International Journal of Artificial Intelligence in Medical Issues

Volume 4 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v4i1.436

Research Article

Depression Risk Prediction Among Teenagers Using Explainable Machine Learning and Imbalanced Behavioral Data

Rudi Setiawan [1,*]; Effan Najwaini [2]; Rexania Agramanisti [3]; Rasmiati Rasyid [4]
[1] Universitas Trilogi, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12760, Indonesia, rudi@trilogi.ac.d
[2] Politeknik Negeri Banjarmasin, Kota Banjarmasin, Kalimantan Selatan 70124, Indonesia, effan@poliban.ac.id
[3] Universitas Bina Darma, Kota Palembang, Sumatera Selatan 30111, Indonesia, rezania.agramanisti.azdy@binadarma.ac.id
[4] Universitas Sembilanbelas November Kolaka, Sulawesi Tenggara 93561, Indonesia, ammy.fti@usn.ac.id

Correspondence should be addressed to Rudi Setiawan; rudi@trilogi.ac.id

Received 12 January 2026; Revised 22 January 2026; Accepted 25 April 2026; Published 30 May 2026

Abstract:
Adolescent depression has become an important public health concern, particularly in relation to increasing digital media exposure, lifestyle changes, and psychosocial pressure. This study proposes an explainable machine learning framework for predicting depression risk among teenagers using social media usage, lifestyle behavior, and psychosocial indicators. The dataset consisted of 1,200 records with 13 variables, including age, gender, daily social media hours, platform usage, sleep hours, screen time before sleep, academic performance, physical activity, social interaction level, stress level, anxiety level, addiction level, and depression label. The target variable was highly imbalanced, with 1,169 samples categorized as non-depression and only 31 samples categorized as depression risk. Several machine learning models were evaluated, including Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting. The experiments compared two feature settings, namely behavioral-only features and full features, combined with three imbalance handling strategies: no imbalance treatment, class weighting, and SMOTE. Model performance was evaluated using accuracy, precision, recall, F1-score, balanced accuracy, ROC-AUC, PR-AUC, Cohen's Kappa, MAE, and RMSE. The results showed that the full-feature setting substantially outperformed the behavioral-only setting. The best performance was achieved by Random Forest using full features without imbalance handling, producing perfect classification results with accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC of 1.0000. Permutation importance analysis identified sleep hours, stress level, anxiety level, and daily social media hours as the most influential predictors. These findings indicate that teenage depression risk in this dataset is strongly associated with sleep behavior and psychosocial conditions, in addition to social media exposure. Although the model achieved excellent performance, the result should be interpreted cautiously due to the small number of positive depression-risk samples and the possibility of highly separable label patterns. Therefore, the proposed approach should be positioned as an early risk screening framework rather than a clinical diagnostic tool.

Keywords: Adolescent Mental Health; Depression Risk Prediction; Explainable Machine Learning; Social Media Usage; Imbalanced Classification; Health Informatics.

Dataset link: -

1. Introduction

Adolescent mental health has become an increasingly important public health concern, particularly as teenagers experience rapid biological, emotional, social, and behavioral changes during a critical stage of development [1]. The World Health Organization reports that approximately one in seven adolescents aged 10–19 years experiences a mental health condition, while depression and anxiety are among the leading causes of illness and disability in this age group.

Untreated mental health problems during adolescence may continue into adulthood and affect educational achievement, social relationships, physical health, and overall quality of life [2], [3].

At the same time, social media has become deeply embedded in teenagers' daily lives [4]. Digital platforms provide opportunities for communication, entertainment, learning, and social connection; however, excessive or problematic use may also be associated with negative psychological outcomes. Recent data from the WHO Regional Office for Europe showed that problematic social media use among adolescents increased from 7% in 2018 to 11% in 2022, indicating growing concern about the relationship between digital behavior and adolescent well-being [5]. Several studies have examined the association between social media use and symptoms of depression, anxiety, and psychological distress among children and adolescents, but the findings remain complex and sometimes inconsistent [6], [2]. Therefore, further investigation is needed to understand how social media behavior, sleep habits, stress, anxiety, and lifestyle-related factors can be used to identify teenagers at potential risk of depression [7], [8].

