Analysis of Psychiatric Disorders from EEG Signals Using Machine-Learning Techniques

by

Md. Makam Ul Hasan Chowdhury (180021129)

Nadim Ahmed (180021134)

Hasib Arman Chowdhury (180021140)

A Thesis Submitted to the Academic Faculty in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering

Islamic University of Technology (IUT)

Gazipur, Bangladesh

May 2023

Analysis of Psychiatric Disorders from EEG Signals Using Machine-Learning Techniques

Approved By

Mr. Fahim Faisal
Supervisor and Assistant Professor,

Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Boardbazar, Gazipur-1704.

Date: 07/06/2023

Declaration of Authorship

We, Makam Ul Hasan Chowdhury (180021129), Nadim Ahmed (180021134), Hasib Arman Chowdhury (180021140), declare that this thesis titled, 'Analysis of Psychiatric Disorders from EEG Signals Using Machine-Learning Techniques' and the works presented in it are our own.

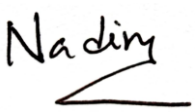
We confirm that:

- This work has been done for the partial fulfillment of the Bachelor of Science in Electrical and Electronic Engineering degree at this university.
- Any part of this thesis has not been submitted anywhere else for obtaining any degree.
- Where we have consulted the published work of others, we have always clearly attributed the sources.

Submitted By

06-05-2023

Md. Makam Ul Hasan Chowdhury (180021129)



Nadim Ahmed (180021134)

Hasib Arman Chowdhury (180021140)

Table of Contents

List of Tables ..... **vii**

List of Figures ..... **viii**

List of Acronyms ..... **ix**

Acknowledgements ..... **xi**

Abstract ..... **xii**

Chapter 1 ..... **1**

Introduction **1**

1.1 Problem Statement and Motivation **1**

1.2 Research Objectives **3**

1.3 Literature Review **3**

Chapter 2 ..... **5**

Machine Learning Algorithms **5**

2.1 Random Forest (RF) **6**

2.2 Light Gradient Boosting Machine (LightGBM) **7**

2.3 Support Vector Machine (SVM) **9**

2.4 Extreme Gradient Boosting Random Forest (XGBRF) **11**

Chapter 3 ..... **12**

Methodology **12**

3.1 Dataset Description **12**

3.1.1 Dataset Overview **12**

3.1.2 Data Collection Procedure **12**

3.1.3 Participant Demographics **13**

3.1.4 EEG Data Format and Preprocessing **15**

3.1.5 Annotations and Ground Truth Labels **15**

3.1.6 Data Availability and Usage Restrictions **15**

3.2 Pipeline **16**

3.3 Data Preprocessing **17**

3.3.1 Label Encoding **17**

3.3.2 Up Sampling of Data **18**

3.3.3 Feature Extraction **19**

3.4 Model Training and Evaluation **20**

3.4.1 Cross Validation **20**

3.4.2 Model Selection **21**

3.4.3 Model Evaluation **21**

3.4.4 Hyper-parameter Tuning **23**

Chapter 4 ..... **24**

Performance Metrics **24**

4.1 Accuracy **26**

4.2 Recall **27**

4.3 Precision **29**

4.4 F1- Score **30**

4.5 Receiver Operating Characteristics (ROC) **31**

Chapter 5 ..... **33**

Results and Discussion **33**

5.1 Simulation Outcomes **33**

5.2 Results **38**

5.3 Comparative Analysis **41**

5.4 Feature Importance **42**

Chapter 6 ..... **43**

Potential Future Development **43**

6.1 Gender Specific Classification **44**

6.2 Application with Different Datasets or Merging Multiple Datasets **44**

6.3 Implementation of Deep Learning **44**

6.4 Age-Specific Classification **45**

6.5 Efficiency Improvement **45**

6.6 Implementation of More Hybrid Algorithms **45**

Chapter 7 ..... **46**

Conclusion **46**

References **48**

List of Tables

Table 4.1 Confusion Table format 25

Table 4.2 Performance Parameter and Corresponding Equations..... 25

Table 5.1.1: Confusion matrix for binary classification (5 fold) 33

Table 5.1.2: Confusion matrix for binary classification (10 fold) 33

Table 5.1.3: Comparison between 5 fold and 10 fold..... 34

Table 5.1.4: Confusion matrix for MCC (5 fold)..... 34

Table 5.1.5: Confusion matrix for MCC after SMOTE and SPCA (5 fold)..... 35

Table 5.1.6: Confusion Matrix for Specific Disorder using **SPCA** in case of **SVM**..... 36

Table 5.1.7: Confusion Matrix for Specific Disorder using **ANOVA** in case of **SVM** 37

Table 5.2.1: Binary Class Classification Performance Metrics 38

Table 5.2.2: Performance Metrics of Main Disorders Using SPARSE PCA 38

Table 5.2.3: Performance Metrics of Main Disorders Using ANOVA 39

Table 5.2.4: Performance Metrics of Specific Disorders Using SPARSE PCA..... 39

Table 5.2.5: Performance Metrics of Specific Disorders Using ANOVA..... 40

Table 5.3.1: Comparison with Relevant Works..... 41

List of Figures

Fig 2.1.1: Structure of a Random Forest 6

Fig 2.2.2: Leaf-wise Tree Growth 8

Fig 2.3.1: SVM Classified Using Decision Boundary (Hyperplane) 10

Fig 3.1.1: Coverage of Different Main Disorders 14

Fig 3.1.2: Coverage of Different Specific Disorders 14

Fig 3.2.1: Proposed Pipeline 16

Fig 5.4.1: Important Features for Main 42

Fig 5.4.2: Important Features for Specific Disorders 42

Fig 6.1: Overview of Future Development 43

List of Acronyms

ADHD	Attention Deficit Hyperactivity Disorder
ANOVA	Analysis of Variance
AUC	Area Under the ROC Curve
BAD	Behavioral Addiction Disorder
CV	Cross Validation
CNN	Convolutional Neural Network
DD	Depressive Disorder
DL	Deep Learning
DMN	Default Model Network
DSM-5	Diagnosis and Statistical Manual of Mental Disorders, Fifth Edition
DT	Decision Tree
EDF	European Data Format
EEG	Electroencephalographic
EC	Effective Connectivity
FMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
GB	Gradient Boosting
GBDT	Gradient Boosted Decision Trees
GBM	Gradient Boosting Machine
GSCV	Grid Search Cross Validation
HC	Healthy Control
IENN	Improved Elman Neural Network
IRB	Institutional Review Board
KNN	K Nearest Neighbors
LDA	Linear Discriminant Neural Network
LGBM	Light Gradient-Boosting Machine
LR	Logistic Regression

MCC	Multi-Class Classification
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NEXUS	Neurophysiology Data Transfer Standard
NI	Neuroimaging
OSF	Open Science Framework
PCA	Principal Component Analysis
PD	Psychiatric Disorder
PTSD	Post-traumatic Stress Disorder
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
RSCV	Random Search Cross Validation
SAD	Social Anxiety Disorder
SMOTE	Synthetic Minority Over-Sampling Technique
SPCA	Sparse Principal Component Analysis
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WHO	World Health Organization
XGB	Extreme Gradient Boosting
XGBOOST	Extreme Gradient Boosting
XGBRF	Extreme Gradient Boosting Random Forest

Acknowledgements

First and foremost, we would like to thank the Almighty for this opportunity and would like to express our gratitude to our beloved parents for their unending support and pushing us to come this far in our career. This research is dedicated to them as an attempt to pay tribute to their sacrifices and contributions.

We would like to express our heartfelt gratitude to our thesis advisor, Mr. Fahim Faisal, for his motivation and instructions. Without his keen sense of attention to detail, this research would never have come into fruition and we, as a team, could not ask for a better mentor. His responsiveness and enthusiasm was an inspiration to all three of us and it provided the drive required to finish this thesis in time and for that, we are eternally grateful.

And lastly, we acknowledge the department of Electrical and Electronic Engineering, Islamic University of Technology for letting us conduct a research that contributes to medical sciences and real world problems based on machine-learning which is becoming more of a necessity rather than a fascinating field of study.

Abstract

Over the past few years, Psychiatric Disorders (PD) have had a significant impact on global health and their prevalence has been leading towards major adversities like functional disabilities and even suicide. These disorders can be divided into some major and specific categories which show different symptoms and have different remedies accordingly. This study makes an attempt to tackle this problem head on and detect said disorders to allow the patients to take necessary actions before the point of no return. For efficient and trustworthy detection of PD, a Machine Learning (ML) approach has been taken and different algorithms were run on the chosen dataset. The dataset that was used for this study was collected from Neuroimaging (NI), Electroencephalography (EEG), tests which have a variety of distinctive features. It was observed that EEG is a reliable and effective way of collecting brain signals which can later be used in different studies like this one. Judging by the magnitude of the samples taken by the EEG device, it was decided that ML would be a very useful tool in this regard and with good accuracies, an acceptable structure can be created. The goal of this study to make a contribution to application of machine learning algorithms in medical sciences and also to call attention to the capabilities of EEG in the prompt detection of PD.

From the findings of this study, it can be observed that very high accuracy was obtained for both the binary and multiclass classifications. The results were tabulated taking samples by using feature selection and feature extraction methods. The highest accuracy for main disorder for multiclass classification was 80.69% and that of the specific disorder was 87.52% both of which used SPARSE PCA feature extraction method. This study makes a solid attempt at addressing this rising issue with a very satisfactory approach and thus makes a fruitful contribution to the medical and data science field for addressing similar adversities in the process.

Chapter 1

Introduction

The purpose of this chapter is to provide an introduction to detection of PD from EEG by implementing some ML algorithms.

1.1 Problem Statement and Motivation

Psychiatric disorders are, in fact, a force to be reckoned with when it comes to functioning normally in everyday life and can have a wide range of adverse effects on individuals as well as communities as a whole. When investigated, it has been observed that such disorders have effects from multiple dimensions as opposed to having a linear impact. According to sources like the World Health Organization (WHO), about 10% of the entire population of the world is suffering from one or more types of PD and about 25% get affected by these disorders at some point in their lifetime. Being diagnosed with PD can make an individual isolated from society which can result in discriminations in family as well as workplace environments as there is a substantial amount of social stigma associated with such disorders. Besides, the overall quality of life is hampered greatly since the patients face impairment in performing everyday tasks. Moreover, the chances of causing physical health problems and functional disabilities are increased greatly which can be triggered by the psychological effects of the disorder itself. Even extremities like risk of suicide have to be taken into consideration as it is not uncommon in such scenarios.

During and after the COVID-19 pandemic, the cases of PD have risen significantly as well as increased their severity. Effects such as extended depression, anxiety and substance abuse are some of the few aftereffects which prevails within the general masses. Due to the disruption caused by it in mental health services, the newly affected ones were not attended to properly which

sustained till today. It has been found out from studies that among the Chinese people, stress disorder increased to 73.4%, depression to 50.7% and anxiety to disorder to 44.7% as a result of this outbreak [1]. This goes to show how such an unanticipated event can spike the rate of PD and with that, their aftermaths and makes a compelling case on how it is so important to identify those disorders and work on them as promptly as possible.

