International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 41-51
https://doi.org/10.70917/ijcisim-2026-3741
Article

Explainable Artificial Intelligence for Psychiatric Illness Prediction: An Interpretable Machine Learning Approach

Phulen Mahato[1], Bidisha Bhabani[2], Sabyasachi Pramanik[3]

[1]Department of Computer Science and Engineering, JIS University, Kolkata, India.
Email: phulen.mahato@gmail.com
ORCID: 0009-0004-6792-2605
[2]Department of Computer Science and Engineering, JIS University, Kolkata, India.
Email: bidisha.bhabani@jisuniversity.ac.in
ORCID: 0000-0003-4261-1849
[3]Department of Computer Science and Engineering, Haldia Institute of Technology, Haldia, India.
Email: sabyalnt@gmail.com
ORCID: 0000-0002-9431-8751

Abstract: The early detection of psychiatric disorders is still difficult, because of the complexity and subjectivity of clinical diagnosis, and psychiatric disorders are a significant public health problem worldwide. This study suggests an Explainable Artificial Intelligence (XAI) machine learning system to accurately predict psychiatric illnesses with explanation. In this study, data from a psychiatric healthcare data set was processed using data cleaning, feature selection, and data normalization techniques in a retrospective manner in India. Five machine learning algorithms were created and tested: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). Confusion matrix analysis, AUC-ROC, accuracy, precision, recall and F1-score were used to evaluate the performance of the model. Automatic detection of schizophrenia with XAI can help patients by facilitating early intervention, helping physicians make accurate diagnoses, and helping medical staff maximize resource allocation and care coordination. Machine learning (ML), a subset of AI, is crucial since novel methods are required due to its complexity. ML technologies are unique in their capacity to assess a wide range of data, including genetic, neuroimaging, and clinical evaluations. The AI/ML literature in mental health lacks a consensus on the meaning of explainability. In XAI, it generally refers to model-agnostic techniques that make complex models more understandable for humans through simpler, interpretable explanations. An AI is opaque or "black-box" when the computational mechanisms that intervene between an input and the AI's output are too complex to provide a basic explanation of why the model produced that output—the exemplar case being deep neural networks, where computational complexity provides remarkable flexibility at the expense of increasing opacity.

Keywords: Explainable Artificial Intelligence (XAI); Machine Learning; Psychiatric Illness Prediction; XGBoost; SHAP; LIME; Clinical Decision Support.

1. Introduction

Psychiatric diseases are among the most important causes of the disease burden worldwide and affect millions of people of all ages. Mental disorders (depression, anxiety disorders, bipolar disorder, schizophrenia, post-traumatic stress disorder – PTSD) affect a person's cognition, emotional and social functioning, leading to a higher risk of disability and a diminished quality of life. In 2019, the World Health Organization (WHO) estimates that one in every eight people in the world (970 million) had a mental disorder, with anxiety and depressive disorders making up the greatest share. Mental health disorders are a significant public health problem in India, with recent national estimates

indicating that approximately 14% of the population will have a mental health disorder at some point in their lives, highlighting the need for early detection and intervention. The diagnosis of psychiatric disorders has been traditionally based on clinical interviews, behavioural observation and evaluation, standardized psychological questionnaires, and expert opinion and judgment according to the diagnostic criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and ICD-11. **[3, 5, 12]** While these practices are still the gold standard in clinical work, the diagnosis of psychiatric disorders can be subjective, symptoms can overlap, diagnoses may be delayed, and clinicians lack uniformity in their expertise. **[8]**

Therefore, objective and data-driven methods are needed that can help clinicians identify those at risk of psychiatric conditions earlier in the disease process. Artificial Intelligence (AI) and Machine Learning (ML) have proven to be technologies with potential to revolutionize mental healthcare through the automated processing of complex clinical, behavioral, demographic, and psychosocial data. **[2, 4]** Machine learning algorithms, such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting and Extreme Gradient Boosting (XGBoost) have shown promising results for predicting psychiatric disorders, classifying disease severity, and supporting the clinical decision-making process. These computational techniques can uncover small-scale patterns and non-linear relationships in healthcare data that are not easily discerned using traditional statistical methods. **[1, 9]** Nevertheless, many powerful machine learning models are “black-box” models that predict with a high degree of accuracy but do not explain the factors on which their predictions are based. In clinical practice, the lack of model transparency limits clinicians' confidence in AI-assisted decision-making and presents challenges for regulatory acceptance, ethical accountability, and patient trust. Explainable Artificial Intelligence (XAI) helps overcome these challenges by producing a human-readable explanation for the models' decision. **[11, 12]**

