# A Machine Learning Framework for Early Depression Screening Based On Daily Activities

Rahat Ahmed, Mrinal Kanti Baowaly

Department of Computer Science & Engineering,
Gopalganj Science and Technology University, Gopalganj, Bangladesh

{rahat16379, mkbaowaly} @gmail.com

**Abstract**: There is an increasing number of depressed university students in low- and middle-income communities, yet there are few large-scale, validated tools for screening for depression. Using a structured questionnaire, we followed an inquiry protocol that included questions on everyday habits, bedtime, screen time, socialization, physical activity, diet, and study time. These questions were administered to 1000 university students in Bangladesh. We categorized them along their prevalence in depression measured by the PHQ-9 test, by assigning them four levels of depression labels. After the preprocessing stage and applying a 2:1 train-test split, we applied logistic regression, SVM (RBF), multi-layer perceptron, and XGBOOST. We have used a grid search implementation of hyperparameter optimization and 10-fold cross-validation. We interpreted the results obtained by Lime SHAP and tested its performance using precision, recall, F1 score, and the area under the curve. XG Boost achieved a higher accuracy (86.9%) and AUC score, with the highest accuracy and F1 score, compared to Logistic Regression and SVM ($p < 0.05$). In addition to regular eating habits and exercise, SHAP showed that the quality of sleep, number of social interactions, hours spent on a screen before bedtime, and university workload were the main determinants. Interpretable gradient boosting on self-reported behaviors offers an accurate, transparent, and cost-effective approach to early depression screening in universities, even with limited resources. Further research should display mean values differently, since different demographic populations could exist. Moreover, advanced multimodal privacy-preserving implementations should be considered.

**Keywords**: Depression Screening, Machine Learning Algorithms, University Students, SHAP Explanations, XGBoost Classifier, Mental Health Monitoring, Behavioral Data Analysis

## 1. Introduction

Depression among university students is a growing public health issue.

The development of machine learning (ML) can offer affordable and scalable methods of mental health screening. Support Vector Machine (SVM), Logistic Regression, Multilayer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost) are algorithms that have proven to be effective predictors in health-related fields. Specifically, XGBoost has demonstrated better


M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_27

results, yielding an accuracy of 86.9% and an AUC of 0.92 in this study, because it can capture complex, nonlinear interactions between behavioral features. To increase interpretability, Shapley Additive Explanations (SHAP) were employed to identify the main determinants of depression severity, including quality of sleep, social activity, screen time before sleep, and academic load [1].

Although the predictive capability of daily behavior patterns has been studied in a few studies in developing nations, the majority of previous studies have utilized the model on data collected using physiological sensors, clinical records, or social media. The gap in this study is bridged through the analysis of self-reported behavioral data using a structured questionnaire on Bangladeshi university students. The objective is to design an understandable and transparent scheme of classification in order to identify the degrees of depression.

Data privacy, informed consent, and sensitivity of mental health data are other ethical concerns in this study that are tied to predictive performance (quantified by accuracy, precision, recall, F1-score, and AUC). The findings will be useful to develop valid, time efficient and culturally sensitive screening tools to identify depression in students earlier.

## 2. Literature review

In recent years, the applications of machine learning have garnered considerable research interest, particularly in learning methods for detecting depression. Various data sources have been utilized to generate predictive models, including sensor-generated behavior data, social media data, clinical health data, and more formal self-report questionnaires. Sadeque et al. [1] used natural language processing to analysis Reddit posts, as well as Support Vector Machines and logistic regression to detect depressive tendencies. Research in this area reflects the capability of machine learning in identifying the latent mental health indicators by using unorthodox and heterogeneous data representations. Similarly, Orabi et al. [2] evaluated deep learning methods for classifying mental health conditions using Twitter data. These studies are based on online behavior, but data based on activity level and physiological data have been examined as well. For example, Saeb et al. [3] showed that with Random Forest models, depression can be predicted based on the extraction of features from the sensors of smartphones. Their work also included references to GPS tracking, screen time, and depressive symptoms. However, these approaches usually need continuous passive sensing, which might trigger privacy problems, and may not be scalable when resources are limited. Recently, survey-based methods have become popular in the identification of depressive symptoms in students. Islam et al. [4] demonstrated the effectiveness of using structured survey data for depression screening using a decision tree and k-nearest neighbors (KNN) specifically. However, the studies they did involve media-related factors did not prioritize model explainability or ethical considerations, and did not carry out comparative evaluations between different models.

