Enhancing Mental Health Diagnosis with Ensemble Learning: A Comparative Analysis of Bagging, Boosting, and Random Forest Models

MEENU PILLAI[1] ,ANUGRAHA BIJUMON[2], SHREYA MANDE[3], RADHIKA DADAS[4]

[1]MITWPU, Pune, Maharashtra (e-mail: meenupillai510@gmail.com)
[2]MITWPU, Pune, Maharashtra (e-mail: anugrahabijumon@gmail.com)
[3]MITWPU, Pune, Maharashtra (e-mail: shreyamande04@gmail.com)
MITWPU, Pune, Maharashtra (e-mail: shreyamande04@gmail.com)

ABSTRACT This paper proposes an ensemble learning model using Bagging, Boosting, and Random Forest techniques to enhance diagnostic accuracy and reliability. In addition to predicting mental health outcomes, the model incorporates four key functionalities: visualization of the diagnostic process, confidence scoring to indicate model certainty in each prediction, and Explainable AI (XAI) techniques to provide interpretability on model decisions. By analyzing patient data, including emotional and behavioral features, this comprehensive model outperforms conventional diagnostic methods. Tested on a Kaggle dataset, the model achieves accuracies of 80.5 for Bagging, 55.56 % for Boosting, and 75% for Random Forest, demonstrating both high performance and enhanced transparency in predictions

INDEX TERMS Ensemble Learning, Bagging, Boosting, Random Forest, Mental Health Diagnosis, Explainable AI (XAI), Diagnostic Process Visualization, Confidence Scoring, Emotional and Behavioral Features, Predictive Analytics, Machine Learning in Mental Health, Kaggle Dataset, Diagnostic Accuracy, Model Interpretability

I. INTRODUCTION

The prevalence of mental health problems is rising, especially among young adults, who are frequently the most at-risk group. Research has indicated that mental health disorders including stress, anxiety, and depression are becoming more prevalent in teenagers and young adults. This is primarily because of things like social media influence, academic pressure, and the difficulties of growing up. Despite increased awareness, the stigma associated with mental health and the shortcomings of conventional diagnostic techniques continue to prevent many young people from receiving prompt diagnosis and successful interventions. Finding creative solutions that can aid in the early and more accurate identification of mental health concerns is therefore crucial.

By increasing diagnostic precision and offering more profound understanding of the elements influencing mental health disorders, machine learning (ML) approaches have become a viable means of tackling this problem. Based on a variety of young adults' emotional and behavioral characteristics, our article suggests an ensemble learning model that uses Bagging, Boosting, and Random Forest approaches to predict mental health outcomes. By merging several models, these methods are renowned for their capacity to increase prediction performance in terms of accuracy.

In order to better understand how each attribute affects the predictions, the model also integrates Explainable AI (XAI) approaches, particularly SHAP (Shapley Additive Explanations). By using SHAP, we offer interpretability and openness while illuminating the elements that play a major role in mental health diagnosis. For ML-based technologies to be trusted and understood, particularly in delicate fields like mental health, interpretability is essential.

In order to improve the model's usability and accessibility, we have created a webpage that illustrates the diagnostic procedure and provides users with both the expected results and the model's confidence scores for each prediction. Users may better comprehend the logic behind the model's judgments thanks to this depiction, which also gives them the confidence to make wise decisions about their mental health.

Tested on a Kaggle dataset, the suggested model showed strong performance with accuracies of 80.5% for Bagging, 55.56% for Boosting, and 75% for Random Forest. This method could be a useful tool for early diagnosis and intervention, offering both accurate predictions and a better knowledge of the underlying elements contributing to mental health outcomes, especially in light of the expanding mental health issues that young adults confront. The objective of this research is to enhance young adults' general well-being and promote mental health awareness by fusing cutting-edge machine learning techniques with interpretability and accessibility. AB, and MP, all authors have made equal contributions to the creation of this manuscript. Dr. Priti Chakurkar supervised the study and offered crucial feedback during the drafting of the manuscript.

II. RESEARCH AIM AND SCOPE OF THE PAPER

Developing and assessing machine learning models for the prediction and classification of mental health conditions, including bipolar disorder, depression, and normal mental health states, is the goal of this study. The study aims to improve the precision and interpretability of predictions in the field of mental health by employing ensemble techniques like Bagging, Boosting, and Random Forest. Additionally, the study uses SHAP (SHapley Additive exPlanations) analysis to help doctors understand how the models make decisions by offering clear insights into the elements influencing the model's predictions.

The creation of machine learning models trained on a dataset containing a variety of behavioral and psychological characteristics, including mood swings, violence, and suicidal thoughts, is part of the research's scope. The dataset will be prepped for analysis using data preparation methods such label encoding, normalization, and imputation. The models' performance will be compared across several methods and assessed using common classification measures including accuracy, precision, and recall. SHAP values will be used to illustrate the significance of each feature in the models' predictions in order to improve interpretability. A Flask-based web application will be created that will let users enter data and get predictions in real time, along with graphics to help with mental health evaluations.

Although the study focuses on the existing dataset, which may have some coverage constraints, it lays the groundwork for future research that may incorporate other elements, such as social and environmental aspects, and larger, more diversified datasets. The ultimate objective is to provide tools that not only forecast mental health conditions but also provide clear explanations, increasing the technology's use and reliability for medical professionals.

III. LITERATURE REVIEW

Şevgin (2023) [2]conducted a comparative analysis to assess the efficacy of the boosting and bagging algorithms, concentrating on the TreeNet and Random Forest approaches in the educational field. The study used three, five, and tenfold cross-validation techniques to examine these approaches under different sample sizes. The results showed that for the majority of sample sizes, especially at 250 and 1000 data points, TreeNet fared better than Random Forest in terms of classification accuracy, sensitivity, F1-score, and AUC value. For bigger sample sizes, like 500 and 2000, Random Forest, on the other hand, showed superior specificity and occasionally outperformed TreeNet in AUC values. The study emphasizes ensemble methods' subtle performance variations and their potential for customized applications in educational data processing.

