Selection of Relevant Features to Support Automatic Detection of Epileptiform Events

MARITZA FERNANDA MERA GAONA

PhD Thesis in Telematics Engineering

Supervisors:
PhD. Diego Mauricio López
PhD. Rubiel Vargas Cañas
PhD. Maria Eugenia Miño

Universidad del Cauca
Faculty of Electronics and Telecommunications Engineering
Department of Telematics
eHealth
Popayán, June 2021

MARITZA FERNANDA MERA GAONA

Selection of Relevant Features to Support Automatic Detection of Epileptiform Events

Dissertation submitted to the Faculty of Electronics and Telecommunications
Engineering of the Universidad del Cauca, Colombia
for granting the academic degree of

Doctora en:
Ingeniería Telemática

Supervisors
PhD. Diego Mauricio López
PhD. Rubiel Vargas Cañas
PhD. Maria Eugenia Miño

Popayán
2021

To the people who suffer from
any disease and to the scientists
who dedicate their lives to find
treatments, cures, and vaccines.

Acknowledgements

Foremost, I would like to thank the almighty God, who has let me to experience his grace in every important moment of my life. His abundant blessings gave me the knowledge and wisdom to be able to complete this thesis.

> ***“I can do all things, through Christ who strengthens me.”***
> ***-Philippians 4:13***

Second, this research is presented as my work, nevertheless, I could not have completed it without the support of those who have believed in me throughout my life. Although mentioning names may be unfair because it is impossible to mention all those who in one way or another contributed to my education and life. So, I don’t want to miss the opportunity to express my gratitude to some of them but not before apologizing to those that I don’t mention.

To my supervisor, Diego M. López, and the professors Rubiel Vargas Cañas, and Maria Eugenia Miño, who gave me the opportunity to do this project on the topic Automatic Reading of EEG to support Epilepsy Diagnosis. Without their assistance, disposition, and patience, this thesis would never have been accomplished.

To PhD Ursula Neumann, and Professor Peter Hussar, who were involved in this research as advisors. Their support during my research stays in Germany was fundamental in the successful completion of my doctoral studies.

To Minciencias and the Colombian government for the studentship that allowed me to conduct this thesis under the call 647-2015.

To the University of Cauca for giving the opportunity to grow as an engineer, master, and doctor. A special acknowledgment to my professors and friends Sandra Roa, and Carolina Gonzalez, who contributed their knowledge and advice to my professional growth.

To my colleagues and friends who supported me during this challenge and encouraged me to make this doctoral dream possible.

To my high school teachers in Liceo Bello Horizonte, especially Maria del Carmen Garzón, Luz Angela Garzón and Sofia Garzón. Thank you for believing in me and giving me the opportunity to receive a high-quality education that led me to university.

To my family for their support throughout my life and education: my mother, Maria Gaona; my father, Eduardo Mera; my brothers, Andrea, and Jason. Thank you for encouraging me to study.

To Microsoft, the company which believed in me when I was just an engineering student. Thanks for encouraging me to achieve more and helping me to make my dreams come true.

To Taylor A. Swift, Katheryn E. Hudson, William H. Gates III, Maria Sharapova and Roger Federer; Although they are people that I may never meet, they inspired me throughout the duration of my studies with their genius, talent, perseverance, and courage.

Lastly, but perhaps most importantly, to the memory of my grandmother Nubia, who always trusted me and although she couldn´t see me as a doctor, she was my biggest motivation to finish this thesis.

Structural abstract

Background: Identifying relevant data to support the automatic analysis of electroencephalograms (EEG) has become a challenge. In the literature, there are many proposals built to support the diagnosis of neurological pathologies. However, the current challenge is to improve the reliability of the tools to classify or detect the abnormalities. Thus, the Ensemble Feature Selection approach allows the integration of the advantages of several Feature Selection algorithms to improve the identification of features with high power of differentiation in the classification of normal and abnormal EEG signals. Feature Selection has attracted the attention of many researchers in the last years due to the increasing sizes of datasets. In many cases, the datasets contain hundreds or thousands of columns. However, not all columns contain relevant information, which leads to the weak performance of classifiers. Besides, several Feature Selection Algorithms have been proposed in the literature to analyze datasets and determine their subsets of relevant features and remove irrelevant or redundant features from the classification process. Those Feature Selection algorithms are typically classified according to their design, which is related to how they find the subset of relevant features and the complexity to calculate them. There are three main types of feature selection algorithms: filters, wrappers, and embedded. The implementation of wrappers and embedded algorithms are complex because its implementation requires including at least a classification algorithm to calculate the relevance index of each feature; the index relevance could change when instances are added or removed from the dataset. Likewise, the filter-based feature selection algorithms can be computationally simpler than the other approaches (envelopes and embedded).

Objectives: the main objective of this thesis is to propose a mechanism for selecting relevant features for the classification of electroencephalograms segments to support the automatic detection of epileptiform events. For this, a conceptual framework was designed following a quantitative method in order to represent a structure that provides an understanding of how to improve the performance of machine learning algorithms by using the consensus of several feature selection algorithms.

Methods: to achieve the main objective of this thesis, a conceptual framework was designed to understand the key concepts and relationships in the aggregation of a set of feature selection algorithms. Based on the conceptual framework, a Development Framework was implemented to validate the theoretical proposal and select a set of relevant features for the classification of electroencephalogram segments. The selected features were used to train classification algorithms and build a classification model, which was used to detect automatically epileptiform events in electroencephalograms. For training the classification model, a dataset was built by applying feature extractors from raw records of electroencephalograms and creating a representation of electroencephalogram based on extracted data.

Results: we obtained three main results: (i) A dataset of epileptiform events described by applying feature extractors on raw records of electroencephalograms (ii) A ensemble feature selection framework to identify relevant features for classifying epileptiform events in EEG signals. To evaluate the quality of the set of relevant features, we measured the stability of the set of relevant features selected by the proposed framework. The results showed a perfect stability in 100% of the cases evaluated. (iii) A classifier trained by using the relevant features selected from the dataset of epileptiform events. For this, we evaluated the performance of the classifiers: Decision Tree, Logistic Regression, Random Forest, and SVM. These algorithms were assessed by using all the features and the set of relevant features for their training. The results evidenced that the models improved their performance when trained with the EFS approach's features. In addition, the classifier of epileptiform events built using the features selected by the EFS method achieved an accuracy, sensitivity, and specificity of 97.64%, 97.32%, 96.27%, respectively. Hence, the stability of the EFS method evidenced a reliable subset of relevant features. Moreover, the accuracy, sensitivity, and specificity of the EEG detector are equal to or greater than the values found in the literature

Conclusions: The design of a conceptual framework allowed to guide the development of an implementation framework to support the assembly of feature selection algorithms, thus overcoming the limitations of individual feature selection algorithms. The relevant features selected by using the framework allowed to train a classifying model of segments extracted from electroencephalograms of children. Finally, to describe the electroencephalograms, a set of feature extractors were applied on the signals extracted from the records.

Key words: Ensemble Feature Selection, Framework, EEG, epileptiform events.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1. Problem

The incidence rate of neurological diseases worldwide reports that there are currently about 450 million people suffering from multiple type of mental disorder [1]. One difficulty in treating patients with this type of disorder is the lack of specialized personnel and the high costs involved in the diagnosis [2]. In low- or middle-income countries such as Colombia, the care provided to patients with this sort of disorder is not of high quality, because of the lack of resources that the national government invests in attention to Mental Health problems.

Additionally, the specialists in charge of diagnosing and treating these diseases are usually found in specialized medical centers, which are difficult to access for the populations from rural areas. This situation is one reason early detection of neurological disorders is difficult in low- and middle-income countries. According to statistics, in Colombia, the total number of neurologists in 2011 was 231, which represented a distribution of one neurologist for every 199,327 inhabitants [3]. This number is too low if we compare it with the availability of neurologists in developed countries such as the United States and Spain. In those countries, the ratio is one neurologist for every 22,880 inhabitants and one neurologist for every 23,500 inhabitants, respectively [4]. On the other hand, in small and medium-sized cities the problem increases because the number of specialists and clinical centers decreases. For example, in Popayán there are only four neurology specialist centers to serve a population of 267,976 inhabitants [5]. Considering that, in the rest of the department of Cauca, there are no centers with these characteristics, the 4 centers of the city of Popayán must provide their services to the entire population in the department of Cauca, which is about 1,051.007 inhabitants [5]. Therefore, and considering the importance of diagnosing patients at an early age to get better results in the applied treatments, this project will focus on supporting the diagnosis of epileptiform events in EEGs of underage patients.

Moreover, the limited availability of professionals specialized in neurology generates long times to obtain the reading of Electroencephalograms (EEG) and an excessive cost of the services. This is because the visual inspection of an EEG signal can be too complex and turn into an extended task due to the duration of the EEG and the presence of abnormal patterns in the EEG. Hence, tools that automatically detect epileptic episodes in an EEG signal would help to reduce the time used in the visual inspection of an EEG signal, especially when long-lasting EEGs (24, 48, or 72 hours) are analyzed. The possibility of reducing the inspection time of a specialist could represent the opportunity to attend to more patients, something especially useful in countries similar to Colombia, due to the low availability of specialists.

In recent years, research on the development of systems for capturing and analyzing biomedical signals has increased to find new mechanisms for the clinical diagnosis of specific pathologies. Through EEG signal processing, neuronal activity in the brain can be monitored and information that describes useful information for the detection of neurological pathologies can be extracted. However, the challenge of characterizing EEG signals is associated with the complexity of extracting the relevant information capable of describing the presence or absence of the disease. This difficulty results from the complexity of knowing a priori, which information from the signal should be considered relevant for the detection of the patterns that characterize each pathology [6].

To process and analyze an EEG signal, we must extract a set of features from the signal and store them in a vector. The vector should describe as much information as possible to gather a complete representation of the signal. In this way, when the feature extraction process is applied on multichannel signals such as an EEG, the descriptors used to extract information could generate hundreds of features because they are applied on each channel of the signal. In a feature extraction procedure, the features can represent noise, relevant information, or redundant information [7]. Considering the above, Feature Selection (FS) is recommended task when we must handle high-dimensional vectors. FS allows to identify the features with useful information for the detection of pathologies, reduce the complexity and the computational cost generated by the calculation of redundant features, noise, and handling of vectors with enormous size (n) [8].

In recent years, the diagnosis of diseases such as epilepsy through digital analysis of EEG signals has been one of the key areas of research in neuroscience. Through different feature extraction mechanisms, it has been possible to obtain information to classify a signal as normal or abnormal [6]. Likewise, other studies based on the analysis of EEG signals have been carried out to analyze brain activity [7][8] and support clinical diagnostic processes. Some signal classification mechanisms have used neural networks, decision trees, rules based on domain knowledge, and clustering mechanisms to classify a new signal [9][10]. Although in the literature some works have reported feature selection methods to identify features with greater power of differentiation in the classification or detection of epileptic patterns, the vast majority of the reviewed works in the state-of-the-art were focused on the identification of specific patterns by using a set of features without considering the real relevance of each feature [9].

In the literature, there are many mechanisms used to characterize EEG signals and detect or classify events associated with epilepsy, however, the great research challenge in this area is focused on improving the performance of the classification in terms of precision, accuracy, and recall, to provide reliable tools that support and help specialists during the Epilepsy diagnosis process. Considering the above, one of the main strategies to improve the classification models in machine learning or data mining is to train the models with relevant features, that is, those features that do not represent noise for the learning model and, on the contrary, have a high power of differentiation between classes. Commonly, the identification of relevant features becomes a very useful process in scenarios in which there is a dataset with high dimensionality [10]. Likewise, in scenarios in which the training datasets have more features than instances, the feature selection becomes a mandatory task to reduce the dimensionality of the datasets and support a robust classification [11]. This scenario coincides with the classification of abnormalities in EEG if we consider the considerable number of feature extractors reported in the literature and the low availability of datasets with instances that describe abnormalities or epileptiform events.

Also, there are three types of feature selection algorithms: filters, wrappers, and embedded. The implementation of wrappers and embedded algorithms is complex because its implementation requires including at least a classification algorithm to calculate the relevance index of each feature; the index relevance could change when instances are added or removed from the dataset [12]. Likewise, the filter-based feature selection algorithms can be computationally simpler than the other approaches (envelopes and embedded). However, these types of algorithms commonly analyze the relevance of the features individually, in such a way that two features are correlated could help to classify better, although these same features separately could have been considered irrelevant.

Recently, several studies have focused on improving the performance of feature selection algorithms, in the authors proposed to identify correlations between features and classes to [12] improve the effectiveness and maintain a low computational cost in the feature choice process. Another proposal incorporated techniques such as Bootstrap to select features using samples from the original dataset and integrating the subsets of features generated [13][14][15]. Although these works presented promising results, the methods can be sensitive to the balancing of the datasets and the treatment of continuous data. In this sense, authors have proposed assembly feature selection algorithms to improve the identification of relevant features through the consensus of FS algorithms with different approaches [15].

Considering the above, we consider that the ensemble of feature selection methods can improve the effectiveness of the choice of relevant features and enhance the classification of epileptiform events in EEG signals. This approach is based on the premise of multi-classifiers: “several classifiers classify better than one”, which would be applied to the feature selection, where we intended to prove that “several feature selectors select better than one”.

The complete description of the research proposal can be reviewed in Appendix A.

1.2. Objectives

General

To propose a mechanism for the choice of relevant features in the automatic detection of epileptiform events in EEG signals.

Specifics

- To build a dataset of EEG signals and clinical data of patients with epilepsy.
- To construct an ensemble feature selection method to determine relevant features in the classification of epileptiform events.
- To evaluate the performance of the feature selection mechanism in the classification of epileptiform events in pediatric EEG signals.

1.3. Chapters

This section describes the structure of the document that details the definition of the Framework to support Ensemble Feature Selection in the feature selection of EEG.

Chapter 1: Introduction

This chapter describes the problem, the objectives of the research, and the structure of the rest of the document.

Chapter 2: Context and State-of-the-Art

The chapter explains the key concepts and context of the research and the conduction of a systematic mapping to describe the state-of-the-art.

Chapter 3: Construction of the Dataset

This chapter describes the process of construction of a dataset of EEG signals following some stages and tasks of CRISP-DM methodology.

Chapter 4: Design of the Framework

This chapter presents the design of the Conceptual Framework of Ensemble Feature Selection following a quantitative approach.

Chapter 5: Evaluation

This chapter describes the evaluation carried out to validate the selection of relevant features for the automatic classification of epileptiform events.

Chapter 6: Conclusions, Recommendations, and Future Work

This chapter describes the conclusions obtained from the research carried out and proposes some recommendations and improvements to consider for future work.

Chapter 2
Methods and State of the Art

This chapter is divided into four sections. The first one introduces the fundamental concepts and definitions related to the context and theory concerning brain activity and electroencephalography. The second section presents a general description of the classification process of electroencephalograms and some techniques used in its stages. The third section presents the design and results of a systematic mapping conducted to identify relevant advances in ensemble feature selection for the classification of physiological signals. Finally, the fourth section describes the main conclusions and research gaps identified in the state of the art.

2.1. The Central Nervous System

The Central Nervous System (CNS) comprises nerve cells, each nerve cell comprises axons, dendrites, and cell bodies. These cells respond to stimuli and can transmit information over long distances. Activities in the CNS are related to synaptic currents transferred between the junctions of axons and dendrites or dendrites and dendrites, which generate potential changes in the membranes of neurons [16].

The information transmitted by a nerve is known as Action Potential (AP), an AP is generated from the exchange of ions through the membrane of a neuron. PAs are temporary potential changes that are transmitted along the axon. When the potential becomes positive, a peak occurs, when the potential becomes negative, the membrane potential returns to normal, as can be seen in Figure 1.

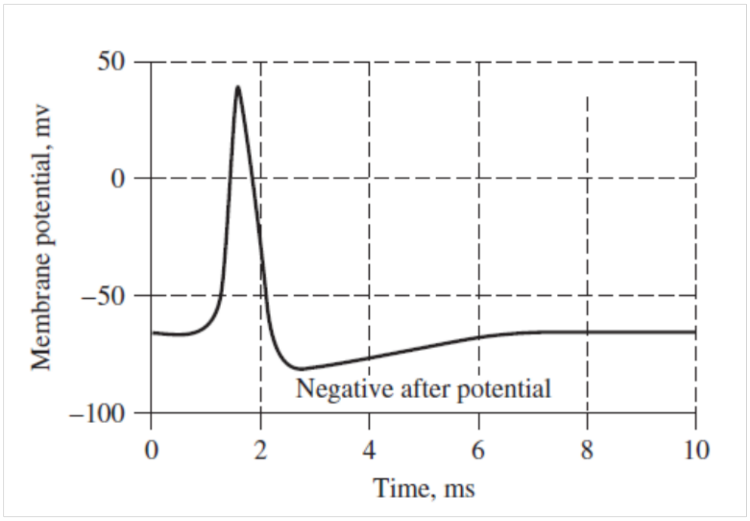


Figure 1. Example of section potential, Source [17].

2.1.1. Electroencephalogram

An Electroencephalogram (EEG) is a tool used by neurologists to measure the electrical activity of the human brain generated by the currents that flow during synaptic excitations. Thereby, an EEG represents the electrical brain activity from the electric field on the scalp generated when the neurons are activated, and synaptic currents are activated in the dendrites [17].

Visual inspection is how specialists analyze the electrical activity of the brain described in the EEGs. The information in the EEGs allows the analysis of the brain functions over time and detects neurological disorders. The EEGs are represented in recordings of multichannel signals taken through the digitization of the data captured from electrodes in the scalp or intracranially [18].

Electroencephalograms are used mainly to support the diagnosis of neurological diseases that affect brain activity, such as:

- Epilepsy
- Brain tumors
- Head injuries
- Sleep disorders
- Dementia
- Monitoring of anesthesia during surgeries

In the recording of an EEG, the electrodes capture small electrical discharges that are generated from the activity of cells in the brain. These charges are amplified and digitized to create a graphic representation described by waves. To take an EEG, a different number of electrodes can be used. Generally, each pair of electrodes generates the information of a channel (See Figure 2). Depending on the number of electrodes used, more channels will be obtained in the EEG and more information.

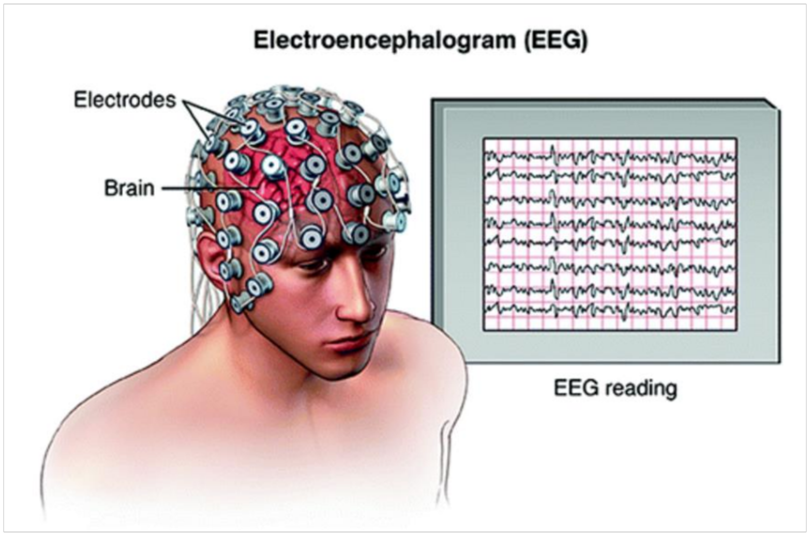


Figure 2. EEG Acquisition Scheme, Source [19]

2.1.2. 10-20 System

In the context of an EEG exam, the 10-20 system or international 10-20 system is a well-known method to describe the location of scalp electrode [20]. The location of the electrodes is related to the lobes of the brain, thereby, the electrical activity captured by using the electrodes describes the brain activity associated with the lobes.

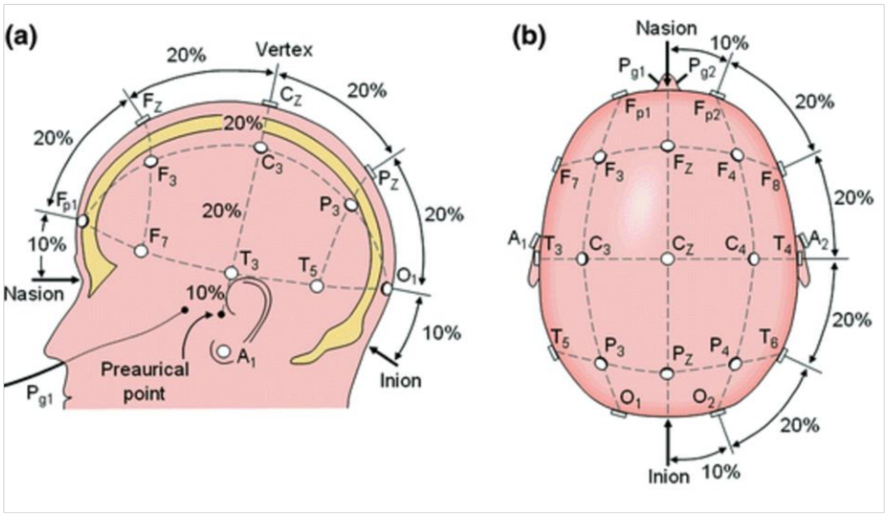


Figure 3. Locations of the electrodes according to with 10-20 system. Source [19]

Figure 3 describes the standard mounting used to position each electrode. The location of each electrode is determined by measuring the distances between the adjacent electrodes, which could be 10% or 20% of the distance calculated from the front part of the skull (Nasion) to the back of the skull (Inion) or the distance measured between the right and left preauricular points. The Nasion is the intersection between the nose and the forehead, and the Inion is the point defined by the base of the skull, which the ligament nuchae and trapezius muscle join. Also, the preauricular points are skull landmarks that describe the indentations just above the tragus which covers the external ear openings.

According to the 10-20 system, each site has a letter to identify the lobe and a number to describe the hemisphere location, where is each electrode [21]. The letters used are:

- F: Frontal
- T: Temporal
- C: Central (The central lobe does not exist; however, the letter ‘C’ is used to complement the scheme)
- P: Parietal
- O: Occipital

Additionally, when the letter "z" is used, it shows that the electrode is in the midline. To differentiate the electrodes in the right hemisphere from the left hemisphere, even numbers or odd numbers are used, respectively. Once the electrodes are on the patient's scalp, the voltage differences between a pair of electrodes are measured and shown to the specialist as information from an EEG channel.

Considering the above, the study of EEGs becomes a very useful tool for the diagnosis of different neurological disorders and other abnormalities that alter brain electrical activity.

2.1.3. Brain Rhythms

Brain rhythms define patterns of massed neuronal activity in the Central Nervous System and are related to specific behaviors, arousal levels, and sleep states. Usually, brain rhythms are measured by an electroencephalogram (EEG) and classified according to the neuronal oscillations. The brain rhythms range from very slow oscillations to very fast oscillations. Figure 4 presents the five main frequency bands of the brain [22], which are described below:

- Alpha

These types of waves are generated in the occipital region of the brain. They are in the 8-13 Hz range and appear as a round or sinusoidal signal. The state of the brain to which these waves are associated is a state of relaxation (with no concentration or attention). Some patients can produce these types of waves with their eyes closed.

- Theta

Theta waves appear in a range of 4-7.5 Hz. Their presence normally shows that consciousness is about to go into a state of drowsiness. Some activities associated with this type of wave are a creative inspiration and deep meditation.

- Beta

This type of wave appears in a range of 14-26Hz, they are associated with waves generated in awake subjects who perform analysis tasks, active attention, or solving specific problems. High beta waves could be associated with a state of panic in the individual.

- Delta

Delta waves are described by high amplitude brain waves and frequencies of oscillation between 0.5 and 4 Hz. This kind of wave is mainly associated with the waves produced during deep stage 3 of NREM (non-rapid eye movement) sleep.

- Gamma

Gamma waves present a neuronal oscillation above 30Hz. Normally, the amplitudes of this type of wave are very low and can be increased through meditation. The gamma activity is correlated with cognitive processes like attention and working memory and the detection of this type of waves can identify some neurological disorders.

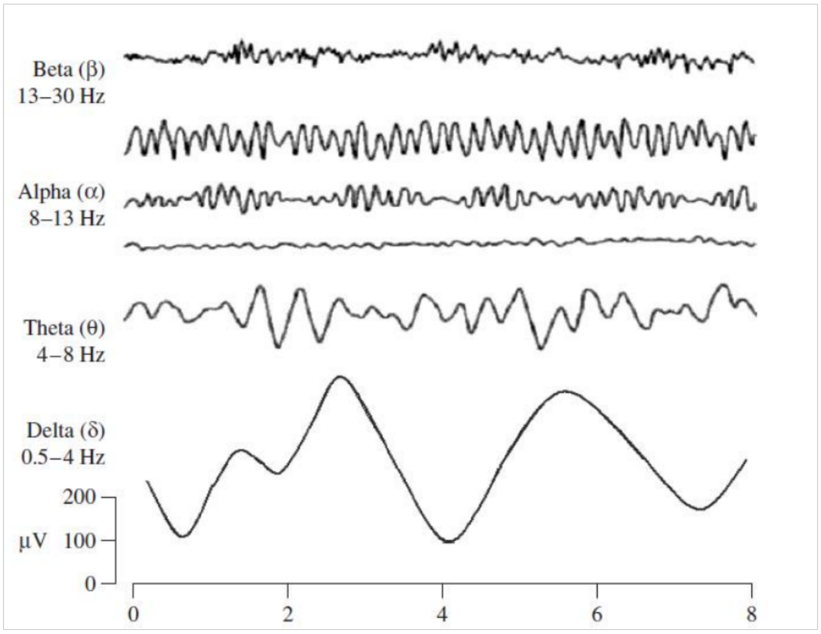


Figure 4. Brain rhythms. Source [17].