One of the most efficacious noninvasive NI techniques is EEG which gives the user the option of observing the neurobehavioral networks with the help of time and frequency scales by placing electrodes on the scalp. EEG is unique for its higher temporal resolution which allows higher accuracy when it comes to neural activities [4]. In numerous medical studies, EEG modules were used to investigate the cognitive models and behavioral trends in performing tasks when it came to psychiatric disorders [2]. In comparison to other neural diagnosis and prognosis devices, EEG is cheaper and physician friendly. It allows ambulatory diagnosis for the patients allowing them to receive healthcare at a reasonable cost [3]. Though the whole process is fairly effective and inexpensive and does not require outrageous footprint, the specialists are needed to be trained for an extensive period [5]. Overall, EEG can be considered to be a reliable source of data for the neural data for a number of demographics of patients and can be used to build a dataset in the process for future applications.

A dataset made up of EEG data was selected and the tool that was relied upon throughout this study was machine learning for a variety of upsides like its reliability and efficiency. This study in particular deals with both main disorders (mood disorder, addictive disorder, trauma and stress related disorder, schizophrenia, anxiety disorder and obsessive compulsive disorder) as well as specific disorders (depressive disorder, schizophrenia, alcohol use disorder, behavioral addiction disorder, bipolar disorder, PTSD, Social Anxiety Disorder (SAD), obsessive compulsive disorder, acute stress disorder and adjustment disorder) incorporating binary and Multi-Class Classification (MCC) in machine learning which can be considered as a unique feature. A number of machine learning algorithms such as random forest (RF), light gradient-boosting machine (LGBM), support vector machine (SVM) and hybrid Extreme Gradient Boosting Random Forest (XGBRF) were used on the chosen dataset to come to a conclusive and satisfactory set of results.

1.2 Research Objectives

The goal of this research is to distinguish between Healthy Control (HC) subjects and unhealthy patients and to further narrow down on their disorders. Besides separating the two, the objective of this work is to categorize the test subjects into some main disorders and then some subcategories, namely specific disorders, with satisfactory accuracy. For feature selection and extraction, Analysis of Variance (ANOVA) and SPARSE Principal Component Analysis (PCA) were used respectively and later Synthetic Minority Over-Sampling Technique (SMOTE) was used for data augmentation and up sampling. After the preprocessing and manipulation of data, certain metrics were considered for evaluating the impact of the models such as confusion matrices, F1 score, recall and Receiver Operating Characteristics (ROC) were observed. To sum up, this study makes an attempt at a very swift and accurate diagnosis of any of the prevalent PD using ML-based alternatives.

1.3 Literature Review

A plethora of research has been conducted on PD identification with many using a ML approach. Watts et al. [6] has proposed a binary classification model for predicting the chances of committing criminal activities by psychiatric patients. It was concluded that machine learning models can show greater accuracy than gold-standard risk assessment tools. [2] depicts eight electroencephalographic (EEG) regions of interest were used to examine SAD-related changes in brain connections within the default mode network (DMN). This was the first study to use EEG-based EC to categorize the severity of SAD (control mild, average, and severe) and to look into the topological structure of the effective brain connectivity network in SAD as well as the neural correlates of the DMN resting-state. The main finding was that SAD patients had increased EC between the DMN regions compared to HCs in the resting state. For the diagnosis of psychiatric illness, a statistical ML methodology-based automated diagnosis procedure based on electroencephalography (EEG) data is proposed [7]. This paper illustrates two binary classifications and a multiclass classification. Accordingly, the results are 88.3%, 92.7%, and

87.1%. The feature selection method of Peng [8] and the MFA classification procedure [9], which make up the proposed EEG-based methodology, are found to be extremely effective for the diagnosis of psychiatric disorders.

Park et al. proposed in their study [10], a machine learning (ML) classifier is being created with the intention of using electroencephalography (EEG) to detect and compare the main psychiatric disorders. The support vector machine, random forest, and elastic net machine learning methods were used in this study to make predictions. Then, they compared the results of each prediction performance analysis.

The model with the highest degree of precision used the elastic net technique along with an IQ adjustment. According to a different study [11] by Rahman et al., SVM has the highest accuracy for multiclass classification (95.3%). For analysis, they also used EEG data that was openly accessible.

Additionally, Ranjith et al. used the Improved Elman Neural Network (IENN) to detect stress levels in a dataset of only six people, with an accuracy of 94.12%. Ashford et al. proposed a method with a different approach that performs CNN as those image classifiers and represents the statistical features as 2D images, achieving up to 89.38% accuracy [13]. Iscan et al. in his research stated that EEG systems are inexpensive and can be used anywhere, including private medical practices, in contrast to structural and functional MR imaging, which is very expensive and can only be carried out in specialized facilities. Additionally, resting state EEG is thought to have higher test-retest reliability than traditional resting-state fMRI sequences and measures and is at least on par with the reliability of anatomical MRI measures. [14].

For this study, these researches were studied that provided the core inspiration. It was considered by the team that the approach for this study will be a combination of all the works stated above and prove to be conclusive and trustworthy when it comes to problem-solving in this domain.

Chapter 2

Machine Learning Algorithms

Analysis of PD from EEG signals using ML techniques involves detecting patterns in EEG data that indicate abnormal brain activity associated with specific psychiatric conditions. Anomaly detection plays a crucial role in this task, as it aims to identify instances where the EEG signals deviate from the normal patterns observed in healthy individuals. However, developing effective models for anomaly identification in psychiatric disorders is challenging due to the difficulty in defining comprehensive normal ranges that encompass all possible variations in EEG signals. Moreover, the presence of noise in the data can mimic true anomalies, further complicating the detection process.

In the context of psychiatric disorder analysis, large datasets of EEG signals are often available, with the majority of the data representing normal brain activity. This poses an additional challenge in developing accurate detection algorithms, as the imbalance between normal and abnormal instances can impact the algorithm's performance. Nevertheless, researchers have devoted significant efforts to developing machine learning algorithms specifically tailored for detecting anomalies in EEG signals associated with psychiatric disorders.

This research integrates a range of machine learning techniques to analyze psychiatric disorders using EEG signals. These techniques leverage advanced algorithms and models discussed in the following sections to identify patterns of abnormal brain activity associated with specific psychiatric conditions. By applying machine learning to EEG analysis, this research aims to contribute to the understanding, diagnosis, and treatment of psychiatric disorders.

2.1 Random Forest (RF)

Random Forest (RF) is a machine learning algorithm that combines regression and classification methods. It is an ensemble technique that improves prediction accuracy by combining the predictions of multiple decision trees [15]. During the training period, RF constructs multiple decision trees and generates output classes for each individual tree. RF leverages the concept that a large ensemble of statistically independent systems can outperform individual algorithms. Even without hyper-parameter tuning, RF generally yields reasonable prediction results.

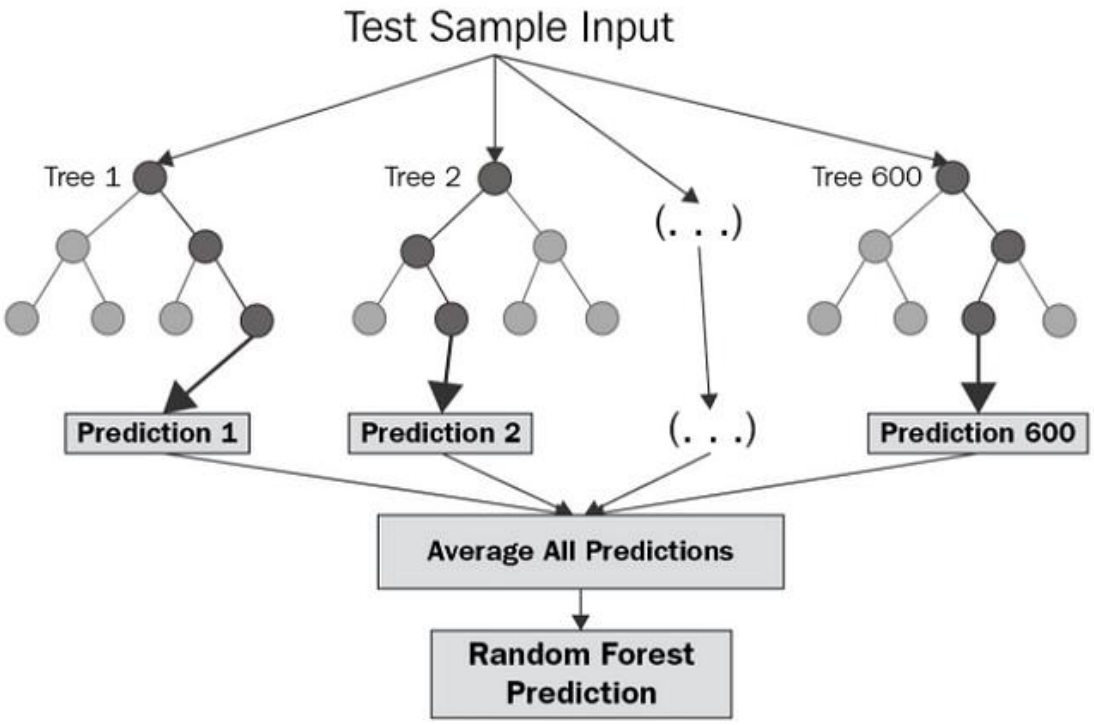


Fig 2.1.1: Structure of a Random Forest

To enhance the diversity and robustness of the individual trees, RF incorporates a technique called bagging. Bagging involves creating new sets of training data by resampling with replacement from the original dataset. Each bootstrap sample is used to train a separate decision tree. By doing so, RF introduces variation among the trees and reduces overfitting.

During the training process, RF also employs a technique called feature sampling. It randomly selects a subset of features from the original set of characteristics to be considered at each node of the decision tree. This further promotes diversity and helps prevent the dominance of certain features in the ensemble.

The prediction of unknown inputs can be determined by the formulae below.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{1}$$

Here, B= Optimal number of trees

Also, uncertainty of the prediction can be written as:

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B}(f_b(x') - \tilde{f})^2}{B-1}} \tag{2}$$

The variance of Random Forest is calculated as follows:

$$\rho\sigma^2 + \frac{1-\rho}{K}\sigma^2 \tag{3}$$

Here σ^2 denotes tree variance, ρ denotes correlation between trees, K represents total trees. [16]

2.2 Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage. It builds upon the traditional gradient boosting

algorithm by introducing innovative techniques that improve performance and scalability. LightGBM employs a leaf-wise growth strategy, unlike other algorithms that grow trees in a level-wise or horizontal manner. This means that LightGBM selects the leaf node with the maximum delta loss to expand during tree growth. In contrast to level-wise algorithms, the leaf-wise approach has the potential to attain superior loss reduction while expanding identical leaves.

