

A Comprehensive Investigation into Detecting Schizophrenia from EEG Signals Using a Machine Learning Approach

by

A.Abdur Rahman Akib – 180021122

S M Mehedi Zaman – 180021128

Fabiha Farzana – 180021202

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)

The Organization of Islamic Cooperation (OIC)
Board Bazar, Gazipur-1704, Bangladesh

June 2023

CERTIFICATE OF APPROVAL

The thesis titled “Machine Learning Approach to Detect Schizophrenia from EEG Signal” submitted by A.Abdur Rahman Akib (180021122), S M Mehedi Zaman (180021128), and Fabiha Farzana (180021202) has been found as satisfactory and accepted as partial fulfillment of the requirement for the degree of Bachelor of Science in Electrical and Electronic Engineering on 5th June, 2023.

Approved by:



(Signature of the Supervisor)

Mirza Muntasir Nishat

Assistant Professor

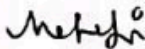
Department of Electrical and Electronic Engineering
Islamic University of Technology

DECLARATION OF CANDIDATES

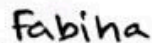
It is hereby declared that this thesis or any part of it has not been submitted elsewhere for award of any degree or diploma.



(Signature of the Candidate)
A. Abdur Rahman Akib
Student ID: 180021122



(Signature of the Candidate)
S M Mehedi Zaman
Student ID: 180021128



(Signature of the Candidate)
Fabiha Farzana
Student ID: 180021202

DEDICATION

We would like to dedicate this thesis to our family members and everyone who have given us unwearied support throughout the entirety of our existence and every situation of our life. They have always been a source of motivation for us. They pushed us ahead and showed us how to make the correct decisions. They never fail to inspire us to work hard and move forward to overcome life's difficulties. They have provided us with the protection, wisdom, and fortitude we need to face difficult situations.

ACKNOWLEDGEMENTS

First, we would want to express our heartfelt gratitude to Almighty Allah, our Creator, for creating and instilling in us the intellect to educate ourselves with worldly knowledge and, therefore, complete our thesis research. **Mr. Mirza Muntasir Nishat**, Assistant Professor, Department of EEE, IUT, is our respected supervisor. We owe him a debt of gratitude for his continuous advice, care, and support in our pursuit of a career in electrical and electronic engineering. He has always encouraged us to learn new things and broaden our horizons to keep our minds sharp. We would not be exploring the power electronics area if it were not for his motivation. He has encouraged us to learn the fundamentals of the specific field and shown us how to proceed in the right direction

We would also like to express our gratitude to **Prof. Dr. Mohammad Rakibul Islam**, Head, Department of EEE, and all the faculty members of the EEE Department, IUT, for their unwavering support, encouragement, and assistance.

Finally, we owe a debt of gratitude to our family for encouraging and assisting us in overcoming life's challenges, as well as enchanting us with their wonderful words. Last but not least, we would like to express our gratitude to our friends for their unconditional support and for keeping our spirits upbeat throughout this journey.

ABSTRACT

Schizophrenia is a prevalent psychiatric condition that places significant clinical demands on both patients and their caregivers. An accurate and expeditious diagnosis is essential for the effective treatment of schizophrenia. In this regard, the identification of classification biomarkers has the potential to enhance comprehension of the neural underpinnings of schizophrenia and supplement clinical assessments. Recent years have seen an increase in research into the diagnostic and prognostic utility of Machine Learning (ML) techniques in schizophrenia. Several such studies have attempted to classify individuals with schizophrenia from healthy controls using neuroanatomical features. However, the range of neuroanatomical measures utilized in these investigations has been limited thus far. The objective of this study was to detect schizophrenia at an early stage using the largest EEG signal dataset to date, consisting of 193 patients. To compile this dataset, the three largest open-source EEG signal datasets were merged and processed. For the most accurate detection of schizophrenia from EEG signals, an ML array was utilized. With the Gradient Boosting Classifier (GBC) method, feature engineering, and model tuning, this research achieved one of the highest classification accuracies to date, 93.3%, among the other supervised ML models used in the study. In addition, the study's results demonstrated that precision, recall, and f1 score were, respectively, 84.6%, 80%, and 82%. The obtained results from this thesis surpasses all previous works using EEG signal in terms of accuracy and number of subjects considered and the results were obtained only using supervised model which is computationally lighter than typical signal analyzing Convolutional Neural Network (CNN) models. This thesis concentrates on the robustness and significance of larger EEG signal datasets, as contemporary studies have implemented prediction strategies on relatively smaller datasets.

TABLE OF CONTENTS

	Page No.
Certificate of Approval	
Declaration of Candidates	
Dedication	
Acknowledgements	
Abstract	
Table of contents	
List of Tables	
List of Figures	
List of Acronyms	
CHAPTER – 1: INTRODUCTION	
CHAPTER – 2: LITERATURE REVIEW	
CHAPTER – 3: RESEARCH METHODOLOGY	
3.1 Dataset	
3.2 Data Pre-Processing	
3.3 Feature Engineering	
3.4 Machine Learning Models	
3.5 Tuning Models	
3.6 Prediction Metrics	
CHAPTER – 4: PREDICTIVE RESULTS	
CHAPTER – 5: DISCUSSION	
CHAPTER – 6: CONCLUSION	
REFERENCES	

LIST OF TABLES

No.	Title	Page No.
2.1	Previous works on EEG Signal Dataset	
2.2	Previous works on Structural MRI Signal Dataset	
2.3	Previous works on fMRI Signal Dataset	
3.1	Dataset Description	
4.1	Prediction Metrics of Base Models	
4.2	Optimal Hyperparameters after Optimization with Optuna	
4.3	Results after Hyperparameter Tuning	
4.4	Comparison among contemporary studies	

LIST OF FIGURES

No.	Title	Page No.
1.1	64 EEG Electrodes layout	
3.1	Overall workflow of the study	
3.2	Data Pre-processing steps	
3.3	Distribution of the output 'Group' for Schizophrenia patients	
3.4	Johnson SU Distribution of the output 'Group' for Schizophrenia patients	
3.5	Normal distribution of the output 'Group' for Schizophrenia patients	
3.6	Log distribution of the output 'Group' for Schizophrenia patients	
3.7	Skewness of the output 'Group' for Schizophrenia patients	
3.8	Kurtosis of the output 'Group' for Schizophrenia patients	
3.9	Correlation among the features and output 'Group' for Schizophrenia patients	
3.10	MATLAB code snippet for finding out the six statistical features	

4.1	Results after Hyperparameter Optimization	
4.2	Comparison of Results with Contemporary Studies	

LIST OF ACRONYMS

Abbreviated Form	Description
PTSD	Post-Traumatic Stress Disorder
ADHD	Attention Deficit Hyperactivity Disorder
EEG	Electroencephalography
LFPs	Local Field Potentials
ERP	Event-Related Potential
SVM	Support Vector Machines
RF	Random Forests
ANN	Artificial Neural Networks
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
MRI	Magnetic Resonance Imaging
fMRI	Functional Magnetic Resonance Imaging
AI	Artificial Intelligence
ML	Machine Learning
SVC	Support Vector Machine Classifier
SZ	Schizophrenia
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
KNN	k-Nearest Neighbors
MVPA	Multi-Voxel Pattern Analysis
MKL	Multiple Kernel Learning
ROC	Receiver Operating Characteristic
K-fold	k-Fold Cross-Validation
L2	L2-Regularized

Chapter 1

Introduction

The concept of mental illness incorporates several distinct aspects into its overall definition. The term “diagnosis” refers to the classification of a patient’s condition based on a set of distinctive symptoms, each of which corresponds to a specific disease or malfunction. In a perfect world, these diagnoses would be based on data visible from the patient’s body, neurobiology, or genetics. However, we do not live in a perfect world. On the other hand, the underlying biological causes of many mental disorders are not yet fully understood. Mental disorders with unknown causes are categorized according to the patterns of symptoms that manifest themselves because these patterns have a greater likelihood of co-occurring and progressing in a comparable manner. Behind the specific concepts of disease and dysfunction lies a more general concept, and this concept is of vital importance to our daily lives. This notion lies dormant beneath the particular notions of illness and dysfunction. This concept serves as the foundation for our understanding of what it is like to be ill, regardless of the specific diagnosis that may or may not be given. In the course of our daily lives, we frequently encounter situations in which we exhibit the signs and symptoms of a medical condition, but the vast majority of the time, we are unable to identify the underlying cause. Pain, fever, physical weakness, irritability, feeling depressed, disturbed sleep, and other forms of cognitive impairment are among the frequent symptoms that may be indicative of a mental disorder and that can be detected. There is a likelihood that certain diseases have an early stage known as the prodromal stage, which manifests before a sufficient number of typical symptoms appear to warrant a formal diagnosis. If so, this is referred to as the prodromal stage. The progression of diseases over time is one characteristic that distinguishes them from one another. This trend could result in a decline, an improvement, a recovery, or even a fatal conclusion. Each of these outcomes is conceivable. Syndromes that have been observed for an extended period of time without improvement are no longer considered mental disorders. Alternatively, they may be referred to as personality disorders or disabilities (such as mental impairments), given that they primarily impact cognitive ability [1].

Mental health issues are extremely prevalent on a global scale. Approximately one in eight people worldwide suffers from mental illness. Depending on a person’s age and gender, they may be more susceptible to developing a particular mental disorder [2]. In 2019, more than 970 million people worldwide, or one out of every eight people, suffered from a mental illness. It has been demonstrated that anxiety and depression are the most prevalent forms of mental illness [3]. The COVID-19 pandemic resulted in a significant increase in the incidence of anxiety and depressive disorders, with a notable increase observed in 2020. An increase in the number of people with these conditions was what caused this surge. In just one year, there was a 26% increase in the number of people diagnosed with anxiety disorders and a 28% increase in the number of people diagnosed with severe depressive disorders, according to some preliminary estimates [4].

A person is diagnosed with a mental illness if they exhibit significant cognitive, emotional, or behavioral disturbances on a clinical level. The majority of the time, these disturbances cause discomfort or impairment in crucial aspects of functioning. There are numerous classifications that can be used to describe the spectrum of conditions that are collectively referred to as mental illnesses. Occasionally, these conditions are also referred to as mental health disorders. The latter term encompasses mental disorders, psychosocial impairments, and other mental states associated with extreme discomfort, functional impairment, or the likelihood of self-harm [5].

Anxiety disorders, depression, and bipolar disorder, to name a few, are prevalent forms of mental illness in contemporary society. The hallmarks of anxiety disorders are excessive fear, worry, and abnormal behavior, whereas the hallmarks of depression are persistent melancholy and a diminished interest in activities that one typically enjoys when things are normal. Depression may also be characterized by a loss of interest in activities that a person enjoys doing under normal circumstances. Throughout their condition, a person with bipolar disorder will cycle between depressive and manic symptoms. Traumatic events are the cause of post-traumatic stress disorder, also known as PTSD, whereas perceptual and behavioral deficits are the hallmarks of schizophrenia. Disorders of disruptive behavior include conduct disorder and oppositional defiant disorder. Eating disorders focus on abnormal patterns of eating. Neurodevelopmental disorders encompass a wide range of conditions, such as intellectual development issues, autism spectrum disorders, and attention deficit hyperactivity disorder (ADHD), to name a few examples. These issues relating to mental health are treatable, and effective treatments are available. These treatments include psychological interventions, medication, and individualized therapy tailored to the specific needs of each individual patient [6].

Schizophrenia, one of these severe mental disorders, is the focus of our research. Schizophrenia is a complex mental disorder with diverse causes that manifests during the early stages of brain development. Schizophrenia is characterized by subtle pathogenic alterations in certain populations of neural cells and in cell-to-cell communication, whereas massive brain disease is not. Schizophrenia is a mental disorder that affects thought and behavior and is related to how the brain processes information. Neuroimaging studies have revealed that individuals with both first-episode and persistent schizophrenia exhibit abnormalities in information processing [7].

People with schizophrenia frequently struggle in academic and social settings. Together, psychotic symptoms, such as hallucinations, delusions, and disorganization, as well as motivational and cognitive deficits, comprise the schizophrenia expression. Patients may have experienced similar sensations; however, the veracity of this phenomenon is currently uncertain, and it may or may not exist. Positive, negative, and cognitive symptoms are the three categories into which the most prominent symptoms of schizophrenia fall. Positive symptoms are those that a doctor can spot and are absent in healthy people. The intensity of hallucinations, delusions, and other forms of aberrant motor behavior may vary. The morbidity rate is high, and it is difficult to identify negative symptoms. The most frequently reported negative symptoms were avolition, alogia, anhedonia, and less emotional expressiveness. “Cognitive Symptoms” is the most recent diagnostic category. These eventually hinder the person’s ability to communicate by making it difficult for him to speak clearly and pay attention [8, 9].

Symptoms typically manifest in early adulthood and persist indefinitely. Typically, schizophrenia manifests in men in their early to mid-20s. In women, the onset of symptoms typically occurs in their thirties. People younger than 25 have a significantly lower incidence of schizophrenia than those older than 45 [10], and those younger than 25 have an even lower incidence than those older than 45. Schizophrenia affects approximately 24 million people worldwide, or approximately 0.32% of the total population. This disorder affects approximately 0.45% of adults, or roughly one in every 222 individuals [11].

According to [12], the prevalent explanation for the causes of schizophrenia is strikingly similar to the explanation for cancer. It is believed that a combination of non-genetic environmental factors and genetic traits passed down from one's parents causes schizophrenia. These factors are known as "hits," and it is believed that schizophrenia is caused by a number of hits. These external factors may either directly harm the brain or influence the regulation and expression of the genes responsible for brain function. It is possible that some individuals have a genetic predisposition to the condition, but in the majority of cases, the symptoms of schizophrenia may not manifest until a convergence of environmental and psychological factors. This convergence causes irregularities in the brain's growth and maturation, which is a continuous process during the first two decades of a person's life.