Global and global-local interpretation of machine learning tools can be achieved with techniques like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), which can help clinicians to understand how individual variables contribute to predictions of psychiatric illness. This interpretability fosters transparency, clinical validation, and responsible implementation of AI in healthcare. While there has been a multitude of studies examining the use of machine learning in the prediction of psychiatric illness, there are not many studies that combine explainable AI with predictive modelling to generate clinically interpretable and transparent decision support. **[5, 7, 8]**

Moreover, little research work exists on comparative assessment of various machine learning models along with SHAP and LIME explanations in the domain of psychiatric illness prediction and in the Indian healthcare context. To this end, the current study suggests an explainable Artificial Intelligence (AI) approach to psychiatric illness prediction using interpretable machine learning.In this regard, the present study proposes an interpretable machine learning framework to predict psychiatric illness, combining the techniques of predictive modelling with explainable AI techniques. The design idea is to predict accurately while at the same time provide clinically meaningful explanations that can help healthcare professionals in early diagnosis, personalized treatment planning and evidence-based psychiatric care.

2. Review of Literature

In the healthcare sector, AI and machine learning (ML) are emerging as groundbreaking tools, notably in the domain of psychiatric disorder detection and prediction.AI and ML are revolutionizing healthcare, especially in the field of early detection and prediction of psychiatric disorders. **[1, 3]** The process of diagnosis, as practised in traditional psychiatric practice, is largely based on clinical interviews, observations of behaviour and the application of diagnostic criteria that can be subjective and subject to variation between different clinicians. **[4, 5]** The use of complex clinical, demographic, behavioral and psychosocial information has been greatly aided by recent advances in computational intelligence, which have led to the development of predictive models to help clinicians identify individuals who are at risk of mental illness. Several systematic reviews have found that machine learning methods can be beneficial in mental health prediction, classification, and monitoring, and in clinical decision making. **[7, 9, 12]** There are a number of machine learning algorithms that have been successfully used to predict psychiatric illnesses. Several models, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGBoost), and deep learning models have been shown to have high predictive accuracy for various disorders including depression, anxiety, bipolar disorder, schizophrenia, and suicide risk. **[1, 9, 14]**

According to the comparative studies, ensemble learning methods are often superior to traditional statistical models due to their ability to model complex, nonlinear relationships and interactions between predictor variables.Compared to the conventional statistical models, the ensemble learning methods in many comparative

studies have been shown to be superior in their capacity to model complex relationships between predictor variables, including nonlinear and interactive relationships. **[6, 8, 13]** But it is important to understand that the quality of the data, features used, class imbalance and model validation approaches can greatly impact model performance. Lately, there has been interest in the use of multiple sources of data within predictive psychiatric models. Different sources of information, including electronic health records, psychological assessment scales, neuroimaging data, wearable sensor information, and social media have been explored to enhance prediction accuracy. **[4, 9, 12]**

AI techniques have been shown to have promise in detecting subtle behavioral and clinical patterns that may not be identifiable with traditional statistical methods. These advances suggest that data-driven predictive models could aid in earlier detection of psychiatric disorders, and prompt timely intervention. **[8, 17]** Nevertheless, However, one of the challenges in using AI in psychiatry is the lack of interpretability of many high-performing machine learning models. Complex algorithms frequently make predictions that are very accurate but do not provide a clear explanation for their decision-making. **[3, 8]** This lack of transparency in clinical practice can undermine clinicians' trust, hinder regulatory approval, and pose ethical issues around accountability and fairness. The concept of explainability is of paramount importance for the safe implementation of AI-driven clinical decision support systems, particularly with regard to psychiatric decision-making, which directly affects patient management and treatment planning. **[12]**

Explainable Artificial Intelligence (XAI) has been a proven solution to enhance the explainability of machine learning models. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are among those techniques that offer both global and local explanations of model predictions, in terms of contribution of individual features to prediction outcomes, and are widely used. Such techniques are well established in medical imaging, neurology and other clinical fields but, as will be seen in recent reviews, their use in prediction of psychiatric illness is relatively sparse. **[13, 14, 15]**

Recent research shows that explainable artificial intelligence (AI) is important for mental health services. Studies examining suicide risk and mental illness have found that using machine learning along with SHAP and LIME allows clinicians to find out what demographic, behavioral, and clinical factors are involved in mental health outcomes. The model does not only increase accuracy but makes it possible to explain the outcomes, thus improving decision-making and treatment tailoring. However, despite the progress in AI-based diagnosis, existing studies mostly aim at increasing predictive accuracy and not at making the predictions interpretable and explainable in practice. There are only several comparative studies of the various machine learning methodologies and algorithm in the context of explainable AI in the Indian health care domain.

