pre-processing we applied to these data is minimal, however, for simplification purposes we vertically summed (as for ERPs) the sinus rhythms of the same channel. This proved to be problematic in this context, as the information about the heart rate is lost which is known to be detrimental to the diagnosis of rhythm disorders.

For the generated SDFs, even though a value of $w = 0.5$ was set to give equal importance to the free and forced regime, the model by itself used only the U component (forced regime) for the fit of Y (as the x_0 was not solicited). The procedure to set the number of modes m and the value of α_f has proven to be working quite well. The sparsity level variation was not needed for this application, as ECG are not very noisy compared to EEG, thus, a perfect fit is preferable in order to capture as much information as possible.

An intriguing pattern (curves) appeared on the visual of the active modes. This pattern motivated us to visualise the SDFs across all categories to determine whether there was any valuable information emerging that could separate all the categories. Adding together the SDFs of the subjects belonging to the same category revealed very noticeable differences for the RBBB and LBBB categories, but only subtle differences were observed for the other classes. To make these observed differences more apparent, to each sum of the SDFs for each category, the value of the healthy validation SDFs has been subtracted, leaving a kind of differential image that shows the difference between the SDFs for each category and those of the healthy.

For each category, the SDFs where this difference is the most remarkable were selected.

These selected SDFs were used in an AdaBoost classifier after being filtered with a gaussian spatial filter. The results obtained are preliminary and can be improved. A test accuracy of approximately 75% was obtained for each of the three abnormalities of rhythm disorders (AF, SB, and ST). The test accuracy obtained for the conduction disturbances (1dAVb, LBBB, and RBBB) was approximately 10% to 20% higher than that of the rhythm disorders (between 89% and 95%). This difference in results was expected as conduction disturbances alter the shape of the sinus rhythm, and this information could be retrieved by our model. However, since rhythm disorders affect mainly the heart rate, and as underlined in Remark 1, that important information was lost during the creation of the signal windows which explains this low accuracy. The testing accuracy obtained using all the categories at once is in the order of 52%. We believe that this is due to the fact that: (1) the information necessary and important for the classification of rhythm disorders has not been included, which considerably reduces the accuracy. (2) the SDF information that was selected was designed to separate a single abnormality vs healthy, however, specific SDFs that not only separate each disease from healthy but also distinguish between different diseases must be included.

Perspective

An important perspective is already presented at the end of Section 5.3.3, which consists of using a cascade of classifiers, each of which has a specific classification task. As a consequence, the features that are relevant to the specific task are used only for that classifier and therefore do not contaminate the other classifiers with unnecessary information. Moreover, the information needed to classify rhythm disorders such as the RR interval, and heart rate should be included in the classification. It is important to remember that this information is already available following the use of the algorithm that locates the R-peaks. The fact of using all the dataset and not just a sub-part of it is a good way of adding information and improving the classification and its robustness.

Another approach that we consider interesting to study is the characterisation of the curves obtained during the generation of the SDFs (such as the one illustrated in Figure 5.14). These curves could be characterised by just two or three coefficients, and it could be that these coefficients are the new bio-markers that allow the six conditions to be separated.

A more in-depth study of the features that can be derived from the visuals produced by the SDFs is a possible and interesting option (like we did in Chapters 3 and 4).

Conclusion

In this thesis, a new feature generation methodology for time series has been proposed. The fundamental concept of the method consists of parsimoniously decomposing the signal into dynamical modes that can be activated and/or deactivated at the appropriate time with the appropriate amplitude. The excitations applied to these modes to fit the signal of interest, will then be used as features. Inspired by brain function, the method remains generic and versatile as it can be applied to a wide range of time series. The features generated, which we have named Sparse Dynamical Features (SDF), provide a new point of view on the data, which gives access to informative features that are faithful to brain functioning.

The versatility of the method comes not only from its design, but also from the various parameters that offer a wide range of options. One important parameter used to manage the level of sparsity of the solutions produced is "α". This parameter allows our model to fit the signal of interest tightly capturing all the information and noise as well, or not fitting tightly the signal of interest with the risk of not picking up the important information. This parameter has proved very useful in the case of high-noise signals such as EEG.

Another useful parameter is the weighting parameter "w", which is used to favour/guide the choice of modes that are used by the model and whether the model puts more emphasis on the free or the forced regime. It is important to note that this weighting only favours the regime used, however, the most important decision lies in the optimisation, so if there is no free response in the signal, as in the case of ECG applications, the model will by itself only use the forced response.

The last important parameter is the number of modes "m" contained in the system. Once the frequency support used in our model matches the frequency support of the signal (the same range), a question arises: what should be the frequency resolution of our model? The parameter m can be used to change the frequency resolution of the model and therefore, by its design, the model's fitting capabilities. There is a trade-off between model complexity and fit capacity, and this parameter is used to fine-tune this trade-off. With regard to the frequency distribution, we have found that a linearly spaced distribution is the most effective for the SDF generation. With regard to the frequency distribution, we have found that a linearly spaced distribution is the most effective for the SDF generation.

For all the parameters mentioned above, their setting is procedural in order to reduce the effect of human subjectivity on the one hand, and on the other hand, makes the method more explicable and repeatable compared to purely algorithmic (heuristic) selection methods.

It is important to highlight that the computation of a sort of Power Spectral Density (PSD) based on the generated SDFs resembles the PSD (conventional one) of the signal of interest.

The methodology we have proposed has been tested on three applications, producing encouraging results while using very few features (one or two for brain diseases) and relatively simple classifiers (linear). It is important to remember that the aim is not to replace clinicians but rather to offer diagnostic aid tools, as we are aware of the gap between a clinical application that works and tests that have been carried out on databases, even if they contain real signals. The applications studied encompass diagnostic scenarios related to: (1) Parkinson's Disease, (2) Schizophrenia, and (3) six Cardiac Abnormalities. For all three applications, the data is publicly available and for repeatability and transparency purposes, our source code is publicly available (with guides to repeat the findings).

In the literature on the first two applications (i.e. PD and SZ), data leakage problems have often been encountered. These problems are well known in the machine learning community and are: (1) group leakage (where correlated data is present in both the training and test set), and (2) optimisation of hyper-parameters and even parameters on the test set (absence of a validation set). In addition, we noted a lack of studies of the statistical significance of the findings and often this is in problematic cases such as being subject to the curse of dimensionality where the number of features used is relatively high for the few number of data instances.

It is important to note that our aim was to obtain good results without undermining their significance. For this reason, various validity tests were performed in addition to the simplicity and explainability of the findings.

The possibility of making a temporal and/or frequency selection of the SDFs generated, which in turn can be used directly or combined with a feature extraction layer, makes it possible to reinforce the significance of the results by incorporating different prior knowledge established in the studied domain. Moreover, once feature selection/extraction has been performed, simple classifiers can be used for classification encouraging the explicability of the results. Moreover, using a few number of features reinforces the robustness of the classifier to the data quantity, as the model would perform just as well if only half of the data was used. To further strengthen the explanability, a fairly intuitive way of visualising the SDFs in the form of an image is proposed where one can see the activation and deactivation instants of each mode, their frequency, and their amplitude. This visualization allows either to find areas of interest or to tailor the choice of features to be extracted. In addition, we hope that this will enable new biomarkers to be identified and perhaps shed light on the unknown biological reasons that cause the different symptoms of the disease.

The proposed methodology has been packaged as a library that can be easily used by other researchers in other contexts. All the data used in this thesis are publicly available through the links provided in the corresponding data sections within each chapter addressing specific applications. Our codes employed in the three applications are also publicly accessible via the following links, accompanied by a step-by-step guide to reproduce the results:

- Parkinson's Disease: https://github.com/HoussemMEG/SDF_PD
- Schizophrenia: https://github.com/HoussemMEG/Schizophrenia-Detection
- Various Cardiac abnormalities: https://github.com/HoussemMEG/ECG

Perspectives

It is important to note that the perspectives set out here only concern the methodology. The perspectives concerning the applications have been covered in their corresponding chapters.

- It would be interesting to evaluate the performance of the method in a task that involves only separating the forced regime and the free regime (very used in brain-computer interfaces).
- We have noticed that the methodology often suffers in the case where the signal of interest is long (L is large). One should think of ways of slicing the signal, generating the SDFs on each portion, and then reassembling the generated SDFs in a coherent manner.
- We find it very interesting to try out a new functional basis in the model, which would be more suitable for the problem being addressed (in our case the functional basis used is the sinusoidal one).
- Another interesting possibility for EEGs is, for example, to use the sinusoidal basis in the model (as we have done), in addition to which we can add the dynamical models of eye blinks, and cardiac rhythm, which are known to be contaminants of the EEG. It is possible to add weightings to the model to make it more sensitive or less sensitive to pick up this kind of noise. Subsequently, to the SDFs generation, the contribution of these "unwanted" signals can be removed, leaving only useful information. This approach may be a substitute for the ICA algorithm that was needed for the EEG signals, offering an end-to-end approach.
- Adding other information that is useful in addition to SDFs during classification can be a good proposal.

Appendix

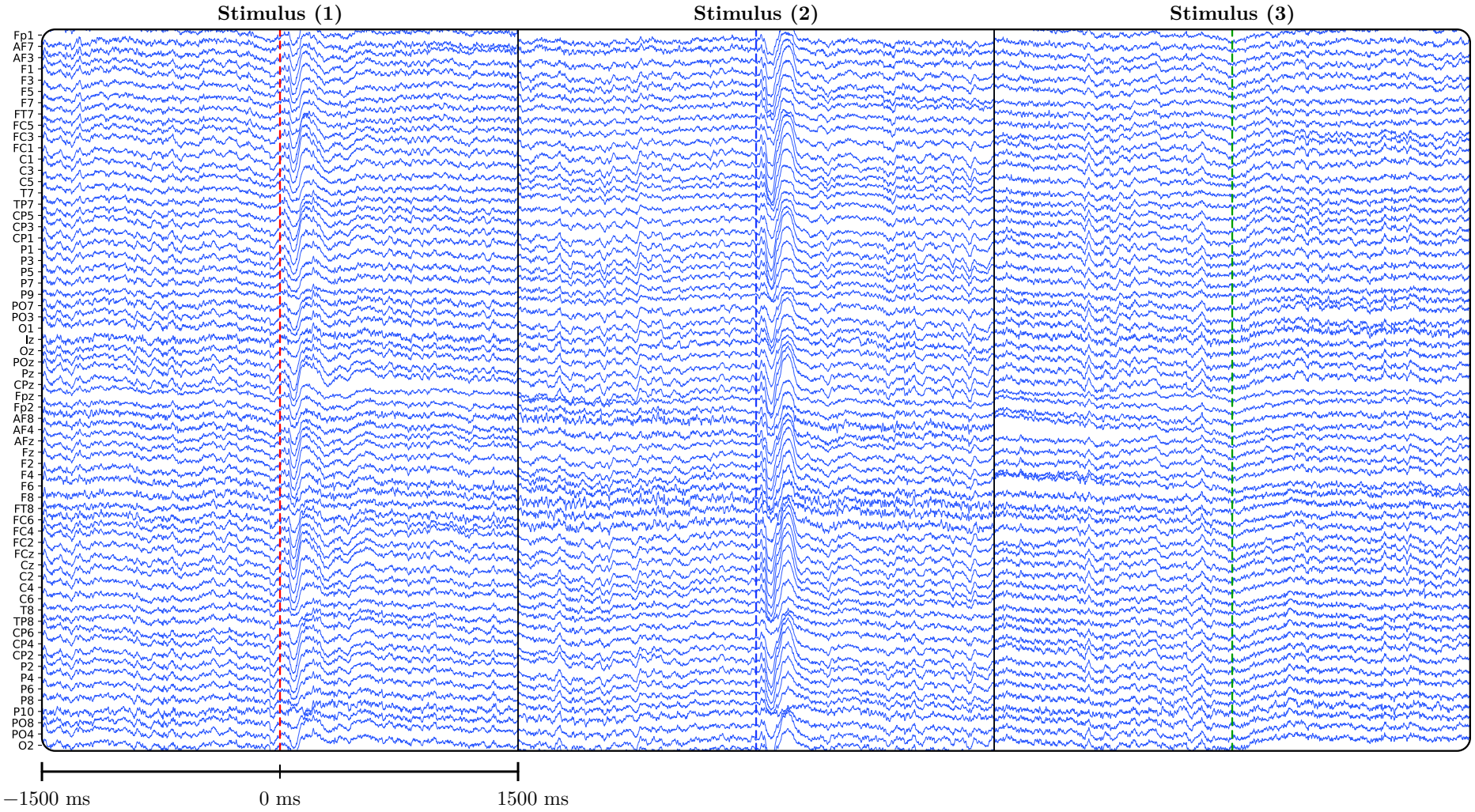
6.1 Appendix A: Schizophrenia data cleaning

In this Appendix, we have grouped together the signals from several subjects in order to compare them visually. We have taken the example of a subject with signals that do not visibly contain a lot of noise, in order to compare it with subjects' signals that are noisy, and this, even after the various pre-processing operations that took place before they were made available online.