In brain circuits and neurotransmitter systems, dispersed anomalies are more prevalent than localized ones. When there is a breakdown in connection and communication within brain circuits, patients experience a wide array of symptoms and cognitive deficits. Despite its diversity, the disease is characterized by a single pathogenesis. The dysfunctional control of cerebral information processing is what causes schizophrenia [13].

Patients with schizophrenia have a 10- to 13-fold higher lifetime risk of suicide than the general population, which is estimated at around 1%. Although schizophrenic people are more likely to commit suicide at any point in their lives, the risk is greatest before middle age and decreases afterward [14].

Even though the vast majority of schizophrenics are not violent, the risk of violence rises when the disorder is untreated [15]. In a 2006 study, 19.1% of schizophrenia patients were found to engage in any form of violence, and 3.6% of participants reported engaging in serious violent behavior within the previous six months [16].

Schizophrenia is notoriously difficult to diagnose based solely on observation or symptom listening, for a variety of reasons. Individuals with schizophrenia may exhibit a diverse array of symptoms. These signs and symptoms may or may not be readily apparent to others, and they may manifest differently between individuals. Due to the absence of a universal set of behaviors that can unambiguously indicate the presence of schizophrenia, it is difficult to diagnose the condition based solely on visual observation of individuals with the disease. Some of the most prominent symptoms of schizophrenia, such as auditory hallucinations and delusions, are subjective mental states that are difficult for outside observers to detect. Due to their origin in the patient's internal

experience and cognition, these symptoms may be difficult to detect at first glance. People with schizophrenia may also attempt to conceal their symptoms or use alternative coping mechanisms to appear “normal” to others. This can make it more difficult for those who rely solely on visual cues to identify the presence of the disease. It is possible that the negative symptoms of schizophrenia, such as avolition (lack of motivation), alogia (limited speech), anhedonia (inability to perceive pleasure), and impaired emotional expression, are more subtle and difficult to detect. Negative symptoms are more difficult to detect than positive ones because they involve the absence or deterioration of normal behavior. Schizophrenia is characterized by fluctuating symptom severity over time, with possible remission periods. Due to the fact that a person with this disorder may appear “normal” at times, it may be difficult to detect the disorder solely through visual inspection.

In the early stages of the disease, it is possible that the symptoms of schizophrenia will not be readily apparent, making it difficult to diagnose through external observation. However, it is of the utmost importance that those exhibiting symptoms of schizophrenia receive treatment as soon as possible. The majority of these symptoms manifest early in a person’s life and can become progressively more severe over decades. A prompt and accurate diagnosis of schizophrenia paves the way for individualized treatment that can slow or halt the progression of the disease. However, in order to achieve this objective, psychiatrists must frequently make individualized diagnoses for their patients. If the early symptoms are misdiagnosed or the initial medication is ineffective, it can result in the loss of an ideal window for disease control and treatment, which could cause the patient to experience undesirable side effects [17]. In addition, studies indicate a substantial increase in healthcare expenditures by 2050, which would represent a sizable portion of the GDP. These results are based on previous statistical evaluations. Communities and governments may therefore feel compelled to implement certain measures, such as reducing the number of patients treated in intensive care units (ICUs) [18]. These conditions are probably going to have an impact on treatments that are not only expensive but also time-consuming. Due to the limitations of using eye examination to diagnose schizophrenia, it is becoming increasingly important to automate a variety of intensive care unit (ICU) operations, such as the use of differential diagnostic algorithms to identify different disorders [19].

An electroencephalogram, also known as an EEG, is a representation of the electrical impulses generated by the coordinated activity of brain cells. More specifically, it depicts the temporal patterns of extracellular field potentials that are generated when these cells collaborate. Electrodes can be placed on the scalp or directly on the surface of the brain to record an EEG or an electrocorticogram (ECoG), which is another name for an EEG. Figure 1 shows the location of electrodes on the scalp during EEG.

In the study [21], the researchers analyzed the resting-state EEGs of 121 schizophrenia-diagnosed patients and 75 control participants. To obtain all 194 EEG characteristics, numerous signal processing techniques were utilized. The fact that 69 out of 194 characteristics demonstrated a statistically significant difference between patients and controls demonstrates that these characteristics are capable of detecting key components of schizophrenia.

Due to the fact that individuals with schizophrenia exhibit distinct readings in a variety of aspects, there has been a great deal of research conducted over the past few years to diagnose schizophrenia using an EEG-based Machine Learning (ML) method. This is due to the fact that patients with schizophrenia exhibit different readings. ML ensures that the detection is both automated and cost-effective. In addition, this permits the detection of the earliest stages of schizophrenia with a high degree of accuracy and precision.

Numerous studies have previously investigated the application of ML to the diagnosis of schizophrenia; however, the available datasets have been restricted to a maximum of 84 test subjects. In this paper, we aim to overcome this limitation by creating a larger dataset of 193 test subjects by combining existing EEG datasets of schizophrenia patients and healthy individuals. Using a larger dataset increases confidence in the accuracy of detection and validation of previously published works that employ ML to detect schizophrenia. The findings of this study make a significant contribution to ongoing efforts to detect schizophrenia using ML.

Chapter 2

Literature Review

Schizophrenia is a severe mental disorder that impairs normal thinking, feeling, and behavior. For effective diagnosis and treatment of schizophrenia, a prompt and accurate diagnosis is essential. There has been a recent increase in research into the applicability of machine learning techniques to the diagnosis and prognosis of mental health disorders such as schizophrenia. This study seeks to provide a summary of the existing literature regarding the use of machine learning (ML) algorithms for the diagnosis of schizophrenia. This analysis will discuss the difficulties, new insights, and promising directions for future research. The primary objective of this study is to develop reliable methods for diagnosing schizophrenia. ML techniques, such as feature selection and extraction, classification algorithms, and the integration of multiple data modalities, have produced positive results in a number of areas of schizophrenia research. How well machine learning models select and extract features determines their performance.

Multiple techniques have been applied to neuroimaging (structural and functional MRI), electroencephalography (EEG), and behavioral assessments [22] in order to identify crucial characteristics. To differentiate between individuals with and without a schizophrenia diagnosis, extracted features are classified using classification techniques. Several machine learning (ML) techniques, including support vector machines (SVM), random forests (RF), artificial neural networks (ANN), and deep learning models (CNN, RNN), have been utilized in this field [23]. Using the retrieved information, algorithms construct classification models that accurately classify individuals as healthy or schizophrenic.

In academic studies aiming at identifying schizophrenia, neuroimaging techniques have garnered considerable interest. In a variety of studies, structural MRI has been used to learn about the brain's anatomy. ML algorithms applied to structural MRI data have yielded promising results [24] in detecting brain abnormalities associated with schizophrenia and distinguishing healthy participants from sufferers. Scientists have used functional magnetic resonance imaging (fMRI) to study schizophrenia in order to measure brain activity and connectivity. Using fMRI data and ML techniques, it has been demonstrated that it is possible to distinguish between individuals with schizophrenia and healthy people. Researchers must first isolate individual patterns of neural activity and connectivity in the brain before achieving this goal. EEG has been demonstrated to aid in the diagnosis of schizophrenia. The electroencephalogram (EEG) is a highly accurate instrument for monitoring brain activity; it can detect the erratic patterns frequently observed in schizophrenia patients. By evaluating unique properties such as event-related potentials (ERPs) or spectral power [25-27], ML algorithms applied to EEG data have demonstrated promising results in differentiating between individuals with a schizophrenia diagnosis and healthy individuals. An EEG analysis yielded the aforementioned results. Furthermore, it has been demonstrated that multimodal techniques, which combine data from multiple modalities such as neuroimaging and EEG, can improve the diagnostic accuracy of schizophrenia. It has been demonstrated that using ML techniques on multimodal data substantially improves performance compared to using a single

modality alone. The aforementioned approaches provide insight into the complex neurobiological mechanisms at work in schizophrenia.

Using ML techniques, the capacity to identify schizophrenia has significantly improved. Nevertheless, there are still numerous obstacles to be overcome. Both the introduction of standardized data collection techniques [28] and the need for more comprehensive and diverse datasets are significant obstacles that must be overcome. In addition, there is an urgent need for the development of interpretable models that cast light on the disease's underlying causes and enhance clinical decision-making. Future research is encouraged into the utility of machine learning in predicting treatment outcomes and disease progression, as well as the validation of produced models in real-world clinical settings. It is imperative that future research prioritizes this factor.

Neuroimaging studies have uncovered anomalies in the cortical regions (such as the frontal region), subcortical regions (such as the hippocampus and thalamus), and network connections of the brains of individuals with schizophrenia. Neuroimaging investigations [29–33] have uncovered anomalies related to structural and functional alterations in the brains of individuals with schizophrenia. These anomalies have been identified in cortical regions (e.g., the frontal region), subcortical regions (e.g., the hippocampus and thalamus), and alterations in network connectivity. Integrating structural neuroimaging characteristics with ML methods has been shown in an increasing number of studies to improve the diagnostic accuracy of schizophrenia. Guo et al. utilized amygdala and hippocampal subregion characteristics to differentiate between healthy controls and those with a schizophrenia diagnosis. This information was initially published in a peer-reviewed academic journal. Using Support Vector Machine Classifier (SVC/SVM) and sequential backward elimination, the features of interest were narrowed down. As a result, reports indicated an accuracy of 81.75 percent and a sensitivity of 84.1 percent. In a study by Yassin et al., 64 individuals with schizophrenia and 106 healthy controls were classified. Subcortical volume and cortical thickness served as the classification's foundation. A random forest classifier combined with subcortical volumes as features produced the highest accuracy (76.4%), according to the study. In addition, the accuracy of a decision tree analysis using cortical thickness as a feature was 70.5%. In conjunction with variables representing subcortical volume and cortical thickness, the use of logistic regression as a classifier increased accuracy to 70.5%. Xiao and colleagues conducted a classification study on 163 individuals, consisting of 163 patients with drug-free first-episode schizophrenia and 163 healthy controls. Researchers were able to achieve 81–85 percent accuracy and 77–83 percent sensitivity when measuring cortical thickness and cortical surface area [34].

Numerous studies have attempted to diagnose various mental disorders, including schizophrenia. Several of these studies have utilized EEG [35–38]. Electroencephalography (EEG) is a non-invasive technique for measuring brain function that involves placing electrodes on the cranium to record brainwave activity. The results demonstrate that electrical activity is generated when cranial nerves exchange signals [39]. Electroencephalography (EEG) is a neurophysiological instrument used to diagnose mental disorders by detecting abnormalities in normal brain function. EEG recordings can be utilized to classify individuals as mentally healthy or mentally ill by analyzing patterns of brain activity. Numerous scientific studies support the utilization of EEGs in the diagnosis and treatment of mental disorders. In addition, there is a recent proposal to combine AI

with more conventional diagnostic methods. The overwhelming majority of EEG-based research projects utilize EEGs as part of a diagnostic strategy that incorporates AI models that visualize or tabulate EEG data. However, this methodology may reduce the classification's precision. Methodological advancement is essential for obtaining the highest diagnostic precision possible.

Below are tables depicting previous diagnoses of schizophrenia based on EEG signals, MRI images, and fMRI data.

This section provides an overview of the current state of the art and prospective directions for detecting schizophrenia (SZ) through the analysis of EEG data employing diverse artificial intelligence (AI) approaches and machine learning (ML) algorithms.

Table 2.1: Previous works on EEG Signal Dataset

Study	Year	Subjects	Accuracy	AI/ML Technique
Latou et al. [36]	2014	54	84.7%	Naïve Bayes, SVM, Decision Tree, Adaboost, Random Forest
Neuhaus et al. [37]	2014	144	74%	LDA, QDA, SVM, Naïve Bayes, KNN, Mahalanobis classification
Johannesen et al. [38]	2016	40	87%	1-norm SVM

Study	Year	Subjects	Accuracy	AI/ML Technique
Shim et al. [39]	2016	34	88.24%	SVM
Taylor et al. [40]	2017	21	80.84%	SVM, Gaussian processes classifiers, MVPA
Krishnan et al. [41]	2020	14	93%	Various, SVM (Radial Basis Function)
A. Shoeibi et al. [26]	2021	28	93.75%	CNN-LSTM Models
L.Zang et al. [27]	2019	81	88%	Random Forest

The following works provide an overview of research efforts and projections pertaining to the identification of SZ through the utilization of structural MRI data via diverse AI techniques and ML algorithms.

Table 2.2: Previous works on Structural MRI Signal Dataset

Study	Year	Subjects	Prediction	AI/ML Technique
Schnack et al. [42]	2014	46/47	90%	SVM
Cabral et al. [43]	2016	71	69.7%	SVM, MVPA
Lu et al. [44]	2016	41	91.9% (sensitivity), 84.4% (specificity)	SVM, Recursive Feature Elimination (RFE)
Yang et al. [45]	2016	40	77.91%	MLDA, SVM
Squarcina et al. [46]	2017	127	80%	SVM

Study	Year	Subjects	Prediction	AI/ML Technique
Rozycki et al. [47]	2018	440	76%	Linear SVM
de Moura et al. [48]	2018	143, 32	77.6% (sensitivity), 68.3% (specificity)	MLDA
Liang et al. [49]	2019	98, 54	75.05%, 76.54%	Gradient Boosting Decision Tree
Deng et al. [50]	2019	65	76.9% (sensitivity), 75.0% (specificity)	Random Forest

The following works provide an overview of research efforts and projections pertaining to the identification of SZ through the utilization of fMRI data via diverse AI techniques and ML algorithms.