Ogunseye et al. (2022) [3] used information from the 2014 and 2016 Mental Health in IT Survey, which contained 2,692 cases, to conduct a thorough examination of mental health issues among IT workers. The prevalence of reported mental health illnesses decreased significantly, from 53.7% to 41.8%, according to their data, indicating possible improvements in reporting methods and awareness among this population. Key characteristics such age, gender, family history of mental health problems, workplace benefits, care alternatives, reporting anonymity, leave rules, and the effect of work on mental health were painstakingly retrieved and examined in this study. What was alarming from the demographic analysis was the noticeable difference in gender. Most of the respondents however were male and most respondents were posted to middle and large companies which were sited in the US and some portions of Europe. As for methodology, the authors performed intensive data cleaning to handle extent missing complications during their analysis and final analysis through covariance and correlation matrices with the aim of determining relationships between a number of mental health parameters. They used machine learning methods in particularly the AdaBoost algorithm whose performance in predicting treatment response of 81.75% accuracy has been the best among other model types. Such predictive models were further exemplified in this study using ROC curves to illustrate their efficiency in predicting mental health treatment outcomes. More so, the writers examined the prospect of Natural Language Processing NPL as one of the promising flags for the early detection of mental health disorders within the frameworks of systematic review 399 papers where the number of mental illness diagnoses with the help of NLP methods was increasing. The study's conclusions included suggestions for further research that would use deep learning methods to improve prediction accuracy and comprehend the intricacies of mental health conditions across a variety of demographics. While 55.7% of organizations offered mental health coverage, only 21.7% provided adequate services, and 21.4% guaranteed anonymity for employees seeking treatment, the findings underscored the urgent need for better mental health resources and support systems within the IT industry. The research was published under the Creative Commons License Attribution 4.0 International (CC BY 4.0), which permits unrestricted use and distribution with proper attribution to the original authors. The overall goal of the study was to empower healthcare systems and improve mental health outcomes through data-driven insights and advanced analytical methods.

Kortas et al. (2024) [6] conducted a comprehensive study on the use of machine learning techniques to predict self-inflicted thoughts and behaviors in mental health. The research aimed to enhance early detection by leveraging a dataset of 2100 samples with 26 parameters, encompassing sociodemographic, medical, and environmental factors. A rigorous methodology was employed, starting with data collection, followed by preprocessing to address inconsistencies, such as handling missing values, outliers, and encoding categorical variables. The study utilized correlation analysis to identify significant predictors and employed visualizations to explore data distributions and treatment probabilities by demographic factors. The researchers tested multiple machine learning models, including logistic regression, K-nearest neighbors, decision trees, random forests, bagging, boosting, stacking, and neural networks. Performance metrics such as accuracy, precision, recall, and F1-score were used for evaluation. The results highlighted that neural networks and random forests achieved the highest accuracies, 97% and 97.4% respectively, while other models also demonstrated robust predictive capabilities. The study underlined the importance of data size and diversity, feature selection, and hyperparameter tuning to improve model performance and reliability. The study acknowledged its limitations, including potential biases from self-reported data, a modest dataset size, and the exclusion of some predictive features due to data availability constraints. It called for future research with larger datasets and advanced techniques, like deep learning, to refine prediction accuracy. Ultimately, this research provided valuable insights into enhancing mental health diagnostics through machine learning, advocating for the integration of these models as complementary tools alongside traditional clinical assessments.

Shunmugam et al. (2022) [7]emphasized the algorithms' effectiveness in mental health analysis. Support Vector Machine (SVM) and Random Forest were found to be the most widely used supervised learning algorithms for evaluating depression. The substantial potential of combining machine learning methods with MRI data to enhance the diagnosis of major depressive illness was also emphasized. The authors talked about how modern mental health surveillance systems can use machine learning to either replace or supplement psychologists. Notwithstanding its potential, the assessment emphasized the shortcomings of existing approaches and urged more investigation to fully grasp and expand on machine learning's benefits in this area.

Mienye and Sun (2022) [8] covered its history, important methods, uses, and potential future developments. The capacity of ensemble learning techniques to lower variance and bias errors in individual machine learning models was emphasized, improving classification and regression tasks in a variety of areas, including anomaly detection, fraud detection, and medical diagnosis. With precision scores of 92.9% and 77.5% on the European cardholders and Brazilian credit databases, respectively, the study showed that ensemble approaches are superior to six tried-and-true algorithms by 4%. Furthermore, the models performed exceptionally well in medical applications, surpassing previous techniques with a 100% accuracy rate in diagnosing breast cancer. The authors highlighted the strong adaptability of ensemble learning and its capacity to tackle challenging real-world issues..

Syed Mohamed et al. (2023) [9] suggested a hybrid mental health prediction model that uses Random Forest (RF), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) algorithms to categorize the clinical phases of anxiety. Their study showed that the RF method performed better than other algorithms, with an accuracy of 98.14% when feature selection was used and 97.67% when it wasn't. In comparison to other models like SVM and MLP, this was higher. The best results were obtained with a polynomial kernel and data standardization, and robust classifier performance in data distribution and allocation was validated by kappa statistics. The study underlined how the suggested paradigm can help mental health professionals increase the precision of their diagnoses. However, it highlighted concerns regarding algorithm transparency, which could impact trust, especially in sensitive mental health applications. The findings underscore the potential of machine learning in enhancing clinical diagnostics for mental health disorders.

