EPILEPSY SEIZURE DETECTION USING EEG SIGNAL

A PROJECT REPORT

Submitted by

ASWATHY R S (TKM19MCA007)

to

The APJ Abdul Kalam Technological University

In partial fulfillment of the requirements for the award of the Degree of

MASTER OF COMPUTER APPLICATIONS

Thangal Kunju Musaliar College of Engineering
Kerala

DEPARTMENT OF COMPUTER APPLICATIONS

MAY 2022

DECLARATION

I undersigned hereby declare that the project report ”EPILEPTIC SEIZURE DETECTION USING EEG SIGNAL”, submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Prof. Jasmin M R. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Kollam

Date:14-05-22

Aswathy R S

DEPARTMENT OF COMPUTER APPLICATIONS

TKM COLLEGE OF ENGINEERING

C E R T I F I C A T E

This is to certify that, the report entitled ***"EPILEPTIC SEIZURE DETECTION USING EEG SIGNAL"*** submitted by **ASWATHY R S**, to the **APJ Abdul Kalam Technological University** in partial fulfillment of the requirements for the award of the Degree of **Master of Computer Applications** is a bonafide record of the project work carried out by her under our guidance and supervision.This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor Head of the Department External Examiner

ACKNOWLEDGEMENT

First and foremost I thank GOD almighty and my parents for the success of this project. I owe sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of my project.

I am extremely grateful to **Dr.Fousia M Shamsudeen,** Head of the Department, for providing us with best facilities.

I would like to thank my project guide **Prof.Jasmin M R,** Department of Computer Applications, who motivated me throughout the project.

I profusely thank all other faculty members in the department and all other members of TKM College of Engineering, for their guidance and inspirations throughout the course of study.

I owe my thanks to our friends and all others who have directly or indirectly helped us in the successful completion of this project.

Aswathy R S

ABSTRACT

Epilepsy is one of the most common neurological diseases that affects millions of persons all over the world. The disease has always been of great importance in the biomedical field, due to the health risks it causes. It is characterized by recurrent, unprovoked seizures and can be assessed by the electroencephalogram (EEG). Electroencephalogram (EEG) is one of the most powerful tools that offer valuable information related to different abnormalities in the human brain. One of these abnormalities is the epileptic seizure.EEG measures the electrical activity in the brain, and one important aspect of the epilepsy research includes analyzing the EEG data in order to detect epileptic seizures in early stages. A framework is proposed for detecting epileptic seizures from EEG signals recorded from normal and epileptic patients. The suggested approach is designed to classify the abnormal signal from the normal one automatically. This work aims to improve the accuracy of epileptic seizure detection and reduce computational costs.To address this, the proposed framework uses five machine learning (ML) classifiers:Logistic Regression,Decision Trees,Random Forest Gradient Boosting classifier,Extremely Random Trees.However, the Extremely Random Trees classifier achieved the best accuracy and it outperformed the other examined classifiers.

Contents

List of Figures

List Of Abbreviations

Chapter 1

INTRODUCTION

The word epilepsy originates from the Latin and Greek word ‘epilepsia’ which means ‘seizure’ or ‘to seize upon’. It is a serious neurological disorder with unique characteristics, tending of recurrent seizures . The context of epilepsy, found in the Babylonian text on medicine, was written over 3000 years ago . This disease is not limited to human beings, but extends to cover all species of mammals such as dogs, cats and rats. However, the word epilepsy does not give any types of clues about the cause or severity of the seizures; it is unremarkable and uniformly distributed around the world.

Several theories about the cause are already available. The main cause is electrical activity disturbance inside a brain , which could be originated by several reasons such as malformations, shortage of oxygen during childbirth, and low sugar level in blood . Globally, epilepsy affects approximately 50 million people, with 100 million being affected at least once in their lifetime . Overall, it accounts for 1 percentage of the world’s burden of diseases, and the prevalence rate is reported at 0.5–1 percentage. The main symptom of epilepsy is to experience more than one seizure by a patient. It causes a sudden breakdown or unusual activity in the brain that impulses an involuntary alteration in a patient’s behaviour, sensation, and loss of momentary consciousness. Typically, seizures last from seconds to a few minute(s), and can happen at any time without any aura. This leads to serious injuries including fractures, burns, and sometimes death.

Recently, a lot of research work have been carried out to detect the epileptic and non-epileptic signals as a classification problem . From the traditional diagnosis point of view, recognition of epileptic and non-epileptic EEG signals is a challenging task. Usually, there is a small amount of epilepsy data available for training a classifier due to infrequently happening of seizures. Further, the presence of noise and artifacts in the data creates difficulty in learning the brain patterns. The existing automatic seizure detection techniques use traditional signal processing (SP) . Many of these techniques show good accuracy for one problem but fail in performance due to less availability of labeled data.Thus there is a need of generalized automatic system that can show good performance even with fewer training samples.

Epilepsy can have an effect on long term memory causing the individual to eventually forget things like their personal information and people around them.There are many kinds of seizures, each with different symptoms, such as losing consciousness, jerking movements, or confusion. Some seizures are much harder to detect visually; the patients will usually exhibit symptoms such as not responding or staring blankly for a brief period. Seizures can happen unexpectedly and can result in injuries such as falling, biting of the tongue, or losing control of one's urine or stool. Hence, these are some of the reasons why seizure detection is of utmost importance for patients under medical supervision that are suspected to be seizure prone. Epilepsy can be identified by analyzing the patterns of Electroencephalogram (EEG) signals, which is a popular technique that is used to determine the abnormality of the brain. Hence, EEG signals are widely used by medical doctors and researchers to study epilepsy .

The figure 1.1 below shows the normal and abnormal EEG signal reading.