Table 2.3: Previous works on fMRI Signal Dataset

Study	Year	Subjects	Prediction	AI/ML Technique
Mikolas et al. [51]	2016	63	74.6% (sensitivity), 71.4% (specificity)	Linear SVM
Peters et al. [52]	2016	18	91%	SVM, Leave-one-out cross-validation
Yang et al. [53]	2016	40	77.91%	MLDA, SVM
Skaatun et al. [54]	2017	182	80%	Multivariate regularized LDA

Study	Year	Subjects	Prediction	AI/ML Technique
Chen et al. [55]	2017	20 (SZ), 20 (depression)	60% (sensitivity), 90% (specificity)	Linear SVM, MVPA
Kaufmann et al. [56]	2017	90 (SZ), 97 (bipolar)	60% (sensitivity), 90% (specificity)	5-class regularized LDA, k-fold cross- validation model
Guo et al. [57]	2017	28	96.43% (sensitivity), 89.29% (specificity)	SVM, Receiver operating characteristic (ROC) curve
Iwabuchi and Palaniyappan [58]	2017	71	80.32%	MKL

Study	Year	Subjects	Prediction	AI/ML Technique
Yang et al. [59]	2017	446	86%	Multi-task classification, 10-fold cross-validation
Bae et al. [60]	2018	21	92.1% (SVM)	Various (5 types), 10- fold cross-validation
Li et al. [61]	2019	60	76.34% (LDA)	KNN, Linear SVM, Radial basis SVM, LDA
Chatterjee et al. [62]	2019	34	94% (SVM), 96% (1-NN)	SVM, k-nearest neighbors
Kalmady et al. [63]	2019	81	87%	L2-regularized Logistic regression

The use of diverse data sources and the extraction of distinctive features from such sources are fundamental components of ML methodologies, which have the potential to significantly improve the identification and evaluation of schizophrenia. Neuroimaging and EEG-based methodologies have elucidated the neural underpinnings of schizophrenia, either independently or in conjunction. These observations can be identified independently or in conjunction with one another. The ongoing development of ML algorithms has the potential to revolutionize early detection and clinical decision-making in the field of schizophrenia research. In spite of the progress made thus far, additional progress is required to effectively address the existing obstacles and ensure the viability of ML in the identification of schizophrenia.

Chapter 3

Research Methodology

Any research endeavor's success is heavily dependent on the methodology employed. This is especially true when it comes to machine learning research, as classification and regression techniques are typically quite similar. Consequently, it is of the utmost importance to adapt the methodology to the diverse requirements of the datasets. Nevertheless, the general methodology pipeline for ML remains essentially unchanged for studies of similar nature. A 'Binary Classification' study's methodology follows a particular pattern, whereas 'Regression' studies take a different approach. The 'Dataset' is the initial component of any research methodology for machine learning studies. How the dataset was compiled or from where it was extracted are the primary issues that a researcher must address first. The dataset is then normalized using particular pre-processing techniques, such as locating missing values, filling in those missing values with data, and encoding the label of string-type features. After feature extraction and engineering, the EDA is used to propose new or modified features. Implementing baseline models first aids in determining which machine learning models perform better. After identifying the relatively superior models, the research should refine the models' parameters to achieve more precise results. These are the steps that have also been taken for this thesis, and Figure 3.1 provides a summary of the entire methodology.

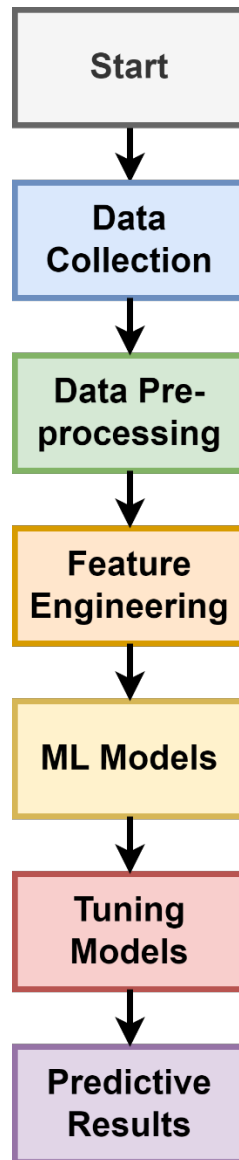


Fig 3.1. Overall workflow of the study

3.1 Dataset

For this research, there are primarily three well-established datasets, namely (one from the Northern California Institute, one from Moscow State University, and one from a group of Polish researchers). These features were combined to produce a larger database of 193 patients, which is larger than any other contemporary dataset for the detection of schizophrenia [64–66]. The description of the original dataset can be found in Table 3.1.

Table 3.1. Dataset Description

Features	Data Type	
F4	Continuous	
F8		
C3		
Cz		
C4		
P3		
Pz		
P4		
O1		
O2		
F7		
F3		
Output		
Group		Binary

These features are the distinct electrodes placed in various regions of the schizophrenic and healthy subjects' brains. The output 'Group' is of binary data type, where 0 indicates a healthy patient and 1 indicates a patient with schizophrenia. The output group comprises 193 instances, as the merged dataset contains 193 patients.

The datasets were gathered from diverse repositories, including Kaggle and RepOD. To verify the authenticity of these datasets, authentic research articles published at conferences and in academic journals were also verified.

3.2 Data Pre-processing

For data preprocessing, several steps were taken into account. They can be summarized through the following points:

- As all the features of each dataset did not match, only the common features i.e. (F4, F8, C3, Cz, C4, P3, Pz, P4, O1, O2, F7, F3) were taken into consideration.
- The input features contained many missing values and outliers which were normalized through averaging and removing outliers.
- Several EDA were implemented to further visualize the relationship between the features and how one would act in the presence of outliers.
- The correlation between the features was also determined through the ‘Seaborn’ library of python and because of the fact that there were many features removed from the dataset due to mismatch between the experiments, the correlation came out to be quite tedious and so in the feature engineering part, several other features were introduced to mitigate this effect.

The pre-processing portion can be visualized from the following figure 3.2.

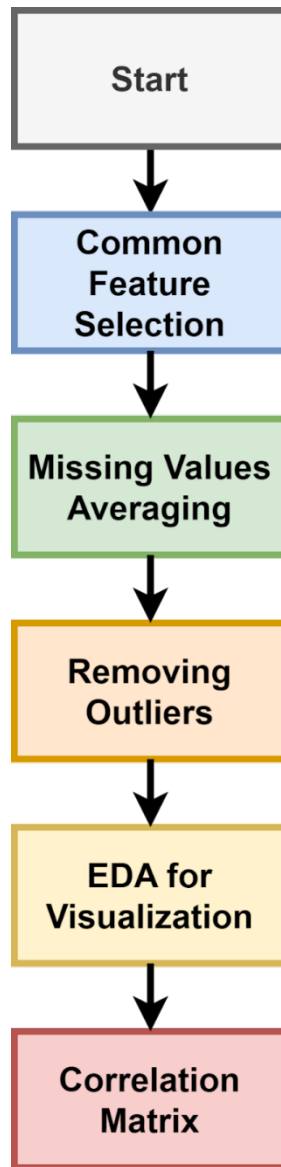


Fig 3.2. Data Pre-processing steps

The density distribution of the schizophrenia patients can be depicted from figure 3.3.

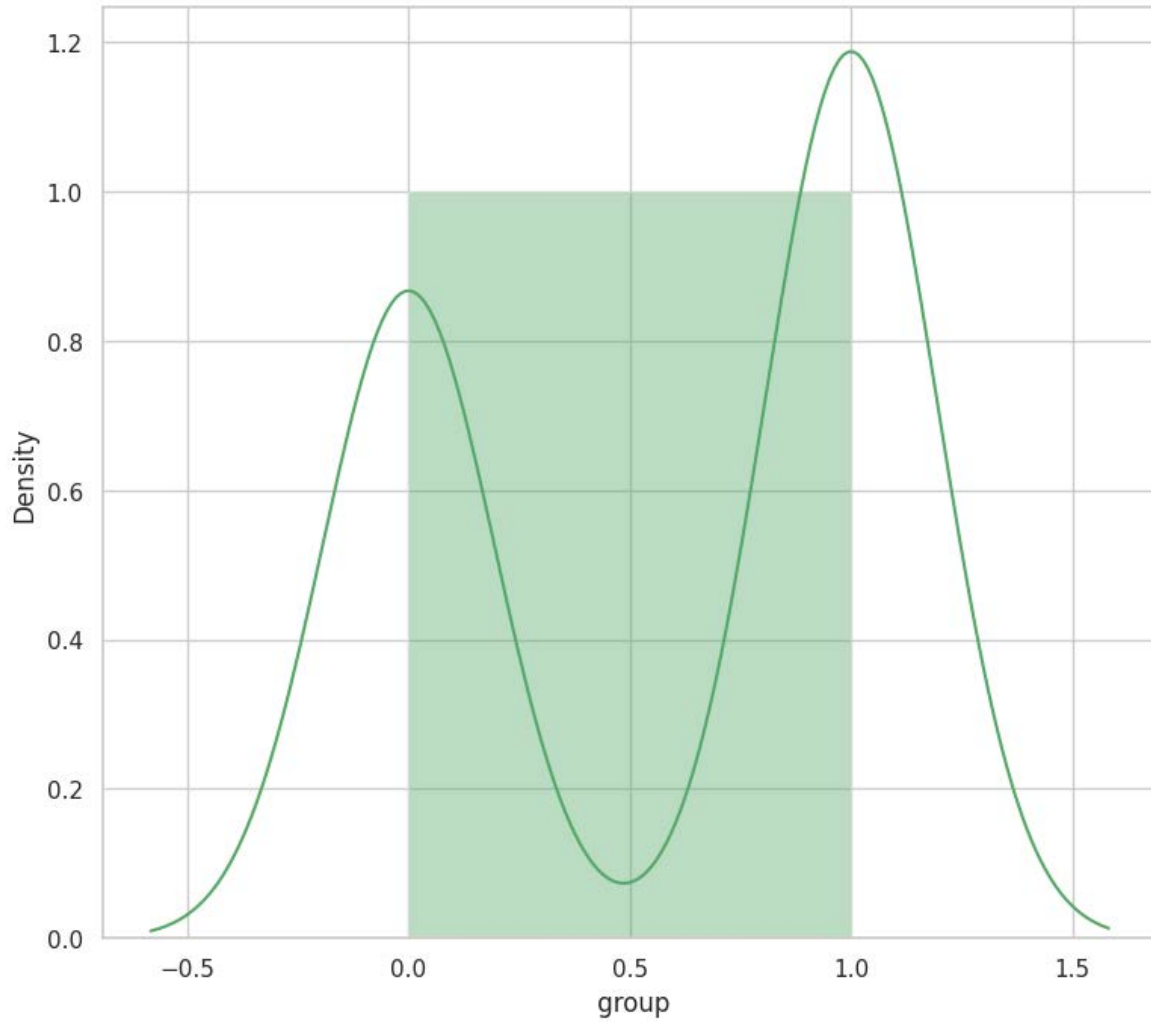


Fig 3.3. Distribution of the output 'Group' for Schizophrenia patients

Now, the figures containing the Johnson SU, normal and log distributions are depicted in the figures 3.4, 3.5 and 3.6 respectively.

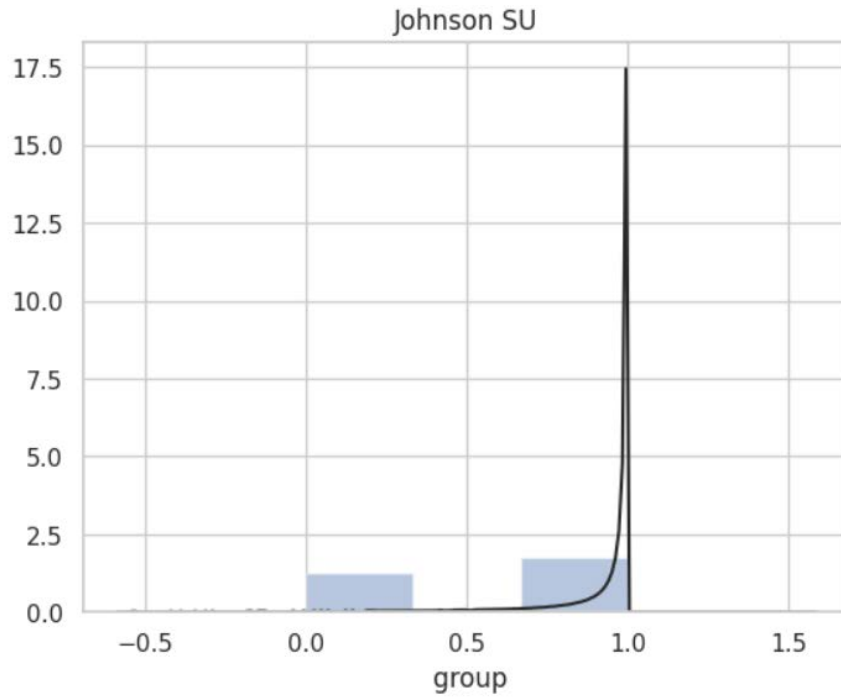


Fig 3.4. Johnson SU Distribution of the output ‘Group’ for Schizophrenia patients

