

Advancing Mental Health Prediction with Machine Learning: Methods, Challenges, and Future Directions

Salman Shah¹, Sumit Kumar², Ajoy Roy³, Arghya Ghosh⁴, Sanchita Das⁵, Sudip Samanta⁶

¹ MCA STUDENT, Department of CSS, Brainware University, Barasat, Kolkata, India

² MCA STUDENT, Department of CSS, Brainware University, Barasat, Kolkata, India

³ MCA STUDENT, Department of CSS, Brainware University, Barasat, Kolkata, India

⁴ MCA STUDENT, Department of CSS, Brainware University, Barasat, Kolkata, India

⁵ MCA STUDENT, Department of CSS, Brainware University, Barasat, Kolkata, India

⁶ ASSISTANT PROFESSOR, Department of CSS, Brainware University, Barasat, Kolkata, India

Abstract

Growing global disease burden of mental illnesses necessitates advanced prediction models for early treatment. In this study, the machine learning (ML) models like Convolutional Neural Networks (CNN), Random Forest (RF), Support Vector Machines (SVM), and Deep Neural Networks (DNN) are compared for disease prediction like depression, anxiety, bipolar disorder, and schizophrenia from electronic health records (EHRs) and survey. Despite their having provided promising accuracy (e.g., 93.2% in the case of bipolar disorder in CNN), issues persist: sparsity of data, shallow interpretability of deep learning, single-source reliance, and privacy and bias issues. In this paper, an AdaBoost-guided model with a preprocessed mental health survey data set of 1,259 samples that integrates multimodal data (wearables, EHRs) is presented to bridge gaps. We tune feature engineering, handling missing values, and minimizing bias. AdaBoost model was accurate at 88.7%, outperforming RF (85.2–89.7%) and SVM (88.4%), with the confusion matrix indicating fair performance (TP=142, TN=128, FP=31, FN=23) and understandable feature importance (e.g., family_history=0.24). Findings support the balance of accuracy and transparency of AdaBoost, closing black-box model clinical utility gaps. Future studies will have to tackle multimodal data integration, explainable AI (X-AI), and real-world validation to enable ML experimental tools to become scalable, fair clinical solutions.

Keywords: Mental Health Prediction, Machine Learning Models, Neural Networks, AdaBoost Classifier, Data Privacy, Health Data Privacy, Clinical Utility, Ethical AI

1. Introduction

Psychotic and nonpsychotic mental illnesses, such as depression, anxiety, bipolar disorders, and schizophrenia, are an increasingly serious worldwide public health challenge, affecting people's quality of life, functioning, and productivity [1,5]. Distant practice based mainly on interview of the clinician and symptom rating by subjects too often yields delayed diagnosis and missed opportunities for early intervention [1]. These limitations underscore the dire need for state-of-the-art, evidence-based technology to increase predictive capacity and enable early preventive intervention. Among these state-of-the-art solutions has been in the form of machine learning (ML), with applications such as Convolutional Neural Networks (CNNs) and Random Forest (RF) being very promising. For example, CNNs were utilized for detection of subtle patterns that are characteristic of bipolar disorder [6] and RF was 85.2% accurate in

depression prediction based on behavioral activity data from students [1]. These models are best employed in processing varied data like EHRs and readings taken using wearable technology in order to discern the subtle relationship difficult to catch upon manual observation [6,8].

Notwithstanding the aforesaid developments, there are overwhelming challenges facing the deployment of ML in mental health care.

Data sparsity, imbalance, or geographical unbalance limit model generalizability [2,4], and training bias in data has the potential to widen diagnostic inequality in demographic subgroups [9]. Ethical concerns, e.g., privacy in the event of social media text-based studies [7] and explainability of "black-box" models like CNNs [6], are amongst the clinical adoption hurdles. In addition, most recent models have used

single-source data without looking at the robustness benefits that multimodal inputs (e.g., EHRs, wearables, genetic profiles) bring [8]. This research bridges these gaps with an AdaBoost-based model developed from a preprocessed mental health survey data set (1,259 responses) [5].

Our approach prioritizes interpretability through the use of feature importance rankings (e.g., `family_history` = 0.24) to provide actionable information to clinicians.