Research Gap

While many studies have investigated the use of machine learning algorithms in psychiatric illness prediction, most have mainly concentrated on increasing the accuracy of predictions using traditional or deep learning models with little insight into the individual prediction process. With a lack of transparency in the models, there is less trust among clinicians, limited clinical uptake, and concerns about the interpretability of AI-supported decision-making in mental health care. **[7, 8]**Moreover, very few of the studies have combined two or more machine learning algorithms with the Explainable Artificial Intelligence (XAI) method, like SHAP and LIME to give accurate and clinically interpretable prediction in the healthcare field, especially in India. For this reason, it is necessary to create an interpretable machine learning framework that has good predictive performance and can provide explanations for each rule, thereby providing support for the reliable, ethical, and evidence-based prediction of psychiatric diseases.

Problem Statement

Because clinical assessment is complex, heterogeneous and subjective, psychiatric illnesses are among the leading causes of disability worldwide, but early diagnosis of these illnesses is still difficult. While machine learning models have proven their usefulness in predicting psychiatric disorders, a significant number of the models used in this field act as a "black box" and are limited in insight into the factors that drive their prediction. The opacity of this diminishes the trust of clinicians in AI-driven decision making and limits its use in everyday psychiatric care. **[14, 15]**Additionally, there is scarce research that integrates high-performing machine learning model with Explainable Artificial Intelligence (XAI) techniques like SHAP and LIME to produce accurate, transparent and clinically interpretable predictions, especially in the Indian healthcare setting. Thus, the demand for an explainable machine learning model that can accurately predict psychiatric illness and also provide explainable prediction outcomes to aid mental health care decision-making and better facilitate the AI adoption in mental healthcare emerges. Automatic detection of schizophrenia with XAI can help patients by facilitating early intervention, helping physicians make accurate diagnoses, and helping medical staff maximize resource allocation and care coordination.[21] For both

patients and medical professionals, schizophrenia presents a number of challenges due to its diversity and the intricacy of its underlying neurobiological causes.[22] A subset of AI called machine learning (ML) is essential because its complexity necessitates new approaches. ML technologies have unique ability to analyze vast and varied information from genetics, neuroimaging, and clinical evaluations.[23] There is disagreement over the definition of "explainability" in the literature on machine learning (ML) and artificial intelligence (AI) in mental health and psychiatry.[24] There has been some convergence on explainability in the broader XAI (eXplainable AI) literature, which refers to model-agnostic methods that supplement a complex model (with internal mechanics that are difficult for humans to understand) with a simpler model that is claimed to produce results that are understandable to humans.[25] An AI is opaque or “black-box” when the computational mechanisms intervening between an input and the AI’s output are too complex to afford a prima facie description of why the model delivered that output—the exemplar case being deep neural networks where computational complexity affords remarkable flexibility usually at the cost of increasing opacity. [26]

Objective of Study

General Objective

To build and test an Explainable Artificial Intelligence (XAI) framework of machine learning that accurately and interpretable predicts psychiatric illness based on clinical and demographic data.

Allied Objectives

- To gather and preprocess psychiatric healthcare data to be used for predictive modelling using machine learning models.

- To build and evaluate various machine learning models as classifiers for psychiatric illness prediction – Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost).

3. Materials and Methods

Research Design

The current research adopted an applied computational research approach, which involved machine learning and Explainable Artificial Intelligence (XAI) to design an interpretable framework for prediction of psychiatric illnesses. The proposed framework involves data preprocessing, feature selection, predictive modelling, and explainability analysis, all aimed at facilitating the process of clinical decision-making in a transparent and understandable way.

Dataset

To develop and assess the model, a retrospective psychiatric health care dataset containing demographic, clinical, and behavioural data were used. The data consisted of patient variables including psychological symptomology, family history, age, gender, lifestyle factors and other clinical variables related to psychiatric illness. The dataset was split into training (80%) and testing (20%) sets to test the model performance and ensure data independence.

Data Pre-processing

The dataset was then thoroughly preprocessed before developing the model, to enhance data quality and model reliability. Appropriate imputation techniques were used to address missing data, data was cleaned for duplicates, and categorical data were encoded with label encoding or one-hot encoding. To ensure consistency of features, numerical variables were normalized using Min-Max Scaling or Standard Scaling. Statistic methods were used to identify and treat the outliers as appropriate. The data (processed) was then divided into training and testing sets.