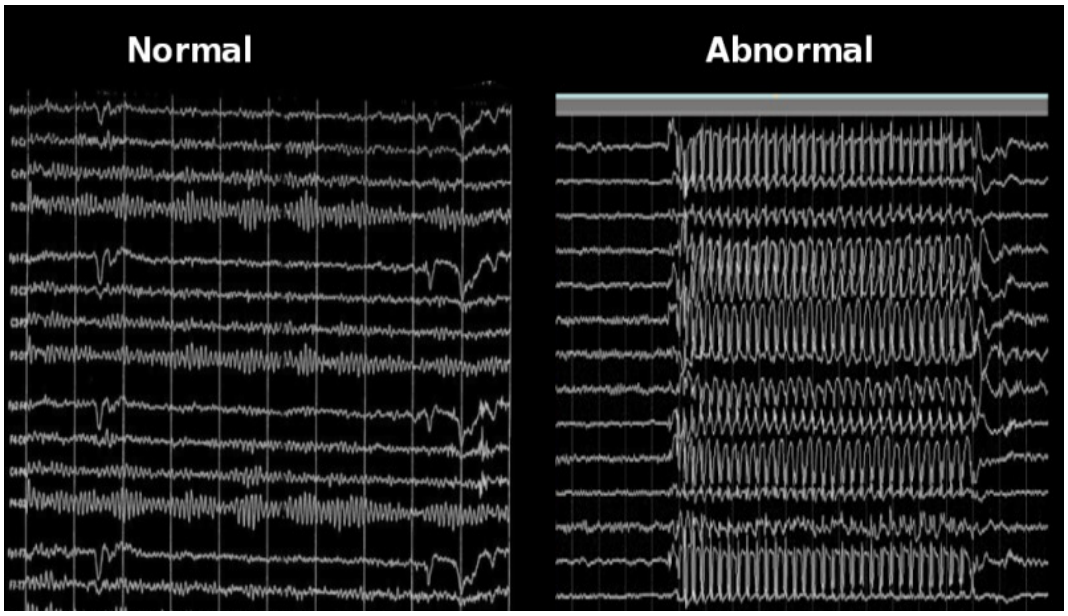

Figure 1.1: Normal and abnormal EEG signal

1.1 Objective

- The main objective of the proposed model is identify the arrival of epileptic seizure in a fully automated way.
- Make easy to diagnose whether corresponding EEG signals indicate the presence of seizure or not without human intervention.
- To classify the abnormal signal from the normal one automatically.
- Minimizing time consumption and human error.
- Save money and resources regarding the early detection and prediction of epilepsy.
- Develop a new approach of seizure classification model in a high performance environment.

Chapter 2

RELATED WORK

Many researchers have paid attention to EEG signals classification for epilepsy detection. This section describes a review on a set of recent related works to epileptic seizure detection from EEG signals.

D. Chen et al. [11] proposed a new approach based on the 54- DWT mother wavelets divided into seven families to divide the EEG data into different sub-bands to extract the statistical features. Then, an SVM classifier is used to categorize the EEG signals based on the extracted features. The experimental results display that the accuracy is mainly sensitive to the level of decomposition, and 40 percentage of the redundancies were removed from the resulting features.The main demerit of this model is that,it is less accurate.

A.Sharmila et al. [12] primarily rely on an analysis of EEG signals by making use of discrete wavelet transforms (DWT) to decompose the EEG data into different sub-bands, and then extract the statistical features. The derived statistical features from DWT are used to train the classifier. After that, two classifiers are used to determine the signals if they have epileptic or not. The two classifiers are KNN and Naive Bayes classifiers [18]. This research measures the performance of the 14 numerous combinations of two class epilepsy detection. The experimental consequences defined that, for the detection of epileptic seizure abnormality, the NB classifier achieves higher accuracy for most combinations of the dataset with less computation time, and the other classifier attains better accuracy for just 4 data sets combination.This model is not suitable for huge datasets.

S.Madan et al. [13] presented in their research an outline of the definition of epileptic seizure prognosis with the aid of way of making use of Hurst Exponent (HE) that primarily based on discrete wavelet for features functions extraction from EEG records. These features are gained through the ictal and pre-ictal stages of affected patients. The categorizing process of EEG indicators was applied using SVM and KNN Classifiers. In their research, the HE is defined to differentiate the EEG signals in terms of the more potent relative consistency, and less dependence on data length. The main consequences that appeared from this research are; the DWT-primarily based non-linear features coupled with SVM have given vital effects.It is very time consuming process.

D.Selvathi et al.[14], used a wavelet transform and Support Vector Machine (SVM) classifier to identify seizure to identify the rate of seizure in a patient from the EEG signals. They also aim to avoid aggressive situations during a patient seizure. They use seven levels of decomposition to obtain sub bands. Narrow sub bands used to detect the seizure and the other two sub bands used for extracting the statistical features and then for the classification of EEG signal using an SVM classifier. A normal EEG data set and a seizure data set during a seizure period have been used.The proposed model was less accurate and time consuming.

B.Harender et al. [15] used recorded EEG signals for a healthy and epileptic patient to develop a new framework used for the detection of an epileptic seizure. The simulation tool used Simulink. The statistical feature extracted for epilepsy detection with k-Nearest Neighbor (k-NN) classifier is: Mean Absolute Value (MA), Standard Deviation (SD), and Average Power (AP). Highly complex and time consuming model with low accuracy.

A novel automated detection system was developed by S.Lahmiri et al.[16] to distinguish between intracranial EEG time courses with seizures and the seizure-free ones based on complexity measures. An estimate of multi scaling properties with a large spectrum measured by using the generalized Hurst exponent to characterize the EEG signals. These estimates were able to correctly classify the seizure intervals tested on a given data set and using the k-nearest neighbor classifier and with tenfold cross-validation.Less accurate model.

G.Wang et al.[17] aim to improve the treatment and diagnosis of medically refractory epilepsy patients. Using directed transfer function (DTF), they developed a new algorithm for epileptic seizure detection. The authors used the sliding window technique for EEG recording segmentation. The DTF algorithm used to calculate cerebral functional connectivity. Then, the total information outflow based on the DTF-derived connectivity was calculated by adding up the information flow from a single EEG channel to other channels. Finally, the information outflow was assigned as the features of the support vector machine (SVM) classifier to discriminate interictal and ictal EEG segments.Time consuming is the major drawback of this model.

For features extraction and classification of EEG signals, a reference model is introduced by P.Jahankhani et al.in [18] identifies a region of interests from the set of time series. These regions (also known as events) encodes the most relevant information for the classification task, hence, no need to process the whole time series. Then, the time-frequency analysis is conducted using a Discrete Wavelet Transform (DWT). Finally, the Adaptive Fuzzy Inference Neural Network System is used for the classification.

A two-phase system was proposed by V.Kodogiannis et al. [19],where in the first phase (preprocessing phase), a wavelet transformation is used to extract essential features from the EEG signals. A learning based technique is then applied in phase 2 to classify the extracted features into the correct classes. Due to the large complexity of the extracted features, multiple sub-classifiers were combined together to perform the classification task. Each sub-classifier is an “expert” subdomain.High computational complexity reduced the system performance.

Chapter 3

MATERIAL AND METHODOLOGY

3.1 Problem Definition

Each classifier has its own merits and demerits, depending on the dataset attributes and requirements . In general, it is very difficult to point out which classifier was the most effective for brain datasets. To identify the capable classifier, several classifiers have been tested on EEG datasets and their performance has been evaluated, and the one which performs well is to be considered in solving seizure detection.This approach is designed to classify the abnormal signal from the normal one automatically. It aims to find the best machine learning classifiers for improving the accuracy of epileptic seizure detection. Aims to reduce computational costs.This approach uses various five machine learning models such as Logistic Regression,Decision Trees,Random Forest, Gradient Boosting classifier,Extremely Random Trees.This work aims to find which is the best accurate for epilepsy detection.

3.2 Proposed System

To overcome the limitations of traditional epilepsy diagnosis, the method of automatically detecting epilepsy is developed with better accuracy these years. Traditionally, epilepsy diagnosis was performed by expert radiologists and/or doctors by analysing EEG of the brain.Automated diagnostic methods have been developed in recent years as an alternative to expert diagnosis. Machine learning-based EEG signal processing has been shown to be effective in automated diagnosis of

epilepsy. However, traditional diagnosis is time consuming and expensive to obtain in medical applications as it requires the input of human experts.Here the proposed system uses various type of machine learning classifiers .There are less prone to human errors and less time consuming and computational costs.