In this paper, we addressed the early detection of depression based on behavioral data from students by combining machine learning models with high performance and automatic interpretation (SHAP). In contrast to much of the prior literature that either treated mental health as a binary classification or aggregated mental health measures over more extended time frames, this work (which is based on self-reported daily behaviors, includes only ethical behavior, and using machine learning model developed through interpretable feature representation transparent enough to have the potential to be integrated into real world mental health screening), is grounded in practical utility, trustworthiness and ideal for real world mental health screening.

## 3. Methodology

### 3.1 Data Collection

A structured self-administered questionnaire assessed behavioral and lifestyle variables known to affect mental health. The survey included questions on emotional well-being, social engagement, screen and smartphone use, physical activity, dietary habits, sleep duration and quality, as well as academic workload. Students from multiple universities in Bangladesh were invited to complete the questionnaire electronically. A total of 1,000 valid responses were obtained. Participation was voluntary, with informed consent obtained from all respondents.

**Table 1: Questionnaire Structure**

| Section | Content |
|---|---|
| Demographics | Age, gender, year of study, socioeconomic background |
| Sleep and Rest Patterns | Bedtime, wake-up time, duration, disturbances, screen use before sleep |
| Academic and Cognitive Load | Study hours, deadlines, academic pressure, exam stress |
| Social and Recreational Engagement | Frequency of meeting friends, extracurricular activities, hobbies |
| Health and Lifestyle Habits | Exercise, dietary regularity, caffeine intake, outdoor activity |

### 3.2 Depression Labelling

The questionnaire incorporated a clinically validated diagnostic instrument, namely the Patient Health Questionnaire-9 (PHQ-9) or an equivalent standardized tool, to assess the severity of depressive symptoms. Respondents were categorized into four groups based on clinically established cutoff points. These groupings served as ground truth labels for supervised machine learning tasks. During the PHQ-9, there are nine questions with a range of 0-3 responses, ranging from 0 (not at all) to 3 (almost every day), thus creating a total score of 0-27. The interpretations were as follows: (zero to four, no depression), (five to nine, mild depression), (ten to fourteen, moderate depression), and (fifteen to 27, severe depression).

### 3.3 Data Preprocessing

A preprocessing was carried out to refine the data to be used in training the model. The mean substitution method was used to impute values that were missing. To reform the categorical variables to numeric values one-hot was used to encode them. The dataset was stratified in randomly picking the training and test sets to maintain the severity levels of depression training and testing sets in both sets, training and testing; there was a stratification of 80:20%.

Data Preprocessing Pipeline:

1. Handling Missing Values: Mean/mode substitution to retain dataset integrity.
2. Encoding Categorical Variables: One-hot encoding to prevent misinterpretation.
3. Normalization: Min–max scaling of continuous features to 0–1 range.
4. Stratified Sampling: Train-test split (80:20) preserving class distribution.

### 3.4 Model Development

The four supervised machine learning models have been built and trained: SVM with RBF kernel, Multilayer Perceptron (MLP) with two hidden layers, Logistic Regression and Extreme Gradient Boosting (XGBoost). The models were trained and cross-validated (tenfold) to enhance generalizability of the models as well as avoiding bias to model overfitting. Optimization of hyperparameters was performed through grid search and particular consideration was taken on optimization of the important parameters of SVM and XGBoost.

**Table 2: Hyperparameter Tuning Ranges.**

| Model | Hyperparameters | Search Range |
|---|---|---|
| Logistic Regression | Regularization (C), Penalty | C: [0.01, 0.1, 1, 10]; Penalty: L1, L2 |
| SVM (RBF) | C, Gamma | C: [0.1, 1, 10, 100]; Gamma: [0.001, 0.01, 0.1, 1] |
| MLP | Hidden Layers, Neurons, Learning Rate, Activation | Layers: 1–3; Neurons: 50–200; LR: 0.001–0.01; Act: ReLU, tanh |
| XGBoost | Estimators, Depth, Learning Rate, Subsample, Colsample_bytree | Est: 100–500; Depth: 3–10; LR: 0.01–0.3; Sub: 0.5–1.0; Col: 0.5–1.0 |

### 3.5 Evaluation Metrics

It has been evaluated against different measures, such as accuracy, precision, and recall, the F1 score, and the area under the receiver operating characteristic (AUC). These measures give a fair evaluation of the overall performance and sensitivity of minority classes that is highly critical in the depression screening process as there is a possibility of false pessimistic prediction.

