Predictionof Mental HealthUsing Machine Learning

Yashwant Kumar[1], Sagar Mishra[2], Yogesh Kumar[3], Shashank Kumar Singh[4],Anirban Jyoti Hati[5]

[1]NIELIT,
Aurangabad, India,
yashwantkumar1928@gmail.com

[2]NIELIT,
Aurangabad, India,
sagarmishra9876@gmail.com

[3]NIELIT,
Aurangabad, India,
yogeshkumar@nielit.gov.in

[4]NIELIT,
Aurangabad, India,
shashank@nielit.gov.in

[5]NIELIT,
Aurangabad, India,
anirban@nielit.gov.in

***Abstract*:**
Mental health disorders have emerged as a paramount problem worldwide, affecting millions and presenting substantial hurdles to healthcare systems. In this research, a method for predicting mental health disorders utilizing extensive machine learning (ML) and artificial intelligence (AI) models is proposed. The proposed system integrates a range of sophisticated machine learning algorithms to analyze user inputs and predict potential mental health issues. By selecting the feature importance, we can select the best suitability model for prediction with high accuracy.

The algorithms, K-Neighbors Classifier, Decision Tree Classifier, Random Forest, Boosting and Stacking, are used to predict mental health. Among them, Boosting appears to be the best model based on its highest F1 score. The anticipated likelihood condition was evaluated to make an appropriate recommendation. We focus on college students and older adults, specifically adults older than 18 years. This demographic is at higher risk of mental health challenges due to academic and workplace stress, making early intervention crucial.

***Keywords*: Mental health, Machine learning Prediction, Stacking, Random Forest, Boosting.**

(Article history: Selected from 3rd NICEDT 2025, Ropar, 14-15 Feb 2025)

I. INTRODUCTION

The mental well-being is one of the major indices that can define the overall personality and mental status of an individual. An imbalance in neurological chemistry may trigger a number of mental illnesses, including anxiety disorders and depression. Screening is therefore important in the application of appropriate interventions in patients who manifest symptoms that define psychological distress. Nonetheless, the acute increase in mental health issues was seen globally, with an estimated 970 million people affected in 2019 by these disorders, mainly anxiety and depression [1]. The increasing trend in its incidence, coupled with an apparent shortage of mental health professionals at large—0.5 psychiatrists for every 100,000 people in low-income countries, as compared to 8.59 per 100,000 in high-income countries—is one of the paramount challenges for mental health care globally [2].

The results of mental health disorders do not end with the suffering of a person; they almost involve all spheres of life, including work performance, family life, social involvement, and community functions. In addition, social unrest, aggression, and widespread fear could further complicate the delivery of clinical services properly [2]. With increasing mental illness burden, there will also be a growing need for a valid and efficient diagnostic tool. Conventional diagnostic approaches often depend on patients' self-reported data and clinical assessments, which may be time-consuming to collect and subject to bias.

Recent developments in AI and ML open up opportunities for faster and more objective assessment through the quick analysis of large datasets without human biases. Such technologies can offer increased speed and accuracy in mental health diagnostics and provide scalable solutions with the ability to improve consistently but assess consistently. This transformation is fundamentally altering conventional diagnostic techniques, facilitating early intervention and customised treatment strategies that have the potential to avert the emergence or reoccurrence of mental health disorders [3].

In the present research, we have utilised an openly available dataset from the Open Sourcing Mental Illness initiative, which largely comprised data related to working individuals. This data set provides a very useful contribution to the understanding of the relationship between mental health and the workplace environment, benefiting both employers and employees by creating greater awareness and encouraging proactive intervention strategies. The current research uses a machine learning methodology to augment prediction precision, incorporating the model into an accessible web-based platform. The tool works in a very helpful way for early intervention purposes in mental health: it gives a probability score and suggests solutions as per the data inserted by the users.

II. LITERATURE SURVEY

The complexity involved in the aspects of this particular research challenge demands more deliberation. In a bid to identify the prevailing gaps in research, as well as the methods and approaches applied in previous studies, this section reviewed various related scholarly articles. During the last years, new achievements in AI and ML methodologies allowed the construction of more accurate predictive and decision-making systems.

Madhvan Bajaj et al. (2023) used a dataset comprising different individual demographic behaviours and clinical characteristics to predict mental health outcomes. They created their predictive models using four different machine learning algorithms: logistic regression, K-nearest neighbor, decision tree, and random forest. Using demographic data, the researchers were attempting to increase the accuracy of their mental health predictions [4].

Yifan Li (2023) followed up on that by using the Grid Search algorithm to find the optimum model parameters. In addition, the study focused on an analytical approach toward interpretable machine learning, enabling the field of feature importance assessment by means of a technique called permutation importance [5]. CH.M.H. Saibaba et al. show an association of lifestyle factors with mental health. The work utilizes the real-world data from social media augmented into a machine learning model to predict the outcomes for mental health [6].

