Predicting Psychiatric Disorders Using EEG Data: A Machine Learning Approach

Abstract

This study investigates the classification of mental illnesses from EEG data using machine learning. We assessed four supervised algorithms: K-Nearest Neighbors, Random Forest, Support Vector Machine, and Logistic Regression on a Kaggle dataset. Focusing on learning curves, confusion matrices, test accuracy, and cross-validation scores, the Random Forest algorithm showed superior performance with 88% test accuracy, highlighting its potential for clinical use, and emphasizing the importance of machine learning in improving psychiatric treatment and diagnostic precision.

Keywords—component;machine learning,classification,EEG,K-NearestNeighbour , Psychiatric Disorders, classification, Random Forest, Support Vector Machine, Logistic Regression.

The source data and Jupyter Notebooks used for the analysis can be accessed here:

word count: 2228

I. Introduction

One of the biggest challenges in providing mental health care is accurately diagnosing and categorizing psychiatric diseases. Conventional approaches frequently rely on subjective patient reports of symptoms and clinical evaluations, which might differ greatly amongst practitioners. Patient outcomes could be negatively affected because of this inconsistency, which may lead to inaccurate diagnosis and treatments.

This study aims to improve the diagnostic accuracy of psychiatric disorders by using EEG data; a non-invasive method of recording brain activity, to capture brain activity. In order to find patterns that indicate mental disorders, this study analyzes complex, high-dimensional EEG data using machine learning algorithms. The selection of machine learning is based on its ability to process big datasets, identify underlying patterns, and produce accurate predictions.

- Problem statement

This study examines the challenge of classifying psychiatric disorders using EEG data. Psychiatric disorders, including schizophrenia, bipolar disorder, and others, are reflected in EEG signals. the key challenge is accurately differentiating these disorders using the high-dimensional EEG signals data. The model must handle noise and variability in the data and provide interpretable results that can aid in clinical decision-making.

- Choosing learning algorithms

Four learning algorithms had been evaluated: KNN, Random Forest, SVM, and Logistic Regression. they were chosen to classify psychiatric disorders using EEG data due to their unique benefits and applicability for different aspects of the problem.

further information regarding the particular benefits and functionality of each algorithm will be covered Later in the study.

II. Exploratory Data Analysis

A. Data overview

The project's dataset, which was obtained from Kaggle, addresses EEG recordings connected to mental illnesses. It contains 919 observations and 1,149 attributes in it. The dataset includes a wide range of characteristics, including demographic data (age, IQ, sex, and level of education), disorder information; main disorder (e.g., addictive disorder, mood disorder, healthy control) and specific disorder (e.g., alcohol use disorder, bipolar disorder) , and EEG features (different measurements taken from different brain regions, which includes delta, theta, alpha, and gamma wave features). The collection also contains correlation features, which show the cohesiveness levels among various brain areas.

B. Data cleaning

The process of data cleaning includes handling missing values. Particularly, there were some missing values in the education and IQ columns with 15 missing values in the education column and 13 in the IQ column, it was addressed by removing the rows with the missing values. Additionally, 919 missing data were also found in an unnamed column, which was eventually eliminated because it had no significance on our analysis.

Given that most machine learning algorithms depend on numerical data to function, these categorical features were encoded numerically using Label Encoding. Through this method, the categorical variables have been transformed into integer values while retaining their unique categories but in a numerical format that algorithms can process effectively.

A standardization of the dataset has been done to ensure that each attribute had an equal contribution to the analysis. In this step, the data was scaled using MinMaxScaler, which converted the numerical features into a standard range of 0 to 1. Additionally, due to the high dimensionality of the dataset, we performed Data Scaling and Feature Selection to have better model performance. With 1,149 properties, the dataset was highly dimensional, therefore it was essential to narrow down the number of features down. Using this method, the top k features have been selected by evaluating the features according to their statistical significance relative to the target variable ('main.disorder'). The top 50 most essential traits were selected in this instance. By deleting features that are redundant or unnecessary and do not increase the model's predictive capacity, this selection method helps to improve model performance.

