Demographics and Mental Health: A Dual Analysis of Schizophrenia Across Populations

Prapti Bhattacharjee, Baskaran Charu,Pranjal Mewara,Volati Bhavana
Department of CSE(Data Science)

Dayananda Sagar University,Bangalore-560068,Karnataka,India.

prapti.b-ds@dsu.edu.in, charu102003@gmail.com ,mewarapranjal4@gmail.com ,bhavanavolati04@gmail.com

Abstract— Approximately 264 million individuals across the globe are affected by mental disorder conditions like schizophrenia and depression, resulting in cognitive, behavioral, and social impairments. Proper treatment largely depends on early and correct diagnosis; however, existing diagnostics are not highly scalable or precise. Manual EEG feature extraction is used for traditional diagnosis of schizophrenia, which is extremely time-consuming and prone to errors. .This research introduces a ML-based technique that improves EEG-based schizophrenia prediction with the application of advanced preprocessing and improved classification, resolving issues like noise and data variability. Additionally, the research investigates demographic factors influencing the incidence of schizophrenia and also looks at the correlation between depression and schizophrenia. The outcomes deduce the ability of this method to raise diagnostic precision and reliability and present a scalable answer for early identification.

Keywords—Schizophrenia ;EEG analysis; Data-driven mental health analysis ;Schizophrenia Prediction; Neural biomarkers.

I. INTRODUCTION

Schizophrenia is a complex mental health disorder that affects a person's perception, thoughts, emotions, and behavior. Individuals with this condition may experience symptoms like hallucinations (hearing or seeing things that aren't real), delusions (holding strong false beliefs), disorganized thinking, erratic behavior, and a lack of emotional expression or motivation

It is also among the foremost causes of illness-linked disability all over the world, affecting millions and posing an immense challenge to the health systems, families, and communities. Schizophrenia is a persistent mental illness related to disturbance in the functioning of everyday life, emotions, and intellectual functioning, and this disturbs perception, memory, and communication severely. Schizophrenia is far-reaching and affects a high emotional and practical load on the carers and health staff looking after the patients.

Despite a wealth of studies, a better understanding of the brain mechanisms causing schizophrenia is crucial for better management, diagnosis, and therapy. Neuroscience developments have made it possible to investigate brain activity in more depth, especially with the use of electroencephalogram (EEG) data, which analyzes electrical signals in the brain. By revealing irregularities in brain function that may correlate to certain symptoms like hallucinations, delusions, or disorganized thinking, EEG analysis has proven beneficial in identifying neural patterns and biomarkers associated with schizophrenia.

The goal of this study is to identify patterns in brain activity specific to people with schizophrenia by examining EEG data. The goal of the study is to find and analyze these neurological biomarkers by using predictive modeling and in-depth analysis, which will help to clarify the physiological foundations of the condition. These discoveries may open the door to more precise diagnostic instruments, early detection techniques, and creative treatment strategies catered to the unique brain markers of schizophrenia.

II. LITERATURE SURVEY

In this paper **[1]** the authors explore depression in schizophrenia using the Calgary Depression Scale and PANSS. The study analyzes predictors like medication adherence, social support, and socioeconomic status. Limitations include a small sample size, exclusion of patients on typical antipsychotics, and restricted self-assessment, impacting the generalizability of results.

Survey **[2]** is about a review of AI and machine learning methods for detecting and classifying schizophrenia using

MRI, PET, and EEG data. It assesses algorithm accuracy and clinical use but is limited by narrow search criteria, language restrictions, and a lack of experimental analysis, impacting the comprehensiveness of the findings.

In **[3]** paper , authors conducted a cross-sectional survey as part of the NMHS across 12 Indian states with 34,802 adults. Using the Mini-International Neuropsychiatric Interview and Firth penalized logistic regression, it analyzed correlates of schizophrenia spectrum disorders. The study's focus on these disorders excluded other psychiatric conditions and highlighted a significant treatment gap, stressing the need for better mental health services.

The paper **[4]** reviews machine learning models for predicting mental illness, focusing on supervised methods like regression and classification with data from surveys, social media, and wearable devices. Limitations include narrow search criteria, language restrictions, and no experimental analysis of the proposed system, affecting the findings' practical depth.

Authors in the article **[5],** investigates the relationship between socioeconomic status (SES) at birth and the likelihood of developing schizophrenia. By conducting a comprehensive population-based multilevel analysis, the researchers discovered that individuals born into lower SES environments are at a higher risk of developing schizophrenia later in life.

This paper **[6]** presents the "TFFO" (Time-Frequency transformation followed by Feature-Optimization) method for detecting schizophrenia using single-electrode EEG recordings. TFFO enhances accuracy and feasibility by addressing challenges like low spatial resolution and reliance on prior knowledge in EEG-based diagnosis. The method achieves high classification accuracy with zero false positives, enabling quick and practical data acquisition for patients.