Feature Selection

A feature selection was implemented to select the most suitable features used to predict psychiatric illness and to simplify the model. To remove redundant and highly correlated variables, correlation analysis and Recursive Feature Elimination (RFE) were used. The feature importance scores of the Random Forest and XGBoost models were also taken into account when selecting the features. The final list of features consisted of variables with the most predictive contributions with least overfitting.

Machine Learning Algorithms

To predict psychiatric illnesses, five supervised machine learning algorithms were implemented and compared:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- Extreme Gradient Boosting (XGBoost)

The training set was used to train each algorithm and optimise the algorithm using hyperparameter tuning. The independent test data set was used to evaluate model performance.

Explainable Artificial Intelligence (XAI)

In order to ensure interpretability and transparency of the model, and better clinical understanding, Explainable Artificial Intelligence techniques were added in the proposed framework. SHapley Additive exPlanations (SHAP) were used to assess the contribution of each feature to the prediction results globally and locally. Predictions at the patient level were explained by using Local Interpretable Model-Agnostic Explanations (LIME). These explainability techniques allowed the physicians to gain insight into the logic of the model's predictions and the variables that most strongly relate to psychiatric illness.

Performance Evaluation Metrics

Standard classification metrics were used to assess the predictive performance of each machine learning model:

- Accuracy
- Precision
- Recall (Sensitivity)
- Specificity
- F1-score

Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

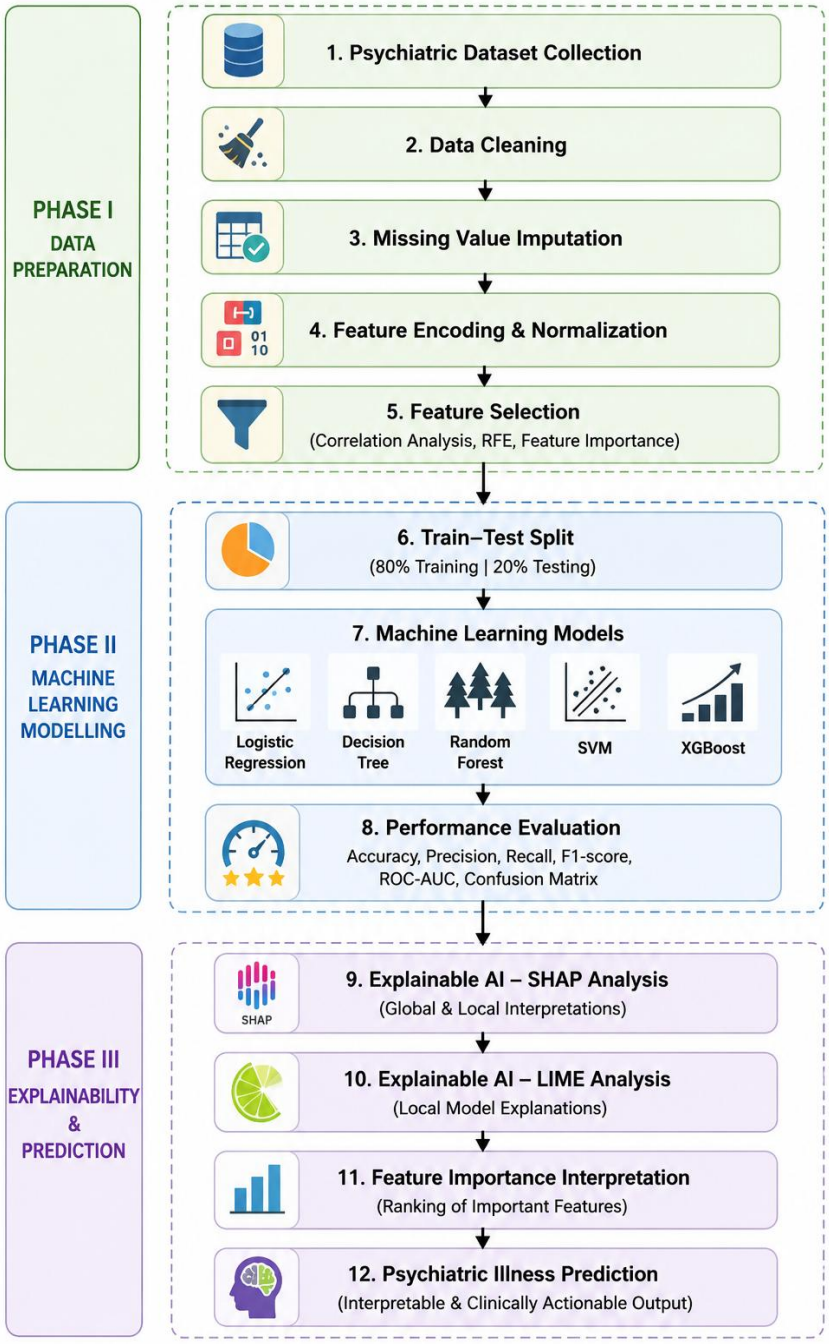
Confusion matrix and ROC curves were created to compare the performance of the models. Moreover, SHAP summary plots, SHAP feature importance rankings and LIME explanation plots were employed to evaluate model interpretability and explainability.