3.3 System Architecture

The proposed method uses four classifiers to classify the EEG signals for epilepsy seizure detection. Figure 3.1 shows the flow chart of the proposed methodology.

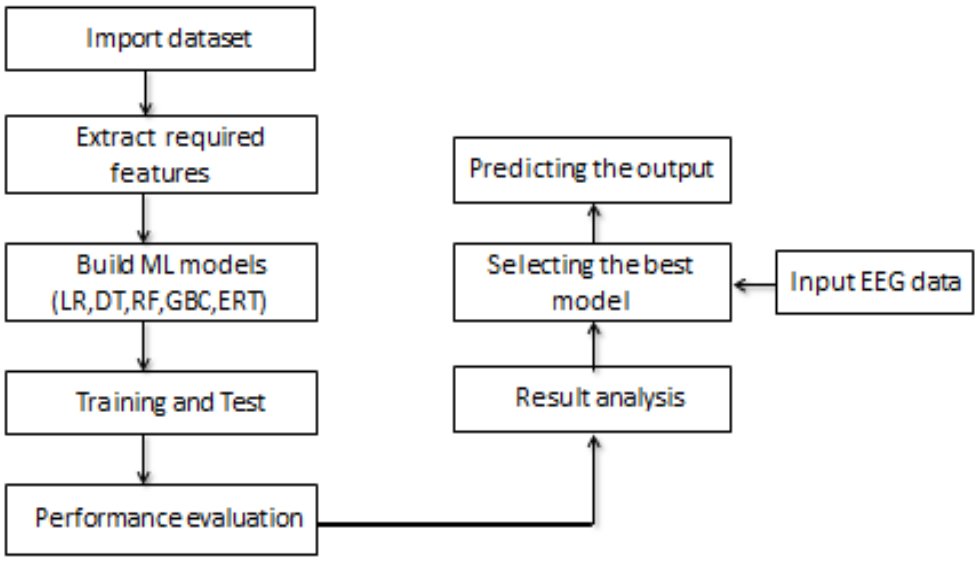

Figure 3.1: System Architecture

The proposed model uses publicly accessible EEG data from Bonn University, where the data include five sets (1, 2, 3, 4, and 5). The feature extraction stage aims to extract the statistical features from the EEG signals .At next stage various ML classifiers are applied and proposed system was tested under different measurement metrics such as Accuracy, and Specificity.Based

on the measurement analysis of various classifiers the best one was selected for further disease detection process.

3.4 Dataset

This EEG database is publically available database provided by the Bonn University .This dataset includes 4097 electroencephalograms (EEG) readings per patient over 23.5 seconds, with 500 patients in total. The 4097 data points were then divided equally into 23 chunks per patient; each chunk is translated into one row in the dataset. Each row contains 178 readings that are turned into columns; in other words, there are 178 columns that make up one second of EEG readings. All in all, there are 11,500 rows and 180 columns with the first being patient ID and the last column containing the status of the patient, whether the patient is having a seizure or not. Therefore,the last column represents the label y 1,2,3,4,5.

The response variable is y in column 179, the Explanatory variables X1, X2, ..., X178

y contains the category of the 178-dimensional input vector. Specifically y in 1, 2, 3, 4, 5:

5 - eyes open, means when recording the EEG signal of the brain the patient had their eyes open.

4 - eyes closed, means when recording the EEG signal the patient had their eyes closed..

3 - Identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area.

2 - The EEG from the area where the tumor was located.

1 - Recording of seizure activity.

All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure.

Only subjects in class 1 have epileptic seizure.

The below figure 3.2 shows the dataset description.

Classes	The Class Description	The Patient State	The Number of cases
1	Seizure activity is recorded from epileptic patients	General epilepsy (with seizures)	2300
2	The tumor was observed in epileptic patients	Partial epilepsy (without seizures)	2300
3	The E.E.G. signal was recorded from a healthy brain region of epileptic patients	Partial epilepsy (without seizures)	2300
4	eyes closed	Healthy	2300
5	eyes opened	Healthy	2300

Figure 3.2: Dataset Description and number of cases in each class

3.5 Data Preprocessing

One of the most critical and required stages in machine learning is data preparation. This technique is essential for reliable, accurate, and successful prediction outcomes when using machine learning algorithms in a data set . Data preparation is a methodology that includes turning raw and unprocessed data into a suitable format for the classification process. Data from the real world is frequently insufficient, untrustworthy, and/or lacking in specific behaviors or patterns, as well as including various mistakes. Pre-processing data is a tried-and-true way to solve such issues. Raw data is pre-processed to make it ready for subsequent processing. In this dataset, there are no missing values (NAN). According to figure 3.3 , there is an unbalanced class distribution issue; to avoid this, sampling approache is used. Sampling is a term that refers to a series of strategies

for balancing the class distribution in a classification dataset with a skewed class distribution.The figure 3.3 shows sample view of epileptic dataset.

	Unnamed	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
0	X21.V1.791	135	190	229	223	192	125	55	-9	-33	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	4
1	X15.V1.924	386	382	356	331	320	315	307	272	244	...	164	150	146	152	157	156	154	143	129	1
2	X8.V1.1	-32	-39	-47	-37	-32	-36	-57	-73	-85	...	57	64	48	19	-12	-30	-35	-35	-36	5
3	X16.V1.60	-105	-101	-96	-92	-89	-95	-102	-100	-87	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
4	X20.V1.54	-9	-65	-98	-102	-78	-48	-16	0	-21	...	4	2	-12	-32	-41	-65	-83	-89	-73	5

Figure 3.3: The epileptic seizure dataset in a sample view

3.6 Feature Extraction

Feature extraction is useful in data visualization and comprehension. In addition, it reduces the requirement for data calculation and storage as well as the time for training and application. Numerous signal feature extraction algorithms are used in practice.In an automated seizure detection system, the distinctiveness of the EEG signals before, during, and after a seizure has to be determined and evaluated. Several features have been identified to better describe the behavior of seizures. Selecting features that best describe the behavior of EEG signals is important for seizure detection and classifier performance. Many types of features and processing techniques have been proposed.Here Time–frequency-domain features are used as a feature extraction method.