Prithvipal Singh et al. (2022) used a large dataset for his study; hence, he first pre-processed data to prepare the data for further processing. For developing a system dependent on questionnaires, which estimated levels of anxiety, depression, and stress, they used two different kinds of machine learning algorithms: Random Forest and Decision Tree Classifier. The result obtained showed that the accuracy of the Random Forest algorithm is higher than that of the Decision Tree Classifier in predicting these disorders [7]. Dr. S. Vidya et al. propose a broad approach to mental health care, including health monitoring, AI chatbots, and emotional analysis extracted from social media sites. The system shall avail the use of the Random Forest algorithm to enhance the precision in disorder prediction for revolutionary treatment of mental health by providing continuous evaluations and instant notification to the caregivers or psychiatrists, hence bridging the gap between the patients and health care providers [8].

Laijawala et al. conducted an investigation into the current methodologies utilized for forecasting mental health disorders, which encompass the application of chatbots and internet-based surveys. These approaches capitalize on user feedback and behavioral assessments, utilizing the Random Forest algorithm to predict mental health conditions [9].

TABLE I. COMPARISION OF EXISTING MODELS

S. no.	Paper	Algorithm Used	Scope
1	Madhvan Bajaj et al.	logistic regression, KNN, decision tree, and random forest	The current study will combine demographic and behavioural data with machine learning algorithms in predicting the outcomes of mental health. Although promising, the methodology needs further work to improve the level of accuracy and decrease the number of false positives.
2	Yifan Li	Random forest.	The grid search technique was used to find the ideal model parameters. However, the model's accuracy need further refinement.
3	CH.M.H. Saibaba et al.	Random forests, Bagging, Boosting, Stacking, Logistic Regression, Decision tree classifier, KNN	Social media data was used to predict mental health outcomes. However, the reliability of this data remains a concern.
4	Prithvipal Singh et al.	Random Forest and Decision Tree Classifier	A very large dataset was used in the development of an ADS-based questioner system. However, the dataset size was not really well-defined, and it used only two algorithms, which calls for more improvement.

5	Dr. S. Vidya et al.	Random Forest	The accuracy of mental health predictions is highly dependent on the quality of data obtained from smartwatch sensors.
6	Laijawala et al.	Random Forest	This study primarily focuses on working individuals, which may limit the generalizability of the findings to other populations.

The present study fills these shortcomings using sophisticated ensemble methods and a high-quality dataset. Limited scope in previous research, current research tends to concentrate on either certain algorithms or small sets of data. By pointing this out, the paper legitimizes the requirement of a wider, more general solution. Generalizability problems most studies do not control for various demographics, which is the most important thing for making findings relevant to practical situations. This research deals with this by being based on both college students and working professionals.

Current contribution XGboost and Stacking have been found to provide a solid method of dealing with complex datasets, guaranteeing enhanced accuracy and scalability. This illustrates how the work addresses gaps in scalability and real-world application.

III. MATERIALSAND METHODS

A. Dataset

The dataset from Open Sourcing Mental Illness (OSMI) represents an organization committed to improving mental health within the technology sector by advocating for open-source solutions. OSMI engages in surveys and research initiatives aimed at collecting extensive data concerning mental health challenges experienced by professionals in the tech field. The dataset lacks sufficient representation of teenagers or children, limiting the generalizability to these age groups.

The data used is from the Open-Source Mental Illness (OSMI) dataset, collected from over 1,200 responses. This dataset covers many features, including the age of the subject, gender, self-employed, family history, treatment, work environment, anonymity degree, perceived advantages, job leave policies, and 17 other supplement features. Given the large size of the dataset, we began with extensive cleaning. The process followed included dealing with any missing values (NAN) and handling them through suitable imputation techniques or eliminating them. This was followed by the use of label encoding to transform categorical variables into numerical representations, a step crucial for efficient analysis. Following the encoding process, an elaborate validation had been done to ensure that no residual missing values existed within the dataset, hence confirming the integrity and adequacy of the dataset for further analysis.

The data encoding process focuses on transforming categorical variables into numerical formats that machine learning models can interpret. Typically, label encoding is used for this, assigning a distinct integer value to each category. This transformation is crucial for models that require numerical input, enabling them to process and learn from the categorical features effectively. The encoded information is then used in the modelling process, hence ensuring that all the features are represented in a form that is suitable for the algorithms used.