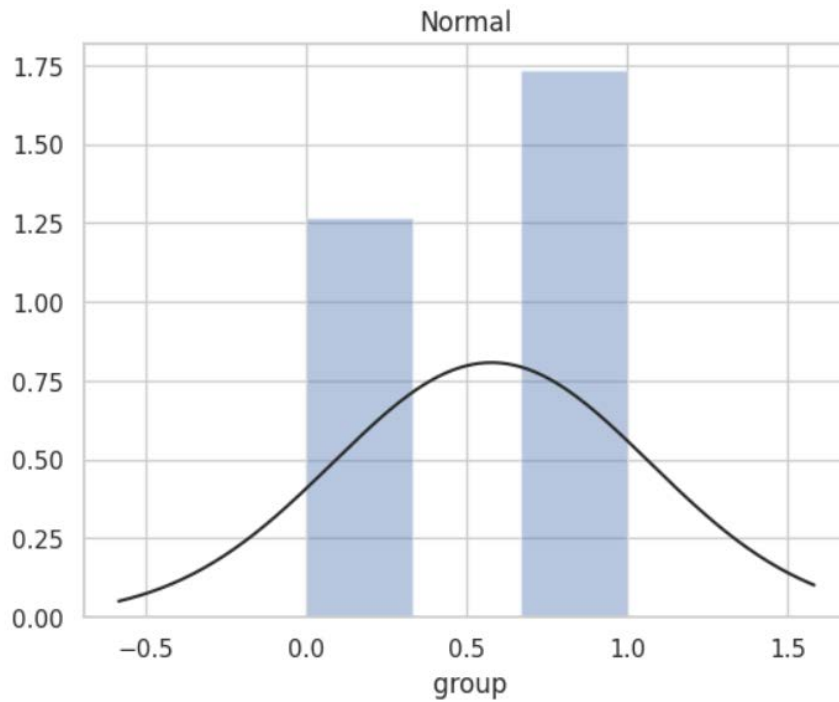


Fig 3.5. Normal distribution of the output ‘Group’ for Schizophrenia patients

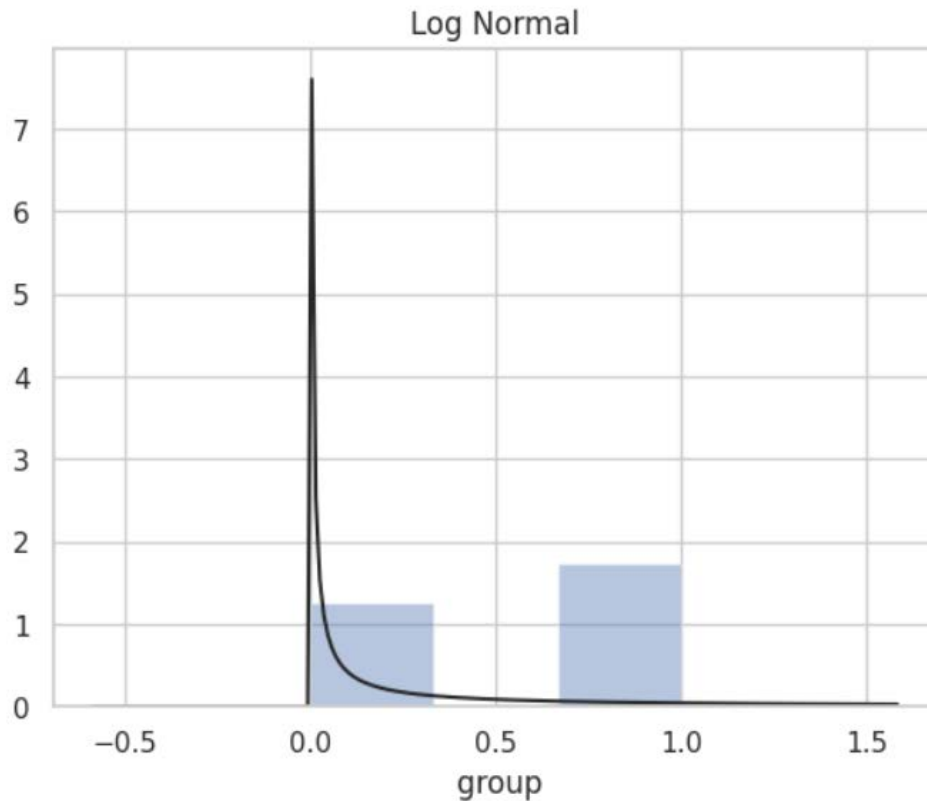


Fig 3.6. Log distribution of the output ‘Group’ for Schizophrenia patients

The skewness of a distribution indicates how much the data deviates from a symmetric distribution. A skewness value of 0 indicates a distribution that is perfectly symmetrical. When the skewness value is positive, the distribution has an extended right tail, indicating that it is right-skewed. In contrast, a negative skewness value indicates a left-skewed distribution with an extended left tail [67].

In contrast, kurtosis quantifies the peakedness or flatness of a distribution, evaluating the concentration or dispersion of data values in the ends. A kurtosis value of 0 indicates a normal distribution with the same degree of peaking as the standard normal distribution. Positive kurtosis, also known as leptokurtic, denotes a distribution with heavier tails and a more pronounced apex than a normal distribution. In contrast, negative kurtosis, also known as platykurtic, denotes a distribution with thinner tails and a flatter apex than a normal distribution [68].

The skewness and kurtosis of the output variable can be portrayed from the following figures 3.7 and 3.8 respectively.

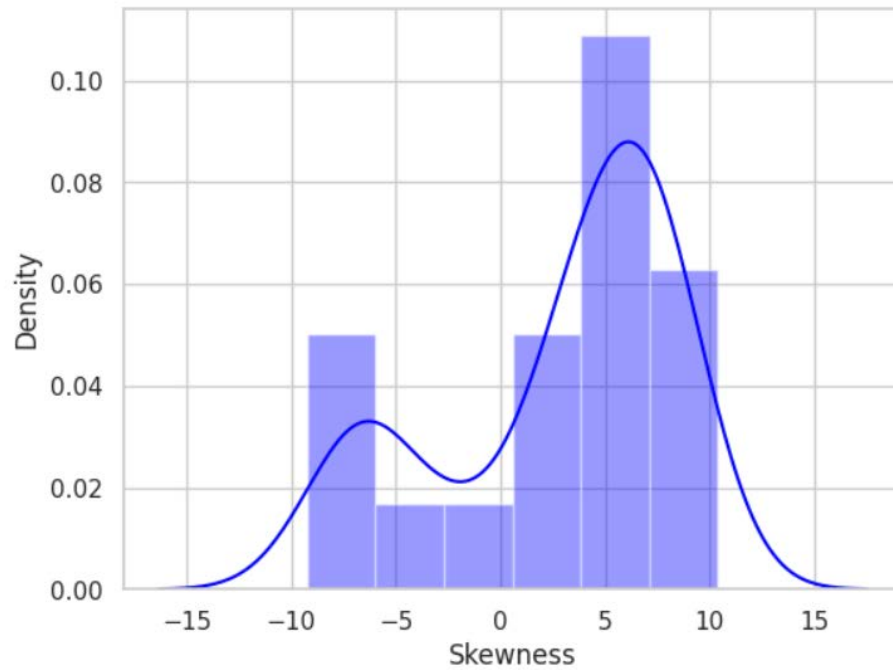


Fig 3.7. Skewness of the output 'Group' for Schizophrenia patients

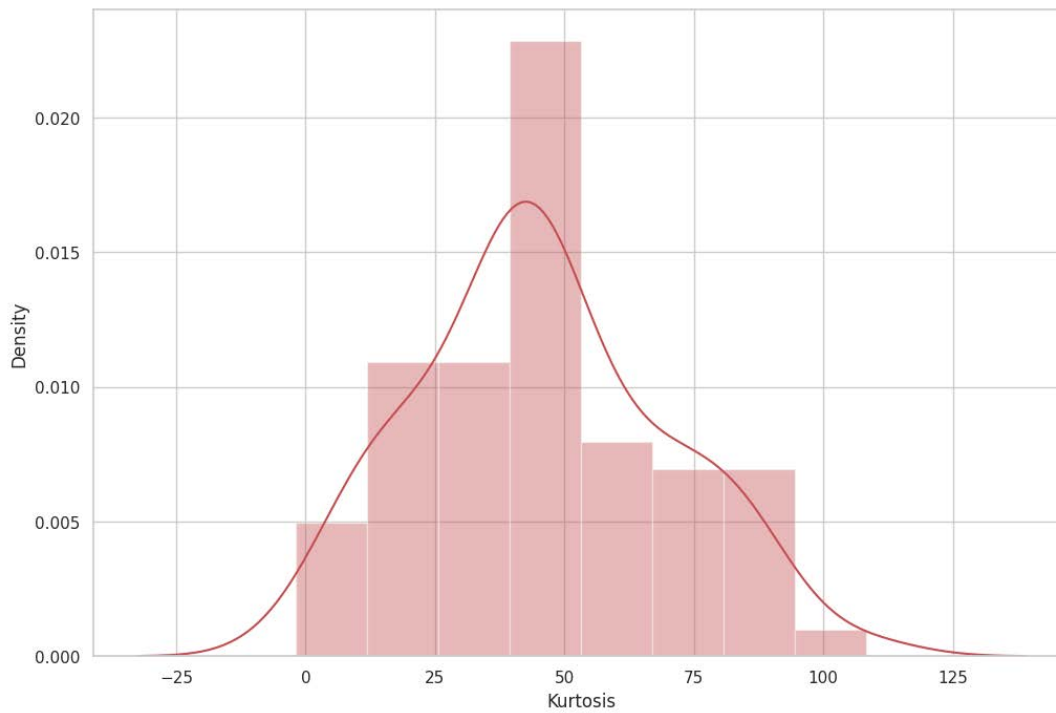


Fig 3.8. Kurtosis of the output 'Group' for Schizophrenia patients

Now comes the correlation part where the feature columns are correlated with the output group column. The following figure 3.9 shows the correlation.

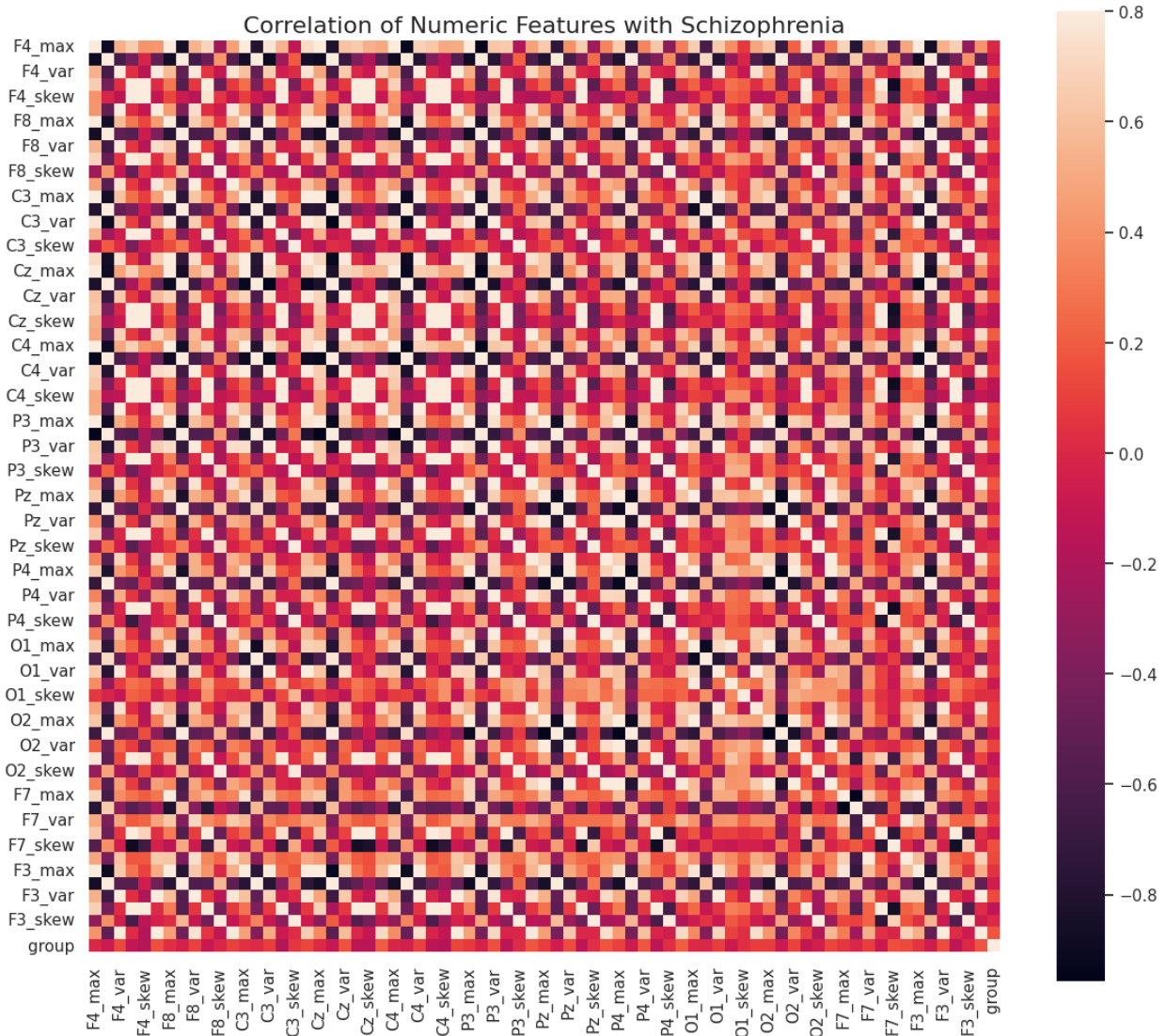


Fig 3.9. Correlation among the features and output 'Group' for Schizophrenia patients

3.3 Feature Engineering

Feature engineering comes after feature extraction, where this study has seen six statistical features introduced. They are: maximum, minimum, standard deviation, variance, kurtosis, and skewness. These features were extracted from the common features mentioned in the previous section. The

extraction process included MATLAB scripts where the six statistical figures mentioned above were found for every feature column. The code snippet can be seen in Figure 3.10.

```
close all
clear all
clc
folder = 'G:\Thesis\All';
csvFiles = dir(fullfile(folder, '*.csv'));
csvFiles = natsortfiles(csvFiles);
numfiles = length(csvFiles);
%main = zeros(numfiles, 1);
maximum = zeros(numfiles, 1);
minimum = zeros(numfiles, 1);
variance = zeros(numfiles, 1);
kurtosis1 = zeros(numfiles, 1);
skew = zeros(numfiles, 1);
std_dev = zeros(numfiles, 1);
for k = 1:numfiles
    M = csvread(fullfile(folder, csvFiles(k).name), 1, 67);
    maximum(k) = max(M(:,1));
    minimum(k) = min(M(:,1));
    variance(k) = var(M(:,1));
    kurtosis1(k) = kurtosis(M(:,1));
    skew(k) = skewness(M(:,1));
    std_dev(k) = std(M(:,1));
    %plot(M(:,1));
    %figure()
end
csvwrite(fullfile(folder, 'O2.csv'), [maximum, minimum, variance, kurtosis1, skew, std_dev]);
```

Fig 3.10. MATLAB code snippet for finding out the six statistical features

Now there are six features for each of the twelve feature columns, totaling 72 feature columns. These columns were again analyzed through a correlation matrix. In this study, the highly correlated features (>95%) were dropped as they introduced multicollinearity, which can result in overfitting [69]. As a result, thirty features were found to be highly correlated, and they were dropped to finally get a dataset with 42 input features.