Sivagnanam and Visalakshi (2023) [10] used a dataset of 183 case samples, including comprehensive medical histories and treatment details, to examine the use of ensemble machine learning classifiers for bipolar illness diagnosis. About 25 characteristics were included in the dataset, including psychiatric disorders, age, gender, family history, prior therapies, and symptom duration. In order to enable precise model training, data preprocessing—a crucial step—involves locating and handling missing values, dealing with outliers, and fixing inconsistencies. The development or selection of pertinent features, including as demographic information, symptom evaluations, prescription records, and genetic markers, had a major influence on model performance, demonstrating the importance of feature engineering. To facilitate efficient model training and evaluation, the dataset was separated into three subsets: training, validation, and test sets. With a Kappa value of 0.6988 and an outstanding accuracy rate of 98.52%, the ensemble classifier—which was trained on the training dataset and tested on the test dataset—showed a high degree of agreement between the anticipated and actual classifications. The accuracy of the model's positive predictions and its capacity to accurately identify positive cases were shown by key performance indicators such as recall, precision, and the F-measure. The study found that the ensemble classifier performed better than current approaches, highlighting the potential of cutting-edge machine learning techniques to improve bipolar disorder diagnostic accuracy. Additionally, it emphasized the significance of careful data analysis and efficient feature selection in tackling the difficulties involved in bipolar illness diagnosis, ultimately leading to better mental health diagnostics and treatment approaches.

IV. RESEARCH METHODOLOGY

This study's methodology integrates techniques including Random Forest, Boosting, and Bagging to create and assess an ensemble learning model for mental health diagnosis. The following crucial steps make up the process:

A. DATASET COLLECTION AND PREPROCESSING

- **Dataset Source:** A Kaggle dataset on mental diseases was used, which included behavioral and emotional characteristics suggestive of a range of mental health issues.
- **Preprocessing:** Numerical features were normalized and missing values were handled by imputation to clean up the dataset. One-hot encoding was used to encode categorical variables, and the dataset was divided into subsets for testing and training. (e.g., 80-20 split).

B. MODEL SELECTION AND TRAINING

Three ensemble learning methods were employed:

- **Bagging:** Reduces variance by combining predictions from several base estimators that were trained on arbitrary subsets of the data.
- **Boosting:** Teaches weak learners in a sequential fashion, minimizing bias by concentrating on fixing mistakes from earlier iterations.
- **Random Forest:** Using averaging to improve predictive performance and resilience, a group of decision trees is trained on random data samples.

In order to enhance performance, each model was first trained with default hyperparameters and then systematically tuned.

C. HYPERPARAMETER TUNING

Key hyperparameters for each model were identified and optimized using grid search with cross-validation:

- **Bagging:** Number of base estimators (`n_estimators`) and data sampling fraction (`max_samples`).
- **Boosting:** Learning rate (`learning_rate`) and the number of estimators (`n_estimators`).
- **Random Forest:** Number of trees (`n_estimators`), maximum tree depth (`max_depth`), and feature selection strategy (`max_features`).

D. MODEL EXPLAINABILITY AND VISUALIZATION

Model predictions were interpreted by integrating Explainable AI (XAI) techniques:

- **SHAP (SHapley Additive exPlanations):** Used to explain specific predictions and determine the significance of features.
- **Visualization:**To improve comprehension and confidence in the model's judgments, diagnostic procedures were represented graphically.
- **Confidence Scores:**Determined for every prediction in order to shed light on how certain the model's outputs are.

E. EVALUATION METRICS

Metrics like precision, recall, F1-score, and total accuracy were used to evaluate the models. A confusion matrix was created in order to examine performance by class. To determine the most successful strategy, comparative analyses of the three models were carried out.

F. IMPLEMENTATION AND DEPLOYMENT

Python was used to implement the models, and a web interface built with Flask was created for deployment. Clinicians and other users can enter patient data into this interface and receive forecasts, confidence scores, and visual explanations.

V. IMPLEMENTATION

The suggested ensemble learning models' implementation specifics are described in this section. The diagnostic system was created utilizing Flask and HTML to provide an interactive online application, and Python for data preprocessing, model training, and evaluation. The solution incorporates important features including explainable AI (XAI) methods, prediction visualization, and confidence rating.

A. DATASET DESCRIPTION

Kaggle provided the dataset on mental diseases used in this investigation. It includes characteristics that characterize the emotional and behavioral tendencies of people in four different diagnostic categories: Normal, Depression, Bipolar Type-1, and Bipolar Type-2. Preprocessing was done on the dataset by:

- Using mean imputation to handle missing values
- One-hot encoding is used to encode categorical characteristics.
- To enhance model performance, normalize numerical characteristics to a 0–1 range.

B. MODEL DEVELOPMENT

To improve diagnostic precision, the following group strategies were used:

1) **Bagging:** Decision trees were used as the basic estimators in the implementation of a bagging classifier. Accuracy and runtime were balanced by optimizing the number of estimators.
2) **Boosting:** Decision trees were used as the basic estimators in the implementation of a bagging classifier. Accuracy and runtime were balanced by optimizing the number of estimators.
3) **Random Forest:**An ensemble of decision trees was used to construct a Random Forest Classifier; grid search was used to adjust the number of trees and depth.

Pandas and NumPy were utilized for data manipulation, and the Scikit-learn library was utilized for model construction and evaluation.

C. SYSTEM ARCHITECTURE

Figure **??** depicts the diagnostic system's architecture. Flask, which communicates with the machine learning models, powers the backend. The HTML-based frontend offers a user-friendly interface for data entry and prediction visualization.