1	Age	Gender	Country	state	self_empl	family_his	treatment	work_inte	no_emplo	remote_w	tech_com	benefits	care_opti	wellness_	seek_help	anonymity	leave	mental_he	phys_heal	coworkers	supervisor	mental_he	phys_heal	mental_v
2	37	Female	United Sta	IL	NA	No	Yes	Often	Jun-25	No	Yes	Yes	Not sure	No	Yes	Yes	Somewha	No	No	Some of t	Yes	No	Maybe	Yes
3	44	M	United Sta	IN	NA	No	No	Rarely	More thar	No	No	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	Maybe	No	No	No	No	No	Don't kno
4	32	Male	Canada	NA	NA	No	No	Rarely	Jun-25	No	Yes	No	No	No	No	Don't kno	Somewha	No	No	Yes	Yes	Yes	Yes	No
5	31	Male	United Kin	NA	NA	Yes	Yes	Often	26-100	No	Yes	No	Yes	No	No	No	Somewha	Yes	Yes	Some of t	No	Maybe	Maybe	No
6	31	Male	United Sta	TX	NA	No	No	Never	100-500	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Don't kno	No	No	Some of t	Yes	Yes	Yes	Don't kno
7	33	Male	United Sta	TN	NA	Yes	No	Sometime	Jun-25	No	Yes	Yes	Not sure	No	Don't kno	Don't kno	Don't kno	No	No	Yes	Yes	No	Maybe	Don't kno
8	35	Female	United Sta	MI	NA	Yes	Yes	Sometime	01-May	Yes	Yes	No	No	No	No	No	Somewha	Maybe	Maybe	Some of t	No	No	No	Don't kno
9	39	M	Canada	NA	NA	No	No	Never	01-May	Yes	Yes	No	Yes	No	No	Yes	Don't kno	No	No	No	No	No	No	No
10	42	Female	United Sta	IL	NA	Yes	Yes	Sometime	100-500	No	Yes	Yes	Yes	No	No	No	Very diffic	Maybe	No	Yes	Yes	No	Maybe	No
11	23	Male	Canada	NA	NA	No	No	Never	26-100	No	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	No	No	Yes	Yes	Maybe	Maybe	Yes
12	31	Male	United Sta	OH	NA	No	Yes	Sometime	Jun-25	Yes	Yes	Don't kno	No	No	No	Don't kno	Don't kno	No	No	Some of t	Yes	No	No	Don't kno
13	29	male	Bulgaria	NA	NA	No	No	Never	100-500	Yes	Yes	Don't kno	Not sure	No	No	Don't kno	Don't kno	No	No	Yes	Yes	Yes	Yes	Don't kno
14	42	female	United Sta	CA	NA	Yes	Yes	Sometime	26-100	No	No	Yes	Yes	No	No	Don't kno	Somewha	Yes	Yes	Yes	Yes	Maybe	Maybe	No
15	36	Male	United Sta	CT	NA	Yes	No	Never	500-1000	No	Yes	Don't kno	Not sure	No	Don't kno	Don't kno	Don't kno	No	No	Yes	Yes	No	No	Don't kno
16	27	Male	Canada	NA	NA	No	No	Never	Jun-25	No	Yes	Don't kno	Not sure	Don't kno	Don't kno	Don't kno	Somewha	No	No	Some of t	Some of t	Maybe	Yes	Yes
17	29	female	United Sta	IL	NA	Yes	Yes	Rarely	26-100	No	Yes	Yes	Not sure	No	No	Don't kno	Somewha	No	No	Yes	Some of t	Maybe	Maybe	Don't kno
18	23	Male	United Kin	NA	NA	No	Yes	Sometime	26-100	Yes	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Very easy	Maybe	No	Some of t	No	Maybe	Maybe	No
19	32	Male	United Sta	TN	NA	No	Yes	Sometime	Jun-25	No	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Maybe	No	Some of t	Yes	No	No	No
20	46	male	United Sta	MD	Yes	Yes	No	Sometime	01-May	Yes	Yes	Yes	Not sure	Yes	Don't kno	Yes	Very easy	No	No	Yes	Yes	No	Yes	Yes
21	36	Male	France	NA	Yes	Yes	No	NA	Jun-25	Yes	Yes	No	No	Yes	No	Yes	Somewha	No	No	Some of t	Some of t	Maybe	Maybe	Don't kno
22	29	Male	United Sta	NY	No	Yes	Yes	Sometime	100-500	No	Yes	Yes	Yes	No	No	No	Somewha	Maybe	No	Some of t	Some of t	No	No	No
23	31	male	United Sta	NC	Yes	No	No	Never	01-May	Yes	Yes	No	No	No	No	Yes	Somewha	No	No	Some of t	Some of t	No	Maybe	Yes
24	46	Male	United Sta	MA	No	No	Yes	Often	26-100	Yes	Yes	Yes	Yes	No	No	Don't kno	Don't kno	Maybe	No	Some of t	Yes	No	Maybe	No
25	41	Male	United Sta	IA	No	No	Yes	Never	More thar	No	No	Don't kno	No	No	Don't kno	Don't kno	Don't kno	Maybe	No	No	No	No	Yes	Don't kno

Fig. 1. Dataset description in samples

Fig. 1. consists of 1,259 observations and 27 fields, including responses to surveys about workplace mental health. The main information is age, gender, country, working status, history of mental illness in the family, treatment received, and how work is impacted by mental health. It also includes workplace items such as firm size, benefits, and mental health policy. A few columns, e.g., state, self_employed, work_interfere, and comments, have missing data, with comments having mostly null values. The data is useful in learning about mental health issues and support structures in workplaces.