3.4 Machine Learning Models

In the beginning, five supervised models were implemented. These models were referred to as CatBoost, XGBoost (XGB), LightGBM (LGBM), Extra Trees Classifier (ETC), and Gradient Boosting Classifier (GBC). These boosting classifiers were selected because, on average, they perform better for binary classification of the target class. These algorithms can also easily handle nonlinear interactions between attributes [70]. The descriptions of the models are as follows:

- **CatBoost:** CatBoost is a well-known machine learning algorithm used to solve classification and regression problems. Yandex, a Russian search engine, created it with the intention of enhancing gradient boosting. It is optimized for working with discrete data types, making it suitable for categorical variables.

CatBoost is an ensemble technique that employs the gradient boosting machine (GBM) procedure. It integrates several weak predictive models, typically decision trees, into a single, more robust model. This method enables CatBoost to manage both numerical and categorical data with varying degrees of complexity, which is one of its notable strengths.

The CatBoost model has several distinguishing features and qualities, such as:

- CatBoost is capable of handling categorical variables without the need for preprocessing or one-hot encoding. Ordered Target Statistics (OTS) is an advanced algorithm used to compute categorical features. OTS takes into account the statistical properties of the target variable when converting categorical values to numerical representations, allowing for the extraction of valuable information from these characteristics.
- Gradient boosting is the principal algorithm utilized by CatBoost. This technique sequentially trains an ensemble of decision trees, with each successive tree incorporating the enhancements made to the previous trees. CatBoost produces a potent predictive model by iteratively minimizing the loss function.
- CatBoost incorporates numerous regularization techniques to prevent overfitting and improve generalization. It employs gradient-based one-hot encoding to reduce the number of divisions for categorical variables and applies L2 regularization by including a penalty term in the loss function, thereby discouraging complex models. CatBoost efficiently manages missing values regardless of whether the feature is numerical or categorical. During training, it can acquire the ability to implicitly manage missing values, eliminating the need for explicit imputation.
- CatBoost is performance-optimized, providing rapid execution and low memory consumption. It employs parallelization techniques to speed up the training process and supports multithreading, allowing it to operate efficiently on computers with multiple processor cores.
- CatBoost supports multiple evaluation metrics, including precision, log loss, area under the ROC curve (AUC), and mean squared error (MSE), for classification and

regression tasks. These metrics are useful for evaluating the performance of the model and refining its parameters.

CatBoost is a robust ML algorithm that provides competitive performance in a variety of classification and regression problems, and it excels at dealing with categorical variables. Its robust regularization techniques and its ability to handle heterogeneous data have made it a favorite among data scientists and machine learning professionals [71-73].

- **XGB:** The ML algorithm XGBoost, an abbreviation for “Extreme Gradient Boosting,” is a potent member of the gradient boosting family of techniques. It has acquired popularity due to its high efficiency and scalability, making it a common tool for addressing classification and regression issues.

The XGB model is characterized by the following features and characteristics:

- XGB is based on the Gradient Boosting framework, which integrates multiple unreliable predictive models, typically decision trees, into a single, more accurate ensemble model. Each successive model is trained to minimize the errors of its predecessors.
- XGB utilizes numerous regularization techniques to improve generalization and prevent overfitting. L1 and L2 regularization elements are incorporated into the objective function to regulate the model’s complexity. Regularization reduces the model’s propensity to suit noise in the training data and unnecessary tree nodes.
- XGB builds decision trees using a level-wise strategy, with each tree expanding horizontally from its root node outward. This means that at each node, all possible divides are considered, and the optimal split is chosen based on a scoring criterion such as information gain or objective function improvement.
- XGB automatically resolves missing values with its built-in mechanisms. It uses training data statistics to determine where missing values should be inserted in the tree-building process. Consequently, no additional preprocessing or explicit imputation is necessary.
- XGB’s feature importance metric quantifies the importance of each feature in the model’s predictions. It evaluates the importance of a feature by counting the number of times it is used to divide nodes across all ensemble trees. This information can aid in the selection of features and disclose hidden data patterns.
- XGB is designed to be scalable and efficient. It allows for faster training by supporting parallel processing across numerous CPU cores on a single machine. It can also be deployed on clusters of computers using frameworks such as Apache Spark to effectively manage large datasets.
- XGB provides an extensive selection of evaluation metrics for classification and regression tasks. These include precision, log loss, AUC (area under the receiver operating characteristic curve), and F1 score for classification, as well as MSE (mean squared error) and RMSE (root mean squared error) for regression. These metrics permit evaluation of model efficacy and hyperparameter tuning.

XGB has garnered popularity in data science competitions and industry applications due to its high performance, scalability, and resilience. It is well-known for its adaptability to diverse datasets, efficient feature selection, and dependable predictive abilities [72-74].

- **LGBM:** LGBM is a gradient boosting framework developed by Microsoft that is specifically designed for efficient and high-performance training of gradient boosting models. Light Gradient Boosting Machine is an acronym for this term. LGBM is well-suited to large-scale datasets and real-time applications due to its speed and scalability.

The LGBM model is distinguished by the following salient features and characteristics:

- LGBM's Gradient-Based Tree Construction is distinct from that of other gradient boosting frameworks like XGB in that it employs a leaf-wise tree growth strategy. A faster convergence and higher performance can be achieved with the leaf-wise approach because the algorithm selects the split points that minimize the loss the most. However, LGBM includes mechanisms to manage tree depth and apply regularization in order to prevent overfitting.
- LGBM is implemented in a way that minimizes the amount of time and memory it uses. Big data sets with millions of rows and thousands of features are no problem for it. It accomplishes this by employing methods like exclusive feature bundling, which groups categorical features to speed up training, and histogram-based gradient computation, which reduces memory consumption.
- LGBM naturally supports categorical features, with no additional one-hot encoding or preprocessing required. Methods like "Gradient-based One-Side Sampling" (GOSS) and "Exclusive Feature Bundling" are used to effectively incorporate categorical variables into the gradient boosting procedure.
- LGBM is scalable and well-suited for large datasets because it supports parallel and distributed training. It leverages multi-threading to utilize all available CPU cores efficiently. Additionally, it can be distributed across multiple machines and integrated with distributed computing frameworks such as Apache Hadoop or Apache Spark.
- Classification and Regression Evaluation Metrics LGBM supports a wide variety of evaluation metrics. Classification metrics such as accuracy, log loss, area under the curve (AUC), and F1 score are included. MSE, MAE, and RMSE are all metrics that can be used in regression. Model performance can be evaluated and hyperparameters can be tuned with the help of these metrics.

Because of its efficiency, scalability, and adaptability to large datasets, LGBM has become increasingly popular. Its effectiveness in both research and production settings [75-78] stems from its efficient implementation, support for categorical features, and sophisticated tree construction strategies.

- **ETC:** The Extra Tree Classifier, also known as the Extremely Randomized Trees or ExtraTrees Classifier, is an ensemble learning model that belongs to the family of decision tree-based classifiers. It is an extension of the Random Forest algorithm and shares similarities with it. However, the ETC introduces additional randomness in the tree construction process, making it even more diverse and potentially reducing overfitting.

Some key features and characteristics of the ETC model are:

- **Random Feature Selection:** In the ETC, random subsets of features are considered at each split point during the tree construction process. Unlike RF, which evaluates a subset of features and selects the best split among them, the ETC randomly chooses split points without considering optimal thresholds. This random feature selection adds diversity to the trees and makes them less prone to overfitting.
- **Random Split Point Selection:** In addition to random feature selection, the ETC also introduces randomness in choosing split points. Instead of searching for the best split among all possible thresholds, it randomly selects split points without evaluating all possible thresholds. This further enhances the diversity of the trees and contributes to reducing overfitting.
- **Ensemble of Trees:** The ETC creates an ensemble of decision trees, where each tree is grown using a different subset of features and split points. The final prediction is made by aggregating the predictions from all the trees, typically using majority voting for classification problems.
- **Handling Missing Values:** The ETC handles missing values by considering them as a separate category during the split point selection process. This allows the algorithm to make use of the information carried by missing values instead of discarding them.
- **Regularization and Pruning:** While the ETC is already randomized, additional regularization techniques like tree depth and maximum number of leaf nodes can be applied to control the model's complexity and prevent overfitting. These limit the depth and size of the trees, which can improve generalization performance.
- **The average reduction in impurity that each feature provides when splitting trees is used by the ETC as a measure of feature importance.** This importance score can be used for feature selection or to determine which features contribute the most useful information to the dataset.
- **To speed up training and prediction on multi-core systems, the ETC's tree construction can be parallelized.**

The ETC is particularly useful when dealing with high-dimensional datasets or datasets with noisy features. It uses randomization's advantages to cut down on overfitting and boost generalization results. However, due to its high randomness, it may require a larger number of trees in the ensemble to achieve similar accuracy compared to other algorithms like RF [79-81].

- **GBC:** The Gradient Boosting Classifier is an ML model that belongs to the family of gradient boosting methods. This effective algorithm is applied to the resolution of classification issues. To build a robust predictive model, we train an ensemble of low-quality classifiers—typically decision trees—in sequence. It is a powerful algorithm Here are some of the model's most notable features: it is used for solving classification problems. The model is constructed by sequentially training an ensemble of weak classifiers, typically decision trees, to create a strong predictive model. The significant characteristics of the model are as follows:
 - Following the gradient boosting framework, each weak classifier in the GBC is trained to minimize the loss function by focusing on the mistakes of the previous

classifiers. Each successive weak classifier in the GBC is trained to minimize the loss function by zeroing in on the errors of the preceding classifiers, as per the gradient boosting framework. Each successive weak classifier in the GBC takes the output of several low-quality classifiers and combines them into a single estimate. GBC is trained to minimize the loss function by zeroing in on the errors of the preceding classifiers, as per the gradient boosting framework. Gradient Boosting Framework: The GBC follows the gradient boosting framework, where each subsequent weak classifier is trained to minimize the loss function by focusing on the mistakes made by the previous classifiers. It combines the predictions of multiple weak classifiers to make a final prediction.

- The GBC typically employs decision trees as the underlying estimation framework. Decision Tree Base Estimators: The GBC predominantly uses decision trees as the base estimators. Decision trees are constructed based on feature splits that optimize the reduction in the loss function. By sequentially adding decision trees, the model adapts to the complex relationships in the data and improves its predictive power.
- Gradient Optimization: The gradient boosting algorithm optimizes the loss function by computing the gradients of the loss with respect to the predicted values. It updates the model parameters in the direction that minimizes the loss, employing techniques like gradient descent. This optimization process enhances the model's ability to capture complex patterns and improve prediction accuracy.
- Learning Rate: The learning rate is a hyperparameter that controls the contribution of each weak classifier to the overall ensemble. A lower learning rate reduces the impact of each classifier, making the learning process more conservative. Conversely, a higher learning rate allows individual classifiers to have a larger influence, potentially leading to overfitting. The learning rate is typically tuned to balance model complexity and generalization performance.
- Regularization Techniques: The GBC includes regularization techniques to prevent overfitting. It applies regularization through parameters like tree depth, minimum samples per leaf, and minimum improvement in the loss function required for a split. These regularization techniques help control the complexity of the model and improve its ability to generalize to unseen data.
- Feature Importance: The GBC provides a measure of feature importance based on the contribution of each feature in reducing the loss function across all the weak classifiers. Feature importance allows for identifying the most informative features and understanding their impact on predictions. It can be used for feature selection, dimensionality reduction, or gaining insights into the data.
- Evaluation Metrics: The GBC supports various evaluation metrics to assess its performance on classification tasks. Common metrics include accuracy, precision, recall, F1-score, and AUC.

The GBC is known for its ability to handle complex relationships in data, adaptiveness, and high predictive accuracy. It is widely used in various domains, such as finance, healthcare, and e-commerce, where accurate classification is crucial for decision-making [82-83].

3.5 Tuning Models

The Hyperparameter tuning of the five models was done through the optimization technique named 'Optuna'. Optuna is an open-source hyperparameter optimization framework for ML tasks. It provides a flexible and efficient solution for automating the process of finding the optimal set of hyperparameters for a given model. Optuna uses the concept of Bayesian optimization to intelligently search the hyperparameter space and guide the search based on past evaluations.

Here are the key features and characteristics of Optuna:

- **Hyperparameter Optimization:** Optuna focuses on optimizing hyperparameters, which are the adjustable settings that determine the behavior and performance of machine learning models. Hyperparameters include learning rate, batch size, regularization strength, number of layers, and more. Optuna automates the search process by iteratively proposing and evaluating different combinations of hyperparameters to find the optimal configuration.
- **Bayesian Optimization:** Optuna utilizes Bayesian optimization, a sequential model-based optimization approach, to efficiently explore the hyperparameter space. It models the relationship between hyperparameters and their corresponding objective functions using a probabilistic model (typically Gaussian process regression). It then uses an acquisition function, such as Expected Improvement or Upper Confidence Bound, to guide the search toward promising regions of the hyperparameter space.
- **Automatic Pruning:** Optuna supports automatic pruning, a technique that stops unpromising trials early in the optimization process. Pruning helps to save computational resources by terminating trials that are unlikely to yield better results than the current best ones. Optuna integrates with various machine learning frameworks and libraries, such as PyTorch, TensorFlow, and Scikit-Learn, to leverage early stopping capabilities for efficient pruning.
- **Flexible and Extensible:** Optuna offers a high degree of flexibility in defining the search space for hyperparameters. It supports both discrete and continuous hyperparameters, as well as conditional hyperparameter search spaces, allowing the optimization of complex hyperparameter configurations. Optuna also allows the definition of custom optimization objectives and search algorithms, enabling users to tailor the optimization process to their specific needs.
- **Integration with ML Libraries:** Optuna seamlessly integrates with popular ML libraries, providing an easy-to-use interface for hyperparameter optimization. It can be used with frameworks like PyTorch, TensorFlow, scikit-learn, and XGB, among others. This integration simplifies the process of incorporating Optuna into existing ML workflows and experiments.
- **Visualization and Analysis:** Optuna provides visualization tools to analyze the results of hyperparameter optimization experiments. It generates visualizations like parallel coordinate plots, scatter plots, and optimization history charts to help understand the relationship between hyperparameters and performance metrics. These visualizations aid in gaining insights into the optimization process and identifying the best-performing hyperparameter configurations.