Principal Component Analysis (PCA) was taken into consideration as a dimensionality reduction method to improve the machine learning models' effectiveness even more. PCA seeks to maximize data variance while minimizing the number of features. However, after applying PCA, it was shown that the models' accuracy reduced This drop in performance indicated that some important information required for precise predictions might have been removed after applying PCA.as a result, we chose not to use it for the model training and evaluation.

These data cleaning and preprocessing made the dataset more effective to start with the machine learning modelling giving more accurate and reliable predictions of psychiatric disorders based on EEG data.

C. Data visualisation

Figure 1 displays the correlation matrix for the top 5 features selected from the dataset. This matrix highlights the relationships between different features by showing their correlation coefficients. Notably, the correlation between "AB.C.alpha.b.FP2" and "AB.C.alpha.a.FP1" is the highest, almost reaching 1, indicating a very strong positive correlation. Conversely, the features "F" and "M" show a perfect negative correlation of -1, as expected since they represent binary gender categories. The other features, such as "COH.C.alpha.e.Fz.r.O1," reveal relatively low correlations with the other features, indicating minimal linear relationships. This visualization helps in understanding the degree of association between various features in the dataset.

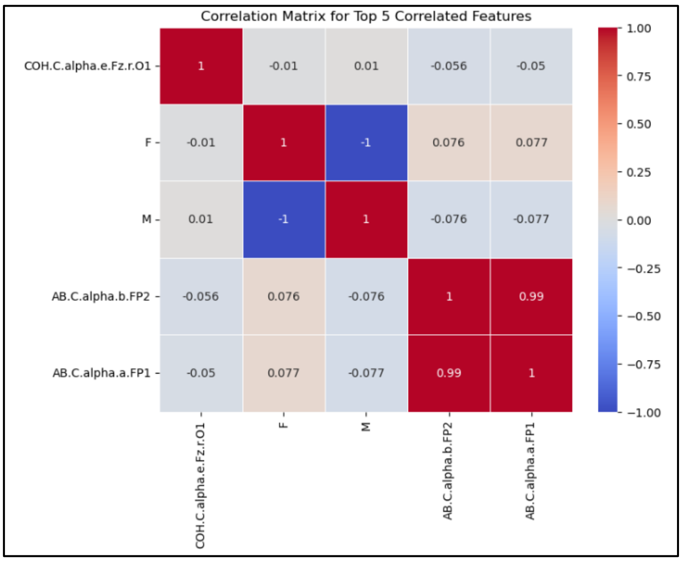


Fig. 1. Correlation Matrix for Top 5 Features.

Figure 2 demonstrates bar plots showing the distribution of the main psychiatric disorders in the dataset, such as obsessive-compulsive disorder, bipolar disorder, personality disorder, anxiety disorder, addictive disorder, and schizophrenia. With over 250 cases, schizophrenia is the most frequent psychiatric disorder. Anxiety Disorder comes in second place, with about 175 cases. Bipolar disorder has roughly 120 cases, Addictive disorder, and Personality disorder about 125 each. There are about 100 cases of Depressive Disorder, and there are less than 50 cases of obsessive-compulsive disorder, which is the least common. This distribution demonstrates how common schizophrenia and anxiety disorder are in the dataset.

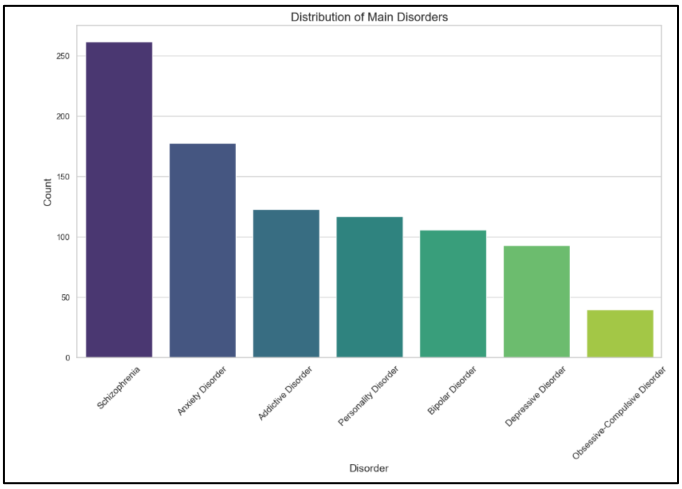


Fig. 2. Distribution of Main Psychiatric Disorders in the Dataset

III. Bivariate and Multivariate analysis

1) *Bivariate analysis*

Bivariate analysis focuses into how two variables are related to one another. Its main purpose is to identify the two variables' actual relationship and level of connection. Finding connections, trends, and patterns between them can be enhanced by this kind of study. Bivariate analysis is commonly performed using, correlation coefficients, and scatter plots, etc.