Many neurological disorders are detected from visual inspection of EEGs. For this purpose, specialists analyze the behavior patterns of different rhythms of the human brain. For example, in adult patients, it is known as the amplitude and frequency of an EEG should be, likewise, how they change from a state of full consciousness to a state of sleep. Another crucial factor that neurologists analyze is an age since this can also influence the behavior of amplitude and frequency.

2.1.4. Epileptic Seizures

It is a neurological disorder generated by electrical discharges in one or more areas of the brain. The diagnosis is given through the visual inspection of EEGs performed on the patients together with clinical data that describe a history associated with possible brain injuries.

There are different symptoms that patients with this disease present, the most common are epileptic seizures. These generate abrupt changes in frequency and amplitude in the EEG. The desynchronization of electrical activity is reflected in the presence of disordered waves that after a while, rearrange themselves. Some changes that can be observed in the EEG waves during an epileptic seizure are sharp spikes or the appearance of slow waves [23].

Sometimes, EEG peaks appear because of artifacts produced by muscle movement, opening, and closing of the eyes, neck movements, among others. These artifacts can be easily eliminated thanks to the fact that EEGs commonly include information of two channels that allow knowing if the patients move or close their eyes.

2.2. EEG Classification

In recent years, research on the development of systems for capturing and analyzing biomedical signals has increased to find new mechanisms for the clinical diagnosis of specific pathologies. Through the processing of EEG signals, the neuronal activity of the brain of patients can be monitored and abnormalities can be detected that allow diagnosing neurological diseases. In this way, the classification of EEGs is modeled as a machine learning problem, and data is extracted from the signals emitted by the brain that describe useful information from them. However, the challenge of characterizing EEG signals is associated with the complexity of extracting the relevant information capable of describing the presence or absence of disease. This difficulty is owing to the complexity of knowing a priori what information is contained in the signal that must be relevant for the detection of the patterns that characterize each pathology.

2.2.1. Segmentation

Because of the extensive amount of information included in EEGs, the process of automatically analyzing this kind of record requires dealing with a large volume of data and the traditional mechanisms of pattern recognition and data mining cannot be applied. Hence, different proposals implemented for the automatic detection of abnormalities in signals have proposed analyzing the EEG channel by channel and performing segmentation in each channel, to analyze the changes suffered by the signal over time [24]. However, this solution faces the difficulty of defining the size of the segments into which each channel should be divided and also determining if the segments are independent or if they should overlap in some part (Figure 5).

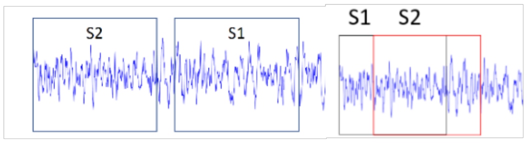


Figure 5. Segmentation: a. independent b. overlapping

This mechanism is known as windowing. The solution proposes to review a signal segment by segment, that is, the segment will be classified as normal or abnormal. When all segments of the signal are checked, the list of all abnormal segments is gained.

Some solutions that have tested the implementation of window analysis have included mechanisms that allow the random definition of the size of each segment that is built.

2.2.2. Feature Extraction

The feature extraction process allows characterizing an EEG signal through the calculation of mathematical operators. The result is represented by a vector that contains information about the signal with which it can be classified [25]. This stage is one of the most important tasks within pattern recognition because performing the classification depends on the extracted features. If the features are not well chosen, the precision and accuracy of the classification will be compromised.

A basic scheme is described in [26] to explain the stages of automatic reading of EEG signals (Figure 6). The scheme presents the feature extraction process as a previous task of the classification and it is influenced by the segmentation technique and the mechanism used to capture the EEG signal. The key methods used to analyze the EEG signals are based on an analysis in the frequency and time-frequency domain.

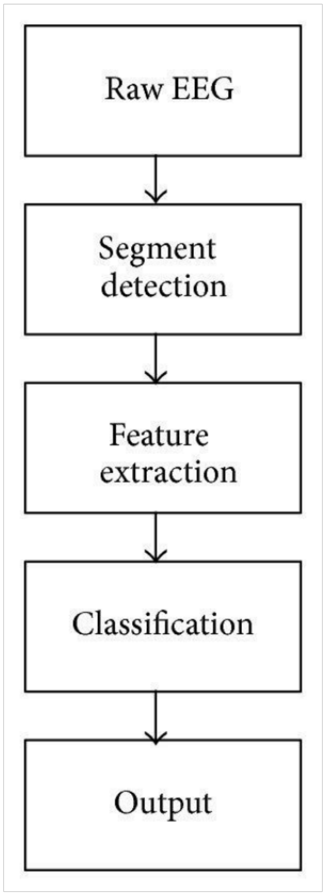


Figure 6. Stages of EEG signal reading. Source [26].

2.2.3. Relevance Analysis

The relevance analysis stage is also known in the literature as feature selection and dimensionality reduction. The three terms are used to describe the development of reducing the original set of data generated by the feature extraction process into a minimum set that represents the greatest amount of useful information in the classification process [27]. During the relevance analysis process, 3 types of data are analyzed: relevant, not relevant (noise), and redundant. Through elimination and transformation methods, the original set of features is transformed into a smaller set, which only contains the features identified as relevant.

- Elimination

  The elimination methods allow creating subsets of features from the original set, the smallest subset with the best performance in the classification (or any criterion) is selected to represent the signal, therefore, the other features are discarded and are not considered in the final set.

- Transformation

The transformation methods create combinations of the original features. One of the most common methods is Principal Component Analysis (PCA). This method creates a linear combination of the input features and the additional features are projections of the original features in a new feature space. The additional features are called components, which are reduced by selecting those with larger variance.

2.2.4. Classification

Classification is one of the most important stages in fields such as data mining, machine learning, pattern recognition, among others. Through a set of variables that describe instances of a dataset, it is possible to create a model capable of classifying an unclassified instance using a classification algorithm.

2.2.5. Evaluation

Different metrics have been reported in the literature to evaluate solutions for automatic EEG reading. Some of these metrics are widely used in fields such as pattern recognition and are consequently accepted in the medical field. To evaluate the solution from different important perspectives to consider a reliable classification, in [19] the following metrics are proposed:

- Accuracy
- Sensitivity
- Specificity
- FAR
- ROC
- K-fold Cross-validation

2.3. State of the Art

This section describes the execution of a systematic mapping focused on identifying studies related to the assembly of feature selection algorithms. The systematic mapping was conducted following some guidelines described by Petersen in [28].

2.3.1. Research Questions

The study was guided by a set of research questions focused on answering how to improve the effectiveness of the selection of relevant features in the classification processes of epileptiform events in EEG signals. The questions are described below.

RQ1: How to improve feature selection in EEG signal datasets through ensemble feature selection?

RQ2: How to assemble feature selection algorithms?

RQ3: How to design and implement a framework to select relevant features by an ensemble feature selection?

2.3.2. Search

The search protocol was led by a search string, which was built with a set of keywords related to the research questions. The search string was used to perform queries on the databases and the results were analyzed to determine a set of studies with important contributions to answering the research questions. Besides, the relevant works referenced by the selected studies in the search were included in the last set of studies, with important contributions (backward snowball sampling). Thus, the search string is described below:

("Framework" and "ensemble") and ("dimensionality reduction" or "feature selection") and ("EEG" and "automatic") and ("detector" or "reading" or "recognition" or “analysis”)

The databases selected to perform the systematic mapping were IEEE, Science Direct, and PubMed. These databases were selected considering their relevance in the fields of science computer, signal processing, and health information systems. The following table describes the keywords used and their synonyms as alternative words:

Table 1. Keyword synonyms.

Word	Synonym
Feature selection	Feature relevance Dimensionality reduction
Detection	Detector Reading Recognition Analysis

2.3.3. Studies Selection

The searches performed on each selected data based retrieved many studies, which were conducted to define a protocol to include or exclude articles of the final analysis. The exploration of the papers was led following review criteria based on the quality of their contributions and available information about their results and conclusions.

The following inclusion and exclusion criteria were applied in distinct steps:

- Inclusion:
  - Studies that describe the use of ensemble feature selection algorithms.
  - Studies that describe the construction of frameworks to support ensemble learning.
  - Studies that describe Ensemble Feature Selection algorithms and their evaluation results.
  - Studies related to automatic reading of electroencephalograms.
- Exclusion:
  - Studies that do not describe techniques and algorithms used.
  - Studies in the abstract do not describe information related to the research questions.
  - Studies that do not present the results of the evaluation.

The inclusion and exclusion criteria were applied first, only considering titles, abstracts, and keywords. The studies that appeared to be important but did not meet all the inclusion criteria were evaluated in a second stage to review the results and conclusions sections. In the ultimate step, we performed a full-text reading to determine the important contributions of each paper.

2.3.4. Analysis and Classification

2.3.4.1. Research Focus Areas

Considering that an electroencephalogram (EEG) represents the electrical activity of the human brain [29] and is used to diagnose different pathologies through the detection of abnormalities or patterns in the EEG signals, the exploration of the literature in this study considered not only the review of feature selection and the ensemble techniques but also the type of diseases or activities that can be identified through the identification of patterns in the electrical activity of the brain.

From the Epilepsy diagnostic process, we know EEGs are widely used as one of the most reliable sources in the detection of seizures. A seizure occurs when neurons generate abnormal electrical shocks in brain cells. These alterations or abnormalities are identified by doctors during the reading of an EEG. An EEG not only allows the identification of voltage changes in the brain but also offers information that allows characterizing the type of epilepsy, which is useful to support the treatment of the patient [30]. Similarly, EEGs allow the detection of other types of abnormalities associated with other pathologies or brain activity performed by the person.

Considering the above, the studies retrieved in the bibliographic database searches were classified in the following research areas:

- Feature Selection
- Ensemble Feature Selection
- Disease/Activity

2.3.4.2. Research Type Classification

Wieringa et al. [31] propose a set of research types to classify works and research in articles according to the engineering development cycle. This classification is described below:

Table 2. Research types

Category	Description
Solution proposal	The study proposes a solution or technique without applying a robust evaluation. Although, the potential benefits and the applicability of the solution are shown.
Validation research	The work describes a novel solution that has not been implemented in a proper environment. The techniques have been evaluated in laboratory environments.
Evaluation research	It refers to a solution to a real problem tested in a proper environment. The study also describes the consequences of the implementation (benefits and drawbacks)
Conceptual proposal	This category describes studies with a novel way of looking at existing things. The contributions of these studies are taxonomies or conceptual frameworks.

Experience paper	These works contain a set of lessons learned by the authors from their experiences
Opinion paper	These papers contain a set of lessons learned by the author from his experience and opinion about a certain technique (good or bad).

2.3.4.3. Contribution Type Classification

Considering the contributions of the papers evaluated in this systematic research, below, we describe the contributions that we consider classifying the results acquired in each study.

Table 3. Contribution types.

Category	**Description**
Tool	A system or prototype
Model	A theoretical representation of a proposal
Metric	Additional measures to estimate effectiveness, for example.
Enhancement	Modifications on techniques or solutions to improve behavior.
Technique	Describe a new technique

2.3.5. Analysis

In this section, the major results of the systematic mapping carried out are graphically presented. Subsequently, we show an analysis according to the classification scheme used to categorize the retrieved studies.

2.3.5.1. Results

The search guided by the string search allowed for the identification of 33 papers with relevant results regarding the ensemble of algorithms for selecting features in the automatic reading of physiological signals. Besides, two papers referenced in some of the 33 selected papers were considered finally because they described relevant contributions. Table 4 shows the selected papers from each database.

Table 4. Results of the searches.

Source	**Revised papers**
IEEE	6
PubMed	7
Science Direct	20
References	2

Figure 7 shows the distribution of the papers studied by year and source of publication, showing that the vast majority of publications come from research journals and only 17% are from conferences.

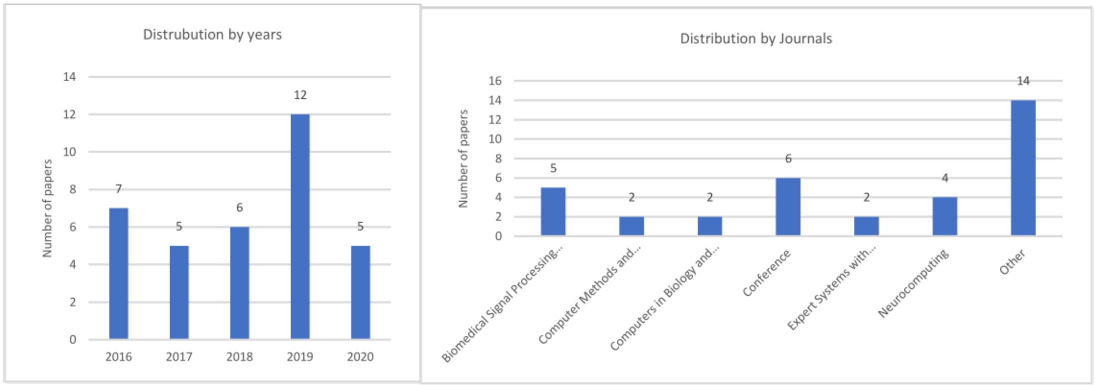


Figure 7. Distributions of papers by year and journal

Figure 8 shows the distribution of the reviewed papers according to the type of research and type of contribution. The results show that most papers present a solution proposal, highlighting a gap in the state of the art related to the lack of validation of research on the subject. On other hand, the most frequent types of contributions are techniques and tools, showing that in the state of the art, important advances have been developed on feature selection algorithms and solutions that assembly them. However, they have not been built as frameworks that support feature engineering processes.

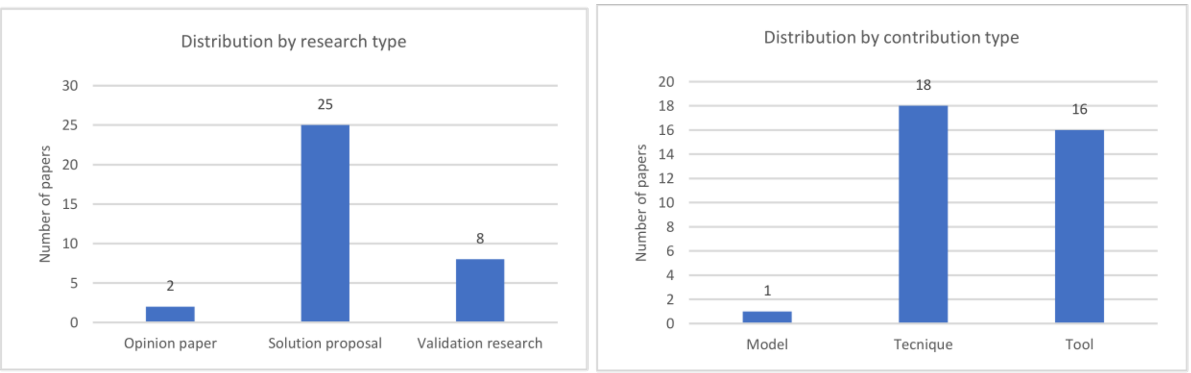


Figure 8. Distribution by research and contribution type.

Table 5 describes the distribution of the recovered articles according to the type of disease or field of application. The review shows the use of different feature selection techniques to support automatic analysis of diverse types of physiological signals. The review carried out shows from general methods to select features on clinical databases [32] to implementations carried out to support the diagnosis of diseases such as Alzheimer's [33], Multiple sclerosis [34], sleep disorders [35] and Epilepsy [36]. Additionally, the list of reviewed papers presents solutions designed for the detection of emotions by analyzing the electrical activity of the brain [37] or the recognition of activities using analysis of physiological signals [38] and external devices [39].

Table 5. Distribution of papers by field of application.

Field	Number of papers
EEG, EOG, and EMG signal	4
Alzheimer	2
Automatic sleep screening	2
BCI	1
Diagnosis system for breast MR	1
DNA micro arrays	2
Emotion Recognition	2
Epilepsy	3
General	8
Human Activity Recognition	3
Multiple sclerosis	1
PCG	1
Other	5

Table 6 presents the distribution of the papers by aggregation methods used to implement the ensemble of the algorithms. Although 42% of the papers do not describe a method based on ensemble learning, in the other papers we can observe that the most frequent method is voting [40]. Additionally, we found aggregations methods based on weighted [41], fusion score [42], averaging scores [43] and cascade [44].

The evidence shows that the methods implemented in most of the reviewed papers considered a parallel scheme in the implementation of the ensemble of the algorithms. This means that several models build their output (individual set of relevant features), which are integrated by an aggregation function to generate a result with the final relevant features. On the other hand, some works considered a cascade mechanism where the result of a feature selection algorithm becomes the input of another FS algorithm. In this last scenario, the first algorithm could skew the result of the second algorithm and as the eventual result, we could get a final set of relevant features without the true relevant features.

Table 6. Distribution by an aggregation method.

Technique	Number of papers
None	15
Averaging scores	2
Cascade	1
Fusion score	2
Heuristic measure	1
Several	1
Voting	11
Weighted	2

2.3.5.2. Related Works

Some of the most relevant studies identified in the systematic mapping are described below:

Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning [42]

In this study, the authors proposed a hybrid feature selection method to improve the identification of relevant features in datasets of physiological signals. The signals are acquired by multi-sensors and are processed to extract features that describe them. Then, the features were evaluated by three FS algorithms based on filters: ReliefF, MRMR Pearson, and MRMR Spearman. The goal of the hybrid method is to support an efficient step for selecting a better feature subset and improve the performance of the task recognition. To aggregate the outcomes of the three FS algorithms, the method used a fusion score to determine the importance of each feature. Finally, the framework built to detect cross-subject driver status used data recorded from real and simulated driving environments and the results showed that the proposed algorithm could help to achieve high recognition accuracy. Although, the idea of ensemble learning is to combine the decisions of different models to compensate the weakness of some algorithms with the strengths of the others, the hybrid method proposed in this study only considered FS algorithms based on filters, which could omit some important features.

Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data [45]

This research describes a hybrid framework designed to support gene selection and improve the classification of DNA microarray data. In this study, five statistical methods based on filters were used to calculate scores for each gene. The scores were combined by a fusion of multiple filters, which pre-selected a set with the most informative genes. Then the selected genes are evaluated by an embedded approach to get the final relevant gene subset. Finally, an SVM (Support Vector Machine) is built to classify the genes using a small subset of genes. Although the authors included an embedded FS algorithm in a second stage to improve the gene selection, they did not consider that the five algorithms used in the first stage could remove informative genes.

Improving classification of epileptic and non-epileptic EEG events by feature selection [46]

In this work, the authors proposed a classification method of generalized epileptic and non-epileptic events by analyzing EEG data. A set of feature extractors were applied to the EEG data from 11 subjects to get patterns in the time domain and the frequency domain. Due to the feature extraction process produced a set of 1155 features, feature extraction was applied using the ReliefF algorithm. To implement feature selection, the authors used a leave-one-out strategy on the data to generate different training subsets and the ReliefF algorithm was applied in each training subset. All rankings calculated by the ReliefF algorithm were combined using two strategies. The first one calculated a total ranking according to the frequency of the specific rank of a feature in each experiment. The second strategy calculated a total rank by adding all weights given by Relief F in each experiment. According to the two total rankings, the authors extracted n-best features to train different classification algorithms and review the accuracy, sensitivity and specificity. Although the leave-one-strategy helps to stabilize the final subset of relevant features, the ReliefF is an FS algorithm based on filters and this kind of FS algorithm is not useful when the features are not independent. Considering that in this study the authors try to classify generalized seizures, the features could be related.

A novel joint HCPMMP method for automatically classifying Alzheimer's and different stage MCI patients [44]

In this work, the authors proposed a method to classify Alzheimer’s and different stage MCI (Mild Cognitive Impairment) patients. For this, they generated 3.260 features to describe the stages. However, the cost of calculation of the features and the presence of noise and irrelevant features in the original set influenced a selection of the relevant features by a process of FS developed in two steps. The first step tested several statistical methods of FS to find the features with the best discrimination. The result of the test showed that the ReliefF algorithm achieved the best discrimination for the analyzed data. By This Means, RefiefF evaluated all features and returned the ranks and weights, the first half of the sorting features with a high score were selected to be evaluated in a second step by the second algorithm of FS. Finally, the Forward Sequential Feature Selection (FSFS) algorithm was used to reduce the dimensionality of the subset of 1620 features generated by ReliefF. The final subset of features (30) was used with different classification algorithms and the best results were achieved by using the Support Vector Machine (SVM) algorithm. Although the FS process designed in this study integrated the power of FS methods based on filters and the wrapper methods, the analysis of the second stage could be biased by the ranks generated by ReliefF.

Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network [40]

Elgin Crist et al. [40] proposed a framework of classification, where they included a new strategy to select features based on an ensemble of bio-inspired algorithms. For the evaluation, the authors used the Hepatitis dataset and Wisconsin Diagnostic Breast Cancer (WBDC) dataset from UCI Machine Learning Repository. Because of the clinical data usually need preprocessing tasks, the proposal implemented hot deck imputation to handle missing values in the datasets. Thus, the imputation data was a preview task to apply before selecting relevant features. The final optimal set of features is selected calculating the correlation values of the features selected by each FS algorithm and removing the features with high similarity from each subset of features, then, they applied majority voting to select the relevant features from the three subsets features. Finally, a backpropagation Neural Network was used to support the classification achieving precisions above 90%. Although the authors proposed a novel feature selection strategy showing excellent results in the classification, the authors did not consider the effect of the imputation data on the FS and classification processes.

Heterogeneous ensemble feature selection based on weighted Borda count [41]

This proposal presents a heterogeneous ensemble feature selection method build on 8 FS methods: Fisher FS, Robust FS, t-test based FS, FS based on Pearson Correlation Coefficient, FS based on Gini index, ANOVA, ReliefF, and Mutual Information Coefficient. The strategy applies on the dataset the eight algorithms and gets the most relevant features according to each method. Finally, the subsets generated by each FS algorithm are aggregated using a weighted Borda count. Although an aggregation based on weighted Borda count could be a dominant method to get a final subset of features, for this study the aggregation could be biased by the subsets of features generated by the 8 FS algorithms.

One of the principal reasons to develop ensemble feature selection is to use the diversity of FS to get a robust set of relevant features. However, in this study, the authors selected FS algorithms based on filters and do not consider wrapper and embedded algorithms.

Hybrid dimensionality reduction forest with pruning for high-dimensional data classification [47]

Chen et al. [47] designed an alternative method for reducing dimensionality in high-dimensional data. For this, they proposed to select the first subset of relevant features by using an FS algorithm and generate a second set based on the unselected features. The second set is calculated by applying PCA to reduce dimensionality and remove noise features in the unselected features of the first stage. Thereby, the second set of selected features are additional features that are used as auxiliary information. Even though the strategy proposed could be a useful tool to get new relevant information from the dataset, in clinical data we should know the meaning of each feature and transform features such as gender, age, etc., could bias the classification models.

A hybrid method for dimensionality reduction in micro array data based on advanced binary ant colony algorithm [48]

In this paper, the authors proposed a method to reduce dimensionality and face the “Curse of dimensionality” problem. The proposal describes a hybrid method based on two stages. The first stage applies several filter methods of FS and then, in the second stage, an advanced binary ant colony (ABACO) meta-heuristic algorithm is applied on the subset of relevant features generated in the first stage. The evaluation conducted in this study used five well-known high-dimensional microarray datasets. The results confirmed the effectiveness of the hybrid method as, they were compared with the results of several states of the art algorithms. However, when we use two or more stages to feature selection and the outcomes of one stage are the inputs of the next stages, we should reduce the risk that the previous stages will not bias the next ones. In this sense, we should make sure that the FS algorithms used in the first stage include algorithms based on filters, wrappers, and embedded to do not bias the next stage with the weaknesses of the unique type of algorithm.

Ensemble feature selection in high dimension, low sample size datasets: parallel and serial combination approaches [49]

Tsai et al. [49] conducted a study to analyze ensemble feature selection in high dimension, low sample size (HDLSS) data by using parallel and serial combinations in FS algorithms. For the evaluation, the authors used three single baseline FS algorithms: Principal Component Analysis (PCA), genetic algorithm (GA), and C4.5 decision tree. Thus, three subsets of relevant features are generated by applying the mentioned algorithms, which are combined using two or three subsets for calculating the union, intersection, and multi-intersection of the subsets. These combinations are evaluated in the parallel analysis. Also, for the serial analysis, the subsets generated by the three algorithms are used to produce six heterogeneous combinations between two subsets and three homogeneous combinations. Finally, the results of the evaluation showed that classification accuracy achieved with features selected by the ensemble feature selection performed marginally better than accuracy achieved with features selected by single feature selection. This study showed promising results in the use of Ensemble Feature Selection regarding Single Feature Selection, however, the evaluation conducted only analyzed 3 algorithms of FS and the results are not conclusive.