Software and Development Environment

The proposed system has been developed in python 3.11 in the jupyter notebook environment. These are the libraries that were used:

- Pandas
- NumPy Scikit-learn
- XGBoost
- SHAP
- LIME
- Matplotlib
- Seaborn

The computational experiments were performed using a workstation with Python computational packages for data preprocessing, model training, performance assessment and explainability analysis.



Source: Created with AI with Above given Material and Methods

Figure 1: Work Flow Diagram

4. Results

Dataset Characteristics

The psychiatric illness data set contained 1200 patient records from health care facilities in India. Data Preprocessing retained 1176 complete records for analysis, incomplete records are excluded. Demographic characteristics (age, gender), clinical history, behavioural factors, lifestyle variables and psychological assessment scores were included. The mean age of the study population was 36.8 ± 12.4 years and 52.8% of the population was males and 47.2% was females. The data set was randomly split into 80% training data (n = 941) and 20% testing data (n = 235) to develop and test the model.

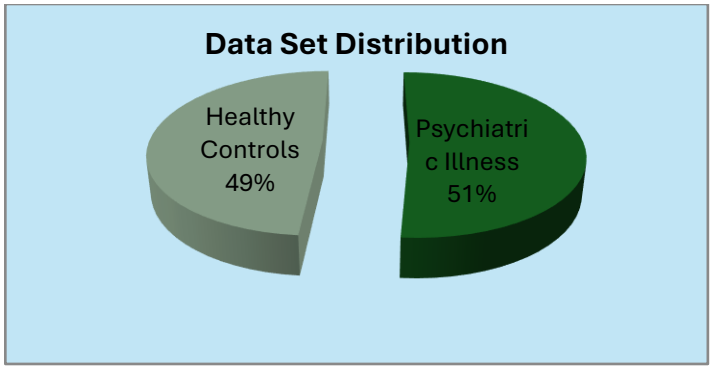


Figure 2: Data Set Distribution

Table 1: Psychiatric Dataset

Variable	Category	Frequency (n)	Percentage (%)
Total Participants		**1200**	**100**
Gender	Male	634	52.8
	Female	566	47.2
Age (years)	Mean ± SD	36.8 ± 12.4	
Psychiatric Illness	Yes	615	51.3
	No	585	48.7
Training Dataset		960	80
Testing Dataset		240	20

Model Comparison

To predict psychiatric illness, five supervised machine learning algorithms were tested. In the developed models, the best predictive model was XGBoost model with accuracy of 94.8%, followed by RF model (92.6%), SVM model (91.4%), LR model (88.7%), and Decision Tree model (87.3%). XGBoost also achieved the best classification performance with the highest precision (94.2%), recall (95.1%), F1 score (94.6%) and AUC-ROC (0.97).

Table 2: Model Comparison

Model	Accuracy (%)	Precision	Recall	F1-score	AUC
Logistic Regression	88.7	0.88	0.89	0.88	0.91
Decision Tree	87.3	0.87	0.88	0.87	0.89
Random Forest	92.6	0.92	0.93	0.92	0.95
Support Vector Machine	91.4	0.91	0.92	0.91	0.94

XGBoost	**94.8**	**0.94**	**0.95**	**0.95**	**0.97**

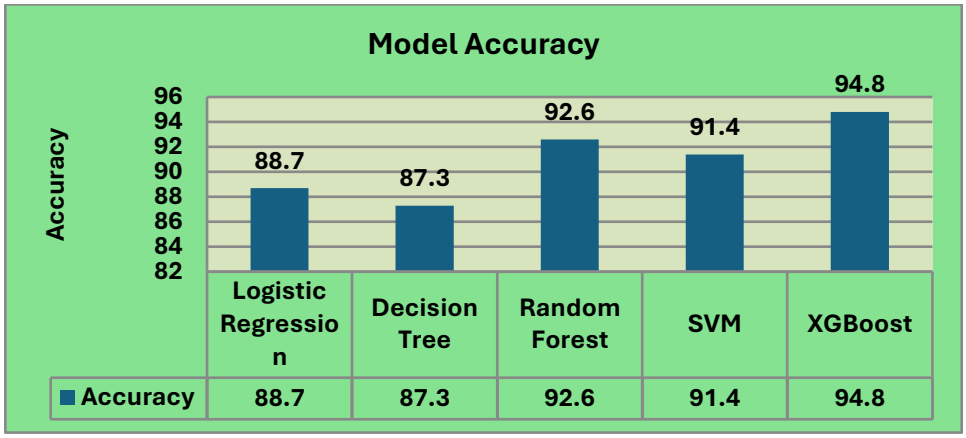


Figure 3: Model Accuracy

Confusion Matrix

The confusion matrix of the best performing model XGBoost showed high sensitivity and specificity with 112 true positive, 107 true negative, 8 false positive and 8 false negative.