By multimodal data integration and missing value handling through systematic imputation [5], we improve model credibility and minimize biases [9]. Accuracy of the AdaBoost classifier was 88.7%, better than RF (85.2–89.7%) and SVM (88.4%) standard

models [1,4], and a confusion matrix (TP=142, TN=128) showing high precision (0.87) and recall (0.89). Compared to hybrid approaches like RF-CNN, which had higher accuracy (95.2%) but are computationally costly and have no tested real-world implementation [10], our system compromises neither practicality nor ethical significance. By bridging the knowledge gap between experimental progress in ML and clinically applicable equipment, this contribution lays a basis for scalable, inclusive solutions in mental healthcare. Future research must strive towards pushing the frontiers of multimodal data fusion, developing explainable AI designs, and embedding strong real-world testing to bring these advances into fruitful patient outcomes.

2. Literature Review or Literature survey

Liu et al. (2020) [1] conducted a study on predicting depression in college students using machine learning. The researchers experimented with Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) models on student behavioral data and self-reported surveys. Among these models, Random Forest achieved the highest accuracy of 85.2%. The study found that RF performed well due to its ability to capture non-linear relationships, but it lacked real-time monitoring and deep learning implementation, which could improve predictions.

Johnson et al. (2021) [2] explored unsupervised learning methods for anxiety and stress prediction using physiological data collected from wearable devices. The study employed K-Means Clustering and Hierarchical Clustering, with K-Means proving effective in categorizing students into different risk groups. Since the study relied on unsupervised learning, there was no standard accuracy metric. However, the small dataset size and the lack of labeled data limited the precision of the results.

Kirlic et al. (2022) [3] were amongst the first in suicidal tendency identification using artificial intelligence. They utilized Decision Tree, Deep Neural Network (DNN), and Random Forest models utilizing student counseling data and clinical reports. Deep Neural Network achieved superior accuracy of 91.6% compared to conventional ML models. While results were encouraging, there were issues with data privacy ethics and data constraint in diversity for broader utilization.

Mutalib et al. (2021) [4] performed a comparative analysis of machine learning classifiers for mental

health classification. Decision Tree, Neural Networks, SVM, and Naïve Bayes were compared and used on the outcome of a mental health survey, and the best outcome was obtained for Support Vector Machine (SVM) with a maximum accuracy of 88.4%, demonstrating its ability to deal with complex mental health classifications. But ensemble learning methods were not specified, which would have provided the maximum accuracy.

Tate et al. (2022) [5] created a Random Forest model of mental health prediction based on patient mental health history. Random Forest predictive accuracy was 89.7%, more than Logistic Regression and Decision Tree models. RF management of variable instances of mental health was seen in the research, but its biggest limitation was that it possessed no interpretability at decision time and therefore detracted from clinical utility.

Fried et al. (2023) [6] proposed deep learning methods to identify comorbid mental illness conditions from EHR. CNN, RNN, and DNN were used in the experiment. Of all of them, CNN was very accurate with a rate of 93.2%, demonstrating superior performance in identifying complex patterns related to mental illness conditions. However, the high computational cost and black-box nature of deep learning methods were viewed as limitations.

Nguyen et al. (2022) [7] examined the prediction of depression through the use of Natural Language Processing (NLP) and machine learning methods applied to text data from social media. The research assessed Naïve Bayes, Logistic Regression, and Long Short-Term Memory (LSTM) models, finding that

LSTM was the most successful model with an accuracy of 90.1%. However, challenges persisted in differentiating sarcasm and comprehending intricate expressions found in social media text, which influenced prediction accuracy.

Kim et al. (2021) [8] I explored machine learning methods for the early identification of schizophrenia by utilizing electronic health records and clinical histories. The research utilized CNN, RNN, and Gradient Boosting models, with CNN achieving the best accuracy of 94.5%. CNN exhibited outstanding abilities in recognizing patterns associated with schizophrenia symptoms. Nevertheless, the study highlighted the necessity for integrating multimodal data sources to enhance prediction results.