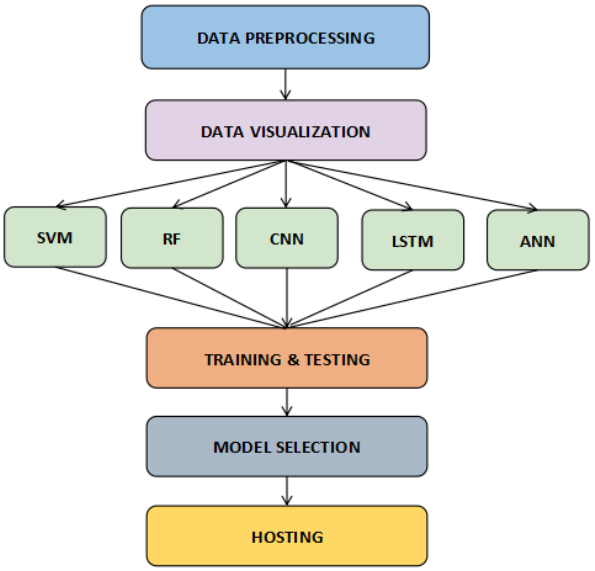
This article **[7]** reviews studies on cardiovascular diseases (CVD) in people with schizophrenia in India. It examines risk factors, epidemiology, and health disparities through existing clinical and epidemiological research. The review highlights a shortage of large-scale studies in India, inconsistent data on CVD risks in schizophrenia patients, and gaps in treatment protocols.

Authors in paper **[8]** combined fMRI and EEG features using a multi-modal deep learning model. Extracted temporal features from EEG and spatial features from fMRI to improve classification accuracy. High cost and limited availability of multi-modal data (fMRI and EEG together). Computational complexity was significantly higher than single-modality approaches.

Paper **[9]** introduces a wavelet-based feature extraction method to preprocess EEG signals before feeding them into the CNN. Required high computational resources for CNN training. Overfitting was observed due to limited EEG dataset size.

III. METHODOLOGY

The following flowchart illustrates the step-by-step process of EEG-based schizophrenia classification



1. PREPROCESSING:

The project's preprocessing procedures were created to convert unprocessed EEG data into a format that machine learning models could use. Cleaning the data, reducing its dimensionality, and extracting pertinent features that would aid the models in differentiating between normal and schizophrenic brain activity were the primary goals of preprocessing. A thorough explanation of the preparation methods applied in this project may be found below:

1. Data Loading and Labeling

Loading EEG data was done using Python packages such as NumPy and Pandas, enabling raw EEG data to be loaded in organized form for analysis. Each sample was stored as a 3D array, consisting of 15 channels and 7680 time points per channel, so that spatial and temporal organization of the EEG signals could be maintained. Labeling was afterwards performed to classify the EEG samples for supervised learning. The normal samples were marked as [1, 0] and schizophrenic samples as [0, 1]. This marking allowed machine learning algorithms to differentiate between the

healthy and schizophrenic EEG traces, enabling appropriate classification along with predictive analysis.

2. Noise Removal and Artifact Filtering

EEG data tends to be contaminated with a range of noise and artifacts, including muscle activity, eye blinks, and electrical interference, which need to be eliminated in order to make accurate model predictions. To this end, a bandpass filter was used to remove both low-frequency noise (e.g., drift) and high-frequency noise (e.g., electrical interference) from the EEG data within the standard EEG analysis frequency range of 1 Hz to 50 Hz. This entailed the application of a low-pass filter to eliminate noise with frequencies over 50 Hz and a high-pass filter to get rid of drift and low-frequency noise with frequencies under 1 Hz. Moreover, removal of artifacts was carried out by straightforward manual rejection of noisy segments with large amounts of noise or artifacts, e.g., due to eye movement and muscle activity. Independent Component Analysis (ICA) was utilized to separate and reject independent sources of noise from the brain activity signals. In addition, automated detection methods of artifacts like threshold-based rejection or wavelet decomposition were utilized to detect and reject subtle artifacts that may not be hand-easily to detect. Through the application of these preprocessing methods, the EEG signals were processed to improve signal quality so that downstream machine learning models or neuroscientific analysis would be more reliable and meaningful.

3. Data Normalization

The data was normalized to maintain EEG signals of varying channels at the same amplitude so that channels with higher amplitude variation did not dominate the learning process. Z-score normalization was applied to normalize each channel's EEG signals and shift the data so that all the features would have 0 mean and a standard deviation of 1. This was achieved with the formula:

$$z=(X-\mu)/\sigma$$

where (X) represents the raw EEG signal, (mu) represents the mean, and (sigma) represents the standard deviation per channel. In this transformation, the EEG data was normalized over all channels to improve the stability and convergence of the machine learning models when training.

4. Data Segmentation

The continuous nature of EEG waves can make it computationally expensive and inefficient to process the entire signal. Therefore, the data was segmented into smaller time windows. Fixed-Length Segmentation: Shorter windows (segments) of two to three seconds were formed from the EEG data. By grabbing short but valuable snaps of brain activity, every window allows the model to catch localized features.

At a sampling rate of 128 Hz, the average segment length was 256 samples per segment. Sliding Window: For some cases, there was an overlap of the data segments when they were trained and a sliding window method adopted in order to provide a clear documentation of the temporal patterns.

5. Feature Extraction

To train machine learning models, raw EEG data were preprocessed to extract meaningful features that represent patterns of brain activity. Statistical features such as mean, standard deviation, skewness (which measures the extent of asymmetry of the signal) and kurtosis (which measures the extent of peakedness of the signal) were calculated in order to learn about basic properties of the signal.