Artificial intelligence and machine learning provide promising approaches for analyzing behavioral and psychosocial data in health-related prediction tasks [9], [10]. In the context of adolescent mental health, machine learning can support early risk screening by identifying hidden patterns in variables such as daily social media duration, sleep hours, screen time before sleep, physical activity, stress level, anxiety level, addiction tendency, and social interaction [11], [12]. Unlike traditional statistical analysis that often focuses on isolated associations, machine learning models are able to learn complex relationships among multiple variables and generate predictive outputs that may assist in early identification of vulnerable individuals.

However, depression risk prediction using behavioral datasets presents several methodological challenges. One major issue is class imbalance, where the number of individuals labeled as at risk of depression is much smaller than those labeled as not at risk. In such cases, conventional accuracy can be misleading because a model may achieve high accuracy by predicting the majority class while failing to detect the minority class. Therefore, evaluation metrics such as recall, F1-score, balanced accuracy, ROC-AUC, and precision-recall AUC are more appropriate for assessing model performance in imbalanced mental health prediction tasks. In addition, model interpretability is also important because predictions in the health domain should not only be accurate but also explainable. Recent studies on explainable machine learning for mental health prediction emphasize that transparency is essential to increase trust and support the responsible use of predictive models in health-related settings.

Based on these considerations, this study proposes an explainable machine learning approach for predicting depression risk among teenagers using social media and lifestyle behavior data [13]. The dataset used in this study contains 1,200 records with variables related to demographic characteristics, social media habits, sleep behavior, academic performance, physical activity, social interaction, stress level, anxiety level, addiction level, and depression label [14]. The target variable is highly imbalanced, with only a small proportion of samples labeled as depression risk. Therefore, this study not only compares several machine learning models but also evaluates the effect of imbalance handling strategies on prediction performance.

The main contributions of this study are threefold. First, this study develops a machine learning framework for early depression risk prediction among teenagers using behavioral and psychosocial indicators. Second, it evaluates the influence of imbalance handling techniques on model performance, particularly in detecting the minority depression-risk class. Third, it applies explainable analysis to identify the most influential factors contributing to depression risk prediction. The results of this study are expected to contribute to AI-driven health informatics by providing a data-driven approach for early mental health risk screening among teenagers.

2. Method

Research Design:

This study employed a quantitative experimental approach using machine learning to predict depression risk among teenagers based on social media usage, lifestyle behavior, and psychosocial indicators [15]. The main objective was to develop and evaluate predictive models capable of identifying teenagers who may be at risk of depression. Since the target class in the dataset was highly imbalanced, this study also examined the effect of imbalance handling strategies on model performance.

The overall research workflow consisted of six main stages: dataset preparation, exploratory data analysis, data preprocessing, model development, imbalance handling, model evaluation, and explainability analysis. The proposed workflow was designed not only to assess predictive performance but also to identify the most influential factors contributing to depression risk prediction.

Data Description:

The dataset used in this study consisted of 1,200 records and 13 variables related to teenagers' demographic characteristics, social media behavior, lifestyle habits, psychosocial conditions, and depression label [16]. The target variable was depression_label, which classified each sample into two classes: non-depression risk and depression risk.

The dataset contained no missing values and no duplicate records. However, the target class distribution was highly imbalanced. There were 1,169 samples labeled as non-depression risk, representing 97.42% of the dataset, and only 31 samples labeled as depression risk, representing 2.58% of the dataset. This imbalance indicates that conventional accuracy alone is not sufficient to evaluate model performance, because a model may achieve high accuracy while failing to detect the minority depression-risk class.

The variables used in the dataset are presented in **Table 1**.