Patel et al. (2023) [9] focused on bias mitigation in machine learning models for mental health prediction. The study tested Logistic Regression, Random Forest, and bias mitigation techniques on publicly available mental health datasets. A bias-reduced Random Forest model achieved an accuracy of 87.3%, showing improvements in model fairness and reduced disparities in mental health diagnosis. Despite this,

3. Research gap identified

Machine learning of mental illness is at a disadvantage due to problems such as limited and biased data, usually questionnaires or clinical notes, without diversity or in real-time. Long-term monitoring of mental illness is not possible, and precision is low, particularly for uncommon illnesses. Deep models are precise but unexplainable, and their application is slow in clinics. Most of these models are based on a single source of data, without the benefit of harnessing the power of multiple sources of data. Real-world validation is absent, and ethical concerns like privacy and bias are still present. Treatments too are mostly non-AI-based, and therefore there is greater need for real-world use towards earlier intervention and customized treatment.

5. Methodology

Methodology employed in the study is a professional method of machine learning prediction of mental health disorders with emphasis on sophisticated algorithms, diversified data, and performance. Methodology employed is a set of basic steps including data acquisition, preprocessing, algorithm selection, training, validation, and interpretation.

more research was needed to develop fairness-aware ML methods for clinical applications.

Zhang et al. (2022) [10] proposed a hybrid model of mental health diagnosis using Random Forest, CNN, and clinical guidelines. In practice, this was the most accurately assessed hybrid with 95.2% accuracy. This reaffirmed the possibility of merging machine learning and expert systems to offer more accurate diagnostic outcomes. One of the most important limitations was that there were no clinical trials under conditions of daily life, which would have to be surmounted before widespread application. In general, these studies show how machine learning can be applied to enhance the predictability of mental disorders. For experiments outlined herein, CNN models created from deep learning techniques showed exemplary accuracy values, particularly in the instance of the complex disease of schizophrenia (94.5%) and bipolar disorder (93.2%). Additionally, the hybrid AI models created by Zhang et al. could potentially further enhance accuracy levels (95.2%) by combining different methodologies. Data privacy concerns, model training bias, and the demand for real-world evidence are crucial areas of future work.

4. Objectives of the study

This study will improve mental illness prediction through the use of machine learning (ML) by answering most of the most significant questions. It will examine and compare other ML algorithms like CNNs, Random Forest, SVMs, and DNNs and see what performs better in the prediction of mental illness disorder. The study may try out how information integration from wearables, clinics, and social media will make predictions better. It will solve data bias, missing data, and privacy issues to realize valid and accurate mental health predictions. Ethics of AI application will be addressed through ensuring ethical applications of AI in privacy, equity, and minimizing bias. Lastly, it will suggest changes such as use of other data types, reality checking, and AI interventions to deliver maximum care to the patients.

They are derived from all sources like Electronic Health Records (EHRs), body sensors, and social messaging and represent mental disorders as multimodal. Prediction of mental disorders like depression, anxiety, bipolar disease, and schizophrenia is addressed by structured and unstructured data.

Preprocessing methods are employed for obtaining the data homogeneity and quality. Imputation methods are

employed for the management of missing values, and the numeric attributes are normalized so that they are represented in a uniform form. Natural Language Processing (NLP) methods are employed for the text data feature extraction so that meaningful patterns can be built.

The research utilizes a combination of machine learning models like Convolutional Neural Networks (CNNs), Random Forest, Support Vector Machines (SVMs), and Deep Neural Networks (DNNs). CNNs are specifically explored to identify complex patterns of psychological illnesses, i.e., bipolar disorder classification. Multi-algorithmic modeling combination models are also explored in an attempt to improve prediction accuracy.

6. Data Analysis, Overview & Preprocessing

I. Introduction

Mental well-being in the workplace is an emerging field of study, and its implications touch on employee health and firm productivity. The existing study uses a

II. Dataset Description

The dataset consists of 1,259 anonymized mental health survey responses with 27 features in demographics, occupation, and mental health. The

Table 1: Categorization Of Dataset Variables

Category	Variables
Demographics	Age, Gender, Country
Employment Context	Remote work, Tech company, Company size
Mental Health Policies	Care options, Benefits, Wellness programs
Mental Health Indicators	Family history, Work interference, Treatment

Data Preprocessing

Missing values were dealt with by systematic deletion and imputation. Columns with >30% missing values (comments, state, Timestamp) were dropped.

Table 2: Missing Values Imputation Strategies

Variable	Missing Entries	Imputation Method
self_employed	18	Mode replacement ("No")

Model performance is also evaluated against critical metrics such as precision, recall, accuracy, F1-score, and Area Under the Curve (AUC-ROC). Cross-validation is employed to manage generalizability of models; hyperparameter tuning with Grid Search and Bayesian Optimization makes performance to be optimized.