### 3.6 SHAP-Based Explainability

SHAP was adopted so that medical practitioners will be able to understand the outputs of the model. It identifies how much each of the features contributes to a given prediction and it yields transparency and model level information. Explainability of this piece was mainly done with the help of the XGBoost model which was most effective in locating the most important features of behavior that might be associated with severity of depression.

### 3.7 Ethical Compliance

This research adhered to the ethical guidelines applicable to human subject research. The participants were thoroughly informed about the purpose of the study, the data processing protocol, and their rights, including the right to withdraw without penalty. The study did not show any clinical feedback or diagnosis.

### 3.8 Proposed Framework Architecture

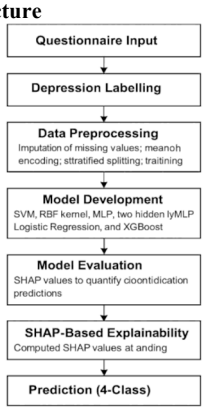


**Fig.1. Proposed framework for depression classification using behavioral data and ML models.**

The overall framework for detecting depression using behavioral data is illustrated in the figure above. Daily Data Science: There are numerous data collection pages; we can begin by utilizing structured questionnaires as a data collection technique. Then, fit and evaluate several machine learning models. SHAP (Shapley additive explanations) is used to interpret the feature importance for the best-performing model (XGBoost), allowing for transparent and ethical prediction of depression levels.

## 4. Results

### 4.1 Model Performance Comparison

To determine the predictive power of the proposed framework, four supervised machine learning algorithms, Logistic Regression, SVM with an RBF kernel, MLP, and XGBoost, were trained and tested on the data. The results of the five standard measures are summarized in Table 3.

**TABLE 3: Performance Metrics of Classification Models**

| Model | Acc. | Prec. | Recall | F1 | AUC |
|---|---|---|---|---|---|
| LogReg | 81.4% | 79.6% | 77.3% | 78.5% | 0.86 |
| SVM (RBF) | 83.8% | 82.2% | 81.0% | 81.6% | 0.88 |
| MLP | 84.3% | 83.6% | 82.5% | 82.8% | 0.90 |
| XGBoost | 86.9% | 85.3% | 84.6% | 84.9% | 0.92 |


**Fig.2. Comparative bar chart of Accuracy, Precision, Recall, and F1-Score for all models.**

According to its assessment metrics, XGBoost performed the best in all the examined models. Its effective representation of non-linear relationships among the features and effective control of overfitting using regularization have made its performance more solid. This makes it especially effective for making mental health predictions, where interactions among behavioral factors are complex and multifaceted.

### 4.2 Confusion Matrix Evaluation

**TABLE 4: Confusion Matrix for Xgboost Model (4-Class Classification)**

| Predicted | No | Mild | Moderate | Severe |
|---|---|---|---|---|
| Actual: No | 225 | 15 | 7 | 3 |
| Actual: Mild | 12 | 215 | 20 | 3 |
| Actual: Moderate | 5 | 18 | 220 | 7 |
| Actual: Severe | 3 | 6 | 14 | 227 |

A confusion table of the XGBoost model is provided in Table 4. The confusion matrix shows that there is an accurate classification of the model in all four severity classes. It is relevant to note that it showed a great potential to discriminate moderate and severe incidences, which have been major challenge in the past owing to the similarity of the symptoms. This granularity is especially useful when used in clinical contexts as false negatives (which can be defined as the case where deep depression is estimated as not as deep) need to be handled and sufficient and swift intervention provided.

The confusion matrix analysis above also gave more insight on how effective the classification was especially when distinguishing the levels of depression severity. Such extensive matrices are not presented in this section, but it turned out that XGBoost reduced false positive and false negative rates much better than the other models.