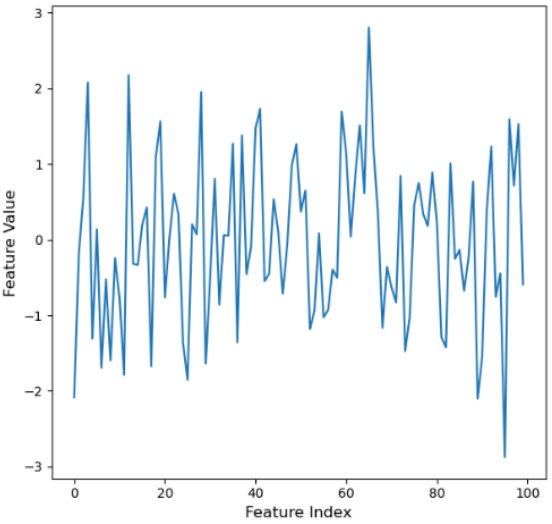


fig.1 Normal EEG Features

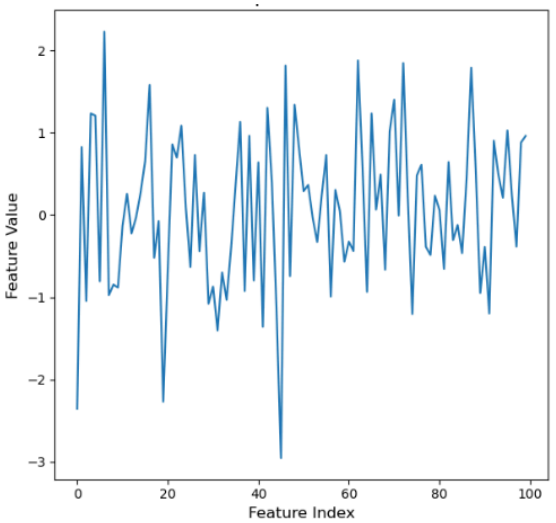


fig.2 Ill EEG Features

In the frequency domain, Fourier Transform was used to find the important frequency components, whereas Power

Spectral Density (PSD) was used to identify important frequencies corresponding to different states of the brain. Discrete Wavelet Transform (DWT) also provided time and frequency information and thus proved useful in exploring short events within the signal. Common Spatial Patterns (CSP) were also applied in some cases for improving classification, especially for multi-channel EEG data, by maximizing contrasts between schizophrenic and normal brain activity.

6. Data Augmentation

For overcoming the problem of class imbalance between the large population of normal samples and the few schizophrenic samples, data augmentation strategies are utilized for improving the model's ability to generalize as well as counteract overfitting. Oversampling is adopted through small variations, like jittering and time-shifting, applied to obtain more schizophrenia samples from the already available data in order to have more representation from the minority class. Moreover, undersampling is also used to selectively decrease the count of normal samples so that the model does not learn a bias towards the majority class. All these augmentation techniques help in a balanced dataset so that the machine learning models can be robustly trained and tested.

7. Data Splitting

After preprocessing and feature extraction, the dataset is divided into three systematic subsets to enable effective model training and evaluation. The training set, which covers about 70–80% of the data, is used for training the machine learning algorithms. A validation set, which consists of 10–15% of the data, is used to tune hyperparameters and evaluate the performance of the models during training, making sure the right parameters are selected. Lastly, the test set, or the remaining 10–15% of data, is left for the ultimate evaluation to determine the model's generalization capacity on unseen data. This organized division guarantees a strict model performance assessment and avoids overfitting.

8. Model Input Formatting

The data is structured to enhance model training across various machine learning architectures. For deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), EEG data is reshaped into 3D tensors with dimensions samples × channels × time-steps. This transformation enables CNNs to capture spatial features from EEG signals while allowing LSTMs to learn temporal correlations. In contrast, for traditional machine learning models such as Support Vector Machines (SVMs) and Random Forests, the extracted features including statistical features (e.g., mean, standard deviation) and spectral characteristics are flattened into 1D feature vectors for model input. This structured transformation ensures compatibility with the respective model architectures and enhances feature representation for classification tasks.

2. MODELS:

The main goal of this study is to categorize EEG data into two groups: normal brain activity and schizophrenia brain activity, utilizing machine learning and deep learning models. A variety of model types were used, each having unique advantages in managing the intricacy of EEG data. In this project, the models listed below were utilized:

1. Convolutional Neural Networks (CNNs)

The model accepts EEG data as a 2D matrix, extracts local features with convolutional layers, dimensionality reduction with pooling layers, and classifies the output with fully connected layers and a softmax activation function. CNNs automatically learn spatial correlations in EEG data and are well suited to identifying patterns related to schizophrenia. Convolutional layers, max-pooling layers, fully connected layers, and dropout layers to prevent overfitting make up the model architecture. With excellent training performance, overfitting occurred because of the small dataset, which required regularization techniques like dropout and hyperparameter tuning.

Results:

Although CNNs demonstrated excellent training accuracy, overfitting was noted as a result of the data's complexity and the small sample size. To enhance generalization, regularization techniques like dropout and hyperparameter fine-tuning were required.