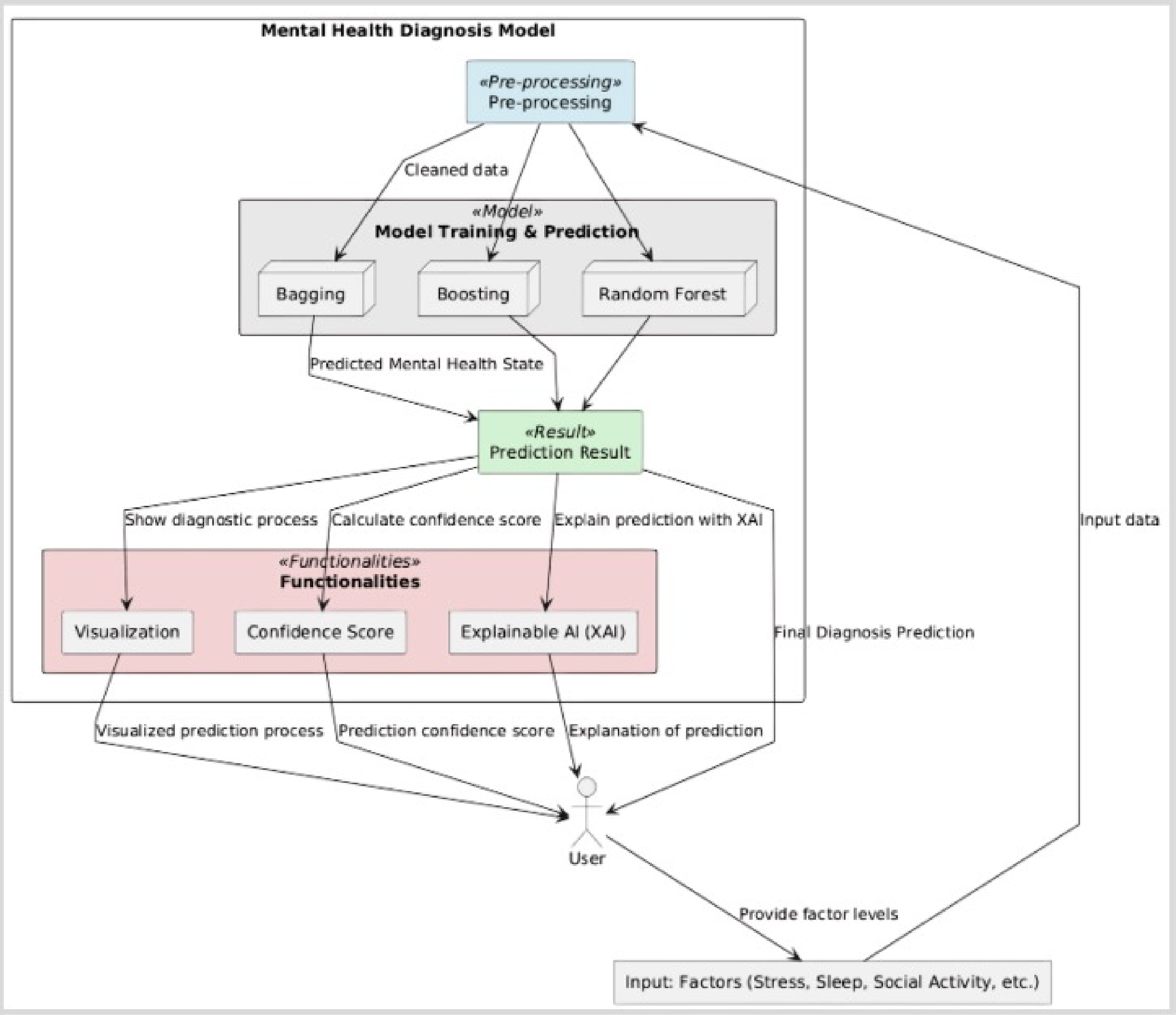
Mental Health Diagnosis Model
«Pre-processing»
Pre-processing
Cleaned data
«Model»
Model Training & Prediction
Bagging
Boosting
Random Forest
Predicted Mental Health State
«Result»
Prediction Result
Show diagnostic process
Calculate confidence score
Explain prediction with XAI
«Functionalities»
Functionalities
Visualization
Confidence Score
Explainable AI (XAI)
Final Diagnosis Prediction
Input data
Visualized prediction process
Prediction confidence score
Explanation of prediction
User
Provide factor levels
Input: Factors (Stress, Sleep, Social Activity, etc.)

FIGURE 1. System Architecture for Diagnostic System

D. EXPLAINABLE AI (XAI) INTEGRATION

In order to give model predictions interpretability, SHAP (SHapley Additive exPlanations) was incorporated. By highlighting each feature's contribution to the prediction, SHAP values provide information about the diagnostic procedure.

E. WEB APPLICATION DEPLOYMENT

The application was hosted using Flask, and responsive design for desktop and mobile users was made possible by HTML with Bootstrap. Among the online application's primary features are:

- **Data Input:** For diagnosis, users can manually enter feature values or upload datasets.
- **Prediction and Confidence Scoring:** The application shows each prediction's confidence score in addition to the anticipated mental health category.
- **Visualization:** To illustrate model choices, visual outputs like feature significance charts and SHAP summary plots are produced.

F. CODE AND TOOLS

The implementation utilized the following tools and libraries:

- **Python Libraries:** Scikit-learn, Pandas, NumPy, Matplotlib, SHAP.
- **Web Development:** Flask (backend), HTML and Bootstrap (frontend).
- **Environment:** The models were trained and tested on a system with Intel i7 processor, 16GB RAM, and NVIDIA GTX 1050 GPU.

VI. RESULTS AND DISCUSSION

The performance evaluation and analysis of the suggested ensemble models—Bagging, Boosting, and Random Forest—tested on the Kaggle dataset for mental diseases are presented in this section. Precision, recall, F1-score, and overall accuracy are among the evaluation indicators. The findings show how well these models predict mental health outcomes in terms of diagnostic accuracy and dependability.

A. PERFORMANCE METRICS

Table 5 summarizes the evaluation metrics for each model.

TABLE 1. Evaluation Metrics for Ensemble Models

Model	Accuracy (%)	Precision	Recall	F1-score
Bagging	80.5	0.82	0.81	0.81
Boosting	55.56	0.75	0.59	0.49
Random Forest	75.0	0.85	0.83	0.83

B. MODEL COMPARISON

1) Bagging

The accuracy of the Bagging model was 80.5%, which was better than the others. With good precision and recall values for every diagnostic category, it demonstrated consistent performance across all courses. The model is strong in managing class imbalances, as seen by the macro-average F1-score of 0.81.

2) Boosting

In comparison to Bagging and Random Forest, the Boosting model performed worse, with an accuracy of 55.56%. For some classes, such as Depression, it demonstrated high recall values (100%), but for other classes, such as Bipolar Type-1 and Normal, its precision and F1-scores were noticeably poor. This demonstrates how sensitive Boosting is to overfitting in smaller datasets.