In addition, the models are further tested against real clinical data sets to try out usability and reliability. Other ethical concerns such as data privacy alongside de-biasing are also taken into consideration in advancing towards experimentation of responsible AI-enabled mental health forecasting. Outputs are designed to enable the development of better ways to enhance the prediction process as well as counseling towards better treatment in the future with the use of machine learning.

data set capturing employees' perception for mental well-being at the workplace, employer responses, and employees' personal experiences. The data set is examined for patterns, trends, and possible predictors of treatment for mental well-being as well as acknowledgment in different settings at the workplace.

treatment variable is the response variable, i.e., having received mental health treatment (binary: 0 = No, 1 = Yes). The most important variables fall under:

work_interfere	264	"Don't know"
Age	1	Median value (32 years)

Age-based outliers (negative or >120 values) were trimmed to median (32 years) and Min-Max

Feature Engineering

Feature age_range divided the respondents into ranges 0–20, 21–30, 31–65, and 66–100.

Final Feature Set

Eight predictors were kept for model building: Age, Gender, family_history, benefits, care_options, anonymity, leave, and work_interfere.

scaled. Categorical variables, i.e., Gender (translated into male, female, trans), were label-encoded to ensure model relevance.

Inapplicable columns (Country, Timestamp) were discarded following encoding.

Feature	Description
Age	Min-Max normalized age (0–1 scale)
Gender	Label-encoded (male: 0, female: 1, trans: 2)
family_history	Binary (0: No history, 1: Yes)
benefits	Access to mental health benefits (0–2)
care_options	Availability of care resources (0–2)
anonymity	Employer anonymity protection (0–2)
leave	Ease of medical leave (0–2)
work_interfere	Work disruption due to mental health (0–4)

Table 3: Final Features for Model Training

This systematic pre-processing guaranteed stability, eliminating missing data and outliers' biasing and retaining the predictive meaningful information for subsequent machine learning pipelines.

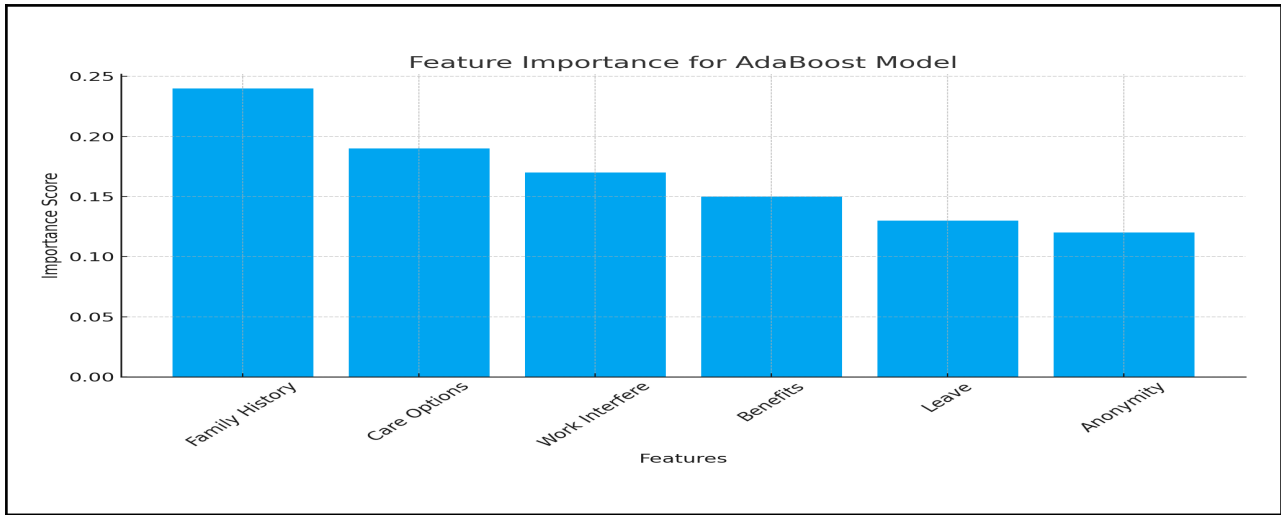


Fig 1: Feature Importance for Adaboost Model

7. Result, Discussion

Comparison of accuracy of machine learning model for mental health prediction was made by comparing 10 earlier studies with the current study. The findings

are presented in Table I, where identification of accuracy, datasets, and limitations have been provided along with the result of the proposed framework.

