People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University M'Hamed BOUGARA – Boumerdes

Institute of Electrical and Electronic Engineering

Department of Electronics

Final Year Project Report Presented in Partial Fulfillment
of the Requirements for the Degree of

MASTER

In **Telecommunications**

Option:
Telecommunication

Title:

EEG Signal Classification and Forecasting for Epileptic Seizure Prediction

Presented by:

AFOUN Laid

ILOUL Zakaria

Supervisor:

Dr.CHERIFI Dalila

Registration Number:……..../2019

Abstract

EEG signal recordings are increasingly replacing the old methods of diagnosis in medical field of many neurological disorders.

Our contribution in this project is the study and development of EEG signal classification and forecasting algorithms for epilepsy diagnosis using machine learning using one rhythm; for classification, an optimum classifier is proposed with only one used rhythm so that both execution time and number of features are reduced; for forecasting, the value of RMSE is minimized when using LSTM where the best hyperparameters are found.

Firstly, we used wavelet packet decomposition (WPD) to extract the five rhythms of brain activity from the public Epilepsy-EEG recordings in order to represent each signal with features vector; then we applied on it the well-known classification methods. Secondly, we implemented forecasting methods for predicting seizures states on the signals using statistics methods and LSTM.

A statistical study is done to validate the different algorithms.

Dedications

I dedicate this work to my big family members

To my parents, sister and brother

And to my beloved friends

Laid

I especially dedicate this work to my family

To all my lovely sisters and brothers

And to my beloved friends

Zakaria

Acknowledgement

We would like to take this opportunity to thank and express our sincere gratitude to our supervised, Dr.CHERIFI Dalila; for her valuable comments , suggestions and sharing of knowledge to make this report more meaningful as well as for her psychological support by encouraging us and spending time in providing instructions during the last months .

We would like to thank all the teachers and the staff of the Institute of Electrical and Electronic Engineering, for their great devotion towards their professions, and for providing a convenient working environment.

Also, we would like to thank Mr. BOUBECHIR Larbi, Mr. BOUKERMA Billal and Mr. ADJERID Chaouki for their help.

Special thanks to our beloved parents for the moral and financial support in making this workbook possible, and most of all, to our Almighty Allah, for giving his endless blessings, knowledge and strength in order to make this modular workbook possible.

Once again, we thank all those who have encouraged and helped us in preparing this workbook, and who have extended to us so much understanding , patience , and support.

Table of Contents

Table of Contents

List of Tables

List of Figures

—

List of abbreviations

EEG	ElectroEncephaloGraphy.
MRI	Magnetic Resonance Imaging.
PET	Positron Emission Tomography.
CT	Computed Tomography.
DTI	Diffusion Tensor Imaging.
WPD	Wavelet Packet Decomposition.
SVM	Support Vector Machine.
k-NN	k- Nearest Neighbors.
LSTM	Long Short Term Memory.
RMSE	Root Mean Square Error.

Introduction

The internet and digitization revolution had led to the presence of huge amount of data in hand, whether the data is previous recordings of forecast in a certain area or stock market prices over few previous hours.

The medical field had his own benefits from this revolution since the previous situation of the patient (age, weight, height and blood pressure..) all is recorded in computers data base which doctors can access and make much more accurate diagnosis and more appropriate treatment, but still some situations where diagnosis is difficult or make predictions about disease evolution is not really accurate because even the existence of huge data is hard to analyze by human being and extract some useful information. This was the situation until artificial intelligence and machine learning algorithms and recently deep learning took the role of data analyst away since it is more accurate and less time consuming.

Machine learning is about algorithms that can analyze and extract useful information and make decisions based on that while deep learning is a branch of machine learning that uses a one or more hidden layer in neural network.

There are many researches about neurological diseases in general and epilepsy in particular and related machine learning and EEG applications, but they take all the signal, this needs a long time execution and large features vector size.

The aim of this project is to study and develop EEG signal classification and forecasting algorithms for epilepsy diagnosis by using machine learning and examining only one rhythm. To carry out this project, the report has been divided into four chapters, briefly described below:

Chapter I introduces a general concepts about brain activity, acquisition systems to detect brain disorders and some familiar neurological diseases such as epilepsy. Chapter II lists some signal features that will be used in this project, and then describes some well-known classification methods. Chapter III illustrates forecasting methods used to predict signal progress, starting by statistics methods and ending with describing LSTM (Long Short Term Memory) method. Chapter IV explores the experimental investigations of this study, statistical analysis for various methods proposed as well as their performance.

Chapter I

EEG signal and brain disorders

This chapter introduces the human brain and lists the disorders that can disturb its proper functionality, it introduces also the main methods used to detect and analyze its activity.

I.1. Introduction

This chapter gives an insight about the function of the human brain; it starts with a small description of brain activity that includes the 5 different rhythms (alpha, beta, gamma, delta and theta) and their relation to different mental and physical situations then moves toward a definition of EEG signals. The chapter ends with a mention of the existing tools used to detect brain function and different diseases that can alter the normal behavior of its activity.

I.2 Brain activity

The brain is the most complex part of the body, it is the part where all neurological actions take place; brain activity is the firing that occurs between neurons in response to certain situation, the brain receives messages from sensory systems as electrochemical signals that can treat and output a decision about what should be done. There are 5 rhythms that can be extracted from an EEG signal; they exist in the brain depending on the situation of the person:

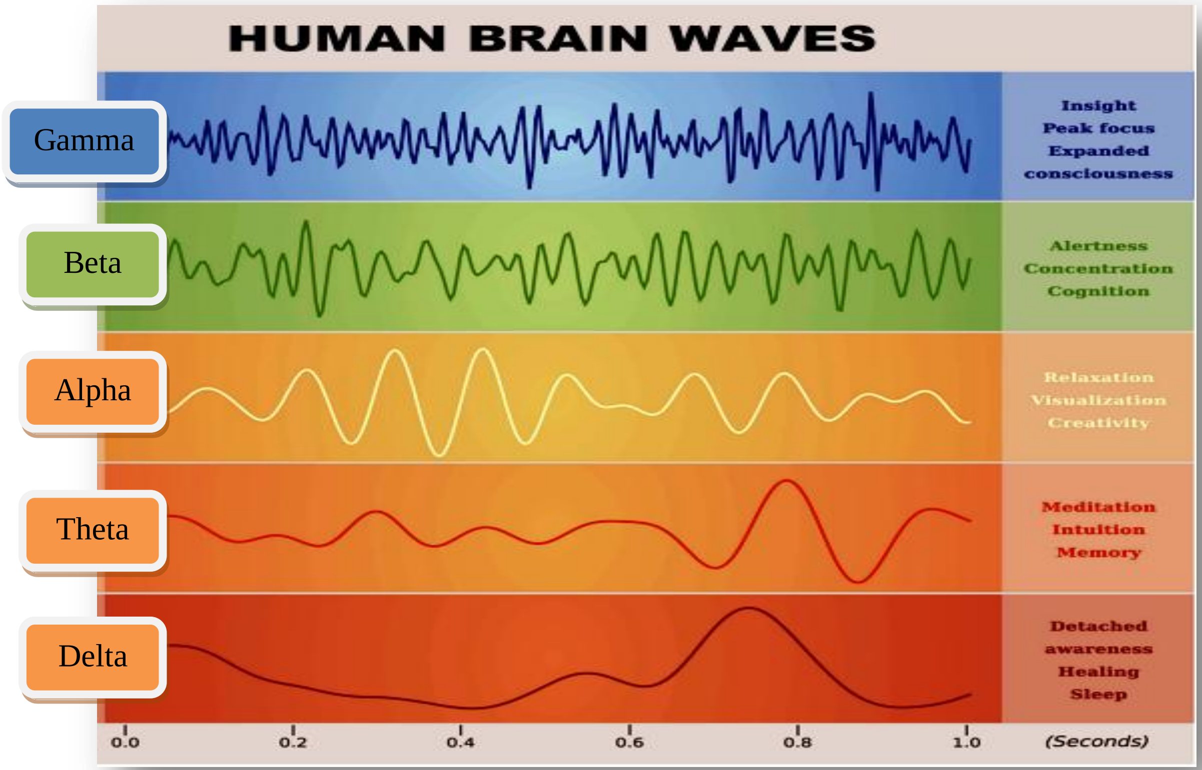

Figure I. 1: *Brain waves illustration* **[1]**.

I.2.1. Delta rhythm: the slowest rhythm that ranges between 0 and 4 Hz but the highest in amplitude, it dominates in case of sleeping and meditation, this tells us that delta waves represent inactive brain situation.

I.2.2. Theta rhythm: it ranges between 4 and 8 Hz, it appears in learning, memorizing and intuition Situations.

I.2.3. Alpha rhythm: the range from 8 to 15 Hz, it is present in case of relaxation, deep meditation and calmness.

I.2.4. Beta rhythm: the range from 12 to 30 Hz, it dominates in case of awareness and active thinking**.**

I.2.5. Gamma rhythm: 30 Hz up to 60 Hz dominates in case of happiness and positive thinking.

I.3 EEG definition

EEG stands for Electroencephalography and it is the measurement of the electrical activity of the brain for a period of time. Since it is a real-time test, it provides valuable information for scientists and physicians to see the brain function and response to various situations whether detecting some unusual wave patterns in case of disorders or to some physical (sports) or mental (meditation) activity [2].

EEG signals are recorded by placing some electrodes on the surface of the head and record the electrical potential.

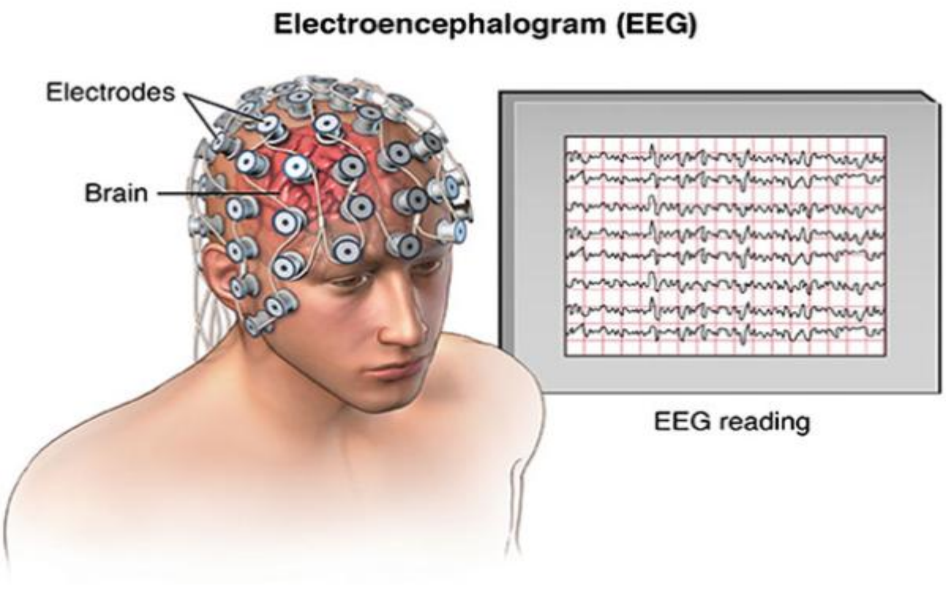

Figure I. 2: *EEG signal recording* [2].

Each two electrodes represent one channel of the recording system and are placed on a specific area of the head, a normal EEG signal amplitude ranges between 0 to 100µV.

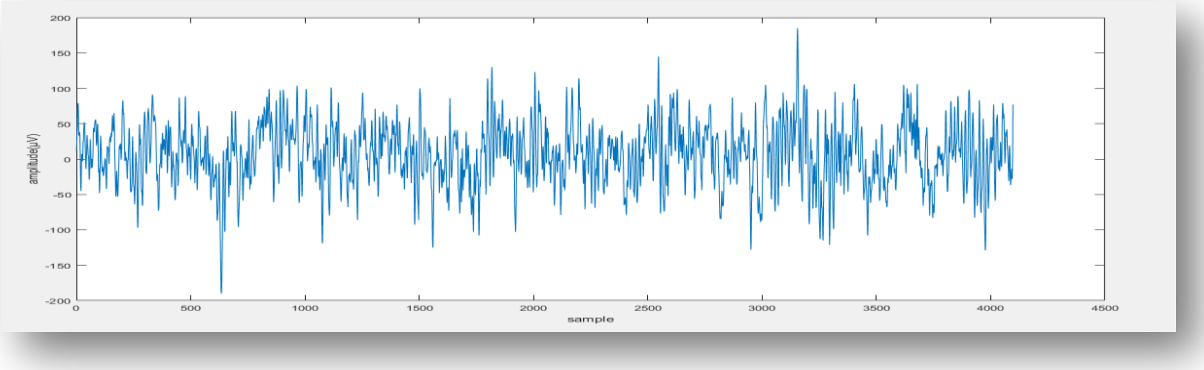

Figure I. 3: *One channel EEG signal.*

I.4 other methods used to study brain activity

- **Magnetic Resonance Imaging (MRI)**

MRI stands for Magnetic Resonance Imaging, it uses a strong magnetic field to create images of the parts of the body in question, water molecules become strictly aligned in the presence of magnetic field since the hydrogen protons absorb energy from the field and when the field is turned off, a radio signal is generated by the molecules when they turn to normal situation, this signal is recorded and turned into image [3].

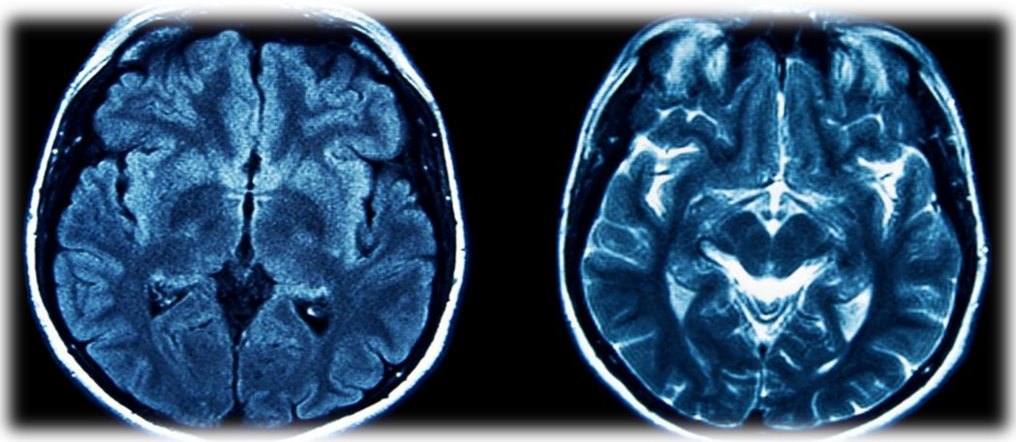

Figure I. 4: *output of an MRI system* [4].