B. Implemented Steps

The following steps can be followed to implement the proposed technique for determining if a person need medical care for their mental health issue.

a) Data Collection and Loading:

First, the data related to mental health will be filtered from the OSMI dataset. Then, it imports the dataset into an environment with the help of a pandas library and does initial exploration to understand how the data looks and what elements it has.

b) Exploring Data:

EDA is done to determine insight from the dataset, which ranges from analysing the data dimensionality to understanding feature distribution through descriptive statistics and missing values or outliers. EDA can be useful in making intelligent decisions about the preprocessing steps required.

c) Data Preparation:

The preprocessing step involves several critical tasks:

i. Handling Missing Values: In this, missing values of the data are handled through data imputation by using mean, median, mode value, or simply removing those records.
ii. Encoding Categorical Variables: Categorical features are numerical values after the use of label encoding, which makes them appropriate for machine learning models.
iii. Feature Scaling: Numerical attributes should be scaled or normalized so that they have the same range. Scaling can be achieved using techniques like MinMaxScaler. This preprocessing is particularly crucial if the model is sensitive to the magnitudes of input features as it improves the performance and convergence.

d) Model Selection:

The following various machine learning algorithms have been selected and then implemented to predict mental health outcomes: These include Random Forest, Stacking, Xgboost, Bagging, K-Neighbours', and Decision Tree classifier.

e) Training and Evaluation Models:

The dataset is then divided into training and test subsets to compare the models fairly. Cross-validation can also be used to check the strength of the models; in this, one trains and tests the data on different subsets. Model performance is quantified with the help of various metrics, such as accuracy, precision, recall, and F1-score.

f) Ensemble Learning and Stacking Techniques:

Ensemble learning methodologies, therefore, including Random Forest and Boost, are employed to enhance these regards for reliability and precision of forecasts. This, in turn, is accompanied by a stacking classifier to combine model predictions with benefits in building a stronger predictive framework.

g) Final Model Selection and Deployment:

After all models are evaluated, the best model-one that shows the best cross-validation results and overall metrics-is fine-tuned and prepared for deployment. This is to be used in predicting mental health outcomes using new data.

h) Development of Web Interfaces Using Flask:

Web Application Configuration: A Flask-based web application has been developed to provide an intuitive interface to interact with the predictive model. It allows users to input data and obtain real-time predictions.

Integration into the Model: In this model, it is integrated with the Flask application. Whenever any data is input by the user via the web application, the application pre-processes the input and feeds it into the model, which makes the prediction.

Deployment: The Flask application is deployed on a server, hence making it accessible to users via the web browser interface. This setupmakes access to the model easier for those who do not know how to write code.

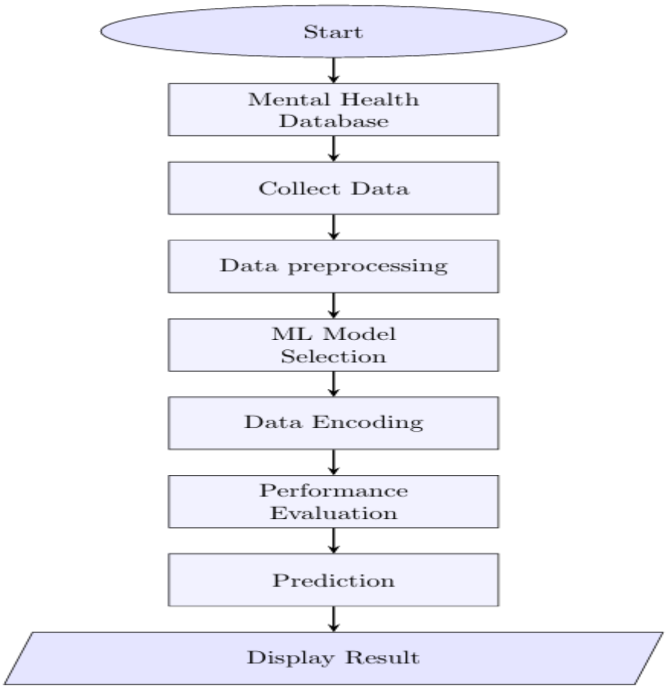

Fig. 2. Flow chart of model construction

Figure 2. demonstrates the implemented system's block diagram. It shows the different system stages and modules as well as the order in which they are completed during the project.