- **Distributed and Parallel Optimization:** Optuna supports distributed and parallel optimization, enabling the execution of multiple trials simultaneously across multiple computing resources. This allows for faster hyperparameter search and efficient utilization of computational power, especially when dealing with large-scale experiments or resource-intensive models.

Data scientists and practitioners of machine learning frequently use Optuna to automate and streamline the hyperparameter optimization process. Its intuitive interface, flexible search space definition, and integration with popular machine learning frameworks make it a powerful tool for improving model performance and accelerating the development of ML models.

3.6 Prediction Metrics

These are used to assess the quality and accuracy of predictions made by an ML model. These metrics provide quantitative measures that help evaluate how well a model is performing and compare different models against each other. The choice of prediction metrics depends on the specific task, such as classification, regression, or clustering [84]. For this study, four commonly used metrics have been used. They are:

- **Accuracy:** The proportion of correct predictions out of the total number of predictions.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

where:

TP = True Positives (correctly predicted positive instances)

TN = True Negatives (correctly predicted negative instances)

FP = False Positives (incorrectly predicted positive instances)

FN = False Negatives (incorrectly predicted negative instances)

- **Precision:** The ratio of true positive predictions to the sum of true positive and false positive predictions. It measures the model's ability to correctly identify positive instances.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

- **Recall (Sensitivity or True Positive Rate):** The ratio of true positive predictions to the sum of true positive and false negative predictions. It measures the model's ability to correctly identify all positive instances.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

- **F1 Score:** The harmonic mean of precision and recall. It provides a balanced measure of both metrics.

$$\text{F1Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Chapter 4

Predictive Results

The results of this study came in two folds: one demonstrated the base model results without any hyperparameter optimization, and the other showed clear improvements with model tuning. At first, the initial results from just the base models were extracted. They are illustrated in Table 4.1.

Table 4.1. Prediction Metrics of Base Models

Model	Accuracy	Precision	Recall	F1 Score
XGB	0.734	0.639	0.698	0.780
LGBM	0.459	0.586	0.876	0.569
ETC	0.746	0.735	0.687	0.831
CATBoost	0.740	0.691	0.789	0.636
GBC	0.740	0.713	0.587	0.787

As can be seen from the above table, the ETC model performed the best for the base models' section with an accuracy, prediction, recall, and F1 score of 0.746, 0.735, 0.687, and 0.831, respectively.

Now, after the hyperparameter optimization, the models' performance improved, and this was possible through the tuning technique Optuna. The hyperparameters that were found to be the best for each model are illustrated in Table 4.2.

Table 4.2. Optimal Hyperparameters after Optimization with Optuna

Model	Hyperparameter
XGB	'alpha': 0.35254014871037614 'lambda': 1.2955565663996598e-05 'colsample_bytree': 0.39247121357632303 'subsample': 0.7465599273062012 'learning_rate': 0.006610819529852856 'n_estimators': 1846, 'max_depth': 34 'min_child_weight': 1.1263371231243005

LGBM	'n_estimators': 2917 'reg_alpha': 3.342210478925749e-08 'reg_lambda': 0.0006540077084941416 'colsample_bytree': 0.9 'subsample': 0.5333679724572413 'learning_rate': 0.9360613793537425 'max_depth': 20 'num_leaves': 402 'min_child_samples': 3
ETC	'n_estimators': 860 'min_samples_split': 17
CAT	'colsample_bylevel': 0.05486904361435098 'depth': 12 'boosting_type': 'Plain' 'bootstrap_type': 'MVS'
GBC	'n_estimators': 105 'max_depth': 4 'learning_rate': 0.40610661612984045

The results improved quite a bit after the optimization of the hyperparameters. The updated results can be seen from table 4.3.

Table 4.3. Results after Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1 Score
XGB	0.760	0.733	0.800	0.719
LGBM	0.633	0.733	0.566	0.606
ETC	0.746	0.735	0.687	0.831
CATBoost	0.740	0.691	0.789	0.636
GBC	0.933	0.846	0.800	0.820

Therefore, from the above table 4, it is clear that the GBC model performed the best among all the other models after hyperparameter optimization through optuna. This can be well visualized in the following figure 4.1.

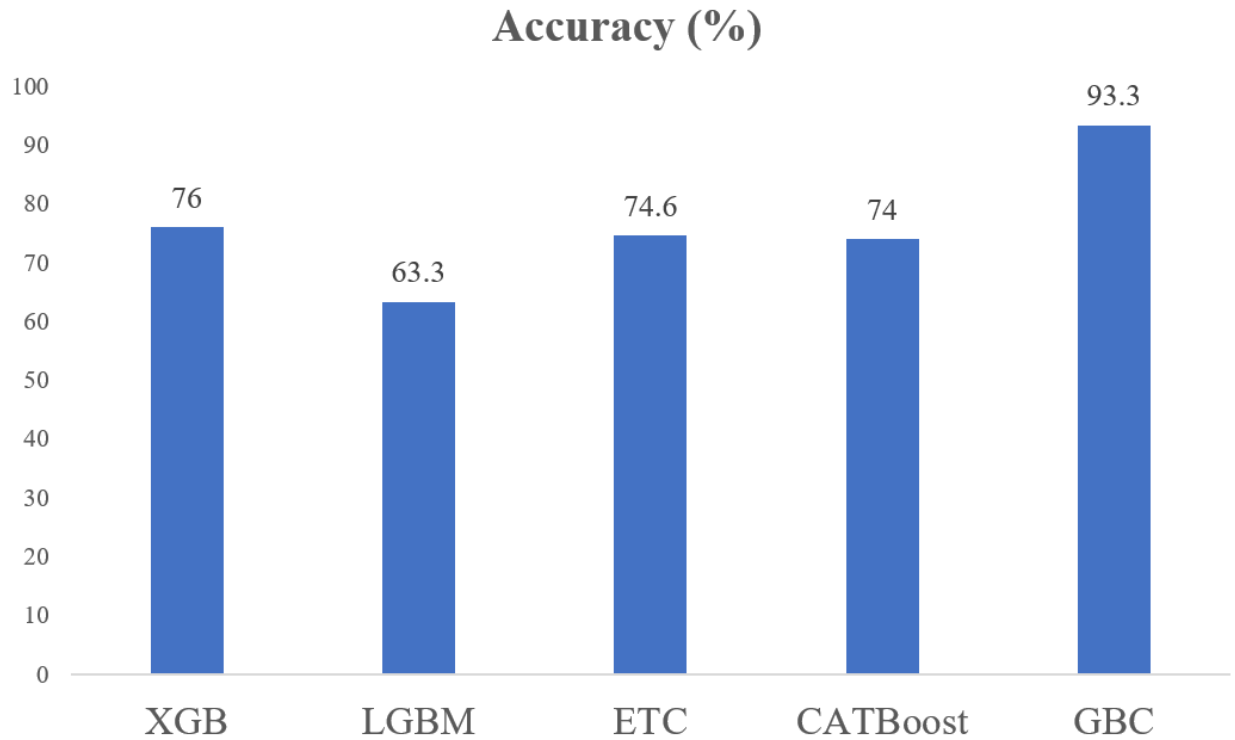


Fig 4.1. Results after Hyperparameter Optimization

The comparison among other contemporary studies is showcased in the following table 4.4.

Table 4.4. Comparison among contemporary studies

Reference	Year	Subjects	Models	Accuracy
Latou et al. [78]	2014	54	Naïve Bayes, SVM, Decision Tree, Adaboost, Random Forest	84.7%
Neuhaus et al. [79]	2014	144	LDA, QDA, SVM, Naïve Bayes, KNN, Mahalanobis classification	74%

Johannesen et al. [80]	2016	40	1-norm SVM	87%
Shim et al. [81]	2016	34	SVM	88.24%
Taylor et al. [82]	2017	21	SVM, Gaussian processes classifiers, MVPA	80.84%
L.Zang et al. [69]	2019	81	Random Forest	88%
Krishnan et al. [83]	2020	14	Various, SVM (Radial Basis Function)	93%
A. Shoeibi et al. [68]	2021	28	CNN-LSTM Models	93.75%

This Study	2023	193	XGB, LGBM, ETC, CATBoost, GBC	93.33%
-------------------	-------------	------------	--------------------------------------	---------------

For better visualization, the following figure 4.2 can be portrayed.

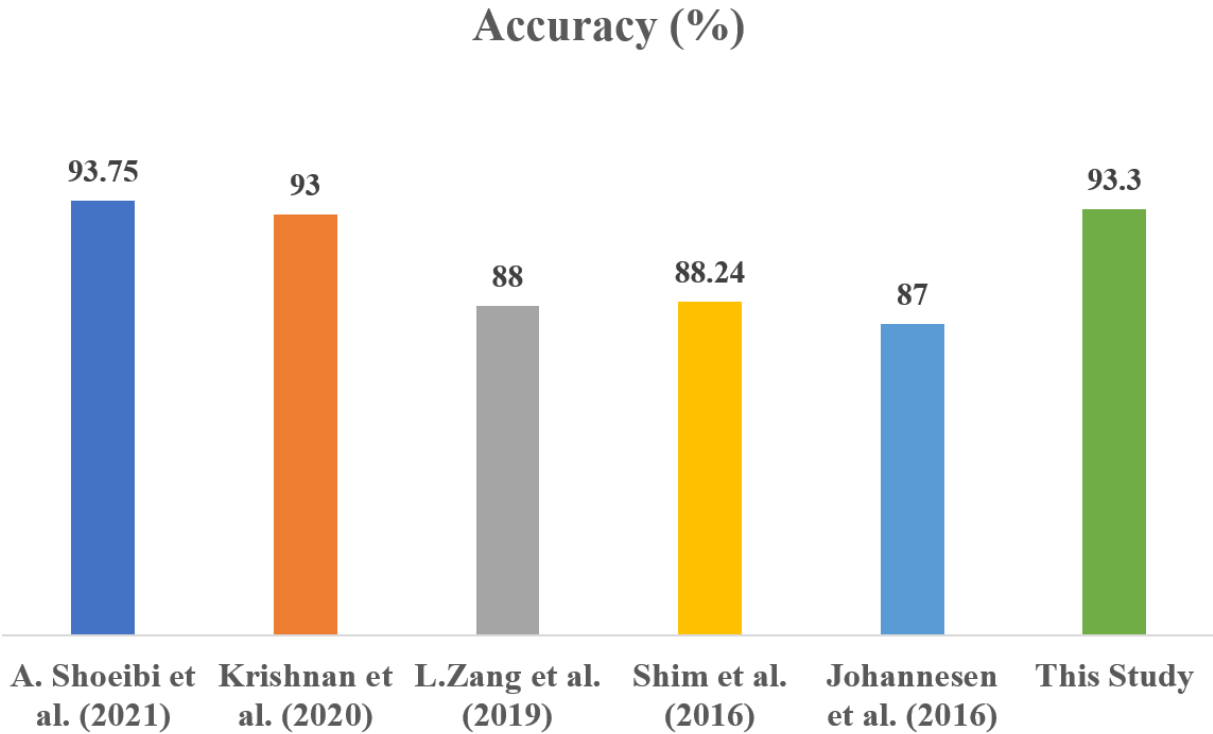


Fig 4.2. Comparison of Results with Contemporary Studies

To summarize, the accuracy and other prediction metrics of the study proved to be the best among all the other studies 93.3%. Though the A. Shoeibi et al. (2021) paper showed better accuracy, the other prediction metrics came out as superior in this study.

Chapter 5

Discussion

Schizophrenia is a severe and chronic mental disorder that affects an individual's beliefs, emotions, and behavior. It is characterized by hallucinations, delusions, cognitive disorganization, and impaired social functioning. It has a significant impact on a person's ability to live a fulfilling and productive existence.

The timely diagnosis of schizophrenia is crucial for multiple reasons. First, it enables early intervention and treatment, which can help manage symptoms, reduce the severity of the disease, and enhance long-term outcomes. Early treatment can reduce the risk of relapse, hospitalization, and functional decline, according to research.

In addition, early detection allows medical professionals to make precise diagnoses. Schizophrenia can be difficult to diagnose because its symptoms overlap with those of other mental health conditions. Early diagnosis enables healthcare providers to differentiate schizophrenia from other disorders and administer the most suitable treatment.

Early detection of schizophrenia also provides individuals and their families with vital information and support. Early diagnosis enables patients to better understand their condition and make informed treatment decisions. It allows them to engage in psychoeducation, acquire coping skills, and gain access to support services that can improve their well-being as a whole.

Additionally, early detection can help reduce the societal and economic burden of schizophrenia. Schizophrenia can result in significant personal and societal costs, including increased healthcare expenses, unemployment, homelessness, and involvement in the criminal justice system, if left undiagnosed and untreated. Early identification and treatment of schizophrenia can reduce the burden on individuals, families, and society.

Multiple factors contribute to the significance of early detection of schizophrenia, including raising awareness among healthcare personnel, the general public, and individuals about the early signs and symptoms of the disorder. Promoting mental health literacy, reducing stigma, and encouraging individuals to seek prompt assistance when they observe disconcerting mental health changes are essential.

Therefore, early identification of schizophrenia is crucial for improving the outcomes of those affected by this mental disorder. It enables timely intervention, accurate diagnosis, access to appropriate treatments and support services, and reduces the burden on individuals, families, and society as a whole. Continue to focus on increasing mental health awareness, enhancing mental health knowledge, and integrating early detection and intervention into mental healthcare systems.