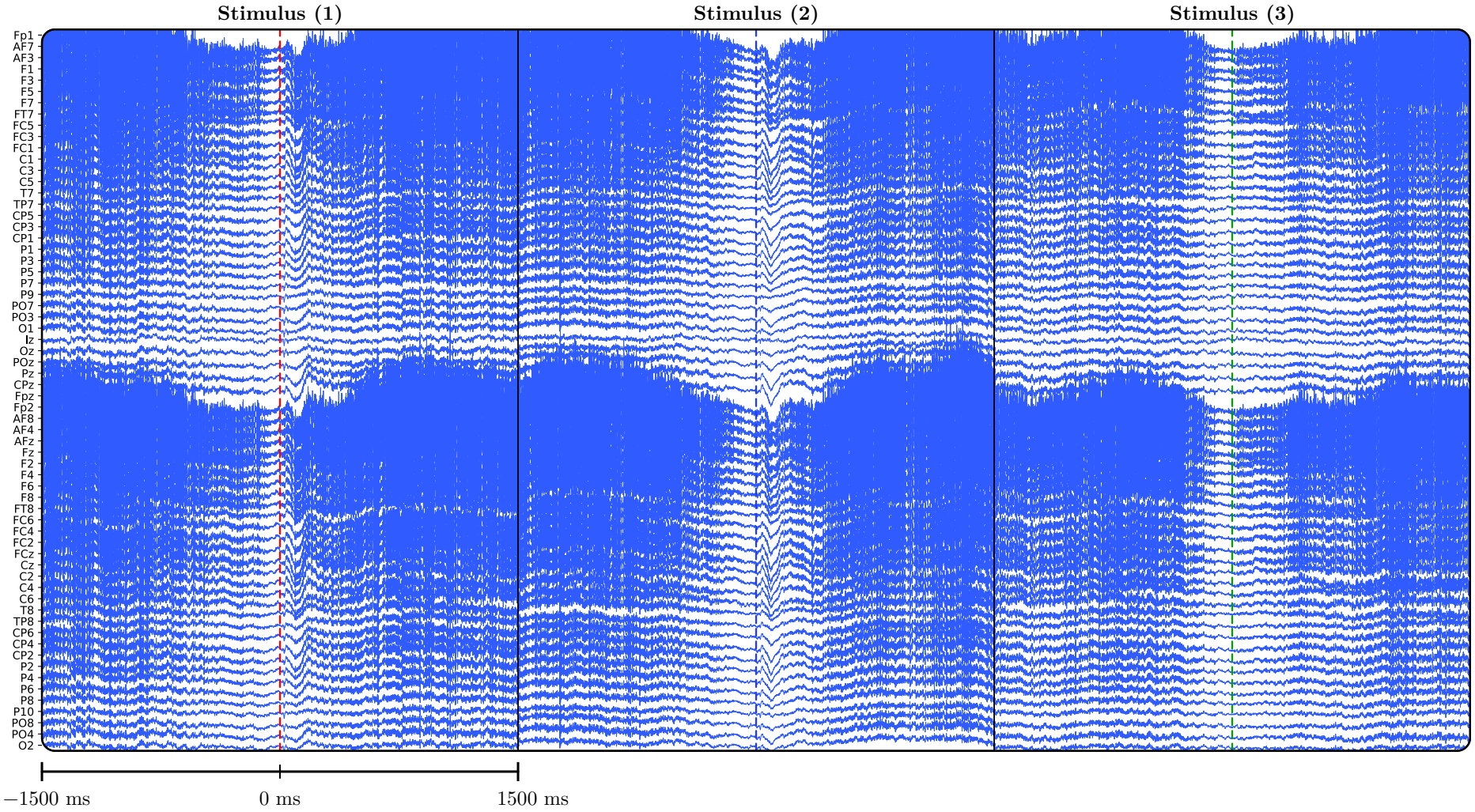
The subjects numbered (5, 17, 23, 30, 33, 57, 78) have been removed from the database only for the section where data cleaning was carried out.

It is important to note that the list of these subjects was established before the experiment was carried out and before the first results were produced, in order to guarantee total impartiality. It is also important to remember that all the signals from the subjects presented in this Appendix have been correctly classified.

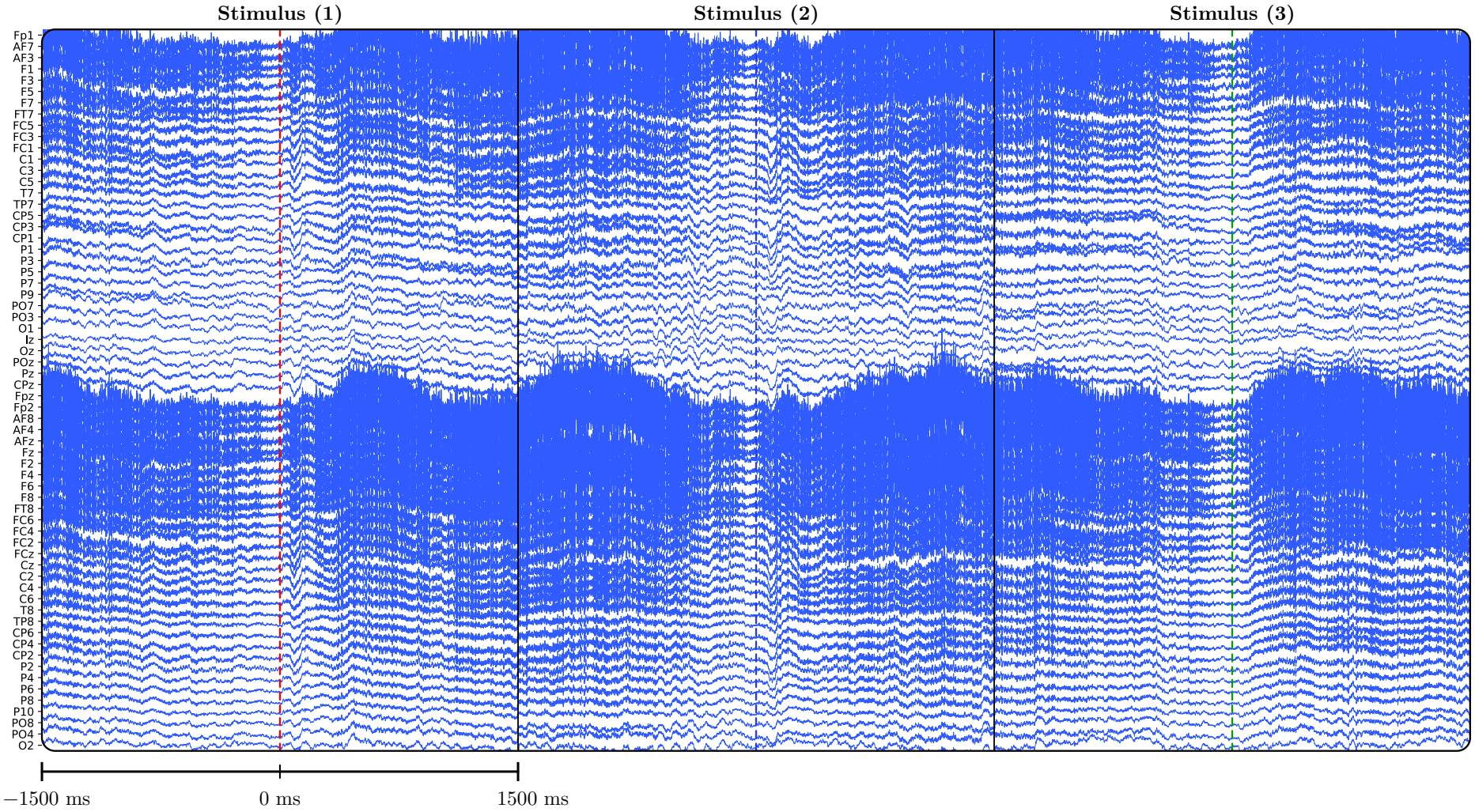
Subject 1 (CTL — good)



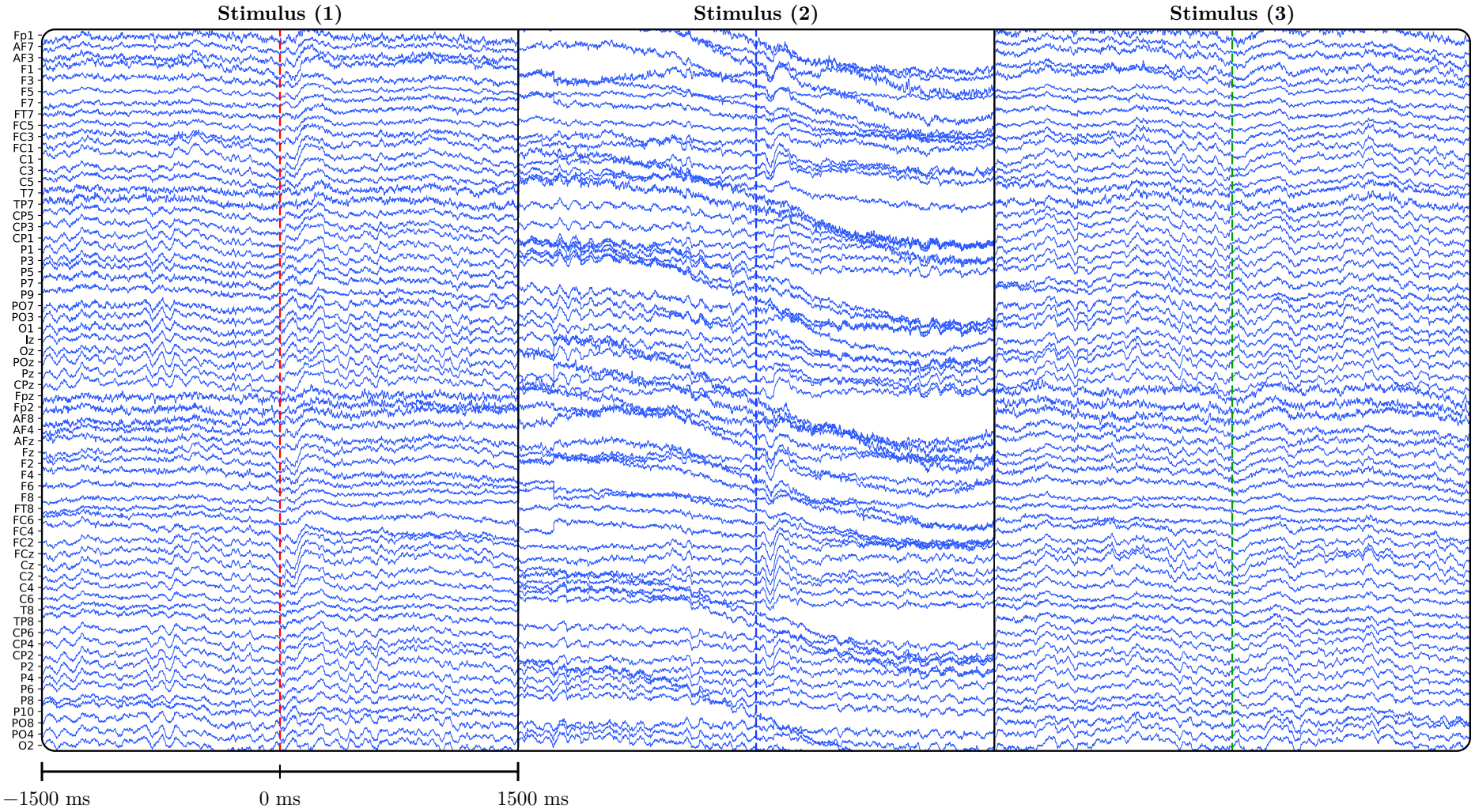
Subject 5 (CTL — bad)