C. Algorithms Attempted for Model Development

i. Decision Tree

The decision tree is a non-parametric supervised learning technique that is adopted for tasks involving classification and regression. This algorithm has a tree-like hierarchical structure having a root node, branches, internal nodes, and leaf nodes [10].A widely used supervised learning algorithm, it is applied for classification and regression problems alike. It looks more or less like a flowchart-pretty similar to a tree structure consisting of nodes, branches, and leaves. The algorithm works by recursively partitioning the data space based on whichever criteria may decide to consider into increasingly smaller subspaces until predefined stopping conditions are met Provides high interpretability, which is essential for understanding feature importance in mental health datasets. However, it is prone to overfitting.

ii. Random Forest

Random Forest is that ensemble learning algorithm where it has been designed for classification and regression purposes. It makes use of constructing many decision trees on different subsets of the dataset and then combining their predictions to enhance accuracy, which helps in minimizing overfitting risk along with having a better model performance through individual outputs [11].Overcomes Decision Tree’s limitations by combining multiple trees to reduce overfitting and enhance performance.

iii. Stacking

Stacking is an ensemble-based technique in machine learning, including multiple weak learners that make a prediction independently. In this technique, different models created using diversity are combined to develop a more accurate and robust prediction system [12].

In stacking, an advanced kind of machine learning, predictions from various kinds of weak learners are used as input data to Meta learners. These Meta learners become important in that they carefully analyse the combined predictions produced by the weak learners. Combines predictions from multiple models for improved accuracy, handling complex feature interactions effectively.

iv. Boosting(XGBoost)

Boosting is a sophisticated, advanced, and respected method of supervised machine learning; it effectively integrates the predictions generated by various weak models, normally referred to as base models, into a strong and highly effective ensemble model. Unlike other classic ensemble techniques, such as bagging or averaging, boosting does not grant the models an equal role [13].

XGBoost technique, in turn, focuses on the sequential training of these basic models. It pays special attention to the misclassified samples of previous iterations, with the aim of making the models learn from their mistakes and improve at each step. Sequentially focuses on correcting previous errors, making it ideal for handling imbalanced datasets often found in mental health studies.

v. K-Neighbors

K-nearest neighbors (KNN) is a classifier that operates on a non-parametric algorithm and is based on supervised learning principles. This specific type of algorithm takes into consideration the proximity of data points in making informed classifications or predictions about the categorization of a specific individual data point in a dataset. KNN is one of the simplest and most commonly used classifiers for both classification and regression problems in today's machine learning domain [14].Effective for smaller datasets and interpretable, it groups data points based on proximity, which is useful for identifying patterns in mental health indicators.

D. Experimental Setup

The execution of the suggested mental health prediction model was performed through Python as the main programming language. Different libraries and frameworks have been used, with Scikit-learn being utilized for machine learning model execution, Pandas and NumPy used for data preparation and numerical calculation, and Matplotlib and Seaborn applied for data plotting and exploratory data analysis. The predictive model has also been coupled with a web application built from Flask to give the model an interactive user interface where users are able to key in data and obtain real-time predictions. Jupyter Notebook run under the Anaconda Distribution served as the laboratory for experiments.

IV. RESULT AND ANALYSIS

The dataset was then divided into two disjoint subsets: 70% of the data for model training and the remaining 30% for testing. The test set will be used to evaluate the performance of the model using metrics such as accuracy, recall, precision, and F1 score. That was one way of determining the best-optimized model that could provide better reliability and overall efficacy in prediction.Accuracy, F1 score, precision, and recall can be computed in machine learning by the formulas below. The following terms are common when one describes a classification model evaluation:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (3)$$

$$F1\,Score = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \quad (4)$$

1. TP: True Positive
2. TN: True Negative
3. FP: False Positive
4. FN: False Negative

TABLE II. THE EFFECTIVENESS OF ALGORITHMS FOR MACHINE LEARNING.

ML Models	**Performance**			
	Accuracy	***Recall***	***Precision***	***F1 Score***
Random Forest	0.8121	0.9304	0.75	0.8305
Decision Tree	0.8068	0.9358	0.7415	0.8274
Stacking	0.8201	0.8502	0.7989	0.8238
XGBoost	0.8174	0.9197	0.7610	0.8329
K-Neighbors	0.8042	0.9037	0.7511	0.8203

Table II is calculated from the equation 1,2,3 and 4. It represents the results of many algorithms. XGBoost turns out to be the best model; its F1 score of 0.8329 is much higher, which means an outstanding balance between precision and recall. Besides, it has a really high value of accuracy, 0.8174, and a recall of 0.9197, which makes it suitable for many scenarios, especially when one needs an optimal trade-off between complete detection of positive cases and minimization of false alarms.

We identified several methodologies and datasets for the prediction of mental health through complete reviews of available literature. Many current methods are insufficient, with decreased accuracy and increased frequency of false alarms. Some of these methods have a limited scope of variables for input, poorly access large datasets, rely heavily on the accuracy of wearable devices, or depend on social media data, which often lacks in reliability or precision. To tackle these challenges, we have used a clear open-source dataset, which does involve more estimation developed over using a traditional dataset. We have experimented with results on many models and selected one among them to be the best due to its performance and efficiency. The model selected was implemented on a web application designed to predict mental health outcomes by giving its users acertain probability.