A new hybrid ensemble feature selection framework for machine learning-based phishing detection system [50]

In this study, the authors proposed a framework based on a hybrid method of ensemble feature selection. The approach considered a novel algorithm CDF-g to define automatically the optimal number of features in an FS process applied on a dataset. The hybrid ensemble strategy comprises two components: data perturbation and function perturbation. The data perturbation implies getting multiple subsets of the dataset and applied the same FS algorithm on them, while the function perturbation implies applying multiple FS algorithms on the same dataset. Thus, the hybrid method includes both strategies to acquire a subset of relevant features to detect phishing websites. The results of the evaluation showed that FS could improve the detection and obtain a more stable subset of relevant features, however, the proposal has not been evaluated with clinical data such as physiological signals.

Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data [51]

The study describes the development of an ensemble strategy of feature selection based on the ensemble logic defined in Ensemble Classification. The research evaluated the effects and implications of the data perturbation in the ensemble feature selection; thus, a core algorithm was applied on different perturbed versions of the original dataset. The evaluation conducted analyzed how the ensemble strategy achieves to improve the overall performance of the selection of informative features, which generated the best results in the predictive accuracy and stability. Although this study showed promising results regard to public genomic benchmarks, these results are not conclusive and are defined by the authors as a source of useful insights into the benefits and limitations of the ensemble approach.

A novel automatic satire and irony detection using ensemble feature selection and data mining [52]

Ravi et al. [52] proposed an ensemble text feature selection to support the automatic detection of satire, sarcasm, and irony in news and customer reviews. In this regard, the authors described news and reviews by using different subsets of features extracted from unigrams, semantic, psycholinguistic, and statistical features of the text. In a second step, a feature subset ensemble process is considered capturing different dimensions of features of a text corpus, thereby, the strategy combined the subsets of features together, and obtained seven subsets of features for each corpus, which generates seven datasets. Although the study proposed the next stage to analyze the relevance of the features on each dataset, this process is based on single feature selection algorithms.

2.4. Conclusions

This chapter allowed us to present the context of the analysis of brain activity and conduct a systematic mapping to show the principal contributions of techniques of ensemble feature selection to identify relevant features in the EEG signals.

The main findings identified in the systematic mapping are described below:

- The most frequent types of contribution found in the review of the systematic mapping are new techniques or tools designed to solve specific problems, which shows important contributions in terms of algorithms, however, these contributions have not been considered into robust frameworks or libraries capable to support feature engineering process in another field.
- In the reviewed literature, we found several studies where they proposed ensemble feature selection under a cascade scheme. Following this scheme, the result of the first algorithm could skew the result of the second algorithm and so on, which could generate a final set of relevant features without the true relevant features.
- Although the idea of ensemble learning is to combine the decisions of different models to compensate the weakness of some algorithms with the strengths of the others, many of the reviewed studies only considered FS algorithms based on filters, which could discard relevant features that usually are identified through the wrapper and embedded methods.
- The most used aggregation method to support ensemble feature selection in the reviewed studies is voting, which shows the need to use more robust functions where we considered in the final subset of relevant features the importance based on the frequency of each feature in the subsets of relevant features.
- The authors proposed novel ensemble feature selection strategies showing excellent results in the classification, however, they did not consider the effect of the missing values and imputation data into the FS and classification processes.
- Finally, in the reviewed studies we did not find an ensemble feature selection method designed to select relevant features and support automatic detection of epileptic seizures.

Chapter 3
Dataset of EEG Signals

3.1. Introduction

This chapter describes the process of construction and annotation of the EEG signals dataset taken from patients suspected of having epilepsy. The CRISP-DM methodology described in Figure 9 comprises a set of phases and tasks that must be carried out in a Data Mining project. According to the method proposed by IBM in [53], CRISP-DM can be considered as a methodology or a process model that summarizes a Data Mining life cycle. As seen in Figure 9, the Life Cycle describes 6 phases with arrows that show the sequence and dependencies between the phases. In some phases, there are bidirectional relationships, which indicate that they can be partially or revised after being carried out.

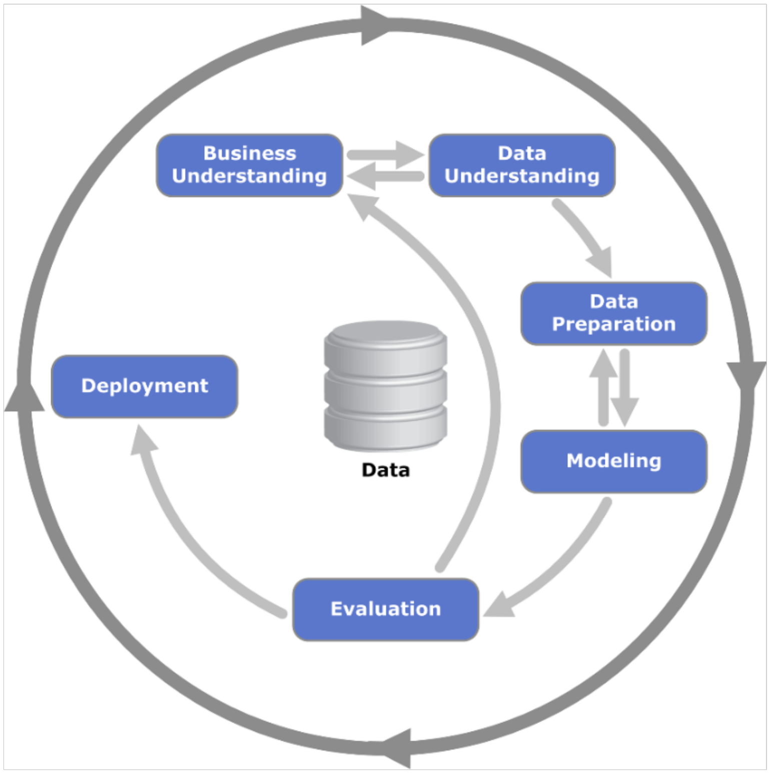


Figure 9. CRISP-DM

Bearing in mind that the aim in this chapter is to describe the dataset construction process and not the model building and evaluation, only the phases of *Busines*s Understanding*, Data Understanding, and Data Preparation* will be carried out.

3.2. Business Understanding

In this phase, CRISP-DM recommends a set of tasks to identify the objectives and requirements without including implementation details of the ultimate solution.

3.2.1. Determining Business Objectives

In this project, we will only consider the tasks associated with the building of the EEG dataset because our goal is to develop a mechanism for selecting relevant features of EEGs. However, it is necessary to define the business objectives to represent the right information to describe the normal and abnormal brain activity in an EEG record. Although we followed CRISP-DM only to build a dataset, the business objectives are associated with the reduction of the time used to read an EEG record and deliver a diagnosis to a patient. In this sense, the final dataset must contain data necessary to train and test the feature selection mechanism and support the automatic reading of EEGs of pediatric patients.

3.2.2. Assessing the situation

The aim of this task is to determine the requirements of the problem in terms of business and data mining. For this, the following questions are answered:

- What is the available prior knowledge of the feature selection to support the classification of epileptic events in an EEG?
- Are there enough EEGs to characterize normal and epileptic signals?
- What is the benefit of the Data Mining application for neurologists?

Considering the need to understand the problem, two experimental studies, [54] and [55], were carried out (Appendixes B and C), to interpret the data, understand the manual process of detecting epileptiform events in EEGs and define a representation of the EEG signals in the dataset. In both studies, the literature was reviewed to contrast existing solutions and identify the usefulness of the proposal.

3.2.2.1. Study 1

A study was conducted to identify public datasets with EEG records of patients with epilepsy. The analysis carried out allowed to identify the structure and type of information available in each EEG. As part of this work, EEG records from Physionet's CHB-MIT database [56] were segmented for the building of an experimental dataset and the construction of a preliminary classifier of abnormal and normal signals [54]. The selected segments corresponded to EEGs with periods of epileptic activity and EEGs with normal brain activity. This implementation was carried out to understand the nature of EEGs and the behavior of brain activity in epileptic patients.

Figure 10 describes the general scheme followed to build the classifier of epileptic seizures. It is worth mentioning that the stages of signal acquisition and signal processing components are commonly included in the software that captures the EEG signals, therefore, they were not addressed in the study. Feature Extraction describes the process used to calculate a representation of the EEG signals while the Classification stage describes how to determine whether the signal is normal or abnormal through Machine Learning and Artificial Intelligence algorithms. Also, if an abnormality is identified, it is possible to classify the signal according to the type of Epilepsy suffered by the patient.

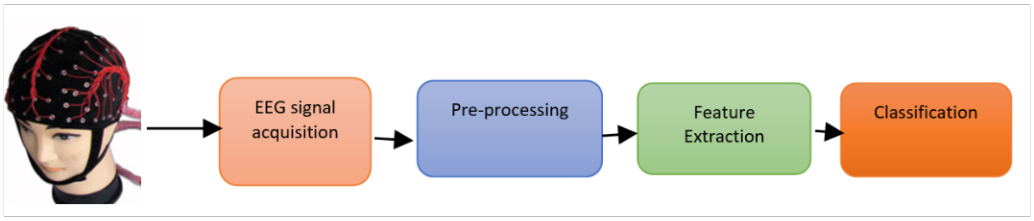


Figure 10. Classification scheme

- Context of the Dataset

To build the preliminary dataset, we studied the CHB-MIT database recorded at Children's Hospital in Boston. This contains recordings from 23 pediatric patients using the international 10-20 system of EEG electrode positions and nomenclature with a sampling rate of 256 samples per second. Each EEG record contains information of 23 channels following the standard European Data Format (.edf) used to exchange and store multi-channel signals of biological and physical origin. In the database, each patient has several EEG records; some of them have epileptic seizures and others describe normal brain activity.

- Feature Extraction

To describe the segments of EEG extracted from the CHB-MIT database, a feature extraction process was conducted. This describes the process of applying several operators to obtain a set of descriptors for each segment. The descriptors represent the information contained in the EEG signal. According to the literature, the following descriptors are identified and calculated: Entropy [57], Maximum Amplitude, Minimum amplitude, Mean, Variance, Maximum Power and Mean Power [58]. The abovementioned features were computed for each EEG channel ending up in 161 descriptors. Moreover, to define the relevant features from the set of 161 ones, we evaluated the Forward Selection, Optimize Selection, and Backward Elimination algorithms.

- Classification

The features calculated in the feature selection process were used to train a machine learning model and determine whether a signal is normal or abnormal. The algorithms Naive Bayes, Rule Induction, Decision Tree, and KNN were combined with the feature selection algorithms Forward Selection, Optimize Selection, and Backward Elimination to determine the relevant features capable to get high accuracy in the classification process. The results of these combinations are shown in Table 7.

Table 7. Accuracy of the algorithms used in the feature selection process.

	Forward Selection	**Optimize selection**	**Backward Elimination**
Naïve Bayes	57.69%	42.30%	48.07%
Rule Induction	70.00%	**63.46**%	40.38%
KNN	71.11%	42.30%	40.38%
Decision Tree	**80.77**%	34.61%	48.07

In the evaluation, the Decision Tree algorithm produced the best results in the classification by using the subset of features selected by the Forward Selection Algorithm (FSA). Figure 11 describes the design built in RapidMiner to select the features. Figure 12 shows the classification scheme inside of the Forward Selection Package (Figure 11).

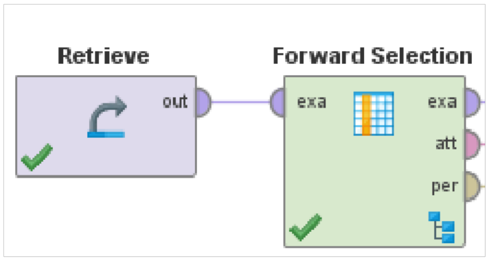


Figure 11. Feature selection scheme

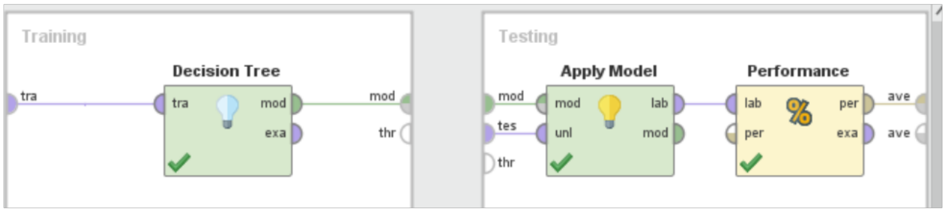


Figure 12. Classification scheme

The subsets of features generated by the FSA according to each classification algorithm tested are described in Table 8. The FSA reduced 161 features to (i) 4 using Naive Bayes, (ii) 8 using Rule Induction, (iii) 4 using Decision Tree, and (iv) 5 using KNN. However, the best accuracy was achieved using the subset of features generated using the Decision Tree as a classifier.

The features selected were a6, a27, a94, and a121. According to the encoding scheme used to name the features, it was observed that the Mean Power of the first channel, Mean Power of the second channel, the Minimum Amplitude of the fourteenth channel, and the Maximum Amplitude of the eighteenth channel are the features that determine whether a signal is normal or abnormal.

This proposal was tested with data from a real database, i.e., data collected in a clinical setting, achieving an accuracy of 80.77% using only 4 features whereas in the literature, other proposals achieved better results using over five features and using experimental data [58][59][60], data collected in non-clinical settings.

Table 8. Features selected using the FS algorithm.

Feature	Naive Bayes	Rule Induction	Decision Tree	KNN
a6			x	
a26	x			x
a27			x	
a32		x		
a42				x
a45	x			
a51	x			
a32		x		
a69		x		
a81		x		x
a82				x
a83		x		
a94			x	
a121			x	
a129		x		
a136		x		
a137		x		
a160	x			x

Although accuracy of 80.77% was reached, we detected that the reason some segments of EEGs were classified badly, is because some segments contained abnormalities only in some channels. This is a consequence of seizures that can be either focal (some channels) or generalized (all channels). In this sense, an enhancement for this study is to implement the classification of channel segments instead of EEG segments. Likewise, an additional explanation of why the accuracy is not better is because abnormal activity could appear in EEGs in different shapes.

The development of this study allowed us to identify needs and challenges in the automatic reading of EEGs. Some of these considerations to be borne in mind are described below.

- The performance of a solution to automatically read an EEGs must process a vast amount of data because of the nature of the recordings. In consequence, feature extraction must include the least number of possible features.
- Characterize types of abnormalities could increase the precision and accuracy of the detector/classifier due to it can apply the right extractors.
- Available, open EEGs datasets do not describe the exact periods of time where a single seizure happens, nor the channels affected. By contrast, they describe periods happening various seizures (normal activity along of them). On the other hand, some datasets describe only single segments extracted from EEGs that describe the normal or abnormal activity.

3.2.2.2. Study 2

Having regard to the above, one of the relevant considerations was the identification of types of abnormalities associated with epilepsy to build characterizations based on identified patterns (shapes). In this sense, a preliminary implementation to detect well-known shapes or patterns of seizures was built. For this, a review of literature recommended the use of Matched Filters, which are briefly used in the field of signal processing to seek wave shapes in a signal.

- Characterization of Patterns

Matched filters are basic signal analysis tools used to extract known waveforms from a signal that has been contaminated with noise [61]. The model used for the extraction or detection of the wave can be seen in Figure 13.

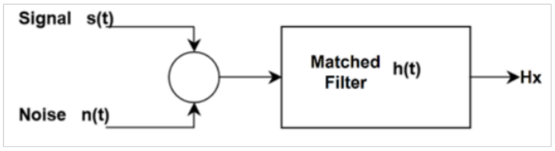


Figure 13. Detection scheme, after [61].

The scheme defined in *Figure 13. Detection scheme, after* [61].Figure 13 describes the implementation of a filter *h(t)* to extract the signal *s(t)* contaminated with noise *n(t)*, because of applying *h(t)*, finding the hypothesisH_x. In this scheme, the null (H_0) and alternative (H_1) hypotheses are considered in equations (1) and (2). If the waveform that is sought is present in the signal, hypothesis H_1 is confirmed, otherwise, the H_0 hypothesis is confirmed. In the context of the detection of epileptic spikes, *x(t)* is a function describing the brain activity measured in an EEG, the noise *n(t)* represents a normal brain activity of a patient (EEG base rhythm), the signal *s(t)* represents the epileptic spike to be found, H_0 *normal* represents the normal EEG activity of the patient and H_1 represents the existence of the epileptic spike.

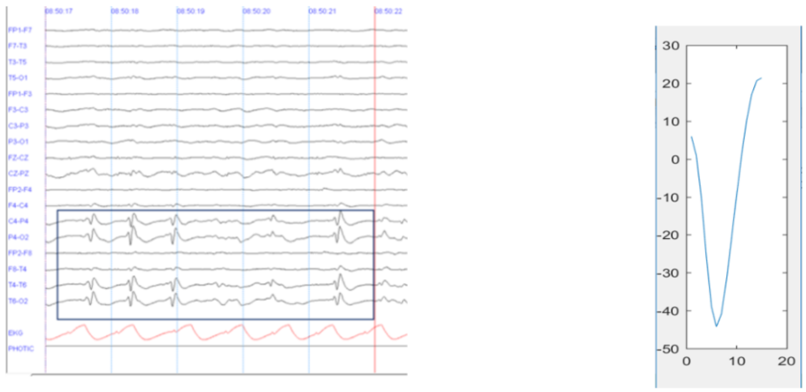
$$H_0 : x(t) = n(t) \tag{1}$$

$$H_1 : x(t) = s(t) + n(t) \tag{2}$$

This mechanism works very well in practice when a known pattern or waveform is sought because the filter allows to maximize the SNR (signal-noise ratio) of the filtered signal and reduce the effect of noise on the original signal [62]. However, when waveforms are not known, the method does not work efficiently.

Considering the importance of spikes detection to diagnose Epilepsy, it was conducted a study to detect them by using a Match Filter. In this work, the aim was to develop a tool to diagnose epilepsy, identifying epileptiform events described by known waves. For this purpose, a review of the literature was made to identify characteristic patterns that describe the presence of an epileptic discharge. Epileptic seizures generate electric shocks in some areas of the brain, generating unexpected changes in the EEG waveforms. Sometimes, the appearance of shocks (waveforms) is identified periodically, semi-periodically, or simply as a disorganization of the electrical activity of the patient. Some of the most wanted patterns by neurologists during the inspection of EEGs correspond to peaks (narrow and broad).

Considering the above, a template that describes the behavior of a spike was defined and used to search similar segments into an EEG signal. The template was constructed by averaging 25 segments diagnosed as spikes by a neuropediatric expert in reading EEGs. Figure 14 (a) shows an example of the appearance of epileptic spikes in the base rhythm of the EEG wave on channels 17, 18, 22, and 23 of the EEG.



(b) Epileptic Spikes in the rhythm base. **(b).** Epileptic spike pattern.

Figure 14. Epileptic spikes in the rhythm base.

- Building the Matched Filter

From a set of epileptic spikes identified by a neurologist, the epileptic spike pattern of Figure 14 (b) was constructed. Considering the visual analysis performed by the neurologist, it was defined that the segments contain 15 samples of data, and a duration of 13.33 ms (to be able to capture the whole epileptic spike from the beginning to the end of the abnormality).

Considering the wave pattern (template) which describes an epileptic spike, a Matched Filter was built to scan an EEG, searching the template by sliding windows over each EEG channel. The algorithm is defined below:

Algorithm 1. Spikes detector.

```
Void SpikesDetector (windowSize, slidingSize, pattern, EEGChannel,
spikesBeginnings, spikesEnds, thresh)
    startIndex = 0
    maxIndex = Lenght (EEGChannel)
    b_matchedFilter = createMatchedFilter(pattern)
    while (startIndex< maxIndex) do
            segment = EEGChannel(:, startIndex: maxIndex + windowSize)
            matches = matchedFilter(segment, template, thresh, b_matchedFilter)
            if (isNotEmpty(matches)) do
                    spikesBegginings.Add (startIndex)
                    spikesEnds.Add (startIndex+windowSize)
                    startIndex = startIndex + slidingSize
            else
                    startIndex = startIndex + windowSize
            end if
End SpikesDetector
```

The algorithm receives 7 arguments: the size of the window, size of sliding, pattern, EEG channel, beginning and end of detected segments, and threshold. The size of the window allows determining the start and end of the segment to be analyzed, as well as the size of the sliding, which establishes how many samples move to the right of the beginning of the segment that has been analyzed. The threshold establishes the percentage of similarity between the window analyzed and the template of spikes. Figure 15 illustrates the previously mentioned process.

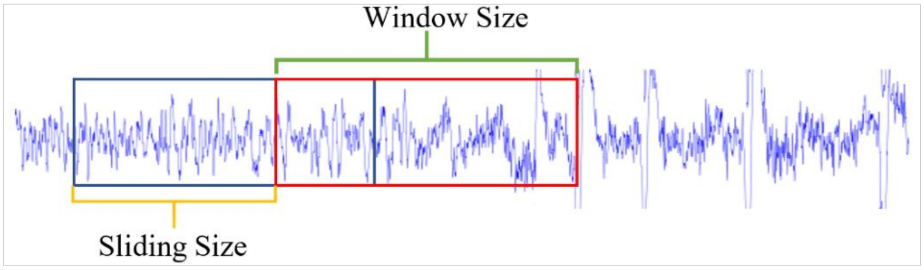


Figure 15. Analysis scheme by window.

The argument pattern corresponds to the template constructed from the epileptic spikes, EEGChannel corresponds to a channel extracted from the EEG in which the pattern will be searched. spikesBeginnings and spikesEnds correspond to the arrangements where the beginning and end of the segments (that have the presence of the pattern of epileptic spikes) are stored, and the function createMatchedFilter creates a matched filter based on the template. The algorithm analyzes the entire EEG channel, while segments can be extracted through the sliding window, and for each window extracted a review is made with the Matched Filter to determine if this window has the presence of the epileptic spike pattern.

The algorithm that describes the Matched Filter is described below:

Algorithm 2. Matched Filter.

```
Var matches MatchedFilter (segment, template, thresh, b_matchedFilter)
    y = FilterSignal(b_matchedFilter, segment)
   u = template.'*template
    matches = ReviewThreshold (y,thresh,u)
    return matches
End MatchedFilter
```

Where, segment describes segment to evaluate, template represents the epileptic spike pattern, thresh sets a detection threshold, which was established empirically in 0.9 by testing values between 0.6 and 1, b_matchedFilter contains the matched filter based on a template, y contains the segment filtered with the matched filter, u stores the autocorrelation matrix of the template, and function ReviewThreshold establishes if y exceeds the threshold. The autocorrelation matrix is used for detecting the appearance of patterns in a signal. In this case, the autocorrelation matrix was used for detecting the pattern of spikes in brain activity.

For the evaluation of the epileptic spike detector, 8 segments extracted from EEGs of voluntarily recruited children were used. For this stage, construction of the dataset aroused as a need to have a set of training data that describes in detail the beginning and end of an epileptic abnormality. This because the open datasets found in the literature, describe periods of time with normal and abnormal activity, it means that the segments show the appearance of an abnormality and then disorganization or new appearances of the abnormality instead of showing the specific segments where epileptiform events happen.

Considering the annotations made on the dataset, the beginning and end of 56 segments in which epileptic discharges occur in the form of a spike are known. In this sense, the spike detector was used for each EEG segment, and the correctly identified, badly identified, and unidentified segments were counted to determine the sensitivity and specificity of the detection. *Figure 16* describes the number of spikes contained by each extracted segment.

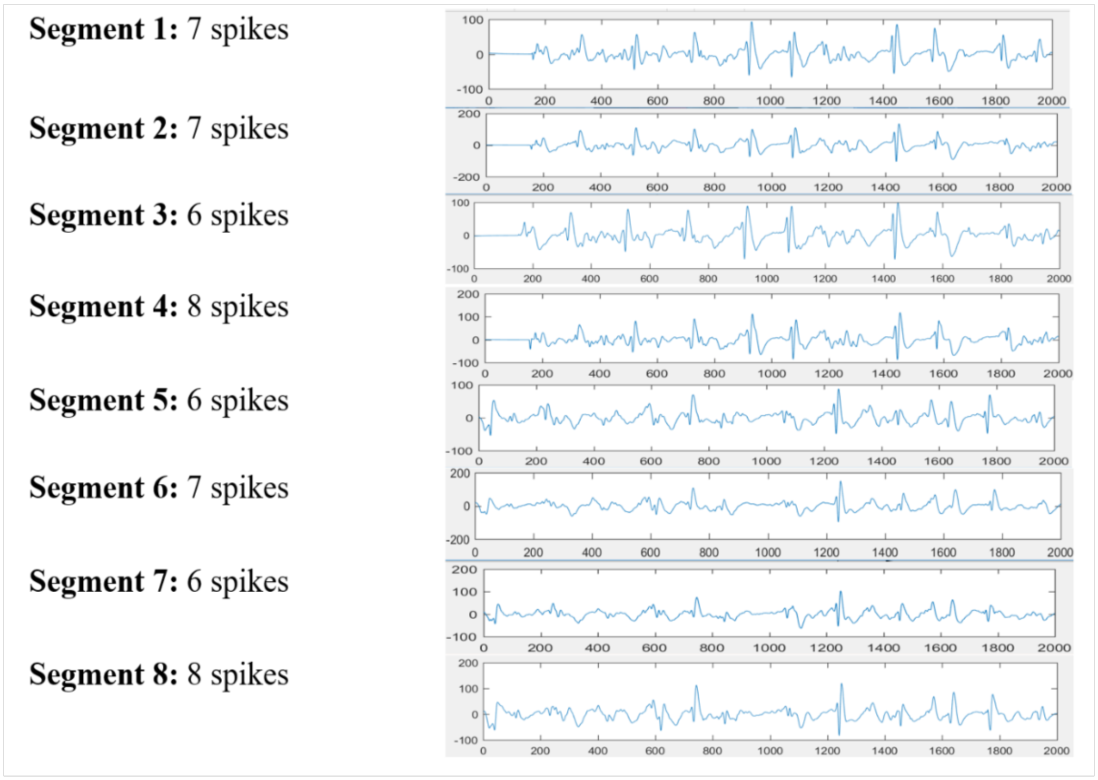


Figure 16. Description of EEGS segments with presence of points.