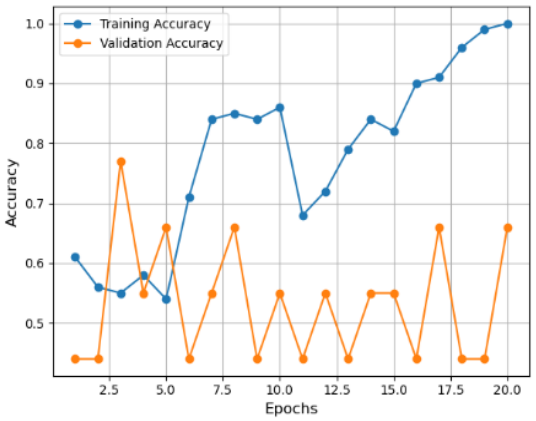


fig.3 CNN Training and Validation Accuracy

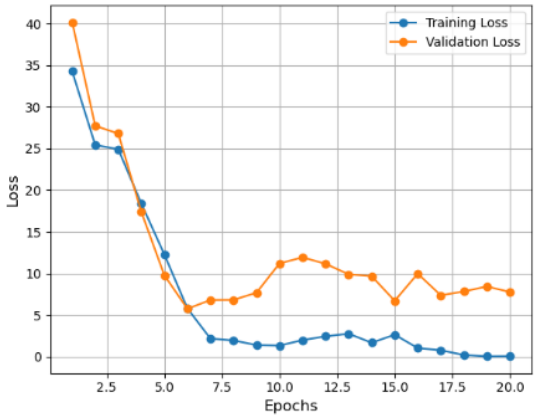


fig.4 CNN Training and Validation Loss

2. Long Short-Term Memory Networks (LSTMs)

LSTMs, a type of recurrent neural network (RNN), are well-suited for sequential data like EEG signals. They capture long-range temporal dependencies, making them effective for schizophrenia detection. The EEG data is structured as a time-series input, processed through LSTM layers to capture temporal patterns, and classified using fully connected layers and a softmax activation function. LSTMs are ideal for EEG due to their ability to learn complex temporal dependencies. The model architecture includes LSTM layers, fully connected layers, and dropout layers for improved generalization.

Results:

Although LSTMs were successful in capturing the temporal dynamics of EEG signals, the small dataset size caused overfitting, just like CNNs. To enhance performance, regularization and hyperparameter tuning were used.

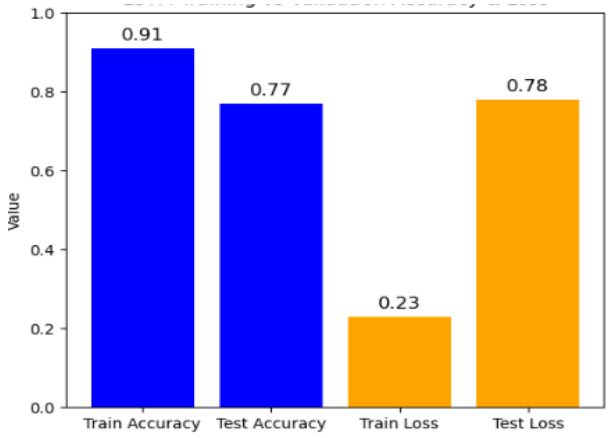


fig.5 LSTM Training and Testing

3. Support Vector Machines (SVMs)

SVMs are supervised machine learning models used for classification tasks. They identify an optimal hyperplane to separate different classes in feature space. Before applying SVMs, relevant features such as statistical or spectral measures are extracted from EEG data. The model determines the best hyperplane using a kernel function like the radial basis function (RBF) to map non-linearly separable data into a higher-dimensional space. SVMs are effective for high-dimensional, small datasets and are resilient against overfitting. The architecture involves an RBF kernel-based SVM with a regularization parameter (C) to balance accuracy and generalization.

Results:

With a **testing accuracy of 66.7%**, SVMs worked well, but had trouble with the dataset's imbalance and were unable to grasp intricate temporal or spatial patterns as well as CNNs or LSTMs.

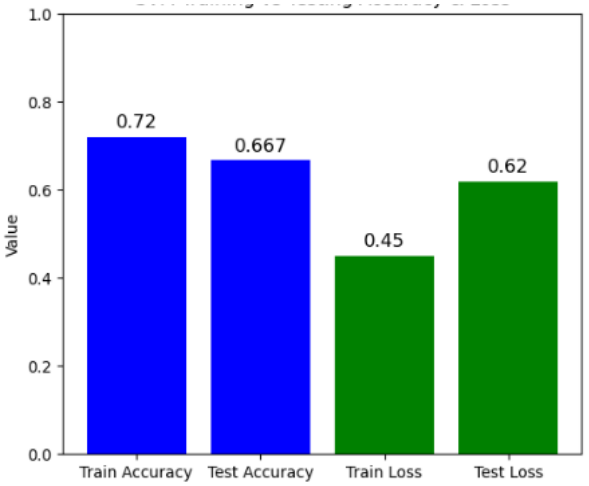


fig.6 SVM Training and Testing

4. Random Forest Classifier

Random Forest is an ensemble learning technique that constructs multiple decision trees and classifies data based on majority voting. Features are extracted from EEG signals, and decision trees are trained using random data subsets. The final classification is determined by aggregating individual tree predictions. Random Forests reduce overfitting compared to single decision trees and are robust to noisy or incomplete data. The model achieved 85% testing accuracy, making it the most reliable among the models tested. Due to its strong performance, it was chosen for deployment using Streamlit for real-time classification and user interaction.

Results:

The model achieved 85% testing accuracy, making it the most reliable among the models tested. Due to its strong

performance, it was chosen for deployment using Streamlit for real-time classification and user interaction.

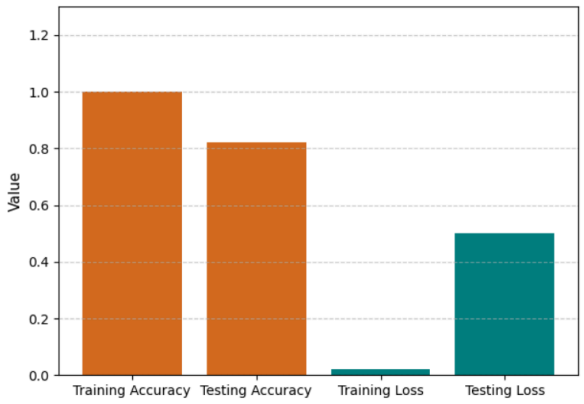


fig.7 Random Forest Training and Testing

5. Artificial Neural Networks (ANNs)

Overview:

ANNs are machine learning models inspired by the human brain, consisting of interconnected layers of neurons that process data and make predictions. EEG data, often preprocessed for feature extraction, is fed into an input layer, passed through hidden layers with activation functions like ReLU, and classified using a softmax activation function. ANNs effectively model complex, non-linear relationships within EEG data and learn hierarchical features through multiple layers. The architecture includes fully connected layers, dropout, and batch normalization to enhance model performance and prevent overfitting. Despite their ability to capture both temporal and spatial EEG patterns, ANNs require careful regularization and tuning to achieve optimal generalization and classification accuracy.