- **Emission Tomography and Computed Tomography (PET & CT scans)**

They stand for Positron Emission Tomography and Computed Tomography, respectively. PET scan is based on a tracer that is injected on the body; this tracer is a gamma rays molecule, the output of the system draws the concentration of the molecules in the body [5]. CT scan is an enhanced version of PET scan system that uses many X-rays measurements to make a cross-sectional image of the specified area [6].

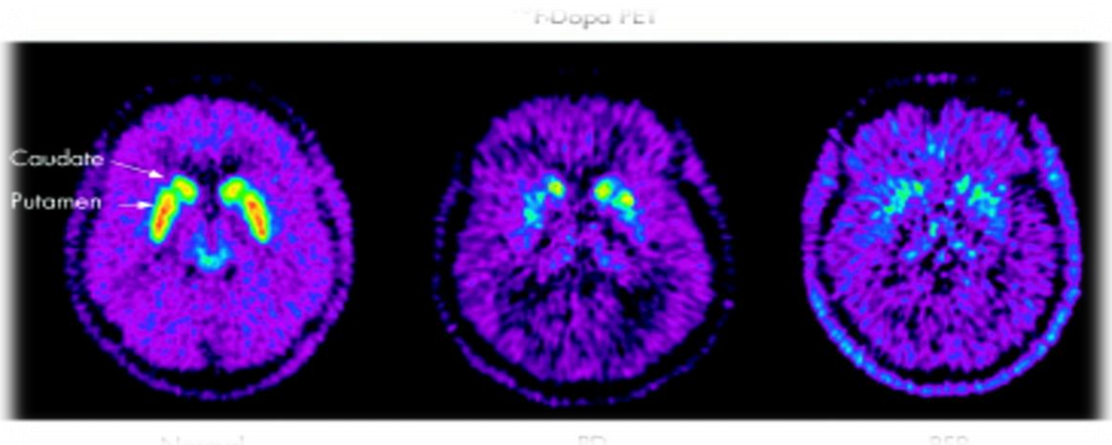

Figure I. 5: *Output of PET scanner* [7].

- **Diffusion Tensor Imaging (DTI)**

DTI stands for Diffusion Tensor Imaging; it is popular for imaging the white matter of the brain, the DTI diffuses water movement within the brain which allows doctors to isolate regions that are not functioning properly. The measured quantity is the diffusivity or diffusion coefficient, a proportionality constant that relates diffusive flux to a concentration gradient and has units of mm^2. Unlike the diffusion in a glass of pure water, which would be the same in all directions (isotropic), the diffusion measured in tissue varies with direction (is anisotropic). The measured macroscopic diffusion anisotropy is due to microscopic tissue heterogeneity. In the white matter of the brain, diffusion anisotropy is primarily caused by cellular membranes, with some contribution from the packing of axons. Anisotropic diffusion can indicate the underlying tissue orientation [8].

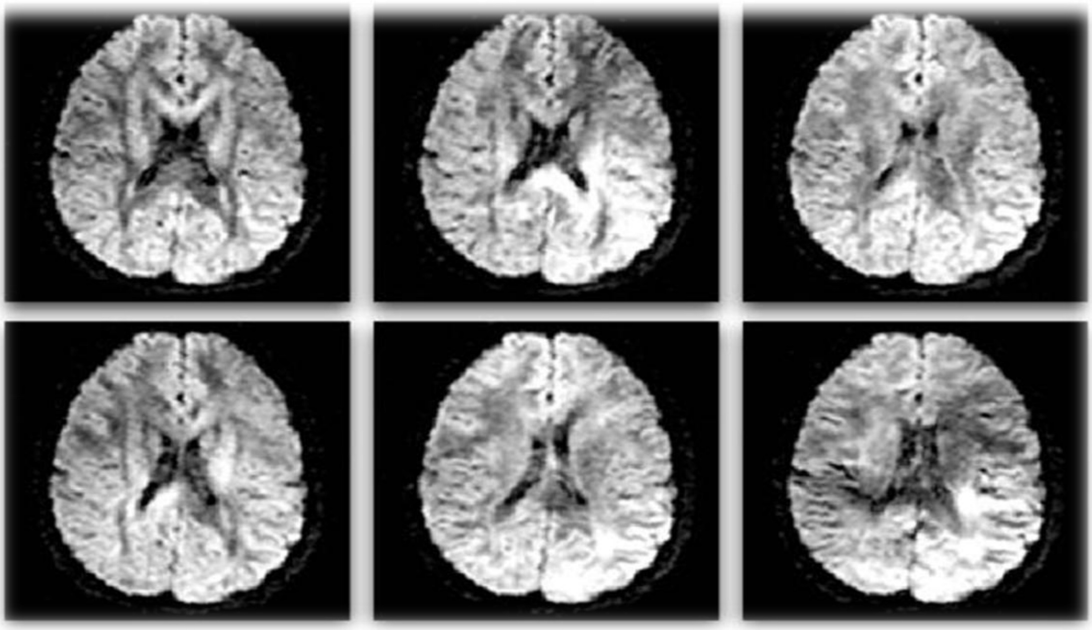

Figure I. 6: *Output of DTI system* [8].

I.5 Advantage of EEG method

All the previous methods use radioactive process or strong magnetic field exposure which can cause damage to the processed part of the body if used frequently unlike EEG that has no side effects on the brain; it is simple and harmless diagnosis.

Scanning methods do not give the appropriate diagnosis results for most of neurological diseases; unlike EEG that gives electrical activity of brain, they output an image of the brain, this is more convenient with tumors situation.

I.6 Brain disorders

I.6.1 Dementia

Dementia refers to neuro-degenerative disorder caused by death of neurons; it is associated with decreasing of brain functions (memory, concentration, executive functions…). It can be classified into several types: Alzheimer’s disease (AD), Parkinson’s disease (PD), vascular dementia and front temporal dementia, and they are very known diseases to old people.

I.6.2 Brain tumor

A tumor is a collection of abnormal cells in the brain that makes parts of the brain dysfunctional since it exerts a huge pressure. The source of the tumor can be either from the brain itself known as internal brain tumor (ex: internal bleeding) or it can start in other part of the body and spread all the way to the brain known as secondary brain tumor (ex: cancer).

I.6.3 Stroke

A stroke happens when the proper blood flow is being interrupted either by broken vessel (hemorrhagic) or blood clot (ischemic), this situation leads to neurons death, the death of huge number of brain cells causes permanent brain dysfunction depending on the area injured, some of them are: loss of memory, paralysis, coma or even death.

I.6.4 Autism

Autism is a neurophysiologic disease that starts in the early years of a person, it is concerned with deficits in social relations and it can stay with the patient all his life if there was no treatment, EEG analysis is the mostly method used to diagnose autism.

I.6.5 Epilepsy

Epilepsy is a brain disorder that takes the form of seizures caused by sudden change in the brain electrical activity that may develop from imbalanced brain wiring, an imbalance in neurotransmitters production in cells or a combination of both. As a result, behaviors such as losing consciousness, jerky movements, temporary loss of breath and memory loss are experienced by the patient, the number of seizures and the duration vary from one patient to another. Epilepsy is one of the most common neurological diseases experienced worldwide [2].

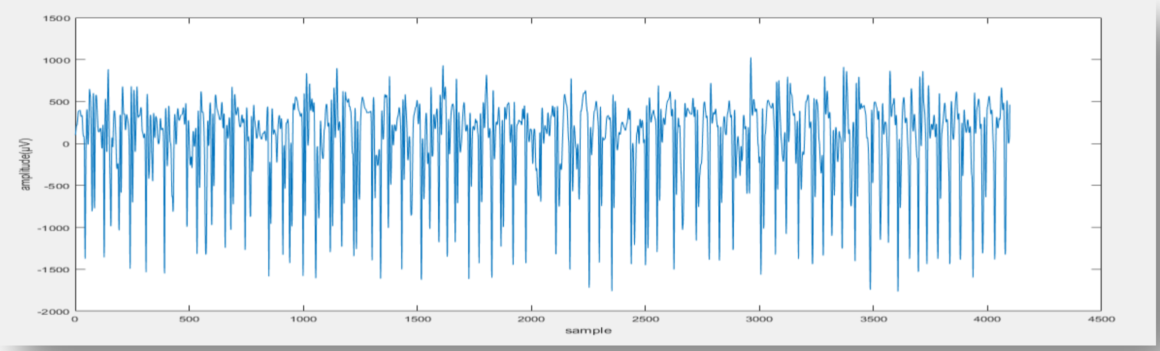

Figure I. 7: *EEG signal having epilepsy.*

I.7 Summary

This section was a small introduction about human brain and the ways we can use to detect its functionality showing the advantage of EEG method. Also some neurological disorders are explained.

Chapter II

Signal features and classification methods

This chapter provides an insight to data classification with different steps involved; we focus on signal features extraction and classification stages; for classification, the classification is done using supervised machine learning algorithms, a detailed description of those algorithms is included.

II.1. Introduction

Classification is the problem of identifying to which class an observation belongs by the classifier. In data classification, the data set may contain 1-D signals like EEG in our case, a 2-D signal like image classification. Before classification takes place, a preprocessing is performed to filter any undesired noise that comes from acquisition system. After preprocessing phase, the way to see the patterns among the data elements is by extracting some features from them such that they demonstrate the differences to the classifier. Classification has many applications in computer vision, medical imaging, speech recognition, biometric identification, pattern recognition…etc.

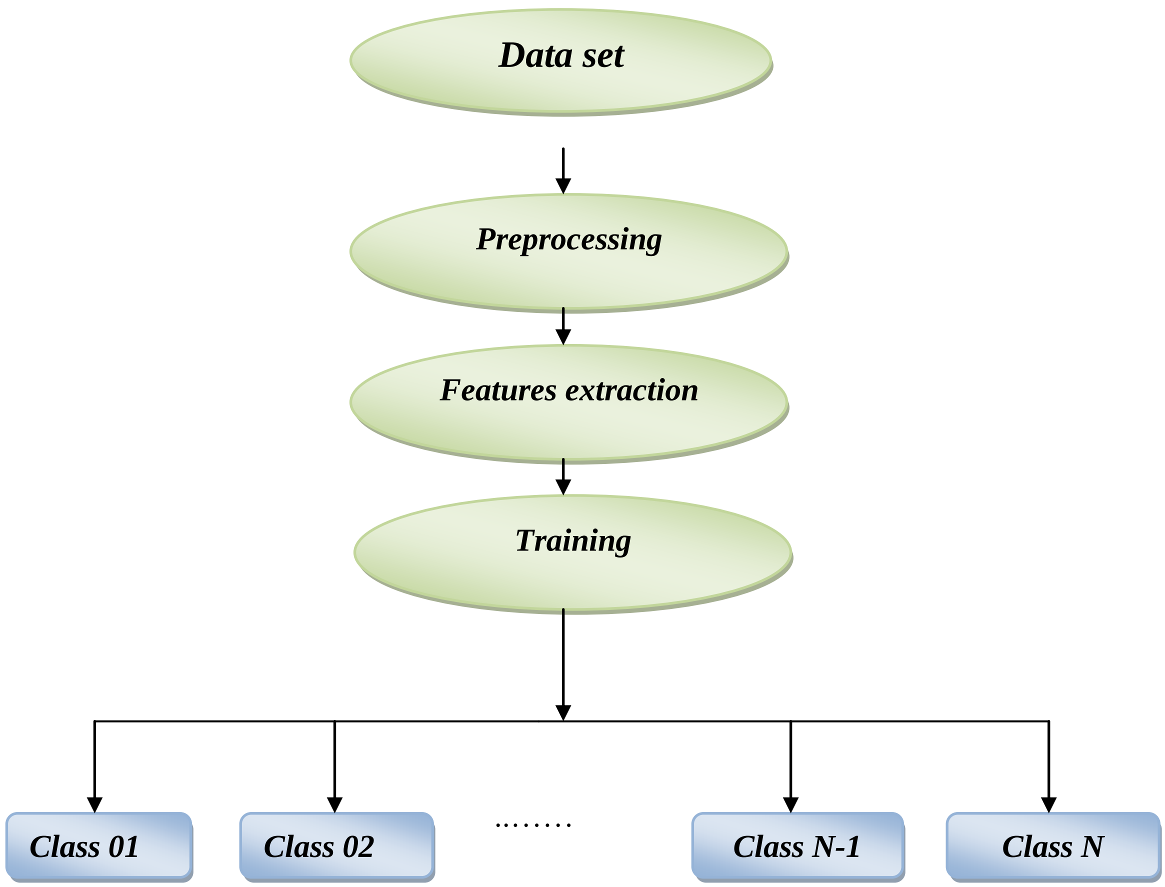

***Figure II. 1**: Illustration of classification process.*

EEG is a 1-D signal type classification problem; this means that the features should be statistical parameters.

II.2signal features

Each signal can be represented by features that define its properties whether its time properties or frequency properties or energy properties [2]:

- **The mean**

It defines the average value of the data in hand, the value of the mean helps to identify where the data points are concentrated in classification, it is the sum of the numbers divided by how many numbers there are.

$$\mu=\sum_{k=1}^{N}\frac{x_k}{N} \qquad \text{(II.1)}$$

Where μ is the mean; x_k are the values in the data set; N is the number of elements.

- **The variance**

In.probability theory and statistics;variance is Expectation.of.squared.deviation of a variable from its mean. Informally, it measures how far a set of numbers are spread out from their average value. Variance has a central role in statistics, where some ideas that use it include descriptive statistics, statistical inference, hypothesis testing. Variance is an important tool in the sciences, where statistical analysis of data is common. Variance is represented by σ^2 and its mathematical expression is [9].

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2 \qquad \text{(II.2)}$$

- **The standard deviation**

It is the measure used to quantify the amount of dispersion in a data set or by how much the data set deviate from the mean and it is the square root of the variance [10].

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2} \qquad \text{(II.3)}$$

Where σ^2 is the variance, x_k are the values in the data set and μ is the mean.

- **First quartile, third quartile and inter quartile (IQR)**

The first quartile is the median of the first half of the data set while third quartile is the median of the second half. The inter quartile is a measure of the dispersion of the data; it is the difference between first and third quartiles.