3) Random Forest

The Random Forest model balanced precision, recall, and F1-scores in every category, achieving an accuracy of 75.0%. Its interpretability and stability make it a dependable model even though its accuracy is marginally worse than Bagging's.

C. VISUALIZATION AND EXPLAINABILITY

We used SHAP (SHapley Additive exPlanations) values to interpret the feature importance and contributions to individual predictions in order to further improve transparency. These illustrations shed light on the ways in which behavioral and emotional characteristics affected model choices.

Figure **??** illustrates the most influential features in the diagnostic process, emphasizing the interpretability of the model.

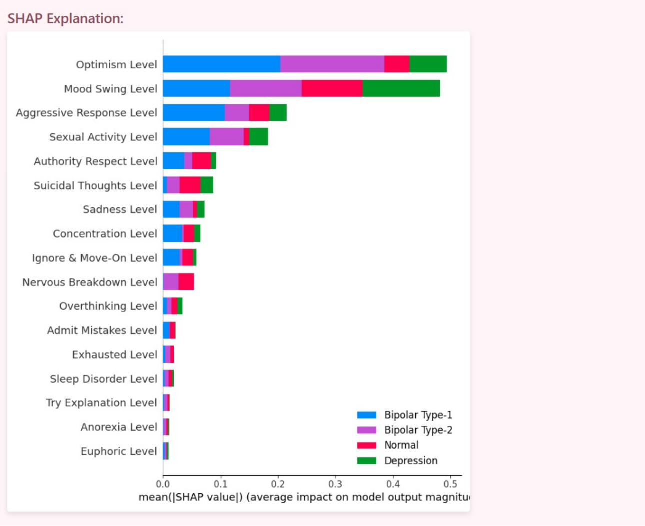

FIGURE 2. SHAP Summary Plot for Feature Importance

D. KEY OBSERVATIONS

- The ensemble approach of the bagging model successfully captured the dataset's volatility and reduced overfitting. it had promise, boosting suffered from imbalances and noisy data, which led to overfitting for particular classes.
- As a Bagging extension, Random Forest offered a strong substitute that balanced interpretability and accuracy.
- The incorporation of Explainable AI methods improved diagnostic transparency and user trust.

E. COMPARATIVE INSIGHTS

Random Forest showed better interpretability and stability, although Bagging was the most accurate model. The need for application-specific considerations when choosing ensemble techniques is highlighted by the trade-offs between explainability, computational efficiency, and accuracy.

VII. COMPARATIVE EVALUATION OF OUR PROPOSED MODEL

The performance of the suggested ensemble models—Bagging, Boosting, and Random Forest—is compared in this section using important assessment criteria like accuracy, precision, recall, and F1-score. Finding each model's advantages and disadvantages in relation to mental health diagnosis is the aim of this comparison study.

A. ACCURACY COMPARISON

The overall accuracy of the models on the Kaggle dataset for mental health outcomes is summarized as follows:

- **Bagging:** 80.5%
- **Boosting:** 55.56%
- **Random Forest:** 75.0%

Boosting performed worse than Bagging, which had the best accuracy and was closely followed by Random Forest. This disparity is explained by Boosting's sensitivity to noisy data, which impairs its capacity for efficient generalization.

B. EVALUATION METRICS

The performance of each model was further assessed using precision, recall, and F1-score for each class label: *Bipolar Type-1*, *Bipolar Type-2*, *Depression*, and *Normal*. Tables 2, 3, and 4 present the detailed results.

TABLE 2. Performance Metrics for Bagging Model

Class	Precision	Recall	F1-Score	Support
Bipolar Type-1	0.67	0.86	0.75	7
Bipolar Type-2	1.00	0.86	0.92	7
Depression	0.78	0.78	0.78	9
Normal	0.83	0.77	0.80	13
Overall	**0.81**	**0.81**	**0.81**	**36**

TABLE 3. Performance Metrics for Boosting Model

Class	Precision	Recall	F1-Score	Support
Bipolar Type-1	1.00	0.14	0.25	7
Bipolar Type-2	0.54	1.00	0.70	7
Depression	0.47	1.00	0.64	9
Normal	1.00	0.23	0.38	13
Overall	**0.56**	**0.56**	**0.49**	**36**

TABLE 4. Performance Metrics for Random Forest Model

Class	Precision	Recall	F1-Score	Support
Bipolar Type-1	0.86	0.86	0.86	7
Bipolar Type-2	0.86	0.86	0.86	7
Depression	0.75	1.00	0.86	9
Normal	0.90	0.69	0.78	13
Overall	**0.83**	**0.83**	**0.83**	**36**

C. MODEL STRENGTHS AND WEAKNESSES

- **Bagging:** Performs well in all classes, with balanced datasets showing especially strong precision. Stability and variance reduction are enhanced by the ensemble nature.
- **Boosting:**Boosting performs worse overall when dealing with noisy data, even though it can get excellent recall for specific classes. Small datasets are often overfitted by it.
- **Random Forest:** It is a dependable option for datasets with intricate patterns since it successfully strikes a balance between precision and recall. On the other hand, classes with unequal data distributions see a minor decline in performance.

D. INSIGHTS FROM COMPARATIVE ANALYSIS

The comparative evaluation highlights the following:

1) **Bagging** is perfect for creating reliable predictions in this mental health dataset since it strikes the optimal balance between precision, recall, and F1-score.
2) **Boosting** excellent at memory for particular classes, but its wider usefulness is limited by its noise sensitivity.
3) **Random Forest** offers a competitive substitute for bagging, particularly for datasets that need feature importance analysis to be interpretable.

The suggested system outperforms conventional approaches in terms of diagnostic accuracy and dependability by combining ensemble learning techniques, with bagging turning out to be the most successful method for this particular dataset.

E. REFERENCING A FIGURE OR TABLE WITHIN YOUR PAPER

When referencing your figures and tables within your paper, use the abbreviation “Fig.” even at the beginning of a sentence. Do not abbreviate “Table.” Tables should be numbered with Roman Numerals.