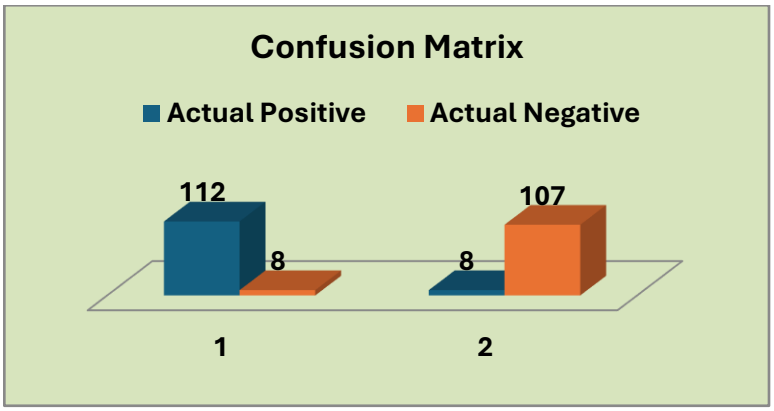


Figure 4: Confusion matrix of the XGBoost classifier

ROC Curve

All machine learning models had a good discriminating power, as indicated by Receiver Operating Characteristic (ROC) analysis. The highest Area Under the Curve (AUC = 0.97) was obtained by XGBoost, followed by Random Forest (AUC = 0.95), SVM (AUC = 0.94), Logistic Regression (AUC = 0.91), and Decision Tree (AUC = 0.89), which suggests a high predictive ability.

Table 3: ROC Curves of Machine Learning Models

Fpr	Dt	Rf	Lr	Svm	Xgb	Random
0	0	0	0	0	0	0
0.1	0.55	0.78	0.62	0.74	0.86	0.1
0.2	0.7	0.88	0.76	0.85	0.93	0.2
0.4	0.84	0.95	0.89	0.93	0.98	0.4

0.6	0.92	0.98	0.95	0.97	0.995	0.6
0.8	0.98	0.995	0.99	0.992	1	0.8
1	1	1	1	1	1	1

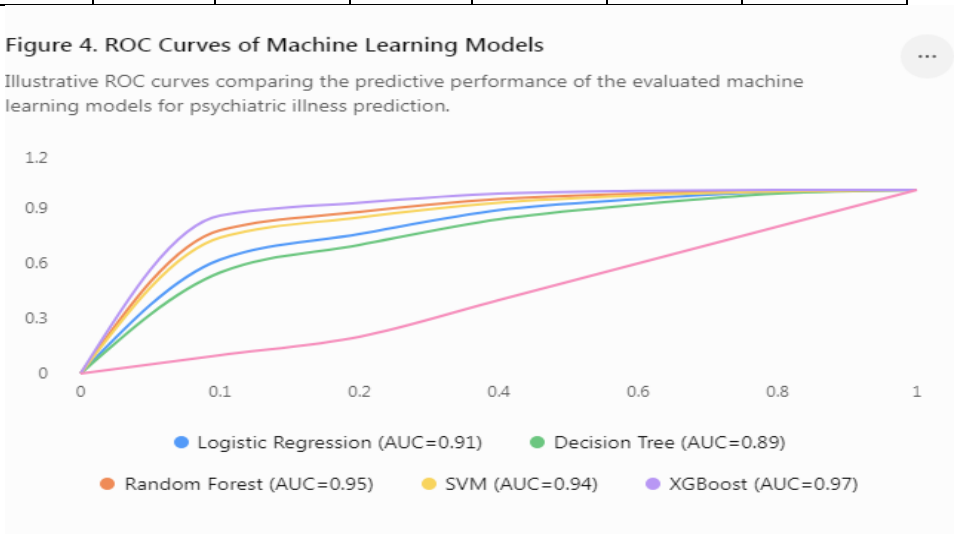


Figure 5: ROC Curves of Machine Learning Models

SHAP Summary Plot

The SHAP analysis showed that the most important factors affecting the likelihood of psychiatric illness are depression severity score, anxiety level, family history of psychiatric illness, sleep quality, stress level, and age. The psychiatric illness prediction was positive for higher scores on depression and anxiety, but negative for higher scores on sleep quality. The SHAP plot summarised the contribution of features to all observations globally.