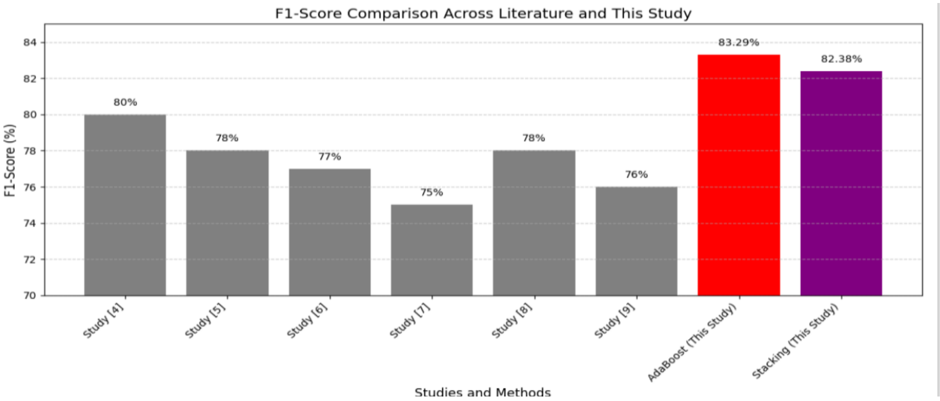

Fig. 3. Comparison with Existing Results.

Most of the studies attained below 80% accuracy. The bar chart illustrates F1-scores of different studies compared with the findings of the research. The y-axis is the percentage of the F1-score ranging from 70% to 84%, and the x-axis is a list of different studies and techniques. The gray bars are F1-scores from previous research ranging from 75% to 80%. On the other hand, the models applied in this study—XGBoost and Stacking—are represented by red and purple bars, respectively. The XGBoost model has the best F1-score of 83.29%, which is higher than all previous studies. The Stacking model is also very good with an F1-score of 82.38%, ranking it as the second-best performing technique. Compared with existing research, both models in this study show a substantial improvement in classification performance. This indicates the effectiveness of the proposed methods, especially XGBoost, which has the highest accuracy among all techniques tested.

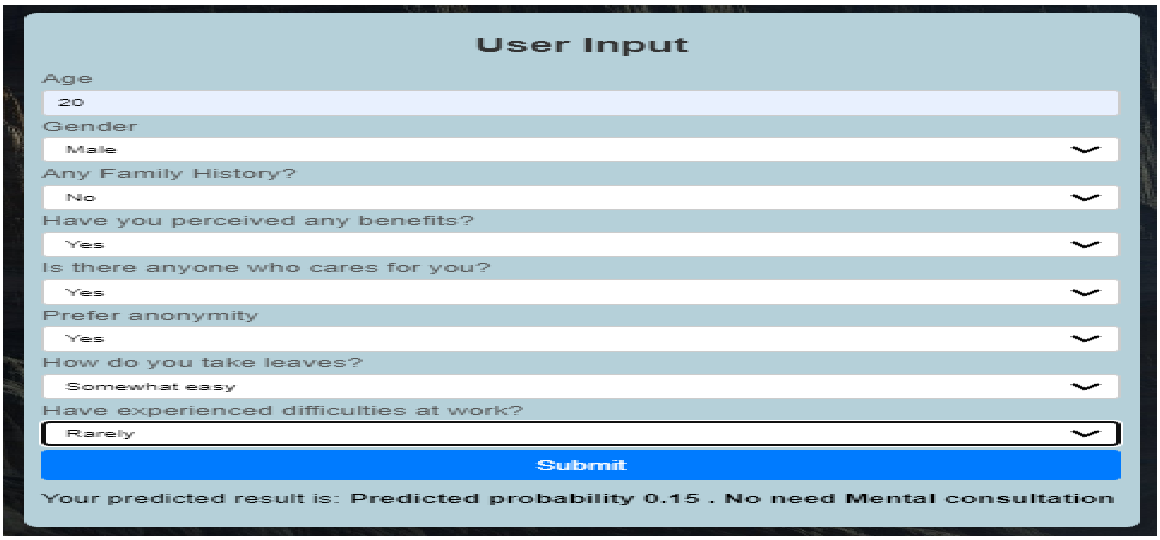

Fig. 4. Outcome with a 0.15 risk forecast.