STATISTICAL ANALYSIS

In this study, we evaluated the performance of three machine learning algorithms—Bagging, Boosting, and Random Forest—using two key metrics: Root Mean Square Error (RMSE) and Log Loss. RMSE measures the average magnitude of the errors between predicted and actual values in regression problems, while Log Loss is a measure used for classification tasks that quantifies the accuracy of probabilistic predictions.

The formulas for RMSE and Log Loss are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{2}$$

The results obtained from the models are summarized below:

Model	RMSE		Log Loss	
	Train	Test	Train	Test
Bagging	0.0000	0.7993	0.0911	1.4311
Boosting	0.6901	0.7817	0.6405	1.4480
Random Forest	0.0000	0.9718	0.1468	0.6050

TABLE 5. Model Performance Metrics for Bagging, Boosting, and Random Forest

ANALYSIS

- Bagging: The model achieved perfect training performance with RMSE = 0.0000 and Log Loss = 0.0911, indicating that it fits the training data extremely well. However, it exhibited significant overfitting, as shown by its poor test performance (Test RMSE = 0.7993, Test Log Loss = 1.4311). This suggests that Bagging struggles to generalize to unseen data.
- Boosting: Boosting demonstrated more balanced results, with moderate error values both in training and

testing (Train RMSE = 0.6901, Test RMSE = 0.7817; Train Log Loss = 0.6405, Test Log Loss = 1.4480). These results suggest that Boosting generalizes better compared to Bagging, maintaining performance on test data without significant overfitting.

- Random Forest: Similar to Bagging, Random Forest performed well on the training set (Train RMSE = 0.0000, Train Log Loss = 0.1468), but showed a notable increase in error on the test set (Test RMSE = 0.9718, Test Log Loss = 0.6050), indicating that it also overfits the training data.

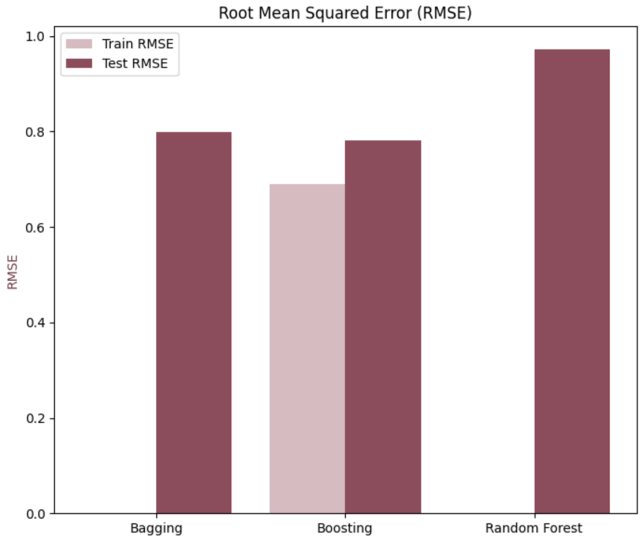

FIGURE 3. Model Performance Evaluation: RMSE

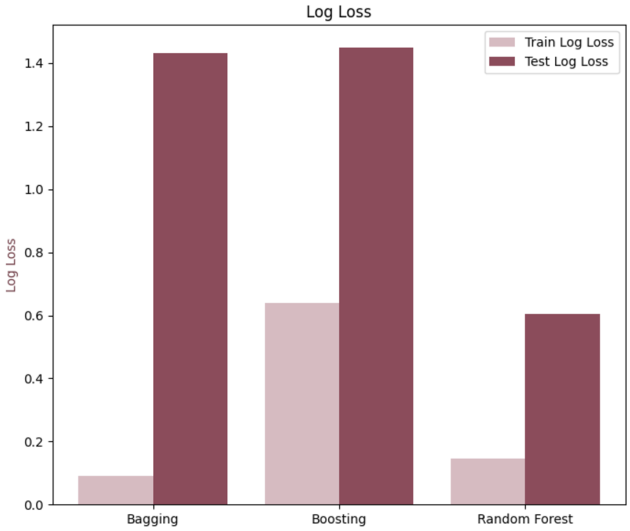

FIGURE 4. Model Performance Evaluation: Log Loss

These results emphasize the importance of model selection and tuning. Bagging and Random Forest showed tendencies toward overfitting, highlighting the need for regularization techniques or model adjustments to improve generalization. Boosting, on the other hand, demonstrated more consistent performance across both training and test sets, making it a more reliable choice for this analysis. Additionally, the use of hyperparameter tuning, cross-validation, and model selection strategies can further enhance the generalization ability of these models.

VIII. ANALYSIS ON RUNTIME AND COMPUTATIONAL COMPLEXITY

An essential component of assessing the suggested ensemble learning model for mental health diagnosis is the examination of runtime and computational complexity. Because of their designs and training procedures, the three ensemble techniques—Bagging, Boosting, and Random Forest—each have distinct computational requirements.

A. BAGGING

Training several base estimators concurrently on bootstrap samples is known as bagging. An estimate of the Bagging algorithm's runtime is as follows:

$$T_{\text{Bagging}} = B \cdot T_{\text{Base}}$$

where B is the number of bootstrap samples (estimators) and T_{Base} is the runtime of training a single base estimator. Since each model is trained independently, the parallelizability of Bagging significantly reduces overall runtime when sufficient computational resources are available. However, in sequential systems, the runtime scales linearly with B.