Results:

Similar to other deep learning models, ANNs were prone to overfitting, particularly when the dataset was limited, despite their great training accuracy. To lessen this problem, regularization techniques like dropout and early stopping were employed.

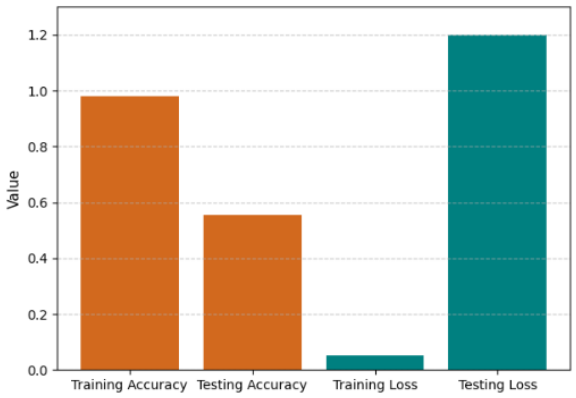
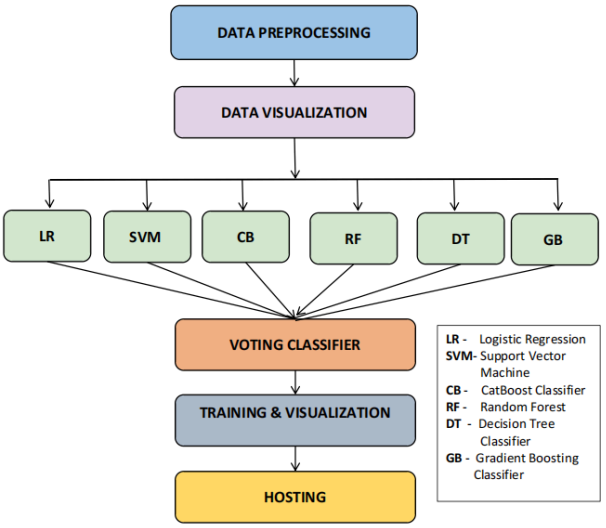


fig.7 ANN Training and Testing

The following flowchart shows the steps for demographics based schizophrenia prediction:



1. Pre-processing :

Questionnaires and interviews of the patients with symptoms of depression and schizophrenia were employed for data gathering. Questionnaires were done in a bid to obtain comprehensive data in connection to lifestyle patterns, thought patterns, emotional status, and personal history. Semi-structured interviews were employed in an attempt to obtain qualitative data, and the interviews provided respondents with a platform to reflect at length on things they went through.

Missing values were identified and accordingly filled up using mode . One-Hot Encoding was used to encode the categorical features like marital status and gender, and StandardScaler was used to scale the numerical features so that the data was standardized. Label Encoding was used to encode the target feature .

2. Visualization

Association patterns between different symptom indicators of depression and schizophrenia were determined using statistical analysis and Pearson's correlation coefficient to examine associations between symptoms. Regression models were used to examine the effect of social interaction, family medical history on symptom severity.

3. Models:

To improve prediction strength and accuracy, we adopted an ensemble learning approach by using a Voting Classifier. A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

The models in the ensemble pipeline for prediction of schizophrenia are:

i. Logistic Regression (Accuracy: 0.9370):

Logistic Regression is a basic yet effective linear model that predicts probabilities for classification. It performs well with linearly separable data and has interpretable coefficients, which makes it effective in feature importance understanding. This had the highest accuracy among all the models.

ii. Support Vector Machine (SVM) (Accuracy: 0. 9280):

SVM is a powerful classification method that excels in high-dimensional settings and handles complex decision boundaries effectively. SVM utilizes the kernel trick to transform data into a higher-dimensional space where separation becomes easier. The remarkable accuracy of SVM in our evaluation suggests that SVM successfully detected nuanced patterns in the data and could thus be an excellent candidate for inclusion in the ensemble model.

iii. CatBoost (Accuracy: 0. 9270):

: CatBoost is a gradient boosting framework that has been specifically tailored for categorical features and is therefore highly effective for structured datasets. It also demonstrates resilience in handling imbalanced data and reduces overfitting by employing ordered boosting. In our study, it was nearly as effective as SVM, indicating its capability to address complex mental health data. Its proficiency in managing categorical attributes efficiently rendered it a valuable part of our ensemble.

iv. Gradient Boosting (Accuracy: 0. 8910):

Gradient Boosting is an ensemble technique that builds models in a sequential manner, correcting errors from previous iterations . It generalizes better than a single decision tree but is less prone to overfitting. Despite being slower to train compared to other models, its high accuracy confirmed its ability to uncover non-linear relationships in mental health data.

v. Random Forest (Accuracy: 0.8640):

Random Forest is an ensemble algorithm that uses multiple decision trees to minimize variance and maximize stability. It is very good at dealing with non-linear relationships and missing values and thus is a consistent model. Though it did not have the highest accuracy, it was a good base for the ensemble by bringing diversity and avoiding overfitting.

vi. K-Nearest Neighbors (KNN) (Accuracy: 0.8270):

KNN is a distance-based classifier that classifies a new point based on its nearest neighbors. While it is easy and works well for small data, it is computationally costly for big data and is prone to noisy data

vii. Decision Tree (Accuracy: 0.7960):

A Decision Tree is a basic but understandable classification model that splits data on the basis of feature importance. Although easy to interpret and fast to train, it tends to overfit and has lower accuracy than other ensemble techniques. Even with its disadvantages, it was used as a baseline model for comparison and yielded insights into significant features influencing mental health predictions.