Finally, the tool developed reached a sensitivity of 89.28% and specificity of 99.96% in the identification of epileptic spikes on our dataset.

3.2.2.3. Conclusions of the Preliminary Experiments

The main conclusions found with these studies are:

- The reading of EEGs implies identifying alterations in the waves that describe the electrical activity of the brain.
- There are two types of epileptic disorders: Focal and Generalized.
- There is a set of abnormal patterns that neurologists try to identify on EEGs to diagnose epilepsy.
- In the literature, no dataset is annotated in such a way that it describes exactly the EEG segments where an epileptiform event occurs, nor does it describe the type of event. The datasets describe periods of abnormal brain activity but not the exact times of each type of abnormality.

- The recording period of an EEG ranges from 20 minutes to 72 hours.
- Manual reading of an EEG can become a tedious and time-consuming task.
- The software tools identified in the literature have shown a potential use of these to reduce the time of reading an EEG by neurologists.
- When the availability of neurologists is limited, the automatic reading of EEGs could be an option to increase the opportunity of the service.

3.2.3. Determining the Data Mining Goals

This task allows identifying the Data Mining objectives of the project. Here, the business aim is to decrease the EEG reading time and the Data Mining objective is to determine how to automatically detect an epileptiform event in an EEG.

3.3. Data Understanding

This stage describes a set of tasks that guide the initial data collection and analysis.

3.3.1. Collecting Initial Data

In this task, the initial data is collected for future processing. Considering that in the exploration of the literature a dataset with precise annotations on epileptiform events was not identified, 200 EEGs were collected from childhood patients with a history of epilepsy.

The stored EEGs were captured using the BW Analysis software from the Neurovirtual company under the 10-20 system. Each EEG contains information in 21 channels with a sampling rate of 200 per second and a duration of approximately 30 minutes. The Axon PED clinic was the medical center where the EEGs were recorded and diagnosed.

3.3.2. Describing Data

Approval for collection and storage was obtained from each EEG record by signing an informed consent by the child's parent or guardian. The collection protocol included a sequential set of steps: (i) obtain the informed consent, (ii) take the EEG, and (iii) code the EEG to eliminate the data that allow the identification of the patient.

Table 9 describes the storage and organization scheme for EEG records. Column 1 represents the coding of each EEG; the next 6 columns describe the data associated with the patient and column 7 the final diagnosis made by the neurologist. Each EEG record was stored according to the code assigned in Table 9 in EDF format.

Table 9. Collection and storage of EEG

Code	**Date**	**Age**	**EPS**	**Clinical Data**	**Medicament**	**Status**	**Diagnosis**
1001	7/07/2017	11	ASMET		No	Awake	Normal

The EEGs documented as abnormal in Table 9 were reviewed and annotated with the help of a neuro-pediatrician. This process was conducted to identify the segments and channels of the EEGs that present abnormalities associated with epilepsy. For this, reviewing sessions were carried out to read each abnormal EEG and to identify the beginning and end of the segments of the epileptic discharges, as well as to make a description of the disorder in the brain activity before and after the epileptic discharge.

Figure 16 presents a photograph taken during an EEG annotation session. At home, during the EEG review session, the neurologist graphically signaled the beginning and end of each of the events of interest, later to extract the segments as individual files that describe the abnormality and the type of abnormality.

Figure 17. EEG annotation session.

Additionally, a review of EEGs registered as normal was made to identify segments with apparent abnormal electrical activity that are artifacts caused by movements of the patient during the examination.

3.3.3. Exploring Data

The data collected for this study include the data of Table 9, the meaning of each column, the EEG stored for each row of the Table, and the annotations of each EEG. To analyze this information, each EEG was inspected by using the annotations made by the neurologist, which allow to count and identify the unique patterns associated with epilepsy. Additionally, we reviewed the clinical data of the patients looking for a relationship between the findings in their EEG and their medical records. However, the data that will support the Data Mining process corresponds to the result of the feature extraction process applied to each abnormal and normal record extracted in the previous stage.

Some of the epileptic abnormalities that were documented during the dataset annotation process are described below.

3.3.3.1. Types of Abnormalities

The segments diagnosed as abnormal by the neurologist were grouped according to the type of abnormality and subsequently characterized by a set of descriptors during the feature extraction process. Next, segments of EEGs with occurrences of the types of abnormalities identified are described:

- Spikes

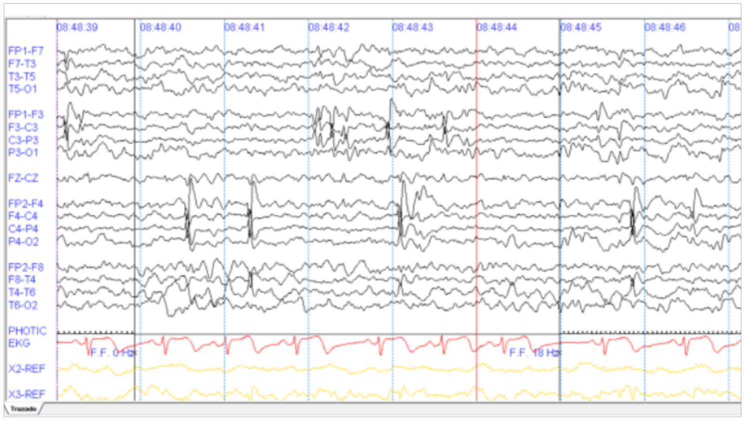


Figure 18. EEG with spikes.

- Phase version

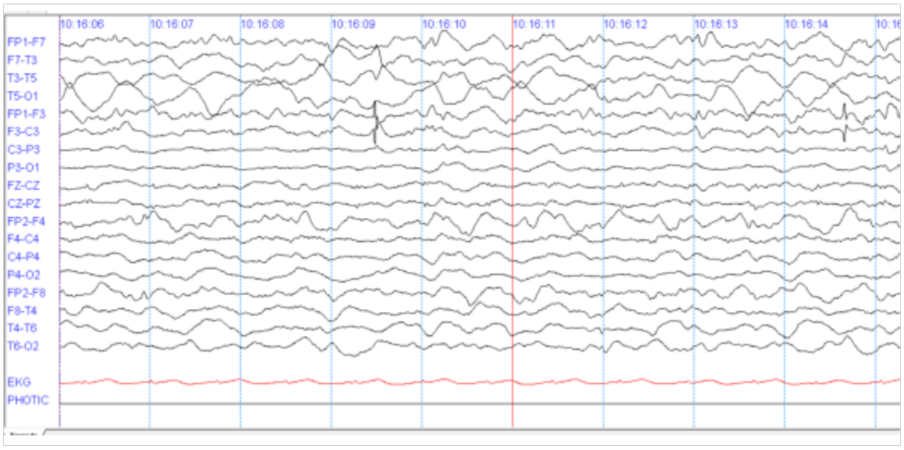


Figure 19. Phase version

- Pseudoepileptic sharp waves

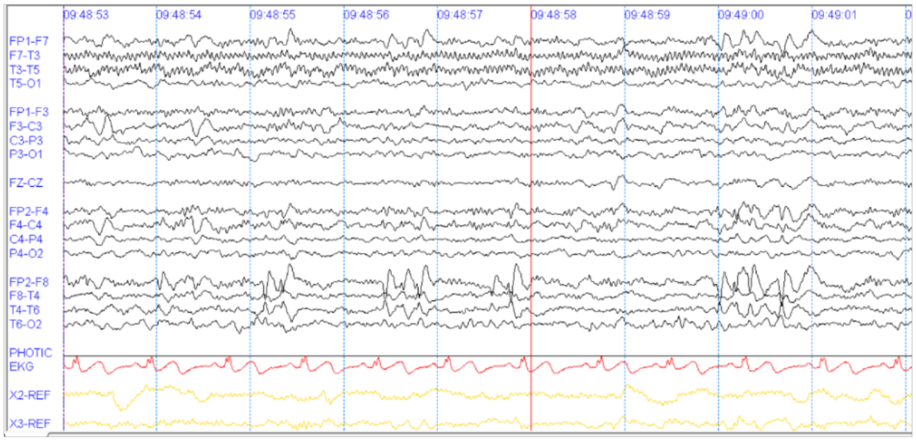


Figure 20. Pseudoepileptic sharp waves

- Polyspikes

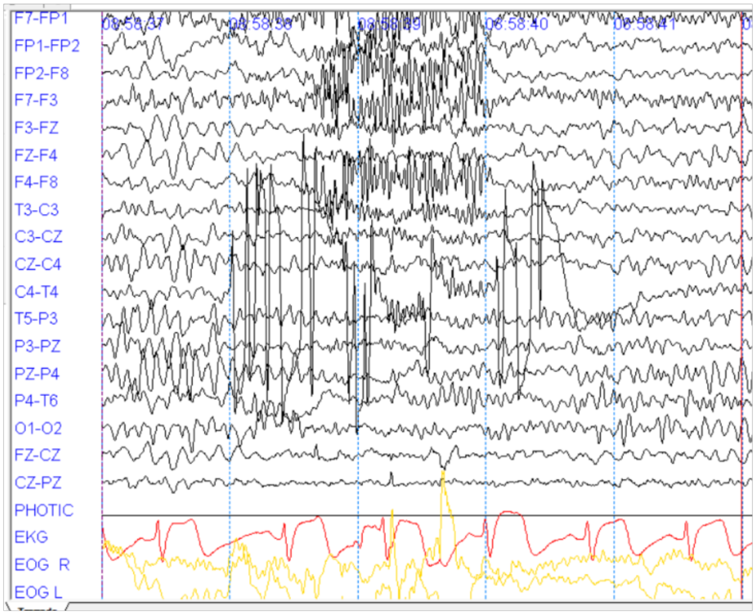


Figure 21. Polyspikes

- Focal discharge

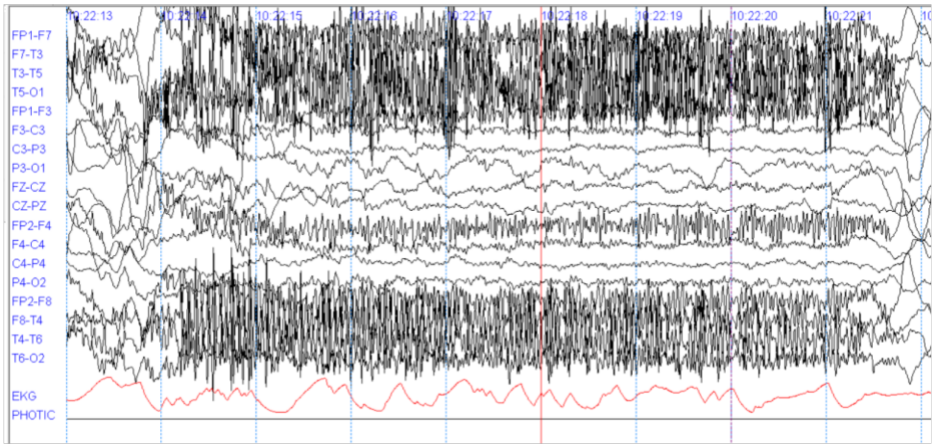


Figure 22. Focal discharge.

For each type of abnormality, the number of appearances in the EEGs collected in the previous task was identified. Table 10 describes the summary of the alterations identified in the EEG repository.

Table 10. Abnormal segments by type.

Abnormality	# Events	# Patiens
Spikes	144	23
Diffuse slow activity	10	10
Phase Version	113	4
Pseudoepileptic sharp waves	6	2
Polyspikes	42	18
Focal discharge	1	1
Generalized discharge	1	1
Photomioclonic response	24	1

3.4. Data Preparation

In this stage, the data is adapted according to the needs of the data mining techniques that are going to process the data. In the previous stage, we combined a repository of EEG signals with clinical information and annotations made by a neurologist. In this state, a set of abnormal and normal segments was extracted from the EEG signals and the segments were characterized by a set of feature extractors to generate a dataset. Then, the dataset was built with the features extracted, which generated a dataset with 142 columns and a target variable.

3.4.1. Integrating Data

The EEG signal repository contains 200 records from 200 patients that, given their structure, cannot be processed by Machine Learning algorithms. Thus, each EEG is decomposed channel by channel, and those segments diagnosed as abnormal are extracted and described using a set of feature extractors. This same process is carried out for a set of segments considered normal.

Finally, the dataset is built with the features extracted from 1344 segments (672 normal and 672 abnormal). Since all the descriptors were applied to all the segments, the dataset does not contain columns with null data.

The descriptors used to represent as vectors the segments extracted from the EEG repository are described below.

- Basic descriptors

Statistical features allow summarizing the values that describe a segment of EEG signal in a single value. The measures of this type that will be applied in the construction of the dataset are:

- Min
- Max
- Mean
- Median
- Low Median
- High Median
- Variance
- Standard deviation

- Entropy

Entropy is considered a family of statistical measures that allow the quantification of the variant complexity in a system. In [63] the authors describe different ways of measuring Entropy, highlighting the following:

  - Shanon Entropy

$$H_\alpha(\wp) = \frac{1}{1-\alpha} log_2 \left\{ \frac{\sum_{k=1}^{n} p_k^\alpha}{\sum_{k=1}^{n} p_k} \right\}$$

- Approximate Entropy

  $$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r)$$

- Renyi Entropy

  $$QSE(x, m) = SampEn(x, m, r) + log(2r)$$

- Kurtosis and Skewness

The skewness and kurtosis are higher-order statistical attributes of a time series.

- Skewness represents the degree of distortion from the symmetrical bell curve or the normal distribution. In other words, the lack of symmetry in data distribution is measured by skewness.
- Kurtosis measures the peakedness of the PDF of a time series. It is used to measure the outliers present in the distribution.

- Energy

The signal is viewed as a function of time and the Energy represents its size. The energy can be measured in different ways but, the area under the curve is the most common measure to describe the size of a signal. It measures the signal strength, and this concept can apply to any signal or vector.

$$E_x = \int_{-\infty}^{\infty} |x(t)|^2 dt$$

- Fractal Dimension - Higuchi

The fractal dimension corresponds to a non-integer dimension of a geometric object. Based on this principle, fractal dimension analysis is used to analyze biomedical signals. In this approach, the waveform is considered as a geometric figure.

$$D = \frac{d\ \log(L(k))}{d\ \log(k)}$$

- Fractal Dimension - Petrosian

This type of analysis provides a quick mechanism to calculate the fractal dimension of a signal by passing the series in a binary sequence. The following describes the equation that calculates the Petrosian fractal dimension.

$$FD_{Petrosian} = \frac{\log_{10} n}{\log_{10} n + \log_{10}\left(\frac{n}{n+0.4N_{\Delta}}\right)}$$

- Hurst exponent

This exponent is a measure of the predictability of the signal. It is a scalar between 0 and 1 which measures long-range correlations of a time series.

- Zero Crossing Rate

The Zero Crossing Rate is a statistical feature that describes the number of times that a signal crosses the horizontal axis.

- Hjort Parameters
  The Hjort Parameters describe statistical properties in the time domain [64]. Usually, these are used to analyze electroencephalography signals.

  - Activity
    Also known as the variance or mean power. Activity measures the squared standard deviation of the amplitude.

$$Activity(x) = \frac{\sum_{n-1}^{N}(x(n) - \bar{x})}{N}$$

  - Mobility

    Mobility measures the standard deviation of the slope concerning the standard deviation of the amplitude.

$$Mobility(x) = \sqrt{\frac{var(x')}{var(x)}}$$

  - Complexity
    This parameter is associated with the wave shape.

$$Complexity(x) = \frac{Mobility(x')}{Mobility(x)}$$

- Discrete Wavelet Transform

The Discrete Wavelet Transform allows the analysis of a signal in a specific segment. The procedure consists of expressing a continuous signal as an expansion of coefficients of the internal product between the specific segment and a Mother Wavelet Function. The discretization of the Wavelet Transform is done by changing from a continuous mapping to a finite set of values. This process is done by changing the integral in the definition by an approximation with summations. Hence, the discretization represents the signal in terms of elementary functions accompanied by coefficients.

$$f(t) = \sum_{\lambda} c_{\lambda} \varphi_{\lambda}$$

The Mother Wavelet functions include a set of scale functions. The parent functions represent the fine details of the signal, while the scale functions calculate an approximation. Considering the above, a function or signal can be represented as a summation of wavelet functions and scale functions.

$$f(t) = \sum_{k} \sum_{j} c_{j,k} \phi(t) + \sum_{k} \sum_{j} d_{j,k} \psi(t)$$

In Wavelet analysis, a signal can be decomposed into various levels from the time domain to the frequency domain. The decomposition is done from the detail coefficients as the approximation coefficients. Figure 23 describes the different encoding paths for n levels of decomposition (signal x). The upper level of the tree describes the temporal representation. As the decomposition levels increase, an increase in the compensation in the time-frequency resolution is obtained. Finally, the last level of the tree describes the representation of the signal frequency.

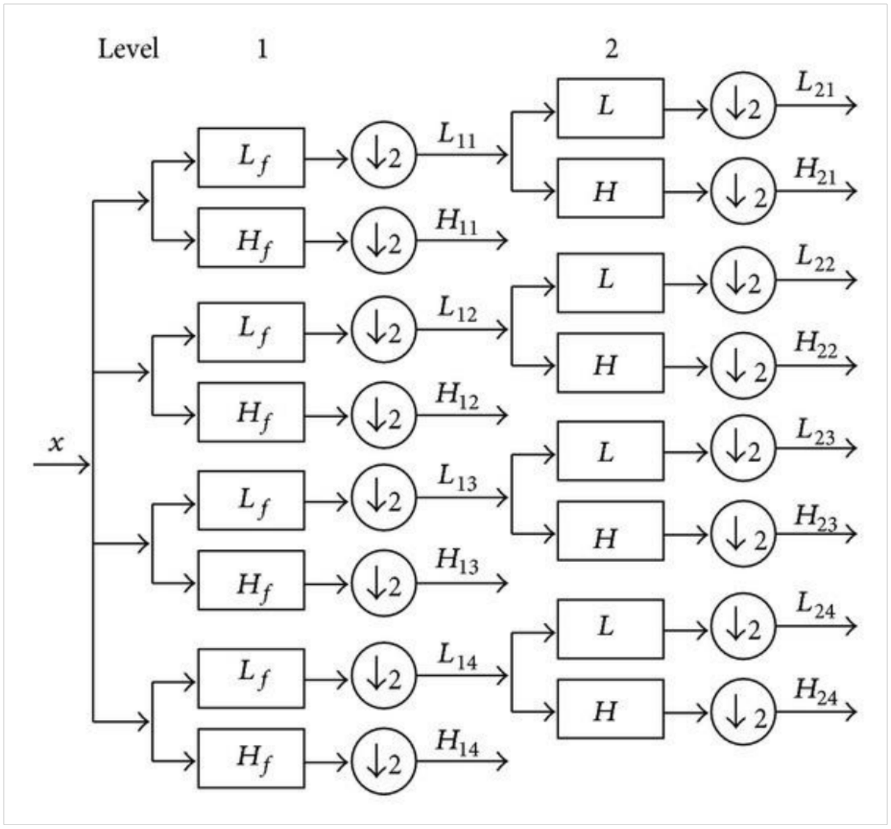


Figure 23. Wavelet decomposition

- Fast Fourier Transform

The Fast Fourier Transform computes a fast version of the Discrete Fourier Transform of a signal by decomposing the original signal into different frequencies (smaller transforms). The decomposed signals are used to calculate the resulting transform signal. FFT is used to convert a signal from the time domain to a representation in the frequency domain or vice versa.

- Features based on Fast Fourier Transform
  - Spectral Centroide

    In digital signal processing, the spectral centroid is a statistical measure used to describe the shape of the spectrum. This defines the spectrum as a probability distribution and represents where the center of mass of the spectrum is located.

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

  - Spectral Flatness

    The Spectral Flatness defines the ratio of the geometric mean to the arithmetic mean of a power spectrum.

$$\text{Flatness} = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} = \frac{\exp\left(\frac{1}{N}\sum_{n=0}^{N-1} \ln x(n)\right)}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)}$$

- Crest Factor

The Crest Factor defines how extreme the peaks are in a signal.

$$C = \frac{|x_{\text{peak}}|}{x_{\text{rms}}} = \frac{\|x\|_\infty}{\|x\|_2}$$

Taking the above into account, 22 descriptors were applied to the normal and abnormal EEG segments and their wavelet coefficients (5) generated from the original segments. 21 descriptors are: Min, Max, Mean, Median, High Median, Low Median, Variance, Standard Deviation, Shanon Entropy, Approximate Entropy, Renyi Entropy, Kurtosis, Skewness, Energy, Higuchi Fractal Dimension, Petrosian Fractal Dimension, Hurst Exponent, zero crossing Rate, Activity Hjort, Mobility Hjort and Complexity Hjort. Besides, the Fatst Fourier Transform (FFT) was calculated, and 15 descriptors were applied to the result of FFT: Min, Max, Mean, Median, High Median, Low Median, Variance, Standard Deviation, Shanon Entropy, Kurtosis, Skewness, Energy, Spectral Centroide, Spectral Flatness and Crest Factor. Also, the Matched Filter was applied to the original segments and a boolean feature was generated with the results.

3.4.2. Formatting Data

To avoid potential conflicts associated with the values in the dataset, an offset was applied to the columns that presented negative values, since some feature selection algorithms do not accept this type of value. On the other hand, a normalization was carried out to guarantee that the data in each column is in a range of [0, 1].

3.5. Conclusions

In this chapter we followed three stages of CRISP-DM methodology: Business Understanding, Data Understanding, and Data preparation, to build a dataset that represents information of EEG signals of patients with normal and epileptic brain activity.

The main findings are described below:

- The studies conducted to understand the business allowed us to understand the business aim: to decrease the EEGS reading time and the Data Mining objective: to determine how to automatically detect an epileptiform event in an EEG.
- The conclusions of studies conducted showed the complexity to detect abnormal brain activity in the EEGs because of the diverse ways that they could appear.
- In the literature, there is no dataset with annotations that describes exactly where an epileptiform event occurs, nor the type of event.
- The duration of EEGs could become the manual reading of an EEG into a tedious and time-consuming task.
- The automatic reading of EEGs could be an option to increase the opportunity of the service.

- The raw data of EEG records cannot be processed by the machine learning algorithms, then, it is necessary to transform the data and extract features from the signals to describe them by using feature vectors.

Chapter 4
Framework for the Ensemble of Feature Selection Methods

4.1. Introduction

The primary objective of this chapter is to describe the process followed to construct a conceptual framework about Ensemble of Feature Selection (EFS) methods. The goal is to represent into a framework, knowledge related to how to improve the FS on datasets with high dimensionality and a few instances. The framework seeks to guide the design of an EFS mechanism to gather the advantages of different FS algorithms, avoid their biases, and compensate for their disadvantages. For this, we designed a conceptual framework to understand the key concepts and relationships in the aggregation of a set of FS algorithms. Based on the conceptual framework, a Development Framework was implemented to validate the theoretical proposal and assessed it by applying the EFS implemented on 4 datasets. The rest of the chapter describes the findings and conclusions of the design and implementation of the framework of FS.

4.2. Context of Feature Selection

A feature is defined as a measurable property of a process or entity that is being observed; in the literature it is also known as attribute, component, variable, column or dimension [65]. In the field of Machine Learning, a set of features describes a domain and classifies, detects, or recognizes patterns. In the past, few Machine Learning applications used over 40 features [27]. However, nowadays the number of features has increased their size from tens to hundreds of features. Consequently, handling this information is costly because of its processing, requiring thus more time and resources. In this sense, many studies in the last two decades faced this problem, especially when datasets have a high number of features and few instances. This problem is called “curse of dimensionality” [66].

Consequently, in 1997 [67] and [68], the authors described the first studies about Feature Selection (FS) in domains where it is common to find datasets with several dozens of features. Nonetheless, in recent years a great deal of techniques has been developed to solve the problems generated by the number of features. According to the findings, most of the features in those datasets are redundant or irrelevant [69]. In view of the above, FS techniques focus on the identification of features with high differentiating power while discarding those considered irrelevant or redundant. In this manner, the primary goal of FS is to avoid features that do not efficiently allow generalization in processes of classification, detection, or pattern recognition.

4.2.1. Dataset's Growth

Recently, datasets with large numbers of features are more frequent in different domains. Three of the most representative examples are microarray classification, text categorization and signal classification. In the first case, developments in DNA (Deoxyribonucleic acid) microarray have generated many datasets with this kind of data. In most of these datasets, the number of instances is not greater than 100 (patients) and the number of features (genes) ranges from 6.000 to 60.000 [70]. Previous studies showed that most of the genes in these datasets do not represent useful information to support a machine learning process. As a result, a preprocessing stage is needed in order to reach an efficient classification of microarray data [11][71]. In [72][73] the authors describe how to reduce computational cost and improve performance in the classification of micro arrays by selecting a representative subset of genes from the original set.