The use of EEG to predict schizophrenia is a promising area of ongoing research with the potential to improve early detection and treatment of this mental disorder. EEG is a non-invasive method for measuring electrical brain activity using cranium electrodes. By analyzing EEG signals, researchers have investigated the possibility of identifying abnormalities or patterns that could serve as biomarkers for schizophrenia.

Multiple investigations have examined various EEG characteristics to predict schizophrenia. Analysis of event-related potentials (ERPs), brain responses that occur during specific tasks or stimuli, is a prevalent method. Individuals with schizophrenia have been observed to exhibit abnormal ERPs, including the mismatch negativity (MMN) component. Researchers have discovered that diminished MMN amplitudes or alterations in other ERP components may serve as potential illness predictors.

Analysis of resting-state EEG recordings is an additional topic of interest. Resting-state EEG measures brain activity when a person is not performing a specific task or being stimulated. Various EEG frequency bands, including alpha, beta, theta, and gamma, have been investigated in studies to identify aberrant patterns in individuals at risk for or diagnosed with schizophrenia. Changes in functional connectivity, coherence, power spectrum, or complexity measures in particular frequency bands have shown promise as potential predictors of the disorder.

Frequently, machine learning techniques are used to analyze complex EEG data and build prediction models. These models use algorithms to discover patterns and associations between EEG feature characteristics and diagnostic outcomes. By training the models on a large dataset of EEG recordings from individuals with schizophrenia and healthy controls, researchers are able to identify patterns that accurately predict the presence or risk of schizophrenia in new individuals.

The use of EEG for predicting schizophrenia has the potential for early detection and intervention, despite the fact that research in this field is still in progress. It is essential to observe, however, that EEG-based prediction models are not yet suitable for routine clinical use. To validate the findings, establish robust and reliable biomarkers, and refine prediction models, additional research is required. In addition, integrating EEG-based prediction with other clinical evaluations and biomarkers may improve accuracy and clinical utility.

Overall, the EEG-based prediction of schizophrenia is an exciting area of study that has the potential to enhance early diagnosis and treatment of this complex mental disorder. By enabling targeted and timely interventions, advancements in this field could contribute to improved outcomes for individuals at risk for or diagnosed with schizophrenia.

This study therefore proposes a reliable EEG-based method for the early detection of schizophrenia patients. The novelty of this research resides in the creation of 193 samples of healthy and schizophrenic patients by combining 193 EEG signal datasets, which is the most samples compared to previous studies. Consequently, this research's methodology is unquestionably solid in terms of authenticity and rigor. Despite the fact that the prediction metrics did not demonstrate significant improvement compared to previous studies, the larger number of data samples, which may have lowered the prediction metrics compared to other studies that only included about 100 samples, is the crucial factor.

Both magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI) are advanced imaging techniques that have been extensively investigated in relation to schizophrenia. fMRI measures changes in blood flow and oxygenation, enabling researchers to investigate brain activity and connectivity.

Individuals with schizophrenia have structural brain abnormalities, as determined by MRI studies. These abnormalities include decreased gray matter volume, specifically in the prefrontal cortex, hippocampus, and temporal lobe. These results indicate that structural MRI can aid in the diagnosis

of schizophrenia by identifying these distinct brain alterations. It is essential to emphasize, however, that structural MRI findings are not unique to schizophrenia and can be observed in other mental disorders.

fMRI has been used to investigate functional brain changes in schizophrenia patients. Studies utilizing fMRI during the resting state have revealed a disruption in the functional connectivity between various brain regions in individuals with schizophrenia. These disruptions are frequently observed in cognitive, emotional, and sensory processing networks. In addition, task-based fMRI studies have revealed aberrant activation patterns during cognitive tasks, indicating deficits in specific brain regions and networks in people with schizophrenia.

Similar to EEG, machine learning techniques have been applied to MRI and fMRI data to create schizophrenia prediction models. Using patterns of brain structure or function, these models classify individuals as either healthy controls or schizophrenia patients. Researchers seek to identify neuroimaging biomarkers that can predict the presence or risk of schizophrenia in new individuals by training models on large datasets.

It is essential to observe, however, that neuroimaging techniques, such as MRI and fMRI, are not used as standalone diagnostic tools for schizophrenia in clinical practice. Primarily used in research settings to improve our comprehension of the disorder's neural aspects.

The translation of neuroimaging findings into clinical practice presents obstacles. The complexity and heterogeneity of schizophrenia, as well as the need for larger and more diverse data sets, make it difficult to develop accurate prediction models. In addition, the high cost and limited availability of MRI and fMRI prevent their widespread application as diagnostic instruments.

MRI and fMRI have provided invaluable insights into the structural and functional brain changes associated with schizophrenia. These techniques hold promise for enhancing early detection and elucidating the disorder's underlying neural mechanisms. Before these techniques can be implemented into routine clinical practice, however, additional research is required to validate the findings, establish reliable biomarkers, and develop robust prediction models. Consequently, the current standard practice for detecting schizophrenia relies on datasets of EEG signals that are presently scarce. In the context of machine learning, conducting more research experiments to acquire larger datasets will prove beneficial.

Chapter 6

Conclusion

Schizophrenia negatively impacts cognitive functions, emotions, and behaviors. Detection at an early stage is essential for effective intervention, accurate diagnosis, and enhanced long-term outcomes. It enables individuals to acquire knowledge about their illness, to make well-informed decisions, and to obtain the necessary support. By identifying and treating individuals with the disorder as early as feasible, the social and economic impact of schizophrenia can be mitigated. There is a growing understanding of the significance of early detection and intervention, and the stigma associated with mental health issues is dwindling.

Electroencephalography (EEG) is a non-invasive technique that analyzes brain activity utilizing implanted electrodes and has the potential to predict schizophrenia. Researchers have examined EEG recordings in the quiescent state and event-related potentials (ERPs) to detect anomalies in individuals at risk for or diagnosed with schizophrenia. Using machine learning techniques, models of the presence or risk of schizophrenia are developed. However, EEG-based prediction models are not yet clinically applicable. Validating results, identifying dependable biomarkers, and enhancing prediction algorithms require additional research.

Combining EEG with other clinical measurements may improve its clinical accuracy and utility. This study combined 193 EEG signal datasets from healthy and schizophrenic participants, demonstrating high levels of authenticity and reliability. In the study of schizophrenia, magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI) are valuable neuroimaging techniques. They can detect structural brain abnormalities and disruptions in interregional connectivity. Although these findings are not unique to schizophrenia, they can aid in its diagnosis. Studies utilizing task-based fMRI indicate that individuals with schizophrenia exhibit aberrant activation patterns, indicating deficits in particular brain regions and networks.

On the basis of MRI and fMRI data, prediction models for schizophrenia are developed using machine learning techniques. Due to their complexity, heterogeneity, and high costs, these methodologies are not used as standalone diagnostic tools in clinical practice at this time. Validating findings, identifying reliable biomarkers, and developing robust prediction algorithms requires additional research.

In this study, the classification of schizophrenic and healthy individuals yielded accuracy, precision, recall, and f1 scores of 93.3%, 84.6%, 80%, and 82.2%, respectively. This demonstrated that the study is preferable to other contemporary works in terms of both the number of subjects and prediction metrics. In addition, the significance of larger datasets is emphasized, as the focus of the research was to emphasize the dearth of larger datasets of EEG signals from schizophrenic patients. This research also sought to pave the way for integrating multiple datasets in order to facilitate the process of conducting robust investigations and educate individuals on how to combat these rare mental disorders.

References

- [1] Häfner, Heinz. “Psychische Krankheit – Ein Mehrregionenbegriff.” *Fortschritte Der Neurologie · Psychiatrie*, vol. 87, no. 12, 17 Dec. 2018, pp. 685–694, <https://doi.org/10.1055/a-0624-9456>.
- [2] ---. “World Mental Health Report: Transforming Mental Health for All.” *Www.who.int*, 16 June 2022, www.who.int/publications/i/item/9789240049338.
- [3] Institute for Health Metrics and Evaluation. “GBD Results.” Institute for Health Metrics and Evaluation, University of Washington, 2019, vizhub.healthdata.org/gbd-results/.
- [4] World Health Organization. “Mental Health and COVID-19: Early Evidence of the Pandemic’s Impact: Scientific Brief, 2 March 2022.” *Www.who.int*, 2 Mar. 2022, www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1.
- [5] Telles-Correia, Diogo, et al. “Mental Disorder—the Need for an Accurate Definition.” *Frontiers in Psychiatry*, vol. 9, no. 64, 12 Mar. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5857571/, <https://doi.org/10.3389/fpsyt.2018.00064>.
- [6] World Health Organization. “Mental Disorders.” World Health Organization, 8 June 2022, www.who.int/news-room/fact-sheets/detail/mental-disorders.
- [7] ---. “Schizophrenia.” *Nature Reviews Disease Primers*, vol. 1, no. 1, 12 Nov. 2015, www.nature.com/articles/nrdp201567, <https://doi.org/10.1038/nrdp.2015.67>.
- [8] Freedman, Robert. “Schizophrenia.” *New England Journal of Medicine*, vol. 349, no. 18, 30 Oct. 2003, pp. 1738–1749, <https://doi.org/10.1056/nejmra035458>.
- [9] Mueser, K.T., Salyers, M.P. and Mueser, P.R., 2001. A prospective analysis of work in schizophrenia. *Schizophrenia bulletin*, 27(2), pp.281-296.

- [10] “Schizophrenia - Symptoms and Causes.” Mayo Clinic, 7 Jan. 2020, www.mayoclinic.org/diseases-conditions/schizophrenia/symptoms-causes/syc-20354443#:~:text=Overview.
- [11] “GBD Results Tool | GHDx.” Ghdx.healthdata.org, ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/27a7644e8ad28e739382d31e77589dd7.
- [12] Armstrong, Este, et al. “The Ontogeny of Human Gyrfication.” *Cerebral Cortex*, vol. 5, no. 1, 1995, pp. 56–63, <https://doi.org/10.1093/cercor/5.1.56>. Accessed 16 Oct. 2021.
- [13] Andreasen, Nancy C. “Understanding the Causes of Schizophrenia.” *New England Journal of Medicine*, vol. 340, no. 8, 25 Feb. 1999, pp. 645–647, www.nejm.org/doi/full/10.1056/NEJM199902253400811, <https://doi.org/10.1056/nejm199902253400811>.
- [14] Hussien, Zebiba Nassir. “Prevalence and Associate Factors of Suicidal Ideation and Attempt among People with Schizophrenia at Amanuel Mental Specialized Hospital Addis Ababa, Ethiopia.” *Journal of Psychiatry*, vol. 18, no. 1, 2015, <https://doi.org/10.4172/psychiatry.1000184>. Accessed 25 Aug. 2019.
- [15] Rasool, Shahid, et al. “Schizophrenia: An Overview.” *Clinical Practice*, vol. 15, no. 5, 2018, <https://doi.org/10.4172/clinical-practice.1000417>.
- [16] ---. “A National Study of Violent Behavior in Persons with Schizophrenia.” *Archives of General Psychiatry*, vol. 63, no. 5, 1 May 2006, p. 490, jamanetwork.com/journals/jamapsychiatry/fullarticle/209569, <https://doi.org/10.1001/archpsyc.63.5.490>. Accessed 22 Oct. 2019.
- [17] Zhang, Lei. “EEG Signals Classification Using Machine Learning for the Identification and Diagnosis of Schizophrenia.” 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), July 2019, <https://doi.org/10.1109/embc.2019.8857946>.

[18] Vincent, J.-L. and Creteur, J. (2019). Critical care medicine in 2050: less invasive, more connected, and personalized. *Journal of Thoracic Disease*, 11(1), pp.335–338. doi:<https://doi.org/10.21037/jtd.2018.11.66>.

[19] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), pp.24–29. doi:<https://doi.org/10.1038/s41591-018-0316-z>.

[20] Blinowska, Katarzyna, and Piotr Durka. “Electroencephalography (EEG).” *Wiley Encyclopedia of Biomedical Engineering*, 14 Apr. 2006, <https://doi.org/10.1002/9780471740360.ebs0418>.

[21] Gordillo, Darío, et al. The EEG Multiverse of Schizophrenia. Vol. 33, no. 7, 27 Aug. 2022, pp. 3816–3826, <https://doi.org/10.1093/cercor/bhac309>. Accessed 4 June 2023

[22] S. L. Oh *et al.*, “A deep learning approach for Parkinson’s disease diagnosis from EEG signals,” *Neural Comput & Applic*, vol. 32, no. 15, pp. 10927–10933, Aug. 2020, doi: [10.1007/s00521-018-3689-5](https://doi.org/10.1007/s00521-018-3689-5).

[23] U. R. Acharya *et al.*, “A Novel Depression Diagnosis Index Using Nonlinear Features in EEG Signals,” *European Neurology*, vol. 74, no. 1–2, pp. 79–83, Aug. 2015, doi: [10.1159/000438457](https://doi.org/10.1159/000438457).

[24] Y. Zhu *et al.*, “Application of a Machine Learning Algorithm for Structural Brain Images in Chronic Schizophrenia to Earlier Clinical Stages of Psychosis and Autism Spectrum Disorder: A Multiprotocol Imaging Dataset Study,” *Schizophrenia Bulletin*, vol. 48, no. 3, pp. 563–574, May 2022, doi: [10.1093/schbul/sbac030](https://doi.org/10.1093/schbul/sbac030).

[25] A. Shoeibi *et al.*, “Automatic Diagnosis of Schizophrenia in EEG Signals Using CNN-LSTM Models,” *Frontiers in Neuroinformatics*, vol. 15, 2021, Accessed: May 27, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fninf.2021.777977>

[26] L. Zhang, “EEG Signals Classification Using Machine Learning for The Identification and Diagnosis of Schizophrenia,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 4521–4524. doi: [10.1109/EMBC.2019.8857946](https://doi.org/10.1109/EMBC.2019.8857946).