2) *Multivariate analysis*

On the other hand, multivariate analysis involves multiple analyses of more than two variables. It is used to comprehend the complex relationships that exist between multiple variables and how those relationships affect the desired result. This kind of analysis can provide a deeper knowledge of the data structure and highlight more complex relationships.

3) *Analysis type chosen*

In this study, multivariate analysis is more suitable due to the high-dimensional nature of the EEG data. EEG datasets are often made up of multiple features capturing a broad range of brainwave activity in various frequencies and regions. The multivariate approach makes it possible to take into consideration each of these variables simultaneously, which proves essential for comprehending the complicated interactions that occur in EEG data.

IV. Learning algorithm selection

A. *Supervised learning*

In this study, the best method for utilizing EEG data to diagnose psychiatric disorders has been found to be supervised learning. This decision is justified based on the nature of the dataset and the goals of the analysis. The labelled EEG data used in this study are observations that are related to a particular psychiatric disorder. Supervised learning makes sense since each instance in the dataset has clear, pre-defined labels. Building a model that can classify new, unseen EEG data into a predetermined category of psychiatric disorders is the main objective of the analysis. This kind of task is well suited for supervised learning algorithms, which use labeled training data to teach how to map input features (EEG signals) to corresponding output labels (psychiatric disorders).

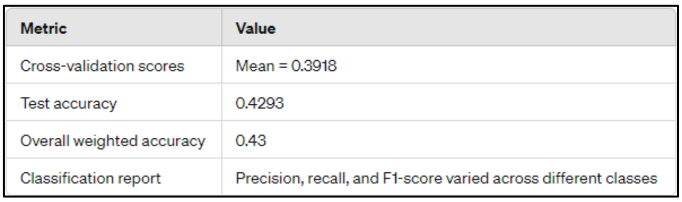
1) *Choice of algorithms*

For this study, several supervised learning algorithms were chosen and put into practice, each of which has unique benefits for the classification task:

- **K-Nearest Neighbors (KNN):** This technique was chosen because it is easy to use and efficient when solving multi-class classification problems. KNN is especially helpful in comprehending the data's local structure.
- **Random Forest:** Selected due to its ensemble learning capability and capacity to manage high-dimensional data and produce reliable predictions. Given the complexity of EEG data, Random Forest can also help in deciding the significance of different aspects.
- **Support Vector Machine (SVM):** it has been implemented Because it is so good at determining the best decision limit across classes.
- **Logistic Regression:** used due to its ease of use and interpretation. It offers a probabilistic framework that makes it easier to comprehend how various features affect the classification result. Equations

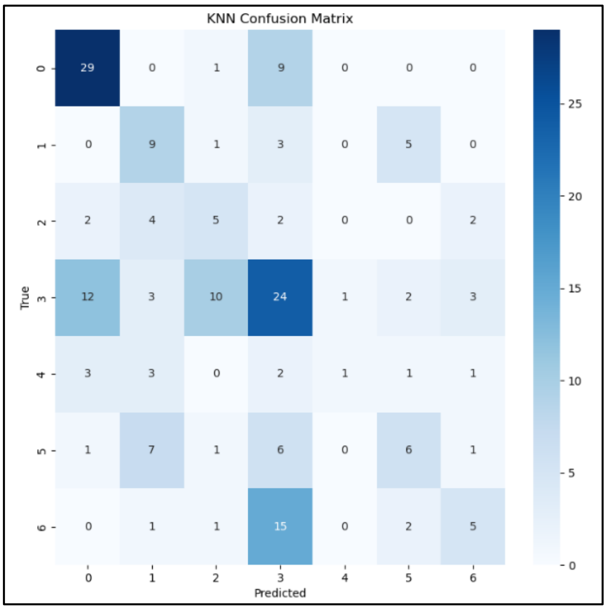
2) *Implementation and evaluation(Model's performance)*

a) *K-Nearest Neighbors (KNN)*



Metric	Value
Cross-validation scores	Mean = 0.3918
Test accuracy	0.4293
Overall weighted accuracy	0.43
Classification report	Precision, recall, and F1-score varied across different classes