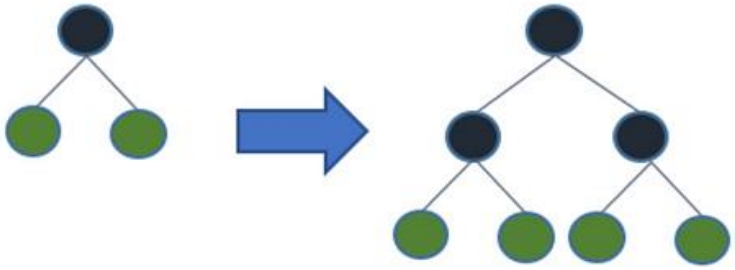


Fig 2.2.1: Level-wise Tree Growth

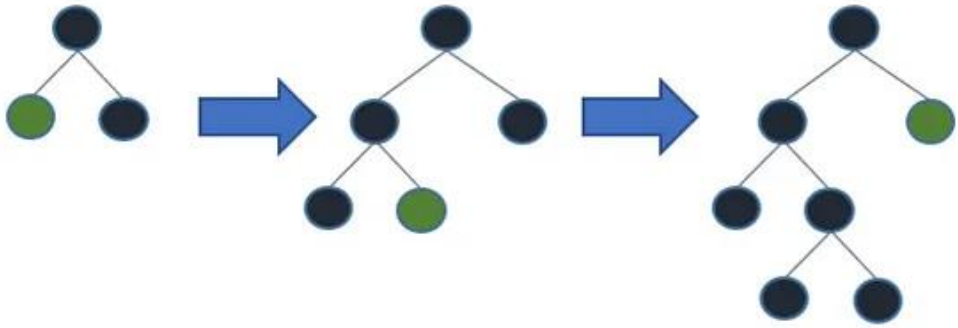


Fig 2.2.2: Leaf-wise Tree Growth

The nomenclature of "Light" in LightGBM is attributed to its exceptional rapidity and efficacy. The escalating magnitude of data poses a formidable challenge to conventional data science algorithms, impeding their ability to furnish expedited outcomes. The Light Gradient Boosting Machine (LightGBM) algorithm is tailored to address the challenge of processing large datasets with limited memory resources. The suitability of LightGBM for addressing large-scale data challenges is evidenced.

The popularity of LightGBM can be attributed to its emphasis on precision in outcome delivery. The objective of LightGBM is to attain elevated levels of prediction accuracy through the utilization of optimization techniques, including the leaf-wise growth strategy. In addition, it is worth noting that LightGBM is equipped with the capability to facilitate GPU learning, thereby empowering data scientists to harness the computational prowess of GPUs, which in turn leads to expedited training and inference. The aforementioned capability has played a significant role in the extensive acceptance of LightGBM across diverse data science application development scenarios.

2.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a robust machine learning methodology that can be effectively employed for both regression and classification tasks. The present study highlights the exceptional performance of the subject under consideration in the domains of intricate data handling and attainment of elevated levels of predictive accuracy. It is based on the concept of finding an optimal hyperplane that best separates different classes or fits the regression line to the data [16]. In the case of binary classification, SVM aims to find a hyperplane in the feature space that maximizes the margin, i.e., the distance between the hyperplane and the nearest data points from each class [17]. This hyperplane serves as the decision boundary for classifying new instances. SVM can handle both linearly separable data and cases where the classes are not linearly separable through the use of a technique called the kernel trick. The kernel trick allows SVM to implicitly map the data into a higher-dimensional feature space, where linear separation becomes possible.

The SVM formulation involves solving an optimization problem to find the optimal hyperplane [10]. The objective is to minimize the hinge loss function, which penalizes misclassified points, while also incorporating a regularization term to control the complexity of the model. Once the optimization problem is solved, the SVM model can classify new instances by evaluating their position with respect to the learned decision boundary.

It selects the critical data points, known as support vectors, to construct the hyperplane that separates different categories. This property of SVM is responsible for its name. To illustrate, let's

consider the following diagram depicting two distinct categories being classified by a decision boundary or hyperplane:

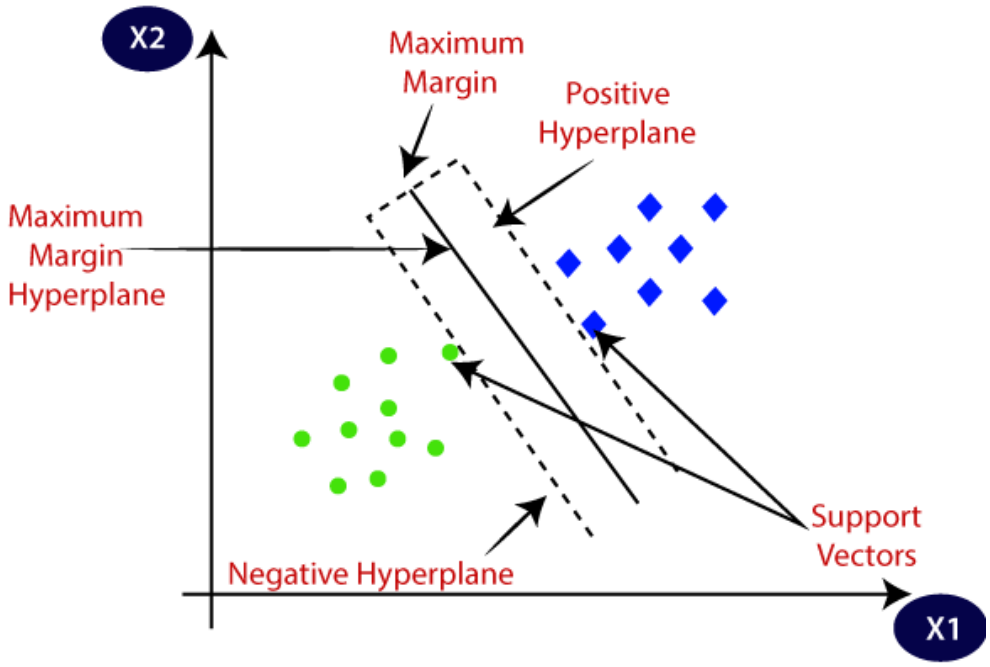


Fig 2.3.1: SVM Classified Using Decision Boundary (Hyperplane)

Its ability to handle complex datasets, support non-linear separation through kernels, and provide strong generalization capabilities makes it a valuable tool in the field of machine learning and data analysis. Based on kernel trick, SVM is as follows:

$$S(x) = sign\left(\sum_{k} \alpha_k y_k k(x_k, x) + b\right) \tag{4}$$

Here $K_x \in R^N$ are support vectors and k (x_k, x) is the kernel function [16]. The decision SVM function is formed on dot product of input feature vector having support vectors which means it has no dimension requirements of the feature vector.

2.4 Extreme Gradient Boosting Random Forest (XGBRF)

XGBRF, which stands for extreme Gradient Boosting Random Forest, is an advanced machine learning algorithm that combines the strengths of gradient boosting and random forest techniques [18]. It is particularly well-suited for solving complex regression and classification problems. In this model, RF just replaces the DT as the base [19].

The XGBRF algorithm is a machine learning technique that integrates the fundamental principles of gradient boosting. This approach entails the iterative inclusion of decision trees into the ensemble, with each subsequent tree rectifying the errors committed by its predecessors [16]. The present study proposes a boosting approach that enhances the predictive efficacy of the model by prioritizing the samples that exhibit a higher degree of complexity in terms of correct classification.

The XGBRF model employs the random forest algorithm, a technique that integrates multiple decision trees to enhance the precision of predictions. The Random Forest algorithm introduces stochasticity into the model building process by selecting subsets of features and training samples for each decision tree in the ensemble. The present study elucidates the creation of a diverse ensemble of decision trees that synergistically collaborate to yield precise predictions and mitigate the backlash of overfitting.

The XGBRF algorithm represents a noteworthy advancement in the domain of machine learning, as it adeptly amalgamates the favorable attributes of gradient boosting and random forest techniques. The amalgamation of various techniques has resulted in the development of XGBRF, which has proven to be a powerful tool for tackling complex machine learning problems. The popularity of a certain tool among data scientists and researchers in diverse domains can be attributed to its capacity to effectively manage voluminous datasets, furnish precise predictions, and offer interpretability.

Chapter 3

Methodology

This chapter presents the methodology used to detect mental disorders using electroencephalography (EEG) using machine learning algorithms. The aim is to provide a comprehensive understanding of the measures taken to accurately and reliably detect mental disorders using EEG signals.

3.1 Dataset Description

3.1.1 Dataset Overview

The dataset utilized in this thesis was available from the Open Science Framework (OSF) repository, especially at the following URL: https://osf.io/8bsvr/. The dataset was composed by a team of researchers from different institutions, and it has been created for the purpose of studying psychiatric disorders using Electro Encephalography (EEG) signal recordings. It is assembled from up of a huge collection of EEG data from people with different kinds of mental health conditions as well as HC participants.

3.1.2 Data Collection Procedure

The EEG data in this dataset was acquired using accepted practices to guarantee consistency and homogeneity among patients. To locate people and gather data, academic institutions, research facilities, and clinical settings were employed. Prior to the data collection, all participants provided their informed consent, and the study's methodology was approved by the relevant ethics committees. [10]

Participants' EEGs were captured throughout the data collecting procedure in meticulously controlled settings with minimum outside interference. Modern EEG equipment was utilized with the correct electrode placements to detect brain activity. The recordings were made either when subjects were at rest or engaged in certain mental activity, depending on the study's design.

3.1.3 Participant Demographics

The data set includes a diverse group of participants of different ages, genders and mental states. Along with EEG data, demographic data of the participants such as age, gender and clinical diagnosis are provided. The sample size and distribution of main and specific mental disorders is as follows:

- Mood Disorder: 266 participants
    - Depressive Disorder (DD): 199 participants
    - Bipolar Disorder: 67 participants
- Schizophrenia: 117 participants
- Addictive Disorder: 186 participants
    - Behavioral Addiction Disorder (BAD): 93 participants
    - Alcohol Use Disorder: 93 participants
- Anxiety Disorder: 107 participants
    - Social Anxiety Disorder (SAD): 48 participants
    - Panic Disorder: 59 participants
- Trauma and Stress Related Disorder: 128 participants
    - Post-traumatic Stress Disorder (PTSD): 52 participants
    - Adjustment Disorder: 38 participants
    - Acute Stress Disorder: 38 participants
- Obsessive Compulsive Disorder: 46 participants
- Healthy Control (HC): 95 participants

The overall coverage of different patients having both main and specific disorders in percentage is shown below in forms of pie charts.