- **The median (second quartile)**

Median is the middle number in a sorted list of numbers. To determine the median value in a sequence of numbers, the numbers must first be arranged in value order from lowest to highest. If there is an odd amount of numbers, the median value is the term that is in the middle, with the same amount of numbers below and above. If there is an even amount of numbers in the list, the middle pair must be determined, added together and divided by two to find the median value. The median can be used to determine an approximate average, or mean. The median is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values. The median of a sequence can be less affected by outliers than the mean [11].

If the data set has odd number of values then:

$$M = x_{\frac{N+1}{2}} \tag{II.4}$$

If the data set has even number of values then:

$$M = \frac{1}{2}\left[x_{\frac{N}{2}} + x_{\frac{N+2}{2}}\right] \tag{II.5}$$

Where M is the median and N is the number of elements in the set.

- **The mode**

The mode is a statistical term that refers to the most frequently occurring number found in a set of numbers. The mode is found by collecting and organizing data in order to count the frequency of each result. The result with the highest count of occurrences is the mode of the set, also referred to as the modal value [12].

The features of minimum, maximum, first quartile, second quartile (median) and third quartile are known as five number summary. They provide a clear representation of the characteristics of a dataset.

II.3. Supervised machine learning algorithms

Supervised machine learning is an interdisciplinary field of study of statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions relying on patterns in data. The algorithms build a mathematical model based on sample data (training data) to make decisions about the testing data without being explicitly programmed.

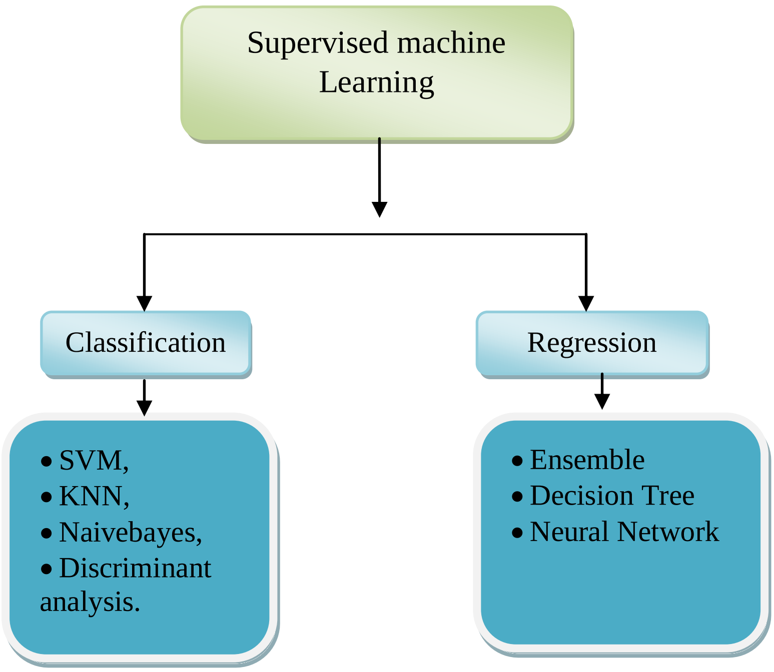

Figure II. 2*: The used supervised Machine learning methods.*

- **Classification**: It predicts discrete number of values. In classification the data is categorized under different labels according to some parameters and then the labels are predicted for the data. Classifying emails as either spam or not spam is example of classification problem.

- **Regression**: It predicts continuous valued output. The Regression analysis is the statistical model which is used to predict the numeric data instead of labels. It can also identify the distribution trends based on the available data or historic data. Predicting a person's income from their age, education is example of regression task.

The used supervised learning algorithms are described below:

II.3.1. Support Vector Machine (SVM)

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The algorithm creates a line or a hyper plane which separates the data into classes, it takes the data as an input and outputs a line that separates those classes if possible, and we find the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. The goal is to maximize the margin. The hyper plane for which the maximum margin is the optimal hyper-plane [14] Consider x_i as the element and y_i is the class label with only two possibilities, the:

$$\boldsymbol{y_i} = \boldsymbol{sign}\,[\boldsymbol{w^T}\,.\,\boldsymbol{\psi}(\boldsymbol{x_i}) + \boldsymbol{b}] \qquad \text{(II.6)}$$

Where w is the weight vector, b is the bias term and $\psi(x)$ is the nonlinear mapping function that maps the input data into a higher dimensional feature space.

The weight vector w and bias term b need to be determined. The weights are expressed as:

$$\boldsymbol{w} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{\alpha_i}\ \boldsymbol{\psi}(\boldsymbol{x_i}) \qquad \text{(II.7)}$$

Here $\boldsymbol{\alpha_i}$ denote Lagrange multipliers; Lagrange multiplier is a strategy for finding the local maxima and minima of a function subject to equality constraints [2].

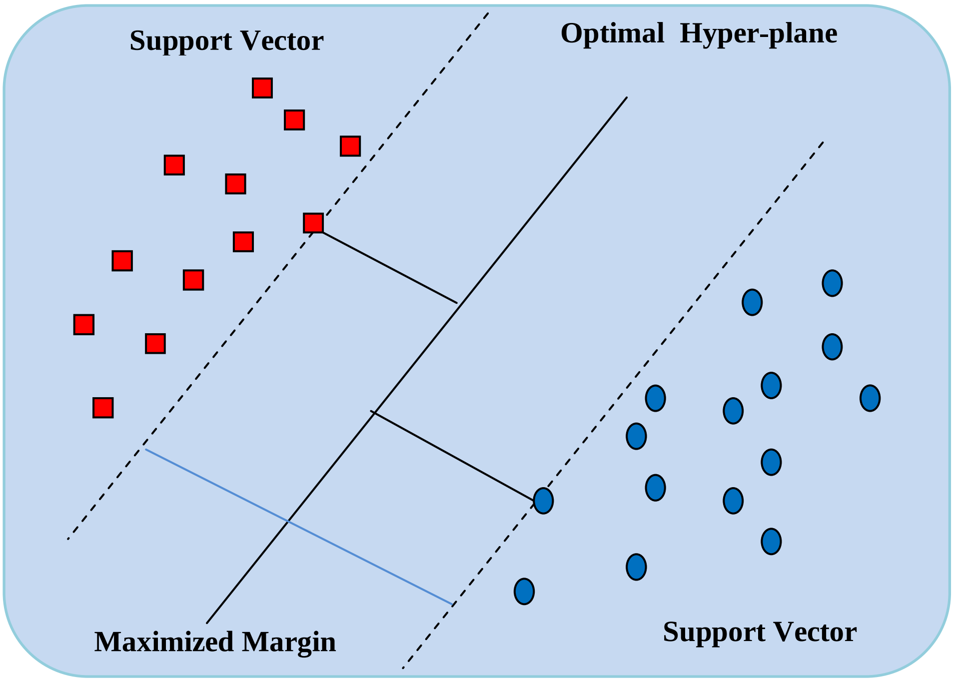

Figure II. 3: *Illustration of SVM*

The bias (Offset) is an extra input and it is always 1, and has its own connection weight. This makes sure that even when all the inputs are none (all 0's) there will be an output. The weight describes the strength of the input x_i.

II.3.2.k-Nearest Neighbor (k-NN)

In pattern recognition, the k Nearest Neighbors algorithm (k-NN) is a method used for classification and regression.

- In classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small and represents the number of neighbors). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

- In regression, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors [15].

Let us suppose that we have a data set comprising N_i points in class C_i with N points in total, if we wish to classify a new point **x**, we draw a sphere centered on x containing precisely *n* points irrespective of their class. Suppose this sphere has volume V and contains K_i points from class C_i then [2]:

$$P\,(x|\,C_i) = \frac{\mathbf{K_i}}{\mathbf{V.N_i}} \qquad \text{(II.8)}$$

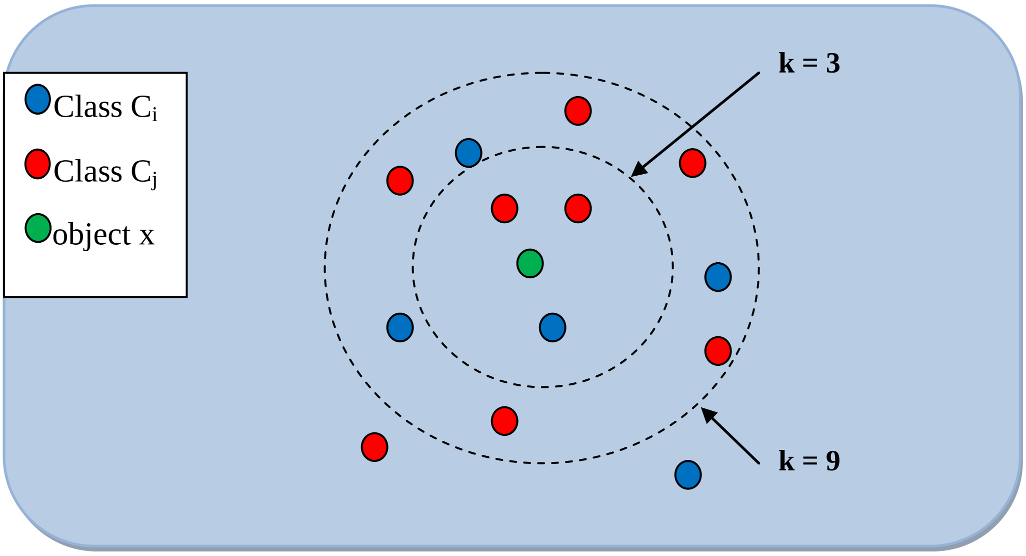

Figure II. 4*: k-NN algorithm for k = 3 and k = 9.*

II.3.3.The decision tree algorithm

Decision tree builds classification models in the form of tree structure; it breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is developed. The final result is a tree with decision nodes and leaf nodes. Decision node has two or more branches. Leaf nodes represent classification or decision. Three steps are required to build the tree:

1) Splitting: partitioning the data set into subsets.
2) Pruning: shortening the branches of the tree.
3) Tree selection: finding the smallest tree that fits the data [16].

II.3.4. Discriminant analysis

Discriminate analysis is a dimensionality reduction technique which aims to reduce the dimensions by removing the redundant and dependent features. This can be achieved in three steps:

1) Calculate the separation between different classes (the distance between the mean of different classes).
2) Calculate the distance between the mean and sample of each class which called within-class variance.
3) Construct the lower dimensional space that maximizes the between-class variance and minimizes the within-class variance [17].

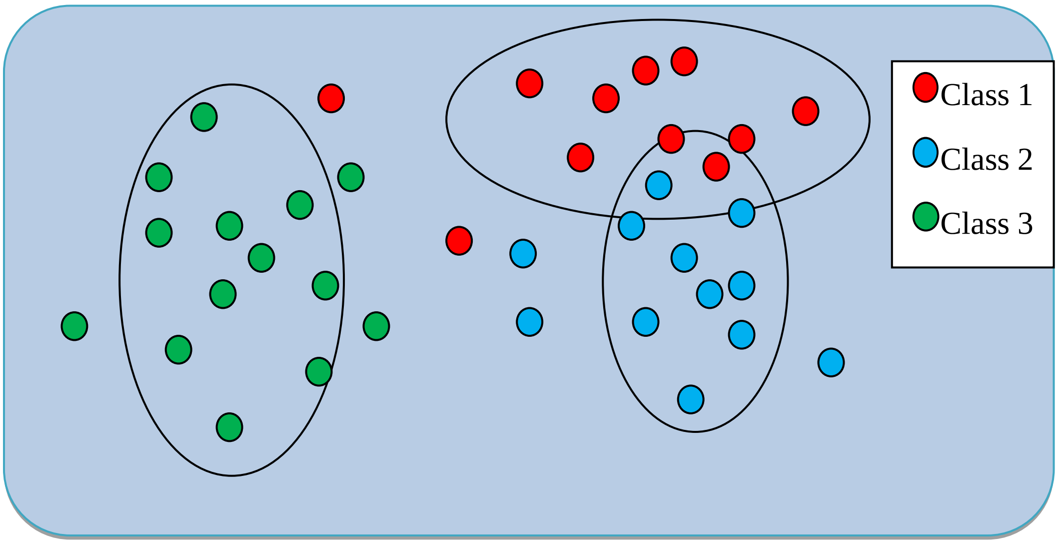

***Figure II. 5**: Illustration of discriminant analysis.*

II.3.5.Naïve Bayes

Naïve Bayes is a classifier that predicts membership probabilities for each class such as the probability that given record data point belongs to particular class based on Bayes theorem. Bayes theorem works on conditional probability. Conditional probability is the probability that something will happen given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge. Below is the formula for calculating the conditional probability.

$$P(H|E) = P(E|H)P(H) / P(E). \quad \text{(II.9)}$$

Where P(H) is the probability of hypothesis H being true. This is known as the prior probability. P(E) is the probability of the evidence(regardless of the hypothesis). P(E|H) is the probability of the evidence given that hypothesis is true P(H|E) is the probability of the hypothesis given that the evidence is true. This is also known as Maximum A Posteriori (MAP) [18].

II.3.6.Logistic regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic/sigmoid function [19]. There are three types of Logistic Regression:

- **Binary Logistic Regression:** the categorical response has only two 2 possible outcomes.
- **Multinomial Logistic Regression:** three or more categories without ordering.
- **Ordinal Logistic Regression:** three or more categories with ordering.

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes; say, if value is greater than threshold, then classify element as class 2. Decision boundary can be linear or non-linear [20].

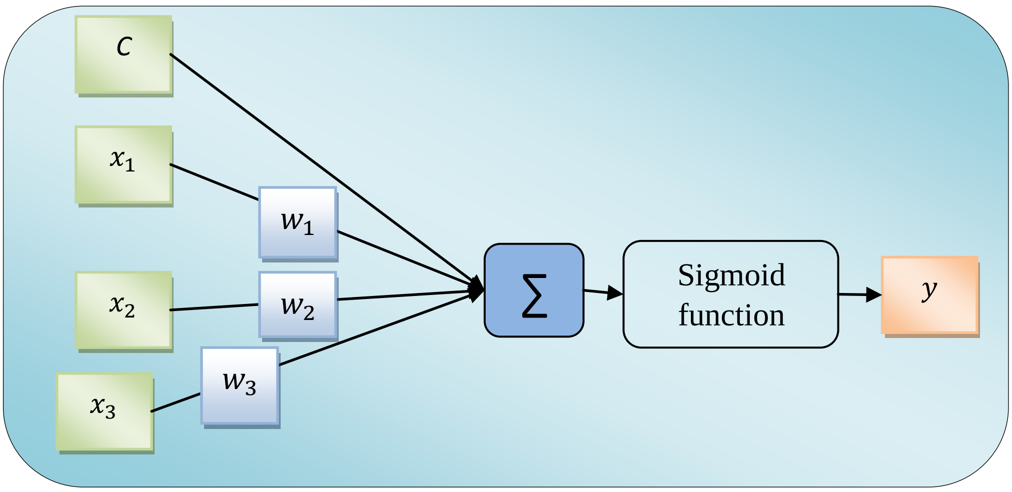

Figure II. 6: *Illustration of logistic regression*

According to the figure the output expression is:

$$y = sign\,(C + x_1.w_1 + x_2.w_2 + x_3.w_3) \quad \text{(II.10)}$$

where x_1, x_2 and x_3 are the independent variable; w_1, w_2, and w_3 are the weights; C is the bias.