Likewise, in text categorization, documents are represented by an array built from their vocabulary and the frequencies of words in such documents. Those vocabulary sets have hundreds of thousands of words, however, in an initial stage vocabulary is pruned to remove the least important words from the documents. Thus, the size of the array that represents the documents is reduced. In literature, there are several collections of documents used in different domain studies, for instance, email analysis [74], detection of articles related to terrorism on the web [75], automatic classification of text [76], opinions [75], feelings [77] [78], and emotions [79], among others. These collections have between 5.000 and 800.000 documents.

In the field of signal classification, previous works have used many mechanisms to process the signal and get a set of features capable of describing the signal. These features are used to classify or detect patterns. In the medical domain, the high availability of devices designed to capture bio signals has supported the diagnosis of diseases by identifying normal and abnormal patterns in the signals. Hence, several authors have developed solutions to support automatic analysis of signals, such as for example EEG and ECG. In the automatic analysis of EEGs, the signals considered are multichannel with (i) information in channels, which range from 12 to 64 in number, (ii) duration between 20 minutes and 72 hours and (iii) a sampling rate between 100 and 256 per second. Considering the above, analyzing an EEG signal is a complex task because it contains a lot of information, thus each channel from the signal is divided into segments and many feature extractors must be applied to describe them. The process to segment each channel of the signal allows the identification of abnormalities, which appear in short periods of time [80].

In [81] the authors analyzed the current context of “Big Data” and “Big Dimensionality”. They introduced those concepts to explain how to handle datasets with unbalanced data, noise, few instances, and high number of features. They found out that the dataset sizes are not growing in both dimensions, columns, and rows. In addition, the most important repositories of datasets used in the experiments of Machine Learning contain datasets with thousands or millions of features and most times, the number of features widely surpasses the number of rows. For example, in the UCI machine learning repository [82], there are 18 datasets with over 5000 features. The LIBSVM database contains datasets with over a million features [83]. Therefore, researchers have focused on developing methods to reduce the size of datasets using a set of objective criteria. This allows them to represent the complex original dataset in a simple dataset.

4.2.2. Context of Ensemble Feature Selection

Depending on the design of the FS techniques, they are classified into three types of methods: filters, wrappers and embedded. Each type defines advantages or disadvantages that are directly related to the context of the dataset. For instance, these algorithms face typical problems, namely, (i) they have a reliable performance on a dataset but by adding or removing instances, the performance decreases, (ii) they allow the removal of features quickly, but they are not capable of detecting redundant features, (iii) they need to have a correctly balanced dataset and (iv) their performance is affected by the presence of noise in the data.

On one hand, there is a large number of FS methods. However, there are no tools or solutions to determine objectively the algorithms, which would work best with the data of a particular domain. In some studies, authors have used a trial and error scheme, that is, they have tested different FS algorithms using one or more classifiers and then chosen the one with best performance in the test.

On the other hand, the results of the Ensemble Learning approach in classification have influenced proposals to select features based on the consensus or aggregation of several FS algorithms. In [84] researchers proposed a classification algorithm based on K-Nearest Neighbors (KNN). They gained and combined several outcomes from the KKN algorithm, and each outcome was found by using a unique set of features. In 1998, the authors proposed an Ensemble Feature Selection (EFS) method designed for decision trees [85]. In 1999 [86] an EFS method based on a genetic algorithm was proposed to improve the quality of the features used in the learners.

Recently, many studies have been conducted about EFS; some involve the use of classifiers, while others do not. In [70] an EFS algorithm aggregates a set of FS algorithms based on filters to classify micro arrays. The scheme in this study aims to use several filters and for each filter to generate a subset of features. The subsets generated are used to train a classifier and, subsequently, the outputs of the classifiers are combined using simple voting. In [87] a mechanism of EFS on micro arrays was built to determine relevant genes in the classification of cancer. A robust feature selection process is conducted in [42], the selection of features is based on EFS and the findings showed that the approach obtained great promise for datasets with many features and few samples. The bi-objective genetic algorithm was used in [43] to develop a method for EFS. The evaluation showed that the proposal reached to acquire robust and noise resilient subsets of features. A method based on the principle of selecting in Random Forest and co-forest was implemented in [44], the method allows to select features in datasets with unlabeled data.

4.3. Methodology

We followed the qualitative method described in [88] to propose a conceptual framework for our EFS. The method establishes a set of phases to design the framework as the development of a plan or network of concepts linked to describe a phenomenon. In this way, it describes a process to select a set of data sources, classify the data found, identify the key concepts related to them, and review and validate the proposal. The primary aim is to highlight the sense and importance of the relationships that associate with the concepts. Because of this, the concepts are not only considered as a collection, but also as a set of entities with a defined role. The methodology describes 8 phases; however, in this study only the first 7 phases were used because the last one corresponds to the reformulation of the framework, which is included in phase 7 for our study.

The implementation built in this study validated the conceptual proposal. Thus, the improvements and adjustments implemented as part of the development process in the phase 7 represented the rethinking of the conceptual framework. Figure 24 shows the phases mentioned.

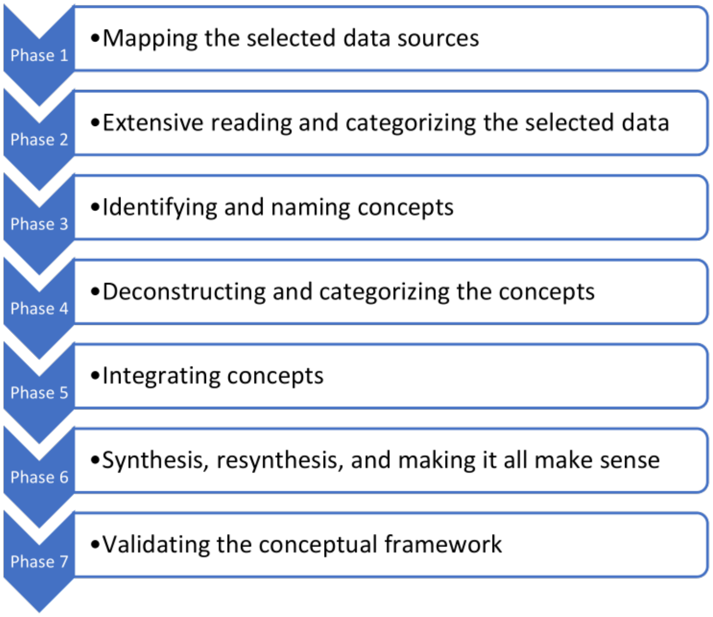

Phase 1
•Mapping the selected data sources
Phase 2
•Extensive reading and categorizing the selected data
Phase 3
•Identifying and naming concepts
Phase 4
•Deconstructing and categorizing the concepts
Phase 5
•Integrating concepts
Phase 6
•Synthesis, resynthesis, and making it all make sense
Phase 7
•Validating the conceptual framework

Figure 24. Methodology to Design Conceptual Framework

4.3.1. Phase 1: Mapping the Selected Data Sources

According to the methodology, in this phase the data sources and literature related to the phenomenon studied are mapped. Here, the mapping considered mainly the theory and research studies about Feature Selection and/or Ensemble Learning. Considering the above, the data source selected was the literature related to the following topics:

- Machine Learning
- Relevance Analysis
- Feature Selection
- Dimensionality Reduction
- Ensemble Learning
- Ensemble Feature Selection
- Consensus and aggregation

The review of literature was conducted searching on the bibliography databases, Science Direct, Scopus and IEEE, about selected topics. Papers with vital information in title, abstract or conclusions about the selected topics were studied in detail in next phases. Criteria defined to specify if a study is relevant considered relevant contributions on theoretical and practical proposals. For this phase, although the quantitative method does not describe how to conduct the process of mapping the selected sources, we recommend following guidelines for conducting a systematic review [89] or systematic mapping [90] in order to identify relevant works.

4.3.2. Phase 2: Extensive Reading and Categorizing the Selected Data

The selected literature was reviewed and analyzed to determine the main findings. According to the analysis carried out, the research studies were classified into the following categories:

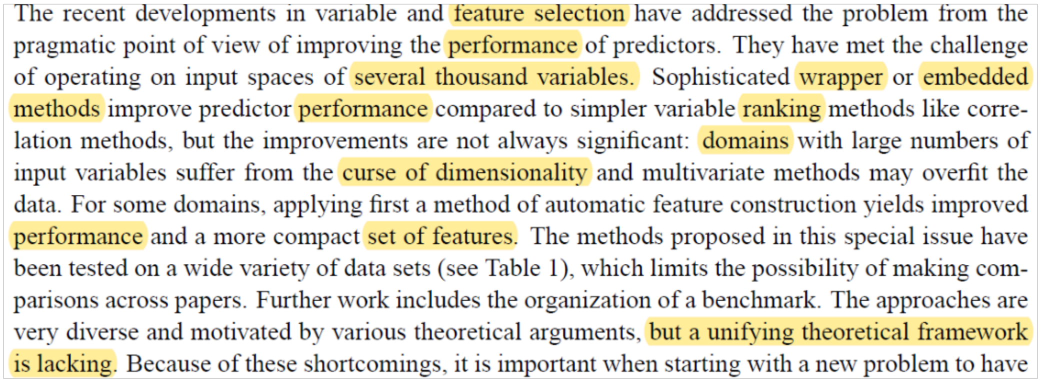
- Feature Selection
- Classification
- Ensemble Learning
- Consensus-Aggregation

Defining categories allowed to break the reviewed literature into specific topics, so papers connected between them were grouped together to analyze and understand in detail each topic. Likewise, in this phase we summarized the evidence by identifying main findings and relevant information in each topic.

4.3.3. Phase 3: Identifying and Naming Concepts

In this phase, the concepts are identified from the studied literature. The criteria included in the final concepts list considered the importance of describing the phenomenon and how these are related to each other.

Process followed to create the list of concepts include reading of all selected papers in the previous phase and highlight the concepts which are suspected to summarize a process, metric, method, measure or something relevant for the topic. After, all concepts highlighted were extracted and documented in a list according to the protocol followed in [91]. Figure 25 describes an example of highlighted concepts in the documents.



The recent developments in variable and feature selection have addressed the problem from the pragmatic point of view of improving the performance of predictors. They have met the challenge of operating on input spaces of several thousand variables. Sophisticated wrapper or embedded methods improve predictor performance compared to simpler variable ranking methods like correlation methods, but the improvements are not always significant: domains with large numbers of input variables suffer from the curse of dimensionality and multivariate methods may overfit the data. For some domains, applying first a method of automatic feature construction yields improved performance and a more compact set of features. The methods proposed in this special issue have been tested on a wide variety of data sets (see Table 1), which limits the possibility of making comparisons across papers. Further work includes the organization of a benchmark. The approaches are very diverse and motivated by various theoretical arguments, but a unifying theoretical framework is lacking. Because of these shortcomings, it is important when starting with a new problem to have

Figure 25. Process of Highlighting in Reading of Papers

All selected concepts were analyzed and filtered to define a final set of relevant concepts. The criteria used to filter were to keep in the list of concepts only those with short and representative name and those with the same meaning were represented in one.

4.3.4. Phase 4: Deconstructing and Categorizing the Concepts

The key attributes, features, assumptions, and roles were identified in this phase. General concepts were deconstructed into specialized concepts and the same procedure was conducted with the relationships or roles. The primary goal in this phase was to identify hierarchical structures of concepts, for example, according to Figure 26, the concept *feature* can be broken into redundant, relevant or noise (irrelevant).

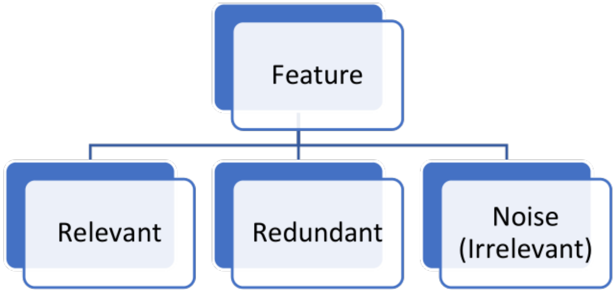


Figure 26. Example of a scheme of categorization.

4.3.5. Phase 5: Integrating Concepts

This phase describes how similar concepts were grouped and related using relationships. Such relationships are named to describe their meaning. The structures created in the previous phase helped to create and name the relationships to define roles. Figure 27 describes an example of a relationship created to associate concepts.

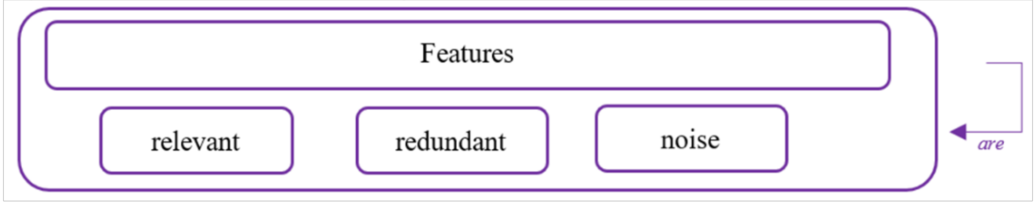


Figure 27. Example of relationship and role

4.3.6. Phase 6: Synthesis, Resynthesizing, and Making It All Make Sense

This phase must be considered as an iterative process to make a synthetization of the framework. The aim is to derive a real and consistent representation of the phenomenon. The Phase describes the design and implementation of a conceptual framework to support our EFS. The design of the conceptual framework considered a set of fundamental concepts to guide the representation of the proposal. Second, the conceptual framework was defined by associating the selected concepts with each other through relationships. Finally, an implementation framework is proposed to represent the conceptual framework in a software tool capable to make EFS.

The graphic representation of the conceptual framework is shown in Figure 28. This contains the concepts, groupings, roles and relationships.

4.3.6.1. Design of the Conceptual Framework

For Designing the Conceptual Framework, we organized all concepts, roles, relationships, and structures identified considering a basic idea of EFS: several FS algorithms generate several subsets of relevant features, and these must be aggregated to generate an unique subset. This idea was the main pillar for organizing the list of elements identified in the last phases. Thus, we described theoretically types of FS algorithms and the consensus process. Likewise, integrating the results of the previous phases is represented in Figure 28.

On the other hand, it is important to mention that a final graphic summarizes, in a visual representation, the content of thousands of papers and studies related to FS and EFS. In this way, we resume the theory around EFS in an easy and understandable view, directed to the designers and implementors of the EFS.

4.3.6.1.1. Main Concepts

The diverse types of FS algorithms are described below: filters, wrappers, embedded and the methods to aggregate experts' opinions.

- Filters: techniques, which are easy to implement, and which can be scaled to use datasets with high dimensionality. Nonetheless, this type ignores the interaction with a classifier. An $R(f)$ function evaluates the relevance of each feature and the output of the filter algorithm corresponds to a ranking that orders the features according to $R(f)$ [92].
- Wrappers: methods which evaluate the relevance of the subsets of features by using a classifier. Thereby, the best subset of features is selected by the learning algorithm. The computational cost of these techniques is high because when selecting the best subset, many subsets must be evaluated [68].
- Embedded: type of mechanism, which combines the advantages of the filters and wrappers. The principal objective is to get the best performance in the learning process from a learning algorithm using a subset of features [27].
- Consensus: In Ensemble Learning it is also called Consensus theory or Aggregation. Widely used in social sciences and administration, its principal objective is to find a way to combine expert opinions through consensus rules [93].

4.3.6.1.2. Conceptual Framework

According to the qualitative method described in [88], a revision of the literature was conducted to select the data sources and to understand the key concepts related to Feature Selection, Ensemble Learning and Consensus. Figure 28 shows the concepts and relationships identified from the data sources selected.
The literature about Relevance Analysis, Feature Selection and Dimensionality Reduction allows the identification of the three types of methods to determine the features with greater differentiation power. Thus, the types of algorithms were analyzed considering their design, aim and performance, which permitted the identification of the key concepts and relationships that describe the framework.

Considering the FS theory, the datasets contain three types of features: relevant, redundant and noise. In addition, some authors state that the relevant features have either strong or weak relevance [94]. Many features can be identified by filters, wrappers, or embedded methods. According to the method, the focus can be on the identification of features with low differentiation power, dependence of features, and relevant features. For instance, to identify features with low relevance, the algorithms analyze columns with low variance.

In general, the methods of FS use measures to evaluate the relevance of the features through statistical tests or cross validation. The results gained in these evaluations define the following: a ranking of feature relevance with the filter-based methods, a subset of relevant features with wrappers, or a subset of features with a learning model with the embedded methods. In the wrappers or embedded methods, the FS is based on the search for a subset of features by evaluating *n* subsets and selecting the one that achieves best performance in the classification.

The outputs of the FS methods are evaluated from several perspectives. Considering the problem to solve, some criteria can be more or less relevant. Because of this, the framework establishes the evaluation of algorithms by reviewing either their performance or their design. Regarding performance, efficiency and effectiveness are evaluated by testing the subset of features selected in a classification process. In terms of design, simplicity and scalability are evaluated by the designers of the algorithms.

In accordance with Ensemble Learning, the consensus of several experts improves the creation of a decision in a context [93]. Thereby, the conceptual framework considers the pooling of several FS algorithms through the consensus of a set of subsets of features selected by each method. This scheme is defined in [95] as a heterogeneous centralized ensemble, where *n* FS methods generate *n* models using the same data.

The principal objective of reaching a consensus among several FS methods is to generate a subset of relevant features capable of representing the advantages and disadvantages of all used methods and to face the biases of single methods.

The design built in Figure 28 describes the results of applying the qualitative analysis showed in Figure 24. The process describes the identification of a set of data sources to read and analyze them to define the concepts related to the topic. In this study, the concepts were selected by highlighting the relevant concepts that explain how to implement the Ensemble Feature Selection. According to phase 4 of the methodology followed, the final list of concepts was analyzed to categorize the concepts and identify their roles in the context of EFS. This phase defines the constraints and special considerations of an EFS method, which are described below:

- Instances must have values in all their columns.
- Instances must not have outliers.
- Values cannot be negative to avoid problems with statistical tests.

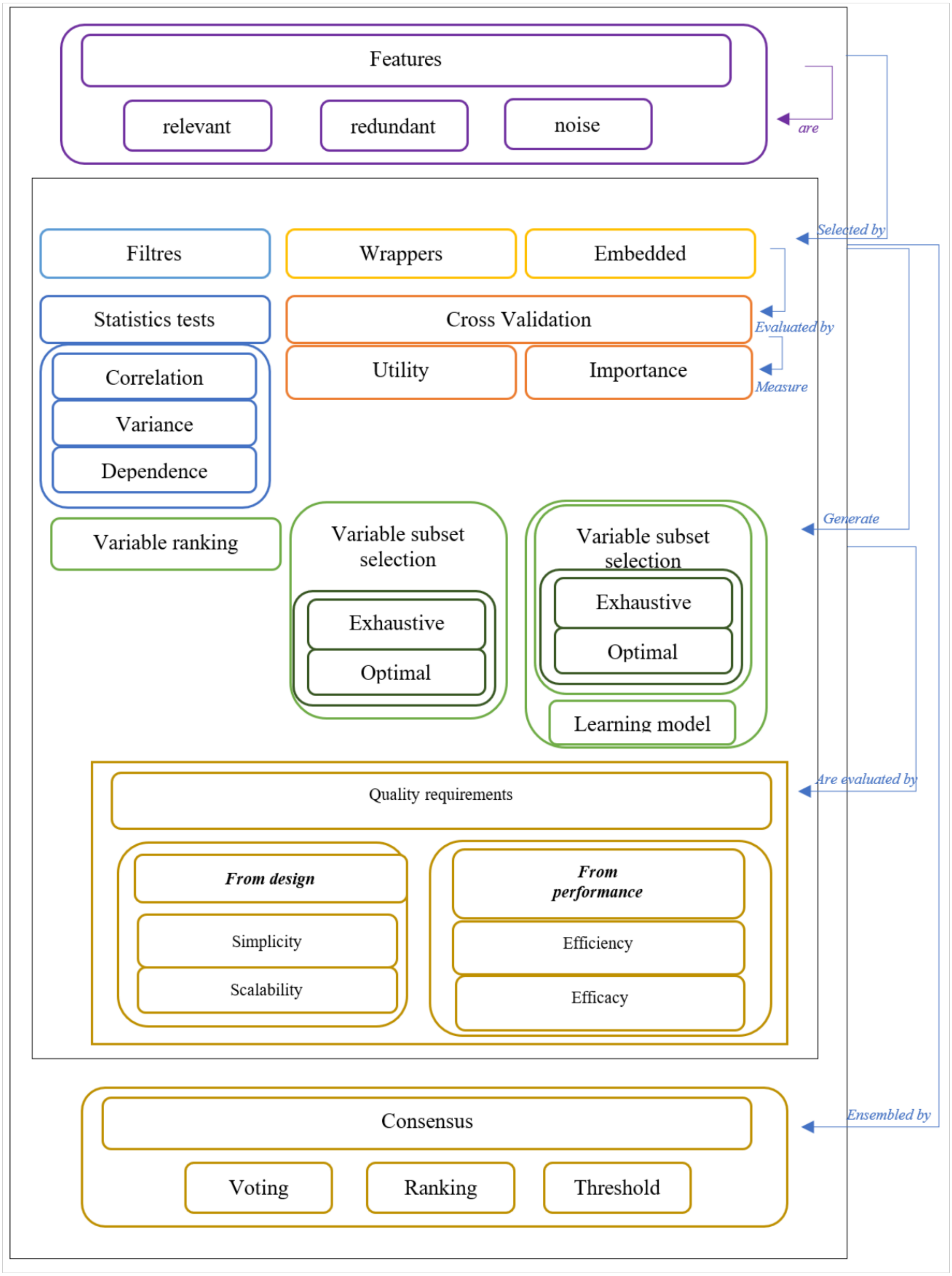


Figure 28. Conceptual framework

Considering the above, datasets must be preprocessed before applying EFS to handle their problems and avoid additional biases in the EFS process. In Data Mining and Machine Learning, these constraints are related to preprocessing and preparing data. One of the most famous methodologies to address Data Mining and Machine Learning projects is CRISP-DM [53]. The methodology breaks the process of Machine Learning into 6 major phases. The phase of data preparation describes a set of tasks to cover all the activities required to build the final dataset. One task of preparation data phase is cleaning data. This task is related to the detection of outliers, handling of missing values and fixing the data in a form suitable for the machine learning models. In this sense, the conceptual framework presented in this work only describes a theoretical explanation of Ensemble Feature Selection and the tasks related to phases of preparation data and feature engineering must be addressed in a previous step. Thus, the Ensemble Feature Selection process is conceived as a task of machine learning that needs a proper input dataset. Likewise, considering that real datasets have many problems and need to be preprocessed, the conceptual proposal represents these needs as restrictions that must be solved.

Concepts with a high relation or similarity were grouped or integrated according to phase 5. Phase 6 generated the graphic representation of the framework (Figure 28), which integrated the concepts and relationships previously identified. The representation was reviewed and analyzed by the authors to guarantee the correct representation of the theory extracted from the data sources. Phase 7 proposes the validation of the conceptual framework by discussing the proposed model with other researchers. However, the model not only aims to be a graphic representation of a research topic, but it also aims to validate that the EFS improves the FS process. Thus, its validation has been designed as implementing a tool to support EFS considering the concepts and relationships of the conceptual framework.

4.3.6.2. Implementation of the Conceptual Framework

The implementation of the conceptual framework was developed to validate the proposal described in Figure 28. For this, the Scikit-learn Machine Learning Library [96] was used to develop a tool that represents the framework. The solution selects features by different FS algorithms and then aggregates their outputs through consensus. The framework developed allows (i) to read, fix, and impute the values of datasets, (ii) to remove dataset features with high correlation, low variance or null values, (iii) to generate *n* subsets of relevant features using *n* FS algorithms, (iv) to aggregate the subsets generated using methods based on voting and (v) to evaluate the performance of the subset of features generated by our EFS.

Considering the above, Figure 29 describes the implementation of the framework in Figure 28. The solution groups the functions and methods in packages according to their objective.

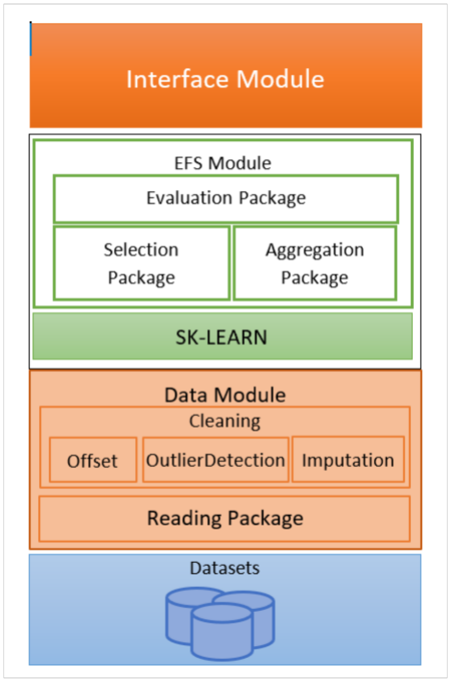


Figure 29. Developed framework

- Interface Module: It describes a class. This exposes the functionalities of the framework to new implementations.
- EFS Module: It is the core of the framework. This includes all the functionalities associated with the FS based on our ensemble method.
- Evaluation Package: This groups a set of functions to evaluate the accuracy and stability of an EFS output.
- Selection Package: It contains a set of methods to select subsets of relevant features.
- Aggregation Package: It integrates the outputs of n methods of FS using a criterion to aggregate the outputs.
- SCIKIT-LEARN: It is a Machine Learning library that supports the implementation of our EFS package.
- Data Module: It includes the functions to read and preprocess the datasets. This allows the reading of data from CVS files to adjust them according to assumptions and constraints considered in the design of the conceptual framework.
- Offset Package: It describes functionalities to calculate an offset dataset in order to avoid negative values.
- Imputation Package: It implements basic methods to handle missing values, a version of Multiple Imputation and some functions to evaluate the performance of the selected methods.
- Outliers Detection Package: This package considers a software component to implement methods of detection of outliers.