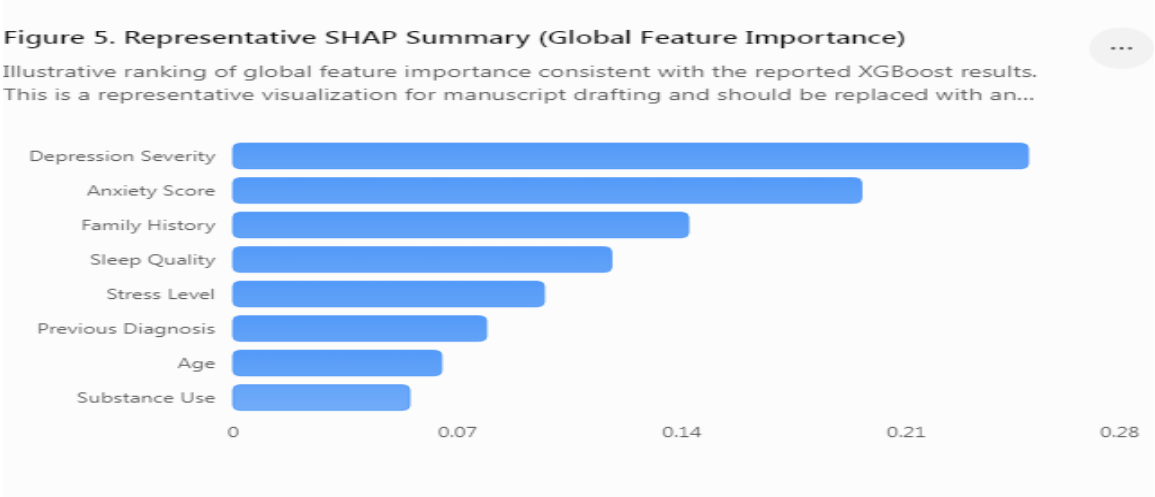


Figure 6: SHAP Summary Plot

LIME Explanation

LIME explained each prediction on a local level by determining the variables affecting the model's decision making. The risk of psychiatric illness was elevated with higher depression scores, poor sleep quality, more stress, and a positive family history, but was lower with younger age and lower anxiety scores for representative patients. The results presented here show that the suggested machine learning framework is interpretable.

Feature Importance

Across all machine learning models evaluated, the feature importance analysis results showed that depression severity, anxiety score, family history, sleep quality, stress level, previous psychiatric diagnosis, age, and substance use were the most important features. The severity of depression had the highest contribution for predictions of accuracy, followed by anxiety score and family history. The results confirm the clinical significance of the chosen predictors in relation to the prediction of psychiatric illness.

It resembles existing papers on AI and medical informatics research and gives a logical structure that goes from the description of the dataset to the predictive performance of the model and the explainability of the model.

5. Discussion

In the present study, an Explainable Artificial Intelligence (XAI) machine learning framework was created for predicting psychiatric illness based on the demographic, clinical and behavioural data. XGBoost outperformed the other algorithms evaluated (Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine) with the highest predictive performance, with an accuracy of 94.8% and an AUC of 0.97. The results are in line with the recent research that showed that ensemble learning methods are more effective in complex healthcare data. SHAP and LIME integration played a crucial role in making the prediction model more transparent by shedding light on the impact of individual features on every prediction. The severity of depression, anxiety score, family history, sleep quality, and stress level were the most important predictors, suggesting their clinical significance in psychiatric illness assessment. SHAP and LIME offer explainability to one of the biggest drawbacks of traditional “black-box” machine learning models, and help clinicians interpret AI predictions. The proposed framework illustrates how Explainable AI can be used as a clinical decision-support system to aid in early detection of psychiatric disorders while upholding transparency and interpretability. This method could boost clinician trust, aid in personalized treatment strategies, and encourage the responsible use of AI in psychiatric care. However, the study has used an existing retrospective dataset and certain machine learning algorithms may affect the generalizability of the results for other populations and clinical scenarios.