II.3.7.Ensemble

Ensemble methods are a combination of several classification methods that already exist in one predictive model. The purpose of that is to decrease the variance (bagging), decrease the bias (boosting) or improve prediction (stacking).Ensemble methods are classified into two categories:

- **Sequential methods:** where the model tends to exploit the dependency between classes.
- **Parallel methods:** to exploit the independencies between them [21].

II.4. Summary

This chapter briefly describes classification process of a signal and walks through different types of machine learning algorithms and then moving to supervised classification with a detailed description of some methods.

Chapter III

Forecasting for EEG signal

In this chapter we highlight the statistics methods and LSTM for signal forecasting, a brief description of treating sequences data using LSTM is also included.

III.1 Introduction

Prediction or forecasting is a process that gives a future values based on previous observations for a certain phenomenon with a certain probability of accuracy; it deals mostly with time series data (1D signal, image or videos). Predictability is very useful for patients with sudden diseases such as epilepsy and heart attacks. Often the patient may be in the stairs or work in a high place and consequently the results are greater than the disease itself. It is therefore good to make a device that predicts the following condition for the patient to avoid dangerous situations. Before deep learning was available, scientists initially treated the signs in order to predict the patient's next condition, but they gave moderate results and a very short prediction period that did not allow the patient to initial work or change his position.

III.2 Forecasting Methods

There are many methods for forecasting; those methods initially were used in and Power Systems [22]**,** after that they were used in weather [23]and other sciences for forecasting the future values.

III.2.1 Statistics methods

Statistics methods are utilize to predict the next samples; prediction is occurred at each channel (channel after channel).There are several statistic methods among them [24]:

- **Naïve forecasting**: From its name the predicted values are the last observed value.
- **Average forecasting:** The expected value equal to the average of all previously observed values.
- **Moving average:** The expected value equal to the average of last few samples only.
- **Holt's Linear Trend method:** In additional to the level of the signal (average) this method takes into account the trend of the signal; so the expected values are composed from the level and the trend.
- **Holt-Winters Method:** In additional to the level of the signal (average) and the trend this method takes into account the seasonality of the signal; so the expected values are composed from the level, the trend and the seasonality.

III.2.2 Artificial neural networks

Inspired by the way biological neural networks in the human brain process information, artificial neural networks are a computational approach for problems in which the solution of the problem, or finding a proper representation, is difficult for traditional computer programs. Such systems can be trained from examples, rather than explicitly programmed [28].

Artificial neural network consists of a set of nodes similar to a human neuron. Each inputs X_i, has its correspondent weight W_i, bias B_0 and outputs Y_o. The node applies a non-linear activation function φ,to the weighted sum of product to produce the node output Y_0 (III.1), the activation function introduce a non-linearity to the output because most of real data are non-linear, so the network can learn from these non-linear representation.

$$Y_o = \varphi(\sum W_i.X_i + B_0) \qquad (III.1)$$

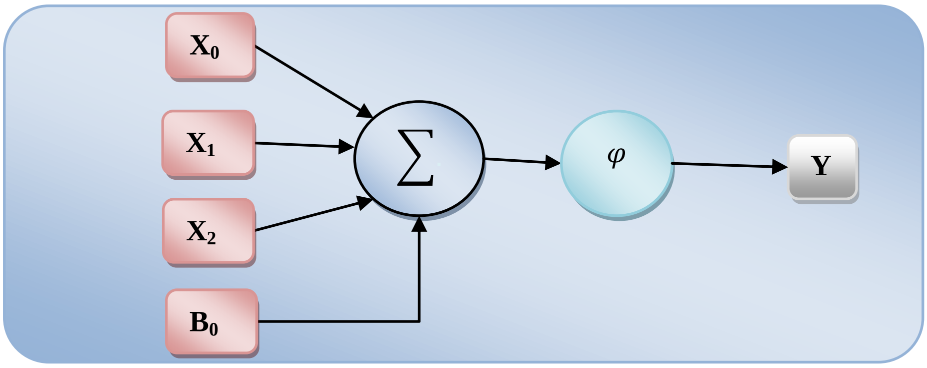

Figure III. 1*: One node consists from three inputs, bias and one output Activation function*

There are many activation functions each one for a specific purpose each one has its advantage and disadvantage here a list of the most important activation functions:

- **Sigmoid:** They have nice interpretation as a saturating "firing rate" of a neuron There are two problems of this function, the first is saturated neurons "kill" the gradients, the second comes from the fact that Exp () is a bit compute expensive.

$$\sigma(x) = \frac{1}{(1-e^{-x})} \qquad (III.2)$$

Tanh: This activation function is very useful in the recurrent network this is because of squashes numbers to range [-1, 1] and zero centered (nice).the problem of this function is that it kills gradients when saturated.

$$H(x) = tanh(x). \tag{III.3}$$

- **ReLU:** This activation function is very useful in the convolution neural network in computer vision. This function does not saturate converges much faster than sigmoid/tanh in practice.

$$F(x) = max\ (0, x) \tag{III.4}$$

A. Neural network architecture

In applied cases, a large group of the nodes are linked in form of layers, input layer, output layer, and the hidden layer, each layer is connected to the next with weights, which are multiplied to the output of the nodes in the previous layer. Additionally each node has a bias. The output of the network Y can be obtained by flowing in the network from the input layer through the hidden layer (Feeding-forward in the network), the output of output layer:

$$\boldsymbol{Y = \sigma\ (\textstyle\sum_{j=1}^{M} W_j h(\sum_{i=1}^{d} W_i X_i + B_0) + B_1)} \tag{III.5}$$

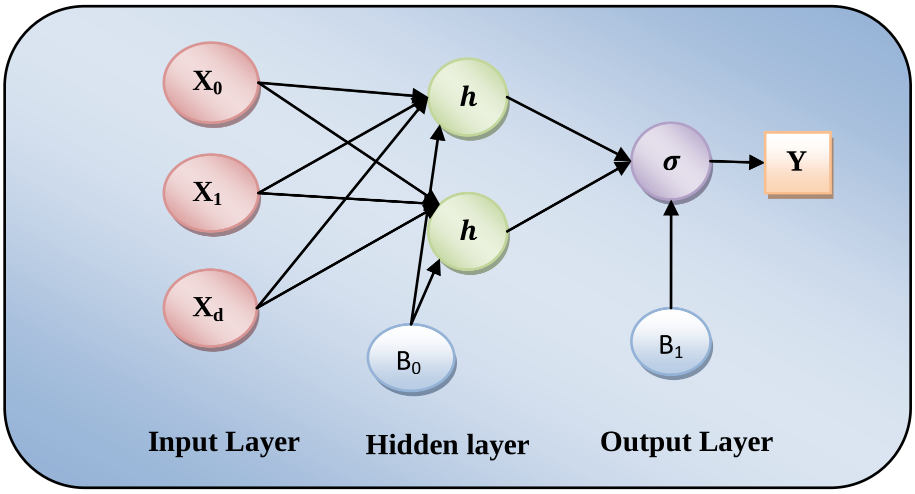

Figure III. 2*: Network diagram for neural network. The input, hidden, and output variables are represented by nodes and the weight parameters are represented by links between the nodes*

B. The loss function

Neural Network is a supervised learning, means that each input data has its correspondent label called the target value T_n that the output should be take for a deferent sample of the same class, the data samples form a certain distribution, so for the same class the output Y of the Neural Network takes various values according to the data distribution. The loss function is the function that is used to calculate the Error between the target value T and the output distribution values Y. there are three main loss functions:

- **Cross-entropy:** it is used for binary classification, means that the output layer contains one output takes a binary target values, the loss function in this case dealing well with logistic sigmoid output activation function [29].

- **Multiclass cross-entropy:** it is used for multiclass classification, means that the output layer contains multi-output nodes, the loss function in this case dealing well with soft-max activation function output activation function [29].

- **Mean square error:** it is used for regression purpose the output activation function should be linear activation function, and mean square error is used in the output layer [29].

C. Training

So far, we have viewed neural networks as nonlinear functions those transform from a vector X of input variables to a vector Y of output variables. A simple approach to the problem is to determining the network weights. If there are two deferent dataset each of them has N samples, then the aim of training is to get the proper boundaries between them by varying the weight such that the Loss function will be as minimum as possible.

The first step in training is to choose random initial values of the weights. The value of Y is calculated using feed forward method; calculate the gradient using Back propagation. Multiply the gradient by a negative real number called learning rate. The new values of the weights are calculating by adding all the previous parameter together.

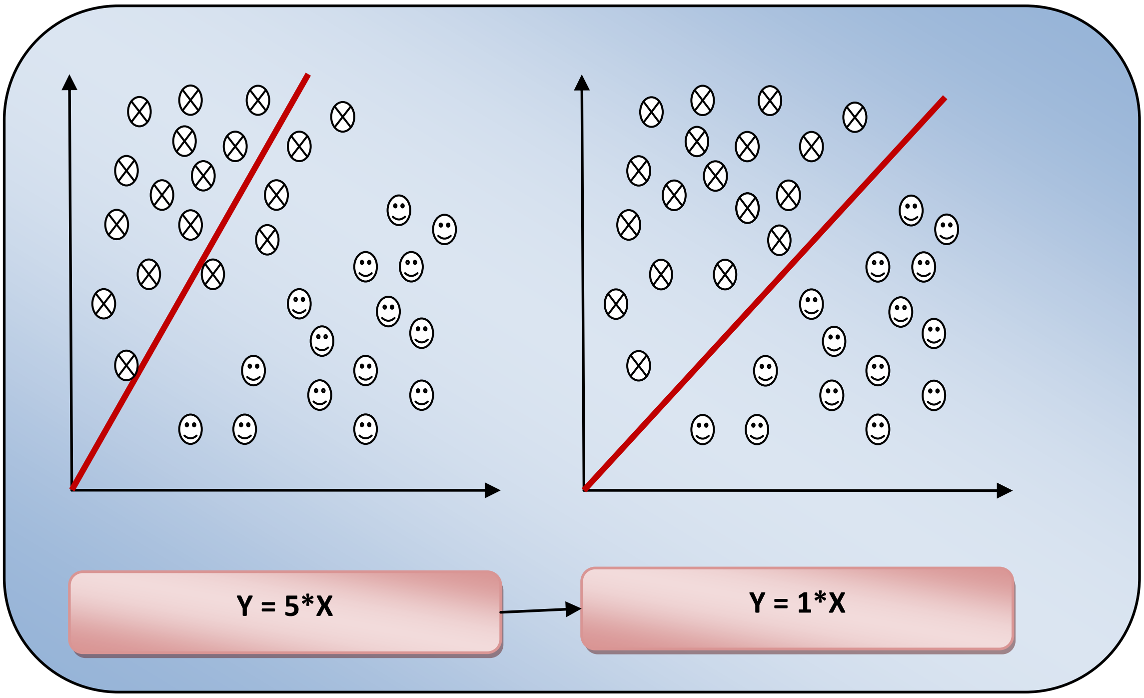

***Figure III.* 3**: *An example of how the weights adjust the boundary between two classes*

Check the value of the gradient if it is null. Repeating the previous steps until the loss will be minimum (the number of repetition is known by the number of epoch.

$$\boldsymbol{W^{i+1}} = \boldsymbol{W^{i}} - \boldsymbol{\alpha} * \boldsymbol{\nabla E(W)} \tag{III.6}$$

A. Optimization algorithm

To get the minimum error the derivative of the error function is calculated with respect to the weights. In multiple dimensions, the gradient is the vector of (partial derivatives) along each dimension. This method is called Stochastic Gradient Descent SGD. Because the high nonlinearity of the network there are many positions where the

gradient will be zero and the loss will not be the minimum (local minimum), the actual minimum position of the loss is called global minimum. For this reason there are many methods to find the global minimum. The most useful optimum methods in recent year is Adam optimum method , The Adam optimization algorithm is an extension to stochastic gradient descent that has recently seen broader adoption for deep learning applications in computer vision and natural language processing [30].

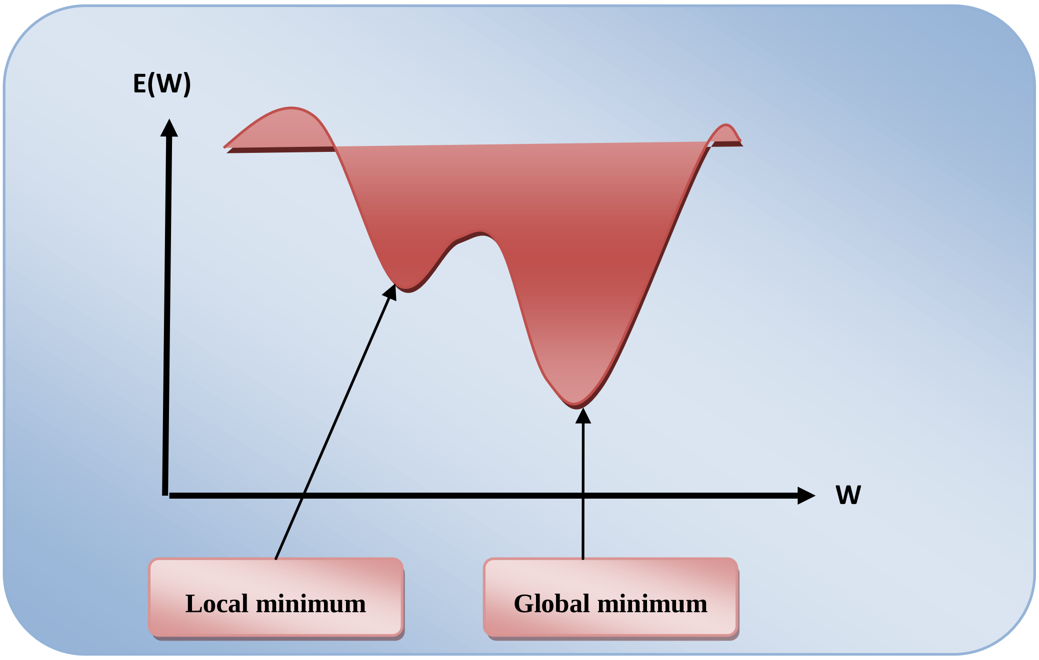

Figure III. **4***: Local and global minimum.*

B. Error Back propagation

So far we have seen that for best training the gradient of the loss function with respect to the network weights should be calculated in order to minimize the loss, unfortunately in a very complex neural network, It is very to calculate the gradient using the analytical way; instead a partial derivative of the loss function is calculated with respect to network weight, after feed forward parameter a_j, and Y_k are calculated. The back-propagation parameter δ_k, $\delta_j$and $\frac{dE_n}{dw_{ij}}$for each node in the network Figure III.5 where n is the nth pattern.

