Yachay University of Experimental Technology Research

School of Mathematical and Computational Sciences

Final Grade Project

An exploratory study on the characterization and classification of electroencephalographic signals for the design of computer-aided epilepsy diagnosis system

Author:
Emil Darío Vega Gualán

Advisor:
Diego Hernán Peluffo Ordoñez

Requirement for obtaining the grade of Information Technology Engineer.

Urcuquí - September 18, 2019

Autoría

Yo, **EMIL DARÍO VEGA GUALÁN**, con cédula de identidad 0706950151, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autor(a) del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Agosto 2019.

Emil Darío Vega Gualán
CI: 0706950151

Autorización de publicación

Yo, **EMIL DARÍO VEGA GUALÁN**, con cédula de identidad 0706950151, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Urcuquí, Agosto 2019.

Emil Darío Vega Gualán
CI: 0706950151

Dedication

"To my parents because they have worked hard every day for me and my siblings to give us the best inheritance that can exist: education. To my little sister Hellen, may this work be an inspiration in her life to achieve her academic goals."

Acknowledgements

“I thank my whole family mainly because they always are supporting me in all my goals. To my old and new friends because they have made this long road more entertaining and easier to travel. To my tutor and friend Diego Peluffo who was the most important piece to achieve this goal because he taught me the skills and knowledge necessary to develop this research. Also, he showed me that human values should never be lost in the professional life. To my new friend Andres Vargas who helped me to solve some problems presented in this research”

Resumen

La epilepsia ocurre cuando la actividad eléctrica de las neuronas sufre un desequilibrio. Esta se ha convertido en el tercer trastorno neurológico más común después del accidente cerebrovascular y la demencia, -se cree que afecta al 0.5 - 1.5% de la población mundial. Afecta principalmente a niños menores de 10 años y personas mayores de 65 años, siendo más común en países en desarrollo y en clases socioeconómicas desfavorecidas. Su posible diagnóstico es a través del análisis de señales electroencefalográficas (EEG). Hoy en día, dado que se debe cumplir tanto su diagnóstico apropiado como la localización precisa de la fuente epiléptica, se utilizan sistemas computacionales para respaldar el procedimiento de diagnóstico. En términos generales, tales sistemas realizan la asistencia de diagnóstico automático en cuatro etapas principales: adquisición de señal EEG, preprocesamiento, caracterización y clasificación. Una vez adquiridas y preprocesadas, las señales EEG deben representarse adecuadamente para posteriormente clasificarse en categorías de diagnóstico (ausencia o cualquier nivel de presencia de actividad convulsiva). A pesar de que existe una amplia gama de alternativas para caracterizar y clasificar las señales de EEG para fines de análisis de epilepsia, muchos aspectos clave relacionados con la precisión y la interpretación fisiológica todavía se consideran cuestiones abiertas. En este sentido, en este trabajo, se propone un estudio exploratorio de las técnicas de procesamiento de señales de EEG, con el objetivo de identificar las técnicas más adecuadas y avanzadas para caracterizar y clasificar las crisis epilépticas. Para hacerlo, se diseña y desarrolla un estudio comparativo sobre varios subconjuntos de características (medidas estadísticas tanto de las señales originales como de la transformación espectral de las mismas), así como algunos clasificadores representativos (clasificador de análisis discriminante lineal (LDC), clasificador de análisis discriminante cuadrático (QDC), k-vecinos más cercanos (kNN) y máquina de vectores de soporte (SVM)). La validación del sistema propuesto se lleva a cabo mediante una configuración experimental exhaustiva sobre una base de datos estándar de UCI Machine Learning Repository, denominado: "Epileptic Seizure Recognition Data Set". Como resultados notables, se demuestra experimentalmente que un proceso de caracterización basado en índices estadísticos de descomposiciones impulsadas por la transformada wavelet y el clasificador de máquina de vectores de soporte son los enfoques más adecuados para diseñar un sistema automático para identificar señales EEG diagnosticadas con epilepsia. Además, el rendimiento general del sistema de reconocimiento de patrones obtenido (para el escenario bi-clase) -en términos de mediciones basadas en matriz de confusión- asciende a 96%, 85% y 98% del rendimiento de clasificación, sensibilidad y especificidad, respectivamente.

Palabras clave: Clasificación de patrones, convulsión, diagnóstico de epilepsia, electroencefalograma (EEG), selección de características, transformada Wavelet discreta (DWT).

Abstract

The epilepsy disorder occurs when the localized electrical activity of neurons suffers from an imbalance. Epilepsy has become the third most common neurological disorder after stroke and dementia -it is believed that affects 0.5 - 1.5% of the world population. It mainly affects children under 10 and people over 65, being more common in developing countries and in disadvantaged socioeconomic classes. Its possible diagnose is via the analysis of electroencephalographic (EEG) signals. Nowadays, since both its appropriate diagnosis and the accurate epileptic source localization must be fulfilled, computational systems are used to support the diagnosis procedure. Broadly, such systems perform the automatic diagnostic-assistance into four main stages, namely: EEG signal acquisition, preprocessing, characterization and classification. Once acquired and preprocessed, EEG signals must be properly represented to be subsequently classified into diagnostic categories (absence or any level of presence of seizure activity). Despite there exists a wide range of alternatives to characterize and classify EEG signals for epilepsy analysis purposes, many key aspects related to the accuracy, computational cost, and physiological interpretation are still considered as open issues. In this connection, in this work, an exploratory study of EEG signal processing techniques is proposed, aimed at identifying the most adequate state-of-the-art techniques for characterizing and classifying epileptic seizures. To do so, a comparative study is designed and developed on several subsets of features (namely, statistical measures on both the original signals and the spectral transformation thereof), as well as some representative classifiers (linear discriminant analysis classifier (LDC), quadratic discriminant analysis classifier (QDC), k-nearest neighbor (kNN) and support vector machine (SVM)). Proposed system validation is carried out by means of an exhaustive experimental setup over a gold standard database from the UCI Machine Learning Repository, so-named: "Epileptic Seizure Recognition Data Set". As remarkable results, it is experimentally proved that a characterization process based on statistical indices from wavelet-transform-driven decompositions, and the support vector machines as classifiers are the most suitable approaches for designing an automatic system to identify epilepsy-diagnosed EEG signals. As well, the overall performance of the obtained pattern recognition system (for the bi-class scenario) -in terms of confusion-matrix-based measurements- amounts 96%, 85% and 98% of classification performance, sensitivity, and specificity, respectively.

Keywords: Discrete wavelet transform (DWT), electroencephalogram (EEG), epilepsy diagnosis, feature selection, pattern classification, seizure.

Contents

List of Figures

List of Tables

Introduction

The epilepsy disorder occurs when the localized electrical activity of neurons suffers from an imbalance. Epilepsy has become the third most common neurological disorder after stroke and dementia -it is believed that affects 0.5 - 1.5% of the world population. It mainly affects children under 10 and people over 65, being more common in developing countries and in disadvantaged socioeconomic classes [7], [8]. Its possible diagnose is via the analysis of electroencephalographic (EEG) signals. Nowadays, since both its appropriate diagnosis and the accurate epileptic source localization must be fulfilled, computational systems are used to support the diagnosis procedure.

Certainly, the human brain is a complex system and unveiling its operation and functioning patterns is still a great open research issue. In this way, there are non-invasive techniques which allow to get data that help to understand in some way how brain works. One of these techniques is the electroencephalography (EEG) which is a record of the electrical potentials generated by the cerebral cortex nerve cells. There are two types of EEG depending on where the signals are taken in the head: scalp or intracranial. For scalp EEG, small metal discs, also known as electrodes, are placed on the scalp with good mechanical and electrical contact. Intracranial EEG is obtained by special electrodes implanted in the brain during a surgery. The recorded EEG provides a continuous graphic exhibition of the spatial distribution of the changing voltage fields over time. [9].

According to the Epilepsy Foundation, the epilepsy is "a chronic disorder, the hallmark of which is recurrent, unprovoked seizures. A person is diagnosed with epilepsy if they have two unprovoked seizures (or one unprovoked seizure with the likelihood of more) that were not caused by some known and reversible medical condition like alcohol withdrawal or extremely low blood sugar" [10]. Besides, this illness can affect to adults and kids because there is not a general cause to prevent it. The seizures -which appear recurrently but infrequently- are the joint reaction of a large number of neurons when going through an excessive and synchronous electrical discharge. There are two types of epileptic seizures: partial and generalized. When the synchronous electrical discharge is produced in a local part of the brain, it is called partial epileptic seizures. Otherwise, when this synchronous electrical discharge is produced in the whole brain, it is called generalized epileptic seizures [11].

Surface (on-scalp) electroencephalography (EEG) is a non-invasive method used to monitor the nonlinear electrical function of the brain's nerve cells. Therefore, EEG is a highly-recommended and useful tool for the evaluation and treatment of epilepsy. Some plottings of EEG signals allow for observing spikes, sharp waves and spike-and-wave complexes not only when a seizure is occurring, but also pre-occurrence- and between- seizures [11]. Old EEG procedures were mostly manual and therefore high time-consuming (up to days), as well as, prone to error. This is the reason why nowadays such procedures are computer-aided, and all current, related-to-epilepsy research works are aimed to develop reliable and accurate computational techniques to detect epileptic activity through EEG recordings.

Developed systems for automatic diagnostic-assistance perform it through five stages: EEG signal acquisition, preprocessing, characterization, classification and in-context interpretation (visualization). During the last years there has been an increasing interest for the improvement of the computational techniques applied for characterization and classification. In this sense, there is a wide range of alternatives to characterize and classify EEG signals. Notwithstanding, many fundamental aspects, such as accuracy, computational cost, and physiological interpretation, are still under improvement and research. In other words, in spite of existing approaches, epileptic seizures diagnosing and prediction is still a challenging and open case of investigation.

In this context, this degree thesis presents a whole EEG analysis framework for evaluating the ability of characterization and classification techniques on the epilepsy diagnose. In other words, this thesis' main goal is to develop a methodology to compare techniques for both characterizing and classifying EEG signals within a epilepsy diagnosing framework. Such a framework includes stages for preprocessing, characterization, feature selection, classification, performance quantification and visualization, and works as follows: As a preprocessing stage, a simple amplitude normalization is used. Subsequently, signals are characterized through statistical measures on both the original signals and the spectral transformation thereof. Afterwards, a set of features are chosen by applying recommended feature selection methods (`Bestfirst` and `Ranker`). Then, selected features are classified by using representative classification approaches, such as: linear discriminant analysis classifier (LDC), quadratic discriminant analysis clasifier (QDC), k-nearest neighbor (kNN) and support vector machine (SVM). Finally, box plots, receiver operating characteristic (ROC) curves and confusion-matrix-based measures are used to quantify the proposed framework performance. Finally, to facilitate an in-context visual interpretation, a friendly-user interface is developed.

For experiments, the "Epileptic Seizure Recognition Data Set" is tested, which is available at the UCI machine learning repository `https://archive.ics.uci.edu/ml`. Such a database was introduced by [6]. Several experiments are carried out to evaluate different key interest, among them: the relationship between the nature of features and classification performance, the behavior of classifiers on the different data structures, the proper classification performance, and the interpretation of classification results in terms of epilepsy diagnosis.

Chapter 1

Preliminaries

1.1 Problem statement

Electroencephalography is an exploratory technique based on recording electrical activity from the brain. Furthermore, since this technique is very common and there are a lot of studies on it, the signal acquisition is simple and accessible. Nowadays, its practice in the diagnosis of epilepsy is very recommended. Taking into account that Epilepsy is the third most common neurological disorder affecting 0.5 - 1.5% of the world population [1], it is important to develop computational systems to support the diagnosis procedure.

Although, several studies have been developed to diagnose epilepsy from EEG signals, determining a method considered as optimum for epilepsy diagnosis through EGG signals remains a great-of-interest open issue, which is difficult to tackle since it involves several aspects, such as accuracy, computational cost, location effectiveness, and reliability. Furthermore, another big problem to take into consideration is the fact that the analyzed signals may be very similar to each other, making their classification a difficult task.

1.2 Justification

The Epilepsy Foundation is investing about 65 millions of dollars in epilepsy research [12]. This foundation is concerned about the epileptic people and is giving grants to new innovative technologies like digital tools and aided systems, for epilepsy. The diagnosis of epilepsy is the first step to start to fight against this disease. In such vein, the present work is been developed to find the best methods and techniques to automatically improve the diagnosis of epilepsy. In this sense, it is necessary to make a comparison of some classifiers, with the aim to find the best balance between accuracy and reliability.

Finding the best method for epilepsy diagnosis is the challenging aim of this work. To achieve this goal, a comparative study is performed, which evaluates techniques for characterization and classification of epilepsy seizure from EGG signals.

Since the difficulty to achieve the right accuracy, computational cost, reliability and location effectiveness, this work is not limited just to the collection of data, also it brings new challenges of research for the scientific community and computational-development engineers who are focused on this field. In this way, it gives the chance to do deeper explorations in medical, industrial or even commercial fields.

1.3 Contribution

Certainly, a computational system to diagnose epilepsy shall be faster than a human at the task of determining whether a person has a disease. Nonetheless, such a diagnosis will be useful, if the systems reaches an admissible or reasonable percentage of accuracy. This work is intended to make an exploratory study to establish important aspects to do a comparison and selection of analysis techniques of EGG signals to diagnose epilepsy. Furthermore, it aims to find a method with a good percentage of security to do the hard task of epilepsy diagnosis through electroencephalogram signals.

To do so, a proper methodology has been developed to choose the optimal model. It will open up ways of knowing where research should focus on the characterization and classification of EGG signals applied to the diagnosis of epilepsy.

Finally, from obtained results, specific information is provided to both epilepsy-diagnosed patients and specialists so that the model meeting their expectations of efficiency and accuracy can be selected. In this sense, it represents an important contribution to the medical field. It will contribute to the development of new research through the understanding of how the epileptic seizure affects the EGG signals.

1.4 Document organization

This work is divided into seven main Chapters named as follows: Preliminaries, Objectives, Theoretical Framework, Methodology, Experimental setup, Result, and Conclusions.

In Chapter 1, the problem statement, the justification of this work and the scientific contributions of this research are presented.

Chapter 2 states the general and specifics objectives.

Chapter 3 presents a brief overview where the reader can find the definition of epilepsy and its relationship with the electroencephalographic signals. Furthermore, an introductory theoretical background is provided, which is a broadly explanation about what entails to work with these signals through machine learning and pattern recognition.

In Chapter 4, the methodology for detecting epileptic seizures trough EGG signals is described. It includes an explanation and brief description of the applied stages for preprocessing, signal decomposition, characterization, feature selection, and classification.

In Chapter 5, the experimental setup is discussed. The performance measures and the tests on the database are shown.

In Chapter 6, the results are discussed and shown through tables and meaningful graphics.

In Chapter 7, the conclusions obtained from this work are presented. Also, future works that can improve the proposed methodology and help to establish open issues are mentioned.

Chapter 2

Objectives

2.1 General Objective

To develop a methodology for the exploratory study through a comparison of characterization techniques and automatic classification of EEG signals in order to detect epilepsy.

2.2 Specific Objectives

- To establish a set of characteristics of EEG signals of temporal, morphological, spectral, representation type, based on information theory and statistics that adequately represent the epileptic seizure area.
- To select and implement classifiers based on models and distances in high-level programming environments to be evaluated with characteristics of EEG signals.
- To design a methodology for comparing characterization and classification techniques of EEG signals aimed at identifying the epileptic seizure area in order to determine which of the techniques achieve a good compromise between accuracy and in-context interpretability.

Chapter 3

Theoretical Framework

In 1970 just began the studies of mechanizing the detection of epileptic seizures [13]. Nevertheless, the algorithms to make the implementation of a physical problem solution were developed 30 years after [14]. Nowadays, there are a lot of studies about the detection of epileptic seizures. Most of these studies are based on electroencephalographic (EEG) signals. The analysis of EEG signals uses the fact that the information processing in the brain is reflected in the EEG as dynamical changes of the electrical activity in time, frequency, and space [15]. These signals are pre-processed and characterized to classify the signals with epileptic seizures from the other signals.

3.1 Epilepsy

Most people confuse an epileptic seizure with a convulsion. Nevertheless, a convulsion is something less serious. Timothy Huzar in his study says that a convulsion "occurs when a person's muscles contract uncontrollably. They can continue for a few seconds or many minutes. Convulsions can happen to a specific part of a person's body or may affect their whole body". [16]. A convulsion is not related to an electrical disturbance in the brain but uncontrollable muscle contractions.

Even though a convulsion is not the same that an epileptic seizure when this last one occurs, it causes convulsions. A serious epileptic seizure always affects the whole body because of the electrical disturbance in the brain. Also, the person who is having an epileptic seizure starts drooling uncontrollably and loses consciousness for a short time [17]. The exact definitions for epilepsy and epileptic seizures still are under discussion, but there is an association of epilepsy who gives some definitions as ILAE and IBE.

The International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE) have defined epilepsy as "a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiologic, cognitive, psychological, and social consequences of this condition. The definition of epilepsy requires the occurrence of at least one epileptic seizure" [18]. Also, they define an epileptic seizure as "a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain" [18]. Notice that an epileptic seizure is related to the electrical activity of the brain that produces convulsions and in the long term, it can turn into a cerebral deterioration. As soon the epilepsy is diagnosed it can receive the right treatment to control the seizure episodes.

3.1.1 Epilepsy recognition

To recognize whether a person has epilepsy, it is necessary a diagnosis from a doctor. In this regard, the doctor needs to review the symptoms and medical history. Furthermore, the doctor has to do several tests for the patient to know if there is a case of epilepsy or not. This evaluation according to the Mayo Clinic website [19], may include:

- **A neurological exam**. The doctor may test your behavior, motor abilities, mental function, and other areas to diagnose your condition and determine the type of epilepsy you may have.
- **Blood tests**. The doctor may take a blood sample to check for signs of infections, genetic conditions or other conditions that may be associated with seizures.

The neurological exam can include tests as follows: Electroencephalogram (EEG), High-density EEG, Computerized tomography (CT) scan, Magnetic resonance imaging (MRI), Functional MRI (fMRI), Positron emission tomography (PET), Single-photon emission computerized tomography (SPECT), Neuropsychological tests [19]. All these tests have to be analyzed by the doctor to determine if the patient has epilepsy and what kind it could be. Notice that depend on the kind of exam it could take a long time to get the results of the test and longer get the diagnosis.

3.1.2 Epilepsy Recognition through EEG signals

The most common test for diagnosing epilepsy is the electroencephalogram (EEG) where some electrodes are attached in the scalp and these record the electrical activity of the brain [19]. An epileptic seizure causes that the normal waves of the brain will be altered. Even when the patient is not having a seizure this alteration remains [19]. This altered waves can be recorded in the electroencephalogram. In this way, EEG signals can help the doctor to make a diagnosis to know what kind of seizure the patient has.

Likewise, the EEG signals help doctors to make a diagnosis. It could be computationally implemented to make the diagnosis faster. Several studies about the classification of EEG signals follows almost the same flow chart of steps showed in Figure 3.1. First, it is necessary to acquire the EEG signal in a file. Second, we have to preprocess the data through normalization and filters to get just the proper signals of the EEG. Third, the signals need to be characterized to find a minimum set of features. Finally, with the selected features it is necessary to apply classification methods to determine when there is epilepsy or not.

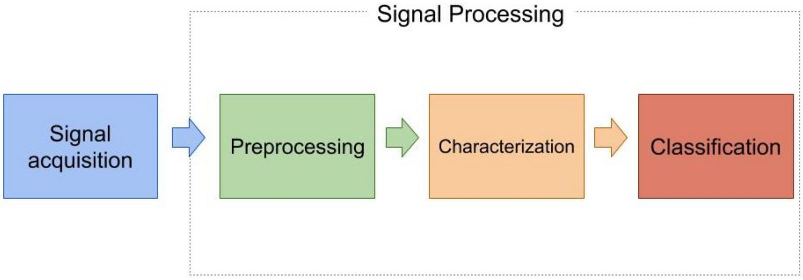


Figure 3.1: The basic operation of a system for diagnosing epilepsy based on EEG signals.

3.1.3 Epilepsy Sources Location

Currently, one of the most effective ways to treat epilepsy is through brain surgery (also called, neurosurgery) procedure. Nonetheless, this procedure can be dangerous if the brain region wherein the epileptic crisis ir originated is not well localized. This brain region is known as the Epileptogenic Region (ER) [1]. Neurosurgery refers to remove the ER in an expert way, it means, the removing must be done accurately and skillfully to minimizing the risk of causing collateral damages [20].