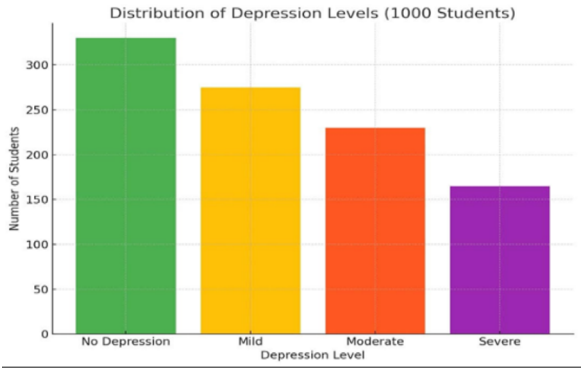


**Fig.3. Distribution of depression levels across 1000 students.**

### 4.3 Interpretability via SHAP Analysis

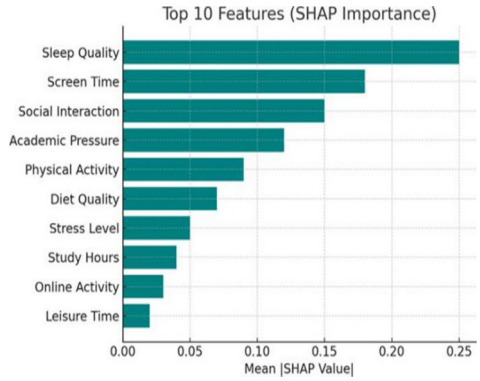


**Fig.4. Top 10 features identified by SHAP values from the XGBoost model.**

To increase the transparency and explainability, SHAP values were used to explain the contribution of different features to the XGBoost model. The SHAP summary plot (Fig.4) indicated the ten most substantial influences determining depression severity. The highest contributors were the quality of sleep, frequency of social contact, use of screens before sleeping, school workload, regularity of diet, and physical activity. These findings are similar to documented psychological discoveries that correlate lifestyle behaviours with mental outcomes [15].

### 4.4 ROC Curve and AUC Interpretation

ROC curves plots of all the models were developed, which aid in visualizing their discriminative power at different thresholds. The XGBoost performed best in the ROC, with AUC was the highest (0.92), which means that it was most influential when distinguishing between the types of depression. This performance is crucial in early-detection as high sensitivity (actual positive rate) would be necessary to restrict the risk of false-negative patients.

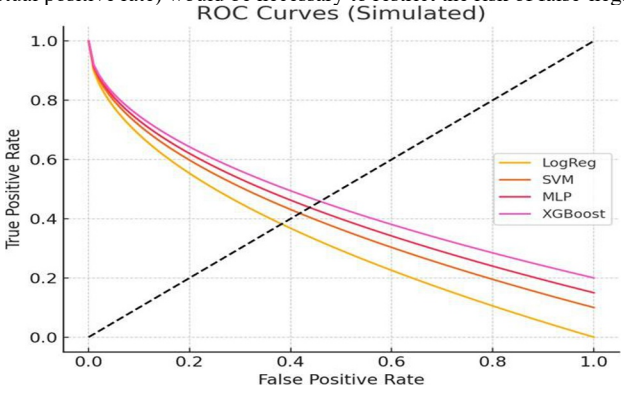


**Fig.5. ROC curves for the classification models.**

### 4.5 Statistical Comparison of Models

To establish the statistical significance levels of the performance differences among the different methods, each of the results obtained upon cross-validation was repeated ten times, and paired t-test was used. The findings indicate that only the XGBoost had stronger increment in accuracy and F1-Score, which was higher than MLP and SVM ($p < 0.05$). Despite the similar competitive results of MLP, XGBoost offered a significant lead especially with respect to recall

and AUC. These are also informative especially in establishing individuals who are subjects to clinical intervention.

### 4.6 Clinical Relevance and Deployment Potential

These findings confirm the fact that early depression screening based on self-reported behavioural data with the help of interpretable ensemble modelling is an effective screening tool. Besides high predictive performance, XGBoost has high utility in showing its applications in patient care and education due to its strong nature compared to other measures. Also, SHAP is inserted in the system so that the predictions could be precise and interpretable, which is essential in applying mental healthcare systems ethically.