The total error E can then be obtained by repeating the above steps for each pattern in the training set and then summing overall patterns [29].

$$\frac{\partial E}{\partial wji} = \sum_n \frac{\partial E_n}{\partial wji} \quad \text{(III.7)}$$

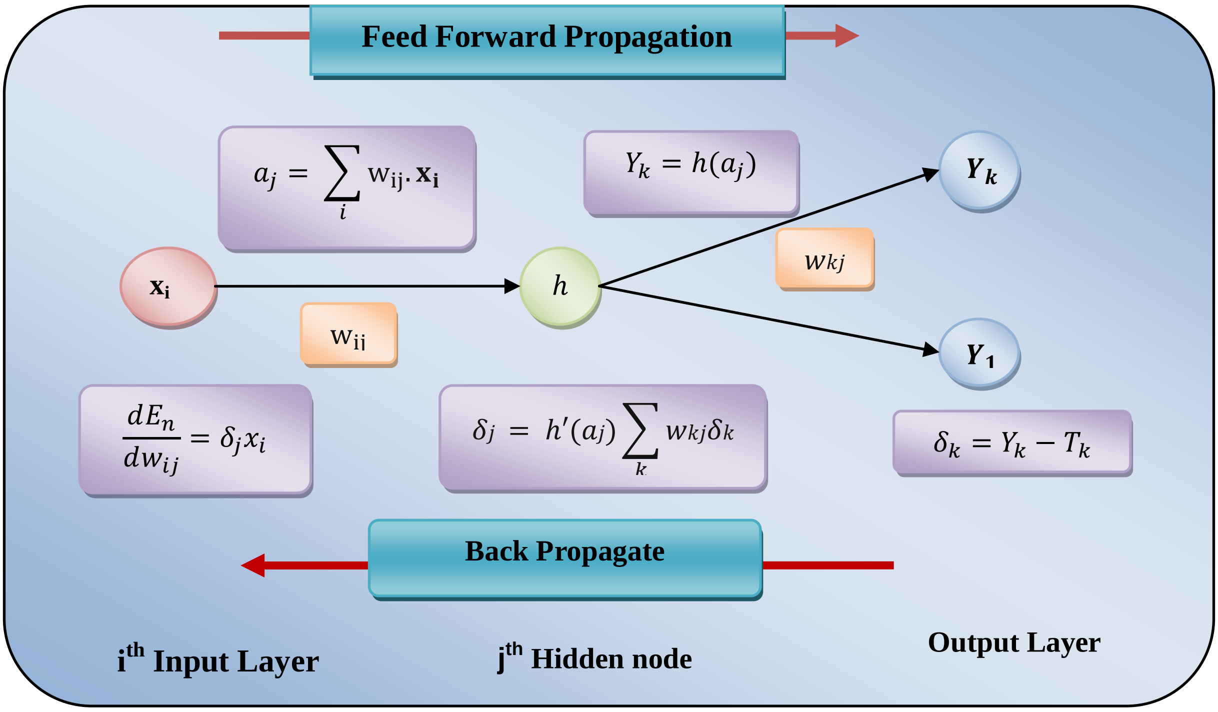

Figure III. **5***: Illustration of forward propagation and back propagate steps in the Back-propagation algorithm.*

III.2.2.1 Long Short Term Memory cell

Long Short Term Memory cell is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed-forward neural networks, LSTM has feedback connections that make it a "general purpose computer" (that is, it can compute anything that a Turing machine can). It can not only process single data points (such as images), but also entire sequences of data (such as speech or video) [31]. LSTM is power full deep learning method for sequential data and time series data. Initially LSTM is a special recurrent neural network (RNN), means that the last time output feed backs to the input, the output is a function (activation function) of the presence input and the last state of the output.

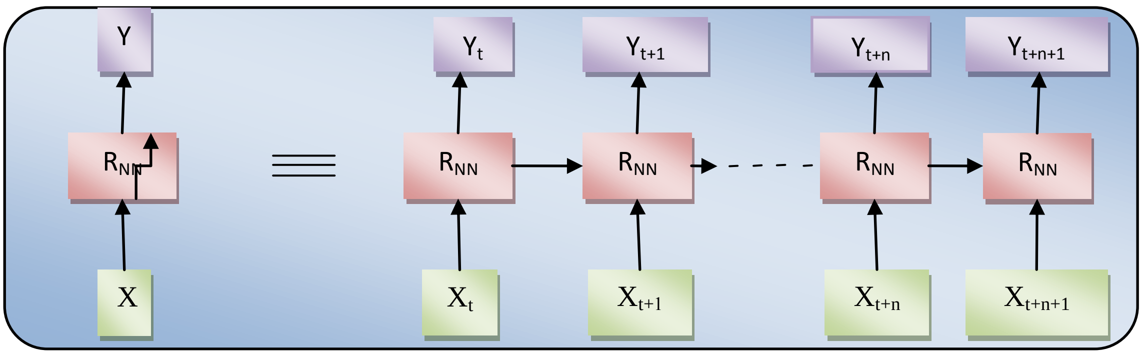

Figure III. **6***: One node recurrent network and its equivalent after t+n+1 iteration*

The idea is to avoid direct derivation with tanh and control the amount of forgetting a data, writing in the output. according to the following equations

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ X_t \end{pmatrix} \quad \text{(III.22)}$$

$$c_t = f \cdot c_{t-1} + i \cdot g \quad \text{(III.23)}$$

$$h_t = o \cdot \tanh(c_t) \quad \text{(III.24)}$$

Where input gate whether to write to cell, f is the forget gate and indicates whether to erase cell or not. While the output gate o indicates how much to reveal cell. The gate indicates how much to write to cell.

The value of the gates (i, f, and o) is between 0 and 1 (like logical gates) for example the forget gate if we want to forget the previous state $c_{t-1}$the value of f will be 0 than the value of c_t represented by equation

$$c_t = o * c_{t-1} + i * g \quad \text{(III.25)}$$

The value of i give how much the value of g will be in the next state c_t. To benefit of this architecture the first is when the back propagate from c_t to c_{t-1} only element wise multiplication by f no matrix multiplication; the second f is varying from one cell to another, therefore gradient vanishing (the gradient becomes small and small) through the network can be avoided.

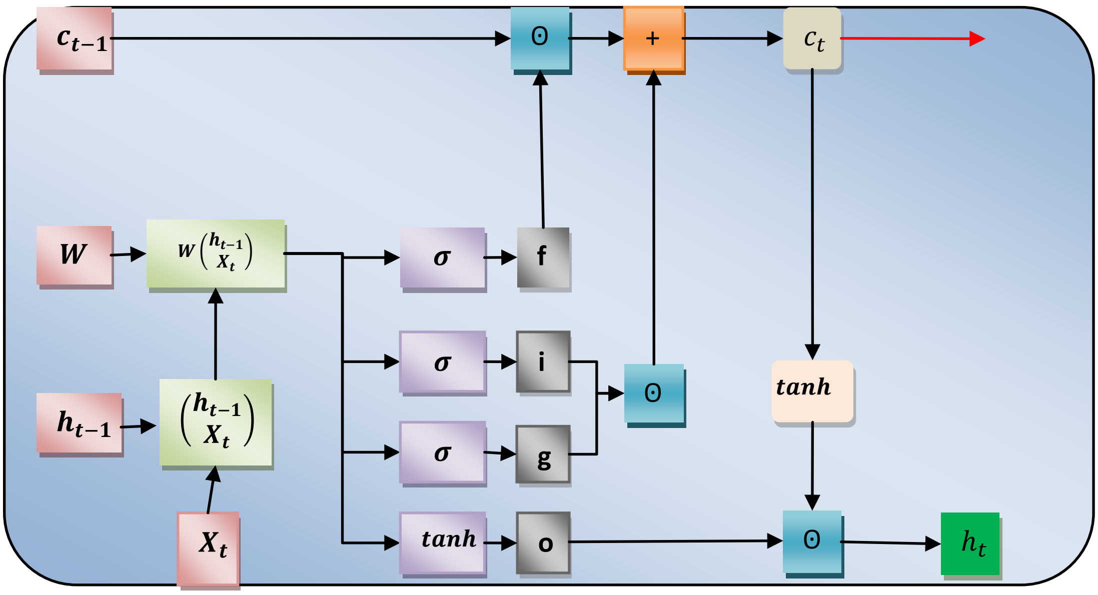

***Figure III. 7**: Deferent gates and activation functions of LSTM cell.*

III.2.2.2 Treating the sequences data using LSTM

In all Deep Learning treating there are three main steps preprocessing, fitting the network, and making prediction,[32].

A. Preprocessing

- **Difference function:** This function makes the data stationary to remove the increasing trend in the data this treating is very important when the data enters into LSTM.
- **Rescaling the values:** The activation function of LSTM is "tanh" so all the values mast be between 1 and -1, this is done using MinMaxScaler function which is part from Sklearn library.
- **The series to supervised () Function:** LSTM is a supervised learning; this means that there are inputs and their correspondents' labels; In time series forecasting the inputs is the sample at time (t) and the label is the next sample (t+1); the following table shows an example for two lagging and two sequences samples.

Table III.* 1: An Example of series to supervised problem*

Presence values	Future values	
T	t+1	t+2
0	10	25
10	25	Nan
25	Nan	Nan

B. Fitting the network (training)

- **Reshaping the data tensors:** The time series data are stored in 3D tensor [samples, time-step, and feature] as it is shown in the following figure

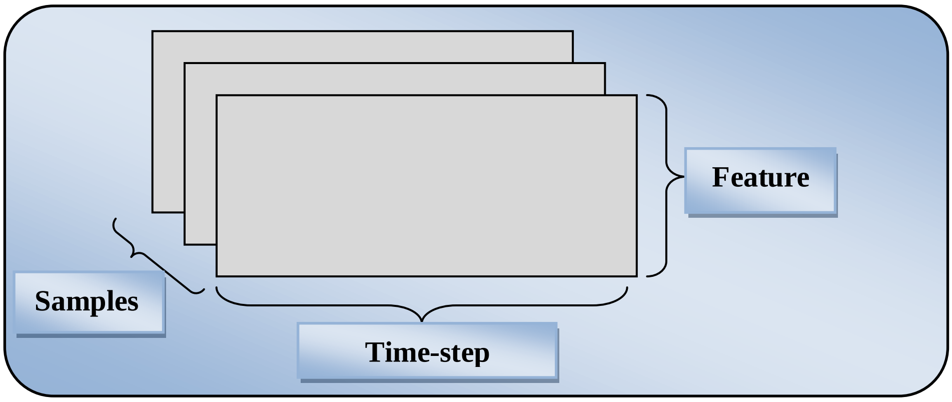

***Figure III.*8**: *Tensor of time series organization*

- **Design the Network**

The Sequential model is linear stack of deep learning method layers, the reason for choosing Sequential model and not API (multiple input multiple output) is because the training occurs only in one channel EEG time series especially with CPU. API with CPU is very time consuming so if we use multiple input multiple output (18 or 100 channel) it will never get the results, therefore the time series of the channels enter manually one after one to the network. After the sequential mode LSTM layer is fitted with deferent parameter we keep the default configurations, the output layer is a Dens layer with one output, the forth layer in this structure is the compile model with mean square Error loss function because the purpose of the network is regression and Adam optimization algorithm [33].

- **Make prediction**

This model is practical model for testing, a segment from the testing data is token and the test is done in that segment after that the testing is done on the next segment and the previous segment is used in the training data, this process is repeated on all testing data, the following figure illustrate this model [33]. For a given 50559 sample 1/3 sample is token as a test dataset, for 500 sample/s 1000 sample is required for 2 second forecasting. That is given presence time observation t forecasts t+1, t+2 and t+1000, after this iteration observation t+1 are token from the test data and forecasts t+2, t+3 and t+1001this process continues until the last sample in test data set. The forecasting samples from 1000 until the end are considered as the 2s forecasted signal; RMSE is calculated at each 1000^{th} forecasted samples the average of them is recorded.

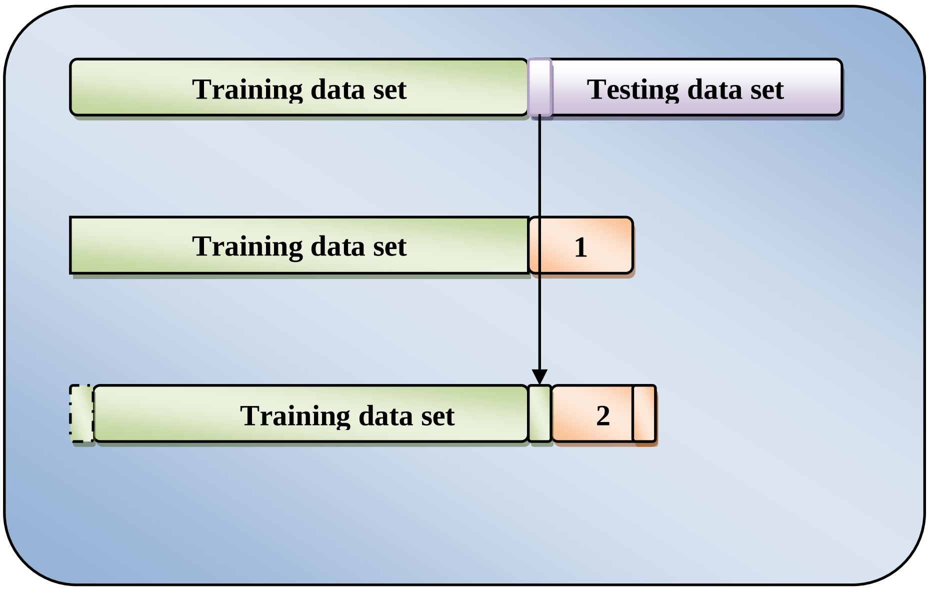

***Figure III.* 9**: *Forward validation model*

The used hardware for classification and forecasting is an INTEL i3 core with 4GHz Random Access Memory (RAM), this took a long time in processing; so, it is preferred to use the more faster GPU.