Fig. 3. Performance Matrix for K-Nearest Neighbors (KNN) Model



True \ Predicted	0	1	2	3	4	5	6
0	29	0	1	9	0	0	0
1	0	9	1	3	0	5	0
2	2	4	5	2	0	0	2
3	12	3	10	24	1	2	3
4	3	3	0	2	1	1	1
5	1	7	1	6	0	6	1
6	0	1	1	15	0	2	5

Fig. 4. Confusion matrix for KNN Model

The confusion matrix reveals that KNN had difficulty distinguishing between certain classes, specifically between elements 3 (Schizophrenia) and 6 (Addictive Disorder), leading to significant misclassifications.

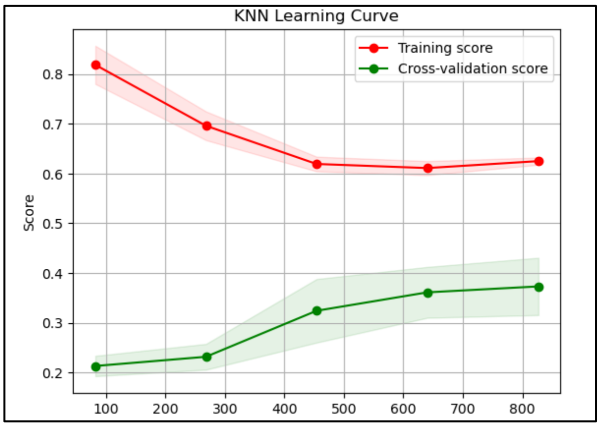
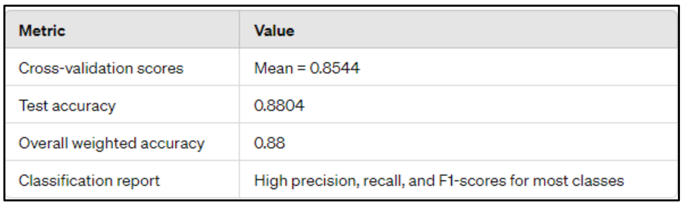


Fig. 5. Learning curve for KNN Model

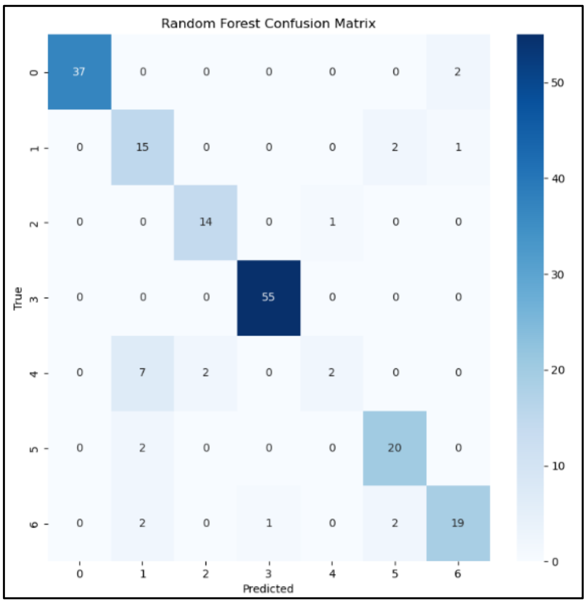
The learning curve for KNN indicates high bias, suggesting the model underfits the data. The training score is higher than the cross-validation score, which gradually improves with more data but remains low overall.

b) Random forest



Metric	Value
Cross-validation scores	Mean = 0.8544
Test accuracy	0.8804
Overall weighted accuracy	0.88
Classification report	High precision, recall, and F1-scores for most classes