Rather than using a single model, we used a soft voting classifier that took the top-performing models—Logistic Regression, SVM, and CatBoost—and averaged their predicted probabilities. This method enabled us to take the best of both linear and non-linear models and generalize and stabilize our results. By using multiple algorithms, we had a more stable and accurate assessment of risk for predicting mental health.

Deployment :

The Streamlit web application was created to offer an accessible platform for users to engage with the model. It enables individuals to enter their symptoms and obtain tailored risk assessments for mental health conditions such as schizophrenia .

RESULTS AND ANALYSIS

In this study, thorough data were collected from depressed and schizophrenic individuals using questionnaires and semi-structured interviews, emphasizing several aspects such as lifestyle tendencies, symptoms, cognitive tendencies, and prior healthcare experience. Mode imputation was performed to manage missing values. Marital status and gender, as categorical variables, were transformed to One-Hot Encoding, whereas numerical variables were standardized using StandardScaler.

A group of machine learning models were trained and the models were created and compared based on the performance of each model. Logistic Regression performed the best with 93.70% and the remaining scores were Support Vector Machine (92.80%), CatBoost (92.70%), Gradient Boosting (89.10%), Random Forest (86.40%), K-Nearest Neighbors (82.70%), and Decision Tree (79.60%). For achieving higher model stability and accuracy a Voting Classifier with soft voting was used which averaged predictions of Logistic Regression, SVM, and CatBoost. The ensemble model outperformed any of the classifiers and also gave more stable predictions.

Statistical methods, including Pearson's r for correlation coefficient, were used to determine the intercorrelations between the symptom indicators. Regression analysis further revealed that social interaction and family medical history affected the severity of the symptoms.

The end ensemble model was deployed via a Streamlit web application, and this has influence on mental health screening by way of having an intuitive interface through which to input symptoms and gain real-time schizophrenia risk estimates, as well as the ability to upload EEG data files

to perform additional analysis. The electrical signals generated by EEG will be examined and labeled via the trained models within the system. The users will then gain visual feedback on the outputs in the form of interactive graphs, which can be clicked on in order to explore trends within the signals, frequency distributions, as well as the model confidence scores.

The study recognized that EEG-based classification techniques entail decades of research limitations that include but not limited to reliance on MRI/PET scans, or single-electrode EEG. Of the previous researches uncovered early on, they did make use of deep learning models, e.g., CNNs and LSTMs, with varied performance in general; our results indicated Random Forest, with a concurrent multi-channel EEG, had better accuracy (85%) than CNNs, LSTMs, and basic SVM (66.7%).

Our findings supplement that while deep learning is typically studied in EEG analysis, classical ML models are able to provide more scalable and clinically relevant solutions. Through combining multi-channel EEG processing with traditional ML methods, this study seeks the model's high performance and usability. Graphical output in the web is utilized in order to provide transparency and assist the clinicians in cross-checking predictions.