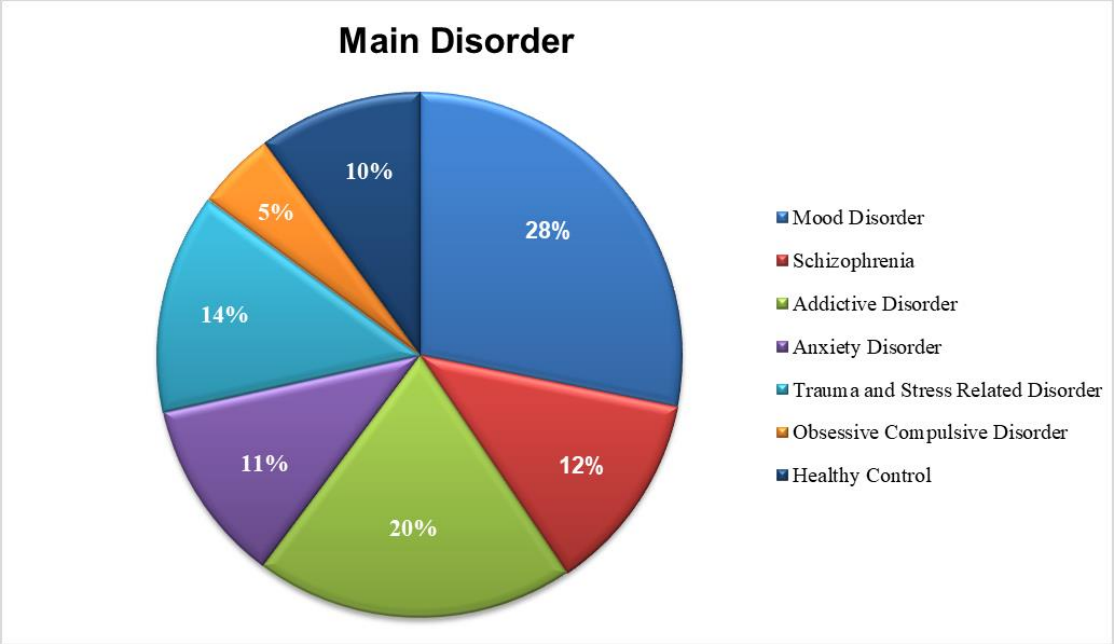


Fig 3.1.1: Coverage of Different Main Disorders

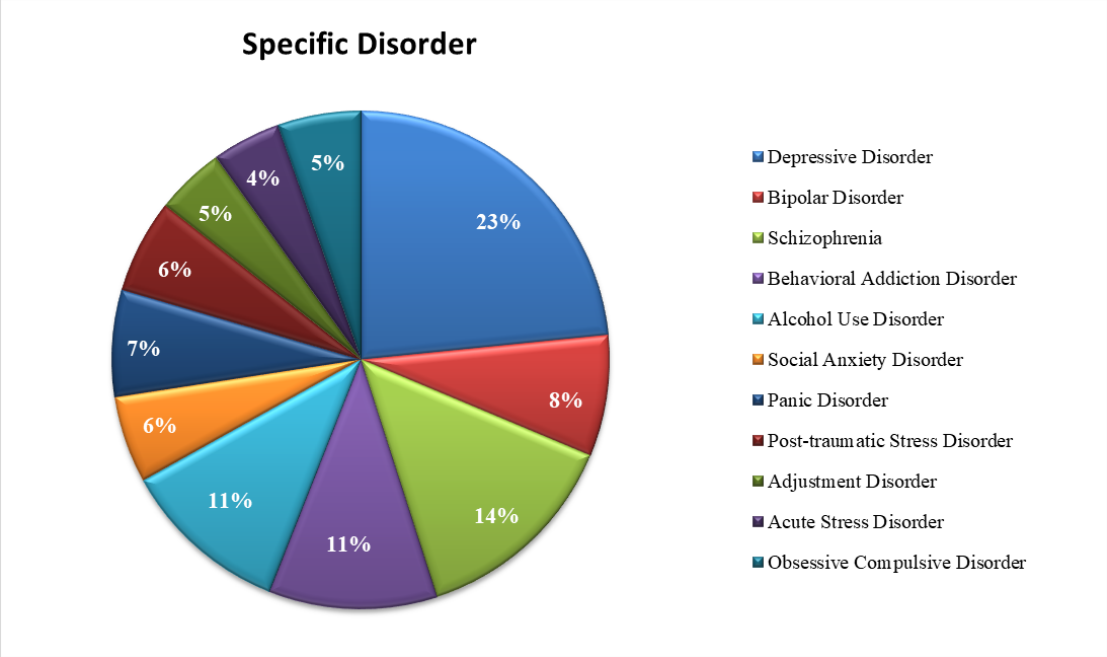


Fig 3.1.2: Coverage of Different Specific Disorders

3.1.4 EEG Data Format and Preprocessing

The EEG data in the data set is kept in a standard format, such as the Neurophysiology Data Transfer Standard (NEXUS) or the European Data Format (EDF). Multiple channels that correspond to various scalp electrode placements make up each EEG recording.

EEG data underwent a number of pre-processing processes to enhance quality and identify pertinent information before analysis. Filtering to remove noise, artifact suppression to get rid of eye blink and muscular artifacts, and down-sampling to simplify computations are common preprocessing methods used on raw EEG data. Epochs were also carried out to divide continuous EEG data into shorter time frames that corresponded to certain activity circumstances or states of rest.

3.1.5 Annotations and Ground Truth Labels

The dataset contains annotations and ground truth labels for each participant to aid in the categorization and evaluation of mental diseases. These labels represent the presence or absence of certain mental diseases that have been identified by clinical professionals in accordance with accepted diagnostic criteria (such as the DSM-5). To guarantee correctness and coherence, annotations are checked by a large number of professionals.

3.1.6 Data Availability and Usage Restrictions

The dataset is freely available for research purposes upon request through the OSF repository. However, it is important to note that the dataset is subject to certain usage restrictions, as specified by the data contributors and outlined in the accompanying documentation. Researchers are encouraged to adhere to the ethical guidelines and data usage agreements when utilizing the dataset.

Of all the participants, there were total 601 male contributors and 344 female contributors aging between 18 years old to 71 years old.

This study was approved by the Institutional Review Board (IRB) (20/16/2019). In line with the retrospective study design, the participants' consent was withdrawn [20].

3.2 Pipeline

What we did was that we first went for binary classification of the dataset where it would detect if a patient were healthy or not. We then multi-classified the main disorder and multi-classified the specific disorder again. It is done so that the detection procedure becomes easier by cross checking the main disorder and specific disorder which would eventually reduce the overall error occurrence of the classifier. The used ML algorithms for these cases are RF, LGBM, SVM, and XGBRF. The overall pipeline diagram is shown in Fig. 3.2.1.

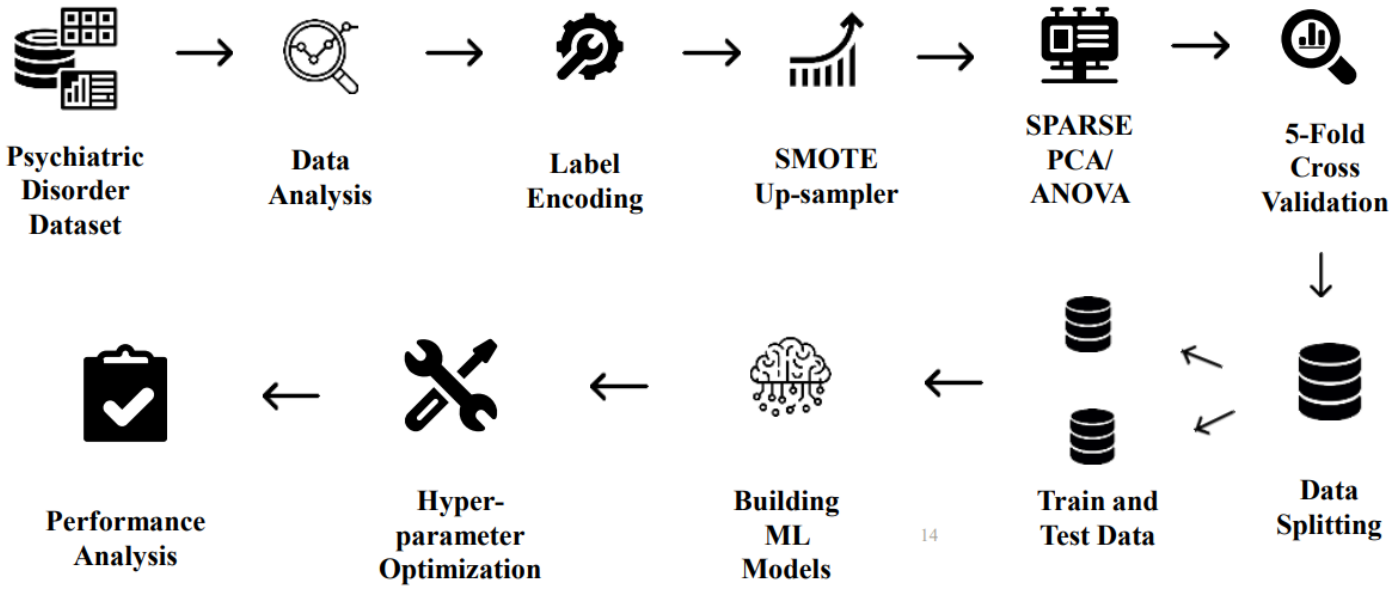


Fig 3.2.1: Proposed Pipeline

3.3 Data Preprocessing

3.3.1 Label Encoding

Label encoding is a crucial step in preparing data for machine learning algorithms. In the context of detecting mental disorders from EEG data, label coding is used to convert categorical labels representing the presence or absence of disorders into numeric values.

Normalization is a technique to ensure that all data in the database have a similar range. In this study, the min-max scaling technique was used to encode the labels. Min-max scaler scales feature values within a specified range, typically from 0 to 1. Min-max Scaler normalization is advantageous in high-dimensional data. [21] Min-Max Scaler is a kind of normalization that can scale all EEG signal values from 0 to 1. Eq. (1) and Eq. (2) shows how Min-Max Scaler is normalized.

$$X_{std} = \frac{(X - X.min)}{(X.max - X.min)} \tag{5}$$

$$X_{scaled} = X_{std} \times (X.max - X.min) + X.min \tag{6}$$

The min-max scaler adjusts the numerical labels' ranges by taking the least value from each label and dividing it by the labels' range of values. In order to facilitate subsequent analysis and model training, this procedure makes sure that the encoded labels are normalized and lie within the desired range.

The categorical classifications for psychiatric diseases are converted into numerical values appropriate for machine learning methods by applying label encoding with the min-max scaler. As a result, the algorithms can successfully identify patterns and generate predictions using the encoded labels. The resultant normalized encoded labels help in the correct identification of psychological conditions from EEG data.

3.3.2 Up Sampling of Data

A critical preprocessing step is balancing the data, especially when working with unbalanced datasets where the proportion of instances in various classes changes greatly. To make sure that the machine learning models are not biased towards the majority class and can accurately capture patterns from all classes, dealing with class imbalance is crucial for diagnosing mental diseases from EEG data.