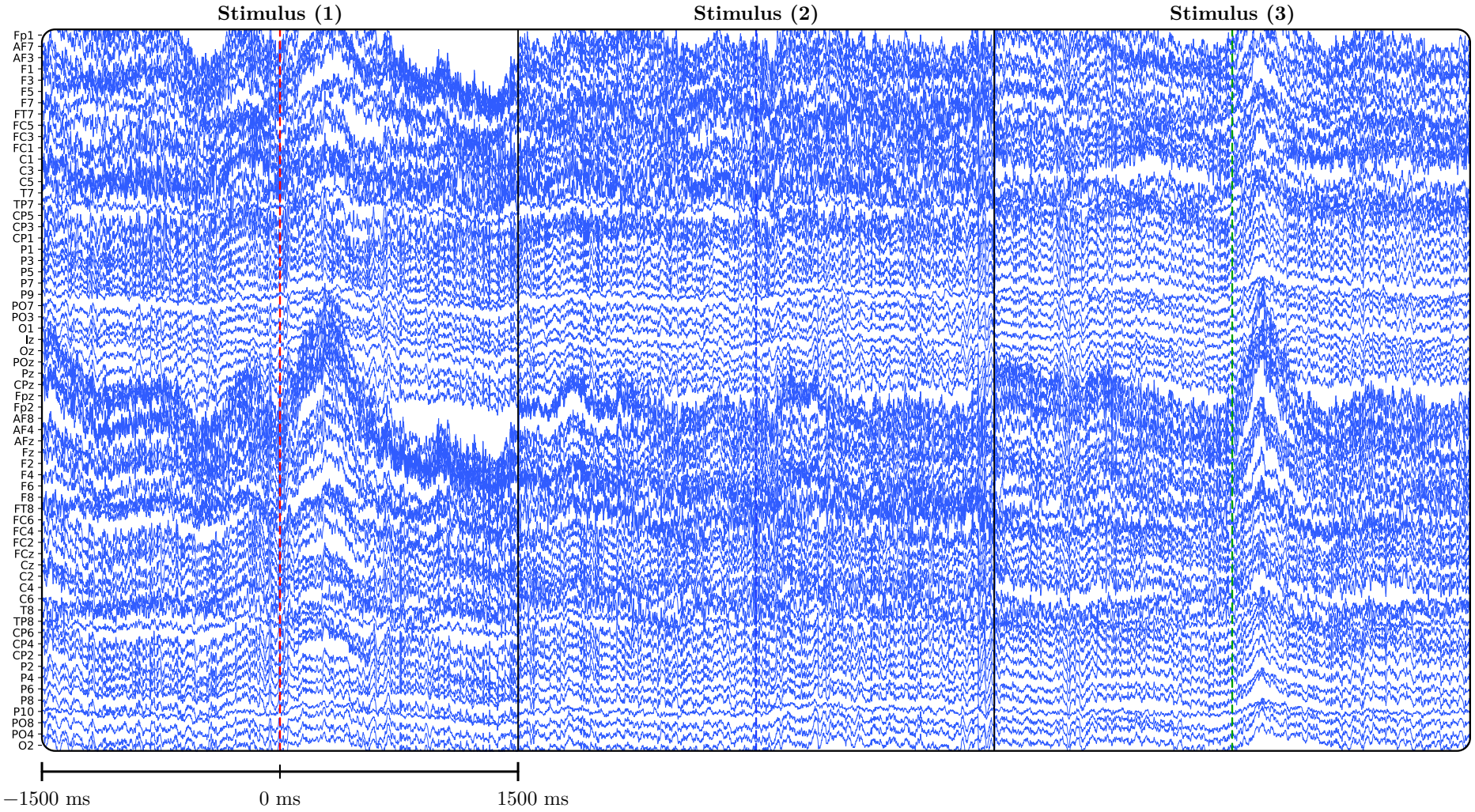
Subject 17 (CTL — bad)



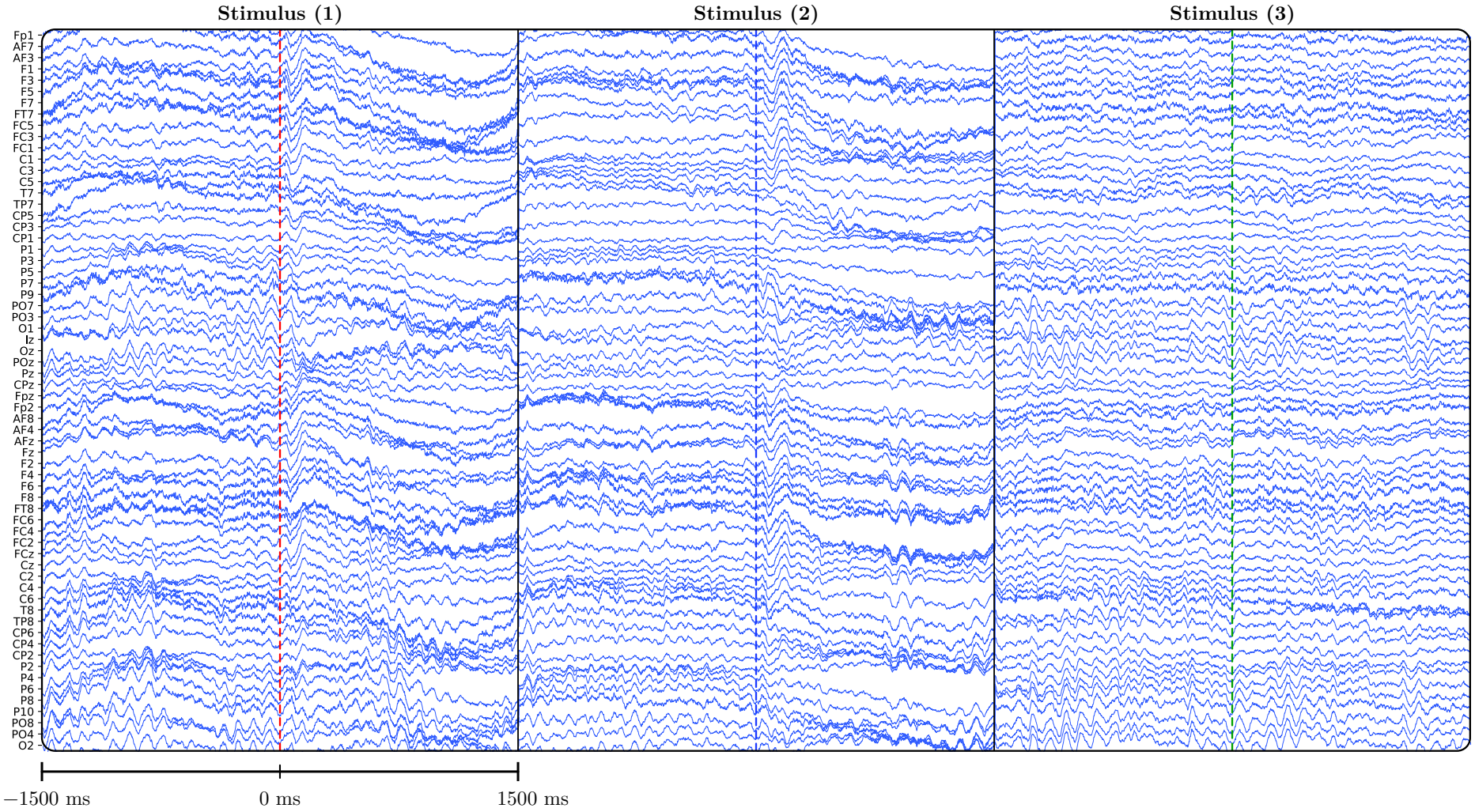
Subject 23 (CTL — bad)



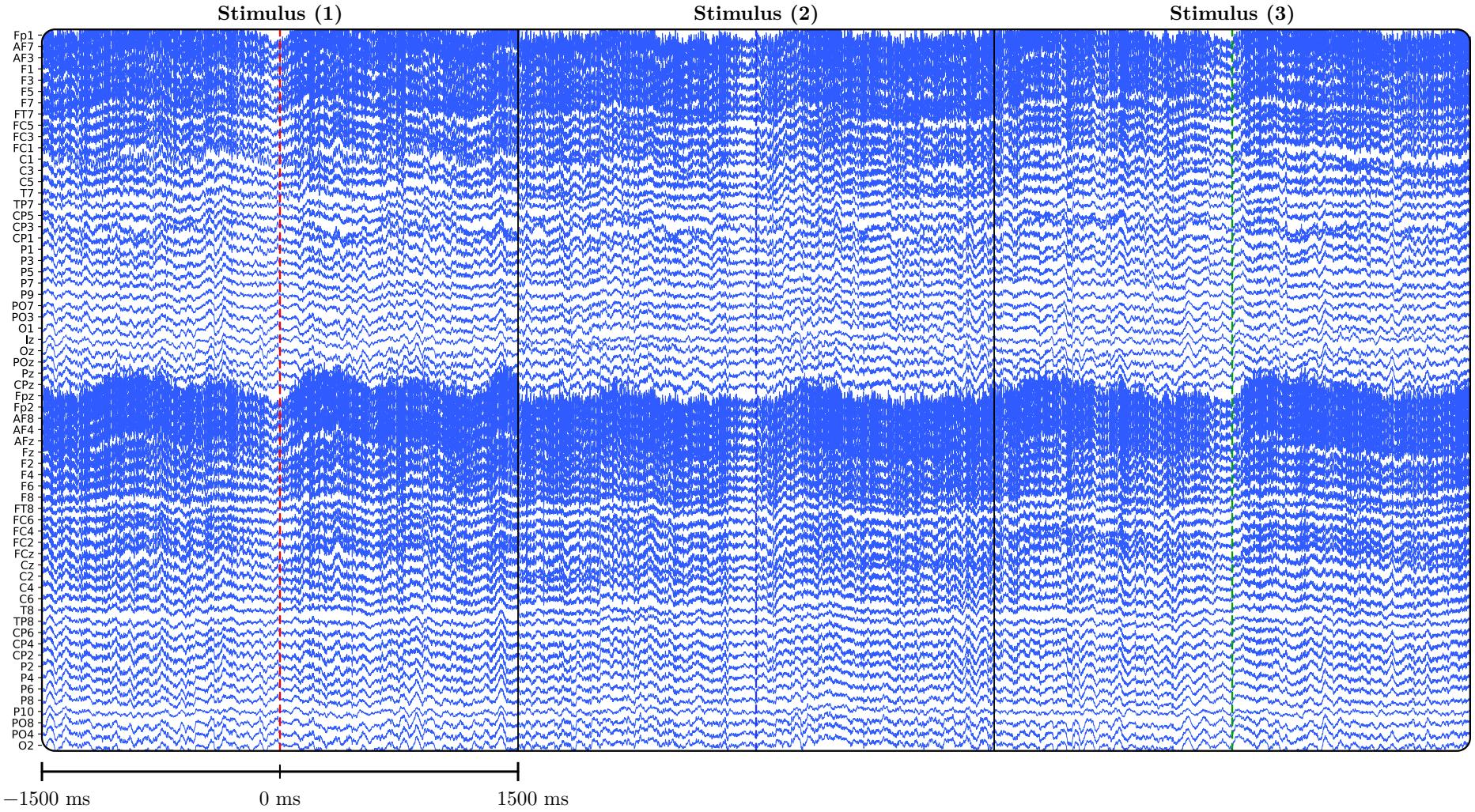
Subject 30 (SZ — bad)