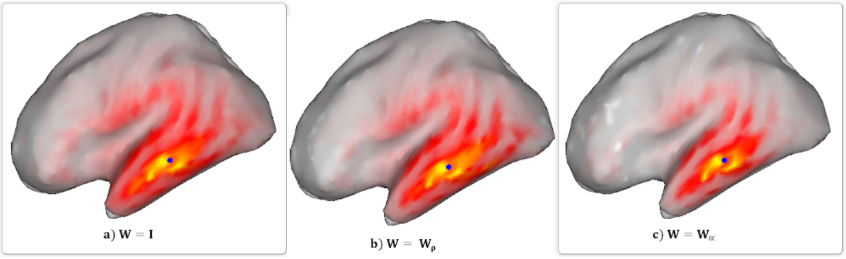

a) $\mathbf{W} = \mathbf{I}$
b) $\mathbf{W} = \mathbf{W}_\rho$
c) $\mathbf{W} = \mathbf{W}_\propto$

Figure 3.2: Source localization results for the three considered approaches. All the methods are tested regarding the same simulated source (blue point). W is the weight used for each approach. *Source*: [1]

A non-invasive way to locate the epileptic sources is through an analysis of electrical potential which can be recorded by EEG signals. These can be mapped onto geometrical coordinates using mathematical models, thus, they can reveal the location of epileptic source [21]. One research on the location of the epileptic source is [1], which uses an inverse problem model to analyze the EEG signals and make a mapping to found the epileptic source. This mentioned research performs an exploratory study of weighted inverse models aimed at identifying the benefit of incorporating weighting factors effect into the solution of the inverse model problem. The results of that work are shown in Figure 3.2, where three different weights are used to find the proper epileptic source.

3.2 Electroencephalography (EEG)

Since the epileptic seizures are electrical activity in the brain, the use of electroencephalography is very useful. Furthermore, it is the technique most used for doctors to diagnose epilepsy. In this way, it is possible to do a deeper observation about what is happening inside the brain through electroencephalography because it allows recording all the brain waves [19]. These waves can be provoked by emotions, slight movements like a blink, thoughts, strong movements, feeling of hunger. Thanks to electroencephalography it is possible to have enough information about brain activity, even it is called "window on the mind" [22]. It is from all this brain activity where it is necessary to filter the signals that indicate an epileptic seizure.

"Electroencephalogram" was the named given by Hans Berger when in 1924 he made the first recording of the electric field of the human brain. This recording was blurred and noisy, even many people considered it meaningless, but it was the starting point of electroencephalography [23]. Conventional electroencephalography measures the electrical activity produced by the brain through an electrode placed in the scalp. The article "Electroencephalography (EEG)" by Picton and Mazaheri, gives an explanation of how the electrodes catch the signal brain. They said that "When the neurons of the brain process information, they do so by changing the flow of electrical currents across their membranes. These changing currents, particularly those caused by the synaptic excitation and inhibition of cortical neurons, generate electric fields that can be recorded using small electrodes attached to the surface of the scalp. The potentials between different electrodes are amplified and displayed as they fluctuate over time" [24].

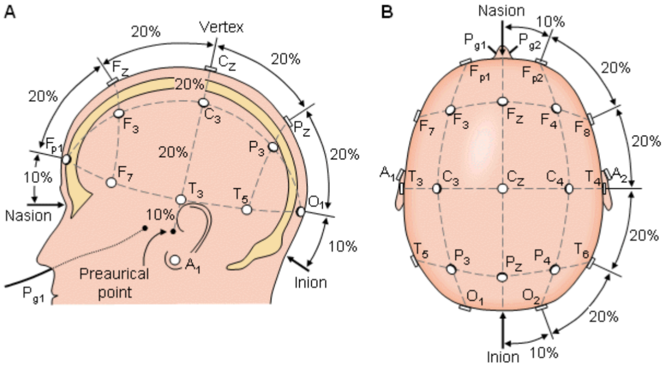


Figure 3.3: The international 10-20 system seen from (A) left and (B) above the head. *Source*: [2]

Since the EGG signal is measured by density, the location of electrodes and distance between them are very important for a good recording [25]. In this way, there exists a standardized system for the placement of electrodes called "10-20 system" [2]. For this system, 21 electrodes are used and Figure 3.3 shows how there has to be placed on the surface of the scalp. Its placement is described in [26], [27] as follows: *Nasion* are reference points, which is the radix (sellion) on the nose (top part), level with the eyes; and *inion*, which is in the occipital bone, placed at the back of the head. After that, the transverse and median planes are used to measure the skull perimeters. These perimeters have to be divided into 10% and 20% intervals to determine the electrode locations. Figure 3.3 part B shows three other electrodes that are placed on each side at the same distance from the neighboring points.

Another system used besides of 10-20 system is the 10% system. The American Electroencephalographic Society (AES) is who standardized the location and nomenclature of these electrodes [28]. The difference is that the names of the four electrodes are changed. These are: T_7, T_8, P_7, and P_8. In Figure 3.4 these are drawn black with white text. Following the principles of 10-20 system, it increases its resolution but has the same approach and electrodes designation [25].

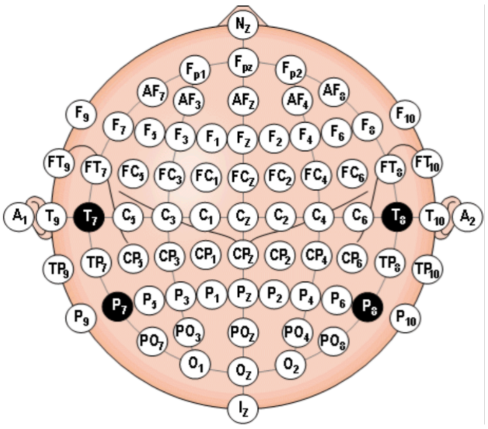


Figure 3.4: Location and nomenclature of the 10% system. *Source*: [2]

According to [25], EEG signals "depends on the degree of cerebral cortex activity measured in voltage as a function of time and are classified according to their frequency, magnitude, wave morphology, spatial distribution and reactivity". In [29] and [3] mention that the classifications of EEG waveform are commonly through their frequency bands which are five: alpha, beta, theta, delta and gamma bands. Figure 3.5 shows all the wave bands of an electroencephalography. These bands in [3], are associated with a mental state which is described as follows:

- **Delta waves** (δ). Their bandwidth is 0.5 - 4 `Hz`, being the slowest waves, normally detected during the deep and unconscious sleep
- **Theta waves** (θ). Their bandwidth is 4 - 8 `Hz`, and these are observed during some states of sleep and quiet focus.
- **Alpha waves** (α). Their bandwidth is 8 - 15 `Hz`, and these are originated during periods of relaxation with eyes closed but still awake.
- **Beta waves** (β). Their bandwidth is 14 - 30 `Hz`, and these are originated during normal consciousness and active concentration.

- **Gamma waves** (γ). Their bandwidth is over 30 Hz, and these are known to have stronger electrical signals in response to visual stimulation.

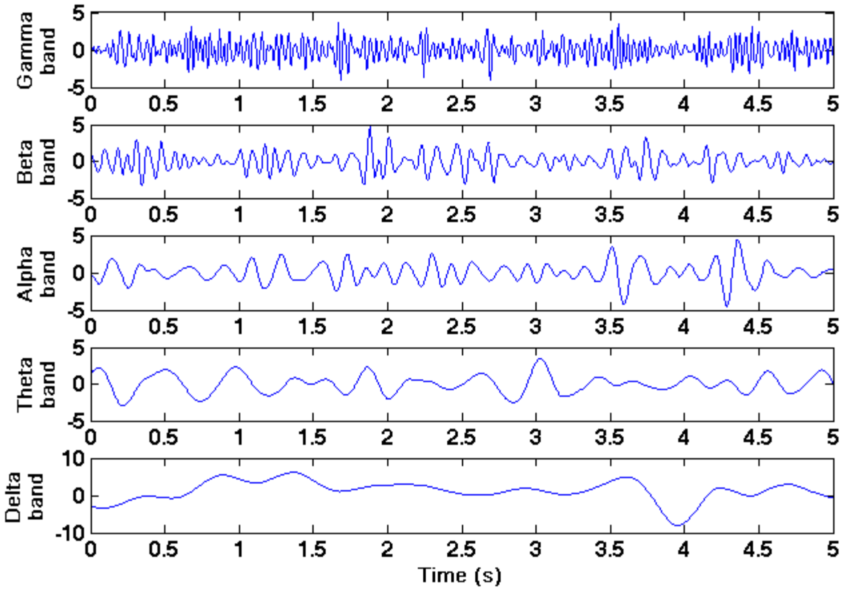


Figure 3.5: The five frequency bands of EEG signal. *Source*: [3]

3.3 Data collection

Technically speaking, data collection is the first step for doing research on EEG signals classification. Since datasets of EEG signals are widely used in brain studies, there exists a wide range of repositories recommended by literature and validated by scientific communities. In [30] are mentioned three interdependent components of a new research resource for complex physiologic signals, which are: PhysioBank, PhysioToolkit, and PhysioNet. Furthermore, another useful repository of databases is "UCI: Machine Learning Repository", which has 476 data sets as a service to the machine learning community [31]. Therefore, this important first step of data collection is already done thanks to the existing repositories that provide the databases needed for researches.

3.4 Machine Learning

When there is a computer problem to solve, an algorithm is necessary. This algorithm is developed based on the needed steps to achieve the correct solutions. Usually, for an algorithm, the input is given to obtain the hoped output. In this way, if you want to know the even numbers of a set, it is necessary to give the set of numbers as the input, and the output will be the even numbers from this set. However, there are problems in which an algorithm cannot be implemented so easily because the processes are not known. For example, when it is necessary to separate spam from other emails. The input will be all the emails and the output will be the spam, but there is not a known process to separate it. The algorithm to be implemented is difficult because the structure of the emails is similar, all have characters and images, so there is not a known specific feature to determine which is spam. However, it is possible to take a set of spam ones and analyze the data to “learn” how they are constituted. If that is possible, an algorithm can be created to more easily separate spam.

For this reason, machine learning is introduced in programming computers to optimize a performance criterion using example data or past experience. Machine learning uses training data or past experience to learn how data is constituted and to make a description or prediction of the input data [32]. It is clear that the output of a machine learning technique will not be 100% accurate, but if it is 90% or greater it can be used as well. Since machine learning helps systems to learn, it is also a part of artificial intelligence.

To classify EGG signals there are different machine learning techniques. Some of these can be more accurate than others. In this way, some of the most popular classifiers are: Support vector machine (SVM), Artificial Neural Network, Fuzzy Inference system, Clustering, K - nearest neighbor (K-NN), Bayesian Model, Linear Discriminant Analysis Classifier (LDC), Quadratic discriminant Analysis Classifier (QDC) [33], [34].

3.4.1 Pattern recognition

In the book [4], says that “the field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories”. In this sense, pattern recognition aims to simulate the cognitive capacity of humans to note the difference between a particular object from others, according to external information. Nonetheless, to understand the whole concept, it is necessary to know what is a pattern. In [35], Watanabe describes a pattern as an entity that can be and that is represented by a set of measured properties and the relationships between them. It is called the vector of characteristics. For example, in the recognition of handwriting digits, as it is showed in the Figure 3.6, each digit corresponds to a 28×28 pixel image and so can be represented by a vector $\mathbf{x}$ comprising 784 real numbers.

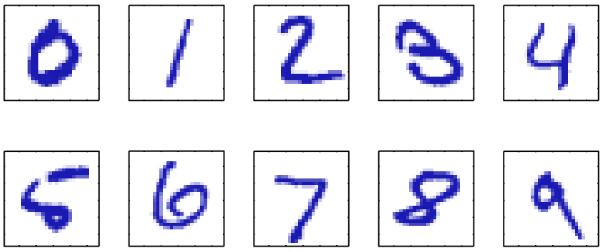


Figure 3.6: Hand-written digits taken from US zip codes. *Source*: [4]

According to how the training data is organized, there are two categories of a pattern recognition system [4], which are described as follows:

- **Supervised learning:** it happens when the training data has its input vector with the corresponding target vector. It means, that there exist defined classes.
- **Unsupervised learning:** when the training data set has its input data without the target vector. There are no defined classes.

3.5 Characterization

Characterization implies meanly to reduce the needed characteristics to describe a huge set of data. It is known as dimensionality reduction [25]. Furthermore, it is defined as the process to create a set of characteristics from the input data [32]. The goals of applying feature extraction will be: provide relevant sets of characteristics to the classifier, reduce redundancy, recover significant latent characteristics, generate greater understanding in the process of generating data, reduce computational costs and improve the rate of generalization [25].

Broadly, during the analysis of certain signals (herein, specifically the given data by the EEG database) is necessary to extract the relevant information for the subsequent classification task. In this sense, the decomposition, so-named, Discrete Wavelet Transform (DWT) is used. DWT is widely used and recommended for dealing with non-stationary and vibration-like signals -just as the EEG signals [36], [37].

3.5.1 Discrete Wavelet Transform (DWT)

In [38], mentions that the EEG signals are not stable because they can change rapidly in time. In this way, the Discrete Wavelet Transform allows modeling these changes in time-frequency.

There are a lot of studies developed under the discrete wavelet transform applied to EEG signals. Also, it is used in many engineering fields to solve real-life problems. According to [39], a wavelet is a short wave and it has its energy intensified in time to give a tool for the analysis of transient, non-stationary signals or time-varying phenomena. Fourier transforms can be applied easily if the signals were stationary, it means if they have not a big change over time. Nonetheless, the EEG signals are non-stationary, so the Fourier transform cannot be applied directly.

Wavelet transform allows getting the individual EEG sub-bands recon-structuring the information accurately. That is possible because DWT has the advantages of time-frequency localization, multi-rate filtering, and scale-space analysis. Therefore, DWT can show in more detail and precision the signals in both time and frequency domain. Also, at the first level, DWT decomposes a specific signal into approximation and detail coefficients [5]. These coefficients obtained has useful information. Furthermore, there is a filter bank

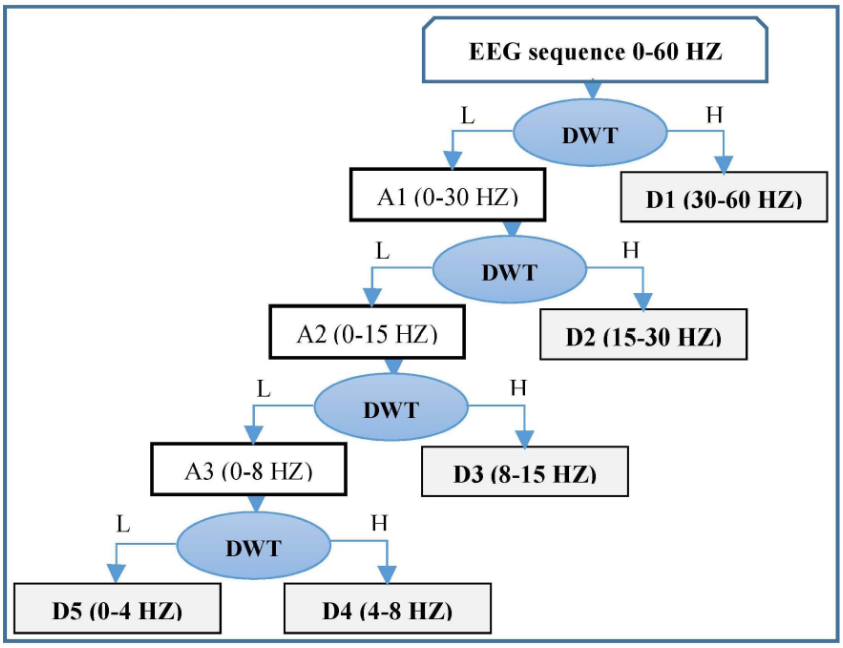


Figure 3.7: Four level wavelet decomposition of EEG. *Source*: [5]

associated with a Wavelet family and decomposition levels. Figure 3.7 shows an example of some levels of DWT. In [5], to achieve better results in feature extraction, wavelet decomposition has been used as a preprocessing level for EEG segments to extract five physiological EEG bands. Table 3.1 shows the levels of decomposition and the frequency range of each level.

Table 3.1: Levels of Discrete Wavelet Transform decomposition. *Source*: [5]

Frequency range (Hz)	Descomposition label	Frequency bands
D1	30 - 60	Gamma
D2	13 - 30	Beta
D3	8 - 13	Alpha
D4	4 - 8	Theta
D5	2 - 4	Delta
A1	0 - 2	Delta

3.6 Feature selection

The dataset of the EEG signal can have many characteristics, some of which are irrelevant. Therefore, these characteristics should be removed from the dataset. For this purpose, it is necessary to apply a feature selection to obtain only the needed features. This is an automatic selection of features from data, mostly founded in the columns of tabular data [40]. It is important to do not confuse feature selection and dimension reduction.

Although both are developed to remove features in the dataset, dimension reduction uses a combination of attributes to reduce it, but feature selection does not change the attributes in the process [40].

The importance of doing feature selection or extraction of a data set, based in machine learning and data mining processes, according to [25] are:

- Because there is irrelevant information, it means, some features can generate overlearning because they do not provide any information to the system.
- Due to redundant information, that is to say, linearly related features which do the same task.
- For the reason of dimensionality problem, in other words, the number of characteristics is greater than the data, meaning that each character represents a dimension.

3.7 Classification

It is the process that is done after performing correct feature extraction, where a feature is assigned to the corresponding class. Classification implies to group by classes the different attributes of a data set. For this process, in machine learning, there are two categories, supervised and unsupervised. A brief description and examples of both paradigms of classification are presented.

3.7.1 Supervised classification

Supervised classification handles the problem of automatically assigning objects to their respective classes on the basis of numerical measurements derived from these objects [41]. To do so, it is necessary to have already a set of classified data in order to assign a class to a second set. This classified set is used for training. In this regard, “training” is the task where the parameters used to classify are estimated for recognizing and label unknown attributes [42]. The classified data for training is previously labeled in order to group in the class that corresponds. Some of the most known supervised classifiers are:

- **k-Nearest Neighbor:** It is based on the belief that similar things exist in close proximity. It means similar things are near to each other [43]. k-NN is based on the idea that the new attributes will be categorized in the class to which the closest neighbors of the training set closest to this belong [44].
- **Linear Discriminant Analysis - LDC:** This classifier is based on the rule of maximum probability and the theorem of Bayesian to estimate probabilities. [45], [44].
- **Quadratic Discriminant Analysis - QDC:** It is closely related to LDC, that is to say, there is a variant from LDC, it is assumed that the measurements are

normally distributed. In this way, an individual covariance matrix is estimated for every class of observations [44].

- **Support Vector Machine - SVM:** It aims to find a hyperplane in an N-dimensional space, where N is the number of features, that plainly categorize the data points [45], [46]. This method can classify linearly separable data, is resistant to overfitting because it seeks a specific decision boundary and is efficient in the non-linear case because it does not explicitly create the transformed space and its non-linear transformation is implicit. Also, it can process a large number of entries [47].

- **Random Forest:** This method of classification is a combination of predictive trees such that each tree depends on the values of a randomly tested vector independently and with the same distribution for each of these [48].

- **Artificial Neural Networks - ANN:** This method is inspired by biological neural networks. Its main characteristics are self-organization and adaptability, its non-linear processing and parallel processing [49] [42].

3.7.2 Unsupervised classification

When the data is not labeled, unsupervised classification is used. It means, that this method tries to give a clustering based on the properties of data. It can separate classes but it can not give a name to these classes. It takes into consideration the similitude between the data, that is to say, it does not need prior knowledge. There are several methods to do so, but the most common is based on the use of clustering algorithms which refers to look for similar attributes and group them [42]. Some examples of unsupervised classifiers are:

- **k Means:** It is also called *migrating means* and *iterative optimization*. This method is based on determining the means of the classes, then in an iterative way, the objects are inserted in the nearest class using the minimal distance technique. Each iteration does a recalculation of the mean class and reclassifies all the objects. The process is repeated as many time as necessary until there are not more movements of objects between clusters. When a large data set is used, the process is not run to completion so it is necessary to use some stopping rules. [42] [50].

- **Isodata clustering:** It is a method based on k means with certain refinements. This method uses in the cluster formed a number of checks. The introduction of this number can be during or at the end of the iterative assignment process. These checks are in charge of make a relation between the number of objects assigned to clusters and their shapes in the spectral domain. [42] [51].

Chapter 4

Methodology

Regarding the general system explained in Figure 3.1 from Section 3.1.2, this work is focused in the last three stages.

The planned methodology for the characterization and classification of EEG signals with the intention to give a diagnosis for epilepsy seizures is summarized in Figure 4.1. In this Figure, there are 5 parts which are named as follows: pre-processing, decomposition, feature extraction, feature selection, and classification.