Fig. 6. **Table 2:** ***Performance Matrix for Random Forest Model***



True \ Predicted	0	1	2	3	4	5	6
0	37	0	0	0	0	0	2
1	0	15	0	0	0	2	1
2	0	0	14	0	1	0	0
3	0	0	0	55	0	0	0
4	0	7	2	0	2	0	0
5	0	2	0	0	0	20	0
6	0	2	0	1	0	2	19

Fig. 7. Confusion matrix for Random Forest Model

The confusion matrix shows that Random Forest performed exceptionally well, correctly classifying most instances across all classes.

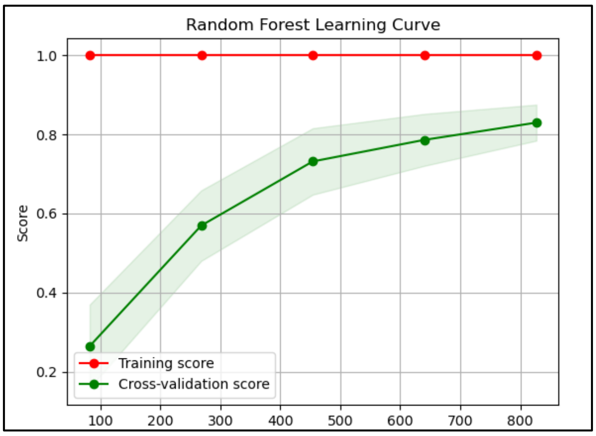


Fig. 8. Learning curve for Random Forest Model

The learning curve indicates low bias and low variance. The training score remains near perfect while the cross-validation score is also high, indicating a well-generalized model.

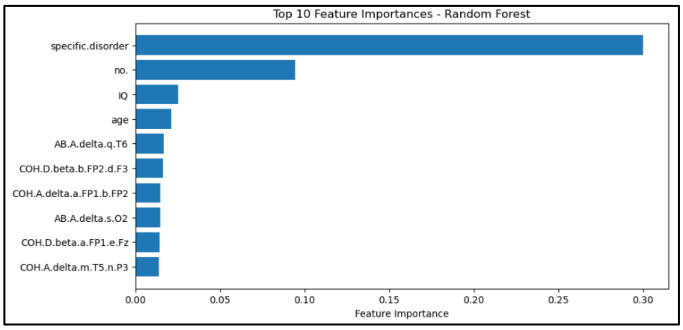
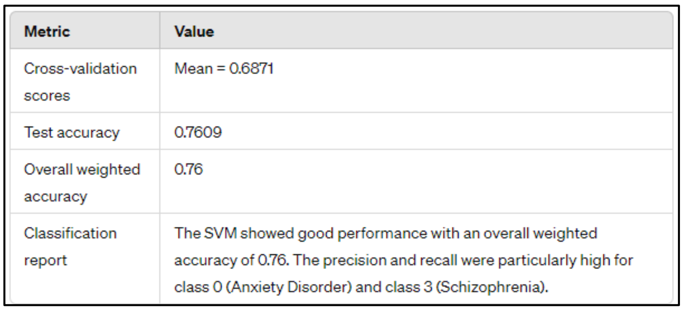


Fig. 9. Feature Importances from Random Forest Model

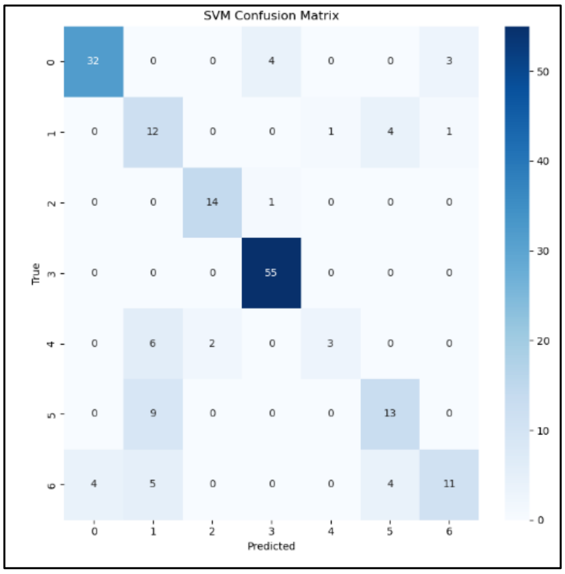
The top 10 features indicate that 'specific. Disorder' and 'no.' were the most important, highlighting the model's ability to leverage significant features effectively.