Our dataset was imbalanced. Thus, we had to balance it. In our case we used the Synthetic Minority Over-Sampling Technique (SMOTE) function in order to balance our dataset. In order to increase the learning algorithm's capacity to predict examples from the minority class, SMOTE was developed to artificially create new instances. SMOTE demonstrated statistically improved performance on the mammography data as well as on a number of other datasets, establishing the groundwork for learning from unbalanced datasets [22].

The SMOTE tool creates synthetic instances by analyzing the feature space and extrapolating between existing minority class samples. These artificial instances are added, up sampling the minority class and increasing its representation in the dataset. By adding this enhancement, the issue of class imbalance is lessened, and it is made sure that the machine learning models have enough data from all classes to develop and produce reliable predictions. The SMOTE function was specifically used in the multi-class classification task to up-sample the minority classes, ensuring a more equitable representation of different psychiatric disorders. This is crucial because the occurrence of specific psychiatric disorders may be relatively rare compared to the overall dataset. SMOTE helps to balance the representation of different disorders by creating synthetic samples which gives the models healthier training data. Before up sampling the dataset there were 945 samples. After excluding HC data, the number of samples became 850. Then using SMOTE function, we made sure that each class had equal number of instances. Hence, while going for multiclass classification for MAIN Disorder total numbers of samples were 1596. In the case of multiclass classification for SPECIFIC Disorder the total numbers of samples after up-sampling were 2189. Because of this, it is more likely that machine learning models developed from balanced data will generalize well and accurately identify psychological disorders from EEG signals.

3.3.3 Feature Extraction

When utilizing machine learning algorithms to analyze EEG data in order to find psychiatric problems, feature extraction is an essential phase. It involves the extraction of the most useful information from the raw EEG signals that can successfully distinguish between various classes. Two feature extraction techniques were used in this work to extract essential features: Sparse Principal Component Analysis (SPCA) and Analysis of Variance (ANOVA).

For the binary classification task of distinguishing between healthy and unhealthy individuals, only the SPCA method was employed. Sparse principal component analysis is an advanced mathematical tool for the analysis of high dimensional data [23]. SPCA is beneficial for generating localized and interpretable patterns of variability, whereas PCA provides little interpretational value [24]. According to empirical findings, sparse PCA did perform as well as PCA in terms of precision and accuracy. By focusing on the most valuable characteristics, the dimensionality reduction method SPCA finds a sparse representation of the data. Applying SPCA reduces the complexity of the EEG data while preserving the important patterns required for the classification problem. This method effectively captures the core mechanisms and distinctive traits that separate persons with psychiatric issues from those who are healthy.

During the multi-class classification tasks of identifying the primary disorder in addition to the specific disorders, both SPCA and ANOVA were utilized for the purpose of feature extraction. The SPCA method was utilized in order to recognize the essential structures within the EEG data, whereas the ANOVA method was employed to zero in on the characteristics that significantly differentiated one group from another.

In the context of ML, ANOVA is an analysis of variance. An effective statistical method for testing hypotheses and determining the significance of differences between groups or levels of a categorical variable is called an ANOVA. ANOVA is a statistical technique that evaluates the importance of group differences. ANOVA techniques are extremely useful in practice. Not only do contemporary methods offer more power, but they also provide new perspectives on how groups of participants differ and by how much [25].

Our dataset was subjected to ANOVA analysis, resulting in the selection of 100 features that exhibited the greatest discriminatory power. The aforementioned characteristics were considered to be the most significant in discriminating among various psychiatric conditions, thereby promoting precise categorization. ANOVA helped to assess the importance of each factor in identifying different psychiatric diseases. It discovers characteristics that differ significantly between classes, making them very discriminating for categorization. The ANOVA technique aids in the selection of the characteristics that are most crucial for differentiating between various psychiatric diseases.

3.4 Model Training and Evaluation

The methodology for detecting psychiatric disorders from EEG data using machine learning algorithms involves critical components of model training and evaluation. The present section delineates the procedures involved in model training, algorithm selection, and performance assessment.

3.4.1 Cross Validation

The technique of cross-validation is extensively employed in the field of machine learning for the purpose of evaluating the efficiency and capacity of models in terms of performance and generalization. The present investigation employed both 5-fold and 10-fold cross-validation techniques to enhance the robustness and dependability of the findings.

In the process of cross-validation, the dataset is partitioned into numerous folds or subsets. The machine learning models are trained on a specific subset of the available data, referred to as the training set. The performance of these models is then assessed on the remaining portion of the data, commonly referred to as the validation or testing set. The aforementioned procedure is iterated numerous times, whereby each fold is utilized as the validation set in a sequential manner. The attainment of a more comprehensive evaluation of the model's efficacy can be accomplished through the computation of the mean value of the outcomes obtained from all the folds.

The integration of cross-validation techniques serves as a possible solution to tackle the plausible partiality that may arise from a train-test partition, thereby furnishing a more exhaustive assessment of the accuracy of the models. Furthermore, it enables the evaluating of the model's ability to generalize to new data.

3.4.2 Model Selection

This study employed a range of machine learning methodologies to detect and diagnose mental disorders through the analysis of EEG data. This study utilizes four ML algorithms, namely RF, LGBM, SVM, XGBRF, for the purpose of binary and multi-class classifications.

The process of model selection enables the evaluation and selection of optimal algorithms based on their efficacy and suitability for the given task. Each algorithm has advantages and characteristics that make it ideal for particular applications.

The study aims to tryout various modeling techniques and their effectiveness in detecting mental illnesses using EEG data. The achievement of this objective is facilitated through the utilization of multiple ML algorithms. This study attempts to identify the algorithms that exhibit the highest potential for accurate classification and prospective medical implementations.

3.4.3 Model Evaluation

The assessment of model performance is an essential component of model evaluation. This study involved an evaluation process that was centered around the confusion matrices and subsequent computation of performance metrics derived from those matrices.

The utilization of confusion matrices in classification scenarios is widely accepted for providing a concrete summary of the classification results. This is obtained by showing the number of true positives, true negatives, false positives, and false negatives for each class which enables a thorough evaluation of the classification performance. The utilization of matrices enables the calculation of diverse performance metrics, such as accuracy, precision, recall, and F1 Score.

For the confusion matrix for n = 2 where n is the total number of classes, the following definitions for each entry:

a is the number of correct negative predictions.

b is the number of incorrect positive predictions.

c is the number of incorrect negative predictions.

d is the number of correct positive predictions. [26]

This matrix can be used to determine the classification error and prediction accuracy as follows:

$$Accuracy = \frac{a + d}{a + b + c + d} \tag{7}$$

$$Error = \frac{b + c}{a + b + c + d} \tag{8}$$

These performance indicators offer insights into the models' capacity to properly categorize occurrences of various psychiatric diseases. They test the accuracy of the models, as well as their sensitivity to recognizing positive cases, their specificity in identifying negative instances, and their overall balanced performance.

The ability of the models to detect mental diseases using EEG data may be evaluated by doing an analysis of the performance metrics that are obtained from the confusion matrices. This assessment method gives significant information on the strengths and limits of the models, which helps in the selection of the most accurate and trustworthy algorithms for the work that is at hand.

3.4.4 Hyper-parameter Tuning

Hyper-parameters are essential components of machine learning models that require configuration to modify the model to the specific dataset. A more effective methodology involves conducting an impartial exploration of various values for model hyper-parameters and subsequently selecting a subset that yields optimal performance of the model on a designated dataset.

In our work we used Scikit-learn Python machine learning library. Using this library Randomize Search Cross Validation (RSCV) has been applied. It defines a search space as a bounded domain of hyper-parameter values and randomly sample points in that domain.

There is another one Grid Search Cross Validation (GSCV). This can be defined as a structured arrangement of hyper-parameter values in a grid format, wherein each position in the grid is systematically evaluated. The grid search method is a valuable tool for evaluating combinations that have a proven track record of high performance. The random search approach is advantageous for the purpose of exploration and identifying hyper-parameter configurations that may not have been intuitively anticipated, albeit its execution may typically demand a greater amount of time.

Chapter 4

Performance Metrics

In the realm of machine learning, performance measures are crucial instruments for determining the efficacy of categorization models as well as their level of precision. They give quantifiable measurements that indicate how well the models perform in diagnosing psychiatric conditions using EEG data. The present chapter is dedicated to the examination of performance measurements employed in the current study. These measurements include the Receiver Operating Characteristic (ROC) curve, Accuracy, Precision, and Recall Scores, as well as the F1 Score. By scrutinizing these metrics, one can acquire valuable insights into the efficacy of the models, their ability to effectively classify diverse categories, and their overall predictive proficiency.

The utilization of performance measures allows for the evaluation of how accurate a prediction model is. Accuracy is the metric that is utilized the most frequently in classification processes. However, for datasets that are not balanced, the accuracy results are skewed toward the dominant group, which might result in significant misinterpretation. [27].

When doing non-equilibrium classification tasks, class-specific counters make the process far more convenient. These ideas may be communicated more clearly thanks to the confusion matrix, which is displayed in Table II. There are two different outcomes that might occur due to the binary categorization problem. The answer is False (0) and True (1).

Table 4.1 Confusion Table format

	Predicted		
Actual	***Categories***	***False (0)***	***True (1)***
	False (0)	TN	FP
	True (1)	FN	TP

Table 4.2 Performance Parameter and Corresponding Equations

Parameter Name	**Equation**
Recall	$\frac{TP}{(TP+FN)}$ (9)
F1 Score	$\frac{(2 \times Precision \times Specificity)}{(precision + Specificity)}$ (10)
Accuracy	$\frac{(TN+TP)}{(TN+TP+FP+FN)}$ (11)
Precision	$\frac{TP}{TP+FP}$ (12)
AUC-ROC	$\frac{TP}{(TP\ +\ FN)} - \frac{FP}{(TN\ +\ FP)}$ (13)

Where,

TP = True Positive;

TN = True Negative;

FP = False Positive;

FN = False Negative

Accurate predictions for positive (1) class instances are denoted by the abbreviation TP, and accurate predictions for negative (0) class instances are denoted by the abbreviation TN. The prediction of negative instances as positive is shown by FP, whereas the prediction of positive cases as negative is shown by FN. In this investigation, unusual occurrences are assigned the number 1 so as to designate a distinct subgroup. Typical instances are counted as zero and make up the majority of the class. These four fundamental concepts serve as the basis for defining the major figures used in evaluation. A more vivid understanding of the confusion matrix format is presented in table 4.1.

4.1 Accuracy

The assessment of accuracy is a widely employed metric within the domain of machine learning, serving as a key performance indicator. The primary purpose of its application is to assess the overall efficacy of a classification model in terms of its accuracy. The present study provides a straightforward evaluation of the model's ability to accurately predict class labels. The present study examines the calculation of accuracy in classification tasks. Specifically, accuracy is operationalized as the ratio of correctly classified instances to the total number of examples in the dataset. The conventional representation of this concept involves the utilization of a numerical value expressed as a percentage, which is bounded by the values of 0% and 100%, and is contingent upon the specific contextual circumstances. A higher accuracy value is indicative of a greater number of accurate predictions generated by the model. The accurate identification of

psychiatric disorders from electroencephalogram (EEG) data is crucial for assessing the efficacy of models in accurately discriminating between healthy and unhealthy individuals.