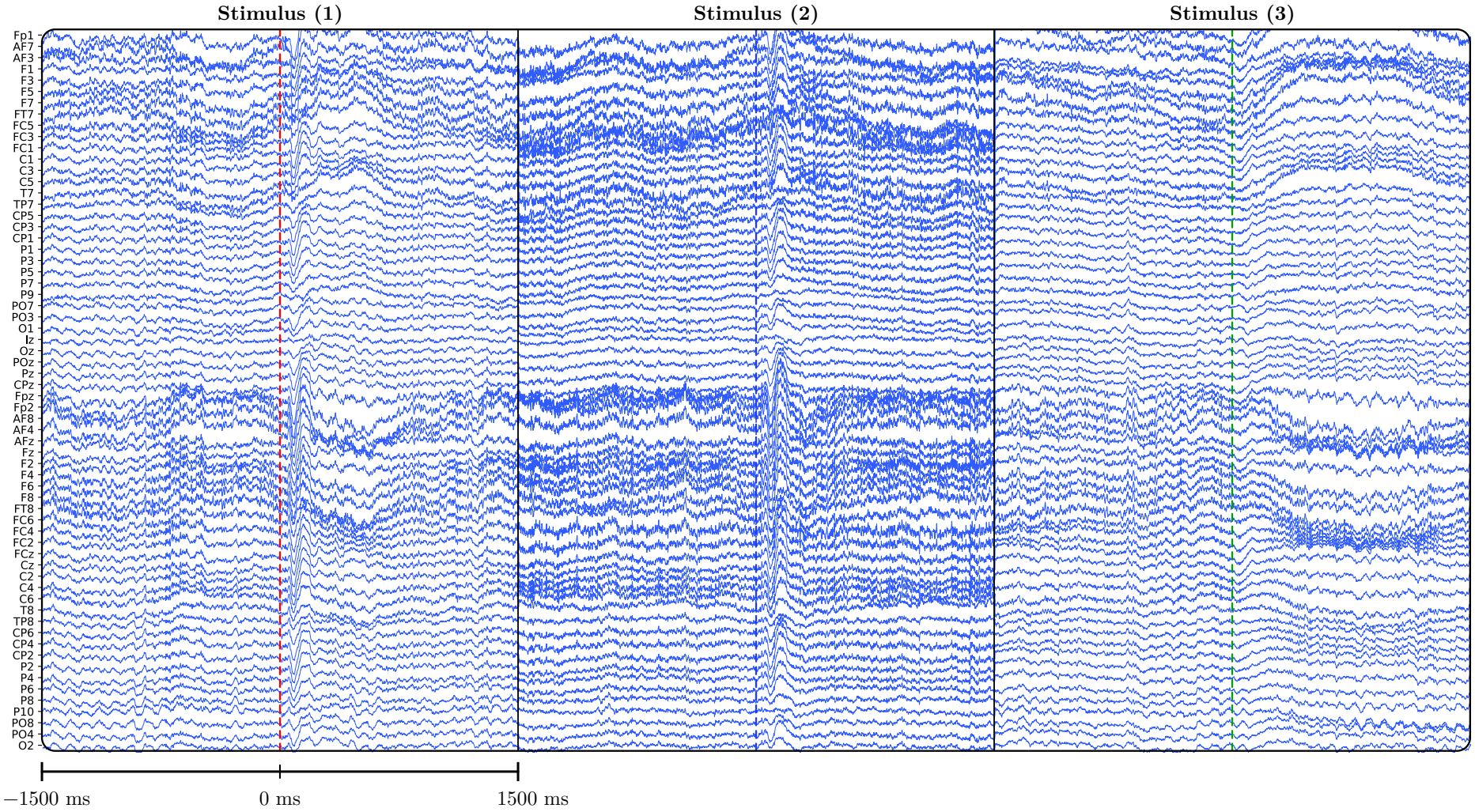
Subject 33 (SZ — bad)



Subject 57 (SZ — bad)



Subject 78 (SZ — bad)



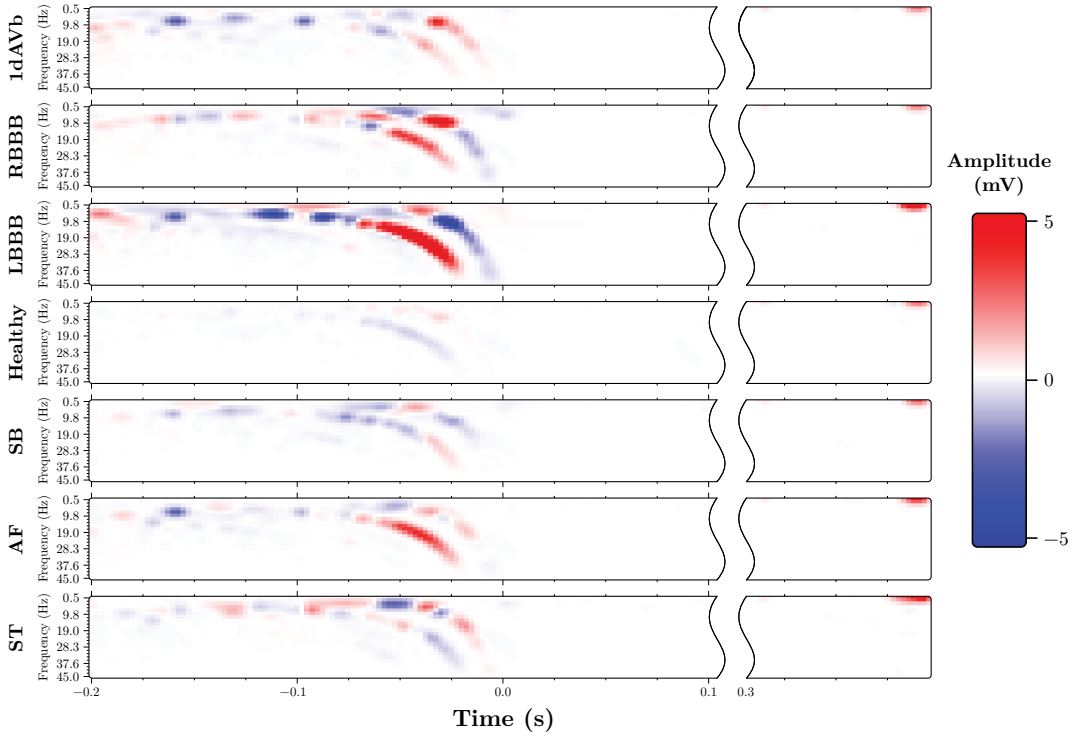
6.2 Appendix B: Cardiac Abnormalities

Abnormality	Lead	Temporal selection	Accuracy (%)							
			no selection no filtering		selection no filtering		selection filtering		no selection filtering	
			learning	test	learning	test	learning	test	learning	test
1dAVb	V3	(37, 79)	82.53	77.57	66.14	55.88	69.59	61.44	82.53	77.57
	V4	(40, 84)	82.20	78.98	68.48	58.18	72.11	65.06	82.20	78.98
	V5	(8, 90)	82.83	80.90	82.83	80.31	85.15	83.05	82.83	80.90
RBBB	V1	(39, 104)	94.25	93.13	94.25	92.47	96.08	94.31	94.25	93.13
	V2	(18, 94)	87.47	83.16	87.57	82.57	90.69	86.12	87.47	83.16
	V3	(20, 100)	83.29	76.81	83.92	77.99	87.82	82.20	83.29	76.81
	V5	(26, 88)	83.81	80.43	83.84	81.39	88.68	83.90	83.81	80.43
LBBB	V1	(6, 84)	95.17	93.43	95.18	93.43	97.02	95.35	95.17	93.43
	V2	(0, 84)	93.89	92.92	93.89	92.92	96.81	94.10	93.89	92.92
	V3	(0, 85)	94.03	92.92	94.03	92.92	96.56	94.17	94.03	92.92
	V4	(0, 85)	92.51	90.26	92.51	90.26	95.77	92.32	92.51	90.26
SB	V2	(27, 81)	74.59	68.09	74.51	69.20	77.78	71.64	74.59	68.09
	V3	(15, 84)	76.31	70.09	75.61	68.69	79.48	71.86	76.31	70.09
	V5	(20, 86)	75.32	68.39	75.03	67.65	78.07	74.30	75.32	68.39
AF	V2	(20, 84)	77.64	71.47	70.53	63.41	76.33	70.36	77.64	71.47
	V3	(20, 84)	79.88	74.35	71.78	63.86	78.01	71.62	79.88	74.35
	V4	(27, 90)	80.57	72.43	73.84	65.34	76.67	72.80	80.57	72.43
ST	V4	(17, 85)	79.89	74.58	79.33	73.69	82.49	77.61	79.89	74.58
	V5	(14, 90)	79.66	74.72	79.23	75.24	82.12	77.75	79.66	74.72

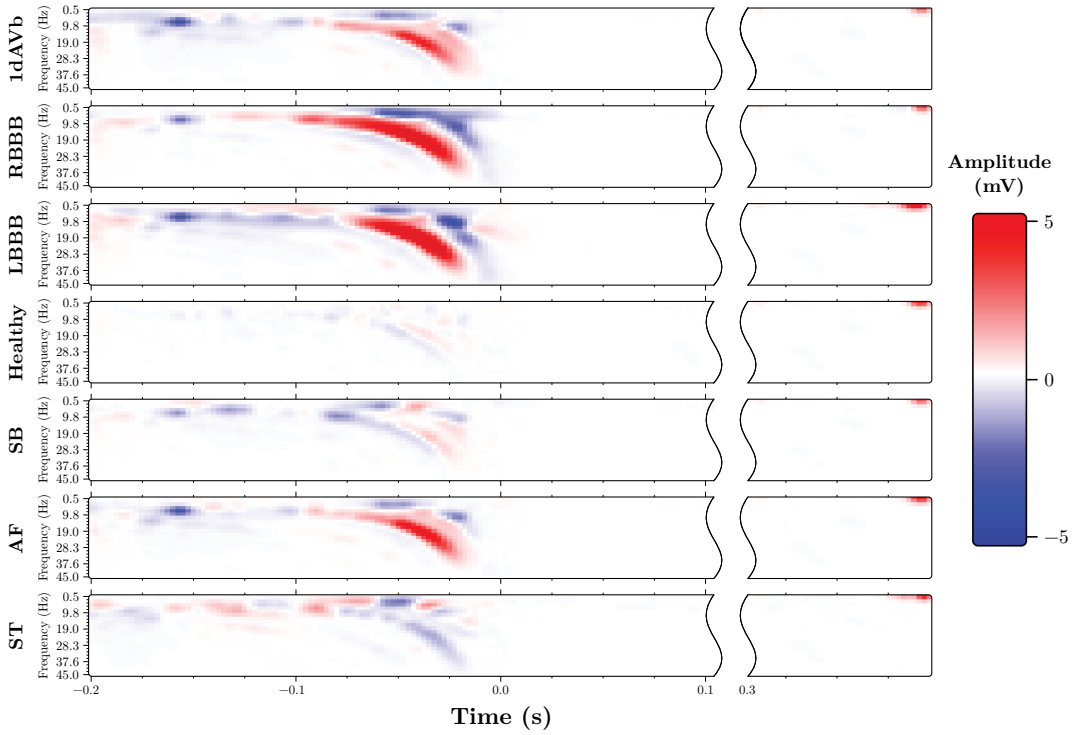
Table 6.4: Result of the classification of all leads that showed visual evidence.

All the visualisation for the leads that were not illustrated previously in Chapter 5, containing the averaged SDFs of all subject within the same category and where the SDFs of healthy validation has been subtracted is shown in the following Figures.