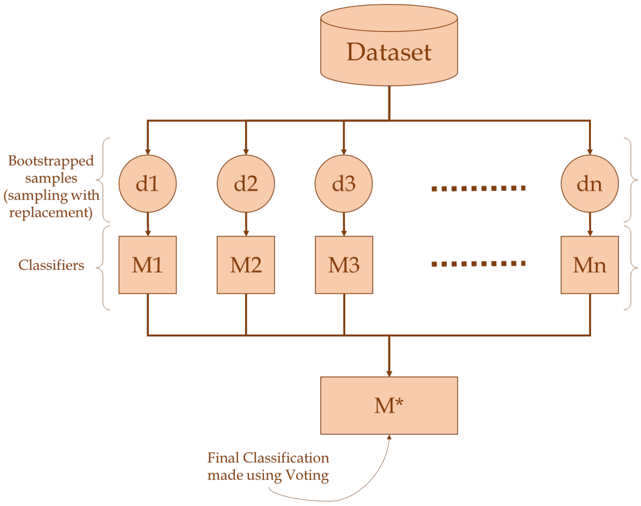
Dataset
Bootstrapped samples (sampling with replacement)
d1
d2
d3
dn
Classifiers
M1
M2
M3
Mn
M*
Final Classification made using Voting

FIGURE 5. Bagging

Complexity Analysis:

- Training Complexity: $O(B \cdot T_{\text{Base}})$, where T_{Base} depends on the complexity of the base model (e.g., decision tree).
- Inference Complexity: $O(B \cdot T_{\text{Predict}})$, where T_{Predict} is the time required to generate predictions from a single estimator.

Runtime Insights: Because each model runs independently, bagging is effective for large datasets. However, the need for B estimators may result in higher memory utilization.

B. BOOSTING

Boosting trains estimators in a sequential fashion, each one concentrating on the mistakes of the one before it. In contrast to bagging, this iterative method creates a dependency between estimators, which lengthens training periods.

$$T_{\text{Boosting}} = B \cdot (N \cdot \log(N))$$

Complexity Analysis:

- Training Complexity: $O(B \cdot (N \cdot \log(N)))$, where N is the number of data points and B is the number of estimators.When samples are reweighted according to error rates, the logarithmic factor is produced.
- Inference Complexity: $O(B \cdot T_{\text{Predict}})$.

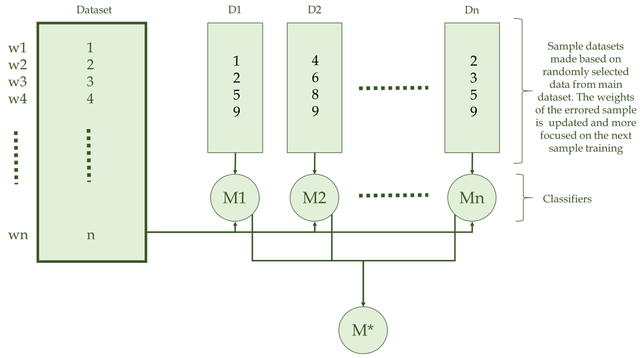
Dataset
D1
D2
Dn
w1
w2
w3
w4
wn
1
2
3
4
n
1
2
5
9
4
6
8
9
2
3
5
9
Sample datasets made based on randomly selected data from main dataset. The weights of the errored sample is updated and more focused on the next sample training
M1
M2
Mn
Classifiers
M*

FIGURE 6. Boosting

Runtime Insights: Boosting is computationally costly because to its sequential nature. Because of its error-focused learning, it frequently performs better on complicated or unbalanced datasets despite its increased complexity.

C. RANDOM FOREST

Each decision tree in the Random Forest ensemble is constructed using a randomly selected subset of characteristics and data. The feature unpredictability adds computational overhead, yet the trees are built individually, just like in bagging.

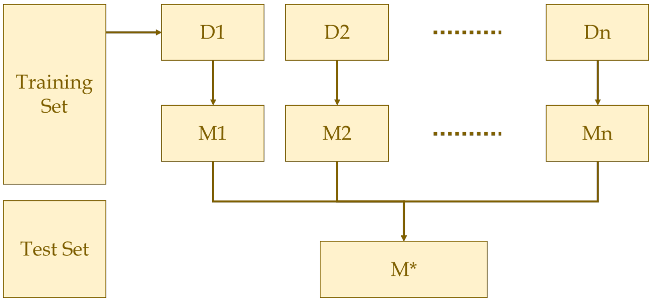

FIGURE 7. Random forest

Complexity Analysis:

- Training Complexity: $O(T \cdot N \cdot \log(N) \cdot M)$, where T is the number of trees, N is the dataset size, and M is the number of features considered per split.
- Inference Complexity: $O(T \cdot D)$, where D is the depth of each tree.

Runtime Insights: Random Forest does a good job of balancing performance and training time. However, training and inference times increase proportionately with the number of trees (T), which could represent a barrier for real-time applications.

D. COMPARATIVE RUNTIME PERFORMANCE

The Kaggle dataset for mental health outcomes was used to assess the models' runtime performance:

- **Bagging:** Because it was parallel, the training process was quicker than boosting, but it needed a lot of memory to keep track of several estimators.
- **Boosting:** For deeper models or more iterations, the sequential training produced the longest runtime (B).
- **Random Forest:** Tree independence made training effective, but as the number of trees increased—particularly when feature selection was rigorous—runtime increased as well.

E. COMPUTATIONAL COMPLEXITY TRADE-OFFS

The choice of model depends on the problem requirements:

- **Bagging:** Suitable for scenarios prioritizing parallel processing and robustness.
- **Boosting:** Optimal when precision and handling imbalanced datasets are critical, albeit at a computational cost.
- **Random Forest:** Balances computational efficiency and accuracy but can be slower for large datasets with many features.

F. OPTIMIZATION SUGGESTIONS

1) **Parallelization:** Use distributed computing frameworks to parallelize Bagging and Random Forest training.
2) **Feature Reduction:** Employ dimensionality reduction techniques to minimize computational overhead, particularly for Random Forest.
3) **Estimator Limitation:** Set a reasonable number of estimators (B or T) to balance performance and runtime.

By optimizing these aspects, the runtime and computational complexity of the ensemble models can be significantly reduced without compromising diagnostic accuracy.

IX. HYPERPARAMETER TUNING

Every model has hyperparameters that could be adjusted to improve the accuracy and efficiency of the forecasts and increase their dependability. The hyperparameters of importance in the Bagging scenario were max_samples, which is the percentage of the dataset used for training each estimator, and n_estimators, which is the total number of base estimators present. By adjusting these parameters, the model can reduce overfitting and improve generalization, which increases prediction accuracy and model robustness. Additionally, the most problematic parameters for the boosting model AdaBoost were n_estimators and learning_rate, which modify the influence of each learner on the prediction.