3.7 Techniques used for classification

3.7.1 Classification with Logistic Regression

Logistic regression(LR) is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In logistic regression, the model predicts if something is true or false, rather than predicting something continuous. The model fits a linear decision boundary for both classes, then is passed through a sigmoid function to transform from the log of odds to the probability that the sample belongs to the positive class. Because the model tries to find the best separation between the positive class and negative class, this model performs well when the data separation is noticeable. This is one of the models that require all features be scaled, and that the dependent variable is dichotomous.The figure 3.4 showing the logistic function:

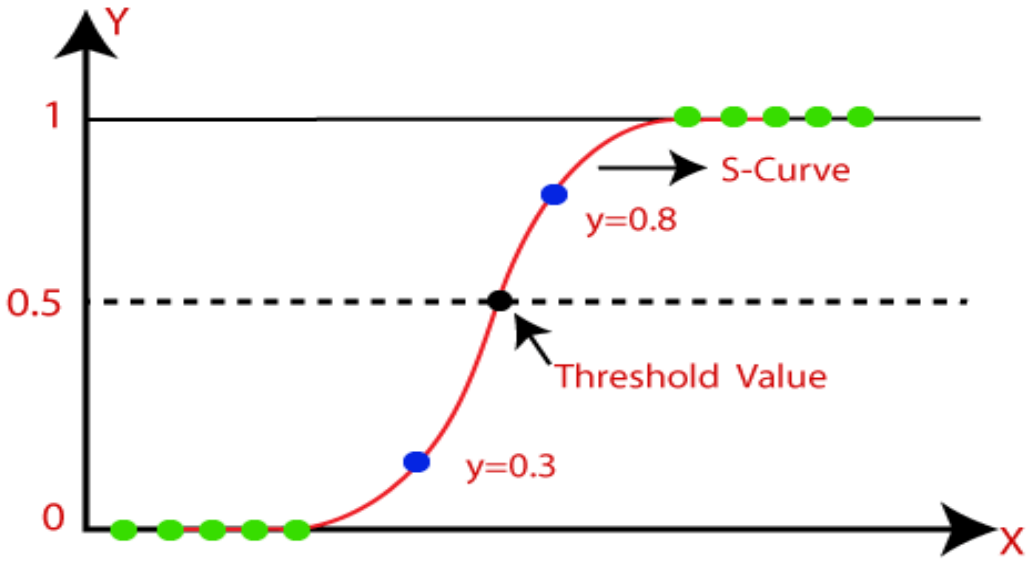

Figure 3.4: Logistic Regression-Classifier model

3.7.2 Classification with Decision Tree

Decision Tree(DT) is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.In order to build a tree, the CART algorithm is used, which stands for Classification and Regression Tree algorithm.A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, it is easy to select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

1.Information Gain

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides us about a class. According to the value of information gain, split the node and build the decision tree. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula.

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

S= Total number of samples
P(yes)= probability of yes
P(no)= probability of no

2.Gini Index

Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits. Gini index can be calculated using the below formula:

$$GiniIndex = 1 - \sum_{j} P_j^2$$

The below figure 3.5 shows decision tree classifier model.

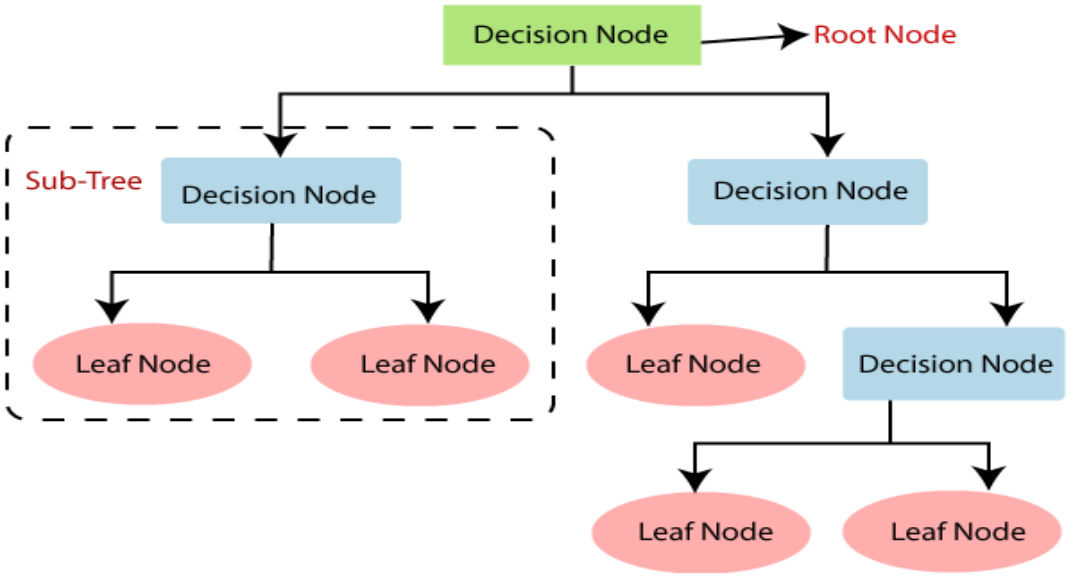

Figure 3.5: Decision Tree-Classifier model

3.7.3 Classification with Random Forest

Random Forest(RF) is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, ”Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.” Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below figure 3.6 shows the random forest classifier model.

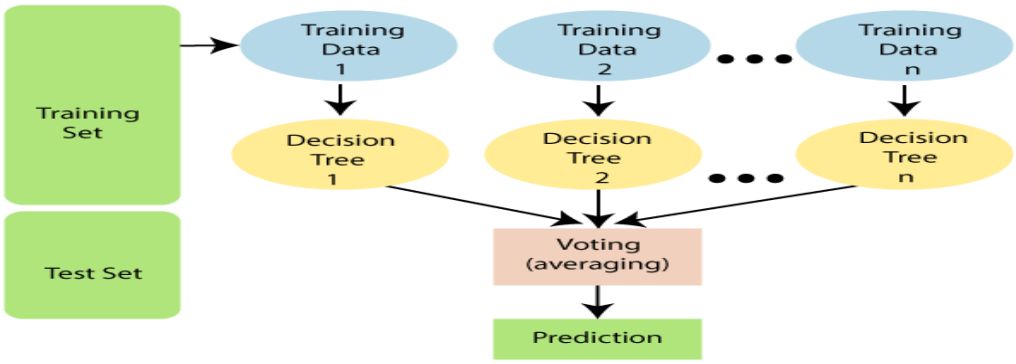

Figure 3.6: Random Forest-Classifier model

3.7.4 Classification with Gradient Boosting classifier

Gradient boosting classifiers(GBC) are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.In Gradient Boosting, each predictor tries to improve on its predecessor by reducing the errors.In gradient boosting decision trees, it combine many weak learners to come up with one strong learner. The weak learners here are the individual decision trees. All the trees are connected in series and each tree tries to minimise the error of the previous tree. Due to this sequential connection, boosting algorithms are usually slow to learn, but also highly accurate. In statistical learning, models that learn slowly perform better.

The weak learners are fit in such a way that each new learner fits into the residuals of the previous step so as the model improves. The final model aggregates the result of each step and thus a strong learner is achieved. A loss function is used to detect the residuals. For instance, mean squared error (MSE) can be used for a regression task and logarithmic loss (log loss) can be used for classification tasks. It is worth noting that existing trees in the model do not change when a new tree is added. The added decision tree fits the residuals from the current model.

The below figure 3.7 represents Gradient Boosting classifier model.