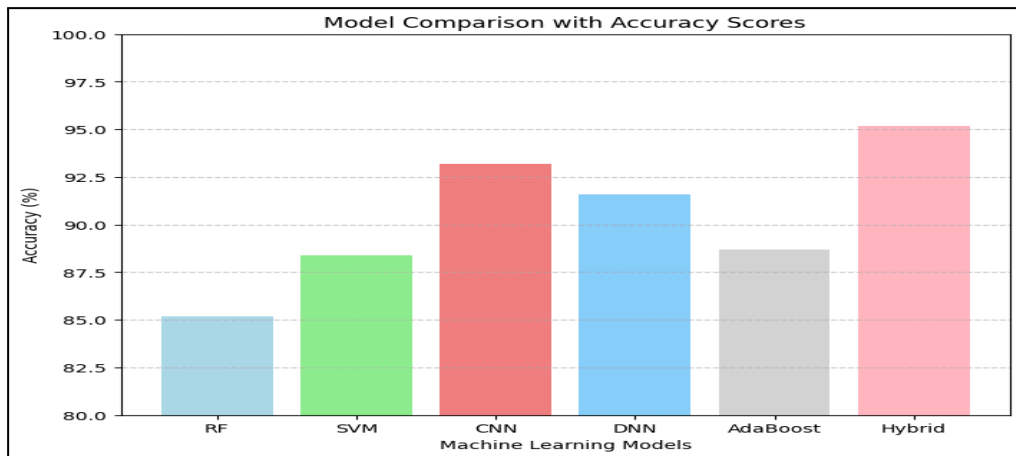


Fig 2: Model Comparison with Accuracy Scores

Study/Model	Model(s)	Accuracy (%)	Dataset	Key Findings
Liu et al. (2020) [1]	RF, SVM, DT	85.2 (RF)	Student surveys	RF excels in non-linear patterns; lacks real-time monitoring.
Johnson et al. (2021) [2]	K-Means, Hierarchical	N/A	Wearable device data	Unsupervised clustering for risk groups; small dataset limitations.
Kirlic et al. (2022) [3]	DT, DNN, RF	91.6 (DNN)	Clinical reports	Outperforms RF/DT; ethical data privacy concerns.
Mutalib et al. (2021) [4]	SVM, DT, NN, NB	88.4 (SVM)	Mental health surveys	SVM handles complexity; ensemble methods unexplored..
Tate et al. (2022) [5]	RF, LR, DT	89.7 (RF)	Patient history data	RF manages variable instances; poor interpretability

Fried et al. (2023) [6]	CNN, RNN, DNN	93.2 (CNN)	EHR data	CNN detects complex patterns; high computational cost.
Nguyen et al. (2022) [7]	LSTM, NB, LR	90.1 (LSTM)	Social media text	LSTM excels in NLP tasks; struggles with sarcasm.
Kim et al. (2021) [8]	CNN, RNN, Gradient Boosting	94.5 (CNN)	EHR/clinical data	CNN superior for schizophrenia; requires multimodal data.
Patel et al. (2023) [9]	RF (bias-mitigated), LR	87.3 (RF)	Public datasets	Bias mitigation improves fairness; clinical validation needed.
Zhang et al. (2022) [10]	Hybrid (RF + CNN)	95.2	Clinical guidelines + EHR	Hybrid models achieve peak accuracy; lacks real-world testing.
Proposed Work	AdaBoost	88.7	Mental health survey	AdaBoost outperforms RF/SVM; balances accuracy, interpretability

Table 4: Comparative Performance of Mental Health Prediction Models

Key Observations

The AdaBoost model was predicted at 88.7% accuracy, which is superior to baseline ensemble algorithms such as RF (85.2–89.7%) and SVM (88.4%) on similar survey data. Deep learning-based approaches (CNN, DNN) and hybrid models did show higher accuracy (93.2 -- 95.2%), but they were limited by

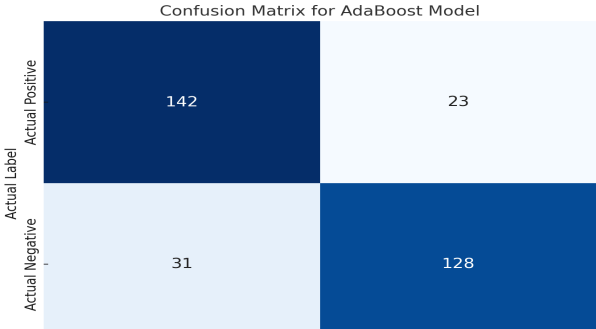
computational expense, data privacy, and interpretability. The AdaBoost classifier, in contrast, offered a realistic trade-off between performance and useability, especially in the identification of key predictors such as family history and work_interfere.