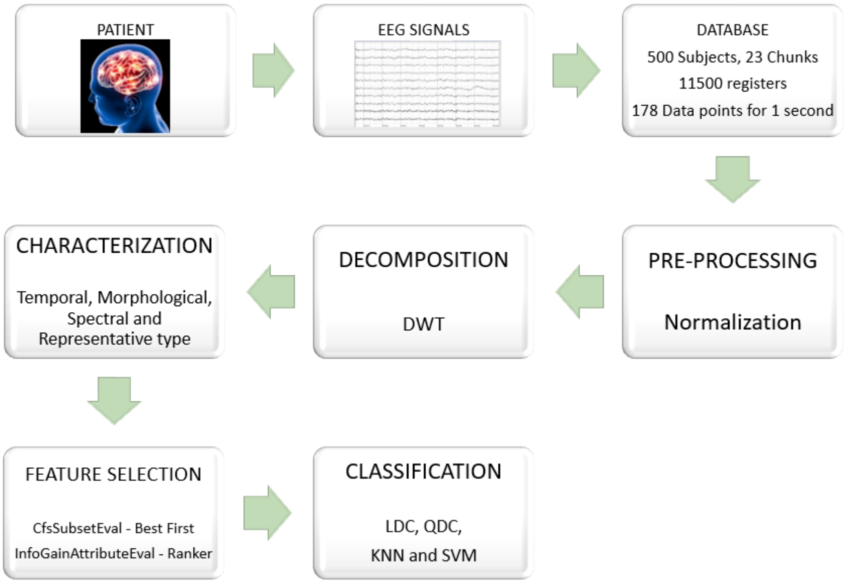


Figure 4.1: Methodology for the characterization and classification of EEG signal for epilepsy diagnosis.

4.1 Database

The EEG signals were obtained from UCI Machine Learning repository, the dataset used is “Epileptic seizure recognition” [52]. In this dataset, there is a total of 500 samples.

Each EEG signal has 4097 data points recorded in 23.6 seconds. However, this signal is divided into 23 chunks. Therefore, the dataset has 11500 records with 178 data points recorded in 1 second. The records were labeled as follows: (1) Seizure activity, (2) EEG signal from the area where the tumor was located, (3) EEG activity from the healthy brain area, (4) EEG signal when the eyes are closed, and (5) EEG signal when the eyes are opened. Figure 4.2 shows a plot of the 5 classes of the dataset marked each class by a different color.

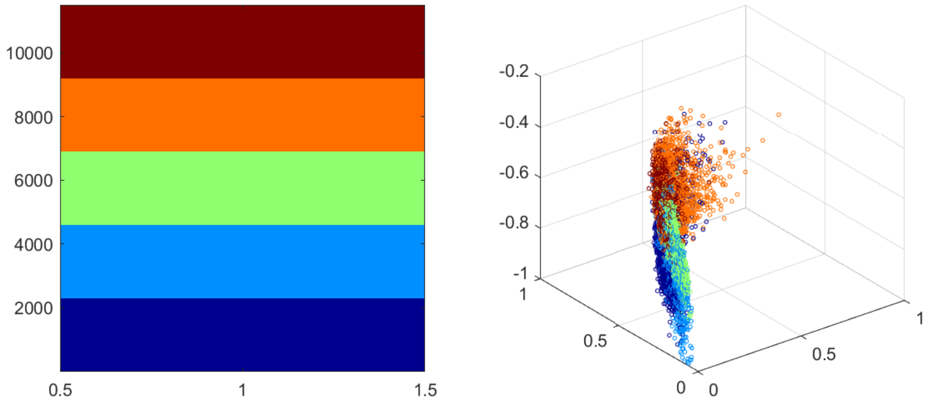


Figure 4.2: Plot of the labels from the dataset where each class is a different color. (1) purple, (2) blue, (3) green, (4) orange, (5) red

4.1.1 Epileptic Seizure Recognition Data Set

According to the UCI Machine learning repository [52], it says that "this dataset is a pre-processed and re-structured/reshaped version of a very commonly used dataset featuring epileptic seizure detection". The original data is found in [6], which consists of 5 different folders, each one with 100 files where each file represents a single person. Also, each file is a recording for 23.6 seconds. The corresponding time-series is sampled into 4097 data points. Therefore, the dataset is composed of 500 individuals with 4097 data points of 23.6 seconds each one.

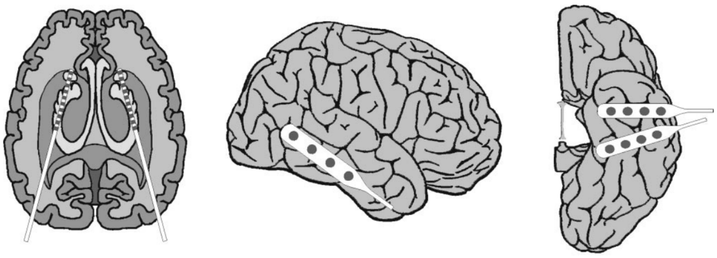


Figure 4.3: Scheme of intracranial electrodes implanted for pre-surgical evaluation of epilepsy patients. *Source*:[6]

This original dataset was divided and shuffled every 4097 data points into 23 chunks which contain 178 data points for 1-second [52]. It is the dataset that is used for this work, thereby it contains 11500 pieces of information which are the rows and 178 data points which are the columns. Additionally, it has the last column which represents the label (1,2,3,4,5). The labels as described as follows:

1. Seizure activity
2. EEG signal from the area where the tumor was located
3. EEG activity from the healthy brain area
4. EEG signal when the eyes are closed
5. EEG signal when the eyes are opened

These segments of 23.6-sec duration were selected and eliminated from continuous multichannel EEG recordings after pass visual inspection for devices, for example, due to muscle activity or eye movements [6]. The sets labeled as 4 and 5 were taken from the scalp that was realized on five healthy volunteers using the standardized 10-20 system scheme. These volunteers were relaxed and in an awake state. The rest of the sets 1, 2 and 3 were taken from presurgical diagnosis archive. These sets were obtained from five selected patients who had achieved complete seizure control after resection of one of the hippocampal formations which were diagnosed as epileptogenic zone [6]. In this way, set 1 only contains the seizure activity and set 2 were recorded from the hippocampal formation of the opposite hemisphere of the brain and the set 3 from within the epileptogenic zone. These last two contain activity measured during seizure-free intervals.

Figure 4.3 shows how the sets 1, 2 and 3 were obtained. To take these EEG signals depth electrodes were implanted symmetrically into the hippocampal formations (top). Segments of sets 2 and 3 were taken from all contacts of the respective depth electrode. Strip electrodes were implanted onto the lateral and basal regions (middle and bottom) of the neocortex. Segments of set 1 were taken from contacts of all depicted electrodes [6].

In [6] mentions one last important thing about the dataset. It says that “EEG signals were recorded with the same 128-channel amplifier system, using an average common reference or strong eye movement artifacts. After 12 bit analog-to-digital conversion, the data were written continuously onto the disk of a data acquisition computer system at a sampling rate of 173.61 Hz”.

4.2 Pre-processing

The EEG signals should be normalized before they will be processed. The normalization is simple, it has to establish the data in a range of $[-1, 1]$. It will be removed offset levels regarding equation (4.1). In this way, the normalization makes sure that the signal does not exceed the given range. It makes easier the processing to follow in the next steps.

$$S = \frac{S - \bar{S}}{max\,|S|} \tag{4.1}$$

Where S is the signal, $max\,|S|$ is the maximum absolute value of the signal, and $\bar{S}$ is the mean of the signal.

The programming for loading the data and doing the normalization process is shown in the appendix A.1.

4.3 Signals decomposition

In order to achieve the objectives proposed in this work, the implementation of a method of decomposition of the signal is needed. It is described below.

4.3.1 Discrete Wavelet Transform (DWT)

In this research, the algorithm of DWT is used as a method of decomposition for the signal. The family of DWT used was Daubechies. This family is based on compactly supported orthonormal wavelets. The order used was 4 and the decomposition level was set to be 5. The order refers to the number of vanishing moments [53]. The application Matlab is used to perform this algorithm.

Discrete Wavelet Transform (DWT) decomposes in a recurrent way a signal into two subsignals with less resolution regarding the frequency. These signals are called as approach and detail, also known as coefficients [54]. The signals $S_i(n)$ and $W_i(n)$ are the approach and detail signals in the i level. The total number of levels depends on the times that the signals are going to be decomposed. In this way, the process will be repeated as many times as necessary, so the signal $S_i(n)$ is decomposed in new others signals $S_{i+1}(n)$ and $W_{i+1}(n)$. These are going to be the new approach and details signals in the $i+1$ level. The equation (4.2) shows how the approach signal can be computed.

$$S_{i+i} = \sum_k g(k) S_1(2n - k) \tag{4.2}$$

The detail signal can be calculated as shows the equation (4.3)

$$W_{i+i} = \sum_k h(k) S_1(2n - k) \tag{4.3}$$

Matlab has an implemented function named `wavedec` which is used to obtained the levels of DWT from the dataset used for this research. The Matlab programming is shown in the appendix A.2 in the source code 4.

4.4 Characterization

EEG signals samples can be treated as a complex and dynamic dataset that can be represented by a feature set. It is possible because EEG signals are extracted as a record of electrical signals from the electric potential generated. Taking into account previous works of EEG signal [25], [55], [56], [57], [58], two sets of several features are taking into consideration. The first one is given by temporal features as follows: absolute mean value, standard deviation, kurtosis, the area under the curve, root mean square (Rms), variance, covariance, entropy, simple quadratic integral, and Shannon entropy, and others [59], [60], [61] . The second one is given by spectral transformation: peak frequency, average frequency and maximum energy of the spectral power [62]. A brief list of different used estimators to characterize the EEG signals is shown in Table 4.1.

Table 4.1 is stated according to the following notation:

- $\bar{x}$: Conventional average.
- $\sigma(x)$: Standard deviation.
- $p(x)$: Given probability.
- $|x|$: Absolute value.
- $log(x)$: Logarithm of x to the base 10.
- $log_2(x)$: Logarithm of x to the base 2.

These features mentioned above were applied directly to the normalized EEG signal. Furthermore, these features were applied to the 5 levels of DWT signals obtained after decomposition. Therefore, the features set is as follows:

i) 36 features were obtained from statistical measures

ii) 192 features were obtained from resultant coefficients of DWT which were pass through 32 statistical measures.

iii) 7 features were removed from the features set because they have wrong values.

Each feature vector is normalized through the equation (4.1) which is the same used for the normalization of the original signal. The final matrix with all features has a dimension of 11500 × 221. Where 221 is the total number of features.

The Matlab programming for the feature extraction is shown in the appendix A.2 in the source code 2 and 3.

Table 4.1: Mathematical formulation of the representative characteristics for EEG

Features	**Mathematical formulation**
Area under the curve	$I = \sum_{n=1}^{N} \lvert x_n \rvert$
Mean	$\bar{x} = \frac{1}{N} \sum_{n=0}^{N} \lvert x_n \rvert$
Root mean square	$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (x_n)}$
Variance	$\frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2$
Standard Deviation	$\sqrt{\frac{\sum_{n=1}^{N} (x_n - \bar{x})^2}{N-1}}$
Log energy entropy	$\sum_{x \in A} log(x^2)$
Squared Integral	$I = \sum_{n=1}^{N} (x_n)^2$
Kurtosis	$\frac{\frac{\sum_{n=1}^{N} (x_n - \bar{x})^2}{N}}{\sigma(x)^4}$
Covariance	$\frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$
Shannon Entropy	$- \sum_{X \in A} p(x) log_2 p(x)$
Average Amplitude Change	$\frac{1}{N} \sum_{n=1}^{N} \lvert x_{n+1} - x_n \rvert$

4.5 Feature selection

The features obtained from the original EEG signal form high-dimensional matrices which have a lot of features that are not very relevant at the time to represent the class you want to classify. On the contrary, these features are causing a bad representation of the class, thus the system is not efficient.

In this work, there are 221 features obtained from characterization. In order to extract the most relevant features, Weka application is used. It is a powerful tool for data mining tasks. It has a big collection of machine learning algorithms. For this research, Weka is used to do the feature selection. In this way, the whole dataset is being passed by

CfsSubsetEval as an attribute evaluator. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred [63]. This attribute evaluator works together with the search method BestFirst. According to [64], BestFirst is an artificial intelligence search strategy that allows backtracking along the search path. It works in the same way that greedy hill climbing, it means, BestFirst is going to make local changes in the current features subset while it is moving along the search space. In contrast with hill climbing, if BestFirst found that this path is no longer promising, it can go back to the best previous subset and continue with the searching.

After using BestFirst, to take just the best of this subset of features, it is useful to use another attribute evaluator. In this case, InfoGainAttributeEval is used. It evaluates the worth of an attribute by measuring the information gain with respect to the class [65]. Then, the search method used is Ranker, which ranks attributes by their individual evaluations [66].

The subset obtained by BestFirst was of 41 features. After applying the Ranker method this subset is reduced to the 9 best, which are the features used for this work. It is important to mention that the best-taken features belong to the subset applied to the decomposition of DWT. At the end of this process, a matrix or subset of data with the selected features is obtained. Also, a vector with the regarding labels is obtained. A representation on that is shown as follows:

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^D \\ x_2^1 & x_2^2 & \cdots & x_2^D \\ \vdots & \vdots & \ddots & \vdots \\ x_N^1 & x_N^2 & \cdots & x_N^D \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix},$$

where N is the number of records and D is the number of selected features.

4.6 Classification

For the classification process the following classifiers are used:

- Linear discriminant analysis classifier (LDC)
- Quadratic discriminant analysis classifier (QDC)
- k-nearest neighbor (kNN)
- Support vector machine (SVM)

This is applied to bi-class problems and 3-class problems with the 9 features obtained in the selection feature process. The Matlab programming for these two kinds of classification problem is shown in the appendix A.3.

Chapter 5

Experimental Setup

In this Chapter, the performance measures of the classifiers applied to the different sets are going to be described.

5.1 Performance metrics

The following performance measures were used to qualify the result of the classification: Mean (Me), Standard Deviation (Std), Sensitivity (Se), Specificity (Sp) and Classification percentage (CP). These measures are calculated with the equations:

$$Me = \frac{\sum_{i=1}^{n} \mathbf{er}_i}{n} \tag{5.1}$$

$$Std = \sqrt{\frac{\sum_{i=1}^{n} (\mathbf{er}_i - \bar{\mathbf{er}})^2}{n-1}} \tag{5.2}$$

$$Se = \frac{Tp}{Tp + Fn} \tag{5.3}$$

$$Sp = \frac{Tn}{Tn + Fp} \tag{5.4}$$

$$CP = \frac{Tn + Tp}{Tn + Fp + Tp + Fn} \times 100 \tag{5.5}$$

The values to calculate these measures are:

- **er**: Errors vector with indexes $i \cdots n > 0$
- n: Length of the vector **er**
- Tp: True positives or cases of interest class correctly classified.
- Tn: True negatives or different cases of the interest class correctly classified.

- Fp: False positives or different cases of the interest class classified as cases of the interest class.
- Fn: False negatives or cases of the interest class classified as different cases of the interest class.

In addition, to have a better visual appreciation, the following graphs will be shown: box plot, and receiver operating characteristic (ROC) curve. Furthermore, confusion-matrix-based measures are presented. These quantifier performance measures will be presented in each experiment.

5.2 Applied Tests on Database

In the proposed methodology the decomposition of the signal is explained. Also, there are two kinds of characterization: on the original signal and on the decomposed signal. All these obtained features are put together in a matrix. In this sense, this feature matrix is used to do 5 tests. The most important class for this research is class 1 which is the indicator of seizure activity. Therefore, the 5 classes of the original dataset were restructured to work with 2 and 3 classes. In order to achieve the best classification of class 1, two tests are a bi-class problem and the three others are a 3-class problem. These 5 tests are useful to analyze the response of four classifiers: LDA, LDC, kNN, and SVM. For each test, four percentual results are obtained which indicate the number of well-classified samples over the total of samples in the classifier by each class. As methods to feature selection first were applied `BestFirst`, then over that was applied `Ranker`.

- **Experiment 1:** It is done using classes 2, 3, 4 and 5 as a single class. Class 1 is the target for classification
- **Experiment 2:** It is done using only class 1 and class 2. The rest of the classes are removed from the dataset. Here the seizure activity and the area where the tumor was located can be distinguished.
- **Experiment 3:** It is done using classes 1 and 2 individually and the rest of the classes are taken as a single class. Here the seizure activity and the area where the tumor was located can be distinguished.
- **Experiment 4:** It is done using the classes 1, 2 and 3 individually. The rest of the classes are removed from the dataset. Class 3 can distinguish the healthy brain area.
- **Experiment 5:** It is done joining classes 1 and 2 as a single class. Class 3 is treated individually and the classes 4 and 5 are joined as another single class.

For all these experiments, the classifiers were applied under 10 iterations using 80% of the data for training and 20% for test. The registered results are those obtained with runnings over the test dataset only. Finally, a user interface is developed, which allows to redo the experiments and recreate the obtained results.

5.3 Classifier settings

For applying the four classifiers is used Matlab and its toolbox called `PRTools`. In this regards, the functions used from `PRTools` are `ldc()`, `qdc()`, and `knnc()`. SVM classifier is imported from the classification learner app of Matlab under the name of `trainClassifierSVM()`. To use these functions, it is necessary to set a configuration showed in Table 5.1.

Table 5.1: Classifiers settings

Classifier	Settings
LDC	No regularization All dimensions
QDC	No regularization All dimensions
KNN	k is optimized with respect to the leave-one-out error on the dataset
SVM	Kernel function: Gaussian Kernel scale: 3 Box constraint: 1 Standardize: true Specify the class names

Chapter 6

Results

The results of each one of the experiments mentioned in Section 5.2 are presented through figures and tables to have a better visual appreciation. In this regard, the tables have statistical measures based on the confusion matrix.

6.1 Experiments

There are a total of 5 experiments in which their results are shown in the following subsections.

6.1.1 Experiment 1

This experiment is developed with 2 classes: (1) Seizure activity and (0) a single class grouping the rest of the classes. Table 6.1 shows the mean and standard deviation from the errors obtained by each classifier. It is clear that the classifier with lower error is SVM, followed by kNN. Furthermore, the standard deviation is low for all classifier with means that the data is almost accurate.

Table 6.2 shows the sensitivity and specificity belongs to the class 1, seizure activity, where the higher sensitivity and specificity is presented with the SVM classifier. Tables 6.3, 6.4, 6.5, 6.6, show the confusion matrices of all classifiers taken in the last iteration. The confusion matrix of SVM shows the best classification followed by kNN.

Figure 6.1 shows the box plots of all classifiers measuring their accuracy of classification. It is clear, that the best is the SVM which is over 95% followed nearly by kNN. Figure 6.2 shows the comparison of the ROC curves of all classifiers. It is noticed that the SVM and kNN curves are so close and are the best of all of them.

Additionally, a significance test of error rates for experiment 1 is presented in Figure 6.3. The test used is Dunn test (Kruskal-Wallis with bonferroni correction).

Table 6.1: Experiment 1 - Comparison of classifier through mean and standard deviation measures.