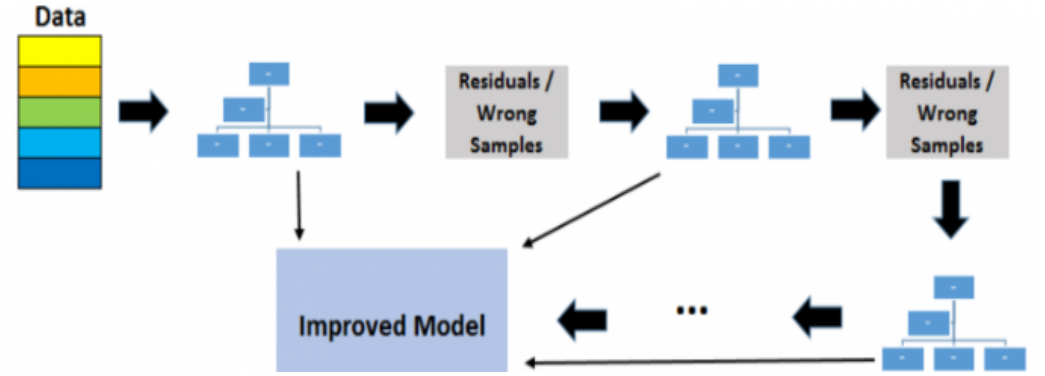

Figure 3.7: Gradient Boosting-Classifier model

3.7.5 Classification with Extremely Random Trees

Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output it's classification result.Specifically, it is an ensemble of decision trees and is related to other ensembles of decision trees algorithms such as bootstrap aggregation (bagging) and random forest.Each Decision Tree in the Extra Trees Forest is constructed from the original training sample.

The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification.

Regression: Predictions made by averaging predictions from decision trees.

Classification: Predictions made by majority voting from decision trees.

In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. The ExtraTrees Classifier is similar to Random Forest except:

- When choosing a variable at the split, samples are drawn from the entire training set rather than bootstrapping samples'

- Node splits are selected at random, instead of being specified like in Random Forest.

This makes the ExtraTrees Classifier less prone to overfit, and it can often produce a more generalized model than Random Forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, Each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to their choice.

The below figure 3.7 represents Extremely Randomized Tree classifier model.

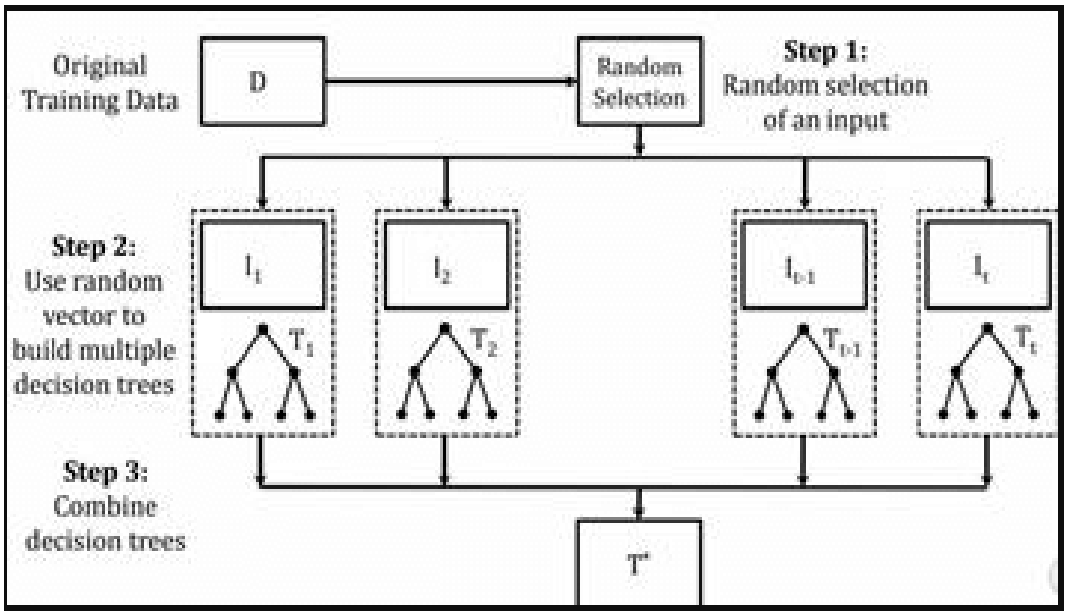

Figure 3.8: Extremely Randomized Tree-Classifier model

3.8 Software Requirement And Specification

3.8.1 Python

Python is an object-oriented programming language created by Guido Rossum in 1989. It's ideally designed for fast prototyping of complicated applications. It has interfaces to several OS system calls and libraries and is protractile to C or C++. several massive corporations use the Python programming language embody NASA, Google, YouTube, BitTorrent, etc. Python programming is widely utilized in AI, natural language Generation, Neural Networks and other advanced fields of computer science. Python is programming language open supply, high-level artificial language developed by Guido van Rossum within the late Eighties and presently administered by Python Software Foundation. It came from the ABC language that he helped produce early on in his career. Python is a powerful language that you can use develop games, write GUIs, and develop web applications. It's a high-level language. Reading and writing codes in Python is far like reading and writing regular English statements. As a result, they're not written in the machine-readable language, Python programs got to be processed before machines can run them. Python is an understood language. This implies

that each time a program is run, its interpreter runs through the code and interprets it into machine-readable byte code. Python is an object-oriented language control users to manage and management data structures or objects to make and run programs. Everything in Python is, in fact, top-notch. All objects, data types, functions, methods, and classes take an equal position in Python. Programming languages are created to satisfy the requirements of programmers and users for an efficient tool to develop applications that impact lives, lifestyles, economy, and society. they assist build lives better by increasing productivity, enhancing communication, and rising potency. Languages die and become obsolete once they fail to live up to expectations and are replaced and superseded by languages that are more powerful. Python programming language artificial language that has stood the test of time and has remained relevant across industries and businesses and among programmers, and individual users. It's a living, thriving, and extremely helpful language that's extremely recommended as a primary programming language for those that want to dive into and experience programming

3.8.2 Google Colaboratory

Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.To be precise, Colab is a free Jupyter notebook environment that runs entirely in the cloud.

The services provided by Google colab includes:

- Write and execute code in Python
- Document the code that supports mathematical equations
- Create/Upload/Share notebooks

- Import/Save notebooks from/to Google Drive
- Import/Publish notebooks from GitHub
- Import external datasets e.g. from Kaggle
- Integrate PyTorch, TensorFlow, Keras, OpenCV
- Free Cloud service with free GPU

Software Requirements

Operating System : Windows 8.1

Python Version : Python 3.6

RAM : 4GB

3.9 Performance Metrics

Performance analysis is done to identify the best model having the highest detection rate.Except for the accuracy, the performance of the classifiers is commonly measured by the following metrics such as precision, recall. These are based on four possible classification outcomes—True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN).

The performance indicators includes:

- True Positives (TP) It is the case when both actual class predicted class of data point is 1.
- True Negatives (TN) It is the case when both actual class predicted class of data point is 0.
- False Positives (FP) It is the case when actual class of data point is 0 predicted class of data point is 1.
- False Negatives (FN) It is the case when actual class of data point is 1 predicted class of data point is 0.

. Derived performance metrics includes:

•Accuracy gives us the percentage of total number of samples that are correctly classified.

Accuracy = T P + T N/T P + T N + F P + F N

•Precision is the ratio of true-positives to the total number of cases that are detected as positive (TP+FP). It is the percentage of selected cases that are correct. High precision means the low false-positive rate.

Precision= T P /T P + F P

• Recall is the ratio of true-positive cases to the cases that are actually positive.

Recall = T P /T P + F N

figure 3.9 shows performane metrics of seizure and non-seizure case.