Table 1. Distribution of Depression Severity Labels

Variable	Description
age	Age of the teenager
gender	Gender of the teenager
daily_social_media_hours	Average daily social media usage duration

Variable	Description
platform_usage	Most frequently used social media platform
sleep_hours	Average daily sleep duration
screen_time_before_sleep	Screen exposure before sleeping
academic_performance	Academic performance score
physical_activity	Level or duration of physical activity
social_interaction_level	Level of social interaction
stress_level	Self-reported stress level
anxiety_level	Self-reported anxiety level
addiction_level	Social media addiction tendency
depression_label	Target label for depression risk prediction

Feature Grouping:

To obtain a more meaningful analysis, this study divided the independent variables into two feature settings. The first setting used behavioral and lifestyle-related features only, while the second setting used the complete feature set by adding psychosocial indicators [17].

The behavioral-only feature setting consisted of age, gender, daily social media hours, platform usage, sleep hours, screen time before sleep, academic performance, physical activity, and social interaction level. This setting was used to evaluate whether behavioral and lifestyle indicators alone were sufficient to predict depression risk.

The full feature setting included all behavioral features and added stress level, anxiety level, and addiction level [18]. This setting was used to evaluate whether psychosocial indicators could improve the model's ability to detect teenagers at risk of depression.

Table 2. Feature Settings Used in This Study

Feature Setting	Included Variables
Behavioral Only	age, gender, daily_social_media_hours, platform_usage, sleep_hours, screen_time_before_sleep, academic_performance, physical_activity, social_interaction_level
Full Features	Behavioral-only features plus stress_level, anxiety_level, and addiction_level

This feature grouping allowed the study to compare the predictive contribution of general behavioral indicators and psychosocial indicators in depression risk prediction.

Data Preprocessing:

Data preprocessing was conducted to ensure that all variables could be properly processed by machine learning algorithms. The preprocessing stage included checking missing values, removing duplicate records, separating features and target labels, encoding categorical variables, and scaling numerical variables.

Categorical variables, including `gender`, `platform_usage`, and `social_interaction_level`, were transformed using one-hot encoding. This transformation was necessary because machine learning models require numerical input. Numerical variables, such as `age`, `daily_social_media_hours`, `sleep_hours`, `screen_time_before_sleep`, `academic_performance`, `physical_activity`, `stress_level`, `anxiety_level`, and `addiction_level`, were standardized using standard scaling. Standardization was applied to ensure that variables with different measurement ranges did not disproportionately influence distance-based or gradient-based learning algorithms.

After preprocessing, the dataset was divided into training and testing sets using an 80:20 ratio. Stratified splitting was applied to preserve the proportion of depression and non-depression classes in both subsets. This was important because the dataset had a severe class imbalance.

Imbalance Handling Strategy:

Since the depression-risk class represented only 2.58% of the dataset, imbalance handling was an essential part of the experimental design. Without proper imbalance handling, machine learning models may become biased toward the majority class and fail to identify teenagers at risk of depression.

This study evaluated three imbalance handling strategies. The first strategy used the original dataset without any imbalance treatment. This setting served as the baseline. The second strategy applied class weighting, where the minority class was assigned a higher penalty during model training. This approach encouraged the model to pay more attention to the depression-risk class. The third strategy applied the Synthetic Minority Oversampling Technique, or SMOTE, to generate synthetic samples for the minority class in the training data.

SMOTE was applied only to the training data to prevent data leakage. The testing data remained unchanged to ensure that model evaluation reflected real class distribution.

Table 3. Imbalance Handling Strategies

Strategy	Description
Without imbalance handling	Original training data were used without modification
Class weighting	Higher weight was assigned to the minority depression-risk class
SMOTE	Synthetic minority samples were generated in the training data

Machine Learning Models:

Several machine learning algorithms were evaluated to compare their performance in depression risk prediction. The selected models included Logistic Regression, Random Forest, Support Vector Machine with radial basis function kernel, and Gradient Boosting [19]. These models were selected because they represent different learning characteristics.

Logistic Regression was used as a baseline linear classifier. Random Forest was used as an ensemble tree-based model capable of capturing nonlinear relationships among variables. Support Vector Machine was included because of its ability to construct decision boundaries in high-dimensional feature spaces. Gradient Boosting was used as a boosting-based ensemble method that sequentially improves weak learners.