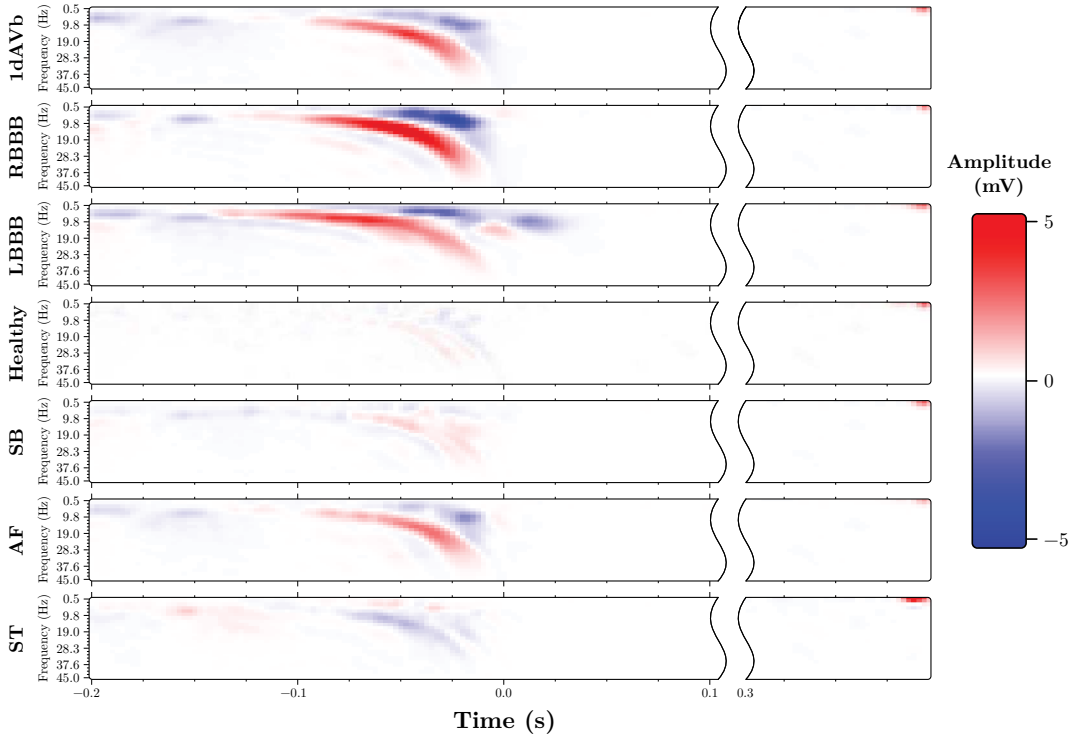
Lead DI



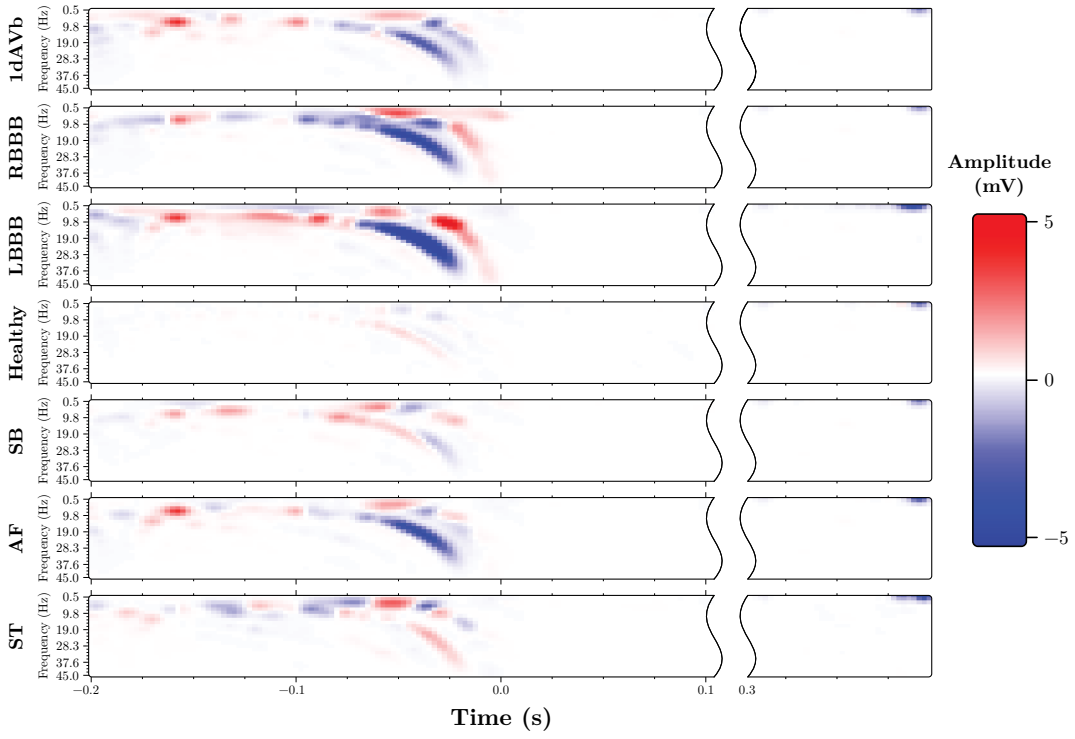
Lead DII



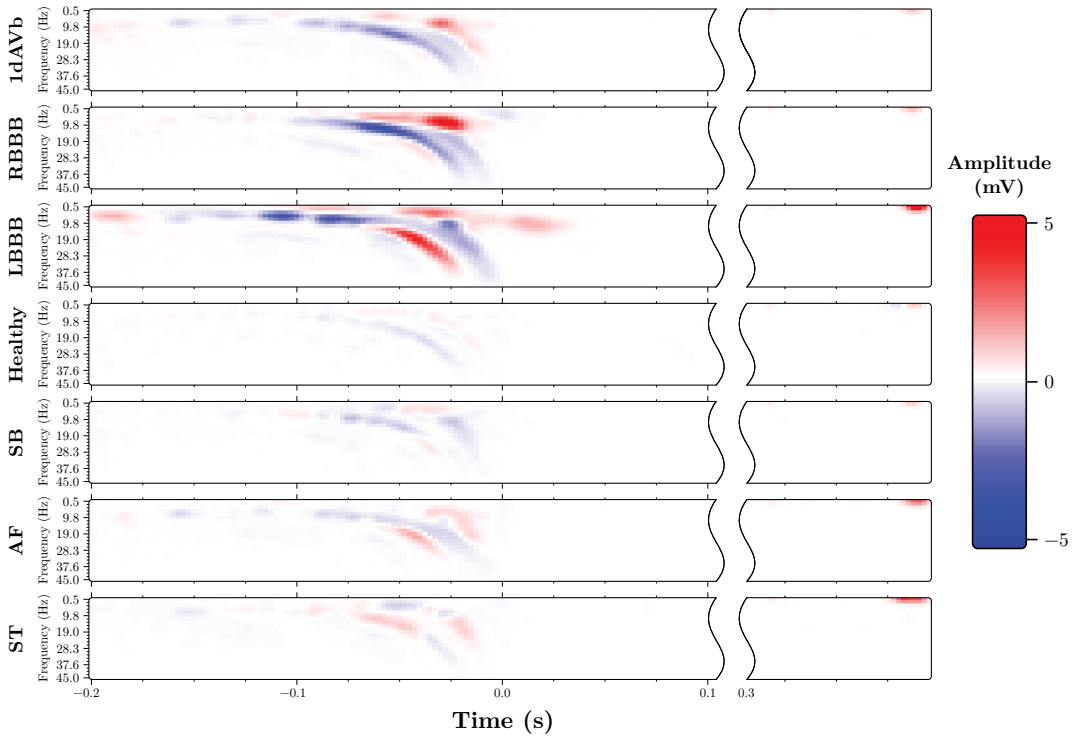
Lead DIII



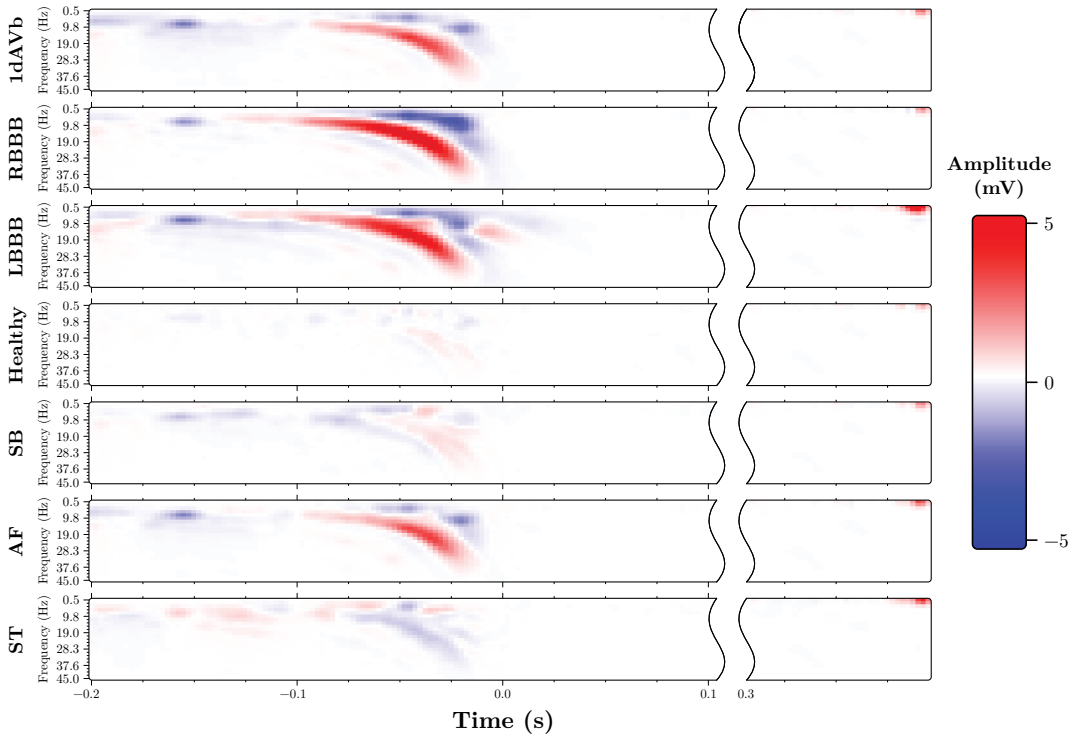
Lead AVR

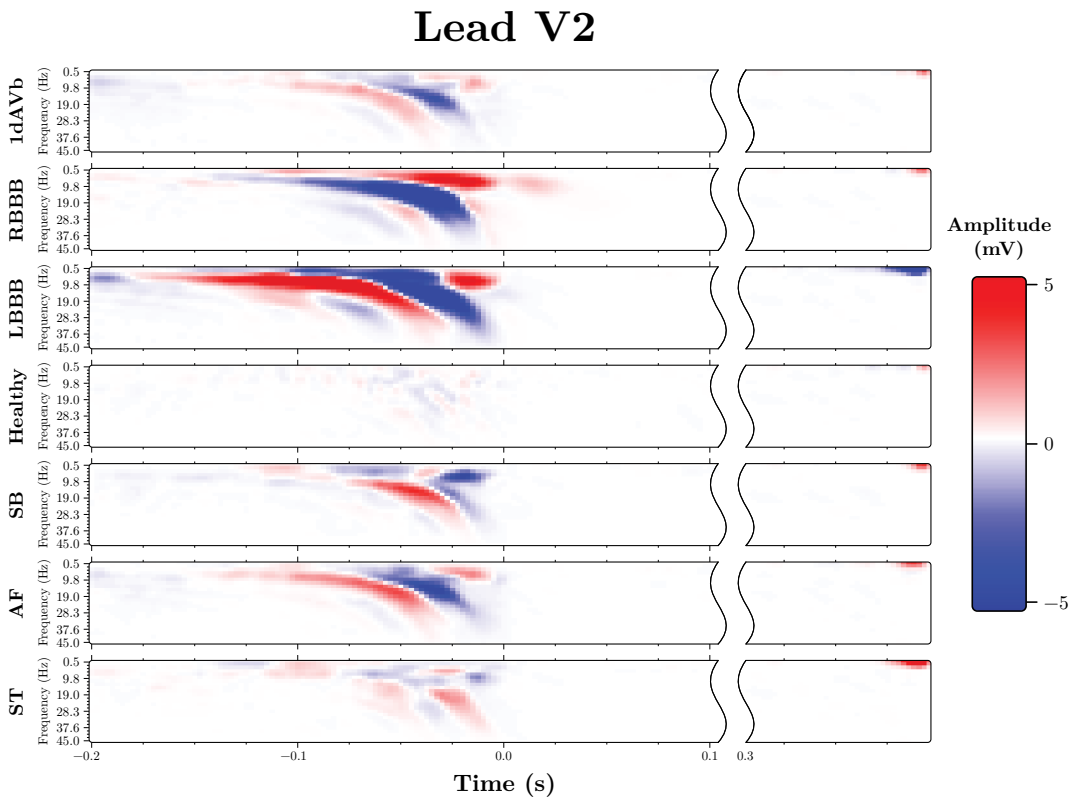


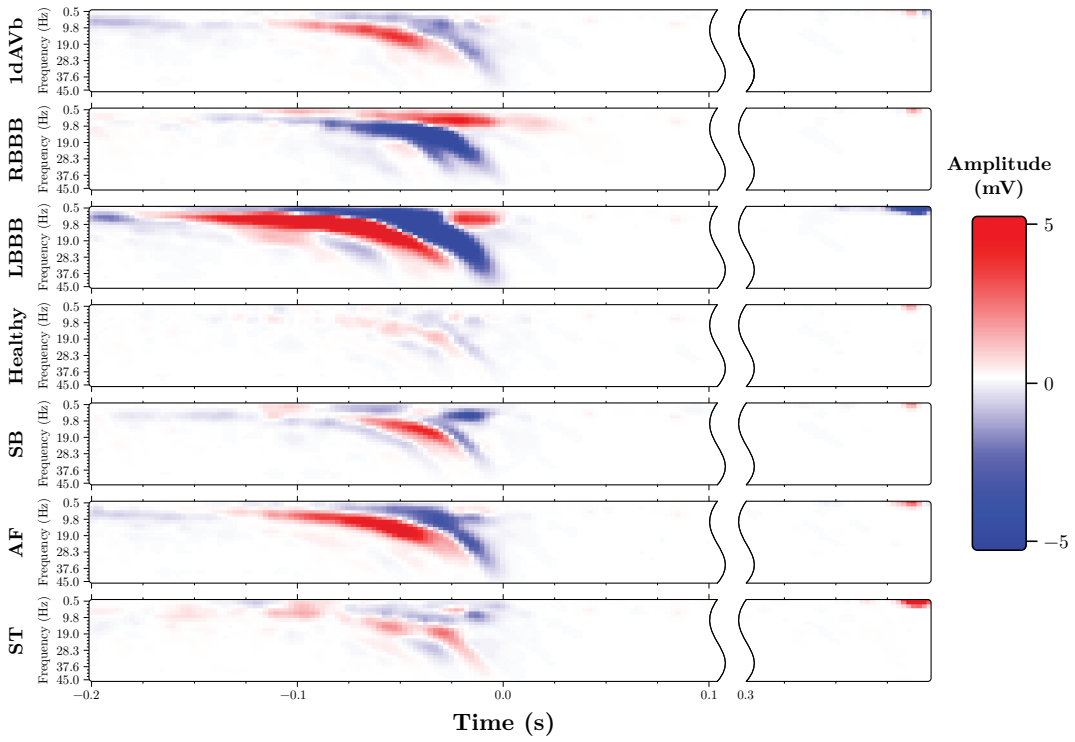
Lead AVL

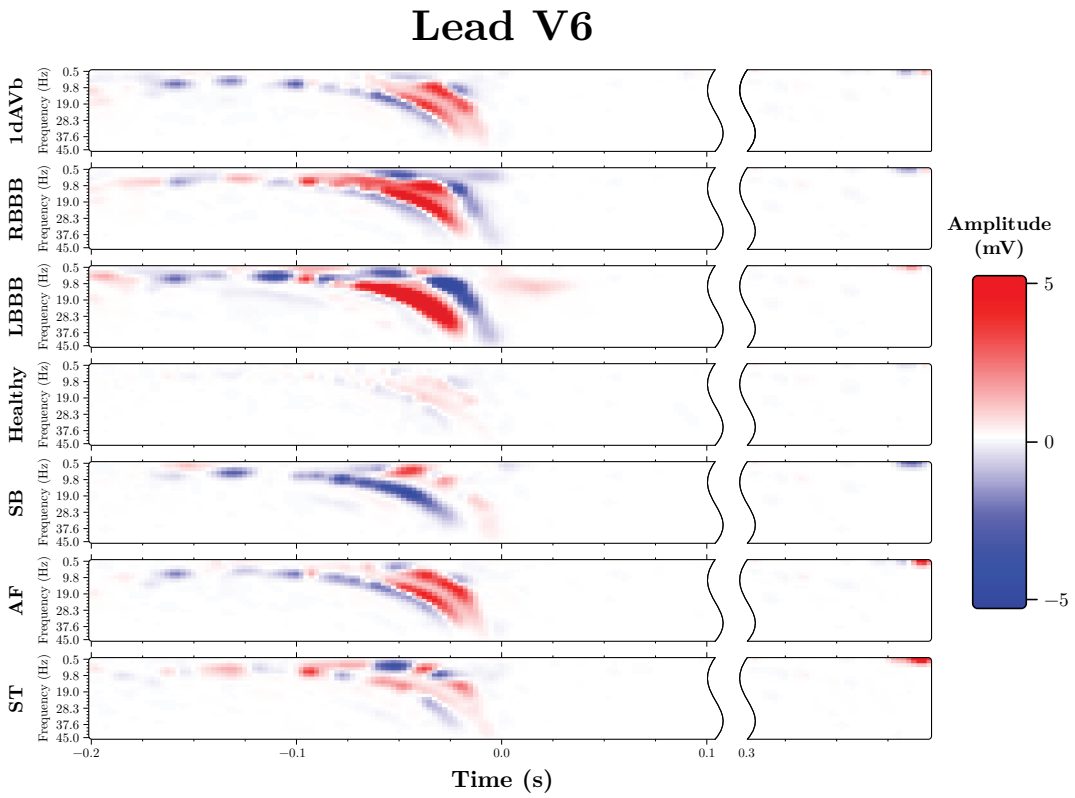


Lead AVF




Lead V2
1dAVb
RBBB
LBBB
Healthy
SB
AF
ST
Frequency (Hz)
0.5
9.8
19.0
28.3
37.6
45.0
−0.2
−0.1
0.0
0.1
0.3
Time (s)
Amplitude (mV)
5
0
−5


Lead V3
1dAVb
RBBB
LBBB
Healthy
SB
AF
ST
Frequency (Hz)
0.5
9.8
19.0
28.3
37.6
45.0
−0.2
−0.1
0.0
0.1
0.3
Time (s)
Amplitude (mV)
5
0
−5



6.3 Appendix C: Effect of data leakage (example)

In this section we will take an example which demonstrates the effect of data leakage; precisely when correlated data is simultaneously contained in the training and test set. To highlight the difference, the same experiment will be carried out for two cases: one with data leakage and the other without. The example concerns the data seen in Chapter 3 which concerns the diagnosis of Parkinson's disease by using a one-dimensional Convolutional Neural Network (1-D CNN). For this example, the data used are the EEG data after being pre-processed and for the *standard* stimulus, as there are more trials for this condition (and therefore more data), which is an important condition for this type of network.

The CNN network will be given as an input a trial (a time window), however, unlike in Chapter 3, here the trial is given in its entirety with its 60 channels. The network will have to learn to differentiate between healthy and parkinsonians on the basis of these signals. As there are around 140 stimuli for each patient, this gives us a total of 7000 data instances for all the patients. (only the off-medication session has been considered).

The type of network used is the same as in [67], except that we selected the size of the network and its architecture ourselves (the number of filters, the size of the filters, etc.). The network used is illustrated in Figure 6.1, where we can observe that the network used is much smaller than that used in [67].