c) Support Vector Machine (SVM)



Metric	Value
Cross-validation scores	Mean = 0.6871
Test accuracy	0.7609
Overall weighted accuracy	0.76
Classification report	The SVM showed good performance with an overall weighted accuracy of 0.76. The precision and recall were particularly high for class 0 (Anxiety Disorder) and class 3 (Schizophrenia).

Fig. 10. Performance Metrics for Support Vector Machine Model



	0	1	2	3	4	5	6
0	32	0	0	4	0	0	3
1	0	12	0	0	1	4	1
2	0	0	14	1	0	0	0
3	0	0	0	55	0	0	0
4	0	6	2	0	3	0	0
5	0	9	0	0	0	13	0
6	4	5	0	0	0	4	11

Fig. 11. Confusion matrix for SVM Model

The SVM confusion matrix demonstrates moderate success in classifying most classes, with some confusion between classes like 3 (Schizophrenia) and 6 (Addictive Disorder).

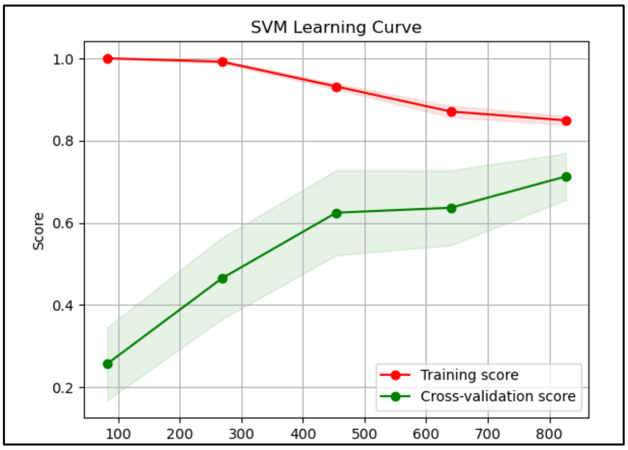
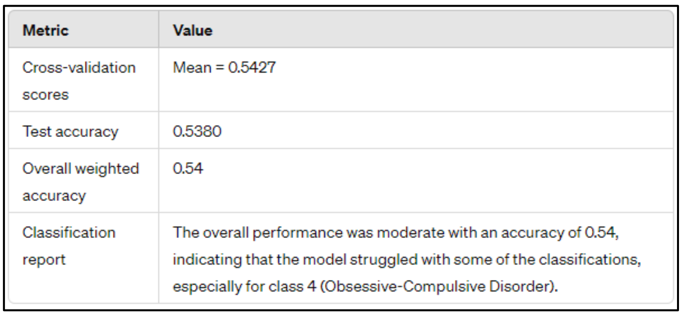


Fig. 12. Learning curve for SVM Model

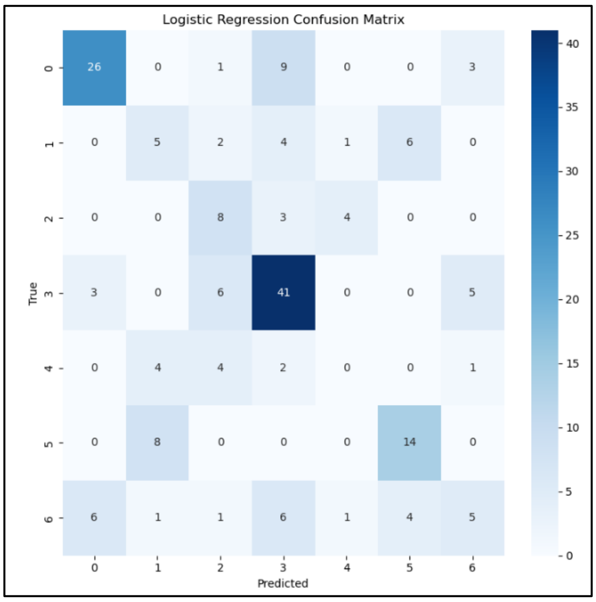
The SVM learning curve shows a decreasing trend in training score as more examples are added, while the cross-validation score improves, indicating the model benefits from more data and is approaching a more generalizable state.