The present study examines the advantages of utilizing accuracy as a metric for evaluating classification models. Specifically, the study highlights that accuracy is a simple and easily interpretable measure that provides a comprehensive assessment of the model's performance. The present study offers a concise and easily interpretable metric for assessing the degree of precision in predictive models, as represented by the proportion of correct forecasts. An evaluation of the model's overall predictive power is provided by accuracy. It measures how well the model can categorize instances accurately across all classes. Thus, Accuracy makes it possible to compare many models or iterations of the same model. When the classes in the dataset are balanced, or when there are an equal number of cases for each class, accuracy performs well.

However, the majority class of the dataset dominates a model's classification accuracy [28]. When working with records that aren't balanced, which means that one class has a lot more cases than the others, accuracy may not tell the whole story. When this happens, the model's accuracy can be skewed toward the class that makes up the majority. This can make it hard to tell how well the model is doing. Another thing to think about with accuracy is that it is based on the idea that the importance of all classes is the same. Nevertheless, incorrectly identifying cases as belonging to one category can have significantly more severe repercussions in certain contexts than in others.

4.2 Recall

The probability of correctly identifying a positive test sample or a minority test sample is indicated by recall [29]. Recall, sometimes referred to as sensitivity or true positive rate, is a performance indicator used to assess a classification model's capacity to find every genuine positive event in a dataset. It gives a numerical value to the percentage of accurately anticipated positive cases relative to the total number of instances that really are positive.

It is possible to calculate the percentage of accurate diagnoses by dividing the total number of correct diagnoses by the number of false negatives in addition to the total number of correct diagnoses. A greater recall value suggests a lower rate of false negatives and a stronger ability of the model to detect positive examples. It can be expressed as a value between 0 and 1, where a higher recall value signifies a lower rate of false negatives. A recall value of 1 indicates perfect recall, which indicates that all actual positive examples are properly detected by the model. This is denoted by the term "perfect recall."

The performance of a classification model can be better understood through the use of recall, which provides numerous significant insights. Recall emphasizes maximizing the detection of true positives in order to make sure that positive instances are not missed. It is crucial in situations like medical diagnosis or anomaly identification, where skipping over good evidence can have serious repercussions. Secondly, Recall measures how thorough or comprehensive a model's predictions are. It evaluates the model's capacity to identify all genuine positive instances, hence reducing the possibility of false negatives and assuring an improved understanding of the underlying data.

There is frequently a trade-off between recall and precision. Precision may decline when recall increases, and vice versa. Depending on the particular needs and priorities of the application, a balance between recall and precision can be achieved.

Some limitations of Recall are it ignores false positive and Class Imbalance Challenges. Recall does not consider false positives (instances incorrectly classified as positive). In some applications, false positives may have negative consequences. Therefore, it is necessary to consider recall in conjunction with other metrics, such as precision and F1 score, to obtain a comprehensive evaluation of the model's performance. Unbalanced datasets can also have an impact on recall. In these circumstances, getting good recall becomes difficult since the model can give the majority class priority. Reliable recall evaluation requires proper treatment of class imbalance.

4.3 Precision

The term "precision" refers to the proportion of discovered minority that corresponds to the correct answers. Due to the fact that it incorporates the findings of both major and small samples, the precision is dependent on the order in which the classes are presented [29]. Precision measures the percentage of positive cases that were accurately predicted out of all the instances that the model classified as positive. Precision is especially helpful in situations like medical diagnosis, fraud detection, or spam filtering when minimizing false positives is crucial.

Precision is determined by dividing the number of true positives (TP) by the total number of true positives and false positives (FP). A higher precision value suggests a lower rate of false positives and a higher accuracy in forecasting positive instances. It is given as a number between 0 and 1. All of the model's successful predictions have a precision value of 1, which denotes flawless precision.

By optimizing precision, the model ensures that positive predictions are reliable and trustworthy, minimizing the risk of incorrect decisions based on false positive results. In situations when the repercussions of false positives are severe, precision is especially crucial. For instance, in the identification of psychiatric disorders, misdiagnosing a healthy person as having a condition may result in unneeded interventions or treatments. High precision aids in preserving the integrity of judgments made using model predictions.

Some limitations of Precision are it ignores false negative and its sensitivity to Class Imbalance. Precision ignores false negatives (cases that are mistakenly categorized as positive) and only concentrates on false positives. False negatives can have serious repercussions in some applications, such the detection of life-threatening illnesses. Therefore, accuracy must be taken into account together with other measures like recall and F1 score. Unbalanced datasets, in which the proportion of positive examples is significantly lower than that of negative instances, can have an impact on precision. In these circumstances, even a little variation in false positive predictions can have a big influence on precision. Precision must be interpreted in light of class distribution.

4.4 F1- Score

F1-score uses the harmonic mean of the two metrics (precision and recall) of a classifier to combine them into a single statistic. The F1-score is used to assess classifications that have a high percentage of false positives and false negatives [30]. It offers a fair assessment of a classification model by taking both false positives and false negatives into account.

The F1 score is computed by utilizing the harmonic mean of precision and recall. The metric ranges from 0 to 1, where a larger value signifies superior model performance. The F1 score is determined by the following formula:

$$\frac{(2 \times Precision \times Specificity)}{(precision + Specificity)} \tag{14}$$

The F1 score offers several noteworthy revelations regarding the performance of the model. Initially, it is noteworthy that the F1 score offers an equitable assessment of the model's efficacy by amalgamating precision and recall. Attaining optimal balance involves considering the compromise between minimizing the occurrence of false positives, which is precision, and false negatives, which is recall. The F1 score is commonly used by practitioners to assess the overall effectiveness of a classification model. A model that exhibits high performance in making both positive and negative predictions is characterized by a superior F1 score. Professionals can utilize this approach to ascertain the optimal equilibrium between recall and precision for a given task.

The F score metric exhibits limitations in its sensitivity to class imbalance and its applicability to specific contexts. Imbalanced datasets have the potential to impact the F1 score. Under certain conditions, the F1 score may exhibit a bias towards the majority class, thereby requiring the implementation of supplementary methodologies such as data resampling or class weighting. The significance of the F1 score may vary depending on the specific application and its associated criteria. Under certain conditions, prioritizing precision or recall may supersede the comprehensive equilibrium that is encapsulated by the F1 score. A comprehensive comprehension of the specific context is necessary for the accurate assessment of the F1 score.

4.5 Receiver Operating Characteristics (ROC)

The Receiver Operating Characteristic (ROC) curve is an indispensable instrument for evaluating the efficacy of classification models. The present study examines the relationship between the true positive rate (TPR) and the false positive rate (FPR) as a means of evaluating the ability of models to differentiate between multiple classes. This analysis is critical for gaining insights into the trade-off between TPR and FPR and its implications for model performance.

The generation of the receiver operating characteristic (ROC) curve necessitates the plotting of the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. The True Positive Rate (TPR) is a fundamental metric in binary classification that quantifies the proportion of positive instances that are correctly classified by a given model. It is defined as the ratio of the number of true positive predictions to the total number of actual positive instances in the dataset. The TPR is a crucial performance measure that is widely used in various fields, including machine learning, statistics, and medical diagnosis, to evaluate the effectiveness of a binary classifier in correctly identifying positive cases. The metric commonly referred to as sensitivity or recall is a measure of the ability of a classification model to correctly identify positive instances. The false positive rate (FPR) is a metric that denotes the ratio of negative cases that are erroneously classified as positive.

The present study aims to investigate the impact of classification threshold adjustments on the model's performance in terms of positive example recognition and negative instance misidentification. This allows us to observe how the model performs. This trade-off is graphically represented by the ROC curve, in which the FPR is shown along the x-axis and the TPR is shown along the y-axis.

One of the major measures that can be produced with the use of the ROC curve is the area under the curve, which is abbreviated as AUC-ROC. The AUC-ROC value can take on any value between 0 and 1, with a greater value signifying improved classification accuracy. A score of 0.5 for the AUC-ROC metric suggests that the classifier could as well be making guesses at random; whilst a score of 1 for the AUC-ROC metric shows that it is faultless.

In the context of diagnosing psychiatric problems using EEG data, the ROC curve gives us the ability to evaluate the models' ability to discriminate between persons who are healthy and those who are not well. In addition to this, it sheds light on how well the models are able to diagnose particular psychiatric diseases. We are able to evaluate the sensitivity and specificity of the models at a variety of classification thresholds by doing an analysis of the ROC curve. This provides us with the ability to make informative choices regarding the classification threshold that is most appropriate. However, the ROC curve isn't the only performance indication that has to be considered; in order to get a comprehensive understanding of the advantages and disadvantages associated with the models, it is necessary to take into consideration other performance measures as well.

Chapter 5

Results and Discussion

5.1 Simulation Outcomes

Table 5.1.1: Confusion matrix for binary classification (5 fold)

Algorithm									
Predicted									
Actual	**Categories**	**RF**		**LGBM**		**SVM**		**XGBRF**	
		0	**1**	**0**	**1**	**0**	**1**	**0**	**1**
	0	172	0	170	0	172	0	171	0
	1	9	8	1	18	17	0	0	18

Table 5.1.2: Confusion matrix for binary classification (10 fold)

Algorithm									
Predicted									
Actual	**Categories**	**RF**		**LGBM**		**SVM**		**XGBRF**	
		0	**1**	**0**	**1**	**0**	**1**	**0**	**1**
	0	86	**0**	87	0	89	0	90	0
	1	4	4	0	8	6	0	0	5

Table 5.1.3: Comparison between 5 fold and 10 fold

Algorithm	5-fold	10-fold
RF	85.42	86.06
LGBM	85.15	85.33
SVM	87.52	88.21
XGBRF	70.53	69.93

Table 5.1.4: Confusion matrix for MCC (5 fold)

Actual	Categories	Algorithm																							
		Predicted																							
		RF						LGBM						SVM						XGBRF					
		0	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**
	0	49	1	4	0	2	1	49	0	6	1	1	0	44	2	6	2	1	0	39	3	0	1	2	2
	1	0	46	6	1	2	2	0	47	6	1	3	0	0	45	3	0	1	0	0	42	3	2	10	1
	2	5	7	28	5	6	7	0	8	29	3	8	10	2	19	15	1	6	5	2	14	24	5	8	6
	3	0	0	1	50	1	0	0	0	1	51	0	0	0	0	0	58	0	0	0	1	1	45	1	0
	4	0	0	1	0	42	1	0	1	4	1	38	0	1	4	3	0	52	0	0	4	2	1	41	9
	5	2	1	2	0	4	42	0	1	5	0	2	43	1	1	5	0	3	39	1	3	4	1	6	35

Table 5.1.5: Confusion matrix for MCC after SMOTE and SPCA (5 fold)

Actual	Categories	Algorithm																							
		Predicted																							
		RF						LGBM						SVM						XGBRF					
		0	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**
	0	39	4	8	0	0	1	36	0	8	1	2	1	37	2	4	2	2	0	41	1	1	2	2	0
	1	0	46	3	0	0	2	0	52	4	0	6	0	1	52	0	0	0	0	1	44	2	2	4	0
	2	6	7	20	3	8	1	2	2	35	0	4	4	6	7	26	1	3	9	3	11	17	4	10	7
	3	0	0	0	53	0	0	0	0	1	60	0	0	0	0	0	56	0	0	1	4	2	49	0	0
	4	0	3	6	0	51	1	0	0	3	0	45	1	0	0	4	0	40	0	0	5	1	2	35	1
	5	1	0	10	0	0	46	0	0	10	0	2	41	6	1	8	0	0	52	3	3	7	3	6	45

The above tables depict the confusion matrices for binary classification and MCC for main and specific disorders. The table 5.1.3 shows the comparison between 5-fold and 10-fold cross validation and it was observed that the change was not that significant when it was compared against the time and computing power utilized.