Acronym	Detection type	Real-world scenario
TP	True-positive	If a person suffers to 'seizure' and also correctly detected as a 'seizure'
TN	True-negative	The person is actually normal and the classifier also detected as a 'non-seizure'
FP	False-positive	Incorrect detection, when the classifier detects the normal patient as a 'seizure' case
FN	False-negative	Incorrect detection, when the classifier detects the person with 'seizure(s)' as a normal person. This is a severe problem in health informatics research

*

Figure 3.9: Metric considering seizure and non-seizure case

Chapter 4

RESULT AND DISCUSSION

4.1 Result of Logistic Regression Classifier

In Logistic Regression Classifier model the rate of AUC at both training and testing phase is very low.The precision and recall values are also very poor.This model posses accuracy of about only 66 percentage.Hence , this model cannot be used for further evaluation.

Below figure 4.1 shows the result of logistic regression model.

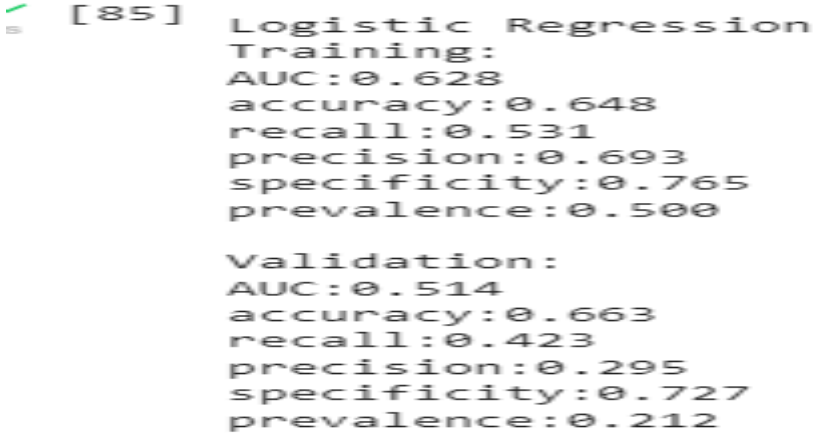

Figure 4.1: Experimental result of Logistic regression

4.2 Result of Decision Tree Classifier

The decision tree classifier attained the AUC 0.986 in training phase and 0.890 in testing phase.The accuracy of this model was only 91. THe figure 4.2 shows the result of decision tree classifier.

```
Decision Tree
Training:
AUC:0.986
accuracy:0.981
recall:0.967
precision:0.995
specificity:0.995
prevalence:0.500

Validation:
AUC:0.890
accuracy:0.911
recall:0.874
precision:0.749
specificity:0.921
prevalence:0.212
```

Figure 4.2: Experimental result of Decision Tree

4.3 Result of Random Forest Classifier

The random forest classifier attained high AUC of 0.998 in training phase and 0.990 in testing phase.The accuracy of this model was 0.967,0.962 in training and testing phase respectively. The figure 4.23 shows the result of random forest classifier.

```
Random Forest
Training:
AUC:0.998
accuracy:0.967
recall:0.946
precision:0.987
specificity:0.987
prevalence:0.500

Validation:
AUC:0.990
accuracy:0.962
recall:0.932
precision:0.895
specificity:0.971
prevalence:0.212
```

Figure 4.3: Experimental result of Random Forest Classifier

4.4 Result of Gradient Boosting classifier

The gradient bosting classifier attained high AUC of 1.00 in training phase and 0.990 in testing phase.All other measurementsin training phase have the value of 1.000.The accuracy of this model was 0.955. THe figure 4.4 shows the result of gradient boosting classifier.

```
Gradient Boosting Classifier
Training:
AUC:1.000
accuracy:1.000
recall:1.000
precision:1.000
specificity:1.000
prevalence:0.500

Validation:
AUC:0.990
accuracy:0.955
recall:0.943
precision:0.860
specificity:0.959
prevalence:0.212
```

Figure 4.4: Experimental result of Gradient Boosting classifier

4.5 Result of Extremely Random Trees Classifier

The extra tree classifier attained the AUC 1.000 in training phase and 0.993 in testing phase.This model attained accuracy about0.961. THe figure 4.5 shows the result of extra tree classifier.

```
Extra Trees Classifier
Training:
AUC:1.000
accuracy:0.997
recall:0.999
precision:0.996
specificity:0.996
prevalence:0.500

Validation:
AUC:0.993
accuracy:0.961
recall:0.970
precision:0.864
specificity:0.959
prevalence:0.212
```

Figure 4.5: Experimental result of Extremely Random Trees classifier

4.6 Model Selection and Validation

The next step is to visualize the performance of all models in one graph. The metric choosed to evaluate the models is the AUC curve. AUC isn't affected by the threshold choosed, so it's a metric that most people use to evaluate their models.Four of the five models have a very high performance, and this is most likely due to the extreme differences in EEG readings between a patient having a seizure and not having one. The decision tree looks like it overfitted as expected, notice the gap between the training AUC and the validation AUC.Among all the other models the extra tree classifier shows high rate of AUC in both training and testing phase.

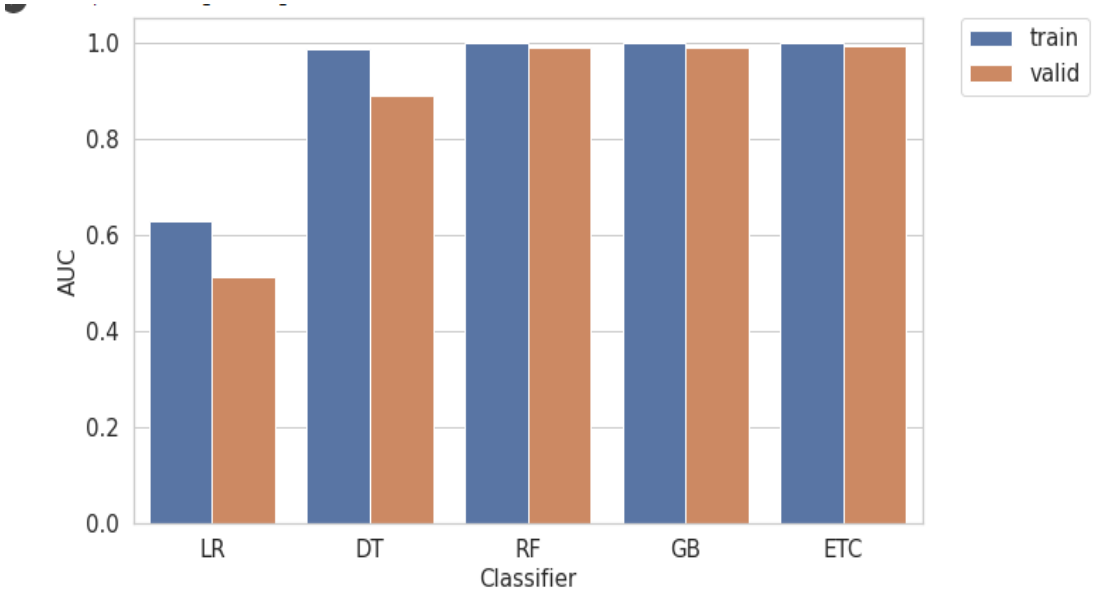

Figure 4.6: Model evaluation using AUC

4.7 Learning Curves

Learning curves are a way to visualize the bias-variance tradeoff in models.Learning curve (or training curve) plots the optimal value of a model's loss function for a training set against this loss function evaluated on a validation data set with same parameters as produced the optimal function.Learning Curves are a great diagnostic tool to determine bias and variance in a supervised machine learning algorithm.The fig 4.7 shows the AUC learning curve for extra trees.