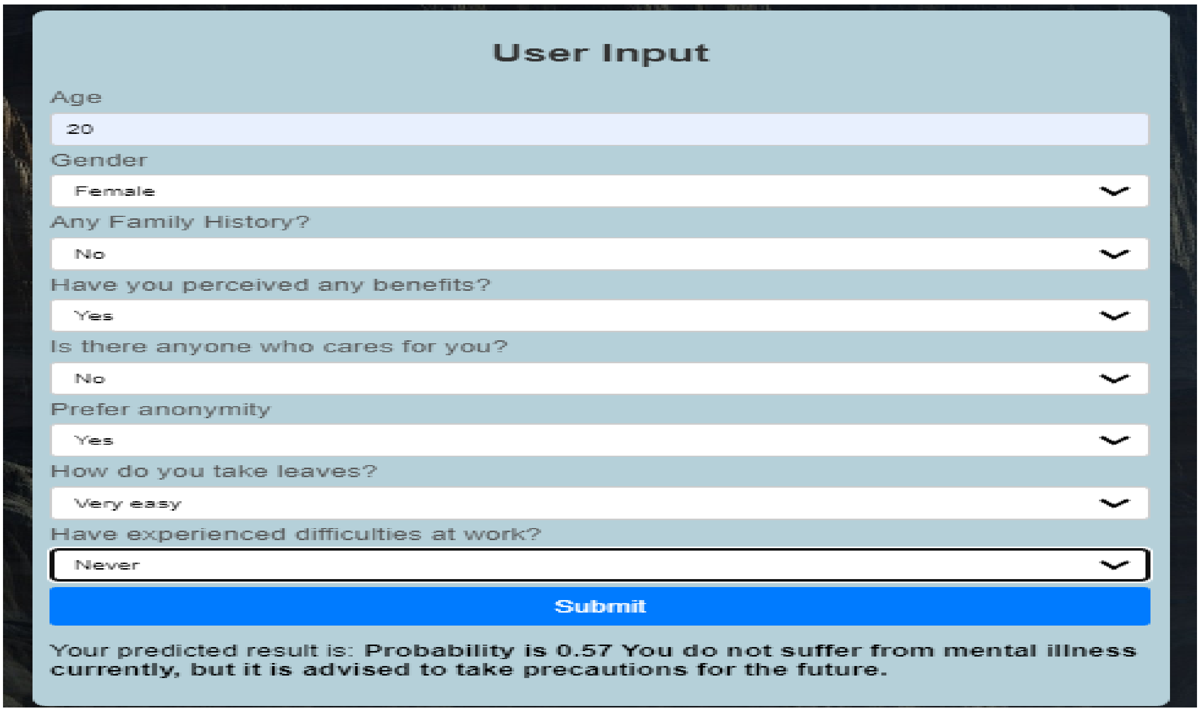

Fig. 5. Outcome with a 0.57 risk forecast.

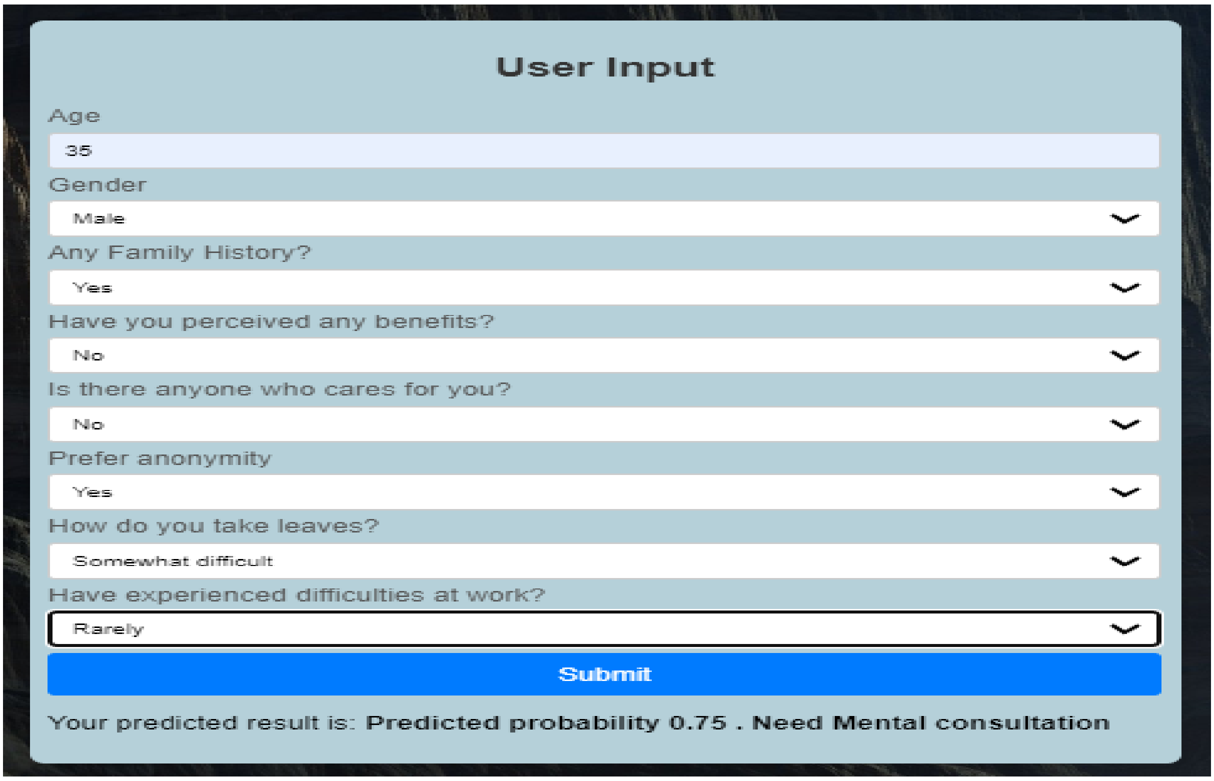

Fig. 6. Outcome with a 0.75 risk forecast.