III.2.2.3 Forecasting using LSTM

When computers appeared to be super-fast in human calculations, at the same time, people were more capable of understanding and understanding images, speech and art. But in the past few years, with the advent of highly sophisticated learning algorithms such as artificial neural networks, computers became capable of comprehension and understanding Such as human perfectly and at the same time quick in the calculations and enable them to deal with signals that cannot be dealt with by humans exceeded. LSTM is a special artificial neural networks this means that all the training steps of artificial neural networks are applied to LSTM. it is power full method for treating sequential data and natural language processing [25]. The interested application of LSTM is its ability to predict a future values based on past and presence values for example it is applied to make prediction in weather [26], hose Electrical power consumption [27], in this work we show how LSTM can applied to treating EEG time series and making forecasting.

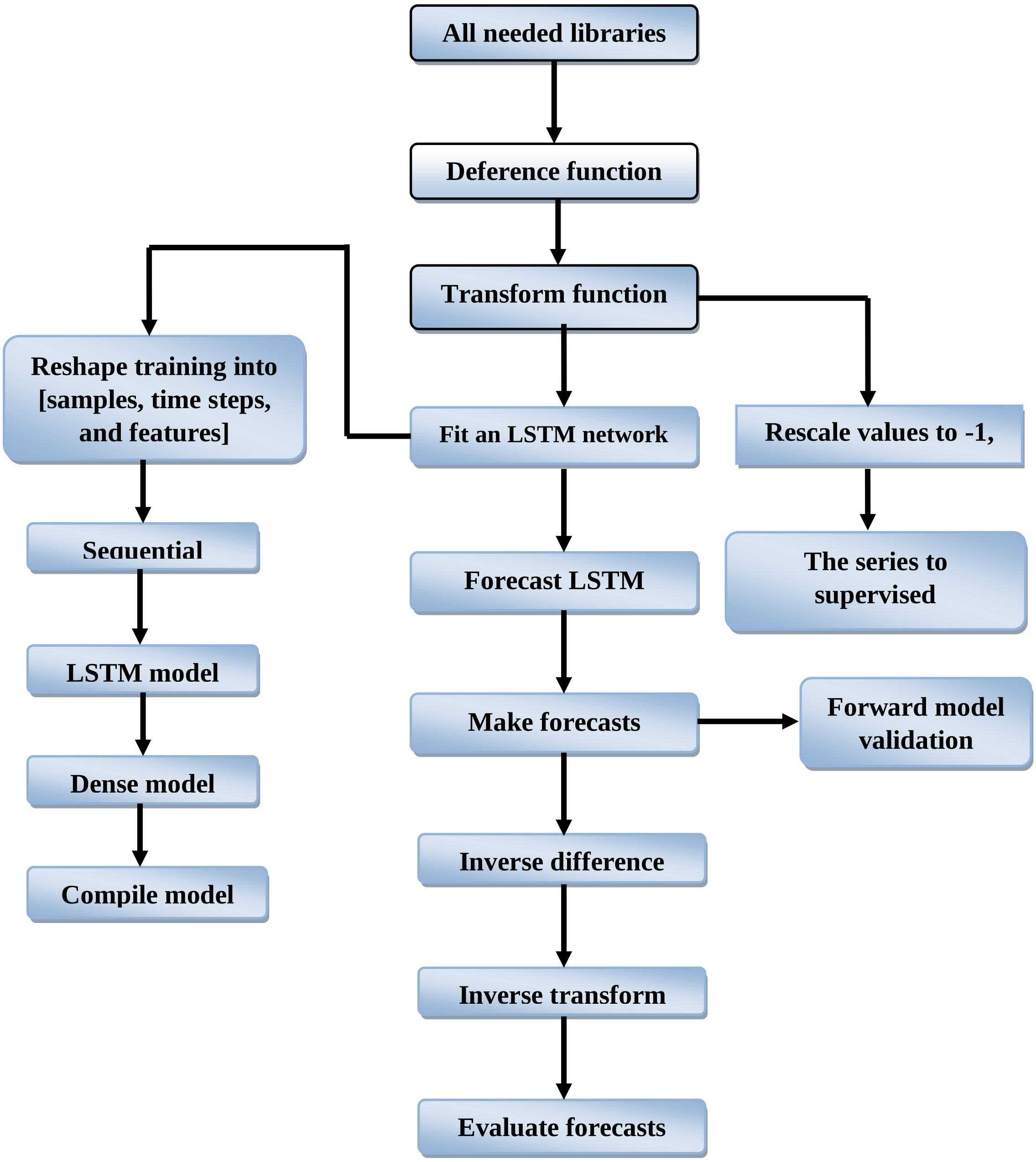

***Figure III.* 10**: *Treating the sequences data using LSTM*

III.3 Summary

This chapter presents a description about statistical methods used for forecasting followed by a brief illustration of different neural networks utilized in sequential (time series); the chapter concludes with treating the data sequences with LSTM network.

Chapter IV

Experimental part

The previous chapters provide a good background to use it in the experimental Work. In this chapter we demonstrate all the software applied to the data and the results gathered in statistical aspect for both classification and forecasting.

IV.1 Introduction

In this chapter, we present the results and discussions of the work for both classification and forecasting, features vector selected for the classification consists from: the mean, standard deviation, minimum, maximum, median, mode, first quartile, third quartile and inter quartile are used in this implementation. For forecasting, the performance of statistics methods is evaluated by calculating RMSE; the LSTM method is evaluated by changing the hyper-parameters (number of neurons and number of epochs) and looking for the best conditions that gives smaller RMSE.

IV.2 Experimental part: Classification

IV.2.1 The data set description

The EEG dataset includes recordings for both healthy and epileptic subjects. The dataset includes five subsets (denoted as A, B, C, D, and E) each containing 100 single channel EEG segments, each one having 23.6-second duration. The subsets A and B have been acquired using surface EEG recordings of five healthy volunteers with eyes closed and open, respectively. Signals in two sets have been measured in seizure-free intervals from five patients in the epileptogenic zone (set D) and from the hippocampal formation of the opposite hemisphere of the brain (set C). Finally, subset E contains seizure activity, selected from all recording sites exhibiting ictal activity.

We use the above-described dataset to create four different classification problems:

1. The first situation, all the EEG segments from the dataset were used and they were classified into three different classes: A and B types of EEG segments were combined to a single class, C and D types were also combined to a single class, and type E was the third class. This set is the one closest to real medical applications including three categories; normal (i.e., types A and B), seizure free (i.e., types C and D) and seizure (i.e., type E).
2. In the second situation, again all the EEG segments from the dataset were used and they were classified into two different classes: A, B, C, and D types are included in the first class and type E in the second class. This is also close to real medical applications, being slightly simpler than the previous, classifying the EEG segments into no seizures and seizures.
3. The third one has similar classes with the first, that is, normal, seizure-free and seizure, but not all the EEG segments from the dataset were employed.

The normal class includes only the A-type EEG segments, the seizure-free class the D-type EEG segments, and the seizure class the E-type.

4. The fourth situation has similar classes with the second, that is, normal and seizure, but again not all the EEG segments from the dataset were employed. The normal class includes only the A-type EEG segments while the seizure class includes the E-type [2].

***Table IV. 1**: Data separation according to classification problem.*

Classification problem	Classes	Number of segment
1	Normal(A-B) Seizure free C-D) Seizure(E)	200 200 100
Total		**500**
2	Non Seizure (A-B-C-D) Seizure (E)	400 100
Total		**500**
3	Normal(A) Seizure free(C) Seizure(E)	100 100 100
Total		**300**
4	Normal(A) Seizure(E)	100 100
Total		**300**

***Table IV. 2**: Detailed description of the data sets.*

	Set A	Set B	Set C	Set D	Set E
Subjects	Five healthy subjects	Five healthy subjects	Five epileptic subjects	Five epileptic subjects	Five epileptic subjects
Patient's state	eyes open	Closed eyes	Seizure free	Seizure free	Seizure activity
Electrode type	Surface	Surface	Intracranial	Intracranial	Intracranial
Electrode placement	International 10-20 system	International 10-20 system	Opposite to: Epileptogenic	Within Epileptogenic	Within Epileptogenic
Number of channel	**100**	**100**	**100**	**100**	**100**
Time duration (s)	**23.6**	**23.6**	**23.6**	**23.6**	**23.6**

IV.2.2 The optimum classifier

The purpose of the first part is to identify which rhythm is mostly affected by epilepsy and which classifier should be used. After answering those two questions, we use a band pass filter to extract only the rhythm in question in order to reduce time execution of WPD. This gives the optimum classifier:

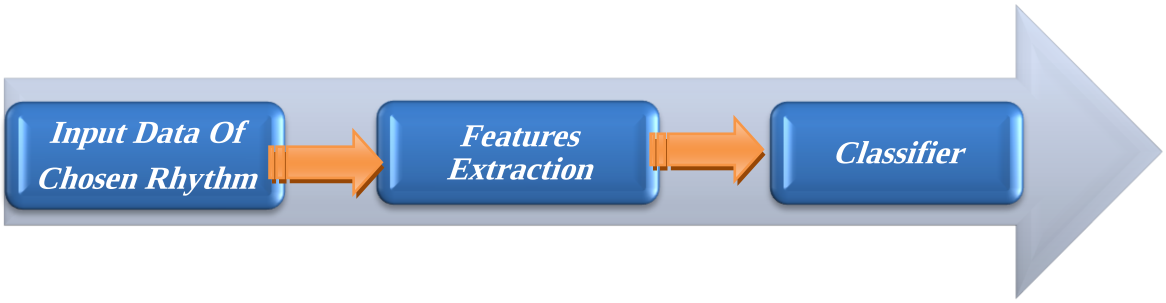

Figure IV. 1: *Optimum classifier.*

First, the EEG signals are decomposed to extract five rhythms; for each rhythm, the features vector selected for the classification is of dimension [1x9]; this gives us a matrix size [9x200] for each rhythm.

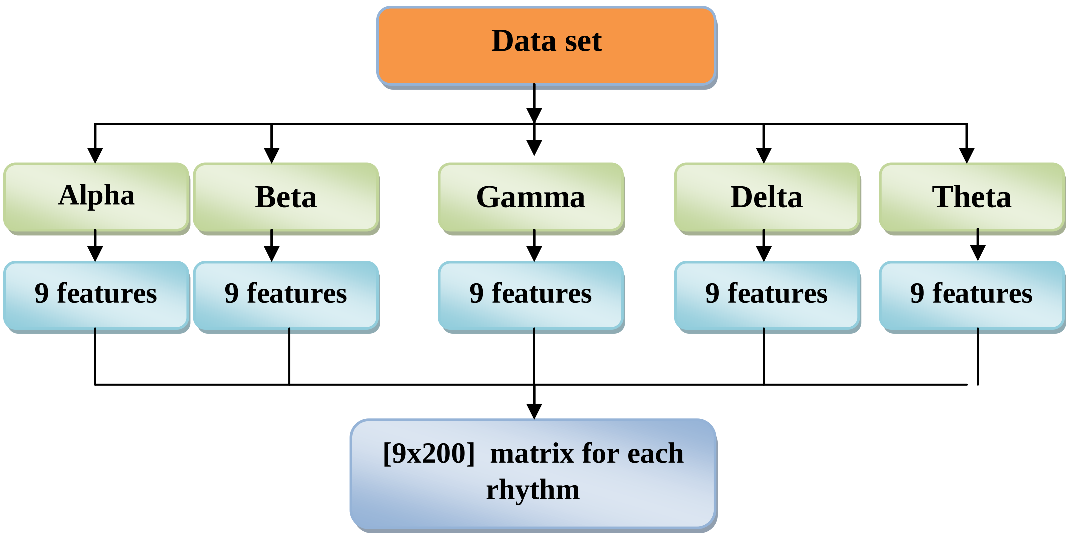

Figure IV. 2*: Rhythms and features extraction from the data set.*

The MATLAB program is decomposed into four parts:

- Filtering the data: the data chosen belongs to the fourth class is filtered using Butterworth filter after sampling. The Butterworth filter is of order 2 low-pass filter with cutoff frequency of 60 Hz with a sample rate of 160 samples per second.

```
for i=1:200
samplerate=160;
cut_off_freq=60;
first_order=2;
[b,a]=butter(first_order,cut_off_freq/(samplerate/2));
M=filtfilt(b,a,B(:,i));
```

Figure IV. 3*: Filtering the input data.*

- The wavelet decomposition algorithm is used to extract the five rhythms that formulate the EEG signals; to enable extracting all rhythms bands, the WPD breaks down the signal into four layers. db1 refers to Daubechies wavelet. After decomposition, features vector is conducted for each rhythm.

```
channel_dec=wpdec(M,4,'db1');
gamma_rhythm=wpcoef(channel_dec,2);
beta_rhythm=wpcoef(channel_dec,4);
alpha_rhythm=wpcoef(channel_dec,8);
delta_rhythm=wpcoef(channel_dec,15);
theta_rhythm=wpcoef(channel_dec,16);
```

Figure IV. 4*: Rhythms extraction with WPD.*

```
gamma_rhythm_ZS(i,:)=[mean(gamma_rhythm),std(gamma_rhythm),min(gamma_rhythm),max(gamma_rhythm),median(gamma_rhythm),mode(gam
beta_rhythm_ZS(i,:)=[mean(beta_rhythm),std(beta_rhythm),min(beta_rhythm),max(beta_rhythm),median(beta_rhythm),mode(beta_rhyt
alpha_rhythm_ZS(i,:)=[mean(alpha_rhythm),std(alpha_rhythm),min(alpha_rhythm),max(alpha_rhythm),median(alpha_rhythm),mode(alp
delta_rhythm_ZS(i,:)=[mean(delta_rhythm),std(delta_rhythm),min(alpha_rhythm),max(alpha_rhythm),median(delta_rhythm),mode(del
theta_rhythm_ZS(i,:)=[mean(theta_rhythm),std(theta_rhythm),min(theta_rhythm),max(theta_rhythm),median(theta_rhythm),mode(the
end
```