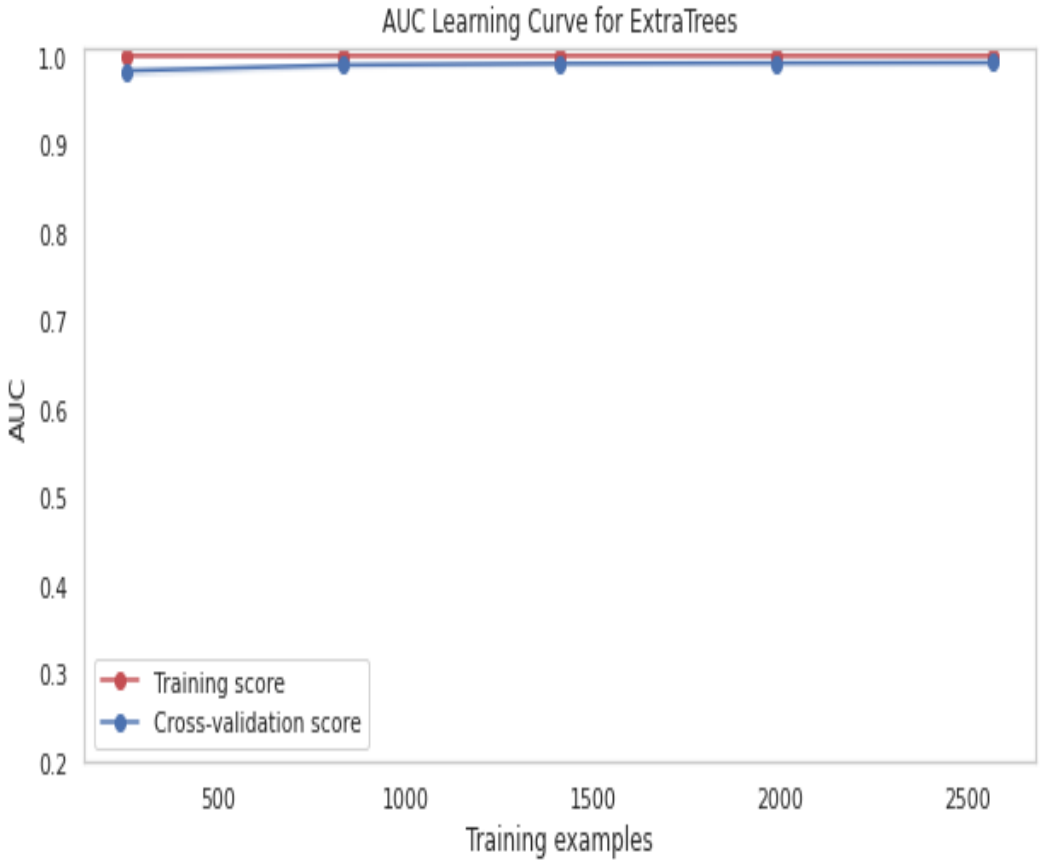

Figure 4.7: AUC Learning Curve for extra trees

4.8 Feature Importance

Loaded both training and testing data with pandas.Obtained the feature score graph. Using the sort() imported from numpy identified the thresholds corresponding to the feature importance (relative importance of each feature in the dataset).The fig 4.8 shows the AUC learning curve for extra trees.

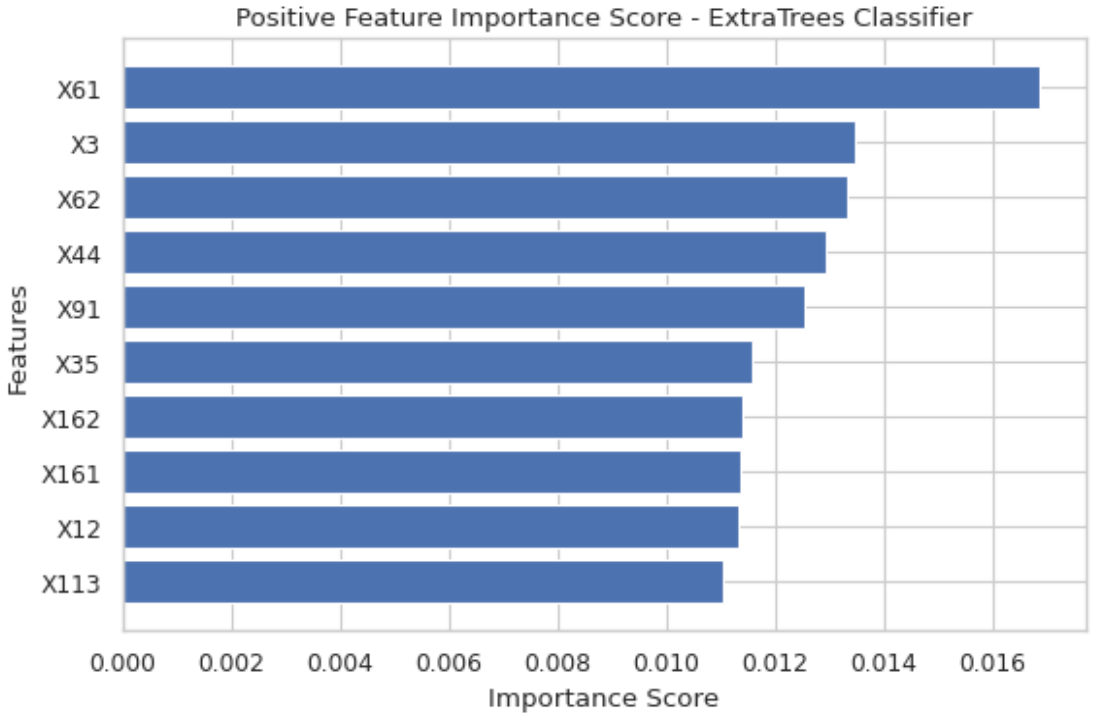

Figure 4.8: Positive Feature Importance Score - ExtraTrees Classifier

4.9 Confusion Matrix

Imported all the required packages and datasets for extra tree classifier.A confusion matrix is a table that is often used to describe the performance of a classification model (or ”classifier”) on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.A confusion matrix is a tabular way of visualizing the performance of prediction model. Each entry in a confusion matrix denotes the number of predictions made by the model.The figure 4.9 shows confusion matrix model and fig 4.10 shows confusion matrix and classification report of extra tree classifier.