d) Logistic Regression



Metric	Value
Cross-validation scores	Mean = 0.5427
Test accuracy	0.5380
Overall weighted accuracy	0.54
Classification report	The overall performance was moderate with an accuracy of 0.54, indicating that the model struggled with some of the classifications, especially for class 4 (Obsessive-Compulsive Disorder).

Fig. 13. Table 4: Performance Metrics for Logistic Regression Mode



	0	1	2	3	4	5	6
0	26	0	1	9	0	0	3
1	0	5	2	4	1	6	0
2	0	0	8	3	4	0	0
3	3	0	6	41	0	0	5
4	0	4	4	2	0	0	1
5	0	8	0	0	0	14	0
6	6	1	1	6	1	4	5

Fig. 14. Confusion matrix for Logistic Regression Model

The confusion matrix for Logistic Regression shows that it was reasonably good at classifying some disorders, like 0 (anxiety disorder) and 3 (Schizophrenia) but had significant misclassifications in others.

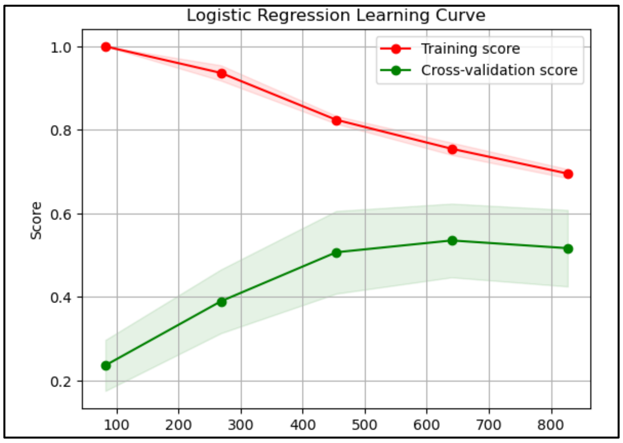
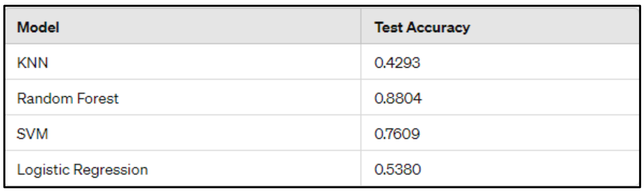


Fig. 15. Learning curve for Logistic Regression Model

The learning curve for Logistic Regression highlights high bias and low variance. The training score is higher than the cross-validation score, which suggests that the model is underfitting.

After evaluating the performance of each model individually, it is crucial to compare their results collectively to draw comprehensive conclusions about their effectiveness in diagnosing psychiatric disorders using EEG data. The comparative analysis provides a clearer understanding of which models perform best and under what conditions.



Model	Test Accuracy
KNN	0.4293
Random Forest	0.8804
SVM	0.7609
Logistic Regression	0.5380