Figure IV. 5: *Features extraction for each rhythm.*

In order to use classification learner app in MATLAB, a table should be constructed; this table contains the values of the features for each rhythm alone added to that a description of whether the signal is for a healthy or epileptic person for all data set.

```
LastName = {'healthy';'healthy';'healthy';'healthy';'healthy';'healthy';'healthy';
T = cell2table(LastName);              %convert cell label to table.
T1= array2table(theta_rhythm_NS);      %convert data matrix to table.
T_theta_rhythm_NS=[T1 T];              %table used for classification.
```

Figure IV. 6*: Tabling output for classification.*

A 10 fold cross validation procedure is used, the cross validation is a modal validation technique that partition the data into training subsets and validation (testing) subsets to reduce variability; each time, nine subsets is used for training and one subsets used for testing. The performance of the system is evaluated by averaging the testing results. The classification app gives the accuracy of many classification algorithms. The accuracy is the number of correct decisions divided by the total number of cases; the accuracy is within ± 0.5% error. The results are shown in tables below:

Table IV. 3: *Results of training algorithms with alpha rhythm.*

Alpha Rhythm					
	E-A	**E-B**	**E-D**	**E-C**	**Mean**
Simple tree	99.5%	94.5%	97.5%	96.5%	97%
Quadratic Discriminant	99.5%	91.5%	95.5%	98.0%	96.13%
Logistic regression	99.5%	95.5%	96.0%	97.5%	97.13%
Fine Gaussian SVM	99.5%	97.0%	96.0%	98.0%	97.63%
Fine KNN	99.5%	96.5%	94.5%	99.0%	97.38%
Ensemble bagged trees	99.5%	96.0%	96.5%	98.0%	97.5%

Table IV. 4: *Results of training algorithms with beta rhythm.*

Beta Rhythm					
	E-A	**E-B**	**E-D**	**E-C**	**Mean**
Simple tree	98.5%	95.5%	97.5%	98.0%	97.38%
Quadratic Discriminant	99.0%	95.5%	94.5 %	97.5%	96.63%
Logistic regression	98.5%	95.5%	98.0%	98.0%	97.5%
Fine Gaussian SVM	99.5%	96.5%	94.5%	97.5%	97.0%
Fine KNN	99.5%	95.0%	92.5%	96.5%	95.88%
Ensemble bagged trees	98.5%	96.5%	95.5%	98.0%	97.13%

__Table IV. 5:__ Results of training algorithms with delta rhythm.

Delta rhythm					
	E-A	**E-B**	**E-D**	**E-C**	**Mean**
Simple tree	99.0%	98.0%	89.5%	95.0%	95.38%
Quadratic Discriminant	100.0%	99.0%	84.5%	93.5%	94.25%
Logistic regression	99.0%	99.0%	92.5%	96.5%	96.75%
Fine Gaussian SVM	99.5%	99.0%	89.5%	94.5%	95.63%
Fine KNN	99.5%	99.0%	90.0%	93.0%	95.38%
Ensemble bagged trees	99.5%	99.0%	92.0%	95.0%	96.38%

__Table IV. 6:__ Results of training algorithms with gamma rhythm.

Gamma rhythm					
	E-A	**E-B**	**E-D**	**E-C**	Mean
Simple tree	97.0%	96.0%	97.0%	98.0%	97.0%
Quadratic Discriminant	99.0%	95.5%	92.5%	98.0%	96.25%
Logistic regression	99.5%	96.0%	98.5%	95.0%	97.25%
Fine Gaussian SVM	99.5%	96.0%	95.0%	98.0%	97.13%
Fine KNN	99.5%	95.5%	91.5%	96.5%	95.75%
Ensemble bagged trees	98.5%	96.5%	96.0%	97.5%	97.13%

Table IV. 7: Results of training algorithms with theta rhythm.

Theta Rhythm					
	E-A	**E-B**	**E-D**	**E-C**	Mean
Simple tree	99.5%	93.0%	96.5%	98.0%	96.75%
Quadratic Discriminant	99.5%	91.0%	93.0%	98.5%	95.5%
Logistic regression	99.0%	95.0%	97.0%	96.5%	96.88%
Fine Gaussian SVM	99.5%	93.5%	95.5%	98.5%	96.75%
Fine KNN	99.5%	92.0%	92.5%	98.5%	95.63%
Ensemble bagged trees	99.5%	94.0%	97.0%	98.0%	97.13%

By summing up the means of each algorithm, we can deduce the rhythm that is mostly affected by epilepsy.

Table IV. 8: Mean of accuracy of methods with each rhythm.

Alpha rhythm	Beta rhythm	Delta rhythm	Gama rhythm	Theta rhythm
97.13%	96.92%	95.63%	96.75%	96.44%

- We see for the above table that the classification using alpha rhythm gives the best results; we see also that the algorithm with the best accuracy is fine Gaussian SVM for the same alpha rhythm.

IV.2.3 Testing the optimum classifier

In this section, we will use the previous results to make a classification model based on obtained results, one purpose from that is to see the time execution difference when using WPD and a band pass filter; another purpose is to reduce the number of features known as dimensionality reduction.

- **Time execution**

The running time for the optimum classifier with WPD is 7.735 s where WPD took 6.018 s to execute

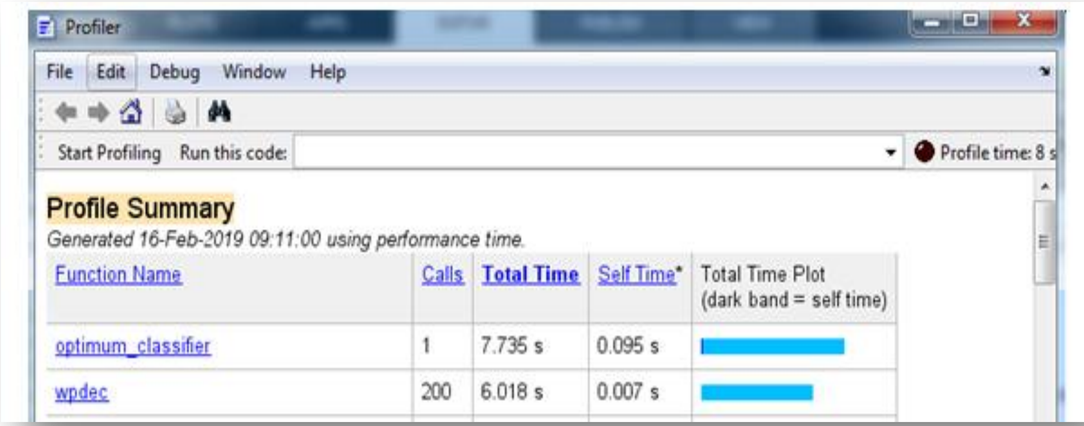

Figure IV. 7: *Time execution of optimum classifier and WPD method.*

Because WPD takes a long time we don't need to calculate all rhythms, we can take just one rhythm using band-pass filter to reduce the Running time. As in the previous part, the data set is filtered but instead of using WPD we use a band pass filter to extract just the range of alpha rhythm, this will decrease the time execution. An SVM model is trained by the data proposed, the training data includes sets A and E from 1 to 90 so that the last 10 channels of each set can be used for testing.

```
for i=1:180

%%filtering .
samplerate= 160;
first_order= 2;
f_low= 8;
f_high= 15;
[b,a]= butter(first_order, [f_low,f_high]/(samplerate/2));
alpha_rhythm= filtfilt(b,a,X(:,i));
alpha_rhythm_2(i,:)= [mean(alpha_rhythm),min(alpha_rhythm),max(alpha_rhythm)];
end
```

Figure IV. 8: *Extracting alpha band from the signals*

- The running time for the optimal classifier is 1.224 s where the filter took only 0.604 s; this means that using the filter reduces time execution compared to WPD.

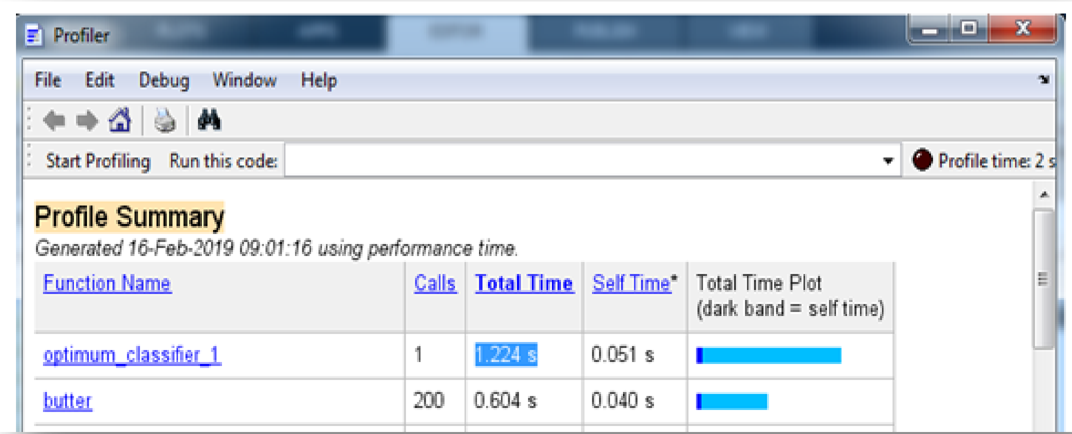

Figure IV. 9: *Time execution of optimum classifier and Butterworth filter.*

- **Dimensionality reduction**

The last point in classification is to train the SVM model and test it to see the accuracy after using a single rhythm with only three features which are the mean, the minimum and maximum; the reason to choose these features is that they are the best to describe the peaks the epileptic EEG has as compared to normal EEG as seen in figure I.3 and figure I.4 of chapter I.

```
%%training the model.
SVMModel1=fitcsvm(alpha_rhythm_2, Y, 'KernelFunction','rbf', ...
     'Standardize', true, 'ClassNames', {'healthy', 'epileptic'});
```

Figure IV. 10*: Training SVM model with alpha rhythm.*

```
%%testing the model.
alpha_rhythm_S099= filtfilt(b,a,transpose(S099));
alpha_rhythm_3=[mean(alpha_rhythm_S099),min(alpha_rhythm_S099),max(alpha_rhythm_S099)];
[label,accuracy]=predict(SVMModel1,alpha_rhythm_3);
label
```

Figure III. 11*: Testing SVM model with alpha rhythm for S099*

```
%%testing the model.
alpha_rhythm_Z099= filtfilt(b,a,transpose(Z099));
alpha_rhythm_3=[mean(alpha_rhythm_Z099),min(alpha_rhythm_Z099),max(alpha_rhythm_Z099)];
[label,accuracy]=predict(SVMModel1,alpha_rhythm_3);
label
```

Figure III. 12: *Testing SVM model with alpha rhythm for Z099..*

- The results are represented by labels, all the tested signals gave the correct label, for the channel S099 and Z099 presented in the programs above, the outputs are:

```
>> test2
>> label
label =
    'epileptic'
>>
```

```
>> test2
>> label
label =
    'healthy'
>>
```

Figure IV. 11: *Output for S099.* ***Figure IV. 14:*** *Output for Z099.*

IV.3 Experimental part: Forecasting

The dataset of this experiment is composed from two parts training and testing parts each part begins by normal patient channel (A dataset) and is completed by abnormal patient channel (E dataset) to simulate a real case. A segment of 1000 sample is used as test data the purpose is to get the loss during 10 second. The walk validation method that was described in the chapter III is used and so a piece of 640 samples (4s) is walking through 1000 sample, firstly 640 samples are generated in the second step a sample from test data is token as one sample from the training data and predict the next 640 samples until all 1000 samples complete as is explained in Chapter III.2.2.4.

IV.3.1: Statistic Methods

Python code is used for evaluating the RMSE, the codes are available in [24] in all methods the best parameter thus the best results are recorded in the below table.

Table IV. 9: *The RMSE values that is resulted from the statistic methods.*

Methods	Naive	Average	Moving Average	Holt's Linear Tred	Holt-Winters
RMSE	364.240	364.624	364.244	363.589	325.532

IV.3.2 LSTM Hyper-parameter

Each experimental scenario will be run 5 times. The reason for this is that the random initial conditions for an LSTM network can result in very different results each time a given configuration is trained. The model will be evaluated on the test datasets at the end of each experiment and the RMSE scores saved.

- **Tuning the Number of Neurons**

The first LSTM parameter we will look at tuning is the number of Neurons that is used in LSTM, there is no analytical way between the data long and the number of neurons for this reason a number from 1 to 5 neurons (even 10,100 neurons are used but they shows very bad results) are used with deferent time forecasting.

Table IV. 10: The effective of the RMSE with the Number of Neurons

	1 neuron	2 neurons	3 neurons	4 neurons	5 neurons
Forecasting after 1s	72.800990	73.373011	73.526215	73.319204	73.661059
Forecasting after 2s	70.026847	70.731082	71.128407	70.884626	71.524755
Forecasting after 3s	70.171021	70.593030	70.893760	70.450140	71.137333
Forecasting after 4s	94.531593	94.734772	95.259619	94.568513	95.322321

- **Tuning the Number of Epochs**

Epoch is an arbitrary cutoff, generally defined as "one pass over the entire dataset", used to separate training into distinct phases, which is useful for logging and periodic evaluation.

Table IV. 11: The effective of the RMSE with the Number of Epoch.

	1 epoch	2 epochs	3 epochs	4 epochs	5 epochs
Forecasting after 1s	72.800990	73.373011	73.526215	73.319204	73.661059
Forecasting after 2s	70.026847	70.731082	71.128407	70.884626	71.524755
Forecasting after 3s	70.171021	70.593030	70.893760	70.450140	71.137333
Forecasting after 4s	94.531593	94.734772	95.259619	94.568513	95.322321

- The statistic methods are very bad to make prediction in epilepsy EEG signal.
- The best hyper parameters of LSTM for epilepsy EEG signal prediction are:
  - The Number of Neurons = 1.
  - The Number of Epochs = 1.

- **Testing in a real data**

This data taken from one person provided by a neurologist is 19 channels; the 19th channel ECG is deleted. The sample rate is 500sample/s. The prediction is made channel after channel (each channel enters to the treating manually).The LSTM model with above Hyper-parameters is used; RMSE is measured and recorded in the below table.