Handling missing values was identified as a critical task, because the result could change the distribution of data. However, there is no relevant study that describes the impact of missing values in FS process. Having regard to the above, an experiment was conducted to analyze evidence about the effect of missing values on datasets and FS process. The results of this experiment were calculated by using *Imputation Package* and are described in the Appendix D.

It is important to mention that the constraints defined by the conceptual proposal were solved in the implementation framework with the development of the offset, imputation, and outlier detection packages. However, with these packages, the framework does not support all tasks of data preparation or feature engineering (extraction and transformation of data). The goal of the packages is only to support common problems associated with the constraints of the FS algorithms.

4.3.7. Phase 7: Validating the Conceptual Framework

In this phase, the conceptual framework must be validated through expert judgment. However, in order to find a quantitative evaluation, a framework of EFS was implemented following the proposal of the conceptual framework and the implementation of the framework was evaluated by testing its performance in the Feature Selection on popular datasets.

4.3.7.1. Evaluation of the Framework

The evaluation used three public datasets, which are available on UCI Machine Learning Repository [82]: *Sonar*, *SPECTF* and *WDBC*. These datasets were used to test the EFS algorithm developed in [97]. The results of the evaluation showed the accuracy and stability of the method.

Performance

For the evaluation, the classifiers *Decision Tree Classifier* and *Logistic Regression* used the subsets of features generated by each FS algorithm and the subset generated by EFS algorithm. Table 11 shows the number of features selected by each FS method.

Table 11. Number of features selected by each FS method.

	Sonar	**SPECTF**	**WDBC**
SelectKBest	5	5	4
RFE	3	3	3
Feature Importance	5	5	4
EFS	10	10	10

To implement the aggregation in the EFS method, the sum of the subsets generated by the ***n*** FS algorithms is calculated. For each feature in the subset SUM, an importance index is computed according to equation (1). Where the importance of feature *i* is determined by the number of times it is present in the subset SUM divided by ***n***. Finally, the features that exceed a threshold will be selected in the last set.

$$IF_i = \frac{FF_i}{n} \tag{1}$$

Table 12 compares the results obtained in this study regarding the results found in [97] of the feature selection in the datasets Sonar, SPECTF and WDBC. Column 2 shows the number of features of each dataset; columns 3 and 4 show the number of features selected by each method and columns 5 and 6 show the percentages of elimination of features obtained in each method. According to the above, the EFS developed in this study was able to select for the three datasets subsets with equal or smaller size than the solution proposed in [97]. In this sense, the percentages of elimination of features are equal or higher. In order to difference the results of EFS of [97] and the results of our proposal, we named our framework: F-EFS (Framework of Ensemble Feature Selection).

Table 12. Comparison of the EFS Method Constructed and the Results Reported in [97]

	Features	**Features selected by F-EFS**	**Features selected in** [97]	**% of elimination of features by F-EFS**	**% of eliminatio n of features by** [97]
Sonar	60	10	24	83.3%	40%
SPECTF	44	10	19	56.7%	43.2%
WDBC	30	10	10	66.7%	66.7%

Table 13 and Table 14 show the accuracy found by both classifiers *Logistic Regression* and *Decision Tree Classifier* when used with each dataset and the subsets generated by FS algorithms and the ensemble method developed.

Table 13. Classification Results Using Logistics Regression

	Sonar	**SPECTF**	**WDBC**
SelectKBest	84.05%	53.22%	90.35%
RFE	85.50%	53.22%	93.65%
Feature Importance	84.05%	59.67%	92.10%
F-EFS	**86.95%**	60.75%	**93.85%**

Table 14. Classification Results Using Decision Trees

	Sonar	**SPECTF**	**WDBC**
SelectKBest	73.91%	63.44%	89.47%
RFE	65.21%	68.81%	87.71%
Feature Importance	78.26%	72.58%	89.47%
F-EFS	73.91%	**74.73%**	92.10%

Subsets of Relevant Features

To facilitate the analysis of the results in the evaluation, the features contained in each dataset were named *Fi* and the target column was named *class*.

Figure 30 shows the subset of features selected by each algorithm on the Sonar dataset. In this test, features F12, F36, and F45 are considered relevant by at least two selection algorithms, feature F11 is selected by the three FS algorithms and the other features are selected by only one algorithm. Thus, if the selection threshold defined by the user is 0, the set of features selected is: {F9, F10, F11, F12, F21, F35, F36, F45, F46, F49}.

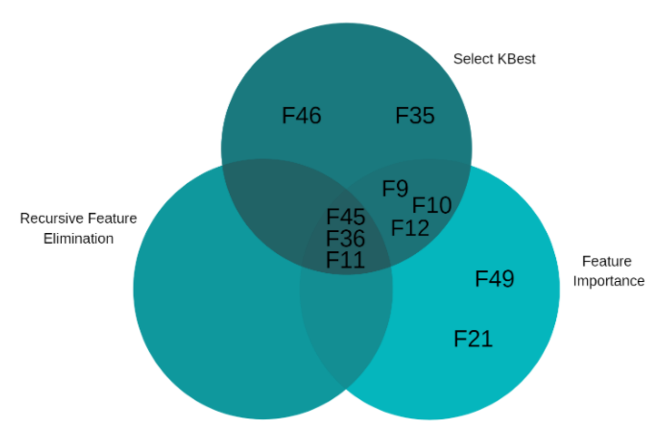


Figure 30. Sonar

Figure 31 shows the subset of features selected by each algorithm on the SPECTF dataset. The Venn diagram shows that the features F25, F26 and F40 are selected by over one algorithm, while the other features were considered relevant by only one algorithm. Thus, assuming that the selection threshold defined by the user is 0, the set of features selected is {F4, F25, F26, F28, F30, F36, F40, F42, F43, F44}.

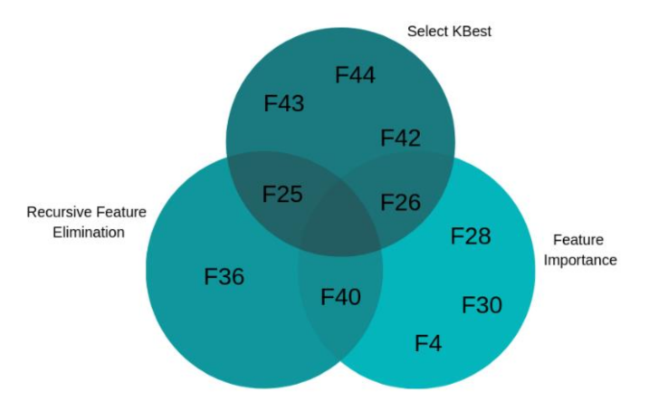


Figure 31. SPECTF

Figure 32 shows the subsets of features selected by each algorithm for the dataset WDBC. In this result, features F23 and F24 were selected by the algorithms Select K Best and Feature Importance, feature F21 by the RFE and the Feature Importance algorithms, while the rest of the features were selected by only one algorithm. In this sense, assuming that the threshold defined by the user is 0, to get all features selected by all FS algorithms, the set of features selected is {F3, F4, F7, F8, F14, F21, F23, F24, F27, F28}.

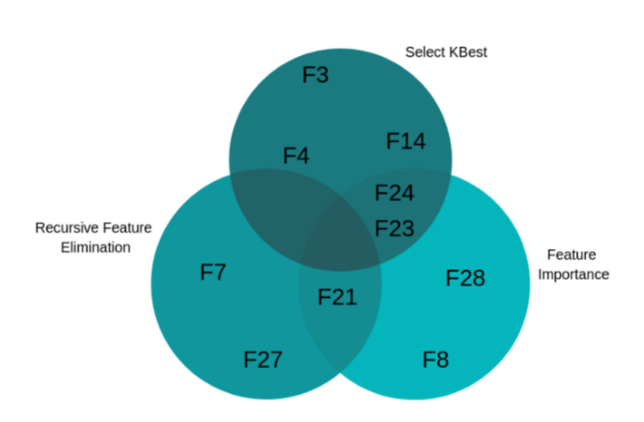


Figure 32. WDBC

The Venn diagrams described in Figures 31, 31, and 32 show that some features are relevant by over one algorithm, while others are only present in one. Taking the above into account, the importance of each feature defined in Formula 1 could also be used as a mechanism based on weight in order to give greater relevance in a classification process to the features that are detected with a high differentiation power by several FS algorithms.

Table 15 shows the results of the stability evaluation of the feature sets generated using our ensemble method implemented in the framework. The results show that the method used achieved perfect stability for the three datasets used in the evaluation. As future work, we propose to improve the framework in order to support the parameterization of new implementations of feature selection methods, increase the number of consensus methods implemented, offer new measures to evaluate the sets of features selected and support the ensemble of FS methods not only in a heterogeneous scheme but also in a homogeneous one.

Stability

In solutions based on Ensemble Learning, it is important to ensure that the outputs of these methods return similar outputs, even if the training data change. This property is known as stability and, according to different studies, there are different measures to evaluate it. The Jaccard index [98] is one of the most famous measures for assessing stability in methods that generate subsets of characteristics. The index is described by equation (2).

$$Jac\ (A, B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

To evaluate stability, for each dataset, the ensemble method developed was executed 10 times using 10 random samples taken from the original dataset. The results showed that in the three datasets used, the set generated by the F-EFS was the same in the 10 iterations.

The results are shown in Table 5.

Table 15. Results of Stability

	Features	Stability
Sonar	F9, F10, F11, F12, F21, F35, F36, F45, F46, F49	1
SPECTF	F4, F25, F26, F28, F30, F36, F40, F42, F43, F44	1
WDBC	F3, F4, F7, F8, F14, F21, F23, F24, F27, F28	1

4.3.7.2. Analysis of Results

In this study, we proposed the design of a conceptual framework to support Ensemble Feature Selection. Considering a set of concepts and relationships, our proposal explains the general behavior of FS algorithms, their techniques and how to improve the performance in classification processes. Additionally, the framework not only summarizes techniques, but it also breaks them down in order to facilitate their understanding and show how we can combine them to compensate for their biases. This allows us to combine outputs of single FS methods and aggregate their outputs by the consensus.

In previous studies, the authors have proposed solutions based on single FS algorithms, which are designed to focus on a special domain and problem, for instance, removing features with low variability or identifying relationships among features [99]. Additional studies have used several FS algorithms to generate a subset of features by each method. These subsets have been tested using classifiers and the subset with best performance in the classification is the selected subset [100][101][102][103]. In the above approach, although they proved different FS algorithms, the final subset of features is influenced by the biases of the algorithm employed to generate it, which is the main difference regarding this study, where the advantages of different methods of FS are considered in the final subset to compensate their biases.

In addition, recent studies have developed tools to support the combination or assembly of FS algorithms. In [104] the authors proposed a tool developed in R. The tool was designed to combine several outputs of FS algorithms into a Random Forest algorithm and was used in [97] to select relevant features in the datasets: Sonar, SPECTF, and WDBC. The results in Table 12 showed that our solution achieved an equal or better performance in dimension reduction. The comparison only was conducted on dimension reduction due to the experiments of classification made in our evaluation were carried out using random datasets extracted from the original datasets and the results showed in [97] were not calculated in the same conditions. Considering this, the classification showed in [97] could outperform the classification showed in our evaluation. However, the goal of the framework is defining a general approach to support EFS independently of the classification process.

Likewise, in [105] the author describes a tool developed in Python and available on GitHub. The solution was implemented as a class under the Object-oriented programming paradigm and its methods implement some of the best-known FS algorithms. Each method receives a set of parameters to configure the FS algorithm. The methods can be used as a single FS method or they can use a special function to get the subsets selected by several FS algorithms. However, this solution does not provide a mechanism to aggregate all subsets generated.

Considering the studies reviewed, and the solution developed, the framework proposed provides an overall scheme to support EFS in order to facilitate the analysis of techniques, biases, disadvantages and advantages of the FS algorithms to determine how we can assemble different techniques to get a subset of relevant features more efficiently. In this sense, the framework is designed for data scientists that handle problems related to high dimensionality and/or few instances.

According to the results in Table 13 and Table 14, the best classification results were obtained for each dataset using the method developed in this study. For the datasets *Sonar* and *WDBC* Table 13, the best performance was achieved using the subsets generated by our method with a Logistic Regression classifier. The evaluation showed that the accuracy was 86.95% and 93.85%, respectively. For the dataset *SPECTF,* the best performance was achieved using our ensemble method with a Decision Tree Classifier with a 74.73% accuracy.

4.4. Conclusions

This chapter described the proposal of a conceptual framework designed to understand with clarity the most relevant concepts in the Feature Selection process and determine how to improve the performance of the classification algorithms through the set of features used during their training process.
The qualitative method followed to build the conceptual framework allowed, through an exploration of literature, the identification of concepts and relationships that describe the process of FS and the consensus among different FS techniques.

The conceptual framework built guided the development of an implementation framework capable of selecting features using an ensemble of FS methods. The method achieved the selection of a set of relevant features with higher performance in the classification regarding the sets of features selected by the single algorithms.

The evaluation allowed to validate that performing a classification process can be improved considering an algorithm of FS based on ensemble instead of single FS methods. Likewise, the performance of our ensemble method achieved 100% of stability for the datasets used in the evaluation.

The major contribution of this work to the field of Machine Learning is the definition of a structure that provides an understanding of how to improve the performance of FS based on the consensus of several techniques. This could guarantee better performance in classification algorithms and increase the reliability in those fields of application in which the reliability of the results must be high.

Chapter 5:

Evaluation of the EFS Framework on the Selection of Relevant Features on EEG Signals

5.1. Introduction

This chapter describes the evaluation of the framework built to support the Ensemble of Feature Selection algorithms. The evaluation considered the performance of the framework to determine the subset of relevant features in the context of classification of normal and abnormal segments of pediatric EEGs. Thus, the dataset used to evaluate the ensemble feature selection corresponds to the result of applying the 16 feature extractors described in chapter 3 on (i) a set of segments extracted from EEGs records and (ii) the resolutions and approximations generated from the set of segments with the Discrete Wavelet Transform.

The evaluation dataset contains a total of 142 features and 1,344 segments of EEG signals recorded from pediatric patients. 672 instances of the dataset correspond to abnormal segments and 672 to normal segments. Thereby, the feature selection process applied to the evaluation dataset identified 27 relevant features from the original set of features. The 27 features were evaluated with different classification algorithms, the best model to classify the segments was an SVM, which achieved an accuracy of 97.42%. Besides, the sensitivity and specificity achieved were 97.32% 96.72% respectively.

Additionally, to evaluate the reliability of the subset of relevant features generated by the ensemble method, we used the evaluation module of the framework to measure the stability of the subset of relevant features generated. The stability obtained after selecting the subset of relevant features using 10 different random samples from the original dataset, allowed to establish a stability of 100%

5.2. Dataset

A set of 1344 segments labeled by a pediatric neurologist were extracted from the 200 electroencephalograms stored in this project. As explained in Chapter 3, the analysis of an electroencephalogram should be considered as the analysis of a non-stationary multichannel signal, which implies breaking down the EEG channel by channel and, in turn, segmenting each channel into small periods of time, in such a way that the generated segments can be considered stationary signals. This is because the wave that represents each channel of an EEG presents changes in its statistical characteristics over time. However, small segments of the signal retain a minimum variation of these characteristics, therefore, they can be considered stationary signals and be characterized through feature extractors such as the 16 that were described in chapter 3.

To build the dataset, a set of normal (672) and abnormal (672) EEG segments with a duration of one second were extracted from the electroencephalograms. For each segment, the Discrete Wavelet Transform was applied in three levels of decomposition. Thus, the segments and their approximations and resolutions were used to get a feature vector by using the 22 descriptors mentioned above, obtaining a set of 142 features.

The process followed to build the dataset is described in Figure 33.

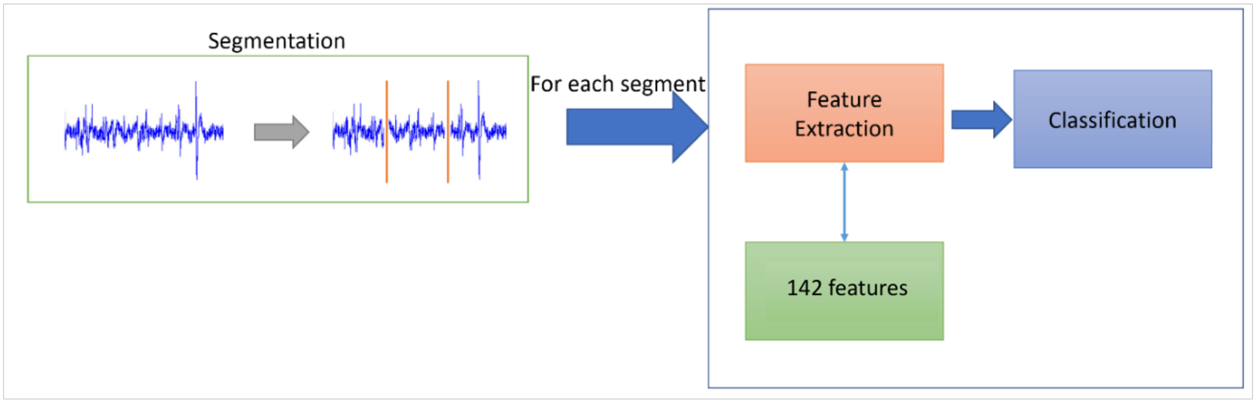


Figure 33. Channel decomposition process.

5.3. Selecting Relevant Features

The evaluation was divided into three stages. The first one compared the performance of several classification algorithms using all features and subsets of relevant features selected by F-EFS. The second stage evaluated the classification algorithm and the subset of relevant features that reached the best performance in the previous stage by applying n-fold Cross Validation. Finally, the stability of the subset of relevant features was calculated.

5.3.1. Evaluating F-EFS

To evaluate the utility of subset features selected by F-EFS, a set of 4 classifiers (Decision Tree, Logistic Regression, Random Forest, and SVM) were configured to determine which one of them gets the best performance for classifying normal or abnormal brain activity. The evaluation considered 70% of data for training the models and 30% for testing them. Table 16, Table *17*, Table *18*, and Table *19* describe the results of the accuracy and standard deviation of accuracy in classification calculated using all features and the subset of selected features by F-EFS. *K* represents the number of features selected by each single FS algorithm used in F-EFS.

Table 16. Results of Accuracy – Decision Tree Classifier.

	Features Selected	**Decision Tree Classifier**	
K		All	EFS
1	3	95.97 +- 1.4	92.03+- 1.8
3	10	95.91+-1.5	93.82+-1.5
5	17	95.71+-1.6	94.48+-2.1
7	23	95.78+-1.4	95.08+-1.9
9	27	95.90+-1.01	96.02+-1.03
11	35	95.97+-1.21	96.12+-1.01

Table 17. Results of Accuracy – Logistic Regression.

	Features Selected	**Logistic Regression**	
K		All	EFS
1	3	97.17 +-0.98	90.31+-1.8
3	10	97.37+-1.1	91.89+-3.1
5	17	97.31+-1.6	92.27+-3.0
7	23	97.39+-1.2	95.16+-2.1
9	27	97.02+-1.4	95.08+-1.6
11	35	97.24+-1.32	95.28+-1.54

Table 18. Results of Accuracy – Random Forest.

	Features Selected	**Random Forest**	
K		All	EFS
1	3	89.17+-2.2	87.84+-2.03
3	10	89.2+-2.76	88.17+-3.13
5	17	89.32+-2.6	88.79+-2.6
7	23	89.24+-2.7	89.24+-2.26
9	27	89.31+-2.8	89.72+-2.6
11	35	89.45+-2.12	89.81+-2.57

Table 19. Results of Accuracy – SVM.

	Features Selected	**SVM**	
K		All	EFS
1	3	87.69+-1.17	94.63+-1.5
3	10	87.54+-1.15	94.93+-1.4
5	17	87.61+-1.2	95.97+-1.2
7	23	87.69+-1.07	**96.61+-1.07**
9	27	87.29+-1.2	**96.79+-1.05**
11	35	87.06+-1.2	**97.31+-1.01**

Results in Table 16, Table 17, Table 18, and Table 19 proved that subsets of selected features could reach a similar performance in classification to the performance achieved using all features. Even accuracy of SVM improved when the classification process used only subsets of features selected by F-EFS. Likewise, the previous tables show Support Vector Machine was the algorithm with the best performance in the classification of abnormal and normal segments of brain activity. For this preliminary test, the dataset was divided into training and test data, 70% of data was used to train and the 30% remaining was used to test de models.

This evaluation was a preliminary study, and it did not represent the evaluation of a final detector of epileptiform events. The only goal was to show that using the features selected by EFS it is possible to achieve a performance equal to or greater than the performance achieved by using all features.

5.3.2. Selecting Relevant Features from the EEG dataset.

This phase used the implementation framework F-EFS described in the previous chapter to analyze the relevance of features on a dataset with descriptions of normal and abnormal segments extracted from EEGs. To validate the subset of relevant features calculated by using the F-EFS, a classification process was built to evaluate the accuracy reached with the subset of features selected. For this, we found the configuration of F-EFS that selects the best subset of features for the EEG dataset.

Algorithm 1 describes the implementation used to evaluate the relevant features on the EEG dataset and the review of performing them by using 4 popular algorithms of classification. The algorithm was coded in python using algorithms from the sci-kit learn library and the EFS Class built to represent the implementation framework F-EFS.

Algorithm 1. Implementation using F-EFS to Select relevant features

```
    obj_efs = EFS()
    data = readData(eeg_dataset.csv])
    goal = dataframe['class']
    X_train, X_test, y_train, y_test = split_data (data)
    names = data.columns
        k_best = 9
        k_importance = 9
    threshold = 0.3
    obj_efs.ensamble_feature_selection(names, data, dataframe, goal, k_best,
          k_importance)
    faetures_selected = obj_efs.select_by_weight_frequency (names,threshold)
    obj_efs. faetures_selected = names_selected
    if (len(faetures_selected) > 0 ):
      print ("Selected features are: ", faetures_selected)
      print('############ DecisionTreeClassifier #########')
      obj_efs.evaluate_accuaricy_by_model (tree, X_train, X_test, y_train,
           y_test, names)
        print('############ LogisticRegression #########')
        obj_efs.evaluate_accuaricy_by_model (model, X_train, X_test, y_train,
            y_test, names)
        print('############ RandomForestClassifier #########')
        obj_efs.evaluate_accuaricy_by_model (RF_predictor, X_train, X_test,
            y_train, y_test, names)
        print('############ SVM SVM SVM SVM SVM #########')
        obj_efs.evaluate_accuaricy_by_model (svcclassifier, X_train, X_test,
             y_train, y_test, names)

    else:
        print ("none feature was selected")
```

To determine the best subset of features, the Algorithm 1 was executed by using different hyper-parameters. Hence, the best results of classification were reached using parameters described in the code. The subset of relevant features calculated by F-EFS contained 27 features: 'F1', 'F11', 'F15', 'F28', 'F32', 'F36', 'F40', 'F53', 'F55', 'F60', 'F65', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106', 'F118', 'F126', 'F132'. This subset was calculated, including the features with an importance index greater than the threshold.

It is important to mention that if the threshold defined by the user was 0, the subset of relevant features would include the union of the subsets generated by each FS algorithm. Here, the final subset of relevant features would contain 39 features: 'F1', 'F11', 'F15', 'F18', 'F19', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F64', 'F65', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106', 'F107', 'F114', 'F116', 'F118', 'F120', 'F126', 'F132', 'F140', 'F141'.

The features selected by each single FS algorithm are shown in Table 20.

Table 20. Subsets of features selected by single algorithms.

Algorithm	Subset
SelectKBest1	'F11', 'F15', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F57', 'F61', 'F70', 'F74', 'F78', 'F95', 'F99'
SelectKBest2	'F1', 'F55', 'F60', 'F65', 'F71', 'F76', 'F77', 'F85', 'F86', 'F92', 'F97', 'F106', 'F118', 'F126', 'F132'
SelectKbest3	'F1', 'F55', 'F60', 'F65', 'F71', 'F76', 'F77', 'F85', 'F86', 'F92', 'F97', 'F106', 'F118', 'F126', 'F132'
RFE	'F11', 'F15', 'F28', 'F32', 'F36', 'F53', 'F70', 'F74', 'F78', 'F95', 'F99', 'F116', 'F120', 'F140', 'F141'
Feature Importance	['F97', 'F76', 'F92', 'F118', 'F85', 'F114', 'F126', 'F77', 'F107', 'F119', 'F106', 'F18', 'F86', 'F65', 'F64'

To facilitate the naming of the features, we named them Fx, where x is a consecutive number that describes the order in which the 16 descriptors were applied on the segment, approximation, or resolution. Table 21 describes the name and type of each feature selected by F-EFS.