- [27] D.-W. Ko and J.-J. Yang, "EEG-Based Schizophrenia Diagnosis through Time Series Image Conversion and Deep Learning," *Electronics*, vol. 11, no. 14, Art. no. 14, Jan. 2022, doi: [10.3390/electronics11142265](https://doi.org/10.3390/electronics11142265).
- [28] J. Oh, B.-L. Oh, K.-U. Lee, J.-H. Chae, and K. Yun, "Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm," *Frontiers in Psychiatry*, vol. 11, 2020, Accessed: May 27, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsy.2020.00016>
- [29] G. S. Chilla, L. Y. Yeow, Q. H. Chew, K. Sim, and K. N. B. Prakash, "Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and Ensemble methods," *Sci Rep*, vol. 12, no. 1, Art. no. 1, Feb. 2022, doi: [10.1038/s41598-022-06651-4](https://doi.org/10.1038/s41598-022-06651-4).
- [30] J. Ruiz de Miras, A. J. Ibáñez-Molina, M. F. Soriano, and S. Iglesias-Parro, "Schizophrenia classification using machine learning on resting state EEG signal," *Biomedical Signal Processing and Control*, vol. 79, p. 104233, Jan. 2023, doi: [10.1016/j.bspc.2022.104233](https://doi.org/10.1016/j.bspc.2022.104233).
- [31] K. Desai, "Using Electroencephalographic Signal Processing and Machine Learning Binary Classification to diagnose Schizophrenia," In Review, preprint, Mar. 2023. doi: [10.21203/rs.3.rs-2715657/v1](https://doi.org/10.21203/rs.3.rs-2715657/v1).
- [32] Y. Xiao *et al.*, "Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI," *Schizophrenia Research*, vol. 214, pp. 11–17, Dec. 2019, doi: [10.1016/j.schres.2017.11.037](https://doi.org/10.1016/j.schres.2017.11.037).
- [33] Laton, J.; Van Schependom, J.; Gielen, J.; Decoster, J.; Moons, T.; De Keyser, J.; De Hert, M.; Nagels, G. Single-subject classification of schizophrenia patients based on a combination of oddball and mismatch evoked potential paradigms. *J. Neurol. Sci.* 2014, 347, 262–267. [Google Scholar] [CrossRef]
- [37] Laton, J.; Van Schependom, J.; Gielen, J.; Decoster, J.; Moons, T.; De Keyser, J.; De Hert, M.; Nagels, G. Single-subject classification of schizophrenia patients based on a combination of oddball and mismatch evoked potential paradigms. *J. Neurol. Sci.* **2014**, 347, 262–267.
- [35] Neuhaus, A.H.; Popescu, F.C.; Rentzsch, J.; Gallinat, J. Critical evaluation of auditory event-related potential deficits in schizophrenia: Evidence from large-scale single-subject pattern classification. *Schizophr. Bull.* **2014**, 40, 1062–1071.
- [36] Johannesen, J.K.; Bi, J.; Jiang, R.; Kenney, J.G.; Chen, C.M.A. Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatr. Electrophysiol.* **2016**, 2, 3–21.
- [36] Shim, M.; Hwang, H.J.; Kim, D.W.; Lee, S.H.; Im, C.H. Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. *Schizophr. Res.*

2016, 176, 314–319.

[38] Taylor, J.A.; Matthews, N.; Michie, P.T.; Rosa, M.J.; Garrido, M.I. Auditory prediction errors as individual biomarkers of schizophrenia. *NeuroImage Clin.* **2017**, 15, 264–273.

[39] Krishnan, P.T.; Raj, A.N.J.; Balasubramanian, P.; Chen, Y. Schizophrenia detection using Multivariate Empirical Mode Decomposition and Entropy Measures from Multichannel EEG Entropy measures from multichannel EEG signal. *Biocybern. Biomed. Eng.* **2020**, 40, 1124–1139.

[40] Schnack, H.G.; Nieuwenhuis, M.; van Haren, N.E.; Abramovic, L.; Scheewe, T.W.; Brouwer, R.M.; Pol, H.E.H.; Kahn, R.S. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *Neuroimage* **2014**, 84, 299–306.

[41] Cabral, C.; Kambeitz-Ilanovic, L.; Kambeitz, J.; Calhoun, V.D.; Dwyer, D.B.; Von Saldern, S.; Urquijo, M.F.; Falkai, P.; Koutsouleris, N. Classifying schizophrenia using multimodal multivariate pattern recognition analysis: Evaluating the impact of individual clinical profiles on the neurodiagnostic performance. *Schizophr. Bull.* **2016**, 42, S110–S117.

[42] Lu, X.; Yang, Y.; Wu, F.; Gao, M.; Xu, Y.; Zhang, Y.; Yao, Y.; Du, X.; Li, C.; Wu, L.; et al. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. *Medicine (Baltimore)* **2016**, 95, e3973.

[43] Squarcina, L.; Castellani, U.; Bellani, M.; Perlini, C.; Lasalvia, A.; Dusi, N.; Bonetto, C.; Cristofalo, D.; Tosato, S.; Rambaldelli, G.; et al. Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques. *Neuroimage* **2017**, 145, 238–245.

[44] Rozycki, M.; Satterthwaite, T.D.; Koutsouleris, N.; Erus, G.; Doshi, J.; Wolf, D.H.; Fan, Y.; Gur, R.E.; Gur, R.C.; Meisenzahl, E.M.; et al. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophr. Bull.* **2018**, 44, 1035–1044.

[45] de Moura, A.M.; Pinaya, W.H.L.; Gadelha, A.; Zugman, A.; Noto, C.; Cordeiro, Q.; Belangero, S.I.; Jackowski, A.P.; Bressan, R.A.; Sato, J.R. Investigating brain structural patterns in first episode psychosis and schizophrenia using MRI and a machine learning approach. *Psychiatry Res. Neuroimaging* **2018**, 275, 14–20.

[46] Liang, S.; Li, Y.; Zhang, Z.; Kong, X.; Wang, Q.; Deng, W.; Li, X.; Zhao, L.; Li, M.; Meng, Y.; et al. Classification of first-episode schizophrenia using multimodal brain features: A combined structural and diffusion imaging study. *Schizophr. Bull.* **2019**, 45, 591–599.

- [47] Deng, Y.; Hung, K.S.; Lui, S.S.; Chui, W.W.; Lee, J.C.; Wang, Y.; Li, Z.; Mak, H.K.; Sham, P.C.; Chan, R.C.; et al. Tractography-based classification in distinguishing patients with first-episode schizophrenia from healthy individuals. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **2019**, *88*, 66–73.
- [48] Mikolas, P.; Melicher, T.; Skoch, A.; Matejka, M.; Slovakova, A.; Bakstein, E.; Hajek, T.; Spaniel, F. Connectivity of the anterior insula differentiates participants with first-episode schizophrenia spectrum disorders from controls: A machine-learning study. *Psychol. Med.* **2016**, *46*, 2695–2704.
- [49] Peters, H.; Shao, J.; Scherr, M.; Schwerthöffer, D.; Zimmer, C.; Förstl, H.; Bäuml, J.; Wohlschläger, A.; Riedl, V.; Koch, K.; et al. More consistently altered connectivity patterns for cerebellum and medial temporal lobes than for amygdala and striatum in schizophrenia. *Front. Hum. Neurosci.* **2016**, *10*, 55.
- [50] Skåtun, K.C.; Kaufmann, T.; Doan, N.T.; Alnæs, D.; Córdova-Palomera, A.; Jönsson, E.G.; Fatouros-Bergman, H.; Flyckt, L.; KaSP; Melle, I.; et al. Consistent functional connectivity alterations in schizophrenia spectrum disorder: A multisite study. *Schizophr. Bull.* **2017**, *43*, 914–924.
- [51] Yang, H.; He, H.; Zhong, J. Multimodal MRI characterisation of schizophrenia: A discriminative analysis. *Lancet* **2016**, *388*, S36.
- [52] Chen, X.; Liu, C.; He, H.; Chang, X.; Jiang, Y.; Li, Y.; Duan, M.; Li, J.; Luo, C.; Yao, D. Transdiagnostic differences in the resting-state functional connectivity of the prefrontal cortex in depression and schizophrenia. *J. Affect. Disord.* **2017**, *217*, 118–124.
- [53] Kaufmann, T.; Alnæs, D.; Brandt, C.L.; Doan, N.T.; Kauppi, K.; Bettella, F.; Lagerberg, T.V.; Berg, A.O.; Djurovic, S.; Agartz, I.; et al. Task modulations and clinical manifestations in the brain functional connectome in 1615 fMRI datasets. *Neuroimage* **2017**, *147*, 243–252.
- [54] Guo, W.; Liu, F.; Chen, J.; Wu, R.; Li, L.; Zhang, Z.; Zhao, J. Family-based case-control study of homotopic connectivity in first-episode, drug-naïve schizophrenia at rest. *Sci. Rep.* **2017**, *7*, 43312.
- [55] Iwabuchi, S.J.; Palaniyappan, L. Abnormalities in the effective connectivity of visuothalamic circuitry in schizophrenia. *Psychol. Med.* **2017**, *47*, 1300–1310.
- [56] Yang, Y.; Cui, Y.; Xu, K.; Liu, B.; Song, M.; Chen, J.; Wang, H.; Chen, Y.; Guo, H.; Li, P.; et al. Distributed functional connectivity impairment in schizophrenia: A multi-site study. In Proceedings of the 2nd IET International Conference on Biomedical Image and Signal Processing (ICBISP 2017), Wuhan, China, 13–14 May 2017; IET: London, UK, 2017; pp. 1–6.

[57] Bae, Y.; Kumarasamy, K.; Ali, I.M.; Korfiatis, P.; Akkus, Z.; Erickson, B.J. Differences between schizophrenic and normal subjects using network properties from fMRI. *J. Digit. Imaging* **2018**, *31*, 252–261.

[58] Li, J.; Sun, Y.; Huang, Y.; Bezerianos, A.; Yu, R. Machine learning technique reveals intrinsic characteristics of schizophrenia: An alternative method. *Brain Imaging Behav.* **2019**, *13*, 1386–1396.

[59] Chatterjee, I.; Kumar, V.; Sharma, S.; Dhingra, D.; Rana, B.; Agarwal, M.; Kumar, N. Identification of brain regions associated with working memory deficit in schizophrenia *F1000Research* **2019**, *8*, 124.

[60] Kalmady, S.V.; Greiner, R.; Agrawal, R.; Shivakumar, V.; Narayanaswamy, J.C.; Brown, M.R.; Greenshaw, A.J.; Dursun, S.M.; Venkatasubramanian, G. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *NPJ Schizophr.* **2019**, *5*, 1–11.

[61] Kirihara, K., Tada, M., Koshiyama, D., Fujioka, M., Usui, K., Araki, T. and Kasai, K., 2020. A predictive coding perspective on mismatch negativity impairment in schizophrenia. *Frontiers in psychiatry*, *11*, p.660.

[62] Gorbachevskaya, K. and Borisov, S., 2019. Eeg of healthy adolescents and adolescents with symptoms of schizophrenia.

[63] Olejarczyk, E. and Jernajczyk, W., 2017. EEG in schizophrenia. *RepOD*.

[64] Seglen, P.O., 1992. The skewness of science. *Journal of the American society for information science*, *43*(9), pp.628-638.

[65] Balanda, K.P. and MacGillivray, H.L., 1988. Kurtosis: a critical review. *The American Statistician*, *42*(2), pp.111-119.

[66] Mela, C.F. and Kopalle, P.K., 2002. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics*, *34*(6), pp.667-677.

- [67] Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), pp.1937-1967.
- [68] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
- [69] Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W. and Zhou, H., 2019. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *Journal of Hydrology*, 574, pp.1029-1041.
- [70] Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S. and Sun, Y., 2021. Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. *Forests*, 12(2), p.216.
- [71] Ogunleye, A. and Wang, Q.G., 2019. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6), pp.2131-2140.
- [72] Dhaliwal, S.S., Nahid, A.A. and Abbas, R., 2018. Effective intrusion detection system using XGBoost. *Information*, 9(7), p.149.
- [73] Ren, X., Guo, H., Li, S., Wang, S. and Li, J., 2017. A novel image classification method with CNN-XGBoost model. In *Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings 16* (pp. 378-390). Springer International Publishing.
- [74] Al Daoud, E., 2019. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1), pp.6-10.
- [75] Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H. and Rehman, M.U., 2019. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting. *Ieee Access*, 7, pp.28309-28318.
- [76] Wang, Y., Chen, J., Chen, X., Zeng, X., Kong, Y., Sun, S., Guo, Y. and Liu, Y., 2020. Short-term load forecasting for industrial customers based on TCN-LightGBM. *IEEE*

Transactions on Power Systems, 36(3), pp.1984-1997.

[77] Liang, W., Luo, S., Zhao, G. and Wu, H., 2020. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics*, 8(5), p.765.

[79] Baby, D., Devaraj, S.J. and Hemanth, J., 2021. Leukocyte classification based on feature selection using extra trees classifier: A transfer learning approach. *Turkish Journal of Electrical Engineering and Computer Sciences*, 29(8), pp.2742-2757

[80] Kaur, K. and Mittal, S.K., 2020. Classification of mammography image with CNN-RNN based semantic features and extra tree classifier approach using LSTM. *Materials Today: Proceedings*.

[81] Sharma, D., Kumar, R. and Jain, A., 2022. Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. *Measurement: Sensors*, 24, p.100560.

[82] Lusa, L., 2017. Gradient boosting for high-dimensional prediction of rare events. *Computational Statistics & Data Analysis*, 113, pp.19-37.

[83] Bahad, P. and Saxena, P., 2020. Study of adaboost and gradient boosting algorithms for predictive analytics. In *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019* (pp. 235-244). Springer Singapore.

[84] Punmiya, R. and Choe, S., 2019. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. *IEEE Transactions on Smart Grid*, 10(2), pp.2326-2329.

[85] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019, July. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

[86] Yacouby, R. and Axman, D., 2020, November. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).