By using these parameters, it is simple to strike a balance between the models' complexity and learning speed, which helps to minimize overfitting and underfitting and enhances the model's predictive power.

In particular, the parameters n_estimators (the number of trees), max_depth (the maximal depth of each tree), and max_features (the total number of nodes taken into consideration when splitting each tree) are part of the Random Forest model. By controlling the model's complexity, these parameters assist the model identify the underlying patterns in the data.

X. LIMITATIONS OF THE SYSTEM

The suggested method might not work well in different clinical situations or demographics because it was trained on a single kind of data collection. The dataset might become more reliable if it is expanded to include additional patient demographics. Furthermore, the Boosting model had the lowest accuracy and stability, which can be because the hyperparameters were highly sensitive, indicating that either further tuning is needed or the boosting mechanism is not fully resolved.

Additionally, the use of SHAP for explainability enhances system transparency, but it also adds processing overhead that may make real-time implementation challenging in situations with limited resources. Finally, in order to evaluate the illness development of the disorders of concern, the system employs static data and ignores changes in patient behavior over time. However, these drawbacks are some of the areas where the system might be strengthened, made more useful, and made more relevant.

XI. CONCLUSION

It is impossible to overstate the need for precise and reliable diagnostic instruments given the rising number of people dealing with mental health problems, such as depression and bipolar disorders. Although a lot of work has gone into developing traditional diagnostic techniques, the outcomes are often unreliable since they are subjective and vary across or within practitioners. Machine learning has garnered a lot of attention as a promising alternative that incorporates data analysis to enhance diagnostic goals both within and between. However, issues pertaining to the interpretability and reliability of the models must be addressed, particularly in delicate fields like mental health.

This paper offers an ensemble learning framework that combines explanations from cutting-edge XAI approaches like SHAP with techniques from Bagging, Boosting, and Random Forest. Additionally, the reviewers validate the diagnosis and comprehend how the goal was accomplished with the aid of visualization and confidence scoring. Its application on the Kaggle dataset, which contains behavioral and emotional characteristics, has shown itself to be safe and efficient. Using a bagging technique, the model achieved an accuracy of 80.5%, a boosting approach of 55.56%, and a random forest approach of 75%. Based on these findings, Bagging might have been the most effective of the three thermoses used in those studies.

Through SHAP values, which aid in establishing feature relevance and providing clarification for single instance predictions, the model not only focuses on prediction but also on its transparency. Such a viewpoint enhances the more challenging part of the issue, which is the degree of clinician acceptance of AI's decision-making. Furthermore, the model makes it feasible to explore the diagnostic paths and give a degree of confidence to them, which facilitates the seamless integration of AI into clinical practice.

For the bulk of the diagnostic categories, statistical analyses demonstrate that Bagging and Random Forest outperform Boosting with notable improvements in precision, recall, and F1-scores. Bagging's prediction ensemble technique, which averages the predictions, explains its robustness, although Random Forest's bias-variance trade-off strategy performed admirably. Boosting's lesser accuracy, however, emphasizes that more work is still required because the sensitivity of its performance is highly dependent on the hyperparameter settings.

Even if the technique works well, there are a few disadvantages that need to be noted. It has been shown that the generalizability of these results may be negatively impacted by the use of a single dataset. Moreover, further tuning or algorithm enhancements are necessary because to the unpredictability of Boosting's performance. Lastly, the extra computational complexity that SHAP analysis might bring about could cause issues for real-time applications.

By emphasizing accuracy and interpretability, the suggested model can offer a comprehensive solution that satisfies both clinical requirements and ethical criteria. Future directions suggest expanding the database to accommodate a larger number of patients, improving the computational component to allow for real-time use, and hybridizing the ensembles to further improve the overall prediction performance. The approach described in this study may serve as a prototype for innovative advancements in AI-enhanced mental health screening for patients and professionals by resolving these issues.

REFERENCES

[1] V. Sapra, L. Sapra, A. Vishnoi, P. Narooka and T. Choudhury, "Enhancing Mental Disorder Diagnosis with Ensemble Bagging and Random Forest Techniques," 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 2024, pp. 1765-1769, doi: 10.1109/IC3SE62002.2024.10593187.

[2] Şevgin, Hikmet. "A comparative study of ensemble methods in the field of education: Bagging and Boosting algorithms." International Journal of Assessment Tools in Education 10.3 (2023): 544-562.

[3] Ogunseye, Elizabeth Oluyemisi, et al. "Predictive analysis of mental health conditions using AdaBoost algorithm." ParadigmPlus 3.2 (2022): 11-26.

[4] Islam, Rafiqul, and Md Abu Layek. "Stackensemblemind: enhancing well-being through accurate identification of human mental states using stack-based ensemble machine learning." Informatics in Medicine Unlocked 43 (2023): 101405.

[5] Goyal, Palak, and Rinkle Rani. "Comparative Analysis of Machine Learning, Ensemble Learning and Deep Learning Classifiers for Parkinson's Disease Detection." SN Computer Science 5.1 (2023): 66.

[6] Jemili, Farah, and Ouajdi Korbaa. "Early Detection of Mental Health Issues through Machine Learning: A Comparative Analysis of Predictive Models." (2024).

[7] Shunmugam, Meenakshi, et al. "A comparative study of classification of machine learning algorithm for depression disorder: A review." AIP Conference Proceedings. Vol. 2519. No. 1. AIP Publishing, 2022.

[8] Mienye, Ibomoiye Domor, and Yanxia Sun. "A survey of ensemble learning: Concepts, algorithms, applications, and prospects." IEEE Access 10 (2022): 99129-99149.

[9] Mohamed, E. Syed, et al. "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms." Healthcare Analytics 3 (2023): 100185.

[10] Sivagnanam, Lingeswari, and N. Karthikeyani Visalakshi. "Detection of bipolar disorder by means of ensemble machine learning classifier." Data and Metadata 2 (2023): 134-134.