Figure 4.9: Confusion matrix model

```
Confusion Matrix:
 [[1327   59]
 [   5  334]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      1386
           1       0.85      0.99      0.91       339

    accuracy                           0.96      1725
   macro avg       0.92      0.97      0.94      1725
weighted avg       0.97      0.96      0.96      1725
```

Figure 4.10: Confusion matrix and classification report of Extra Tree Classifier

Chapter 5

CONCLUSION

Currently, epileptic activity in EEG recordings is mainly examined using a number of traditional and trending technologies. Automation of this process presents many advantages, including a faster diagnosis, continuous monitoring, and reduction in the overall cost of medical treatment . This project, creates a classification machine learning model that can predict whether patients having a seizure or not through EEG readings.This model uses various machine learning classifiers such as logistic regression,decision tree,random forest,gradient boosting and extremely random tree classifier.By evaluating all these classifiers ,the extra trees provide better performance in accuracy of disease detection. This model attains an accuracy of about 96 percentage.

Chapter 6

FUTURE ENHANCEMENT

In recent years, more and more new methods have begun to be applied to the automatic detection of epilepsy. The development of faster and more accurate epilepsy detection models will contribute to epilepsy detection techniques in clinical diagnosis and the development of portable and integrated epilepsy detection equipment. Therefore, a concise and efficient epilepsy detection model will become an inevitable development trend in the future.

- The future work, proposed to investigate the usage of state-of-the-art deep learning networks to overcome the limitations of classical learning models.
- With the development of machine learning, more and more new methods are applied to the feature extraction and classification of epilepsy EEG signals.
- The emergence of deep learning may gradually replace machine learning as the mainstream epilepsy diagnosis method in the future.

REFERENCES

[1] L. Hussain, W. Aziz, A. S. Khan, A. Q. Abbasi, and S. Z. Hassan, "Classification of electroencephlography (EEG) alcoholic and control subjects using machine learning ensemble methods," J. Multidiscip. Eng. Sci. Technol., vol. 2, no. 1, pp. 126–131, Jan. 2015.

[2] A. Hamad, E. H. Houssein, A. E. Hassanien, and A. A. Fahmy, "Feature extraction of epilepsy EEG using discrete wavelet transform," in Proc. 12th Int. Comput. Eng. Conf. (ICENCO), Dec. 2016, pp. 190–195.

[3] U. R. Acharya, S. Vinitha Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated EEG analysis of epilepsy: A review," Knowl.-Based Syst., vol. 45, pp. 147–165, Jun. 2013.

[4] P. Sarma, P. Tripathi, M. P. Sarma, and K. K. Sarma, "Pre-processing and feature extraction techniques for EEGBCI applications-a review of recent research," ADBU J. Eng. Technol., vol. 5, no. 1, pp. 1–8, 2016.

[5] C. Umale, A. Vaidya, S. Shirude, and A. Raut, "Feature extraction techniques and classification algorithms for EEG signals to detect human stress-a review," Int. J. Comput. Appl. Technol. Res., vol. 5, no. 1, pp. 8–14, Jan. 2016.

[6] K. Polat and S. Güneş, "Artificial immune recognition system with fuzzy resource allocation mechanism classifier, principal component analysis and FFT method based new hybrid automated identification system for classification of EEG signals," Expert Syst. Appl., vol. 34, no. 3, pp. 2039–2048, Apr. 2008.

[7] O. Salem, A. Naseem, and A. Mehaoua, "Epileptic seizure detection from eeg signal using discrete wavelet transform and ant colony classifier," in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2014, pp. 3529–3534.

[8] S. A. Aljawarneh, V. Radhakrishna, and A. Cheruvu, "VRKSHA: A novel tree structure for time-profiled temporal association mining," in Neural Computing and Applications. Cham, Switzerland: Springer, 2017, pp. 1–29. [Online]. Available: https://link.springer.com/article/ 10.1007

[9] S. A. Aljawarneh, R. Vangipuram, V. K. Puligadda, and J. Vinjamuri, "G-SPAMINE: An approach to discover temporal association patterns and trends in Internet of Things," Future Gener. Comput. Syst., vol. 74, pp. 430–443, Sep. 2017.

[10] V. Radhakrishna, S. A. Aljawarneh, P. Veereswara Kumar, and V. Janaki, "ASTRA—A novel interest measure for unearthing latent temporal associations and trends through extending basic Gaussian membership function," Multimedia Tools Appl., vol. 78, no. 4, pp. 4217–4265, Feb. 2019.

[11] D. Chen, S. Wan, J. Xiang, and F. S. Bao, "A high-performance seizure detection algorithm based on Discrete Wavelet Transform (DWT) and EEG," PLoS ONE, vol. 12, no. 3, Mar. 2017, Art. no. e0173138.

[12] A. Sharmila and P. Geethanjali, "DWT based detection of epileptic seizure from eeg signals using naive Bayes and k-NN classifiers," IEEE Access, vol. 4, pp. 7716–7727, 2016.

[13] S. Madan, K. Srivastava, A. Sharmila, and P. Mahalakshmi, "A case study on discrete wavelet transform based hurst exponent for epilepsy detection," J. Med. Eng. Technol., vol. 42, no. 1, pp. 9–17, Jan. 2018.

[14] D. Selvathi and V. K. Meera, "Realization of epileptic seizure detection in EEG signal using wavelet transform and SVM classifier," in Proc. Int. Conf. Signal Process. Commun. (ICSPC), Coimbatore, India, Jul. 2017, pp. 18–22, doi: 10.1109/cspc.2017.8305848.

[15] B. Harender and R. K. Sharma, "DWT based epileptic seizure detection

from EEG signal using k-NN classifier," in Proc. Int. Conf. Trends Electron. Informat. (ICEI), Tirunelveli, India, May 2017, pp. 762–765, doi: 10.1109/icoei.2017.8300806

[16] S. Lahmiri and A. Shmuel, "Accurate classification of seizure and seizure-free intervals of intracranial EEG signals from epileptic patients," IEEE Trans. Instrum. Meas., vol. 68, no. 3, pp. 791–796, Mar. 2019, doi: 10. 1109/tim.2018.2855518

[17] G. Wang, D. Ren, K. Li, D. Wang, M. Wang, and X. Yan, "EEG-based detection of epileptic seizures through the use of a directed transfer function method," IEEE Access, vol. 6, pp. 47189–47198, 2018, doi: 10. 1109/access.2018.2867008.

[18] P. Jahankhani, J. A. Lara, A. Pérez, and J. P. Valente, "Two different approaches of feature extraction for classifying the EEG signals," in Engineering Applications of Neural Networks, vol. 363. L. Iliadis and C. Jayne. Berlin, Germany: Springer, 2011.

[19] P. Jahankhani, V. Kodogiannis, and K. Revett, "EEG signal classification using wavelet feature extraction and neural networks," in Proc. IEEE John Vincent Atanasoff Int. Symp. Mod. Comput. (JVA), Sofia, Bulgaria, Oct. 2006, pp. 120–124, doi: 10.1109/jva.2006.17.

[20] V. Radhakrishna, S. A. Aljawarneh, V. Janaki, and P. Kumar, "Looking into the possibility for designing normal distribution based dissimilarity measure to discover time profiled association patterns," in Proc. Int. Conf. Eng. MIS (ICEMIS), May 2017, pp. 1–5