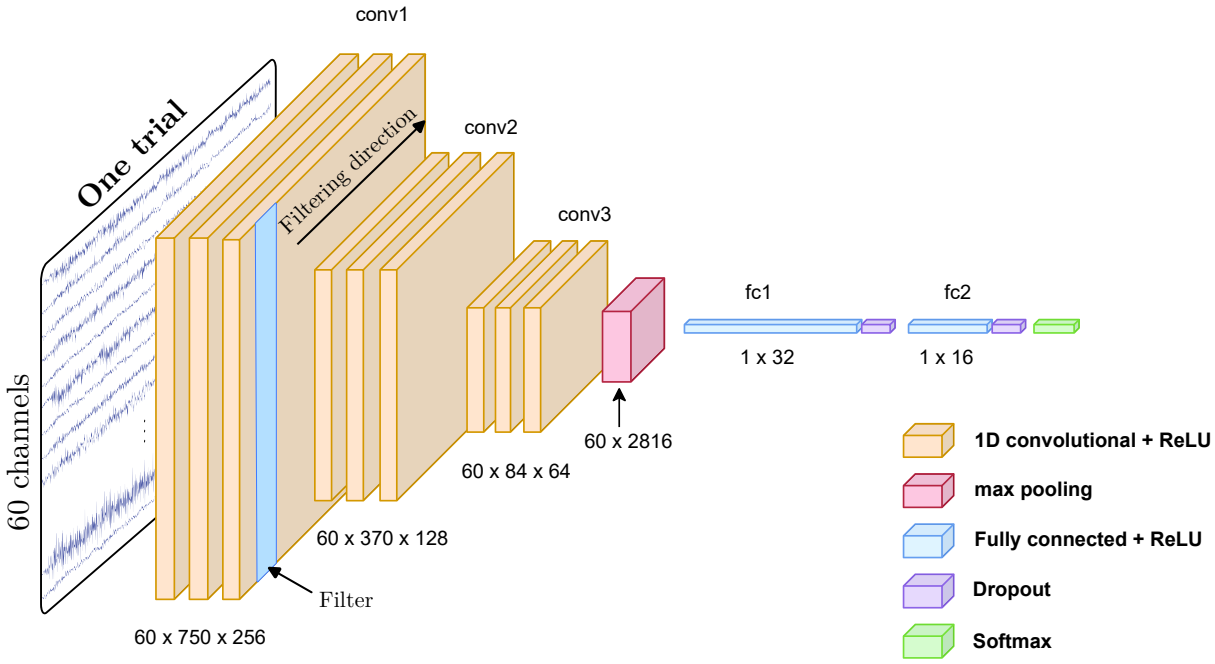


Figure 6.1: 1D-CNN architecture used.

For both cases, the same training parameters and procedure were used. The training, evaluation and splitting of data into training and test were repeated 100 times to obtain statistics given the randomness of the network and the splits. As an indication, the training set corresponds to 90% of the signals and the remaining 10% constitutes the test set.

The first scenario is when there is no data leakage. If attention is paid to ensuring that signals from the same patient are not present in both the training and test sets, the results shown in Figure 6.2 are obtained.

It can be observed that the training accuracy is around 100% and that the test accuracy does not decrease over epochs. This shows that the model does not over-fit the training data, and that the training was stopped at the right time. The final average test accuracy was around 62%.

In the case of the presence of data leakage, in other words, the data instances were shuffled directly and without taking into account that signals from the same patient should not be found in the training and test sets, the obtained results are shown in Figure 6.3.

It can be observed that the training accuracy is practically the same as that of the case without data leakage, which shows that the complexity of the data has not changed and that the network has been able to learn in the same manner in both cases. However, in terms of test accuracy, the final accuracy is around 98%, which is an increase of 36% that is simply due to the fact that the same patient's data can be found in both the

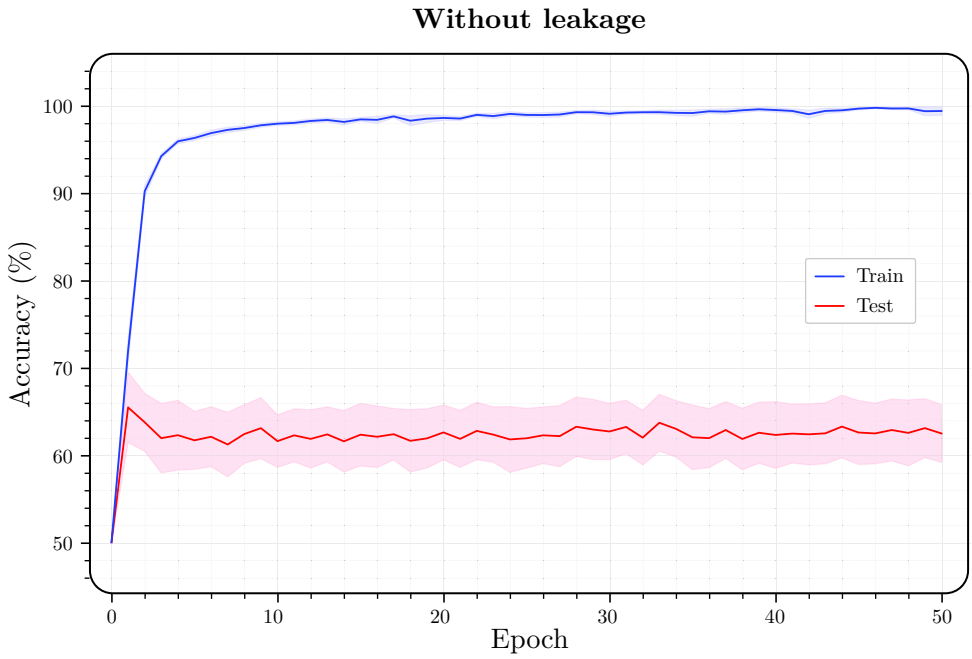


Figure 6.2: Results without data-leakage with 95% confidence interval.

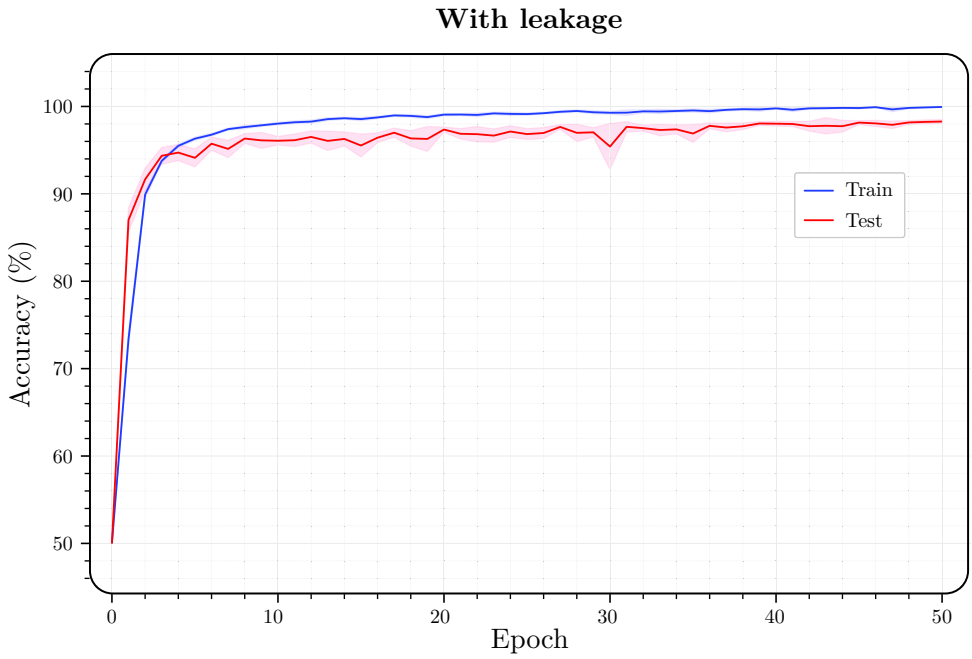


Figure 6.3: Results with data-leakage with 95% confidence interval.

training and test sets. We hypothesise that this is due to the fact that the network is able to recognise the patient's signature because the network has a strong learning capability, in addition to the fact that there is limited amount of data. The effect of this data leakage is mitigated by the availability of more data.