Classifier	Mean	Standard Deviation.
LDC	0.0613043	0.0031752
QDC	0.0651739	0.00211465
KNN	0.0489565	0.00389529
SVM	0.0399565	0.00385327

Table 6.2: Experiment 1 - Comparison of classifier through sensitivity and specificity measures from class 1, seizure activity

Classifier	Sensitivity	Specificity
LDC	0.78913	0.970652
QDC	0.815217	0.9625
KNN	0.795652	0.982065
SVM	0.854348	0.983152

Table 6.3: Experiment 1 - Confusion matrix of LDC classifier

True Labels	Estimated Labels		Totals
	0	1	
0	1786	54	1840
1	97	363	460
Totals	1883	417	2300

Table 6.4: Experiment 1 - Confusion matrix of QDC classifier

True Labels	Estimated Labels		Totals
	0	1	
0	1771	69	1840
1	85	375	460
Totals	1856	444	2300

Table 6.5: Experiment 1 - Confusion matrix of kNN classifier

True Labels	Estimated Labels		Totals
	0	1	
0	1807	33	1840
1	94	366	460
Totals	1901	399	2300

Table 6.6: Experiment 1 - Confusion matrix of SVM classifier

True Labels	Estimated Labels		Totals
	0	1	
0	1809	31	1840
1	67	393	460
Totals	1876	424	2300

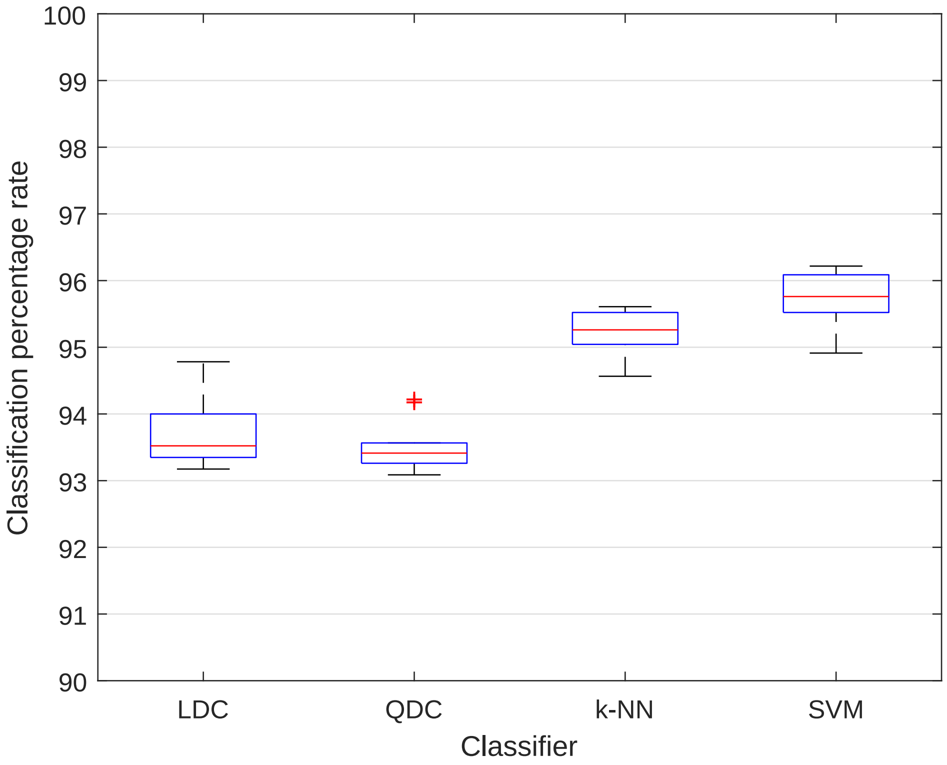


Figure 6.1: Experiment 1 - Comparison of classifiers through their accuracy

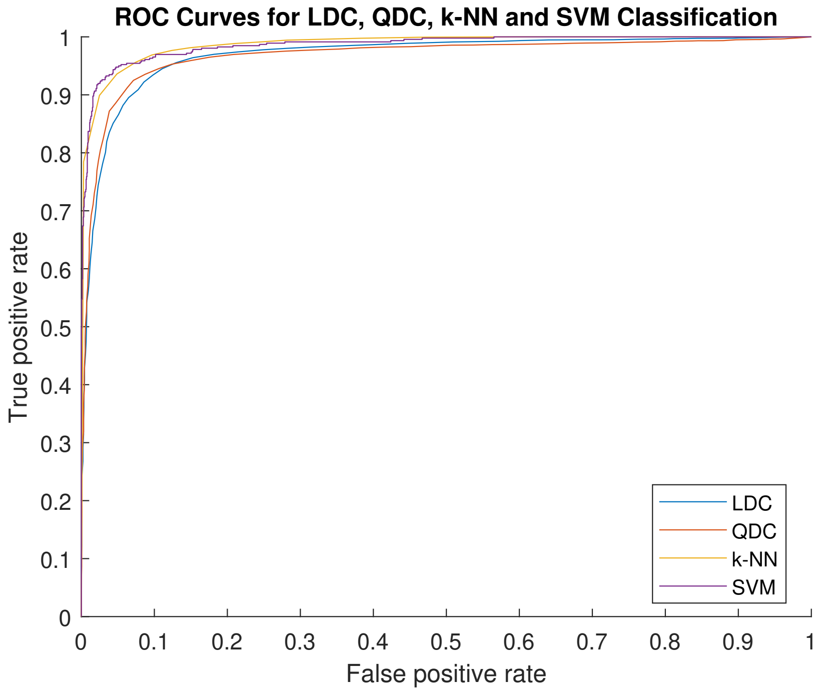


Figure 6.2: Experiment 1 - ROC curve of each classifier belonging to class 1, seizure activity

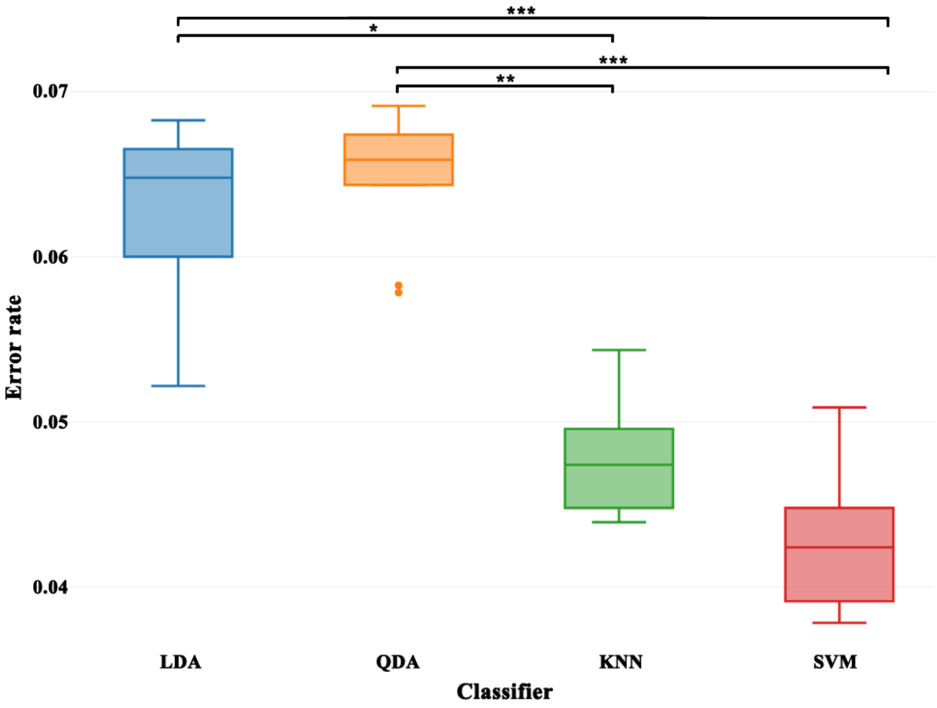


Figure 6.3: Experiment 1 - Comparison of classifiers. $* <= 0.05, ** <= 0.001, *** <= 0.0001$.

6.1.2 Experiment 2

This experiment is developed with 2 classes: (1) Seizure activity and (0) EEG signal from the area where the tumor was located. The rest of the classes were removed from the dataset. Table 6.7 shows the mean and standard deviation from the errors obtained by each classifier. It is clear that the classifier with lower error is SVM, followed by *k*NN. Furthermore, the standard deviation is low for all classifier with means that the data is almost accurate. Table 6.2 shows the sensitivity and specificity belongs to class 1, seizure activity, where the higher sensitivity and specificity is presented with the SVM classifier. In this experiment, if we change the chosen class by class 0 we can obtain the specificity as the sensitivity and vice versa. It indicates that class 0 is also well classified. It is useful to determine the epilepsy source location in the brain region as it is mentioned in Section 3.1.3 because the treatment of neurosurgery can be applied. Tables 6.9, 6.10, 6.11, 6.12, show the confusion matrices of all classifiers taken in the last iteration. The confusion matrix of SVM shows the best classification followed by *k*NN and LDC.

Table 6.7: Experiment 2 - Comparison of classifier through mean and standard deviation measures.

Classifier	Mean	Standard Deviation.
LDC	0.0697826	0.0088572
QDC	0.128478	0.00967873
KNN	0.0578261	0.00906232
SVM	0.0498913	0.00513548

Figure 6.4 shows the box plots of all classifiers measuring their accuracy of classification. It is clear, that the best is the SVM which is near to 95% followed nearly by KNN. The worst of them is QDC. The Figure 6.5 shows the comparison of the ROC curves of all classifiers belonging to class 1. It is noticed that the SVM and kNN curves are so close and are the best of all of them. Besides, Figure 6.6 shows the same curves but they belong to class 0, where it is clear that SVM and KNN are the best. The difference between these two last Figures is the QDC classifier which in Figure 6.6 is the worst curve.

Table 6.8: Experiment 2 - Comparison of classifier through sensitivity and specificity measures from class 1, seizure activity

Classifier	Sensitivity	Specificity
LDC	0.923913	0.915217
QDC	0.786957	0.941304
KNN	0.921739	0.93913
SVM	0.958696	0.958696

Additionally, a significance test of error rates for experiment 2 is presented in Figure 6.7. The test used is Dunn test (Kruskal-Wallis with bonferroni correction).

Table 6.9: Experiment 2 - Confusion matrix of LDC classifier

True Labels	Estimated Labels		Totals
	0	1	
0	421	39	460
1	35	425	460
Totals	456	464	920

Table 6.10: Experiment 2 - Confusion matrix of QDC classifier

True Labels	Estimated Labels		Totals
	0	1	
0	433	27	460
1	98	362	460
Totals	531	389	920

Table 6.11: Experiment 2 - Confusion matrix of kNN classifier

True Labels	Estimated Labels		Totals
	0	1	
0	432	28	460
1	36	424	460
Totals	468	452	920

Table 6.12: Experiment 2 - Confusion matrix of SVM classifier

True Labels	Estimated Labels		Totals
	0	1	
0	441	19	460
1	19	441	460
Totals	460	460	920

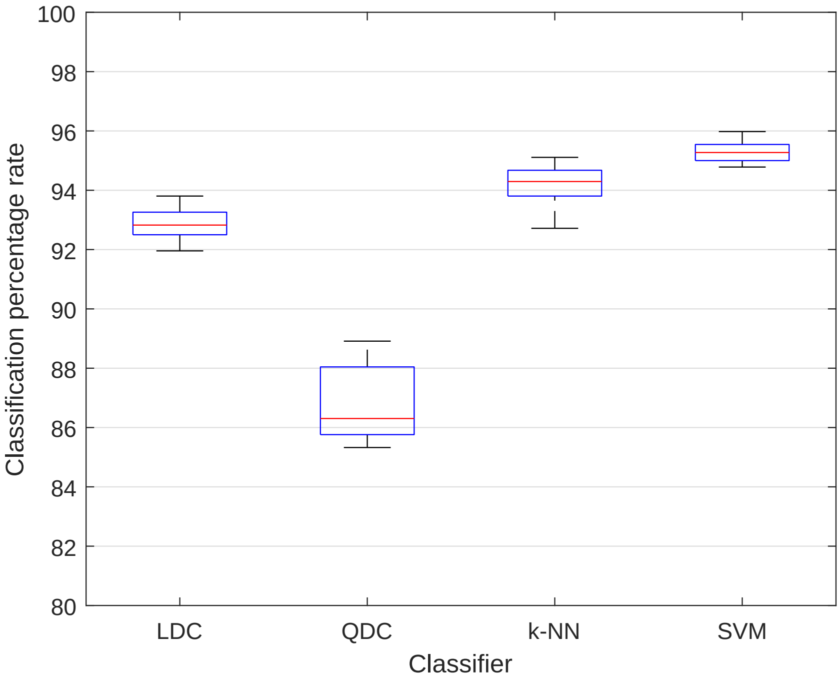


Figure 6.4: Experiment 2 - Comparison of classifiers through their accuracy

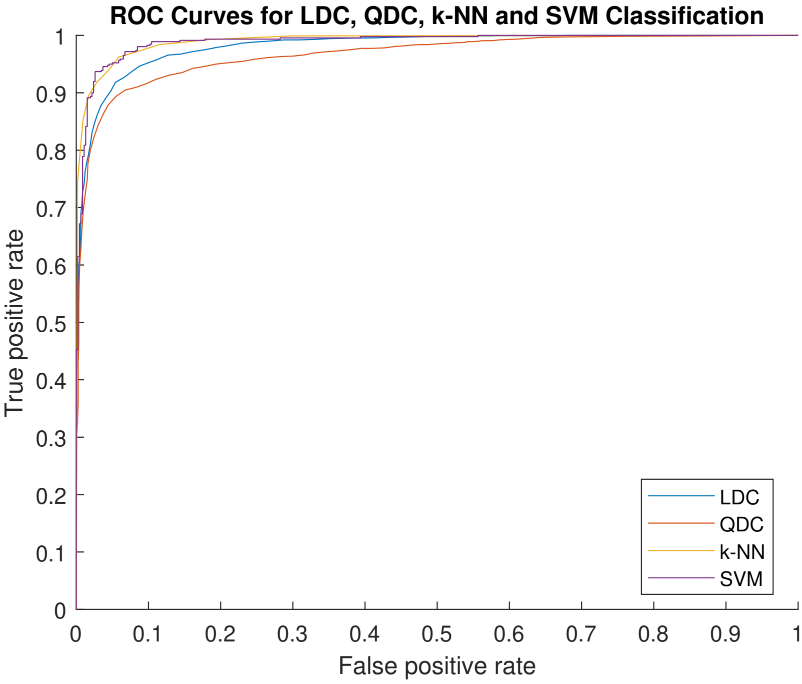


Figure 6.5: Experiment 2 - ROC curve of each classifier belonging to class 1, seizure activity

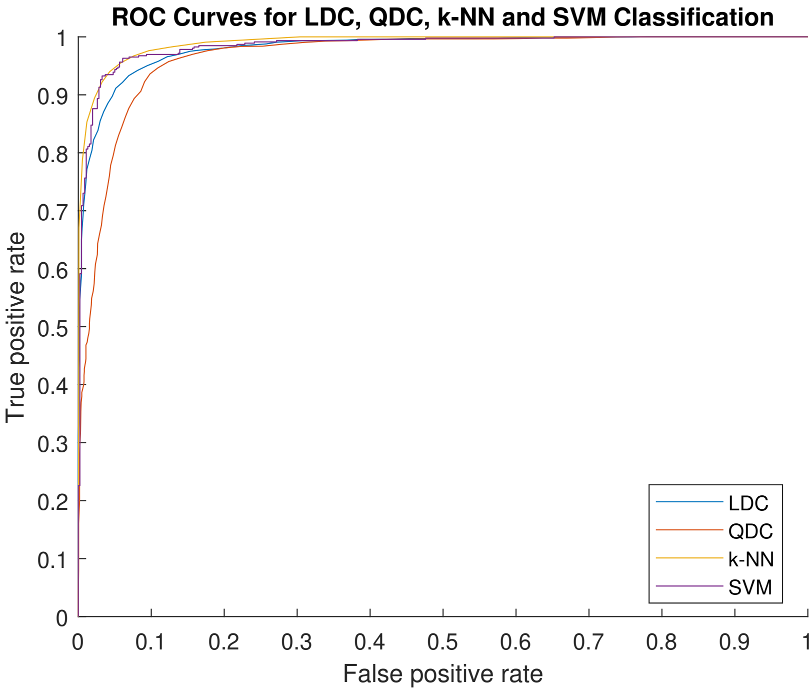


Figure 6.6: Experiment 2 - ROC curve of each classifier belonging to class 2, epilepsy source location

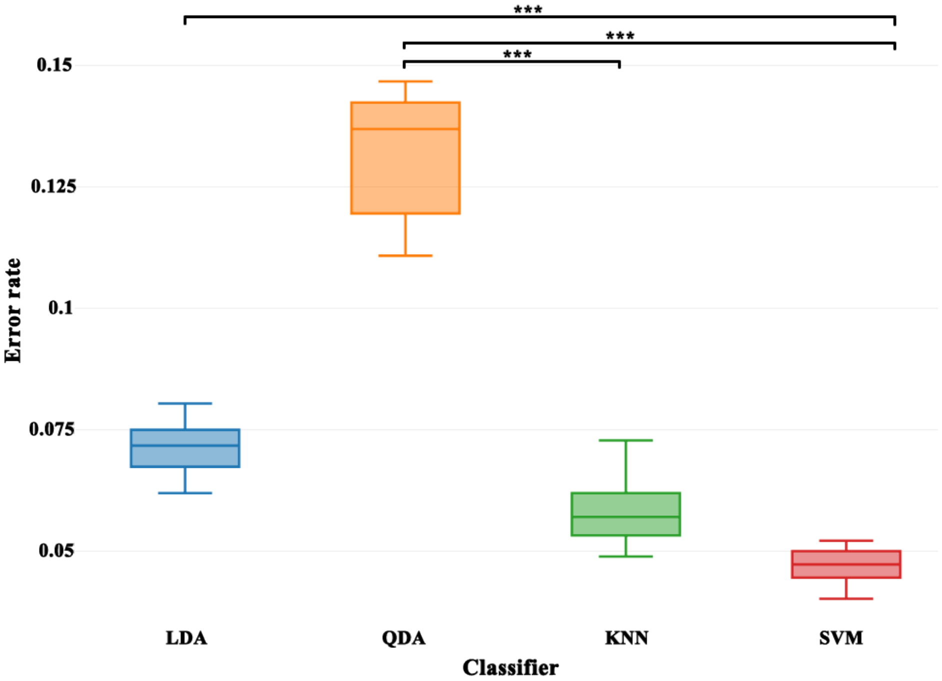


Figure 6.7: Experiment 2 - Comparison of classifiers. $* <= 0.05, ** <= 0.001, *** <= 0.0001$.

6.1.3 Experiment 3

This experiment is developed with 3 classes: (1) Seizure activity, (2) EEG signal from the area where the tumor was located and (3) a single class grouping the rest of classes. Table 6.13 shows the mean and standard deviation from the errors obtained by each classifier. It is clear that the classifier with lower error is SVM, followed by kNN. Furthermore, the standard deviation is low for all classifier with means that the data is almost accurate. Nonetheless, the error is great, that is to say, the possibility of the system fails is over 20% for all the classifiers. It is not good for health works.

Table 6.13: Experiment 3 - Comparison of classifier through mean and standard deviation measures

Classifier	Mean	Standard Deviation.
LDC	0.239652	0.0072051
QDC	0.340522	0.00691959
KNN	0.224783	0.00425998
SVM	0.205261	0.00600864

Table 6.14 shows the sensitivity and specificity belongs to class 1 and 2, seizure activity and source area location, where, in the case of class 1, the higher sensitivity and specificity is presented with the SVM classifier. On the other side, for class 2 the values vary, the higher sensitivity is in QDC and the specificity in KNN. It means that class 1 is better classified than class 2. Tables 6.15, 6.16, 6.17, 6.18, show the confusion matrices of all classifiers taken in the last iteration. It is noticed that class 2 is not well classified in all of them.

Table 6.14: Experiment 3 - Comparison of classifier through sensitivity and specificity measures belonging to class 1 and 2, seizure activity and source area location

Classifier	Class 1		Class 2	
	Se	Sp	Se	Sp
LDC	0.821739	0.969565	0.365217	0.909239
QDC	0.713043	0.976087	0.895652	0.631522
KNN	0.832609	0.97337	0.397826	0.909783
SVM	0.865217	0.978804	0.23913	0.956522

Figure 6.8 shows the box plots of all classifiers measuring their accuracy of classification. It is clear, that the best is the SVM which is near to 80% followed nearly by kNN. The worst of them is QDC. However, this is not enough for saying that this experiment worked good. The percentage is not suitable for this kind of research. Figure 6.9 shows the comparison of the ROC curves of all classifiers belonging to class 1. It is noticed that

the SVM and kNN curves are so close and are the best of all of them. Besides, Figure 6.6 shows the same curves but they belong to class 2, where it is clear that this class cannot be correctly classified.

Additionally, a significance test of error rates for experiment 3 is presented in Figure 6.11. The test used is Dunn test (Kruskal-Wallis with bonferroni correction).