Table 5.1.6: Confusion Matrix for Specific Disorder using **SPCA** in case of **SVM**

		Predicted										
Actual	**Categories**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	0	51	0	0	0	0	0	0	0	0	0	0
	1	0	44	0	0	0	1	0	0	0	0	0
	2	0	1	44	0	0	3	0	0	0	1	1
	3	0	0	0	33	0	1	0	0	0	0	0
	4	0	0	0	0	28	1	0	0	0	0	0
	5	0	1	0	2	4	9	5	3	1	6	2
	6	0	0	0	0	0	0	43	0	0	0	0
	7	0	0	0	0	0	1	0	31	0	0	0
	8	0	0	1	0	0	1	0	0	32	0	0
	9	0	0	2	0	2	5	0	0	0	37	0
	10	0	0	0	0	0	0	0	0	0	0	41

Table 5.1.7: Confusion Matrix for Specific Disorder using **ANOVA** in case of **SVM**

		Predicted										
	Categories	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Actual	**0**	51	0	0	0	0	0	0	0	0	0	0
	1	0	43	0	0	0	2	0	0	0	0	0
	2	2	1	39	0	0	6	0	0	1	1	0
	3	0	0	1	33	0	0	0	0	0	0	0
	4	0	0	0	0	29	0	0	0	0	0	0
	5	2	1	2	2	2	8	2	4	2	5	3
	6	0	0	0	0	1	3	39	0	0	0	0
	7	0	0	0	0	1	0	0	31	0	0	0
	8	0	0	1	0	0	0	0	0	33	0	0
	9	0	0	2	1	3	6	0	1	0	32	1
	10	0	0	0	0	0	0	0	0	0	0	41

5.2 Results

Table 5.2.1: Binary Class Classification Performance Metrics

Performance Metrics					
Algorithm	**ROC**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
RF	0.66	93.12	0.93	0.93	0.91
LGBM	0.98	**99.47**	0.99	0.99	0.99
SVM	0.49	89.84	0.80	0.89	0.85
XGBRF	0.97	99.25	0.99	0.99	0.99

Table 5.2.2: Performance Metrics of Main Disorders Using SPARSE PCA

Performance Metrics					
Algorithm	**ROC**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
RF	0.94	78.01	0.78	0.78	0.77
LGBM	0.96	**80.69**	0.82	0.81	0.81
SVM	0.95	79.57	0.79	0.80	0.79
XGBRF	0.91	68.98	0.69	0.69	0.68

Table 5.2.3: Performance Metrics of Main Disorders Using ANOVA

Performance Metrics					
Algorithm	**ROC**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
RF	0.95	77.31	0.77	0.77	0.77
LGBM	0.96	**78.94**	0.80	0.79	0.79
SVM	0.**92**	77.01	0.76	0.77	0.76
XGBRF	0.91	68.98	0.70	0.79	0.68

Table 5.2.4: Performance Metrics of Specific Disorders Using SPARSE PCA

Performance Metrics					
Algorithm	**ROC**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
RF	0.97	85.42	0.83	0.85	0.83
LGBM	0.98	85.15	0.85	0.85	0.85
SVM	0.98	**87.52**	0.87	0.88	0.87
XGBRF	0.93	70.53	0.70	0.71	0.69

Table 5.2.5: Performance Metrics of Specific Disorders Using ANOVA

Performance Metrics					
Algorithm	**ROC**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**
RF	0.97	83.46	0.83	0.83	0.82
LGBM	0.98	84.10	0.84	0.84	0.84
SVM	0.97	**85.01**	0.84	0.85	0.84
XGBRF	0.93	69.07	0.69	0.69	0.68

In the above tables 5.2.1-5.2.5, different performance metrics has been tabulated for binary classification and multiclass classification by using feature extracting and feature selecting methods for both the main and specific disorders. It can be seen that in case of binary classification, LGBM gave the highest accuracy of 99.47%. Surprisingly, LGBM showed its capability when it comes to multiclass classifications of the main disorders. For main disorders, the highest performance was obtained from LGBM with an accuracy of 80.69% using SPCA. On the other hand, for specific disorders, SVM showed the highest accuracy of 87.52% when it used SPCA as well. Throughout the whole runs, SPCA provided better results compared to ANOVA.

5.3 Comparative Analysis

Table 5.3.1: Comparison with Relevant Works

Reference	Algorithm	Accuracy	Recall	F1-Score	ROC
[4]	RF	90.16	0.97	0.91	-
[10]	SVM	87.59	-	-	-
[31]	LDA	87.00	-	0.85	0.92
[10]	RF	86.16	-	-	-
This Study	SVM	87.52	0.88	0.87	0.98

When compared to the results of relevant works, the accuracy of this study is highly comparable and competitive. Even though the referenced studies were not completely in line with the principles of this study, these were the closest we could find and the metrics are compared above in table 5.3.1.

5.4 Feature Importance

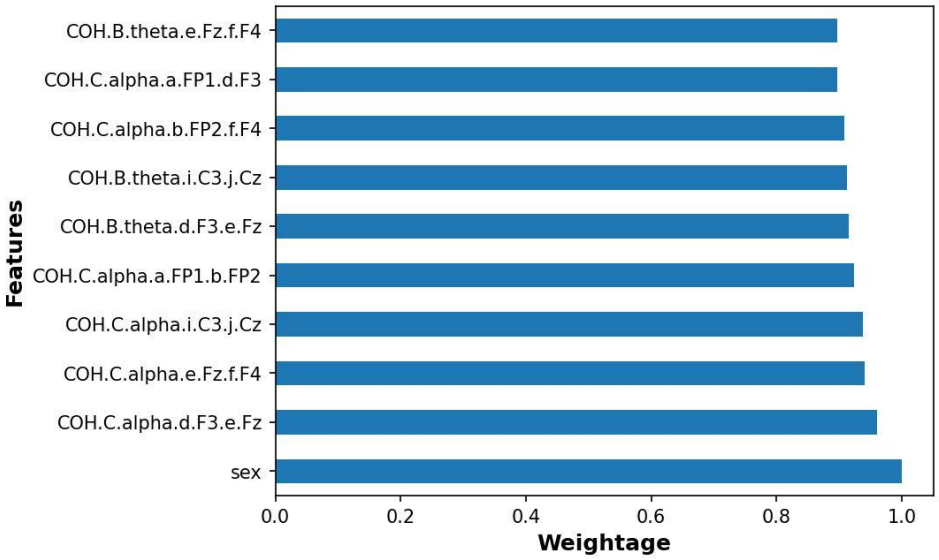


Fig 5.4.1: Important Features for Main Disorders

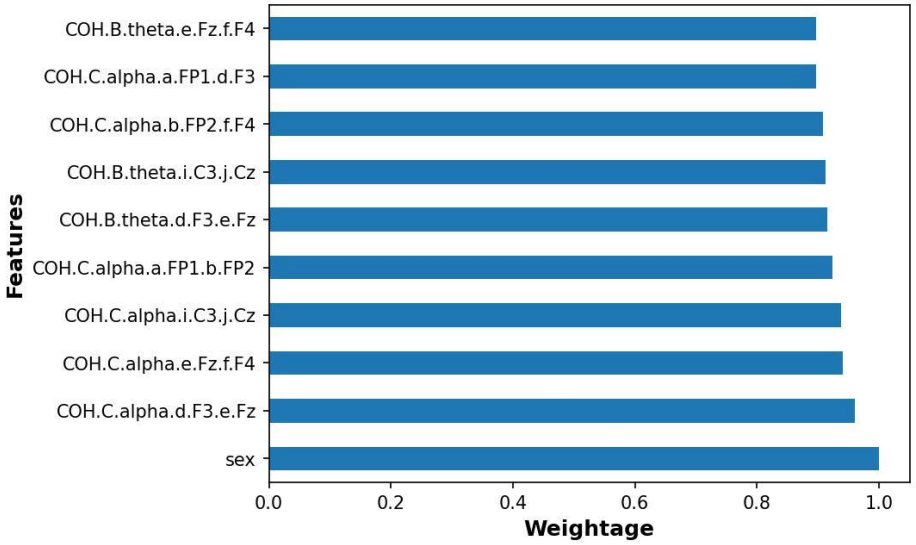


Fig 5.4.2: Important Features for Specific Disorders

In the chosen dataset, the total number of features is 1145. Of them, by using feature selection, 100 features were selected and the feature that was carrying the highest significance or weight was gender/ sex. It was observed that compared to the other features (received brain wave signals), sex had a higher determining factor when it came to classifying the entire subjects' list.

Chapter 6

Potential Future Development

This study has come up with a lot of conclusive results and directions. Rather than closing the chapter, it has opened the path for the future so that the concept can be implemented for even more advanced and prompt detection of psychiatric disorders. The overall blueprint for the future can be observed from fig 6.1 as we move forward.