Table 21. Selected features

Features	Domain
Variance	Time
Energy	Time
Entropy Shannon	Time
Activity A5	Time
Variance A5	Time-Frequency
Energy A5	Time-Frequency
Entropy Shannon A5	Time-Frequency
Activity A5	Time-Frequency
Variance D4	Time-Frequency
Energy D4	Time-Frequency
Entropy Shannon D4	Time-Frequency
Activity D4	Time-Frequency
Variance D3	Time-Frequency
Energy D3	Time-Frequency
Entropy Rényi D3	Time-Frequency
Entropy Shannon D3	Time-Frequency
Activity D3	Time-Frequency
Max D2	Time-Frequency
Variance D2	Time-Frequency
Standard deviation D2	Time-Frequency
Kurtosis D2	Time-Frequency
Energy D2	Time-Frequency
Entropy Rényi D2	Time-Frequency
Entropy Shannon D2	Time-Frequency
Kurtosis D1	Time-Frequency
Energy D1	Time-Frequency
Entropy Rényi D1	Time-Frequency

Where x represents the level of decomposition of the DWT and Ax or Dx represents the approximation or resolution of the level x.

Algorithm 1 used the setting established by default for FS algorithms implemented into the EFS method, except for the values that describe the number of features selected from each FS algorithm. The EFS method used 5 FS algorithms, 3 based on rankings of features, 1 wrapper (importance of features calculated by Decision Trees), and 1 embedded (Recursive Feature Elimination).

Each single FS algorithm generated a subset of relevant features, which were aggregated by the method select_by_weight_frequency in the F-EFS. The setting of the aggregation method returned the features aggregated from the subsets of relevant features generates by each single FS algorithm with an importance index greater than the threshold defined by the user. The values used in the final evaluation were calculated experimentally following the trial-and-error approach. Thus, the features were selected by the F-EFS method if these reached an importance index greater than or equal to the threshold (See Figure 34).

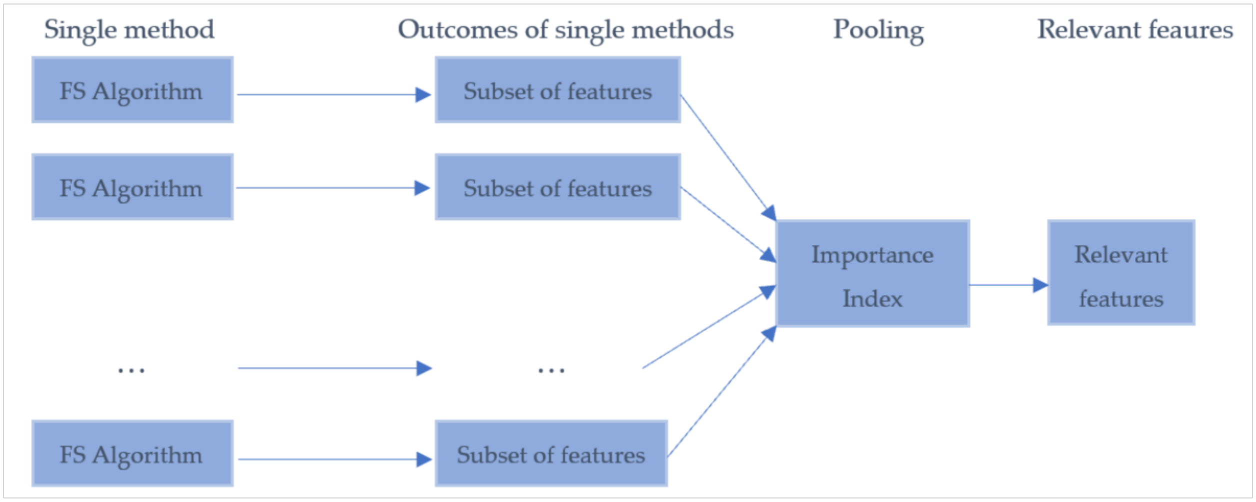


Figure 34. Process of ensemble feature selection.

Table 22 shows the results in the classification of a Decision Tree (DT) algorithm, LR (Linear Regression) algorithm, RF (Random Forest) algorithm, and Support Vector Machine algorithm using all features, and the features calculated by the Select K Best algorithm, Recursive Feature Elimination algorithm, Feature Importance algorithm, and F-EFS method. The subsets of features were calculated according to the configuration described for F-EFS in Algorithm 1. In this experiment, the Select K Best used the Chi-squared metric, which obtained a subset of relevant features better than the subsets generated by ANOVA and mutual information metrics. The comparison was based on the accuracy achieved by each subset of relevant features generated by each metric. We considered 70% of the data for training the models and 30% for testing them for this evaluation.

Table 22.Results of accuracy using different subsets of features.

Algorithm	DT	LR	RF	SVM
SelectKBest	92.79%	94.59%	89.39%	93.43%
RFE	93.01%	85.09%	89.87%	93.43%
Feature Importance	94.56%	94.36%	89.64%	94.18%
All Features	92.89%	95.17%	89.62%	96.75%
EFS	**96.05%**	**95.94%**	**89.79**	**97.46%**

The results showed in Table 22 evidence that the configuration defined for F-EFS allowed identifying the best subset of relevant features used to classify normal and abnormal brain activity.

5.3.2.1. N-fold: Cross-Validation

Cross Validation is an analysis tool that allows the evaluation of the results offered by a model. In this evaluation, this mechanism was used to divide the dataset into the smallest sets to train and evaluate an SVM Classifier. The single-step divides a sample into test and training data. Application of single Cross-Validation consists of classifying the test data with the SVM implemented and trained with training data. However, N-Fold Cross-Validation implied breaking the original dataset into n samples, and for each sample, it tested and trained the sub-samples. Averaging accuracies calculated for all samples allowed us to determine a general accuracy statistically.

Figure *35* describes a general scheme of N-Folds Cross-Validation. The scheme shows how the mechanism divides the sample data into n partitions and performs the traditional Cross Validation process n times, using in each iteration a different partition as a test dataset and the remaining n-1 partitions as a training dataset.

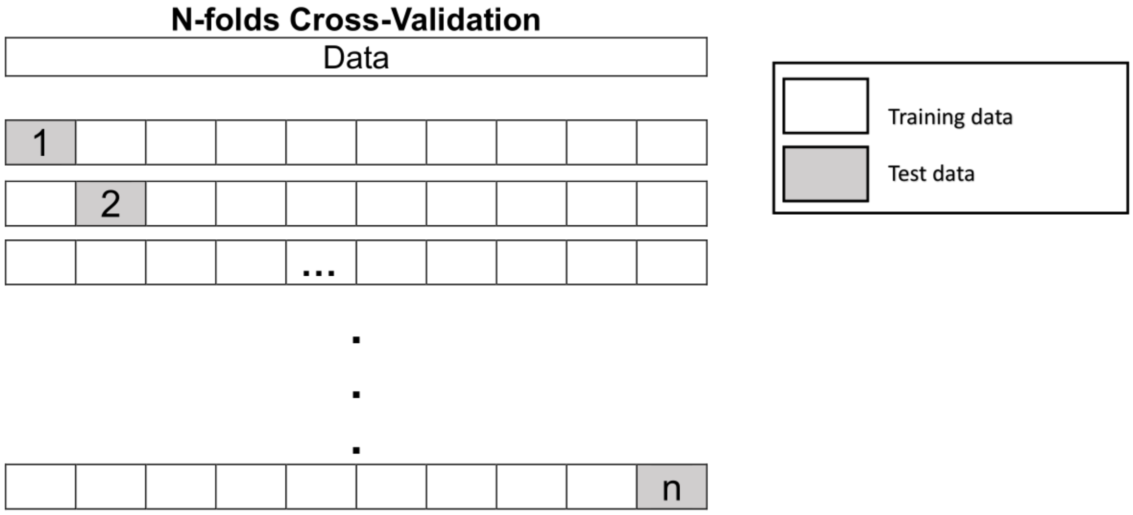


Figure 35. Scheme of n-fold cross-validation

The results of the n-folds cross validation applied for each evaluated k, can be seen in the following table. The value k represents the number of similar cases that are considered constructing the solution of a recent case.

Table 23. Results of N-fold cross-validation.

N	Accuracy (%)
1	97.39
3	97.38 +- 1.100
5	97.45 +- 1.210
7	97.46 +- 1.082
10	97.46 +- 1.080

The value of K in Table 23 corresponds to the value used to determine the number of samples generated in the n-folds validation.

Figure 36 Figure 5 describes the confusion matrix calculated for this evaluation for N=10. The results show that the classifier SVM achieved a rate of true positives of 96.43% and a true negatives rate of 97.96%. Besides, the sensitivity was 96.78.%, and the specificity 97.95%.

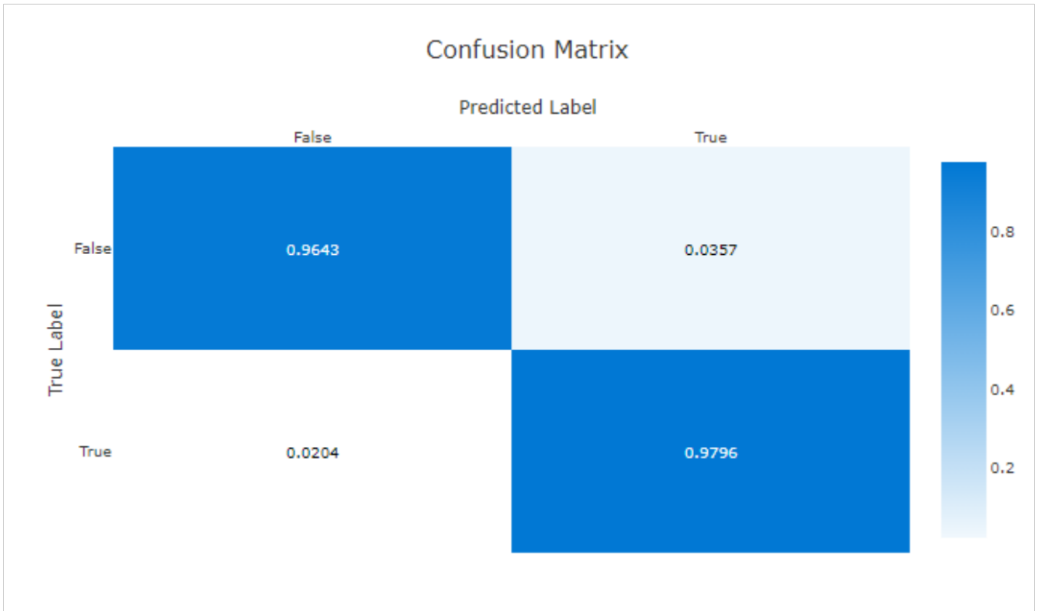


Figure 36. Confusion Matrix

Figure 37 describes the ROC curve with the results for this evaluation.

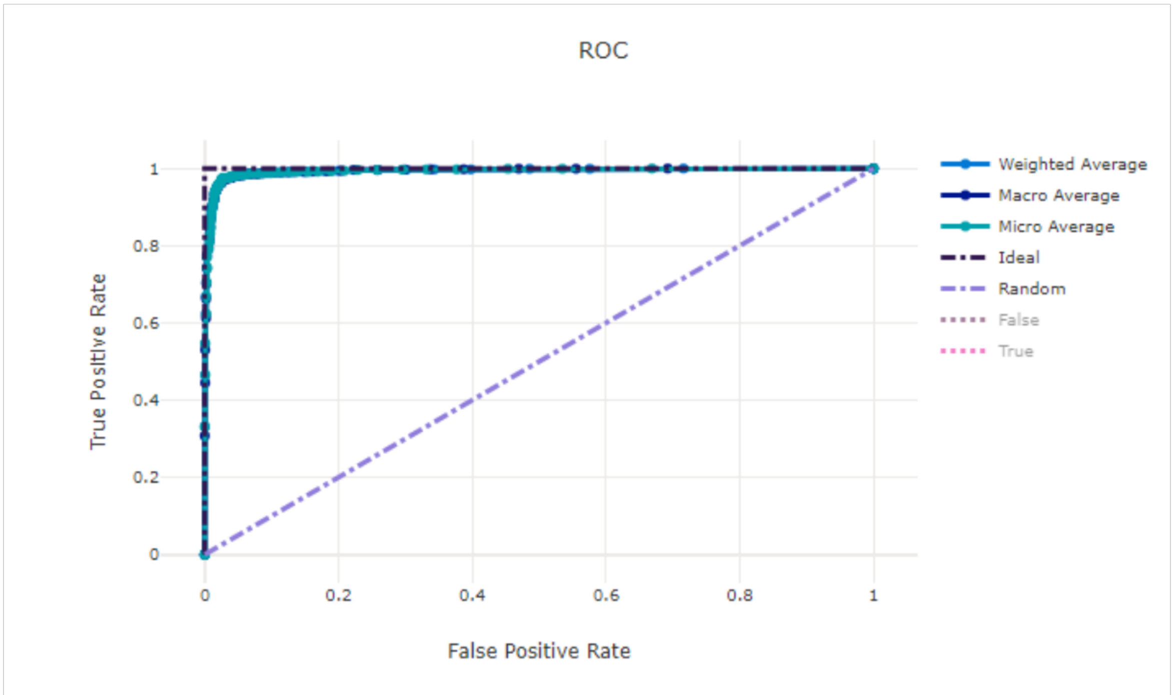


Figure 37. Receiver operating characteristic curve - ROC Curve

Figure 38 describes the graph with the results of the precision and recall.

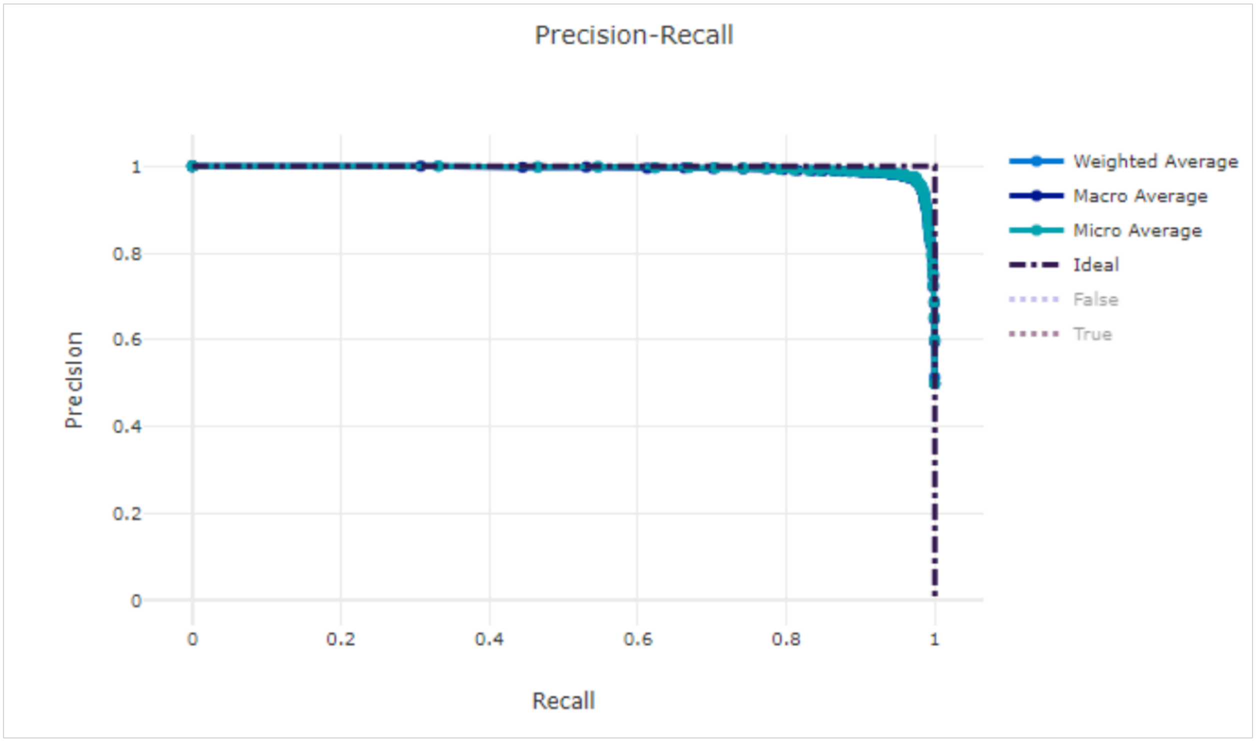


Figure 38. Precision-Recall

5.3.2.2. The Detector of Epileptic Activity

The SVM model built in the previous step was included as part of a detector of epileptic events to support the automatic reading of EEGs. In this approach, the detector decomposes an EEG signal into channels and segments each channel into short periods of time that are classified as normal or abnormal using the SVM model. Figure 39 shows the scheme of reading of the detector.

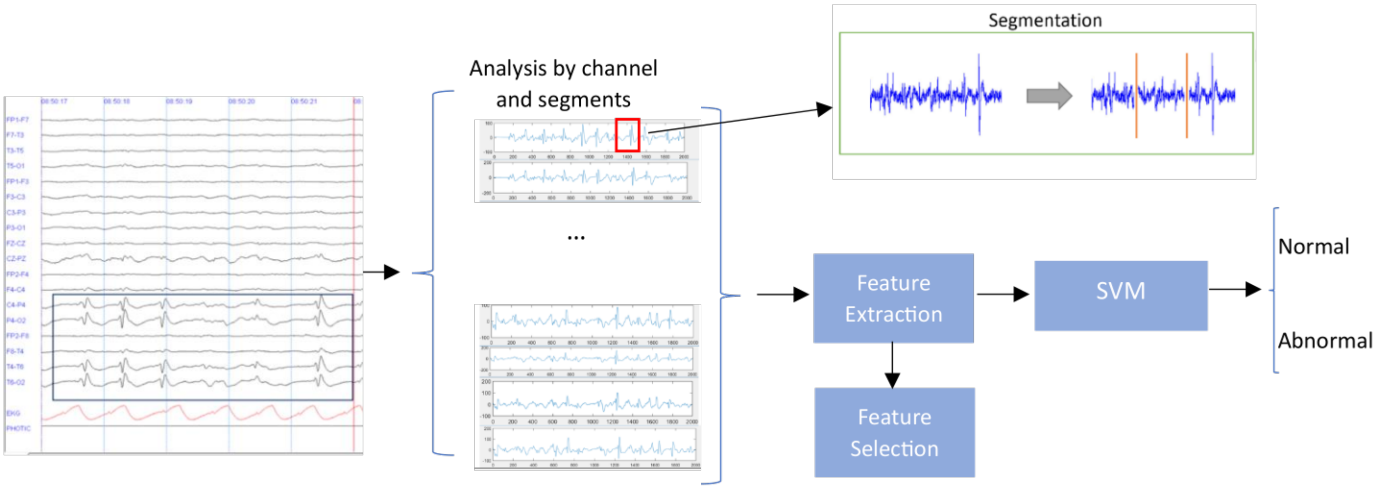


Figure 39. Detector of epileptic events.

The detector has been developed to show the field of the application of the classification of EEG signals. One of the principal reasons that motivated this research was to help in the diagnosis of epilepsy by supporting the automatic detection of epileptic events in EEG signals. To achieve this, we proposed to improve the classification process by including in the learning process only the relevant features that describe an EEG signal.

Validation of the detector

To validate the detector, a set of 100 EEG records taken from 100 pediatric patients were read by the detector. The 100 EEG records are part of the EEG repository built in this research and they were diagnosed by a neuro pediatrician with 20 years of experience in the reading of this kind of exam. For the test, each EEG record with epileptic activity had a description of the beginnings and ends of the epileptic abnormalities. These descriptions were used to validate the detections made by the detector.

Table 24 describes the results of the reading of the 100 EEGs.

Table 24. Confusion Matrix of the detector.

		Predicted	
		Abnormal	Normal
True	Abnormal	6520	286
	Normal	39602	487501

According to the confusion matrix, the detector's accuracy, sensitivity, specificity, NPV, and PPV were 92.53%, 95.57%, 92.48%, 92.49%, and 95.80%. Also, the rate of false negatives was 4.20%, and the rate of false positives was 7.51%.

5.3.3. Stability

To determine the reliability of the implemented EFS method, the subset of features generated to support the classification of epileptiform events was evaluated. The EFS method was used 10 times to generate 10 subsets of relevant features with 10 different random samples from the dataset. The 10 subsets generated were compared according to the Jackar Index to determine the difference between them. However, in the 10 executions, the algorithm selected the same subset of relevant features. The results can be observed in the following table.

Table 25. Results of selecting features on sub-samples

Subset	Selected Features	Stability
1	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
2	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
3	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
4	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
5	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
6	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
7	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
8	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
9	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1
10	['F1', 'F11', 'F15', 'F22', 'F28', 'F32', 'F36', 'F40', 'F49', 'F53', 'F55', 'F57', 'F60', 'F61', 'F70', 'F71', 'F74', 'F76', 'F77', 'F78', 'F85', 'F86', 'F92', 'F95', 'F97', 'F99', 'F106']	1

Considering the previous results, it can be concluded that at least for datasets with complete and correctly balanced data, such as the one used in this test, the EFS method implemented achieved 100% stability.

5.4. Analysis of Results

In this chapter, we evaluated the Framework of Ensemble Feature Selection proposed in the previous chapter. The evaluation considered three aspects: (i) evaluate the impact of the relevant features selected by the EFS method in the classification of segments of EEG signals, (ii) evaluate a classifier of normal or abnormal segments of EEG signals using a set of relevant features selected by the EFS method, and (iii) evaluate the stability of the EFS method for selecting features with different samples of the dataset.

In reviewing the state of the art, several studies proposed approaches for building an ensemble method of feature selection algorithms [44] [48] [45]. However, most of the works were not even applied to EEG datasets and the results are not conclusive. Also, the works that proposed a kind of Ensemble Feature Selection used an approach based on stages, where the first selected the first subset of relevant features and in a second stage, the subset selected in the first stage is evaluated again by another feature selection algorithm. Then, the first stage could bias the second stage.

Likewise, some authors propose solutions to build the ensemble of feature selection algorithms based on filters [15][16][20][26]. Although this kind of algorithm is simple and easy to implement, the algorithms based on filters have many weaknesses. In this sense, if the goal of an ensemble learning scheme is to combine the decision of different models to create robust decisions, the idea to build an ensemble based on a filter could be a wrong decision.

Besides, most of the studies of Ensemble Feature Selection reviewed do not include stability as a metric to evaluate the quality of the feature selection process.

Considering the results showed in Table 22, the best results in the classification were achieved when the classifiers, Decision Trees, Logistic Regression, Random Forest, and SVM, used the subset of relevant features generated by the framework of Ensemble Feature Selection. Besides, the SVM was the algorithm that classified better for the evaluation performed to see the impact of the relevant features in the learning process. Thus, the model built to classify normal and abnormal EEG signals was based on SVM and the relevant features selected by the framework. This classifier achieved an accuracy of 97.46%, a rate of true negatives of 96.43%, a rate of true positive of 97.96%, a sensitivity of 96.78%, and a specificity of 97.25%. These values showed a performance equal to or greater than studies reviewed in the literature.

In the same way, a detector of epileptic was built to show the use of the classifier built in the context of the automatic reading of EEG signals and analyze the performance of the classifier in a scenario where there is not a balanced dataset. An EEG record contains many segments, however, the majority of them are normal segments and a reduced number of segments are abnormal, which generates an unbalanced scenario to evaluate the detection of abnormal EEGs as a binary classification task. Although the classifier used by the detector was trained using a perfect balanced dataset, the results showed an accuracy of 92.52%, sensitivity of 95.79%, and specificity of 92.48%. Considering that early detection of epilepsy is critical to its treatment, the priority for the detector is to increase the probability that a segment detected as normal is a normal segment; this decreases the rate of false-negative and consequently, decreases the probability of putting the patient’s health at risk. Although the tests evidenced a low rate of false negatives, the detector has not been designed to replace the work of an expert and its potential should be used to help the experts to identify abnormalities quickly and optimize their time.

On the other hand, the stability of the Ensemble Feature Selection method was evaluated by generating samples from the dataset, the results showed stability equal to 1, so the EFS method selected the same set of relevant features for all samples generated.

5.5. Conclusions

Because of the evaluation carried out in this chapter, the main conclusions are:

The impact of relevant features in the classifiers evaluated, Decision Tree, Logistic Regression, Random Forest and SVM, allowed demonstrate that machine learning models could improve their performance by discarding the features that are not relevant or represent noise.

The classifier built using features selected by the EFS method achieved an accuracy (97.64%), sensitivity (96.78%) and, specificity (96.27%) equal to or greater than the values found in the literature using only a subset of features selected instead all features. Additionally, the perfect stability achieved in the selection of features on different samples of the original dataset demonstrated the reliability of the feature selection process followed.

Although the detector of epileptic segments decreased almost 5 percentage points in the accuracy (92.52%), and specificity (92.48%) when this was used in unbalanced scenario, it achieved to maintain a sensitivity of 95.97%, which in the medical context is a priority.

The EFS used to select the subset of relevant features allowed decrease the complexity computational in the implementation of the detector of epileptic segments, due to, it is unnecessary calculate all features to describe a signal, only relevant features.