Table 4. Machine Learning Models Used in This Study

Model	Rationale (Alasan Pemilihan)
Logistic Regression	Baseline linear classifier and interpretable model
Random Forest	Ensemble model suitable for nonlinear feature relationships
Support Vector Machine	Effective for complex decision boundaries
Gradient Boosting	Boosting-based model for improving predictive performance

Each model was trained under different feature settings and imbalance handling strategies. This experimental design enabled a comparative analysis of how feature selection and imbalance treatment affected the detection of depression risk.

Threshold Tuning:

In binary classification, the default decision threshold is commonly set at 0.50. However, in imbalanced health prediction tasks, this threshold may not be optimal because the model may produce low sensitivity toward the minority class. Therefore, this study applied threshold tuning to improve the detection of depression-risk samples.

The threshold was selected using validation data by maximizing the F2-score. The F2-score gives more weight to recall than precision, making it suitable for early mental health risk screening, where failing to identify at-risk individuals may be more critical than producing some false-positive predictions.

The F2-score is defined as:

$$F_2 = \frac{(1 + 2^2) \times Precision \times Recall}{(2^2 \times Precision) + Recall}$$

By using F2-score for threshold selection, the model was optimized to prioritize recall for the depression-risk class.

Model Evaluation:

Model performance was evaluated using several metrics, including accuracy, precision, recall, F1-score, balanced accuracy, specificity, ROC-AUC, PR-AUC, Cohen's Kappa, MAE, and RMSE. Since the dataset was highly imbalanced, the main evaluation emphasis was placed on recall, F1-score, balanced accuracy, and PR-AUC rather than accuracy alone.

Accuracy measures the overall proportion of correct predictions. However, in imbalanced datasets, accuracy may be misleading because the model can obtain high accuracy by predicting most samples as the majority class. Therefore, recall for the depression-risk class was considered important because it measures the model's ability to correctly identify teenagers at risk of depression.

Precision measures the proportion of predicted depression-risk cases that are actually positive. F1-score combines precision and recall into a single metric. Balanced accuracy calculates the average of sensitivity and specificity, making it more suitable for imbalanced datasets. ROC-AUC evaluates the model's ability to distinguish between classes across different thresholds, while PR-AUC is especially useful when the positive class is rare.

The confusion matrix was also used to examine the number of true positives, false positives, true negatives, and false negatives. In the context of depression risk prediction, false negatives are particularly important because they represent teenagers who are actually at risk but are not detected by the model.

Explainability Analysis:

To improve model interpretability, this study applied feature importance analysis using permutation importance [19]. This method measures the decrease in model performance when the values of a specific feature are randomly shuffled. A larger decrease indicates that the feature has a stronger contribution to model prediction.

Explainability analysis was conducted to identify the most influential behavioral and psychosocial factors associated with depression risk prediction [20]. This step is important because machine learning models in healthcare-related applications should not only provide predictive outputs but also offer understandable explanations. By identifying influential features, the model can provide insights into whether variables such as social media duration, sleep hours, stress level, anxiety level, addiction level, or physical activity contribute significantly to depression risk prediction.

Experimental Workflow:

The experimental procedure in this study can be summarized as follows. First, the dataset was loaded and checked for missing values, duplicate records, data types, and class distribution. Second, exploratory data analysis was conducted to understand the characteristics of the dataset and the distribution of the target variable. Third, the dataset was divided into behavioral-only and full-feature settings. Fourth, preprocessing was applied using standard scaling for numerical variables and one-hot encoding for categorical variables. Fifth, the data were split into training and testing sets using stratified sampling. Sixth, several machine learning models were trained using different imbalance

handling strategies. Seventh, model performance was evaluated using multiple metrics suitable for imbalanced binary classification. Finally, explainability analysis was conducted to identify the most important features contributing to depression risk prediction.