Fig. 16. Test accuary for each model

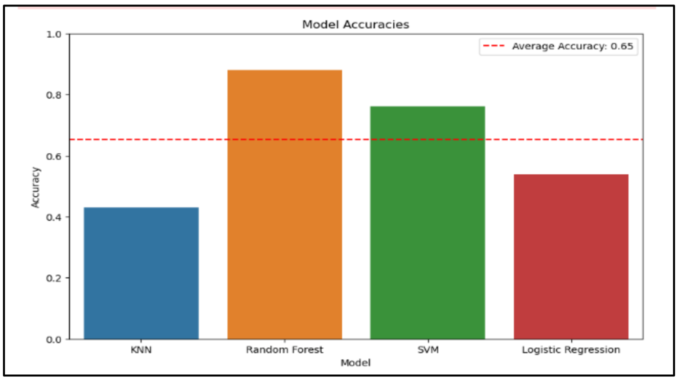


Fig. 17. Model's accuracy comparision using bar plots

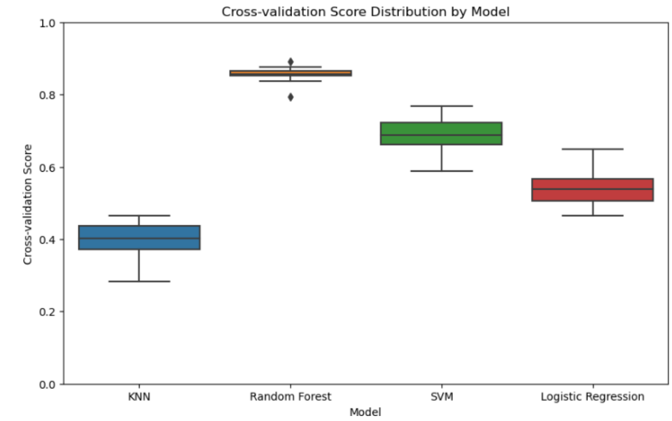


Fig. 18. Cross vaidation distribution box plots

According to fig 17 and 18 ,The Random Forest model was the most effective choice in this analysis because of its strong ability to handle high-dimensional data and good accuracy when diagnosing psychiatric diseases using EEG data. Strong performance was also demonstrated by the Support Vector Machine (SVM), which handled the complicated EEG data with ease. Using confusion matrices, test accuracy, and cross-validation scores, both models were thoroughly assessed to reveal their strengths and weaknesses. These assessments emphasized the significance of accurate mental diagnosis while illuminating the efficacy of each model in clinical settings.

e) Rejected Algorithm

The K-Nearest Neighbors (KNN) approach was shown to be less appropriate in the evaluation of machine learning models for psychiatric disorders prediction from EEG data. This was mostly caused by the dataset's high dimensionality, which has had a negative impact on KNN's effectiveness. Distances between points can become misleading in high-dimensional spaces, which makes it challenging for KNN to reliably detect and classify the closest neighbors. In comparison to other models examined in this work, the KNN algorithm's inherent limitations in processing complicated high-dimensional data, such as EEG signals, resulted in reduced accuracy and effectiveness. This significant restriction was demonstrated by the continually lower accuracy of KNN in our experiments, which led to its rejection in favor of more reliable models that could better handle the complexity of the dataset. Evaluation of models for machine learning
Furthermore, for future research considering deep learning methods like Convolutional Neural Networks (CNNs) may be helpful because of their capacity to manage high-dimensional data and automatically extract features, which may enhance the predictive accuracy and robustness of EEG-based psychiatric disorder classification.

References

[1] Shagass, C., Roemer, R.A. and Straumanis, J.J., 1982. Relationships between psychiatric diagnosis and some quantitative EEG variables. *Archives of general psychiatry*, *39*(12), pp.1423-1435.

[2] Cohen, M.X., 2017. Where does EEG come from and what does it mean?. *Trends in neurosciences*, *40*(4), pp.208-218.

[3] Hosseinifard, B., Moradi, M.H. and Rostami, R., 2013. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer methods and programs in biomedicine*, *109*(3), pp.339-345.

[4] Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[5] Park, S.M., Jeong, B., Oh, D.Y., Choi, C.H., Jung, H.Y., Lee, J.Y., Lee, D. and Choi, J.S., 2021. Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach. Frontiers in Psychiatry, 12, p.707581.

[6] Mumtaz, W., Ali, S.S.A., Yasin, M.A.M. and Malik, A.S., 2018. A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD). Medical & biological engineering & computing, 56, pp.233-246.

[7] Guerrero, M.C., Parada, J.S. and Espitia, H.E., 2021. EEG signal analysis using classification techniques: Logistic regression, artificial neural networks, support vector machines, and convolutional neural networks. Heliyon, 7(6).

[8] Edla, D.R., Mangalorekar, K., Dhavalikar, G. and Dodia, S., 2018. Classification of EEG data for human mental state analysis using Random Forest Classifier. Procedia computer science, 132, pp.1523-1532.