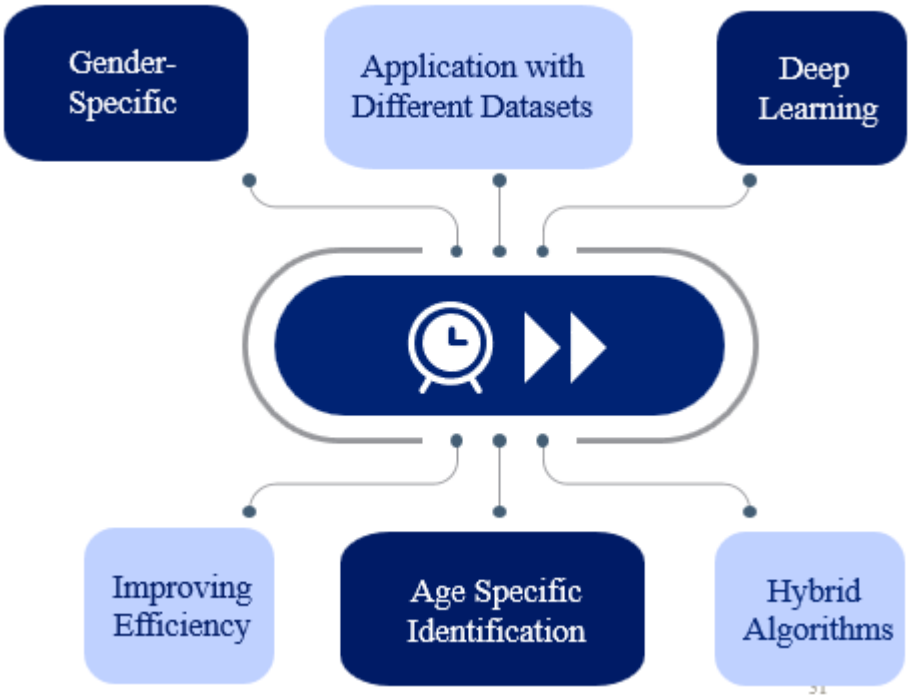


Fig 6.1: Overview of Future Development

6.1 Gender Specific Classification

The feature importance analysis reveals that gender holds greater significance; thus, future research could concentrate on establishing classification models that are specific to gender. We can identify distinctive patterns and traits that may affect the development and diagnosis of psychiatric disorders by taking gender into account as a differentiating factor. The implementation of a customized approach has the potential to result in enhanced precision and personalized medical care.

6.2 Application with Different Datasets or Merging Multiple Datasets

The expansion of the application of proposed models and techniques to diverse datasets is an essential area for future research. This can involve analyzing EEG data from diverse populations, cultural backgrounds, or specific psychiatric disorders. Moreover, the integration of multiple datasets can yield a broader and more standard sample, resulting in reliable and useful findings.

6.3 Implementation of Deep Learning

The utilization of hybrid algorithms has been found to improve performance and robustness in machine learning. This is achieved through the combination of multiple algorithms. Subsequent studies may investigate the utilization of hybrid algorithms that incorporate the advantageous features of diverse methodologies, such as ensemble techniques or hybrid deep learning architectures.

6.4 Age-Specific Classification

The categorization of psychiatric disorders based on age is a critical factor to consider as it influences their symptoms and identification. Further study endeavors may concentrate on developing classification models that are specific to different age groups, to accommodate the distinct attributes and developmental modifications that are present within each group. The precision and applicability of our findings in diagnosing and treating psychiatric disorders can be enhanced by customizing classification algorithms to different age groups.

6.5 Efficiency Improvement

Although our research has yielded encouraging outcomes, there exists potential for further optimization of the suggested algorithms to improve their efficiency. Further study endeavors may investigate methodologies aimed at mitigating computational complexity, augmenting real-time processing capabilities, and enhancing the scalability of the models. The proposed approach aims to facilitate the integration of machine learning-based classification of psychiatric disorders into clinical practice, thereby enabling a seamless and efficient adoption of this technology.

6.6 Implementation of More Hybrid Algorithms

The utilization of hybrid algorithms has been found to improve performance and robustness in machine learning. This is achieved through the combination of multiple algorithms. Subsequent studies may investigate the utilization of hybrid algorithms that incorporate the advantageous features of diverse methodologies, such as ensemble techniques or hybrid deep learning architectures.

Chapter 7

Conclusion

Analysis of Psychiatric Disorders from EEG Signals Using Machine Learning Techniques holds immense potential in revolutionizing the field of mental health diagnosis and treatment. Through the utilization of machine learning, we have conducted an evaluation, and enhancement of the efficacy of diverse ML algorithms to achieve precise identification of anomalies in EEG-based psychiatric disorder classification.

The significance of EEG signals in comprehending the intricacies of psychiatric disorders cannot be overemphasized. Analogous to the way industrial control systems acquire data from sensors to ensure coherence within an industrial setting, EEG signals constitute a valuable data source for capturing neural activity and cognitive processes. This study aimed to measure the effectiveness of machine learning algorithms in a unique context characterized by the absence of temporal similarities, which proved to be a challenge for the detection of abnormalities.

This research approach undertook an investigation into the efficacy of binary and multi-class classification techniques for psychiatric disorders. The study's findings provided a comprehensive understanding of the distinctive characteristics of these methods and the challenges that must be overcome to achieve accurate diagnoses. This research employed various machine learning algorithms, namely SVM, RF, LGBM, and XGBRF, to accurately classify both general and specific PDs. The present study reports on the findings of a series of experiments and performance evaluations aimed at evaluating the efficacy of ML techniques in improving the overall diagnostic process. Our results indicate that these techniques hold significant promise in terms of enhancing classification accuracy, thereby suggesting their potential utility in clinical settings. The present study demonstrates the efficacy of employing sophisticated feature engineering techniques, namely SMOTE and SPCA, in addressing the issue of class imbalance and dimensionality reduction respectively. A higher performance was achieved by using these unique algorithms. The implications of the research possess a wide-ranging purview. The identification of aberrations by

medical professionals and scholars can result in vital intervention, providing improved diagnosis, fine-tuned therapeutic approaches, and ultimately advanced patient prognoses.

The study comes with its inherent limitations. The research highlights the value of high quality EEG data and emphasizes the necessity for additional investigation into the adaptability of ML models within the domain of PDs. The outcome containing all the demographics in a single bubble should be of utmost importance.

References

[1] R. Das, M. R. Hasan, S. Daria, and M. R. Islam, “Impact of COVID-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study,” BMJ Open, vol. 11, no. 4, p. e045727, 2021.

[2] A. Al-Ezzi, N. Kamel, I. Faye, and E. Gunaseli, “Analysis of default mode network in social anxiety disorder: EEG resting-state effective connectivity study,” Sensors (Basel), vol. 21, no. 12, p. 4098, 2021.

[3] M. J. Rivera, M. A. Teruel, A. Maté, and J. Trujillo, “Diagnosis and prognosis of mental disorders by means of EEG and deep learning: a systematic mapping study,” Artif. Intell. Rev., vol. 55, no. 2, pp. 1209–1251, 2022.

[4] Z. Shen et al., “Aberrated multidimensional EEG characteristics in patients with generalized anxiety disorder: A machine-learning based analysis framework,” Sensors (Basel), vol. 22, no. 14, p. 5420, 2022.

[5] N. Langer et al., “A benchmark for prediction of psychiatric multimorbidity from resting EEG data in a large pediatric sample,” Neuroimage, vol. 258, no. 119348, p. 119348, 2022.

[6] D. Watts, H. Moulden, M. Mamak, C. Upfold, G. Chaimowitz, and F. Kapczinski, “Predicting offenses among individuals with psychiatric disorders - A machine learning approach,” J. Psychiatr. Res., vol. 138, pp. 146–154, 2021.

[7] A. Khodayari-Rostamabad, J. P. Reilly, G. Hasey, H. Debruin, and D. Maccrimmon, “Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model,” Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., vol. 2010, pp. 4006–4009, 2010.

[8] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.

[9] Z. Ghahramani, G. Rey, and E. Hinton, “The EM algorithm for mixtures of factor analyzers,” Psu.edu, 1996. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=766f4465747394d304d162197e091f1ae8f7f577. [Accessed: 04-Jun-2023].

[10] S. M. Park et al., “Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach,” Front. Psychiatry, vol. 12, p. 707581, 2021.

[11] A. A. Rahman et al., “Detection of mental state from EEG signal data: An investigation with machine learning classifiers,” in 2022 14th International Conference on Knowledge and Smart Technology (KST), 2022.

[12] C. Ranjith and B. Arunkumar, “An improved Elman neural network based stress detection from EEG signals and reduction of stress using music,” Ripublication.com. [Online]. Available: http://www.ripublication.com/irph/ijert19/ijertv12n1_03.pdf. [Accessed: 04-Jun-2023].

[13] J. Ashford, J. J. Bird, F. Campelo, and D. R. Faria, “Classification of EEG signals based on image representation of statistical features,” in Advances in Intelligent Systems and Computing, Cham: Springer International Publishing, 2020, pp. 449–460.

[14] Z. Iscan *et al.*, “Test-retest reliability of freesurfer measurements within and between sites: Effects of visual approval process: Test-Retest Reliability of FreeSurfer Measurements,” *Hum. Brain Mapp.*, vol. 36, no. 9, pp. 3472–3485, 2015

[15] L. Breiman, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001

[16] S. Bhat, “Detection of polycystic ovary syndrome using machine learning algorithms,” Diss. Dublin, 2021.

[17] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[18] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer, 2009.

[19] K. R. Bhatele and S. S. Bhadauria, “Glioma segmentation and classification system based on proposed texture features extraction method and hybrid ensemble learning,” Trait. Du Signal, vol. 37, no. 6, pp. 989–1001, 2020.

[20] H. Zhang, L. Wan, and Y. Li, “Prediction of soil organic carbon content using sentinel-1/2 and machine learning algorithms in swamp wetlands in northeast China,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., pp. 1–12, 2023.

[21] B. Deepa and K. Ramesh, “Epileptic seizure detection using deep learning through min max scaler normalization,” Int. J. Health Sci. (IJHS), pp. 10981–10996, 2022.

[22] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, “SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary,” J. Artif. Intell. Res., vol. 61, pp. 863–905, 2018.

[23] R. Drikvandi and O. Lawal, “Sparse principal component analysis for natural language processing,” Ann. Data Sci., vol. 10, no. 1, pp. 25–41, 2023.

[24] H. Zou and L. Xue, “A selective overview of sparse principal component analysis,” Proc. IEEE Inst. Electr. Electron. Eng., vol. 106, no. 8, pp. 1311–1320, 2018.

[25] R. R. Wilcox, “Understanding the practical advantages of modern ANOVA methods,” J. Clin. Child Adolesc. Psychol., vol. 31, no. 3, pp. 399–412, 2002.

[26] S. Visa, B. Ramsay, A. Ralescu, and E. VanDerKnaap, “ Confusion matrix-based feature selection,” p. 120, 2011.

[27] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 1, p. 6, 2020.

[28] U. Bhowan, M. Johnston, and M. Zhang, “Evolving ensembles in multi-objective genetic programming for classification with unbalanced data,” in Proceedings of the 13th annual conference on Genetic and evolutionary computation, 2011.

[29] A. P. Bradley, R. P. W. Duin, P. Paclik, and T. C. W. Landgrebe, “Precision-recall operating characteristic (P-ROC) curves in imprecise environments,” in 18th International Conference on Pattern Recognition, 2006.

[30] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, “Thresholding classifiers to maximize F1 score,” arXiv [stat.ML], 2014.

[31] A. Anuragi and D. Singh Sisodia, “Alcohol use disorder detection using EEG Signal features and flexible analytical wavelet transform,” Biomed. Signal Process. Control, vol. 52, pp. 384–393, 2019.