The proposed method is intended as an early risk screening approach and not as a replacement for clinical diagnosis. The model output should be interpreted as a predictive indication of potential depression risk based on behavioral and psychosocial patterns.

3. Result and Discussion

Class Distribution and Data Characteristics:

The dataset used in this study consisted of 1,200 records with two depression-risk classes. As shown in Figure 1, the class distribution was highly imbalanced. The non-depression class contained 1,169 samples, representing 97.42% of the dataset, while the depression-risk class contained only 31 samples, representing 2.58% of the dataset. This indicates an imbalance ratio of approximately 37.7:1 between the majority and minority classes.

Table 5. Distribution of Depression Label

Depression Label	Class Description	Number of Samples	Percentage
0	Non-depression	1.169	97.42%
1	Depression	31	2.58%
Total		**1.2**	**100%**

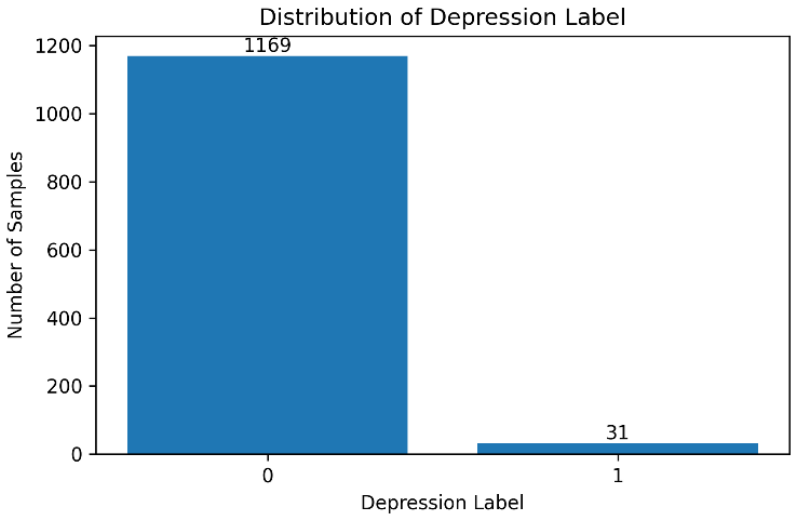

Figure 1. Distribution of depression label.

The severe class imbalance in the dataset is an important methodological issue. If accuracy is used as the only evaluation metric, a model may appear to perform well even when it fails to detect the minority depression-risk class.

For example, a model that predicts all samples as non-depression could still achieve high accuracy because most samples belong to the majority class. Therefore, this study emphasized recall, F1-score, balanced accuracy, ROC-AUC, and PR-AUC to provide a more reliable assessment of model performance.

Descriptive Comparison Between Depression and Non-Depression Groups:

The descriptive statistics showed meaningful differences between teenagers labeled as non-depression and those labeled as depression risk. Teenagers in the depression-risk group had higher average daily social media use, lower average sleep duration, and higher stress and anxiety levels.

Table 6. Mean Values of Numerical Features by Depression Label

Feature	Non-Depression	Depression
Age	15.92	16.06
Daily social media hours	4.48	6.72
Sleep hours	6.49	4.76
Stress level	5.37	8.48
Anxiety level	5.56	8.61
Addiction level	5.57	5.32

The depression-risk group used social media for an average of 6.72 hours per day, compared with 4.48 hours in the non-depression group. The same group also had lower average sleep duration, with 4.76 hours compared with 6.49 hours among non-depression samples. In addition, stress and anxiety levels were substantially higher in the depression-risk group. These patterns suggest that depression risk in this dataset is more strongly associated with sleep reduction, stress, anxiety, and social media exposure than with age, academic performance, or physical activity.

The correlation matrix in **Figure 2** supports this descriptive pattern. The depression label showed a positive correlation with daily social media hours, stress level, and anxiety level, while sleep hours showed a negative correlation with depression label. Although the correlation values were not extremely high, the direction of the relationship was consistent with the mean comparison in Table 6.