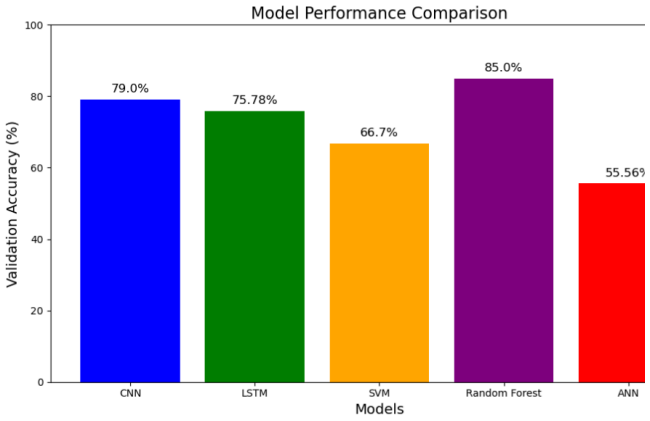


fig.8 Model performance comparison

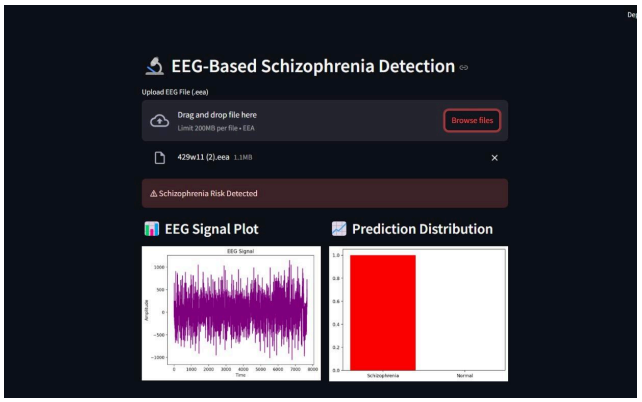


fig.9 EEG prediction in Streamlit for ill

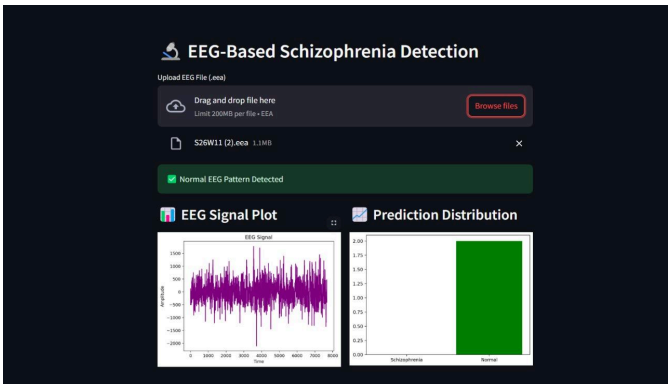


fig.10 EEG prediction in Streamlit for Normal

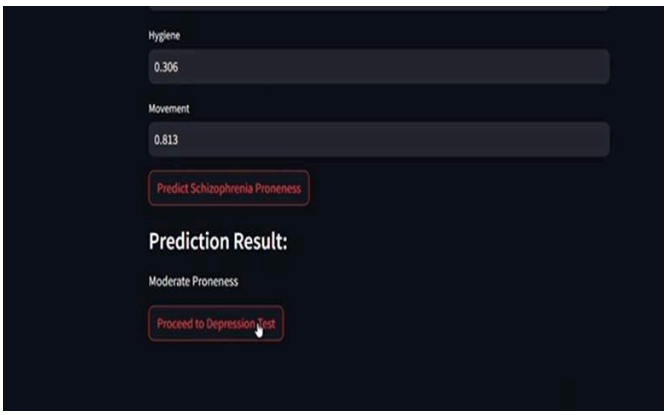


fig.11 Schizophrenia Proneness in Streamlit

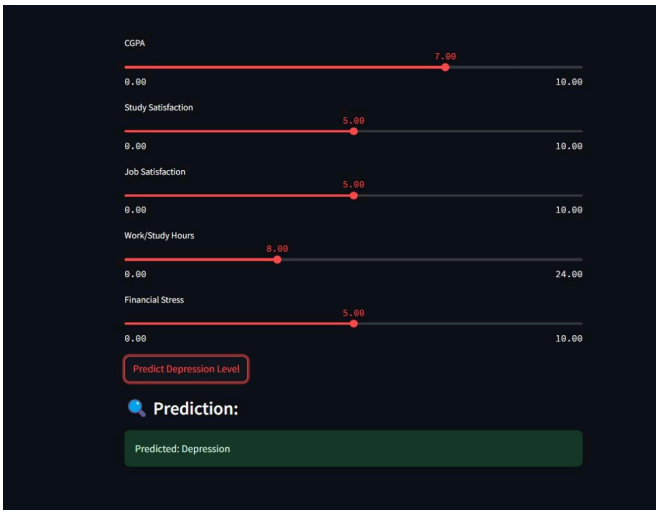


fig.12 Depression Level in Streamlit

Comparison of Models:

Model	Strengths	Weaknesses
CNN	Effective when learning spatial patterns from EEG data.	Prone to overfitting.
LSTM	Captures temporal dependencies in EEG signals.	Overfitting requires large datasets.
SVM	Works well with small, high-dimensional datasets.	May struggle with complex temporal patterns.
Random Forest	Interpretable, handles high-dimensional data well.	Overfitting due to large number of trees.
ANN	Good at capturing complex relationships and patterns	Prone to overfitting with small datasets.

By building an interface and by using demographics, we were able to provide a more personalized risk assessment and also predict levels of schizophrenia in 5 levels . The web application based on Streamlit allows users to input demographic data, refining the model's predictions to offer personalized risk scores.

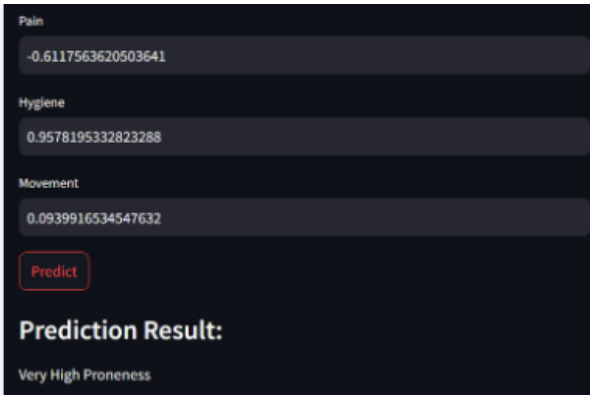


fig.13 Schizophrenia detection interface

Also, using pearson's r score we found an inverse relationship between the symptoms of schizophrenia and depression in certain areas. Based on Pearson's r score, We established that certain symptoms overlapped while others were independent. For example, though depression is commonly associated with enhanced emotional distress and sensitivity to symptoms, schizophrenia presents reduced emotional expression and attenuated awareness of sickness.

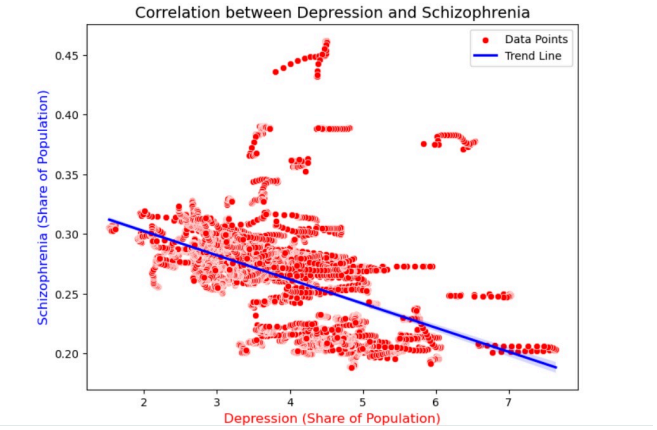


fig. 14 Negative trend line between depression and schizophrenia

IV. CONCLUSION

These experiments helped in the classification of schizophrenia by machine learning and deep learning techniques on EEG signals. Data for analysis were treated with preprocessing processes like feature extraction, segmentation, normalization, and noise removal. The maximum accuracy of 85% was achieved by the Random Forest classifier and an interface using Streamlit for real-time EEG classification.