The outputs depending on various inputs that users provide to the website are shown in Figures 4, 5 and 6. For varying probabilities of mental disease, there is a different guideline. A likelihood of less than 0.30 suggests that the user is not mentally ill, according to the way the system is structured. A likelihood of 0.57 or higher indicates that the user is mentally ill, whereas a probability of 0.3 to 0.57 implies that the user might experience mental illness in the future.

V. CONCLUSIONAND FUTURE SCOPRE

In the present study, we investigated predicting mental health disorders with machine learning methods, including K-Neighbors Classifier, Decision Tree, Random Forest, Boosting (XGBoost), and Stacking. Results reveal that XGBoost was the best performer with an F1-score of 0.8329 and indicates its ability to cope with complex data sets and imbalanced data. The present study fills significant gaps recognized in previous research by integrating ensemble learning techniques, enhancing predictive accuracy, and generalizability. By using boosting and stacking techniques as well, the study copes well with class imbalance, an area of special concern in mental health prediction. Besides, we rationalized the emphasis on college students and adults over the age of 18 as being more prone to mental health issues due to academic and work-related stress. Lack of dataset for children and below 18 years it need to future work aims to enhance the study by expanding the dataset.

The constructed Flask-based web application guarantees ease of use for non-technical users, enabling early intervention. This research enhances current approaches by using strong datasets, transparent model interpretability, and real-world deployment techniques. Future versions will enhance clinical and practical usefulness by differentiating between particular mental health disorders like depression, anxiety, and PTSD. This will be achieved by training models using labelled datasets for these conditions, enabling more accurate and meaningful predictions. Future work can further enhance disorder-specific predictions, incorporate more datasets, and investigate deep learning methods to improve mental health diagnosis. The suggested approach provides a solid foundation for scalable and reliable mental health assessment tools, enabling early intervention and better mental well-being outcomes.

VI. REFERENCES

[1] World Health Organization, “Mental health,” Available at: https://www.who.int/health-topics/mental-health#tab=tab_2.

[2] Pure Edinburgh Repository, “Mental Health and Global Challenges,” discussing the disparity in psychiatrist availability across income levels and its impact on mental health care.

[3] E. Syed Mohamed, Tawseef Ahmad Naqishbandi, Syed Ahmad Chan Bukhari, Insha Rauf, Vilas Sawrikar, Arshad Hussain. "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms", Healthcare Analytics, 2023.

[4] Madhvan Bajaj, Satvik Vats, Priyanshu Rawat, Vikrant Sharma, Diksha, Lisa Gopal. Prediction of mental health treatment adherence using Machine learning algorithm. CICTN 2023.DOI: 10.1109/CICTN57981

[5] Yifan Li. Application of machine learning to predict mental health disorders and interpret feature importance.2023 ISCTIS. DOI: 10.1109/ISCTIS58954.

[6] Ch.M.H. Saibaba, K V K Alekhya, K Yeshwanth, Praveen Tumuluru."Prediction of Public Mental Health by using Machine Learning Algorithms",ICAIS-2022. DOI: 10.1109/ICAIS53314.

[7] Prithvipal Singh, Gurvinder Singh, Sarveshwar Bharti."The Predictive Model of Mental Illness using Decision Tree and Random Forest classification in Machine Learning".ICACITE-2022. DOI: 10.1109/ICACITE53722.

[8] Dr.S. Vidya, S. Inbasekar, U. Vasanthakumaran, A. Venkatesh, "A SMART WATCH APPLICATION TO DETECT PSYCHOLOGICAL STATE OF MENTALLY DISTURBED PATIENTS"ICCPC 2022

[9] Vidit Laijawala, Aadesh Aachaliya, Hardik Jatta, Vijaya Pinjarkar. "Classification Algorithms based Mental Health Prediction using Data Mining", 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020.

[10] IBM. "What is a Decision Tree?" Available at: https://www.ibm.com/topics/decision-trees.

[11] JavaTpoint. "Machine Learning - Decision Tree Classification Algorithm." Available at: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[12] JavaTpoint. "Machine Learning - Random Forest Algorithm." Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm

[13] JavaTpoint. "Machine Learning - Boosting." Available at: https://www.javatpoint.com/machine-learning-boosting

[14] JavaTpoint. "Machine Learning - K-Neighbors Algorithm." Available at: https://www.javatpoint.com/machine-learning-K-Neighbors