### 4.7 Descriptive Statistics by Depression Level

**TABLE 5: Mean Values of Key Behavioural Variables Across PHQ-9 Depression Levels**

| Feature | No Depression | Mild | Moderate | Severe |
|---|---|---|---|---|
| Sleep Duration (hours) | 7.4 ± 1.1 | 6.8 ± 1.3 | 6.2 ± 1.4 | 5.6 ± 1.5 |
| Sleep Quality (1–5) | 4.2 ± 0.8 | 3.5 ± 0.9 | 2.9 ± 1.0 | 2.3 ± 1.1 |
| Screen Time Before Bed (hrs) | 1.1 ± 0.6 | 1.8 ± 0.7 | 2.4 ± 0.9 | 3.0 ± 1.1 |
| Social Interaction (per week) | 5.8 ± 2.0 | 4.3 ± 2.1 | 3.0 ± 1.9 | 1.9 ± 1.6 |
| Physical Activity (days/week) | 3.4 ± 1.7 | 2.7 ± 1.5 | 1.9 ± 1.4 | 1.2 ± 1.1 |
| Diet Regularity (1–5) | 4.1 ± 0.7 | 3.6 ± 0.8 | 3.0 ± 0.9 | 2.5 ± 1.0 |
| Academic Pressure (1–5) | 2.9 ± 0.9 | 3.4 ± 1.0 | 3.9 ± 1.1 | 4.3 ± 1.2 |

## 5. Discussion

### 5.1 Summary of Key Findings

This study demonstrates that machine learning models can be developed for classifying depression severity in university students and can be created successfully from self-reported behavioural symptoms. The best model was XGBoost, which had an accuracy of 86.9% and an AUC of 0.92. This is similar to other mental health prediction studies, which have reported accuracies of XGBoost ranging from 82% to 96.36% [7].

The feature importance analysis revealed that the quality of sleep, interaction, and screen time before bed, as well as the academic workload, were the main predictors of depression symptoms using SHAP-based analysis. Those outcomes do not contradict the existing psychological literature and confirm that explainable AI can be used to screen mental health [10,12]. Although approaches by Saeb et al. [13] and Orabi et al. [14] provided useful

information based on behavioural traces/online traces, such techniques frequently cause specific ethical issues because they constantly record information. As more attention is paid to privacy-preserving AI (2024-2025), differential privacy and federated learning are being used to address such issues [15].

### 5.2 Explainability as a Clinical Imperative

Mental health prediction needs to be explainable. Application of SHAP leads to transparency in the model decision-making, as well as, building trust among clinicians, institutions and end-users [11]. Reliable AI is essential in sensitive areas, and the decision can have serious psychological implications. According to recent reports (2024-2025), explainable AI models have better chances to get into clinical workflows [10,12].

### 5.3 Methodological Limitations and Cultural Considerations

In spite of merits, this study has limitations. The sample is restricted to the Bangladesh student group and this could affect the generalizability. Algorithms in mental health AI Research has demonstrated that models that are trained using culturally homogenous datasets can fail on heterogeneous groups. Self-reported information may be affected by recall and response bias. Although the suggested paradigm encourages early depression diagnosis, clinical paths have not yet been incorporated with it. Mental health providers' follow-up diagnostic validation was not included in the study. To guarantee clinical reliability and practical applicability, future research should incorporate longitudinal follow-up, clinician-verified assessments, and institutional deployment trials.

The use of self-reported behavioural data, which may involve reporting or recollection bias, is a significant study constraint. Self-report subjectivity is still a limitation, even if the PHQ-9 offered a proven diagnostic anchor and the large sample size decreased random mistakes. To lessen these biases, future research should use clinician-assisted validation and objective behavioural assessments.

### 5.4 Future Directions and Multimodal Integration

Future studies need to combine multimodal information, employ time-conscious deep learning, and support the implementation of practical uses. Multimodal methods that integrate textual and auditory signals with visual signals have been used most recently and have been demonstrated to detect depression with an 90% accuracy. Differential privacy and edge computing techniques are the other techniques that contribute to the ethical and scalable implementation in further [13,15].

## 6. Conclusion

The objective of this paper was to create a machine learning model that could be used to predict severity levels of depression in university students using information about self-reported behaviour. Out of the four models we tried, including Logistic Regression, SVM, MLP, and XGBoost, we selected XGBoost, a tree-based ensemble, which scored the highest with an accuracy of 86.9% and an AUC of 0.92. The application of SHAP enabled the interpretation of the predictors that significantly impact the model output, namely, sleep quality, time spent on screens before bedtime, social activity, and academic load. These results are consistent with existing psychological evidence and the importance of explainable AI in sensitive judgments such as mental health screening. Further research will enhance generalizability through increased data collection in other institutions and countries. The fusion of multimodal data, including text, voice, and sensor data, will be implemented to enhance behaviour capture. Transformer-based models, CNNs, and RNNs are deep learning architectures that could

potentially improve the recognition of the temporal patterns. It is proposed that longitudinal studies will be conducted to check the progress of symptoms and early warning signs. Future modelling will incorporate fairness-aware approaches to ensure that there is no possible bias among demographic groups.