Table 6.15: Experiment 3- Confusion matrix of LDC classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	378	14	68	460
2	38	168	254	460
3	18	153	1209	1380
Totals	434	335	1531	2300

Table 6.16: Experiment 3 - Confusion matrix of QDC classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	328	99	33	460
2	11	412	37	460
3	33	579	768	1380
Totals	372	1090	838	2300

Table 6.17: Experiment 3 - Confusion matrix of kNN classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	383	15	62	460
2	32	183	245	460
3	17	151	1212	1380
Totals	432	349	1519	2300

Table 6.18: Experiment 3 - Confusion matrix of SVM classifier

True Labels	Estimated Labels 1	2	3	Totals
1	398	27	35	460
2	20	110	330	460
3	19	53	1308	1380
Totals	437	190	1673	2300

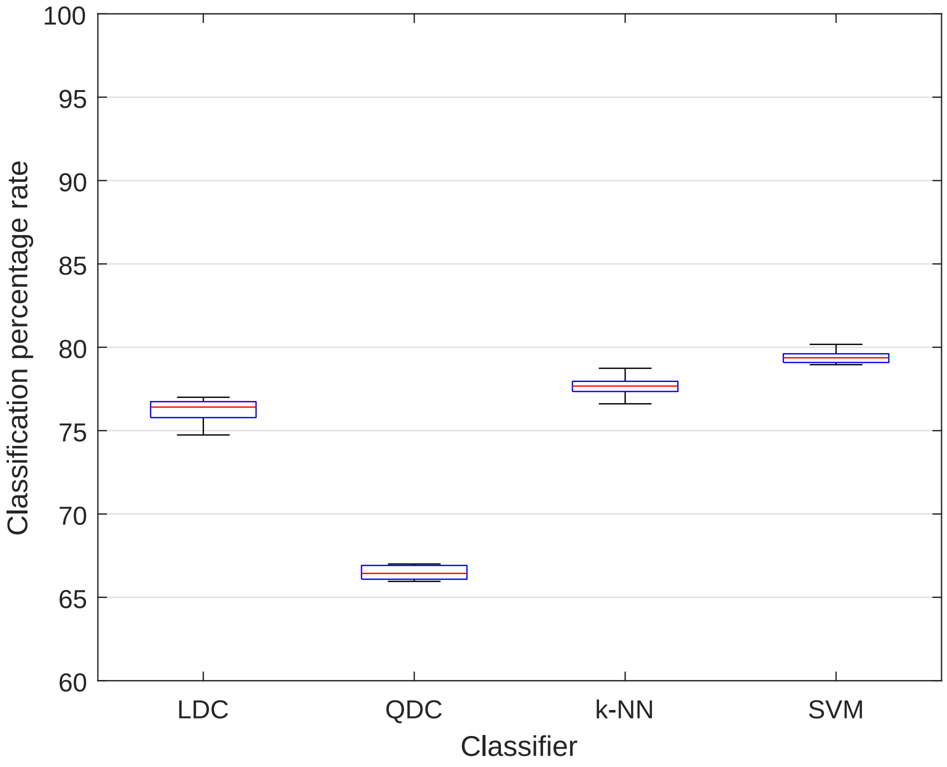


Figure 6.8: Experiment 3 - Comparison of classifiers through their accuracy

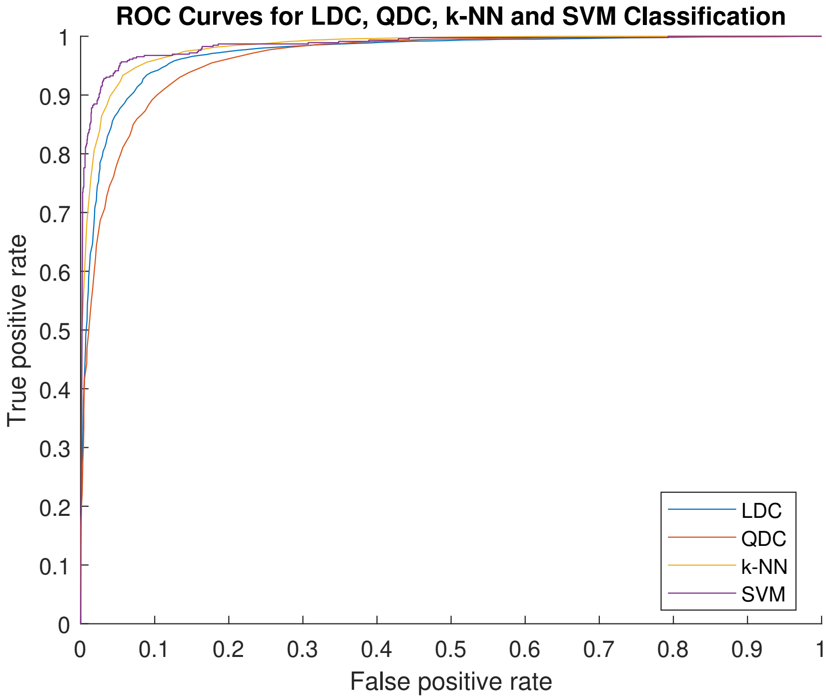


Figure 6.9: Experiment 3 - ROC curve of each classifier belonging to class 1, seizure activity

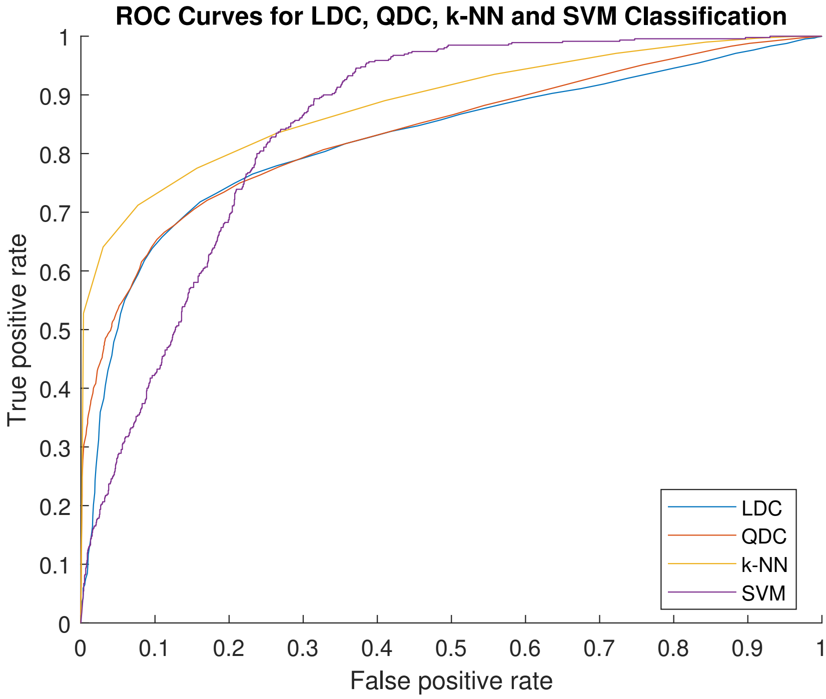


Figure 6.10: Experiment 3 - ROC curve of each classifier belonging to class 2, epilepsy source location

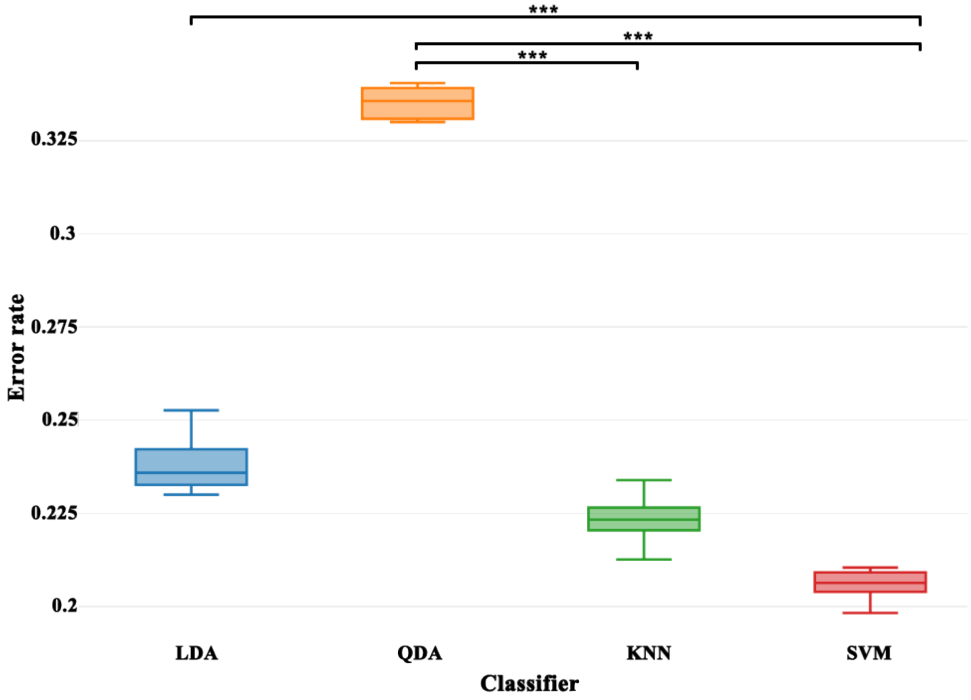


Figure 6.11: Experiment 3 - Comparison of classifiers.
$* <= 0.05, ** <= 0.001, *** <= 0.0001$.

6.1.4 Experiment 4

This experiment is developed with 3 classes: (1) Seizure activity, (2) EEG signal from the area where the tumor was located and (3) healthy brain area. The rest of the classes were removed from the dataset. Table 6.19 shows the mean and standard deviation from the errors obtained by each classifier. It is clear that the classifier with lower error is SVM, followed by kNN. Furthermore, the standard deviation is low for all classifier with means that the data is almost accurate. Nonetheless, the error is great, that is to say, the possibility of the system fails is over 29% for all the classifiers. It is not good for health works.

Table 6.19: Experiment 4 - Comparison of classifier through mean and standard deviation measures

Classifier	Mean	Standard Deviation.
LDC	0.320725	0.0140355
QDC	0.358333	0.0109378
KNN	0.322246	0.0133269
SVM	0.291232	0.00823703

Table 6.20 shows the sensitivity and specificity belonging to class 1, 2 and 3 which are seizure activity, source area location and healthy brain area, where, in the case of class

1, the higher sensitivity is in SVM and specificity is in QDC. On the other side, for class 2 the values vary, the higher sensitivity is in LDC and the specificity is in SVM. On the other hand, for class 3, the greater sensitivity is in QDC and the specificity is in LDC. It means that class 1 is better classified than class 2 and 3. Tables 6.21, 6.22, 6.23, 6.24, show the confusion matrices of all classifiers taken in the last iteration. It is noticed that class 2 is the worst classified in all of them and the best is class 1.

Table 6.20: Experiment 4 - Comparison of classifier through sensitivity and specificity measures of all classes

Classifier	Class 1		Class 2		Class 3	
	Se	Sp	Se	Sp	Se	Sp
LDC	0.936957	0.966304	0.465217	0.78913	0.630435	0.76087
QDC	0.76087	0.975	0.330435	0.806522	0.819565	0.673913
KNN	0.915217	0.968478	0.4	0.81087	0.682609	0.719565
SVM	0.95	0.963043	0.384783	0.881522	0.78913	0.717391

Figure 6.12 shows the box plots of all classifiers measuring their accuracy of classification. It is clear, that the best is the SVM which is near to 70% followed nearly by kNN. The worst of them is QDC. However, this is not enough for saying that this experiment worked good. The percentage is not suitable for this kind of research. Figure 6.13 shows the comparison of the ROC curves of all classifiers belonging to class 1. It is noticed that the SVM, kNN, and LDC curves are so close and are the best of all them. Besides, Figure 6.14 shows the same curves but they belong to class 2, where it is clear that this class cannot be correctly classified. Moreover, Figure 6.15 shows the same curves but they belong to class 3. It is noticed that SVM curve has a totally different behavior regarding the rest of the classifiers. However, it is clear that this class cannot be classified either.

Table 6.21: Experiment 4 - Confusion matrix of LDC classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	431	28	1	460
2	27	214	219	460
3	4	166	290	460
Totals	462	408	510	1380

Additionally, a significance test of error rates for experiment 4 is presented in Figure 6.16. The test used is Dunn test (Kruskal-Wallis with bonferroni correction).

Table 6.22: Experiment 4 - Confusion matrix of QDC classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	350	99	11	460
2	19	152	289	460
3	4	79	377	460
Totals	373	330	677	1380

Table 6.23: Experiment 4 - Confusion matrix of kNN classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	421	31	8	460
2	26	184	250	460
3	3	143	314	460
Totals	450	358	572	1380

Table 6.24: Experiment 4 - Confusion matrix of SVM classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	437	17	6	460
2	29	177	254	460
3	5	92	363	460
Totals	471	286	623	1380

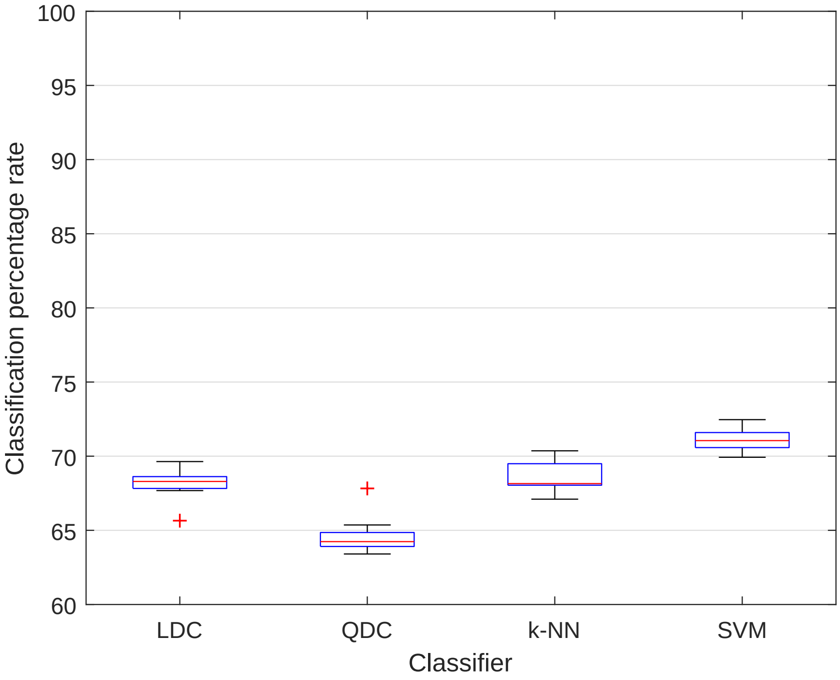


Figure 6.12: Experiment 4 - Comparison of classifiers through their accuracy

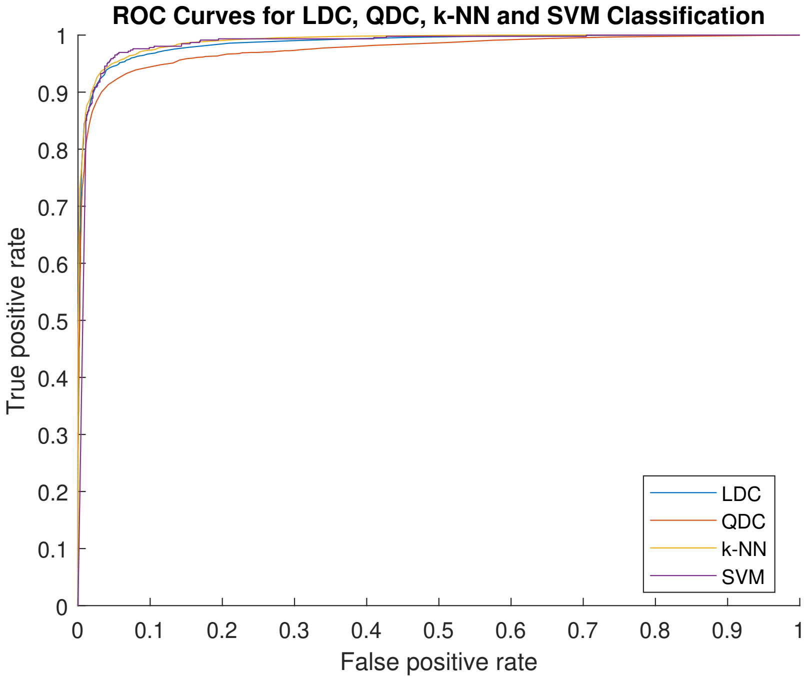


Figure 6.13: Experiment 4 - ROC curve of each classifier belonging to class 1, seizure activity

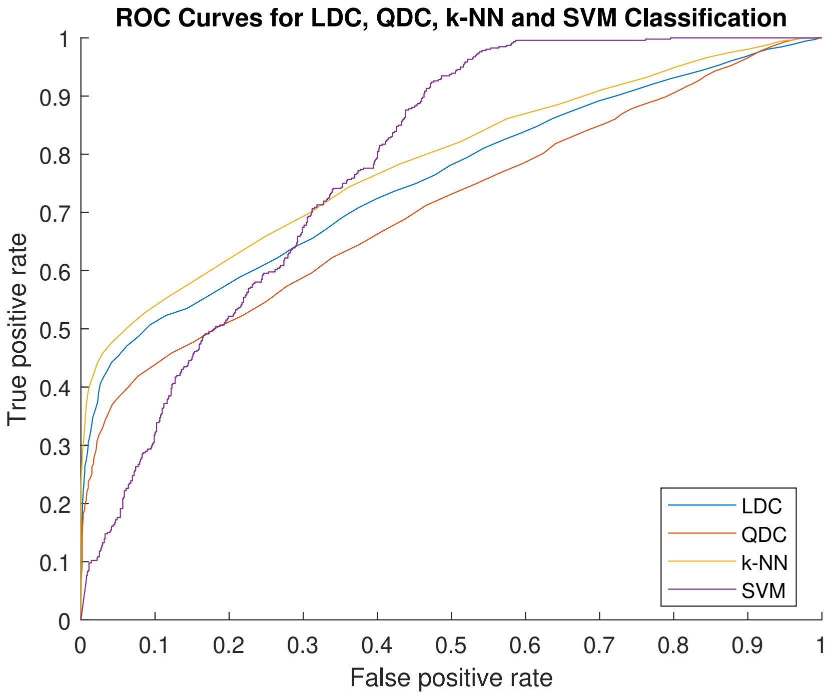


Figure 6.14: Experiment 4 - ROC curve of each classifier belonging to class 2, epilepsy source location

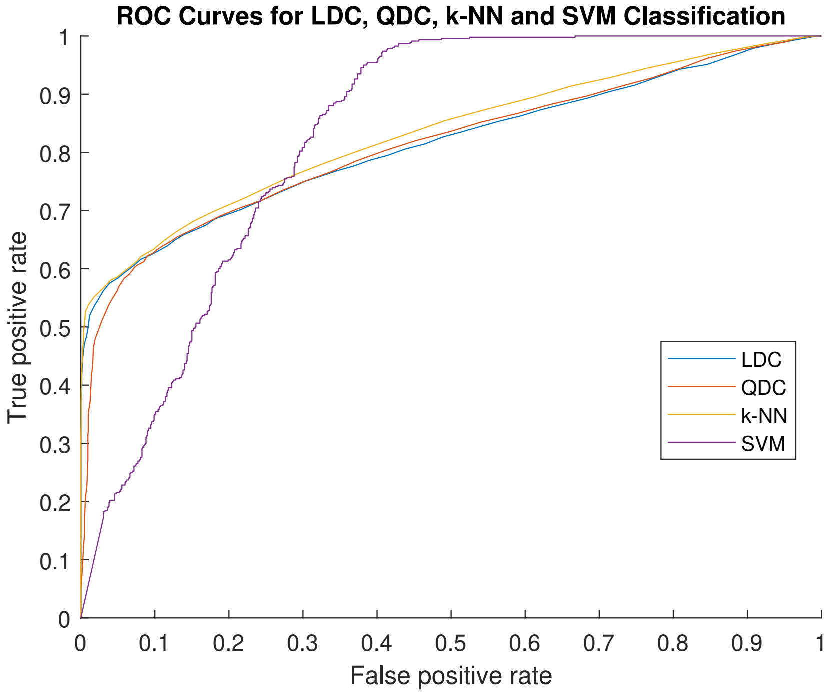


Figure 6.15: Experiment 4 - ROC curve of each classifier belonging to class 3, healthy brain area

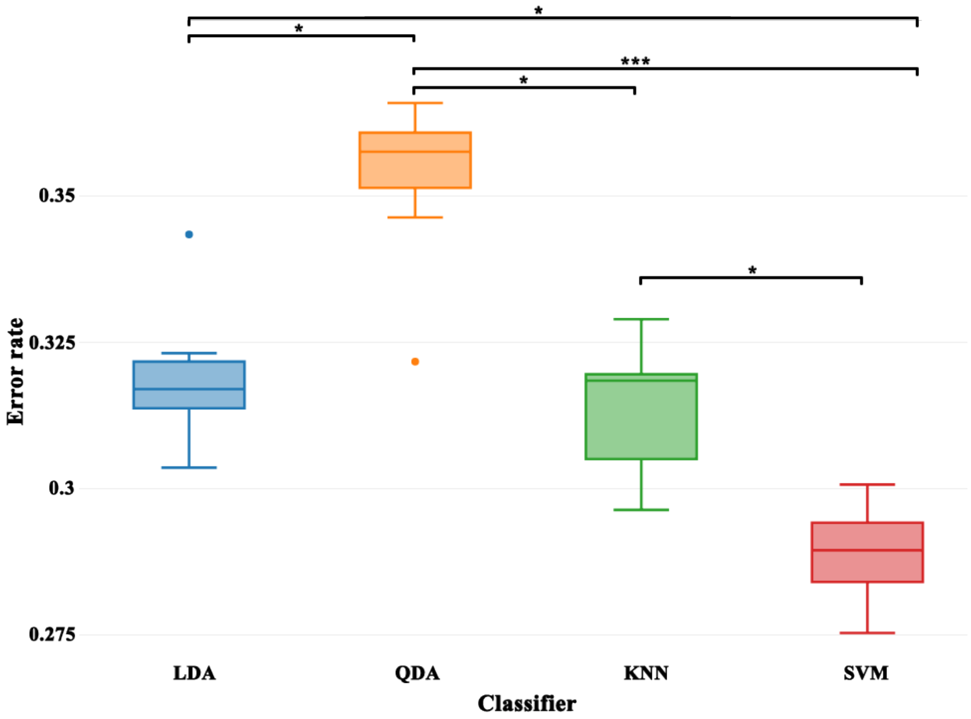


Figure 6.16: Experiment 4 - Comparison of classifiers.
$* <= 0.05, ** <= 0.001, *** <= 0.0001.$

6.1.5 Experiment 5

This experiment is developed with 3 classes, which are a mix of the original classes: (1) Seizure activity and EEG signal from the area where the tumor was located and (2) healthy brain area and (3) a single class grouping the rest of classes. Table 6.25 shows the mean and standard deviation from the errors obtained by each classifier. It is clear that the classifier with lower error is SVM, followed by kNN. Furthermore, the standard deviation is low for all classifier with means that the data is almost accurate. Nonetheless, the error is great, that is to say, the possibility of the system fails is over 23% for all the classifiers. It is not good for health works.