Finally, the chief contribution of this work was to validate the feature selection process carried out by the Ensemble Feature Selection method on a dataset of EEG signals. The results of the evaluation allowed to confirm that the use of EFS could help us improve the reliability of classifiers/detectors of epileptiform events in EEG signals.

Chapter 6:
Conclusions and future work

6.1. Conclusions

The conclusions of this research have been classified according: related works, training data (Electroencephalograms), building of the framework of ensemble feature selection and detection of epileptiform events:

Related works

- Most reviewed studies described new techniques or tools designed to solve specific problems, which shows important contributions in terms of algorithms, however, these contributions have not been considered into robust frameworks or libraries capable to support feature engineering process in another field.
- Most reviewed approaches of ensemble feature selection were designed following a cascade scheme. Following this scheme, the result of the first algorithm could skew the result of the second algorithm and so on, which could generate a final set of relevant features without the true relevant features.
- Many of the reviewed studies of EFS only considered FS algorithms based on filters, this could dissipate the advantages of ensemble learning, whose main goal is to combine the decisions of different models to compensate the weakness of some algorithms with the strengths of the others. Consequently, relevant features that usually are identified through the wrapper and embedded methods could be discarded in the final set of relevant features.
- The methods proposed to support ensemble feature selection strategies showing excellent results in the classification, however, they did not consider the effect of the missing values and imputation data into the FS and classification processes.
- Finally, in the reviewed studies we did not find an ensemble feature selection method designed to select relevant features and support automatic detection of epileptic seizures.

Training Data - Electroencephalograms

- The reading of EEGs implies identifying alterations, well-known patterns, in the waves that describe the electrical activity of the brain in 1 or several channels. This abnormal activity indicates two types of epileptic disorders: Focal and Generalized.
- In the literature, no dataset is annotated in such a way that it describes exactly the EEG segments where an epileptiform event occurs, nor does it describe the type of event. The datasets describe periods of abnormal brain activity but not the exact times of each type of abnormality.
- The raw data of EEG records cannot be processed by the machine learning algorithms, then, it was necessary to transform the data and extract features from the signals to describe them by using feature vectors.
- The recording period of EEGs ranges from 20 minutes to 72 hours, which generates that manual reading of an EEG can become a tedious and time-consuming task. In this sense, the development of tools for reading automatically EEGs could be an option to increase the opportunity of the service for the patients.
- The software tools identified in the literature showed a potential use reduce the time of reading an EEG by neurologists, and the studies conducted, following CRISP-DM, to understand the business aim: to decrease the EEGS reading time and the Data Mining objective: to determine how to automatically detect an epileptiform event in an EEG.
- The studies conducted following CRISP-DM showed the complexity to detect abnormal brain activity in the EEGs because of the diverse ways that they could appear and validate the potential showed by the related works to reduce the time of reading an EEG by neurologists.

Framework of EFS

- The proposal of the conceptual framework was developed following a quantitative method and the results describes with clarity the most relevant concepts in the Feature Selection process and determine how to improve by using EFS the performance of the classification algorithms enhancing the quality of the set of features used during their training process.
- The conceptual approach guided the development of an implementation framework capable of selecting features using ensemble of FS algorithms. The built framework achieved the selection of a set of relevant features with higher performance for classification tasks regarding the sets of features selected by single algorithms.
- The Framework's evaluation allowed to validate how the classification process can be improved considering an algorithm of FS based on ensemble instead of single FS methods. Likewise, the performance of our ensemble method achieved 100% of stability for the datasets used in the evaluation.
- The major contribution of the building of the framework to the field of Machine Learning is the definition of a structure that provides an understanding of how to improve the performance of FS based on the consensus of several techniques. This could guarantee better performance in classification algorithms and increase the reliability in those fields of application in which the reliability of the results must be high.

Detection of epileptiform events

- The evaluation conducted to analysis de automatic detection of epileptiform events evidenced that the models (Decision Tree, Logistic Regression, Random Forest and SVM) could improve their performance by discarding the features that are not relevant.
- The classifier of normal and abnormal segments built using features selected by the EFS method achieved an accuracy (97.64%), sensitivity (97.32%) and, specificity (96.27%) equal to or greater than the values found in the literature using only a subset of features selected instead all features.
- Additionally, the evaluation of the framework of EFS showed a perfect stability in the selection of features on different samples of the original dataset and allowed decrease the complexity computational in the implementation of the detector of epileptic segments, due to, it is unnecessary calculate all features to describe a signal, only relevant features.
- Although the detector of epileptic segments decreased almost 5 percentage points in the accuracy (92.52%), and specificity (92.48%) when this was used in unbalanced scenario, it achieved to maintain a sensitivity of 95.79%, which in the medical context is a priority.
- Finally, the chief contribution of this work was to validate the feature selection process carried out by the Ensemble Feature Selection method on a dataset of EEG signals could help to improve the reliability of classifiers/detectors of epileptiform events in EEG signals.

6.2. Future work

It is essential to mention the main goal of this research was to propose a mechanism to select relevant features to support the automatic detection of epileptiform events in EEG signals, however, we ended up defining a general approach to support EFS independently of the classification process and datasets. That means, data scientists and feature engineers could use our framework to figure out relevant features. As future work, we propose to improve the framework to support the parameterization of new implementations of feature selection methods. Also, to increase the number of consensus methods implemented, to offer new measures to evaluate the sets of features selected and to support the ensemble of FS methods not only in a heterogeneous scheme but also in a homogeneous one.

Besides, we propose to develop a graphic user interface of the framework to increase its use by the users (data scientists, AI engineers, data engineers, etc).

Additionally, we developed a classifier of normal and abnormal EEG segments to support the implementation of an automatic detector of epileptiform events, however, the method was designed to diagnoses recorded EEGs. As a further study, we propose to develop the integration with the devices that record EEGs to support analyzes in real-time.

Finally, we propose to combine EEG data and clinical information extracted from electronic health records to improve the accuracy of the results and use the information of channels where the abnormalities were identified to characterize the type of Epilepsy that the patient suffers.

REFERENCES

[1] O. M. de la Salud, “WHO | Media Centre,” 2010. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs220/en. [Accessed: 06-Jun-2015].

[2] O. M. de la Salud, “The world health report 2002 - Reducing Risks, Promoting Healthy Life,” *World Heal. Rep.*, vol. 1, 2002.

[3] Pontificia Universidad Javeriana, “Estudio de disponibilidad y distribución de la oferta de médicos especialistas, en servicios de alta y mediana complejidad en Colombia,” Boogtá, 2013.

[4] J. F. Ceron, “Encuesta colombiana de neurología - 2011,” *Acta Neurológica Colomb.*, vol. 28, no. 4, pp. 181–186, 2012.

[5] DANE, “Censo 2005- Cauca,” 2005. [Online]. Available: http://www.dane.gov.co/files/censo2005/PERFIL_PDF_CG2005/19000T7T000.PDF. [Accessed: 15-Apr-2015].

[6] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C. M. Hill, and P. R. White, “Signal processing techniques applied to human sleep EEG signals - A review,” *Biomed. Signal Process. Control*, vol. 10, no. 1, pp. 21–33, 2014.

[7] B. Boashash, G. Azemi, and N. Ali Khan, “Principles of time-frequency feature extraction for change detection in non-stationary signals: Applications to newborn EEG abnormality detection,” *Pattern Recognit.*, vol. 48, no. 3, pp. 616–627, 2015.

[8] T. M. Nunes, A. L. V. Coelho, C. A. M. Lima, J. P. Papa, and V. H. C. De Albuquerque, “EEG signal classification for epilepsy diagnosis via optimum path forest - A systematic assessment,” *Neurocomputing*, vol. 136, pp. 103–123, 2014.

[9] P. J. Karoly *et al.*, “Interictal spikes and epileptic seizures: Their relationship and underlying rhythmicity,” *Brain a J. Neurol.*, vol. 139, no. 4, pp. 1066–1078, 2016.

[10] L. Gao, J. Song, X. Liu, J. Shao, J. Liu, and J. Shao, “Learning in high-dimensional multimedia data: the state of the art,” *Multimed. Syst.*, vol. 23, no. 3, pp. 303–313, 2017.

[11] V. Bolón-canedo, N. Sánchez-maroño, A. Alonso-betanzos, J. M. Benítez, and F. Herrera, “A review of microarray datasets and applied feature selection methods,” *Inf. Sci. (Ny).*, vol. 282, pp. 111–135, 2014.

[12] W. Duch, “Feature Extraction,” Berlin: Springer, 2009, pp. 89–117.

[13] H. Yang, A. Gan, S. Shen, Y. Pan, J. Tang, and Y. Li, “Unsupervised ensemble feature selection for underwater acoustic target recognition,” in *Proceedings of the INTER-NOISE 2016 - 45th International Congress and Exposition on Noise Control Engineering: Towards a Quieter Future*, 2016.

[14] J. Meng, H. Hao, and Y. Luan, “Classifier ensemble selection based on affinity propagation clustering,” *J. Biomed. Inform.*, vol. 60, no. February, pp. 234–242, 2016.

[15] Y. Saeys, T. Abeel, and Y. Van De Peer, “Robust feature selection using ensemble feature selection techniques,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5212 LNAI, no. PART 2, pp. 313–325, 2008.

[16] E. Bustamante Z., *El sistema nervioso : desde las neuronas hasta el cerebro humano*, 1st ed. Medellín: University of Antioquia, 2007.

[17] S. Sanei and J. A. Chambers, *EEG Signal Processing*, First Edit. 2007, 2007.

[18] M. Teplan, “Fundamentals of EEG measurement,” *Meas. Sci. Rev.*, vol. 2, no. 2, pp. 1–11, 2002.

[19] S. Sillyy, Y. Li, and Y. Zhang, *EEG Signal Analysis and Classification Techniques and Applications*, First Edit. Melbourne: Springer, 2016.

[20] H. Jasper, “The ten-twenty electrode system of the International Federation,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 367–380, 1958.

[21] K. E. Misulis, *Atlas of EEG Seizure Semiology and Management*, Second Edi. 2014.

[22] F. Lopes da Silva, “EEG: Origin and Measurement,” *EEG - fMRI Physiol. Basis, Tech. Appl.*, pp. 1–539, 2010.

[23] A. T. *et al.*, “Automated Epileptic Seizure Detection Methods: A Review Study,” *Epilepsy - Histol. Electroencephalogr. Psychol. Asp.*, 2012.

[24] S. Siuly, E. Kabir, H. Wang, and Y. Zhang, “Exploring sampling in the detection of multicategory EEG signals,” *Comput. Math. Methods Med.*, vol. 2015, no. August, 2015.

[25] A. A. Kharbouch, “Automatic Detection of Epileptic Seizure Onset and Termination using Intracranial EEG Alaa Amin Kharbouch,” MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2012.

[26] A. S. Al-Fahoum and A. A. Al-Fraihat, “Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time-Frequency Domains,” *ISRN Neurosci.*, vol. 2014, no. February 2014, pp. 1–7, 2014.

[27] I. Guyon, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[28] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, “Systematic Mapping Studies in Software,” *Int. J. Softw. Eng. Knowl. Eng.*, vol. 17, no. 1, pp. 33–55, 2007.

[29] S. Santaniello, S. P. Burns, A. J. Golby, J. M. Singer, W. S. Anderson, and S. V. Sarma, “Quickest detection of drug-resistant seizures: An optimal control approach,” *Epilepsy Behav.*, vol. 22, no. SUPPL. 1, pp. S49–S60, 2011.

[30] Z. Roshan Zamir, “Detection of epileptic seizure in EEG signals using linear least squares preprocessing,” *Z. Roshan Zamir)*, vol. 133, pp. 95–109, 2016.

[31] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, “Requirements engineering paper classification and evaluation criteria: A proposal and a discussion,” *Requir. Eng.*, vol. 11, no. 1, pp. 102–107, 2006.

[32] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Computers in Biology and Medicine*, vol. 112. 2019.

[33] D. Yao, V. D. Calhoun, Z. Fu, Y. Du, and J. Sui, “An ensemble learning system for a 4-way classification of Alzheimer’s disease and mild cognitive impairment,” *J. Neurosci. Methods*, vol. 302, pp. 75–81, 2018.

[34] K. Raeisi, M. Mohebbi, M. Khazaei, M. Seraji, and A. Yoonessi, “Phase-synchrony evaluation of EEG signals for Multiple Sclerosis diagnosis based on bivariate empirical mode decomposition during a visual task,” *Comput. Biol. Med.*, vol. 117, no. September 2019, p. 103596, 2020.

[35] D. JIANG, Y. MA, and Y. WANG, “Sleep stage classification using covariance features of multi-channel physiological signals on Riemannian manifolds,” *Computer Methods and Programs in Biomedicine*, vol. 178. pp. 19–30, 2019.

[36] T. Zhang, W. Chen, and M. Li, “Classification of inter-ictal and ictal EEGs using multi-basis MODWPT, dimensionality reduction algorithms and LS-SVM: A comparative study,” *Biomed. Signal Process. Control*, vol. 47, pp. 240–251, 2019.

[37] O. Dehzangi and V. Sahu, “IMU-Based Robust Human Activity Recognition using Feature Analysis, Extraction, and Reduction,” *Proc. - Int. Conf. Pattern Recognit.*, vol. 2018-Augus, pp. 1402–1407, 2018.

[38] C. Wei, L. lan Chen, Z. zhen Song, X. guang Lou, and D. dong Li, “EEG-based emotion recognition using simple recurrent units network and ensemble learning,” *Biomed. Signal Process. Control*, vol. 58, p. 101756, 2020.

[39] A. K. Chowdhury, D. Tjondronegoro, V. Chandran, and S. G. Trost, *Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry*, vol. 49, no. 9. 2017.

[40] V. R. Elgin Christo, H. Khanna Nehemiah, B. Minu, and A. Kannan, “Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network,” *Comput. Math. Methods Med.*, vol. 2019, 2019.

[41] P. Drotár, M. Gazda, and J. Gazda, “Heterogeneous ensemble feature selection based on weighted Borda count,” *2017 9th Int. Conf. Inf. Technol. Electr. Eng. ICITEE 2017*, vol. 2018-Janua, pp. 1–4, 2017.

[42] L. lan Chen, A. Zhang, and X. guang Lou, “Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning,” *Expert Syst. Appl.*, vol. 137, pp. 266–280, 2019.

[43] B. Lei, W. Hou, W. Zou, X. Li, C. Zhang, and T. Wang, “Longitudinal score prediction for Alzheimer’s disease based on ensemble correntropy and spatial–temporal constraint,” *Brain Imaging Behav.*, vol. 13, no. 1, pp. 126–137, 2019.

[44] J. Sheng *et al.*, “A novel joint HCPMMP method for automatically classifying Alzheimer’s and different stage MCI patients,” *Behav. Brain Res.*, vol. 365, no. January, pp. 210–221, 2019.

[45] E. Bonilla-Huerta, A. Hernández-Montiel, R. Morales-Caporal, and M. Arjona-López, “Hybrid Framework Using Multiple-Filters and an Embedded Approach for an Efficient Selection and Classification of Microarray Data,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 1, pp. 12–26, 2016.

[46] E. Pippa *et al.*, “Improving classification of epileptic and non-epileptic EEG events by feature selection,” *Neurocomputing*, vol. 171, pp. 576–585, 2016.

[47] W. Chen, Y. Xu, Z. Yu, W. Cao, C. L. P. Chen, and G. Han, “Hybrid Dimensionality Reduction Forest with Pruning for High-Dimensional Data Classification,” *IEEE Access*, vol. 8, pp. 40138–40150, 2020.

[48] A. Rouhi and H. Nezamabadi-Pour, “A hybrid method for dimensionality reduction in microarray data based on advanced binary ant colony algorithm,” *1st Conf. Swarm Intell. Evol. Comput. CSIEC 2016 - Proc.*, pp. 70–75, 2016.

[49] C. F. Tsai and Y. T. Sung, “Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches,” *Knowledge-Based Syst.*, vol. 203, p. 106097, 2020.

[50] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong, “A new hybrid ensemble feature selection framework for machine learning-based phishing detection system,” *Inf. Sci. (Ny).*, vol. 484, pp. 153–166, 2019.

[51] B. Pes, N. Dessì, and M. Angioni, “Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data,” *Inf. Fusion*, vol. 35, pp. 132–147, 2017.

[52] K. Ravi and V. Ravi, “A novel automatic satire and irony detection using ensembled feature selection and data mining,” *Knowledge-Based Syst.*, vol. 120, pp. 15–33, 2017.

[53] I. B. M. IBM, “Manual CRISP-DM de IBM SPSS Modeler,” *IBM Corp.*, p. 56, 2012.

[54] M. Mera-Gaona, R. Vargas-Canas, and D. M. Lopez, “Towards a Selection Mechanism of Relevant Features for Automatic Epileptic Seizures Detection.,” *Stud. Health Technol. Inform.*, vol. 228, no. 4, pp. 722–6, 2016.

[55] M. Mera, D. M. López, R. Vargas, and M. Miño, “Automatic detection of epileptic spike in EEGs of children using matched filter,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11309 LNAI, pp. 392–402, 2018.

[56] A. L. Goldberger *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals,” *Components a New Res. Resour. Complex Physiol. Signals*, vol. 101, no. 23, pp. 215–220, 2000.

[57] Mathworks, “Entropy function.” [Online]. Available: http://www.mathworks.com/help/wavelet/ref/wentropy.html. [Accessed: 06-Mar-2016].

[58] M. A. Naderi and H. Mahdavi-Nasab, “Analysis and classification of EEG signals using spectral analysis and recurrent neural networks,” *2010 17th Iran. Conf. Biomed. Eng. ICBME 2010 - Proc.*, 2010.

[59] K. Gopika Gopan, N. Sinha, and J. Dinesh Babu, “Statistical features based epileptic seizure EEG detection - An efficacy evaluation,” *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, pp. 1394–1398, 2015.

[60] A. M. Aldabbagh, T. N. Alotaiby, S. A. Alshebeili, and A.-E. E., “Low Computational Complexity EEG Epilepsy Data Classification Algorithm for Patients With Intractable Seizures,” *Arch. Neurol.*, vol. 27, no. 3, p. 205, 2015.

[61] J. C. Bancroft, “Introduction to Matched Filters,” 2017.

[62] J. Hermand and W. I. Roderick, “Acoustic Model-Based Matched Filter Processing for Fading Time-Dispersive Ocean Channels : Theory and Experiment,” *IEEE J. Ocean. Eng.*, vol. 18, no. 4, pp. 447–465, 1993.

[63] E. María and C. Roldan, “Medidas de entroìa en el procesado de señales biológicas: robustez y caracterización frente a la pérdida de muestras y longitud de los registros,”

Universidad Politécnica de Valencia, 2014.

[64] B. Hjort, “EEG analysis based on time domain propertiesAnalyse EEG basee sur les series temporelles,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, 1970.

[65] A. Gónzalez Pereira, “Selección de características para el reconocimiento de patrones con datos de alta dimensionalidad en fusión nuclear,” UNIVERSIDAD NACIONAL DE EDUCACION A DISTANCIA, 2015.

[66] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd ed. San Diego, 2003.

[67] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artif. Intell.*, vol. 97, pp. 245–271, 1997.

[68] R. Kohavi and H. John, “Artificial Intelligence Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, no. 97, pp. 273–324, 1997.

[69] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[70] N. Sa and V. Bolo, “An ensemble of filters and classifiers for microarray data classification,” *Pattern Recognit. J.*, vol. 45, pp. 531–539, 2012.

[71] C. Lee and Y. Leu, “A novel hybrid feature selection method for microarray data analysis,” *Appl. Soft Comput.*, vol. 11, pp. 208–213, 2011.

[72] Y. Li, G. Wang, H. Chen, L. Shi, and L. Qin, “An Ant Colony Optimization Based Di mension Reduction Method for High-Dimensional Datasets,” *J. Bionic Eng.*, vol. 10, no. 2, pp. 231–241, 2013.

[73] R. Cai, Z. Hao, X. Yang, and W. Wen, “An efficient gene selection algorithm based on mutual information,” *Neurocomputing*, vol. 72, pp. 991–999, 2009.

[74] V. Basto, I. Yevseyeva, J. R. Méndez, and J. Zhao, “A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification,” *Appl. Soft Comput. J.*, vol. 48, pp. 111–123, 2017.

[75] D. Choi, B. Ko, H. Kim, and P. Kim, “Journal of Network and Computer Applications Text analysis for detecting terrorism-related articles on the web,” *J. Netw. Comput. Appl.*, vol. 38, pp. 16–21, 2014.

[76] D. N. Den Hartog, V. Kobayashi, H. Bekers, and G. Kismihók, “Text Classification for Organizational Researchers : A Tutorial,” *Organ. Res. Methods*, vol. 21, no. 3, pp. 1–34, 2017.

[77] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, “Polarity shift detection , elimination and ensemble : A three-stage model for document-level sentiment analysis,” *Inf. Process. Manag.*, vol. 52, no. 1, pp. 36–45, 2016.

[78] A. García-pablos, M. Cuadros, and G. Rigau, “W2VLDA : Almost unsupervised system for Aspect Based Sentiment Analysis,” *Expert Syst. Appl.*, vol. 91, pp. 127–137, 2018.

[79] A. Bandhakavi, N. Wiratunga, P. Deepak, and S. Massie, “Lexicon based Feature Extraction for Emotion Text Classification,” *Pattern Recognit. Lett.*, vol. 93, pp. 133–142, 2016.

[80] M. Mera-gaona, R. Vargas-canas, and D. M. Lopez, “Towards a Selection Mechanism of Relevant Features for Automatic Epileptic Seizures Detection,” *Stud. Health Technol. Inform.*, vol. 228, pp. 722–726, 2016.

[81] V. Bolón-canedo and N. S. A. Alonso-betanzos, “Feature selection for high-dimensional data,” *Prog. Artif. Intell.*, 2016.

[82] D. Dheeru and E. Karra Taniskidou, “UCI Machine Learning Repository,” *University of California, Irvine, School of Information and Computer Sciences*. School of Information and Computer Science, Irvine, CA, 2017.

[83] C. Chang and C. Lin, “LIBSVM : A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 11, pp. 1–39, 2011.

[84] S. D. Bay, “Combining Nearest Neighbor Classifiers Through Multiple Feature Subsets,” in *Fifteenth International Conference on Machine Learning*, 1996, pp. 37–45.

[85] Z. Zheng and G. I. Webb, “Stochastic Attribute Selection Committees,” *Lect. Notes Comput. Sci.*, pp. 321–332, 1998.

[86] D. W. Opitz, “Feature Selection for Ensembles,” in *National Conference on Artificial Intelligence*, 1999.

[87] Y. Piao, M. Piao, K. Park, and K. H. Ryu, “An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data,” *BIOINFORMATICS*, vol. 28, no. 24, pp. 3306–3315, 2012.

[88] Y. Jabareen, “Building a Conceptual Framework : Philosophy , Definitions , and Procedure,” *Int. J. Qual. Methods*, vol. 8, no. 4, pp. 49–62, 2009.

[89] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, “Five steps to conducting a systematic review,” 2003.

[90] K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” in *Information and Software Technology*, 2015, vol. 64, pp. 1–18.

[91] M. Mera, C. González, and D. M. López, “Towards an intelligent decision support system for public health surveillance-A qualitative analysis of information needs,” *Stud. Health Technol. Inform.*, vol. 202, no. July, pp. 44–47, 2014.

[92] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data mining*. Boston: Kluwer Academic Publishers, 1998.

[93] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[94] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[95] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, “Ensemble feature selection: Homogeneous and heterogeneous approaches,” *Knowledge-Based Syst.*, vol. 118, pp. 124–139, 2017.

[96] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2825–2830, 2011.

[97] U. Neumann *et al.*, “Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach,” *BioData Min.*, pp. 1–14, 2016.

[98] A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: A study on high-dimensional spaces,” *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.

[99] D. Lachner-piza, N. Epitashvili, A. Schulze-Bonhage, T. Stieglitz, J. Jacobs, and M. Dümpelmann, “A single channel sleep-spindle detector based on multivariate classification of EEG epochs: MUSSDET,” *J. Neurosci. Methods*, vol. 297, pp. 31–43, 2017.

[100] J. Su, D. Yi, C. Liu, L. Guo, and W.-H. Chen, “Dimension Reduction Aided Hyperspectral Image Classification with a Small-sized Training Dataset: Experimental Comparisons,” *Sensors*, vol. 17, no. 12, pp. 1–20, 2017.

[101] K. Nurnadia M., H. M., Y. S., and B. Shafriza Nisha, “Locality sensitivity discriminant analysis-based feature ranking of human emotion actions recognition,” *J Phys Ther Sci*, vol. 27, no. 8, pp. 2649–2653, 2015.

[102] E. Garbarine, J. Depasquale, V. Gadia, R. Polikar, and G. Rosen, “Information-theoretic approaches to SVM feature selection for metagenome read classification,” *Comput. Biol. Chem.*, vol. 35, no. 3, pp. 199–209, 2011.

[103] L. Mohammad, A. Tajudin, M. A. Al-betar, and O. Ahmad, “Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering,” *Expert Syst. Appl.*, vol. 84, pp. 24–36, 2017.

[104] U. Neuman, N. Genze, and D. Heider, “EFS: an ensemble feature selection tool implemented as R-package and web-application,” *BioData Min.*, pp. 1–9, 2017.

[105] W. Koehrsen, “A Feature Selection Tool for Machine Learning in Python,” *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0. [Accessed: 07-Nov-2018].