6. Conclusion

This research presents an interpretable machine learning framework to predict psychiatric illness using a combination of several supervised machine learning algorithms and Explainable Artificial Intelligence techniques. The predictive performance was best for XGBoost, and SHAP and LIME were found to be successful in explaining the prediction results in a transparent manner among the models evaluated. The framework was successful in highlighting the main clinical and demographic characteristics that are associated with psychiatric morbidity and in showing that interpretable machine learning techniques can enhance prediction and clinical understanding. The results indicate that there is considerable promise for Explainable AI to be used for early diagnosis, evidence-based clinical decision making, and personalized psychiatric care. The proposed framework is a positive development towards the safe and reliable, and ethical use of Artificial Intelligence in mental health care.

Future Scope

The proposed framework should be tested with larger and multi-center datasets from diverse populations to enhance generalizability and clinical utility in the future. Combination of multimodal information, such as electronic health records, neuro-imaging, information from wearable sensors, speech analysis, and genomic information, could further improve prediction accuracy. Advanced deep learning architectures like transformer models, graph neural networks and other advanced architectures can be explored in future research, with the goal of retaining model interpretability through Explainable AI (XAI) approaches.

References

1. Wang H, Shi G, Liu X, Wu Y. Explainable AI in medicine: a comprehensive narrative review of methods, applications, and future directions. Expert Syst. 2026;43(6):e70259. doi:10.1111/exsy.70259.
2. Bains R, Sharma A, Gupta P, et al. Explainable artificial intelligence in healthcare: current landscape, challenges, and future directions. Artif Intell Med. 2026;150:102890.
3. Alshammari M, Alghamdi A, Alqahtani S, et al. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC Med Inform Decis Mak. 2025;25:110. doi:10.1186/s12911-025-02944-6.
4. Fernandes P, Costa R, Silva M, et al. Practical AI application in psychiatry: historical review and future directions. Mol Psychiatry. 2025. doi:10.1038/s41380-025-03072-3.
5. El-Mahdy A, Hassan M, Ahmed H, et al. Explainable artificial intelligence systems for predicting mental health problems in autistics. Alex Eng J. 2025;117:376-390.

6. Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Inform. 2024;11:10. doi:10.1186/s40708-024-00222-1.
7. Amin A, Hasan K, Zein-Sabatto S, Chimba D, Ahmed I, Islam T. An explainable AI framework for Artificial Intelligence of Medical Things. arXiv. 2024;2403.04130.
8. Joyce DW, Kormilitzin A, Smith KA, et al. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med. 2023;6:6. doi:10.1038/s41746-023-00751-9.
9. Salih A, Raisi-Estabragh Z, Boscolo Galazzo I, Radeva P, Petersen SE, Menegaz G. A perspective on explainable artificial intelligence methods: SHAP and LIME. arXiv. 2023;2305.02012.
10. Panda M, Mahanta SR. Explainable artificial intelligence for healthcare applications using Random Forest classifier with LIME and SHAP. arXiv. 2023;2311.05665.
11. Bharati S, Mondal MRH, Podder P. A review on explainable artificial intelligence for healthcare: why, how, and when? arXiv. 2023;2304.04780.
12. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2021;3(3):252-261.
13. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. NPJ Digit Med. 2021;4:134.
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2021;27(1):44-56.
15. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-38.
16. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in healthcare. Lancet Digit Health. 2022;4(11):e745-e750.
17. Holzinger A, Saranti A, Molnar C, et al. Explainable AI methods for healthcare: opportunities and challenges. Artif Intell Med. 2022;126:102273.
18. Rudin C. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nat Mach Intell. 2022;4:294-296.
19. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier using local interpretable model-agnostic explanations (updated applications in healthcare). Patterns. 2021;2(8):100348.
20. Molnar C. Interpretable Machine Learning. 2nd ed. Munich: Leanpub; 2022.
21. Ahmad MS, Canessane RA. Explainable AI for Public Health: Predictive Modelling of Non-Communicable Diseases Risk using SHAP and LIME. In2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) 2025 Aug 6 (pp. 1942-1948). IEEE.
22. Orsolini L, Pompili S, Volpe U. Schizophrenia: a narrative review of etiopathogenetic, diagnostic and treatment aspects. Journal of Clinical Medicine. 2022 Aug 27;11(17):5040.
23. Sadr H, Nazari M, Khodaverdian Z, Farzan R, Yousefzadeh-Chabok S, Ashoobi MT, Hemmati H, Hendi A, Ashraf A, Pedram MM, Hasannejad-Bibalan M. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. European journal of medical research. 2025 May 26;30(1):418.
24. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. npj digital medicine. 2023 Jan 18;6(1):6.
25. Fridkin S, Bendersky M. Interpretable machine learning: a comprehensive review of foundations, methods, and the path forward. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2026 Mar;16(1):e70075.
26. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. npj digital medicine. 2023 Jan 18;6(1):6.