Table 6.25: Experiment 5 - Comparison of classifier through mean and standard deviation measures

Classifier	Mean	Standard Deviation.
LDC	0.263739	0.00764745
QDC	0.319435	0.0103012
KNN	0.253261	0.00407991
SVM	0.233348	0.00568016

Table 6.26 shows the sensitivity and specificity belongs to class 1 and 2, where, in the case

of class 1, the higher sensitivity is in LDC and specificity is in QDC. On the other side, for class 2, the higher sensitivity is in QDC and the specificity is in LDC. Tables 6.27, 6.28, 6.29, 6.30, show the confusion matrices of all classifiers taken in the last iteration. It is noticed that these matrices vary in each classifier but none of them show a good classification.

Table 6.26: Experiment 5 - Comparison of classifier through sensitivity and specificity measures from the new merged classes 1 and 2

Classifier	Class 1		Class 2	
	Se	Sp	Se	Sp
LDC	0.658696	0.858696	0.56087	0.882609
QDC	0.495652	0.978986	0.936957	0.669565
KNN	0.651087	0.87971	0.554348	0.880978
SVM	0.642391	0.905072	0.658696	0.859239

Figure 6.17 shows the box plots of all classifiers measuring their accuracy of classification. It is clear, that the best is the SVM which is near to 75% followed nearly by kNN. The worst of them is QDC. However, this is not enough for saying that this experiment worked good. The percentage is not suitable for this kind of research. Figure 6.18 shows the comparison of the ROC curves of all classifiers belonging to class 1. Besides, the Figure 6.19 shows the same curves but they belong to class 2. It is clear that the SVM and kNN curves are so close and are the best of all of them.

Table 6.27: Experiment 5 - Confusion matrix of LDC classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	606	187	127	920
2	133	258	69	460
3	62	29	829	920
Totals	801	474	1025	2300

Additionally, a significance test of error rates for experiment 5 is presented in Figure 6.20. The test used is Dunn test (Kruskal-Wallis with bonferroni correction).

Table 6.28: Experiment 5 - Confusion matrix of QDC classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	456	390	74	920
2	3	431	26	460
3	26	218	676	920
Totals	485	1039	776	2300

Table 6.29: Experiment 5 - Confusion matrix of kNN classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	599	183	138	920
2	140	255	65	460
3	26	36	858	920
Totals	765	474	1061	2300

Table 6.30: Experiment 5 - Confusion matrix of SVM classifier

True Labels	Estimated Labels			Totals
	1	2	3	
1	591	228	101	920
2	94	303	63	460
3	37	31	852	920
Totals	722	562	1016	2300

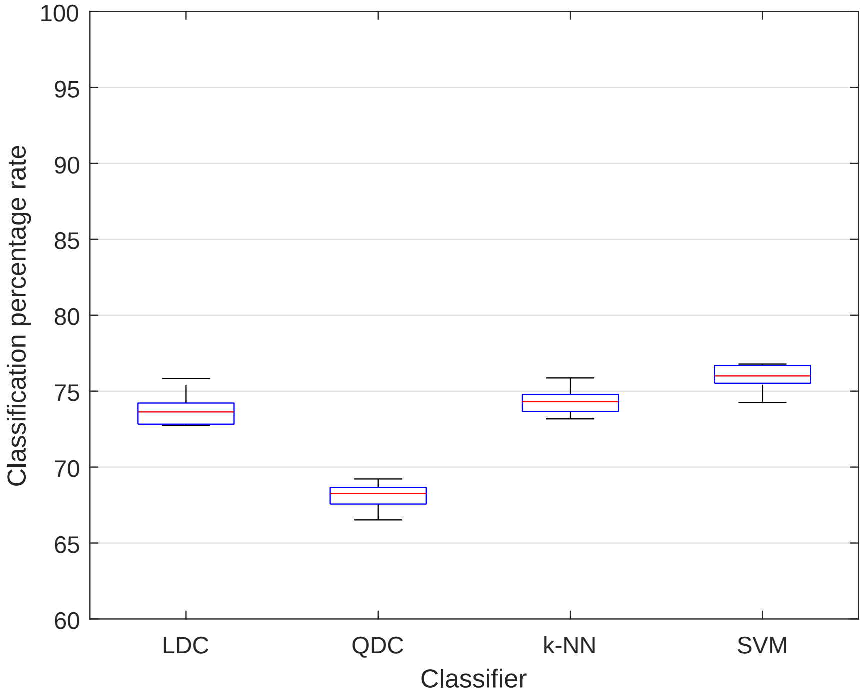


Figure 6.17: Experiment 5 - Comparison of classifiers through their accuracy

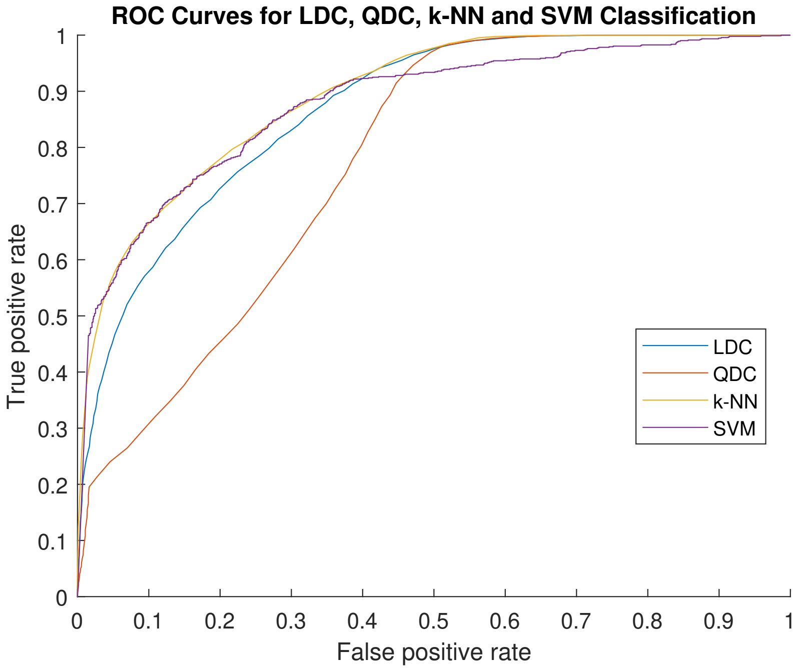


Figure 6.18: Experiment 5 - ROC curve of each classifier belonging to class 1, seizure activity

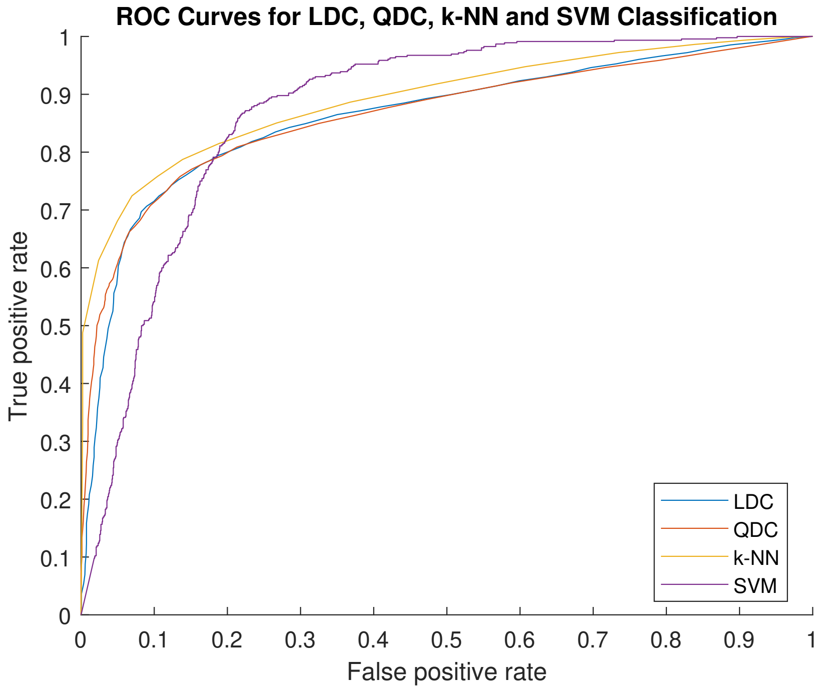


Figure 6.19: Experiment 5 - ROC curve of each classifier belonging to class 2, epilepsy source location

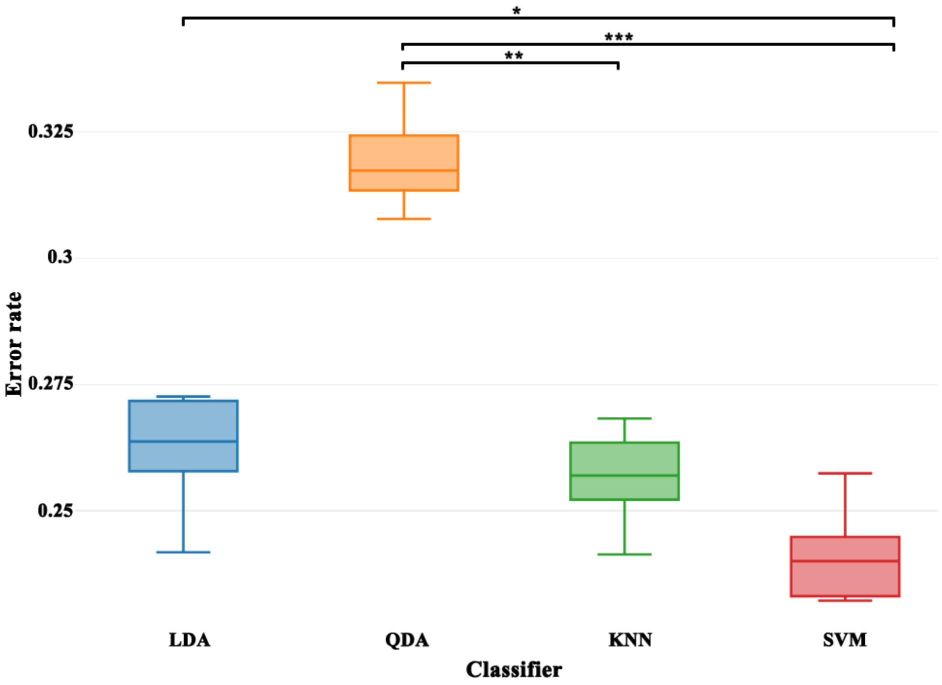


Figure 6.20: Experiment 5 - Comparison of classifiers.
$* <= 0.05, ** <= 0.001, *** <= 0.0001$.

6.2 Additional results

6.2.1 Overall results

To depict the effect of the overall results throughout the 5 experiments, the error rate variations regarding every experiment is plotted in Figure 6.21.

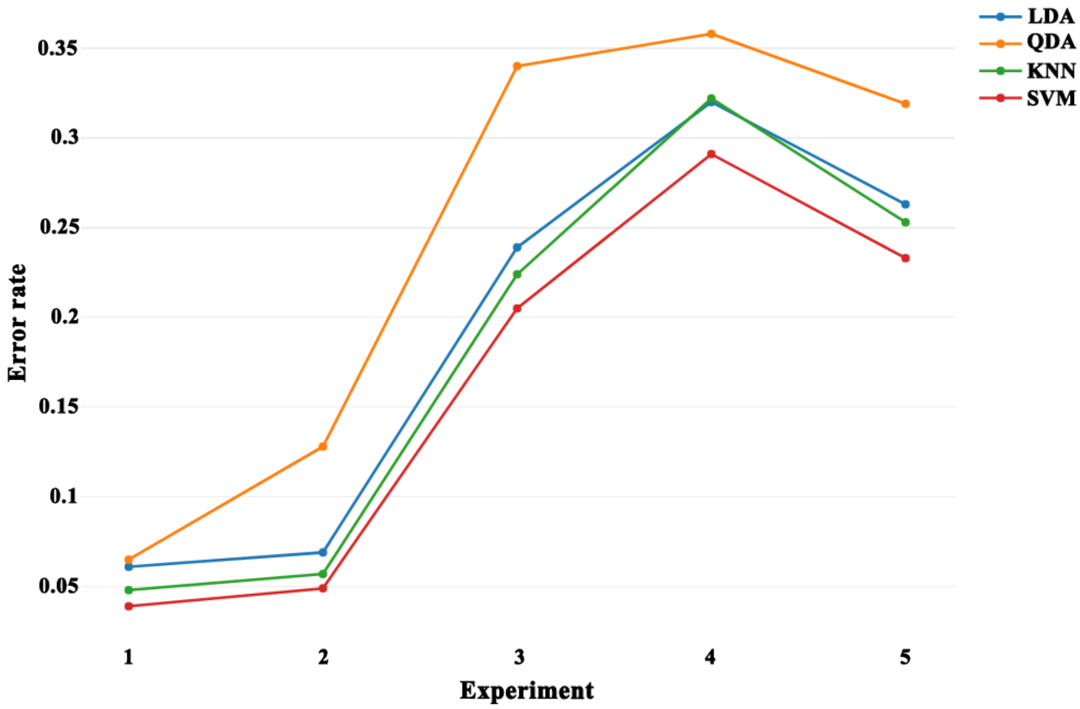
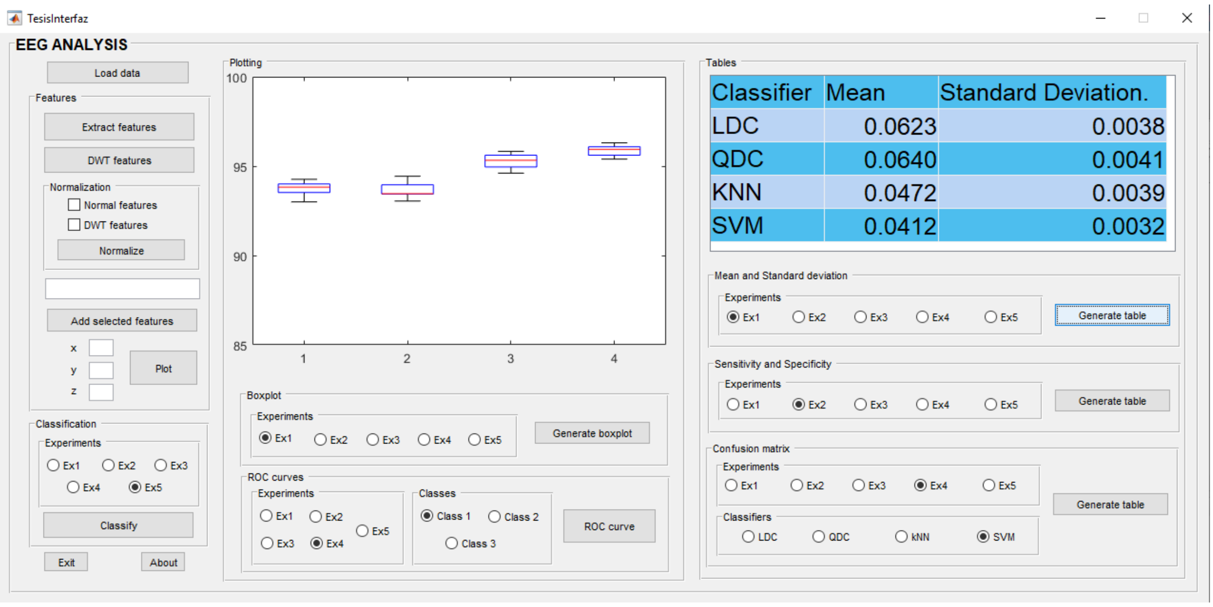


Figure 6.21: Comparison of classifiers through their error rate.

6.2.2 A graphic user interface

In order to replicating the results in an interactive way, as an additional result, an interface application is created, which is very intuitive and practical. It has four main sections: features, classification, plotting, and tables. Also, four external buttons: Load data, exit and about. Figure 6.22 shows a general view of the created graphic user interface.

To replicate the experiments first it is necessary to load the original dataset called “Epileptic seizure recognition data set” from the corresponding path where the file is located. Then, in the feature section, the extraction and selection of features are performed. In the classification section, it is necessary to choose the experiment to replicate. After, with the classification button, the process is started. When the process is finished, in the plotting section you can be able to generate the box plot or ROC curve of any experiment. Finally, in the tables section, you can generate a table of any experiment with the statistical measure that you choose.



Classifier	Mean	Standard Deviation.
LDC	0.0623	0.0038
QDC	0.0640	0.0041
KNN	0.0472	0.0039
SVM	0.0412	0.0032

Normalization: □ Normal features □ DWT features

Classification – Experiments: ○ Ex1 ○ Ex2 ○ Ex3 ○ Ex4 ● Ex5

Boxplot – Experiments: ● Ex1 ○ Ex2 ○ Ex3 ○ Ex4 ○ Ex5

ROC curves – Experiments: ○ Ex1 ○ Ex2 ○ Ex3 ● Ex4 ○ Ex5; Classes: ● Class 1 ○ Class 2 ○ Class 3

Mean and Standard deviation – Experiments: ● Ex1 ○ Ex2 ○ Ex3 ○ Ex4 ○ Ex5

Sensitivity and Specificity – Experiments: ○ Ex1 ● Ex2 ○ Ex3 ○ Ex4 ○ Ex5

Confusion matrix – Experiments: ○ Ex1 ○ Ex2 ○ Ex3 ● Ex4 ○ Ex5; Classifiers: ○ LDC ○ QDC ○ kNN ● SVM

Figure 6.22: Screenshot of the graphic interface

In addition, a Google site is developed with the complete information of this degree thesis. The information of this website is described in the appendix B.

Chapter 7

Conclusions and Future work

In general terms, the main goal for this research is achieved given that state-of-the-art techniques of characterization and classification of EEG signals are properly evaluated and selected accordingly to the performance on the task of epilepsy seizure activity identification. At a more technical and specific level, the following conclusions are stated:

- A set of features was established from previous works on EEG signals. All of them were tested on the dataset. The best subset of features for recognition of seizure activity of these kinds of signals was obtained from the decomposition of the signals through discrete wavelet transform (DWT). On these signals, spectral-representation-based features were calculated. The application that works better for the feature selection was Weka in which two algorithms were used, `CfsSubsetEval` with `BestFirst` and `InfoGainAttributeEval` with `Ranker`. So, by using a characterization based on statistical measures and spectral representation, a suitable set of features was established.

- Classifiers were selected after a revision of works on EEG signals. The classifiers used for this research are: Linear discriminant analysis classifier (LDC), Quadratic discriminant analysis classifier (QDC), k-nearest neighbor (k-NN) and Support vector machine (SVM). All of these classifiers were implemented in MATLAB. The first three ones are used from the toolbox `PRTOOLS` developed for MATLAB. The last one was imported from the application `Classifier Learner` of MATLAB and the code to extract the sets of training and test had to be implemented. All of these classifiers were adapted to be tested with the dataset of this work.

- The classifiers were tested under the same computational conditions. The number of iteration used for each classifier is 10 which shows good results. The comparison of classifiers was through statistical measures like: mean, standard deviation, sensitivity, and specificity. Moreover, it presents figures like box plots and ROC curves. Also, it shows confusion matrices for all classifiers. In the five experiments performed, the classifier Support vector machine (SVM) is which has the higher classification percentage. That is to say, it is the best classifier founded for this research. However, experiments 1 and 2 show a higher classification percentage. It means that for two classes all the classifiers work better. In experiments 1, 2, 3 and 4, the ROC curve, belonging to class 1, shows that this class is well classified, which is the class of interest for this research. Furthermore, in experiment 2, class 2 is well classified which is an indicator of the epilepsy source location.

- The main contribution of this work can be explained into two parts. One hand, the developed experimental framework allows for establishing the general settings of characterization, feature selection and classification for EEG signals. On the other hand, the performed comparison of the explored techniques has given as a remarkable result that the features taken from DWT decomposition along with the Support Vector Machine are the best alternatives to build a Epilepsy-driven EEG analysis computer-aided system. In other words, as a pioneer study in Yachay Tech on this field, the outcomes of this research provide meaningful hints and recommendations to set up the path to follow when aiming at processing EEG signals for epilepsy diagnosis purposes.