Bibliography

[1] C. H. Adler et al. “Low clinical diagnostic accuracy of early vs advanced Parkinson disease: Clinicopathologic study”. In: *Neurology* 83.5 (June 2014), pp. 406–412. DOI: 10.1212/wnl.0000000000000641.

[2] Maria Beatriz Alkmim et al. “Improving patient access to specialized health care: the Telehealth Network of Minas Gerais, Brazil”. In: *Bulletin of the World Health Organization* 90.5 (May 2012), pp. 373–378. ISSN: 0042-9686. DOI: 10.2471/blt.11.099408.

[3] APA. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Association Publishing, Mar. 2022. DOI: 10.1176/appi.books.9780890425787.

[4] Blaine Ayotte et al. “Group leakage overestimates performance: A case study in keystroke dynamics”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2021), pp. 1410–1417. ISSN: 21607516. DOI: 10.1109/CVPRW53098.2021.00156.

[5] R. Balestrino and A. H.V. Schapira. “Parkinson disease”. In: *European Journal of Neurology* 27.1 (2020), pp. 27–42. ISSN: 14681331. DOI: 10.1111/ene.14108.

[6] Erol Başar et al. “Gamma, alpha, delta, and theta oscillations govern cognitive processes”. In: *International Journal of Psychophysiology* 39.2-3 (2001), pp. 241–248. ISSN: 01678760. DOI: 10.1016/S0167-8760(00)00145-8.

[7] Carl C. Bell. “DSM-IV: Diagnostic and Statistical Manual of Mental Disorders”. In: *JAMA: The Journal of the American Medical Association* 272.10 (Sept. 1994), p. 828. DOI: 10.1001/jama.1994.03520100096046.

[8] Richard E Bellman. *Dynamic programming*. en. Princeton, NJ: Princeton University Press, Oct. 1957.

[9] A. Berardelli et al. “EFNS/MDS-ES recommendations for the diagnosis of Parkinson’s disease”. In: *European Journal of Neurology* 20.1 (2013), pp. 16–34. ISSN: 13515101. DOI: 10.1111/ene.12022.

[10] Ankit A. Bhurane et al. “Diagnosis of Parkinson’s disease from electroencephalography signals using linear and self-similarity features”. In: *Expert Systems* June (2019), pp. 1–12. ISSN: 14680394. DOI: 10.1111/exsy.12472.

[11] György Buzsaki and Andreas Draguhn. “Neuronal Oscillations in Cortical Networks”. In: *Science* 304.5679 (June 2004), pp. 1926–1929. DOI: 10.1126/science.1099745.

[12] James F. Cavanagh et al. “Diminished EEG habituation to novel events effectively classifies Parkinson’s patients”. In: *Clinical Neurophysiology* 129.2 (2018), pp. 409–418. ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2017.11.023.

[13] James F. Cavanagh et al. “The patient repository for EEG data + computational tools (PRED+CT)”. In: *Frontiers in Neuroinformatics* 11.November (2017), pp. 1–9. ISSN: 16625196. DOI: 10.3389/fninf.2017.00067.

[14] Mustafa S. Cetin et al. “Multimodal Classification of Schizophrenia Patients with MEG and fMRI Data Using Static and Dynamic Connectivity Measures”. In: *Frontiers in Neuroscience* 10 (Oct. 2016). DOI: 10.3389/fnins.2016.00466.

[15] Fiona J Charlson et al. “Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016”. In: *Schizophrenia Bulletin* 44.6 (May 2018), pp. 1195–1203. DOI: 10.1093/schbul/sby058.

[16] K. Ray Chaudhuri, L. Yates, and P. Martinez-Martin. “The non-motor symptom complex of Parkinson’s disease: A comprehensive assessment is essential”. In: *Current Neurology and Neuroscience Reports* 5.4 (2005), pp. 275–283. ISSN: 15284042. DOI: 10.1007/s11910-005-0072-6.

[17] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[18] Arnaud Delorme and Scott Makeig. “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis”. In: *Journal of Neuroscience Methods* 134.1 (2004), pp. 9–21. ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2003.10.009.

[19] Christ Devia et al. “EEG Classification During Scene Free-Viewing for Schizophrenia Detection”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.6 (June 2019), pp. 1193–1199. DOI: 10.1109/tnsre.2019.2913799.

[20] Bruno Dietsche, Tilo Kircher, and Irina Falkenberg. “Structural brain changes in schizophrenia at different stages of the illness: A selective review of longitudinal magnetic resonance imaging studies”. In: *Australian & New Zealand Journal of Psychiatry* 51.5 (Mar. 2017), pp. 500–508. DOI: 10.1177/0004867417699473.

[21] E. Ray Dorsey et al. “Global, regional, and national burden of Parkinson’s disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016”. In: *The Lancet Neurology* 17.11 (Nov. 2018), pp. 939–953. DOI: 10.1016/s1474-4422(18)30295-3.

[22] Harper Douglas. ““Etymology of schizophrenia””. In: *Online Etymology Dictionary* (). URL: https://www.etymonline.com/word/schizophrenia.

[23] Bradley Efron et al. “Least angle regression”. In: *The Annals of Statistics* 32.2 (Apr. 2004).

[24] Alberto Fernández et al. “Lempel–Ziv complexity in schizophrenia: A MEG study”. In: *Clinical Neurophysiology* 122.11 (Nov. 2011), pp. 2227–2235. DOI: 10.1016/j.clinph.2011.04.011.

[25] Natalia Ferrazoli et al. “The Application of P300-Long-Latency Auditory-Evoked Potential in Parkinson Disease”. In: *International Archives of Otorhinolaryngology* 26.1 (2022), e158–e166. ISSN: 18094864. DOI: 10.1055/s-0040-1722250.

[26] Ronald Aylmer Fisher. *Design of Experiments*. Vol. 1. 1. 1936. Chap. 2.7, p. 554.

[27] Judith M. Ford et al. “Did I Do That? Abnormal Predictive Processes in Schizophrenia When Button Pressing to Deliver a Tone”. In: *Schizophrenia Bulletin* 40.4 (July 2013), pp. 804–812. DOI: 10.1093/schbul/sbt072.

[28] Yoav Freund and Robert E. Schapire. “A desicion-theoretic generalization of on-line learning and an application to boosting”. In: *Computational Learning Theory*. Springer Berlin Heidelberg, 1995, pp. 23–37. ISBN: 9783540491958. DOI: 10.1007/3-540-59119-2_166.

[29] Pascal Fries. “A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence”. In: *Trends in Cognitive Sciences* 9.10 (2005), pp. 474–480. ISSN: 13646613. DOI: 10.1016/j.tics.2005.08.011.

[30] Jim Frost. *Hypothesis testing*. Statistics by Jim Publishing, Sept. 2020.

[31] Wolfgang Gaebel et al. “Trends in Schizophrenia Diagnosis and Treatment”. In: *Advances in Psychiatry*. Springer International Publishing, July 2018, pp. 603–619. DOI: 10.1007/978-3-319-70554-5_35.

[32] GHDx. *Global Health Data Exchange, Institute of health Metrics and Evaluation (IHME) [Online]*. Accessed 25 September 2021. 2021. URL: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/27a7644e8ad28e739382d31e77589dd7.

[33] W R Gibb and A J Lees. “The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson’s disease.” In: *Journal of Neurology, Neurosurgery and Psychiatry* 51.6 (June 1988), pp. 745–752. DOI: 10.1136/jnnp.51.6.745.

[34] W R Gibb and A J Lees. “The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson’s disease.” In: *Journal of Neurology, Neurosurgery & Psychiatry* 51.6 (1988), pp. 745–752. ISSN: 0022-3050. DOI: 10.1136/jnnp.51.6.745.

[35] Alexandre Gramfort et al. “MEG and EEG data analysis with MNE-Python”. In: *Frontiers in Neuroscience* 7.7 DEC (2013), pp. 1–13. ISSN: 1662453X. DOI: 10.3389/fnins.2013.00267.

[36] Michael F. Green et al. “Social Cognition in Schizophrenia, Part 1: Performance Across Phase of Illness”. In: *Schizophrenia Bulletin* 38.4 (Feb. 2011), pp. 854–864. DOI: 10.1093/schbul/sbq171.

[37] P. Hamilton. “Open source ECG analysis”. In: *Computers in Cardiology*. CIC-02. IEEE. DOI: 10.1109/cic.2002.1166717.

[38] Chun Xiao Han et al. “Investigation of EEG abnormalities in the early stage of Parkinson’s disease”. In: *Cognitive Neurodynamics* 7.4 (2013), pp. 351–359. ISSN: 18714080. DOI: 10.1007/s11571-013-9247-z.

[39] Edward C. Hansch et al. “Cognition in Parkinson disease: An event-related potential perspective”. In: *Annals of Neurology* 11.6 (1982), pp. 599–607. ISSN: 15318249. DOI: 10.1002/ana.410110608.

[40] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. New York, NY: Springer, Mar. 2009.

[41] Krawiec C Henning A. *Sinus Tachycardia. [Online]*. Accessed 14 November 2023. [Updated 2023 Mar 5]. URL: https://www.ncbi.nlm.nih.gov/books/NBK553128/.

[42] Vasileios Ioakeimidis et al. “A Meta-analysis of Structural and Functional Brain Abnormalities in Early-Onset Schizophrenia”. In: *Schizophrenia Bulletin Open* 1.1 (Jan. 2020). DOI: 10.1093/schizbullopen/sgaa016.

[43] René S. Kahn et al. “Schizophrenia”. In: *Nature Reviews Disease Primers* 1.1 (Nov. 2015). DOI: 10.1038/nrdp.2015.67.

[44] Lorraine V Kalia and Anthony E Lang. “Parkinson’s disease”. In: *The Lancet* 386.9996 (Aug. 2015), pp. 896–912. DOI: 10.1016/s0140-6736(14)61393-3.

[45] Shachar Kaufman et al. “Leakage in data mining: Formulation, detection, and avoidance”. In: *ACM Transactions on Knowledge Discovery from Data* 6.4 (2012). ISSN: 15564681. DOI: 10.1145/2382577.2382579.

[46] Smith K. Khare and Varun Bajaj. “A hybrid decision support system for automatic detection of Schizophrenia using EEG signals”. In: *Computers in Biology and Medicine* 141 (Feb. 2022), p. 105028. DOI: 10.1016/j.compbiomed.2021.105028.

[47] Fred M. Kusumoto et al. “2018 ACC/AHA/HRS Guideline on the Evaluation and Management of Patients With Bradycardia and Cardiac Conduction Delay”. In: *Journal of the American College of Cardiology* 74.7 (Aug. 2019), e51–e156. ISSN: 0735-1097. DOI: 10.1016/j.jacc.2018.10.044.

[48] Thomas Munk Laursen, Merete Nordentoft, and Preben Bo Mortensen. “Excess Early Mortality in Schizophrenia”. In: *Annual Review of Clinical Psychology* 10.1 (2014), pp. 425–448. DOI: 10.1146/annurev-clinpsy-032813-153657.

[49] Leonard S Lilly. *Pathophysiology of heart disease*. 6th ed. Wolters Kluwer Health, July 2015.

[50] R. J. Linscott and J. van Os. “An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders”. In: *Psychological Medicine* 43.6 (July 2012), pp. 1133–1149. DOI: 10.1017/s0033291712001626.

[51] Simon Little and Peter Brown. “The functional role of beta oscillations in Parkinson’s disease”. In: *Parkinsonism &amp Related Disorders* 20 (Jan. 2014), S44–S48. DOI: 10.1016/s1353-8020(13)70013-0.

[52] Guotao Liu et al. “Complexity Analysis of Electroencephalogram Dynamics in Patients with Parkinson’s Disease”. In: *Parkinson’s Disease* 2017 (2017), pp. 1–9. DOI: 10.1155/2017/8701061.

[53] Carlo J. De Luca. “The Use of Surface Electromyography in Biomechanics”. In: *Journal of Applied Biomechanics* 13.2 (May 1997), pp. 135–163. DOI: 10.1123/jab.13.2.135.

[54] Pion-Tonachini Luca. “ICLabel Tutorial: EEG Independent Component Labeling”. In: *Swartz Center for Computational Neuroscience, University of California, San Diego* (). URL: https://labeling.ucsd.edu/tutorial (visited on 03/21/2021).