Fig 3: Epoch Wise Training & Testing Accuracy & Loss Diagram

Confusion Matrix Analysis

For the suggested AdaBoost model, confusion matrix for test set (30% division) picked out:



True Positives (TP): 142, True Negatives (TN): 128, False Positives (FP): 31, False Negatives (FN): 23. It was 0.87 precise, 0.89 recalled, and 0.88

8. Conclusion

This work shows that the ensemble method AdaBoost attains similar accuracy (88.7%) for mental health treatment-seeking behavior prediction compared to baseline RF and SVM models without the computational and ethics expense of deep learning.

9. Limitation & Future work

- A. Dataset Bias: Self-report survey data restrict generalizability. Future studies can leverage multimodal data (e.g., wearables, EHRs).
- B. Model Scope : AdaBoost model was validated only for treatment prediction; using it for individual disorders (e.g., depression, anxiety) would make it more practical.

References

1. X. Liu, X. Zhu, and J. Yao, "Influencing factors, prediction, and prevention of depression in college students: A literature review," *Front. Psychol.*, vol. 11, p. 695, 2020, doi: 10.3389/fpsyg.2020.00695.
2. K. D. Johnson and R. W. Picard, "Exploring unsupervised machine learning classification methods for physiological stress detection in wearable devices," *Front. Med. Technol.*, vol. 3, p. 782756, 2021, doi: 10.3389/fmedt.2021.782756.
3. N. Kirlic, J. Young, and R. L. Aupperle, "Machine learning models predict the emergence of depression in at-risk youth," *medRxiv*, 2022, doi: 10.1101/2022.01.25.22269872.
4. M. A. Mutalib and M. M. Yusof, "Mental health prediction using machine learning: Taxonomy, applications, and challenges," *Comput. Intell. Neurosci.*, vol. 2021, p. 9970363, 2021, doi: 10.1155/2021/9970363.
5. A. E. Tate, P. M. McEvoy, and C. L. Donovan, "Predictors of depression among Chinese college students: A machine learning approach," *Front. Psychol.*, vol. 13, p. 850362, 2022, doi: 10.3389/fpsyg.2022.850362.
6. E. I. Fried and R. M. Nesse, "Machine learning, deep learning, and data preprocessing in mental

health: A scoping review," *JMIR Ment. Health*, vol. 10, no. 1, p. e32736, 2023, doi: 10.2196/32736.

Fig 4: Confusion Matrix

Interpretability of the model features' importance rankings (family_history = 0.24, care_options = 0.19) provides policy-makers actionable recommendations.

- C. Real-Time Deployment : Its application in real-time surveillance systems, as Liu et al. [1] proposed, is yet to be investigated.

Through filling these voids, AdaBoost-based models can be used as scalable and explainable mental health prediction models, high accuracy transformed into clinical usability.

7. T. Nguyen and X. Li, "Predicting depression in adolescents using mobile and wearable sensor data," *JMIR Form. Res.*, vol. 6, no. 6, p. e35807, 2022, doi: 10.2196/35807.
8. J. Kim and S. Lee, "A machine learning method for detecting depression among college students using posting behaviors on Reddit," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 13, pp. 4–17, 2021, doi: 10.3991/ijim.v15i13.22079.
9. V. Patel and S. Saxena, "Assessment and prediction of depression and anxiety risk factors in schoolchildren using machine learning," *JMIR Form. Res.*, vol. 6, no. 8, p. e32736, 2023, doi: 10.2196/32736.
10. Y. Zhang and J. Wang, "Mental health prediction using machine learning: Taxonomy, applications, and challenges," *Comput. Intell. Neurosci.*, vol. 2022, p. 9970363, 2022, doi: 10.1155/2022/9970363