As future work, since there are a lot of ways to do decomposition of EEG signals in epilepsy research, it is proposed to explore other alternatives different to the ones used in this work, for example: Improved complementary ensemble empirical mode decomposition (ICEEMD), Maximal overlap discrete wavelet transform (MODWT), Hilbert-Huang transform. Also, more characteristics as well as other feature extraction techniques are to be explored.

Furthermore, a former study on the input data might help to assess and determine the suitableness of the features regarding the of-interest classification task. For instance, by following an analysis of relevance of characteristics or applying a method of dimensionality reduction (i.e. Principal Component Analysis – PCA). Besides, it can be useful to add more classifiers for the comparison as a neural network or random forest. Also, a mix of different classifiers can be tested. Also, the results and input signals visualization can be improved by more intelligible graphic representations to be further studied in subsequent researchers.

References

[1] M. E. Acosta-Muñoz, H. A. Paredes-Argoty, E. J. Revelo-Fuelagán, and D. H. Peluffo-Ordóñez, “On the effect of inverse problem weighted solutions for epileptic sources localization,” in *2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*, Sep. 2015, pp. 1–5.

[2] P. Malmivuo, J. Malmivuo, and R. Plonsey, *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields.* Oxford University Press, USA, 1995, pp. 366–375. [Online]. Available: http://www.bem.fi/book/book.pdf

[3] M. Abo-Zahhad, S. Ahmed, and S. N. Seha, “A new eeg acquisition protocol for biometric identification using eye blinking signals,” *International Journal of Intelligent Systems and Applications (IJISA)*, pp. 48–54, May 2015.

[4] C. Bishop, *Pattern recognition and Machine learning.* Singapore: Springer Science+Business Media, 2006, pp. 1–57.

[5] A. Hamad, E. H. Houssein, A. E. Hassanien, and A. A. Fahmy, “Feature extraction of epilepsy eeg using discrete wavelet transform,” in *2016 12th International Computer Engineering Conference (ICENCO)*, Dec. 2016, pp. 190–195.

[6] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, “Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state,” *Physical Review E*, vol. 64, no. 6, p. 061907, 2001.

[7] T. Oliveira, J. Carollo, D. Robertson, Z. Pan, and P. Heyn, “Incidence of epilepsy in adults with cerebral palsy and secondary health outcomes: A review and proposed feasibility study.” *Journal of Neurological Disorders*, vol. 2, p. 188, Oct. 2014.

[8] J. Christensen and P. Sidenius, “Epidemiology of epilepsy in adults: Implementing the ilae classification and terminology into population-based epidemiologic studies,” *Epilepsia*, vol. 53, pp. 14–17, 2012.

[9] A. Subasi and E. Erçelebi, “Classification of eeg signals using neural network and logistic regression,” *Computer methods and programs in biomedicine*, vol. 78, no. 2, pp. 87–99, 2005.

[10] J. Sirven and P. Shafer, *What is Epilepsy?* Epilepsy Foundation, Jan. 2014. [Online]. Available: https://www.epilepsy.com/learn/about-epilepsy-basics/what-epilepsyf

[11] L. Wang, W. Xue, Y. Li, M. Luo, J. Huang, W. Cui, and C. Huang, "Automatic epileptic seizure detection in eeg signals using multi-domain feature extraction and nonlinear analysis," *Entropy*, vol. 19, no. 6, p. 222, 2017.

[12] *Research and New Therapies?* Epilepsy Foundation. [Online]. Available: https://www.epilepsy.com/make-difference/research-and-new-therapies

[13] S. Viglione and G. Walsh, "Proceedings: Epileptic seizure prediction." *Electroencephalography and clinical neurophysiology*, vol. 39, no. 4, pp. 435–436, 1975.

[14] B. Litt and J. Echauz, "Prediction of epileptic seizures," *the LANCET Neurology*, vol. 1, no. 1, pp. 22–30, 2002.

[15] O. Faust, U. R. Acharya, H. Adeli, and A. Adeli, "Wavelet-based eeg processing for computer-aided seizure detection and epilepsy diagnosis," *Seizure*, vol. 26, pp. 56–64, 2015.

[16] T. Huzar, *Everything you need to know about convulsions*, Jan. 2019. [Online]. Available: https://www.medicalnewstoday.com/articles/324330.php

[17] *Epileptic seizures.* Epilepsy Society. [Online]. Available: https://www.epilepsysociety.org.uk/epileptic-seizures

[18] R. S. Fisher, W. V. E. Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel Jr, "Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.

[19] *Epilepsy: Diagnosis and treatment.* Mayo Clinic. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/epilepsy/diagnosis-treatment/drc-20350098

[20] V. Gnatkovsky, M. de Curtis, C. Pastori, F. Cardinale, G. Lo Russo, R. Mai, L. Nobili, I. Sartori, L. Tassi, and S. Francione, "Biomarkers of epileptogenic zone defined by quantified stereo-eeg analysis," *Epilepsia*, vol. 55, Jan. 2014.

[21] B. Erem, D. E. Hyde, J. M. Peters, F. H. Duffy, D. H. Brooks, and S. K. Warfield, "Combined delay and graph embedding of epileptic discharges in eeg reveals complex and recurrent nonlinear dynamics," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2015, pp. 347–350.

[22] P. L. Nunez, R. Srinivasan *et al.*, *Electric fields of the brain: the neurophysics of EEG*, 2nd ed. Oxford University Press, USA, 2006. [Online]. Available: https://brainmaster.com/software/pubs/brain/Nunez%202ed.pdf

[23] M. Tudor, L. Tudor, and K. I. Tudor, "Hans berger (1873-1941)–the history of electroencephalography," *Acta medica Croatica : casopis Hravatske akademije medicinskih znanosti*, vol. 59, no. 4, p. 307—313, 2005. [Online]. Available: http://europepmc.org/abstract/MED/16334737

[24] T. W. Picton and A. Mazaheri, "Electroencephalography (eeg)," *Encyclopedia of Cognitive Science*, 2006.

[25] J. Gomez and O. Ordoñez, *Estudio comparativo de técnicas de caracterización y clasificación automática de emociones a partir de señales del cerebro*, 2018. [Online]. Available: https://sites.google.com/site/degreethesisdiegopeluffo/emotion-detection-via-eeg-analysis

[26] H. Jasper, "Report of the committee on methods of clinical examination in electroencephalography," *Electroencephalogr Clin Neurophysiol*, vol. 10, pp. 370–375, 1958.

[27] R. Cooper, J. W. Osselton, and J. C. Shaw, *EEG Technology*, 2nd ed. London: Butterworth-Heinemann, 2014, p. 275.

[28] F. Sharbrough, G. Chatrian, R. Lesser, H. Luders, M. Nuwer, and T. Picton, "American electroencephalographic society guidelines for standard electrode position nomenclature," *Clinical Neurophysiology*, vol. 8, pp. 200–202, Jan. 1991.

[29] E. Ackerman and L. Gatewood, *Mathematical Models in the Health Sciences: A Computer-Aided Approach*. University of Minnesota Press, 1979.

[30] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[31] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[32] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. London, England: MIT Press, 2010.

[33] M. Heyden, "Classification of eeg data using machine learning techniques," 2016, student Paper.

[34] S. Barua and S. Begum, "A review on machine learning algorithms in handling eeg artifacts," in *The Swedish AI Society (SAIS) Workshop SAIS, 14, 22-23 May 2014, Stockholm, Sweden*, 2014.

[35] S. Watanabe, *Pattern Recognition: Human and Mechanical*. New York: John Wiley & Sons, Inc., 1985.

[36] H. Nagendra, S. Mukherjee, and V. Kumar, "Application of wavelet techniques in ecg signal processing: an overview," *Int J Eng Sci Technol*, vol. 3, no. 10, pp. 7432–7443, 2011.

[37] S. Ari, "An overview of the research work on multispectral imaging, hand gesture recognition, eeg and ecg signal processing," *CSI Transactions on ICT*, pp. 1–5, 2019.

[38] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.

[39] E. Derya Übeyli, "Wavelet/mixture of experts network structure for eeg classification," *Expert Syst. Appl.*, vol. 34, pp. 1954–1962, Apr. 2008.

[40] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[41] M. Loog, "Supervised classification: Quite a brief overview," in *Machine Learning Techniques for Space Weather.* Elsevier, 2018, pp. 113–145.

[42] J. Richards, *Remote sensing digital image analysis*, 5th ed. Springer-Verlag, 2013, pp. 247–342.

[43] T. M. Cover, P. Hart *et al.*, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[44] S. Bhattacharyya, A. Khasnobish, S. Chatterjee, A. Konar, and D. Tibarewala, "Performance analysis of lda, qda and knn algorithms in left-right limb movement classification from eeg data," in *2010 International Conference on Systems in Medicine and Biology.* IEEE, 2010, pp. 126–131.

[45] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert systems with applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

[46] I. Guler and E. D. Ubeyli, "Multiclass support vector machines for eeg-signals classification," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 2, pp. 117–126, 2007.

[47] C. A. González and G. de Sistemas Inteligentes, "Svm: Máquinas de vectores soporte," *Grupo de Sistemas Inteligentes, Departamento de Informática, Universidad de Valladolid, recuperado en Noviembre*, 2011.

[48] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[49] L. Guo, D. Rivero, J. Dorado, J. R. Rabunal, and A. Pazos, "Automatic epileptic seizure detection in eegs based on line length feature and artificial neural networks," *Journal of neuroscience methods*, vol. 191, no. 1, pp. 101–109, 2010.

[50] C. Rodriguez, J. Gallego, I. Mora, A. Orozco-Duque, and J. Bustamante, "Classification of premature ventricular contraction beats based on unsupervised learning methods," *Revista Ingeniería Biomédica*, vol. 8, no. 15, pp. 51–58, 2014.

[51] A. W. Abbas, N. Minallh, N. Ahmad, S. A. R. Abid, and M. A. A. Khan, "K-means and isodata clustering algorithms for landcover classification using remote sensing," *Sindh University Research Journal-SURJ (Science Series)*, vol. 48, no. 2, pp. 315–318, Apr. 2016.

[52] W. Qiuyi and F. Ernest, *Epileptic Seizure Recognition Data Set.* UCI Machine learning repository, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

[53] *Introduction to Wavelet Families.* MathWorks. [Online]. Available: https://es.mathworks.com/help/wavelet/gs/introduction-to-the-wavelet-families.html#f3-1009153

[54] H. Ocak, "Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2027–2036, 2009.

[55] C. Wilmer, *Caracterización y clasificación de señales electroencefalográficas para aplicaciones de interfaz cerebro-computador.* Universidad de Nariño, May 2019.

[56] M. H. Alomari, A. AbuBaker, A. Turani, A. M. Baniyounes, and A. Manasreh, "Eeg mouse: A machine learning-based brain computer interface," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 4, pp. 193–198, 2014.

[57] E. D. Übeyli, "Statistics over features: Eeg signals analysis," *Computers in Biology and Medicine*, vol. 39, no. 8, pp. 733–741, 2009.

[58] V. Bajaj and R. B. Pachori, "Eeg signal classification using empirical mode decomposition and support vector machine," in *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011.* Springer, 2012, pp. 623–635.

[59] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.

[60] D. G. Manolakis and V. K. Ingle, *Applied digital signal processing: theory and practice.* Cambridge University Press, 2011.

[61] B. Medina Salgado and R. Alvarez López, "Characterization of eeg signals using wavelet packet and fuzzy entropy in motor imagination tasks," *Ingeniería*, vol. 22, no. 2, pp. 226–238, 2017.

[62] A. Phinyomark, S. Thongpanja, H. Hu, P. Phukpattaranont, and C. Limsakul, "The usefulness of mean and median frequencies in electromyography analysis," in *Computational intelligence in electromyography analysis-A perspective on current applications and future challenges.* IntechOpen, 2012.

[63] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1999.

[64] E. Rich and K. Knight, *Artificial Intelligence TATA McGRAW-HILL*, 1991.

[65] *Class InfoGainAttributeEval.* Source Forge. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html

[66] *Class Ranker.* Source Forge. [Online]. Available: http://weka.sourceforge.net/doc.stable/weka/attributeSelection/Ranker.html

Appendices

Appendix A

Matlab Code

In this Section, the corresponding source codes for pre-processing, characterization and classification of epilepsy-driven EEG signals are presented.

A.1 Load of dataset

The original dataset is loaded and preprocessed. Furthermore, other files are generated using the needed data for doing the experiments. These files are generated with the selected features obtained from Weka.

Source Code 1

```
% Load data
clc , close all , clear

M=csvread('data.csv',1,1);
y = M(:,end);
for s = 1:500
    for i = 1:23
        EEGDB.(['subject' num2str(s)]).(['chunk' num2str(i)]).data = M((s-1)*23+i,1:178);
        EEGDB.(['subject' num2str(s)]).(['chunk' num2str(i) ]).label = M((s-1)*23+i,179);
    end
end

% Original Matrix Normalization

M2=zeros(11500,179);

for i = 1:11500
    Max_r=max(abs(M(i,:)));
    M2(i,:)=M(i,:)/Max_r;
end

clear i Max_r s

%% Simple features

clc
[lM2,n] = size(M2);
M3=zeros(lM2,36); % Characterication on original data

for i=1:lM2
    M3(i,1:32)=features(M2(i,1:178));
    M3(i,33:36)=features_fs(M2(i,1:178),178);
end

%% DWT
clc

M4=zeros(lM2,192);    % Charactirzation on dwt data
for i=1:lM2
```

```
    M4(i,:)=features_dwt(M2(i,1:178));
end

clear i lM2 n

%% Delete Inf values columns
clc
X = [M3,M4];          % Feature matrix
[i,j]=size(X);
posc=[];
for f=1:i
    for c=1:j
            if( isinf(X(f,c))==1)
                posc=[posc(:,:),c];
            end
    end
end

X(:,posc) = [];       % Final feature matrix without INF values

clear c f i j

% Normalization of features
clc

X_no_norm = X;
X= X./repmat(max(abs(X)),size(X,1),1);

%% Ranking of the best features - Got from Weka

% 132 151 196 88 169 139 165 163 157 150 134 181 216
% 152 198 211 201 221 215 212 217 167 46 185 18 108
% 67 28 69 220 17 58 114 59 146 218 115 107 83 84 178

T=table(X,char(y+97));
writetable(T,’FeatNorm_all.csv’);
clear T

features=[132 151 196 88 169 139 165 163 157];

%% Experiment 1 - 1 vs All
% The classes 2,3,4 and 5 were assigned to a single class

clc

T=table(X,y==1);
writetable(T,’FeatNorm_2.csv’);
clear T

[SVMdata]=test_2classes(’FeatNorm_2.csv’,features);

%% Experiment 2 - 1,2
% The classes 3, 4 and 5 were removed

clc
pos = [];
for i=1:length(y)
    if(y(i)>2)
        pos=[pos,i];
    end
end

X2=X;
X2(pos,:)=[];
y2=y;
y2(pos)=[];

T=table(X2,y2==1);
writetable(T,’FeatNorm_2class_12.csv’);

clear T pos i

[scoresWldc2,scoresWqdc2,scoresWknn2,scoresSVM2]=test_2classes(’FeatNorm_2class_12.csv’,features);

%% Experiment 3 - 1,2 vs all
% The classes 3,4 and 5 were joined in a singles class

clc
y3=y;
for i=1:length(y)
    if(y(i)>2)
        y3(i)=3;
    end
end

T=table(X,y3);
writetable(T,’FeatNorm_3class_12.csv’);
clear T pos i
```

```
[SVMdata]=test_3classes('FeatNorm_3class_12.csv',features);

%% Experiment 4 - 1,2,3
% The classes 4 and 5 were deleted

clc
pos = [];
for i=1:length(y)
    if(y(i)>3)
        pos=[pos,i];
    end
end

X4=X;
X4(pos,:)=[];
y4=y;
y4(pos)=[];

T=table(X4,y4);
writetable(T,'FeatNorm_3class_123.csv');
clear T pos i

[scoresWldc4,scoresWqdc4,scoresWknn4,scoresSVM4]=test_3classes('FeatNorm_3class_123.csv',features);

%% Experiment 5 - 1 and 2 as class1, 3 as class2, 4 and 5 as class3

clc
y5 = y;
for i=1:length(y5)
    if(y5(i)<3)
        y5(i)=1;
    elseif (y5(i)==3)
        y5(i)=2;
    else
        y5(i)=3;
    end
end


T=table(X,y5);
writetable(T,'FeatNorm_3class_mix.csv');
clear T pos i

[scoresWldc5,scoresWqdc5,scoresWknn5,scoresSVM5]=test_3classes('FeatNorm_3class_mix.csv',features);
```

A.2 Feature extraction

The extraction of the features was developed by considering statistical measures and the decomposition of discrete wavelet transformation (DWT). The statistical measures are described in the source code 2 and 3 and the DWT are in the source code 4. These were implemented as functions.

Source Code 2

```
function y = features(x)
n=length(x);
y=zeros(1,32);

y(1)=sum(abs(x)); % Area under the curve                                m
y(2)=mean(abs(x)); % Mean                                                      t
y(3)=rms(x); % Root mean square                                                t
y(4)=VAR(x); % Variance                                                        t
y(5)=std(x); % Standard Deviation                                              t
y(6)=wentropy(x,'log energy'); % Log energy entropy                            t
y(7)=sum(x.^2); % Square integral                                       m
y(8)=kurtosis(x); % Kurtosis                                                   t
y(9)=cov(x); % Covariance                                                      t
y(10)=wentropy(x,'shannon'); % Shannon entropy                          m
y(11)=AAC(x); % Average Amplitud Change                                 m
y(12)= sqrt(sum(abs(x).^2)); % Root Sum of Squares RSSQ                        t
y(13) = sum((abs(x)).^2); % Simple Square Integral SSI                  m
```

```
y(14) = sum((abs(x)).^2)/n; % Simple Square Integral II SSI2 (Potrncia)         m
y(15) = VEEG(x); % calculate the Variance of EEG (VAR)                                  t
y(16) = AsDec(x);                                            %                          t
y(17) = CorPeakNumber(x);                                      %                m
y(18) = Energy1_3Cor(x);                                     %                  m
y(19) = Energy2_3Cor(x);                                    %                   m
y(20) = int_ratio(x);                                       %                           t
y(21) = KurtoEnv(x);                                        %                           t
y(22) = KurtoSig(x);                                        %                           t
y(23) = MaxFFT(x);                                          %                                   s
y(24) = MeanFFT(x);                                         %                                   s
y(25) = MedianFFT(x);                                       %                                   s
y(26) = RappMaxMean(x);                                     %                   m
y(27) = RappMaxMedian(x);                                   %                   m
y(28) = SkewEnv(x);                                         %                   m
y(29) = SkewSig(x);                                         %                           t
y(30) = VarFFT(x);                                          %                                   s
y(31) = median(x); % Median                                                             t
y(32) = entropy(x); % Entropy                                                           t
end
```

Source Code 3

```
function y = features_fs(x,fs)
n=length(x);
y=zeros(1,4);

X2=2*abs(fft(x,n))/n;      % Calculate the FFT, divide the number of samples y multiply by 2
P1=X2(1:n/2+1);            % Leave just the half of FFT
f = fs * (0:(n/ 2))/n;
[m,~]=max(P1);

y(1) = m; % Peak frequency
y(2) = mean(P1);    % Mean Frequency
y(3) = sum(abs(P1).^2)*fs/n;    % Energy
y(4) = FMaxFFT(x,fs); % Max FFT
end
```

Source Code 4

```
function X = features_dwt(x)

y=zeros(1,180);

[CC,LL]=wavedec(x,5,'db4');
A5 = CC(1 : LL(1));
D5 = CC(LL(1)+1 : LL(1)+LL(2) );
D4 = CC(LL(1) + LL(2) + 1 : LL(1) + LL(2) + LL(3) );
D3 = CC(LL(1) + LL(2) + LL(3) + 1 : LL(1) + LL(2) + LL(3) + LL(4) );
D2 = CC(LL(1) + LL(2) + LL(3) + LL(4)+1 : LL(1) + LL(2) + LL(3) + LL(4) + LL(5));
D1 = CC(LL(1) + LL(2) + LL(3) + LL(4) + LL(5) + 1 : LL(1) + LL(2) + LL(3) + LL(4) + LL(5) + LL(6));

XA5 = features(A5);
XD5 = features(D5);
XD4 = features(D4);
XD3 = features(D3);
XD2 = features(D2);
XD1 = features(D1);

X = [XA5,XD5,XD4,XD3,XD2,XD1];

end
```

A.3 Classification functions

Two functions to perform the classifiers were needed because two experiments were implemented only with two classes, while the remaining ones were implemented with three classes. The PRtools MatLab toolbox was used for implementing the LDC, and QDC classifiers. For SVM, the classifier learner application from Matlab was used.