Table IV. 12: *RMSE results for disease patient of 18 channels*

	1s	**1.5 s**	**2 s**
Channel 1	14.632882	24.230715	29.208293
Channel 2	2.392090	27.682805	5.138175
Channel 3	23.924737	36.338896	47.918400
Channel 4	12.004308	20.004786	23.985504
Channel 5	33.692950	51.840071	67.310976
Channel 6	19.104740	32.682895	38.428750
Channel 7	28.561055	43.943205	57.056488
Channel 8	11.888141	19.092864	23.805204
Channel 9	8.252565	15.715273	16.541083
Channel 10	22.541736	38.435212	45.160559
Channel 11	17.226283	26.474379	34.662760
Channel 12	12.318634	19.193771	24.686648
Channel 13	45.027148	70.392196	90.387190
Channel 14	18.652930	29.334848	37.156237
Channel 15	18.387036	30.528500	36.667125
Channel 16	18.277388	29.201231	36.460930
Channel 17	60.376621	91.949636	120.736490
Channel 18	10.552626	17.681794	21.185331

- From the previous LSTM with resulted hyper-parameter works very well with a real time epilepsy EEG signal.

  For real time operation:

- It is preferred to utilize API with multiple input multiple output architecture.

- For fast prediction it is suggested to use nvidia *JETSON NANO* device; it is a small and fast (CPU +GPU) kit it is made specifically for AI applications like image classification, object detection, segmentation and signal processing. It enables running multiple neural networks in parallel.

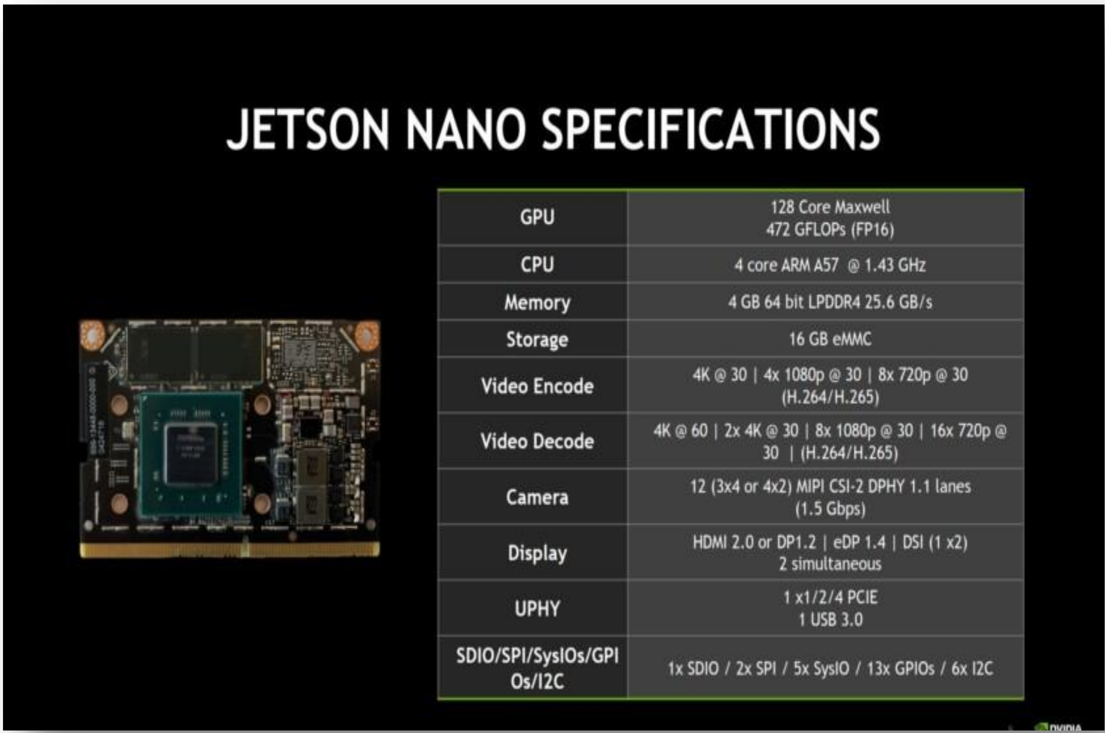

Figure IV. 15: *JETSON Nano Kit Specifications.*

IV.4 Summary

This chapter gathers the statistical study and results of classification and forecasting; from the study, we found that alpha rhythm is the mostly affected rhythm by epilepsy and we utilized this result to propose an SVM classifier characterized with less time execution and a three features (mean, minimum and maximum) vector size compared to nine features. For forecasting, because statistics methods cannot detect the peaks, we found that LSTM is the best solution to replace them and we concluded that the best hyperparameters for LSTM are 1 neuron and 1 epoch.

Conclusion

Brain activity is the response of the brain and neuronal system to exterior excitation, Electroencephalography (EEG) signal is a recording of this activity in certain duration of time, the EEG or brain activity signal is useful tool for neurologists to see the effect of brain disorders on its five main rhythms (alpha, beta, gamma, delta and theta).

Our contributions in this project, we deduced that alpha is mostly affected by epilepsy disorder and that the best classification algorithm that can help neurologists diagnose this disorder is Support Vector machine (SVM) among many others since it gives more accuracy, after finding the rhythm, a band pass filter is used to extract only its band to use it in training the model instead of using the Wavelet Packet Decomposition (WPD), this reduced the time execution since it extracts only the needed rhythm. The number of features also reduced to three (mean, minimum, maximum) since they describe well the differences between healthy and seizure EEG signals, for forecasting process, LSTM hyper-parameter works well in the prediction of seizures since it gives the smallest RMSE value. One disadvantage of this method is that it takes a long time to return the results. For real time applications, we recommend to use API (multi input/output) architecture and JETSON board to forecast.

Based on the results obtained in this work for both classification and forecasting, there is a possibility to develop a device that can predict and alert to the epilepsy patient in a real time situation where the patient can have a sudden seizure; if the patient was in a wrong place or situation, this will make him in severe danger.

After massive development in machine learning and deep learning algorithms, the range can be extended to cover other kind of diseases like ones concerning the heart (sudden heart attack, high and low blood pressure…), predicting the health situation can save lot of lives.

Appendix A: Wavelet packet decomposition

Wavelet packet decomposition (WPD) (sometimes known as just wavelet packets) is a wavelet transform where the signal is passed through more filters than the DWT. Wavelet packets are the particular linear combination of wavelets. They form bases which retain many of the orthogonality, smoothness, and localization properties of their parent wavelets. The coefficients in the linear combinations are computed by a recursive algorithm making each newly computed wavelet packet coefficient sequence the root of its own analysis tree. In the DWT, each level is calculated by passing the previous approximation coefficients though a high and low pass filters. However, in the WPD, both the detail and approximation coefficients are decomposed [34].

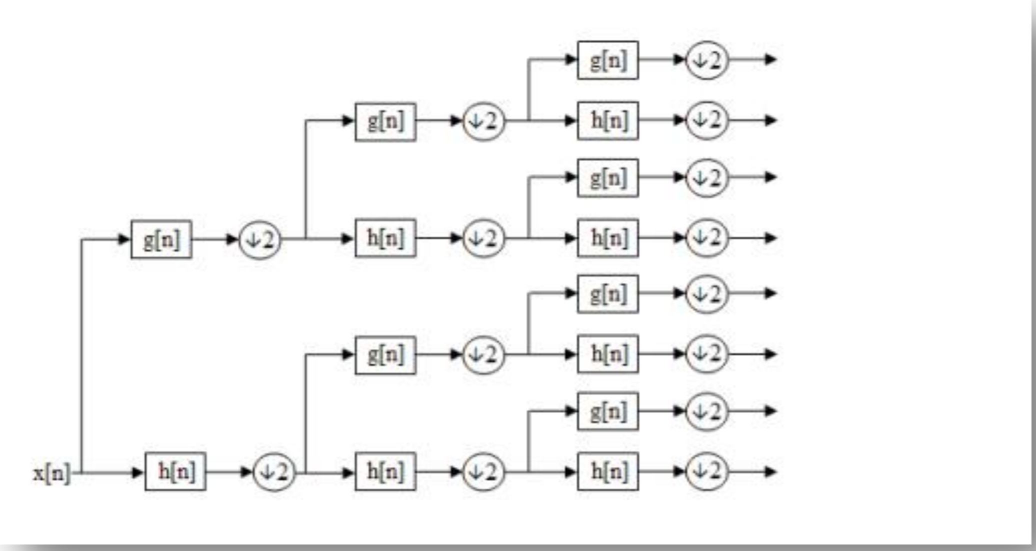

Figure A. 1: Wavelet Packet Decomposition Tree.

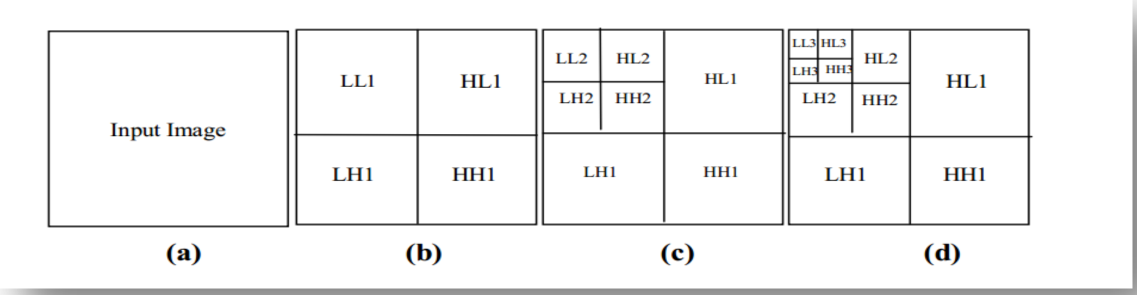

Figure A. 2: Two dimensional discrete wavelet transform

References

[1] [Online]. Available: https://directedwellnesscenter.com/additional-information/.

[2] Y. L. Z. Siuly Siuly, EEG Signal Analysis and Classification Techniques and Applications, Springer International Publishing, 2016.

[3] [Online]. Available: https://www.google.com/amp/s/amp.livescience.com/39074-what-is-an-mri.html.

[4] [Online]. Available: https://www.google.com/search?q=mri+scan&tbm=isch&tbs=rimg:CfVUlkxwhqpPIjh5v3BFyPBLo44GMPJwBmfSScf9dd1zaddV270DxL2zme7r_1dwSU8vysT1Rrh1VWarGVO1pqH5MvyoSCXm_1cEXI8EujEXVjohnlZgsiKhIJjgYw8nAGZ9IRCKwQn-6MsqEqEglJx_1113XNp1xFVjslyMo_1fhCoSCVXbvQPEvbOZEaLTgQob.

[5] [Online]. Available: https://en.m.wikipedia.org/wiki/Position_emission_tomography..

[6] [Online]. Available: https://en.m.wikipedia.org/wiki/CT_scan.

[7] [Online]. Available: https://www.google.com/search?biw=1366&bih=608&tbm=isch&sa=1&ei=zFzhXPayIeLQxgOK5Le4DA&q=output+of+pet+scanner&oq=output+of+pet+scanner&gs_l=img.3...27764.36447..37228...0.0..1.0.0.......8....1..gws-wiz-img.F-JWnL1AcMM#imgrc=Fp1uzp2DHMf4FM:.

[8] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163395/.

[9] [Online]. Available: https://en.wikipedia.org/wiki/Variance.

[10] [Online]. Available: https://en.wikipedia.org/wiki/Standard_deviation.

[11] [Online]. Available: https://www.investopedia.com/terms/m/median.asp.

[12] [Online]. Available: https://www.investopedia.com/terms/m/mode.asp.

[13] [Online]. Available: https://www.quora.com/What-is-the-difference-between-regression-classification-and-clustering-in-machine-learning.

[14] [Online]. Available: https://towardsdatascience.com/https-medium-com-popularushikesh-svm-f4b42800e989..

[15] [Online]. Available: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[16] [Online]. Available: https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248..

[17] [Online]. Available: https://medium.com/@srishtisawla/linear-discriminant-analysis-d38decf48105..

[18] [Online]. Available: dataaspirant.com/2017/02/06/naïve-bayes-classifier-machine-learning/.

[19] [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.

[20] [Online]. Available: https://www.quora.com/How-does-logistic-regression-work-in-machine-learning.

[21] [Online]. Available: https://blog.statsbot.co/ensemble-learning-d1dcd548e936?gi=94d53ff2eae8..

[22] C. García-Martos, "Price Forecasting Techniques in Power Systems," January 2013. [Online]. Available: https://www.researchgate.net/publication/319582458_Price_Forecasting_Techniques_in_Power_Systems.

[23] D. Wood, "Methods & Principles of Weather Forecasting," [Online]. Available: https://study.com/academy/lesson/methods-principles-of-weather-forecasting.html.

[24] G. SINGH, "7 methods to perform Time Series forecasting," 8 FEBRUARY 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/.

[25] v. nigam, "Natural Language Processing: From Basics to using RNN and LSTM," [Online]. Available: https://towardsdatascience.com/natural-language-processing-from-basics-to-using-rnn-and-lstm-ef6779e4ae66.

[26] Z. Karevan, "Spatio-temporal Stacked LSTM for Temperature Prediction in Weather Forecasting," 17 Zahra Karevan 2018. [Online]. Available: https://openreview.net/forum?id=SJgYmxo4jX.

[27] J. Brownlee, "How to Develop LSTM Models for Multi-Step Time Series Forecasting of Household Power Consumption," 10 October 2018. [Online]. Available: https://machinelearningmastery.com/how-to-develop-lstm-models-for-multi-step-time-series-forecasting-of-household-power-consumption/.

[28] A. Nariman, "Face Recognition using Convolution Neural Network," University M'Hamed BOUGRA, Boumerdes, 2017.

[29] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[30] J. Brownlee, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning," 3 July 2017. [Online]. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/.

[31] H. T. Siegelmann and Sontag, "Wikipedia," 1992. [Online]. Available: https://en.wikipedia.org/wiki/Long_short-term_memory.

[32] J. Brownlee, "Multi-step Time Series Forecasting with Long Short-Term Memory Networks in Python," 10 May 2017. [Online]. Available: https://machinelearningmastery.com/multi-step-time-series-forecasting-long-short-term-memory-networks-python/.

[33] F. a. o. Chollet, "Keras Document," 2015. [Online]. Available: https://keras.io.

[34] Vimal Krishnan V.R, Babu Anto P, Features of Wavelet Packet Decomposition and Discrete Wavelet Transform for Malayalam Speech Recognition, International Journal of Recent Trends in Engineering. Vol 1, No 2, May 2009.