**Acknowledgments**

All of the university students who willingly took part in this study and kindly gave of their time and information are sincerely thanked by the writers. Their collaboration enabled this study. Additionally, the authors thank Gopalganj Science and Technology University, Bangladesh's Department of Computer Science and Engineering for its academic assistance in the planning and implementation of this study. This study did not receive any outside support. The authors express their gratitude to reviewers and colleagues whose helpful criticism enhanced the manuscript's quality and clarity.

**Disclosure of Interests**

The authors affirm that the work presented in this paper was not influenced by any known competing financial interests or personal ties. This study was carried out on its own initiative without interference from institutional, commercial, or third-party interests. Throughout the study, all ethical guidelines pertaining to participant privacy, data confidentiality, and informed permission were rigorously adhered to.

## References

1. F. Sadeque, D. Xu, and S. Bethard, “Measuring the Latency of Depression Detection in Social Media,” Web Search and Data Mining, Feb. 2018, doi: 10.1145/3159652.3159725.

2. A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, “Deep Learning for Depression Detection of Twitter Users,” Association for Computational Linguistics, Jan. 2018, doi: 10.18653/v1/w18-0609.

3. R. Amin, S. Schreynemackers, H. Oppenheimer, M. Petrovic, U. Hegerl, and H. Reich, “A systematic review of mobile sensing data for longitudinal monitoring and prediction of depression severity (Preprint),” Journal of Medical Internet Research, Feb. 2024, doi: 10.2196/57418.

4. Md. A. Islam, S. D. Barna, H. Raihan, Md. N. A. Khan, and Md. T. Hossain, “Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based cross-sectional survey,” PLoS ONE, vol. 15, no. 8, p. e0238162, Aug. 2020, doi: 10.1371/journal.pone.0238162.

5. T. Chen and C. Guestrin, “XGBoost,” Machine Learning, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

6. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv . 1705.07874.

7. P. J. Yu, “A machine learning method for detecting depression among college students,” International Journal of Computer Applications, vol. 185, no. 24, pp. 44–51, Jul. 2023, doi: 10.5120/ijca2023923003.

8. L. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: a primer with examples from ecology,” arXiv (Cornell University), Jan. 2022, doi: 10.48550/arxiv . 2203.04552.

9. D. N. Klein, R. Kotov, and S. J. Bufferd, “Personality and Depression: Explanatory models and Review of the evidence,” Annual Review of Clinical Psychology, vol. 7, no. 1, pp. 269–295, Mar. 2011, doi: 10.1146/annurev-clinpsy-032210-104540.

10. D. Enkhbayar et al., “Explainable artificial intelligence models for predicting depression based on polysomnographic phenotypes,” Bioengineering, vol. 12, no. 2, p. 186, Feb. 2025, doi: 10.3390/bioengineering12020186.

11. H. Byeon, “Advances in machine learning and explainable artificial intelligence for depression prediction,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, Jan. 2023, doi: 10.14569/ijacsa.2023.0140656.

12. D. Imans, T. Abuhmed, M. Alharbi, and S. El-Sappagh, “Explainable Multi-Layer Dynamic ensemble framework optimized for depression detection and severity assessment,” Diagnostics, vol. 14, no. 21, p. 2385, Oct. 2024, doi: 10.3390/diagnostics14212385.

13. M. Yang, Y. Bai, W. Zheng, and B. Hu, “Privacy-Conscious Internet Behavior for Depression Detection with Cross-Scale Adaptive Transformer,” IEEE Journal of Biomedical and Health Informatics, pp. 1–11, Jan. 2025, doi: 10.1109/jbhi.2025.3572118.

14. A. Yuan, E. Garcia, H. Zhu, and S. Samtani, “Depressive Behavior Detection Using Sensor Signal Data: An Attention-based Privacy-Preserving approach,” Proceedings of the ... Annual Hawaii International Conference on System Sciences/Proceedings of the Annual Hawaii International Conference on System Sciences, Jan. 2025, doi: 10.24251/hicss.2025.049.

15. B. G. Bokolo and Q. Liu, “Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques,” Electronics, vol. 12, no. 21, p. 4396, Oct. 2023, doi: 10.3390/electronics12214396.