A.3.1 Function for 2 classes

It uses one of the files generated by the source code 1, and the selected features given by Weka. The classifier gives the necessary parameters to generate statistical measures and plots.

Source Code 5

```
function [SVMdata]=test_2classes(filename,columnas)
    %% Initialize variables.
    clc

    delimiter = ',';
    startRow = 2;
    % Format for each line of text:
    formatSpec =
          '%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f
    % Open the text file.
     fileID = fopen(filename,'r');
    % Read columns of data according to the format.
    dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter, 'TextType', 'string', 'EmptyValue', NaN, 'HeaderLines'
          ,startRow-1, 'ReturnOnError', false, 'EndOfLine', '\r\n');
    % Close the text file.
     fclose(fileID);
    % Create output variable
    dataF = [dataArray{1:end-1}];
    % Clear temporary variables
    clearvars filename delimiter startRow formatSpec fileID dataArray ans;

    % Graph representation
    v1=132;
    v2=151;
    v3=196;

    X=dataF(:,1:(end-1));
    Y=dataF(:,end);

    scatter3(X(:,v1),X(:,v2),X(:,v3),5,Y>1);

    clc
    clear res
    j=0.8;              % Training percentage
    for i=1:10
        trainingData = [X(:,columnas) Y==1];
        n=unique(trainingData(:,end));
        extraccionE=[];
        extraccionP=[];
        for m=1:length(n)
            prueba=trainingData(trainingData(:,end)==n(m),:);
            p=j;
            a=size(prueba(:,1));
            Vector1=randperm(a(1));
            tama=round(round(a(1)*p));
            Vector2=Vector1(1:tama);
            Vector1=Vector1((tama+1):end);

            extraccionE=[extraccionE;prueba(Vector2,:)];
            extraccionP=[extraccionP;prueba(Vector1,:)];
        end
        trainingData=extraccionE;

        % Model

        [trainedClassifierSVM, SVMdata]= trainClassifierSVM(trainingData);
        m01=trainedClassifierSVM.predictFcn(extraccionP(:,1:9));
        res(i,1)=sum(extraccionP(:,10)==m01)/length(extraccionP);
    end


    %% Classification

    X=X(:,columnas);
    y=Y==1;

    [A,Wldc,Wqdc, Wknn, error_total, CP_total, Atest]=ldc_qdc_knn(X,y);

    [a,b]=size(CP_total)
    [c,d]=size(res)

    labelLDC=labeld(Atest,Wldc);
    confmat(Atest.nlab,labelLDC);

    labelQDC=labeld(Atest,Wqdc);
    confmat(Atest.nlab,labelQDC);

    labelKNN = labeld(Atest,Wknn);
    confmat(Atest.nlab,labelKNN);

    SVMconf=confmat(extraccionP(:,10),m01);

    %% Plotting boxplot
    figure
    boxplot([CP_total*100 res*100])
    ylim([70 100])

    CP_total2=[CP_total*100 res*100];

    %% ROC curves
    class=input('Ingrese la clase a evaluar: ');

    [scoresWldc,scoresWqdc,scoresWknn,scoresSVM]=curvasROC (A,Wldc,Wqdc,Wknn,extraccionP,SVMdata,class,2)

    %% Tables Mean and Std

    error_total=[error_total 1-res];

    error_table{1,1} ='Classifier'; error_table{1,2} ='Mean'; error_table{1,3} ='Standard Deviation.';
    error_table{2,1} ='LDC'; error_table{3,1} ='QDC'; error_table{4,1} ='KNN'; error_table{5,1} ='SVM';
    for i = 2:5
        error_table{i,2} =mean(error_total(:,i-1));
        error_table{i,3} =std(error_total(:,i-1));
    end
    disp(error_table);

    % To save the table in latex format
    % cell2latextable(error_table,'tablas_latex','errorEx2');

    %% Tables Sensitivity and Specificity

    SpSVM=SVMconf(1,1)/(SVMconf(1,1)+SVMconf(1,2));
    SeSVM=SVMconf(2,2)/(SVMconf(2,2)+SVMconf(2,1));

    LDCres=Atest*Wldc;
    [SeLDC,SpLDC]=testc(LDCres,'sensitivity',1);

    QDCres=Atest*Wqdc;
    [SeQDC,SpQDC]=testc(QDCres,'sensitivity',1);

    KNNres=Atest*Wknn;
    [SeKNN,SpKNN]=testc(KNNres,'sensitivity',1);

    SeSp_table{1,1} ='Classifier'; SeSp_table{1,2} ='Sensitivity'; SeSp_table{1,3} ='Specificity';
    SeSp_table{2,1} ='LDC'; SeSp_table{3,1} ='QDC'; SeSp_table{4,1} ='KNN'; SeSp_table{5,1} ='SVM';
    SeSp_table{2,2} =SeLDC; SeSp_table{3,2} =SeQDC; SeSp_table{4,2} =SeKNN; SeSp_table{5,2} =SeSVM;
    SeSp_table{2,3} =SpLDC; SeSp_table{3,3} =SpQDC; SeSp_table{4,3} =SpKNN; SeSp_table{5,3} =SpSVM;
    disp(SeSp_table);

    % To save the table in latex format
    % cell2latextable(SeSp_table,'tablas_latex','SeSpEx2');

end
```

A.3.2 Function for 3 classes

It, also, uses one of the files generated but the source code 1 and the selected features by Weka. The classifier gives the needed parameters to generate statistical measures and plots.

Source Code 6

```
function [SVMdata]=test_3classes(filename,columnas)
    %% Initialize variables.
    clc

    delimiter = ',';
    startRow = 2;
    % Format for each line of text:
    formatSpec =
          '%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f
    % Open the text file.
    fileID = fopen(filename,'r');
    % Read columns of data according to the format.
    dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter, 'TextType', 'string', 'EmptyValue', NaN, 'HeaderLines'
          ,startRow-1, 'ReturnOnError', false, 'EndOfLine', '\r\n');
    % Close the text file.
    fclose(fileID);
    % Create output variable
    dataF = [dataArray{1:end-1}];
    % Clear temporary variables
    clearvars filename delimiter startRow formatSpec fileID dataArray ans;

    % Graph representation
    v1=132;
    v2=151;
    v3=196;

    X=dataF(:,1:(end-1));
    Y=dataF(:,end);

    scatter3(X(:,v1),X(:,v2),X(:,v3),5,Y>1);

    clc
    clear res
    j=0.8;           % Training percentage
    for i=1:10
        trainingData = [X(:,columnas) Y];
        n=unique(trainingData(:,end));
        extraccionE=[];
        extraccionP=[];
        for m=1:length(n)
            prueba=trainingData(trainingData(:,end)==n(m),:);
            p=j;
            a=size(prueba(:,1));
            Vector1=randperm(a(1));
            tama=round(round(a(1)*p));
            Vector2=Vector1(1:tama);
            Vector1=Vector1((tama+1):end);

            extraccionE=[extraccionE;prueba(Vector2,:)];
            extraccionP=[extraccionP;prueba(Vector1,:)];
        end
        trainingData=extraccionE;

        % Model

        [trainedClassifierSVM, SVMdata] = trainClassifierSVM_3classes(trainingData);

        m01=trainedClassifierSVM.predictFcn(extraccionP(:,1:9));
        m02=string(m01);
        m02=double(m02);
        res(i,1)=sum(extraccionP(:,10)==m02)/length(extraccionP);
    end


    %% Classification

    X=X(:,columnas);
    y=Y;

    [A,Wldc,Wqdc, Wknn, error_total, CP_total, Atest]=ldc_qdc_knn(X,y);

    labelLDC=labeld(Atest,Wldc);
    confmat(Atest.nlab,labelLDC);

    labelQDC=labeld(Atest,Wqdc);
    confmat(Atest.nlab,labelQDC);

    labelKNN = labeld(Atest,Wknn);
    confmat(Atest.nlab,labelKNN);

    SVMconf=confmat(extraccionP(:,10),m02);

    %%
```

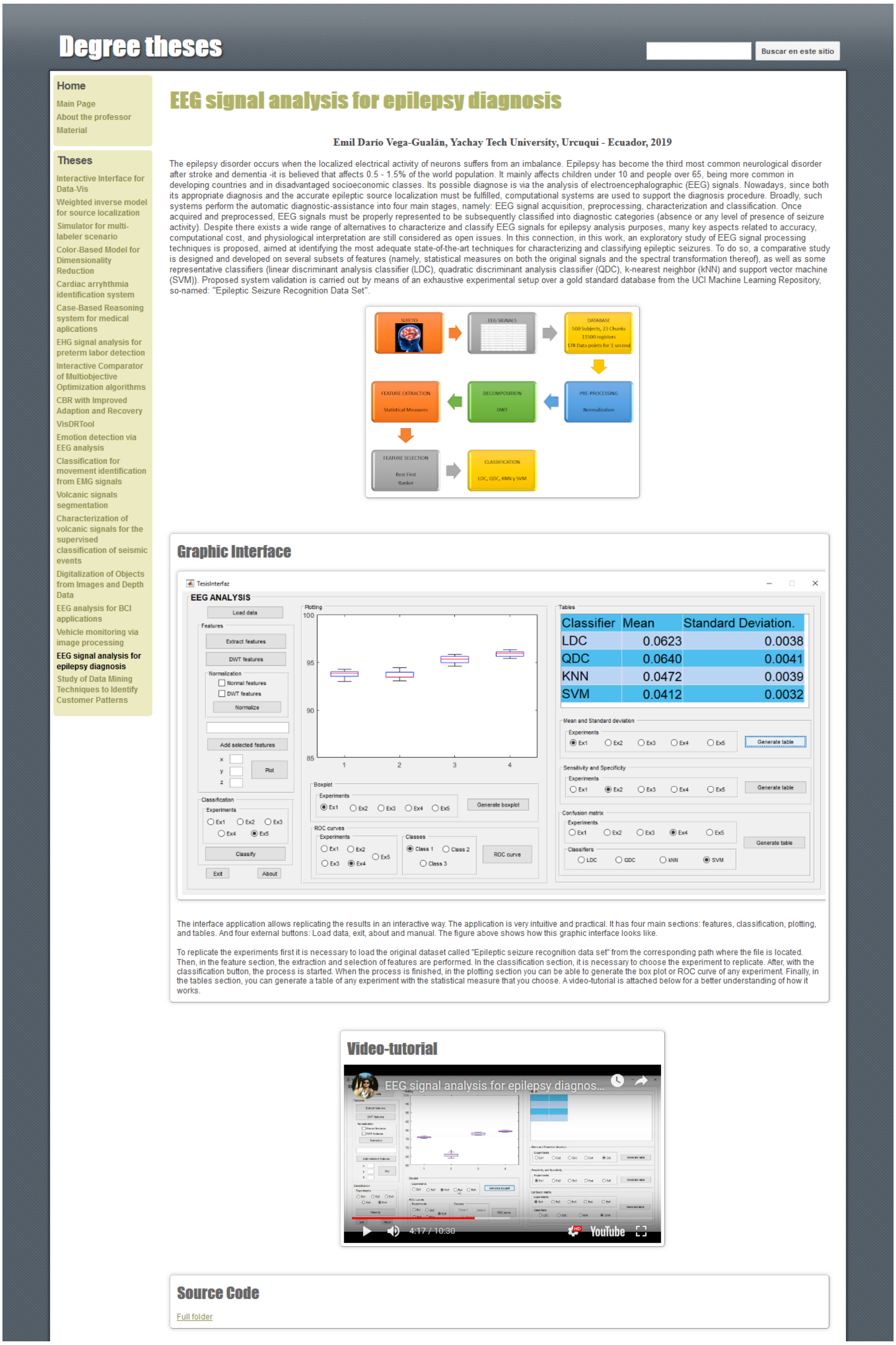
```
        figure
        boxplot([CP_total*100 res*100])
        ylim([50 100])

        CP_total2=[CP_total*100 res*100];

        %% ROC curves
        class=input('Ingrese la clase a evaluar: ');
        [scoresWldc,scoresWqdc,scoresWknn,scoresSVM]=curvasROC (A,Wldc,Wqdc,Wknn,extraccionP,SVMdata,class,3)

        %% Tables Mean and Std

        error_total =[error_total 1-res];

        error_table {1,1} =' Classifier '; error_table {1,2} ='Mean'; error_table{1,3} ='Standard Deviation.';
        error_table {2,1} ='LDC'; error_table{3,1} ='QDC'; error_table{4,1} ='KNN'; error_table{5,1} ='SVM';
        for i = 2:5
            error_table {i,2} =mean(error_total(:,i-1));
            error_table {i,3} =std(error_total (:, i-1));
        end
        disp( error_table );

        % To save the table in latex format
        % cell2latextable( error_table ,' tablas_latex ',' errorEx5');

        %% Tables Sensitivity and Specificity

        SeSVM1=SVMconf(1,1)/(SVMconf(1,1)+SVMconf(1,2)+SVMconf(1,3));
        SpSVM1 = (SVMconf(2,2) + SVMconf(2,3) + SVMconf(3,2) + SVMconf(3,3)) / (SVMconf(2,2) + SVMconf(2,3) +
              SVMconf(3,2) + SVMconf(3,3) + SVMconf(2,1) + SVMconf(3,1));

        SeSVM2=SVMconf(2,2)/(SVMconf(2,2)+SVMconf(2,1)+SVMconf(2,3));
        SpSVM2 = (SVMconf(1,1) + SVMconf(1,3) + SVMconf(3,1) + SVMconf(3,3)) / (SVMconf(1,1) + SVMconf(1,3) +
              SVMconf(3,1) + SVMconf(3,3) + SVMconf(1,2) + SVMconf(3,2));

        SeSVM3=SVMconf(3,3)/(SVMconf(3,3)+SVMconf(3,1)+SVMconf(3,2));
        SpSVM3 = (SVMconf(1,1) + SVMconf(1,2) + SVMconf(2,1) + SVMconf(2,2)) / (SVMconf(1,1) + SVMconf(1,2) +
              SVMconf(2,1) + SVMconf(2,2) + SVMconf(1,3) + SVMconf(2,3));

        LDCres=Atest*Wldc;
        [SeLDC1,SpLDC1]=testc(LDCres,'sensitivity',1);
        [SeLDC2,SpLDC2]=testc(LDCres,'sensitivity',2);
        [SeLDC3,SpLDC3]=testc(LDCres,'sensitivity',3);

        QDCres=Atest*Wqdc;
        [SeQDC1,SpQDC1]=testc(QDCres,'sensitivity',1);
        [SeQDC2,SpQDC2]=testc(QDCres,'sensitivity',2);
        [SeQDC3,SpQDC3]=testc(QDCres,'sensitivity',3);

        KNNres=Atest*Wknn;
        [SeKNN1,SpKNN1]=testc(KNNres,'sensitivity',1);
        [SeKNN2,SpKNN2]=testc(KNNres,'sensitivity',2);
        [SeKNN3,SpKNN3]=testc(KNNres,'sensitivity',3);

        SeSp_table{1,1} =' Classifier '; SeSp_table{1,2} ='Class 1'; SeSp_table{1,4} ='Class 2'; SeSp_table{1,6} ='Class 3';

        SeSp_table{2,2} ='Se'; SeSp_table{2,3} ='Sp'; SeSp_table{2,4} ='Se';
        SeSp_table{2,5} ='Sp'; SeSp_table{2,6} ='Se'; SeSp_table{2,7} ='Sp';

        SeSp_table{3,1} ='LDC'; SeSp_table{4,1} ='QDC'; SeSp_table{5,1} ='KNN'; SeSp_table{6,1} ='SVM';

        SeSp_table{3,2} =SeLDC1; SeSp_table{4,2} =SeQDC1; SeSp_table{5,2} =SeKNN1; SeSp_table{6,2} =SeSVM1;
        SeSp_table{3,3} =SpLDC1; SeSp_table{4,3} =SpQDC1; SeSp_table{5,3} =SpKNN1; SeSp_table{6,3} =SpSVM1;

        SeSp_table{3,4} =SeLDC2; SeSp_table{4,4} =SeQDC2; SeSp_table{5,4} =SeKNN2; SeSp_table{6,4} =SeSVM2;
        SeSp_table{3,5} =SpLDC2; SeSp_table{4,5} =SpQDC2; SeSp_table{5,5} =SpKNN2; SeSp_table{6,5} =SpSVM2;

        SeSp_table{3,6} =SeLDC3; SeSp_table{4,6} =SeQDC3; SeSp_table{5,6} =SeKNN3; SeSp_table{6,6} =SeSVM3;
        SeSp_table{3,7} =SpLDC3; SeSp_table{4,7} =SpQDC3; SeSp_table{5,7} =SpKNN3; SeSp_table{6,7} =SpSVM3;

        disp(SeSp_table);

        % To save the table in latex format
        % cell2latextable(SeSp_table,' tablas_latex ',' SeSpEx5');

    end
```

Appendix B

Web site

The whole work of this thesis is already uploaded in a web site for free access to documentation, algorithms, executable codes, results and other sources as articles and videos. This web site belongs to my tutor Diego Peluffo. The web site was developed in `Google Sites` and the access link for this research is the following:

`https://sites.google.com/site/degreethesisdiegopeluffo/eeg-signal-analysis-for-epilepsy-diagnosis`



Degree theses
Buscar en este sitio
Home
Main Page
About the professor
Material
Theses
Interactive Interface for Data-Vis
Weighted inverse model for source localization
Simulator for multi-labeler scenario
Color-Based Model for Dimensionality Reduction
Cardiac arryhthmia identification system
Case-Based Reasoning system for medical aplications
EHG signal analysis for preterm labor detection
Interactive Comparator of Multiobjective Optimization algorithms
CBR with Improved Adaption and Recovery
VisDRTool
Emotion detection via EEG analysis
Classification for movement identification from EMG signals
Volcanic signals segmentation
Characterization of volcanic signals for the supervised classification of seismic events
Digitalization of Objects from Images and Depth Data
EEG analysis for BCI applications
Vehicle monitoring via image processing
EEG signal analysis for epilepsy diagnosis
Study of Data Mining Techniques to Identify Customer Patterns

EEG signal analysis for epilepsy diagnosis

Emil Darío Vega-Gualán, Yachay Tech University, Urcuqui - Ecuador, 2019

The epilepsy disorder occurs when the localized electrical activity of neurons suffers from an imbalance. Epilepsy has become the third most common neurological disorder after stroke and dementia -it is believed that affects 0.5 - 1.5% of the world population. It mainly affects children under 10 and people over 65, being more common in developing countries and in disadvantaged socioeconomic classes. Its possible diagnose is via the analysis of electroencephalographic (EEG) signals. Nowadays, since both its appropriate diagnosis and the accurate epileptic source localization must be fulfilled, computational systems are used to support the diagnosis procedure. Broadly, such systems perform the automatic diagnostic-assistance into four main stages, namely: EEG signal acquisition, preprocessing, characterization and classification. Once acquired and preprocessed, EEG signals must be properly represented to be subsequently classified into diagnostic categories (absence or any level of presence of seizure activity). Despite there exists a wide range of alternatives to characterize and classify EEG signals for epilepsy analysis purposes, many key aspects related to accuracy, computational cost, and physiological interpretation are still considered as open issues. In this connection, in this work, an exploratory study of EEG signal processing techniques is proposed, aimed at identifying the most adequate state-of-the-art techniques for characterizing and classifying epileptic seizures. To do so, a comparative study is designed and developed on several subsets of features (namely, statistical measures on both the original signals and the spectral transformation thereof), as well as some representative classifiers (linear discriminant analysis classifier (LDC), quadratic discriminant analysis classifier (QDC), k-nearest neighbor (kNN) and support vector machine (SVM)). Proposed system validation is carried out by means of an exhaustive experimental setup over a gold standard database from the UCI Machine Learning Repository, so-named: "Epileptic Seizure Recognition Data Set".

Graphic Interface

Classifier	Mean	Standard Deviation.
LDC	0.0623	0.0038
QDC	0.0640	0.0041
KNN	0.0472	0.0039
SVM	0.0412	0.0032

The interface application allows replicating the results in an interactive way. The application is very intuitive and practical. It has four main sections: features, classification, plotting, and tables. And four external buttons: Load data, exit, about and manual. The figure above shows how this graphic interface looks like.

To replicate the experiments first it is necessary to load the original dataset called "Epileptic seizure recognition data set" from the corresponding path where the file is located. Then, in the feature section, the extraction and selection of features are performed. In the classification section, it is necessary to choose the experiment to replicate. After, with the classification button, the process is started. When the process is finished, in the plotting section you can be able to generate the box plot or ROC curve of any experiment. Finally, in the tables section, you can generate a table of any experiment with the statistical measure that you choose. A video-tutorial is attached below for a better understanding of how it works.

Video-tutorial

Source Code

Full folder

Figure B.1: Web site of EEG signals analysis for epilepsy diagnosis