The paper also included a prediction system that assesses schizophrenia risk based on demographic and survey data and categorizes it into five levels. The implementation of a dual-prediction system enhances accessibility and acts as a non-invasive diagnostic tool.

Techniques like dropout, early stopping, and hyperparameter tuning were used to improve model performance and reduce overfitting. In the future, this could involve data set expansion, model tuning, and the utilization of real-time EEG processing to enhance clinical application and make AI-based early diagnosis and treatment of neuropsychiatric diseases feasible.

V. FUTURE WORK

In subsequent research, we hope to improve the current system by using Natural Language Processing (NLP) methods to support the EEG-based schizophrenia identification. Linguistic anomalies, such as disordered speech, odd word choices, broken sentence structures, and

irregular typing, such as uneven speed or repeated corrections, are common symptoms of schizophrenia. We can find subtle linguistic and behavioral indicators linked to the disease by using NLP techniques to evaluate textual and typing data. Analyzing typing habits, for example, may show delays, unpredictable keystrokes, or hesitation that could be associated with cognitive impairment. Similar to this, examining spoken or written language with sophisticated natural language processing (NLP) models such as sentiment analysis, syntactic analysis, and semantic embedding (e.g., word2vec, BERT) may reveal fragmented language patterns and jumbled thinking.

In order to improve the models' generalization, future research will also concentrate on growing the dataset to include a wider variety of people and language samples. In order to enable the system to dynamically assess behavioral markers and brain activity, we also intend to investigate real-time EEG and NLP data processing. System performance will be further improved by investigating multi-modal fusion strategies and improving the deep learning architectures for both NLP and EEG analysis models. By providing doctors with a cutting-edge, comprehensive tool for assessment and intervention, this combination approach has the potential to completely transform early diagnosis and monitoring of schizophrenia.

REFERENCES

1. Golubović, B., Gajić, Z., & Ivetić, O. (2020). “Factors Associated with Depression in Patients with Schizophrenia.” Acta Clinica Croatica, 59(4), 605–614.
2. Lai, J. W., Ang, C. K. E., Acharya, U. R., & Cheong, K. H. (2020). “Schizophrenia: A Survey of Artificial Intelligence Techniques Applied to Detection and Classification.” Computers in Biology and Medicine, 126, 104044.
3. Hegde, P. R., Nirisha, L. P., & Basavarajappa, C. (2019). “Schizophrenia Spectrum Disorders in India: A Population-Based Study.” Indian Journal of Psychiatry, 61(4), 327–334.
4. Islam, M. M., Hassan, S., & Akter, S. (2021). “A Comprehensive Review of Predictive Analytics Models for Mental Illness Using Machine Learning Algorithms.” Journal of Biomedical Informatics, 118, 103794.
5. Werner, S., Malaspina, D., & Rabinowitz, J. (2007). “Socioeconomic Status at Birth Is Associated With Risk of Schizophrenia: Population-Based Multilevel Study.” Schizophrenia Bulletin, 33(6), 1373–1378.
6. Dvey-Aharon, Z., Fogelson, N., Peled, A., & Intrator, N. (2015). “Schizophrenia Detection and Classification by Advanced Analysis of EEG Recordings Using a Single Electrode Approach.” PLoS ONE, 10(4), e0123033.
7. Padmavati, R., & Balasubramanian, S. (2019). “Cardiovascular Diseases and Schizophrenia in India: Evidence, Gaps, and Way Forward.” Indian Journal of Psychiatry, 61(1), 23–29.
8. Smith, C. J., Johnson, L., & Moore, R. (2018). “Integrating Functional MRI and EEG for Schizophrenia Diagnosis.” Journal of Neural Engineering, 15(6), 066011.
9. Kumar, A., Pratap, R., & Singh, S. (2021). “Deep Learning for Schizophrenia Detection Using EEG Data.” IEEE Access, 9, 18241–18251.
10. Mallat, S. (1999). “A Wavelet Tour of Signal Processing.” Academic Press.
11. Oppenheim, A. V., & Schafer, R. W. (2009). “Discrete-Time Signal Processing.” Pearson Education.
12. Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825–2830.
13. Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5–32.
14. Frid-Adar, M., et al. (2018). "GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification." Neurocomputing, 321, 321–331.
15. Choi, E., Bahadori, M. T., & Sun, J. (2016). "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks." arXiv preprint arXiv:1605.03165.
16. Lundberg, S. M., & Lee, S.-I. (2017). "A Unified Approach to Interpretable Model Predictions." Advances in Neural Information Processing Systems, 30.
17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You? Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
18. Anderson, K. N., & Bradley, A. J. (2013). "Sleep Disturbance in Mental Health Problems and Neurodegenerative Disease." Nature Reviews Neurology, 9(8), 441–451.
19. Garrity, A. G., Pearlson, G. D., et al. (2007). "Aberrant 'Default Mode' Functional Connectivity in Schizophrenia." American Journal of Psychiatry, 164(3), 450–457.