[55] Steven J Luck. *An introduction to the event-related potential technique*. en. Cognitive neuroscience. Cambridge, MA: Bradford Books, Sept. 2005.

[56] Steven J. Luck. “Event-related potentials.” In: *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics*. American Psychological Association, 2012, pp. 523–546. DOI: 10.1037/13619-028.

[57] Aurore Lyon et al. “Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances”. In: *Journal of The Royal Society Interface* 15.138 (Jan. 2018), p. 20170821. ISSN: 1742-5662. DOI: 10.1098/rsif.2017.0821.

[58] Robert A. McCutcheon, Tiago Reis Marques, and Oliver D. Howes. “Schizophrenia—An Overview”. In: *JAMA Psychiatry* 77.2 (Feb. 2020), p. 201. DOI: 10.1001/jamapsychiatry.2019.3360.

[59] “Metacognitive function and fragmentation in schizophrenia: Relationship to cognition, self-experience and developing treatments”. In: *Schizophrenia Research: Cognition* 19 (2020), p. 100142. ISSN: 2215-0013. DOI: https://doi.org/10.1016/j.scog.2019.100142.

[60] A. M. Molinaro, R. Simon, and R. M. Pfeiffer. “Prediction error estimation: a comparison of resampling methods”. In: *Bioinformatics* 21.15 (May 2005), pp. 3301–3307. DOI: 10.1093/bioinformatics/bti499.

[61] Caroline Moreau et al. “Intraventricular dopamine infusion alleviates motor symptoms in a primate model of Parkinson’s disease”. In: *Neurobiology of Disease* 139 (2020), p. 104846. ISSN: 0969-9961. DOI: https://doi.org/10.1016/j.nbd.2020.104846. URL: https://www.sciencedirect.com/science/article/pii/S0969996120301212.

[62] Tumbwene E. Mwansisya et al. “Task and resting-state fMRI studies in first-episode schizophrenia: A systematic review”. In: *Schizophrenia Research* 189 (Nov. 2017), pp. 9–18. DOI: 10.1016/j.schres.2017.02.026.

[63] NHS. *Atrial Fibrilation. [Online]*. Accessed 14 November 2023. 17 May, 2021. URL: https://www.nhs.uk/conditions/atrial-fibrillation/.

[64] NHS. *Schizophrenia. [Online]*. Accessed 31 October 2023. 10 Jan, 2022. URL: https://www.nhs.uk/mental-health/conditions/schizophrenia/symptoms/.

[65] Thomas E. Nichols and Andrew P. Holmes. "Nonparametric Permutation Tests for Functional Neuroimaging". In: *Human Brain Function: Second Edition* 25.July (2003), pp. 887–910. DOI: 10.1016/B978-012264841-0/50048-2.

[66] H. Nolan, R. Whelan, and R.B. Reilly. "FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection". In: *Journal of Neuroscience Methods* 192.1 (Sept. 2010), pp. 152–162. DOI: 10.1016/j.jneumeth.2010.07.015.

[67] Shu Lih Oh et al. "A deep learning approach for Parkinson's disease diagnosis from EEG signals". In: *Neural Computing and Applications* 32.15 (2020), pp. 10927–10933. ISSN: 14333058. DOI: 10.1007/s00521-018-3689-5.

[68] Piotr Olejniczak. "Neurophysiologic Basis of EEG". In: *Journal of Clinical Neurophysiology* 23.3 (June 2006), pp. 186–189. DOI: 10.1097/01.wnp.0000220079.61973.6c.

[69] Toshiaki Onitsuka et al. "Review of neurophysiological findings in patients with schizophrenia". In: *Psychiatry and Clinical Neurosciences* 67.7 (Sept. 2013), pp. 461–470. DOI: 10.1111/pcn.12090.

[70] World Health Organization. *Parkinson disease: a public health approach: technical brief*. World Health Organization, 2022, iv, 21 p. ISBN: 9789240050983.

[71] Lara N. Pantlin and Deana Davalos. "Neurophysiology for Detection of High Risk for Psychosis". In: *Schizophrenia Research and Treatment* 2016 (2016), pp. 1–5. DOI: 10.1155/2016/2697971.

[72] David L. Penn et al. "Social cognition in schizophrenia." In: *Psychological Bulletin* 121.1 (1997), pp. 114–132. DOI: 10.1037/0033-2909.121.1.114.

[73] John Polich. "Updating P300: An integrative theory of P3a and P3b". In: *Clinical Neurophysiology* 118.10 (2007), pp. 2128–2148. ISSN: 13882457. DOI: 10.1016/j.clinph.2007.04.019.

[74] Marios Politis et al. "Parkinson's disease symptoms: The patient's perspective". In: *Movement Disorders* 25.11 (May 2010), pp. 1646–1651. DOI: 10.1002/mds.23135.

[75] Shreya Prabhu and Roshan Joy Martis. "Diagnosis of Schizophrenia using Kolmogorov Complexity and Sample Entropy". In: *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. IEEE, July 2020. DOI: 10.1109/conecct50063.2020.9198472.

[76] Joaquin Quinonero-Candela et al., eds. *Dataset Shift in Machine Learning*. Neural Information Processing series. London, England: MIT Press, June 2022.

[77] Joaquin Quinonero-Candela et al., eds. *Dataset Shift in Machine Learning*. Neural Information Processing series. London, England: MIT Press, June 2022.

[78] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. "The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations". In: *Movement Disorders* 18.7 (2003), pp. 738–750. DOI: https://doi.org/10.1002/mds.10473.

[79] Antônio H. Ribeiro et al. "Automatic Diagnosis of the 12-Lead ECG Using a Deep Neural Network". In: *Nature Communications* 11.1 (2020), p. 1760. DOI: https://doi.org/10.1038/s41467-020-15432-4.

[80] Giovanni Rizzo et al. "Accuracy of clinical diagnosis of Parkinson disease". In: *Neurology* 86.6 (2016), pp. 566–576. ISSN: 1526632X. DOI: 10.1212/WNL.0000000000002350.

[81] Brian Roach. *EEG data from basic sensory task in schizophrenia — button press and auditory tone event related potentials from 81 human subjects. [Online]*. Accessed 06 November 2022. 2021. URL: https://www.kaggle.com/datasets/broach/button-tone-sz.

[82] Gregory A Roth et al. "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (Nov. 2018), pp. 1736–1788. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(18)32203-7.

[83] M Rotman and J H Triebwasser. "A clinical and follow-up study of right and left bundle branch block." In: *Circulation* 51.3 (Mar. 1975), pp. 477–484. ISSN: 1524-4539. DOI: 10.1161/01.cir.51.3.477.

[84] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[85] Sukanta Saha et al. "A Systematic Review of the Prevalence of Schizophrenia". In: *PLoS Medicine* 2.5 (May 2005). Ed. by Steven E. Hyman, e141. DOI: 10.1371/journal.pmed.0020141.

[86] Lorenzo Santos-Mayo, Luis M. San-Jose-Revuelta, and Juan Ignacio Arribas. "A Computer-Aided Diagnosis System With EEG Based on the P3b Wave During an Auditory Odd-Ball Task in Schizophrenia". In: *IEEE Transactions on Biomedical Engineering* 64.2 (Feb. 2017), pp. 395–407. DOI: 10.1109/tbme.2016.2558824.

[87] K.K. Shung, M. Smith, and B.M.W. Tsui. *Principles of Medical Imaging*. Elsevier Science, 2012. ISBN: 9780323139939.

[88] Konstantinos C. Siontis et al. "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management". In: *Nature Reviews Cardiology* 18.7 (Feb. 2021), pp. 465–478. ISSN: 1759-5010. DOI: 10.1038/s41569-020-00503-2.

[89] Yan-Yan Song and Ying Lu. "Decision tree methods: applications for classification and prediction". en. In: *Shanghai Arch. Psychiatry* 27.2 (Apr. 2015), pp. 130–135.

[90] Mahsa Soufineyestani, Dale Dowling, and Arshia Khan. "Electroencephalography (EEG) technology applications and available devices". In: *Applied Sciences (Switzerland)* 10.21 (2020), pp. 1–23. ISSN: 20763417. DOI: 10.3390/app10217453.

[91] S. E. Starkstein et al. "Evoked potentials, reaction time and cognitive performance in on and off phases of Parkinson's disease". In: *Journal of Neurology Neurosurgery and Psychiatry* 52.3 (1989), pp. 338–340. ISSN: 00223050. DOI: 10.1136/jnnp.52.3.338.

[92] M. Stone. "Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 38.1 (1976), pp. 102–102. DOI: 10.1111/j.2517-6161.1976.tb01573.x.

[93] Sigurlaug Sveinbjornsdottir. "The clinical symptoms of Parkinson's disease". In: *Journal of Neurochemistry* 139.S1 (July 2016), pp. 318–324. DOI: 10.1111/jnc.13691.

[94] Rajiv Tandon et al. "Definition and description of schizophrenia in the DSM-5". In: *Schizophrenia Research* 150.1 (Oct. 2013), pp. 3–10. DOI: 10.1016/j.schres.2013.05.028.

[95] Ternimed. "Electrode arrangement according to the international 10/20 system". In: (). URL: https://www.ternimed.de/useful-information/Electrode-arrangement-according-to-the-international-10/20-system (visited on 08/23/2023).

[96] Alaa Tharwat. "Independent component analysis: An introduction". In: *Applied Computing and Informatics* 17.2 (2018), pp. 222–249. ISSN: 22108327. DOI: 10.1016/j.aci.2018.08.006.

[97] Tom N. Tombaugh and Nancy J. McIntyre. "The Mini-Mental State Examination: A Comprehensive Review". In: *Journal of the American Geriatrics Society* 40.9 (1992), pp. 922–935. DOI: https://doi.org/10.1111/j.1532-5415.1992.tb01992.x.

[98] Jose Antonio Urigüen and Begoña Garcia-Zapirain. "EEG artifact removal - State-of-the-art and guidelines". In: *Journal of Neural Engineering* 12.3 (2015). ISSN: 17412552. DOI: 10.1088/1741-2560/12/3/031001.

[99] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7.1 (Feb. 2006). DOI: 10.1186/1471-2105-7-91.

[100] Julie Walsh-Messinger et al. "Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning". In: *Psychiatry Research* 278.November 2018 (2019), pp. 27–34. ISSN: 18727123. DOI: 10.1016/j.psychres.2019.03.048. URL: https://doi.org/10.1016/j.psychres.2019.03.048.

[101] Lawrence M. Ward. "Synchronous neural oscillations and cognitive processes". In: *Trends in Cognitive Sciences* 7.12 (2003), pp. 553–559. ISSN: 13646613. DOI: 10.1016/j.tics.2003.10.012.

[102] WHO, ed. *Mental health systems in selected low- and middle-income countries: a WHO-AIMS cross-national analysis.* Geneva: WHO, 2009.

[103] WHO. *Neurological disorders : public health challenges.* Geneva: World Health Organization, 2006. ISBN: 9789241563369.

[104] WHO. *Schizophrenia. [Online].* Accessed 27 October 2023. 10 Jan, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/schizophrenia.

[105] World Health Organization WHO. "Implementation of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)." In: *Epidemiological bulletin* 18.1 (1997), pp. 1–4. ISSN: 02561859.

[106] World Health Organization(WHO). *Global status report on noncommunicable diseases 2014*. Genève, Switzerland: World Health Organization, Dec. 2014.

[107] Xiao Xing et al. "A High-Speed SSVEP-Based BCI Using Dry EEG Electrodes". In: *Scientific Reports* 8.1 (Oct. 2018). DOI: 10.1038/s41598-018-32283-8.

[108] Rajamanickam Yuvaraj, U. Rajendra Acharya, and Yuki Hagiwara. "A novel Parkinson's Disease Diagnosis Index using higher-order spectra features in EEG signals". In: *Neural Computing and Applications* 30.4 (2018), pp. 1225–1235. ISSN: 09410643. DOI: 10.1007/s00521-016-2756-z.

[109] Lei Zhang. "EEG Signals Classification Using Machine Learning for The Identification and Diagnosis of Schizophrenia". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2019. DOI: 10.1109/embc.2019.8857946.

[110] Chao Zhao et al. "Structural and functional brain abnormalities in schizophrenia: A cross-sectional study at different stages of the disease". In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 83 (Apr. 2018), pp. 27–32. DOI: 10.1016/j.pnpbp.2017.12.017.

[111] Massimo Zoni-Berisso et al. "Epidemiology of atrial fibrillation: European perspective". In: *Clinical Epidemiology* (June 2014), p. 213. DOI: 10.2147/clep.s47385.