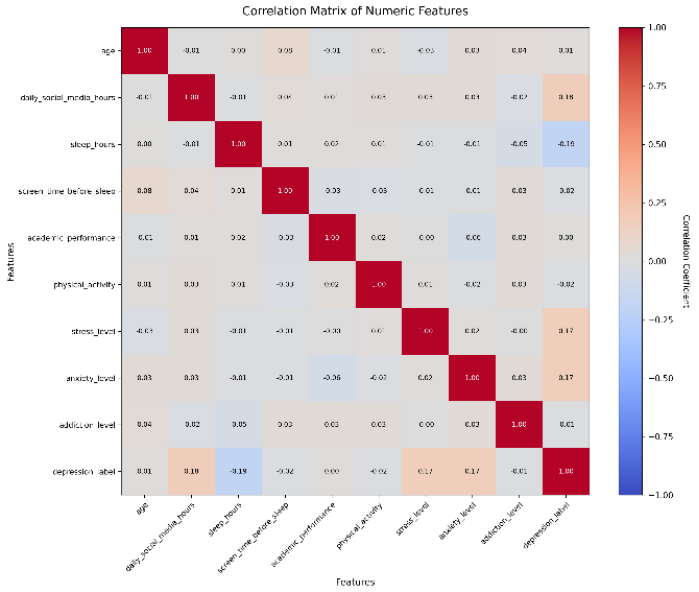

Figure 2. Correlation matrix of numerical features.

Model Performance Based on Feature Setting and Imbalance Handling:

The experimental results showed a clear difference between the behavioral-only feature setting and the full-feature setting. The behavioral-only setting produced relatively weak predictive performance, while the full-feature setting produced substantially higher results. This indicates that psychosocial indicators, particularly stress level and anxiety level, contributed strongly to depression-risk prediction.

Table 7. Average Model Performance by Feature Setting and Imbalance Handling Using Tuned Threshold

Feature Set	Imbalance Method	Accuracy	Precision	Recall	F1-score	ROC-AUC
Behavioral Only	None	9,115	1,204	4,167	1,860	8,608
Behavioral Only	SMOTE	9,198	1,728	5,417	2,570	9,014
Full Features	None	9,927	9,167	7,917	8,333	9,966
Full Features	SMOTE	9,917	8,929	7,917	8,205	9,959

The behavioral-only models had low PR-AUC values, ranging from 0.1508 to 0.1793. This means that social media behavior and lifestyle variables alone were not sufficient to strongly distinguish depression-risk samples from non-depression samples. Although imbalance handling improved recall, the precision and F1-score remained low in the behavioral-only setting.

In contrast, the full-feature setting achieved much stronger performance. The average PR-AUC increased to 0.9208 without imbalance handling, 0.8950 with class weighting, and 0.9182 with SMOTE. This result indicates that the inclusion of psychosocial variables, namely stress level, anxiety level, and addiction level, substantially improved predictive performance.

Best Model Performance:

The best-performing scenario was obtained using the full-feature setting with Random Forest without imbalance handling and a tuned decision threshold. The model achieved perfect classification performance on the test set, with accuracy, precision, recall, F1-score, balanced accuracy, ROC-AUC, PR-AUC, and Cohen's Kappa all reaching 1.0000.

Table 8. Best Model Performance

Component	Result
Feature Setting	Full Features
Model	Random Forest
Imbalance Handling	None
Threshold Strategy	Tuned by validation F2-score
Threshold	0.18
Accuracy	1.2000
Precision	1.2000
Recall	1.2000
F1-score	1.2000
Balanced Accuracy	1.2000
ROC-AUC	1.2000
PR-AUC	1.2000
Cohen's Kappa	1.2000
MAE	0
RMSE	0

The confusion matrix in **Figure 3** shows that the model correctly classified all 240 test samples. It identified 234 non-depression samples as non-depression and 6 depression-risk samples as depression risk. There were no false positives and no false negatives.

Table 9. Confusion Matrix of the Best Model

Actual Class	Predicted Non-Depression	Predicted Depression
Non-Depression	234	0
Depression	0	6

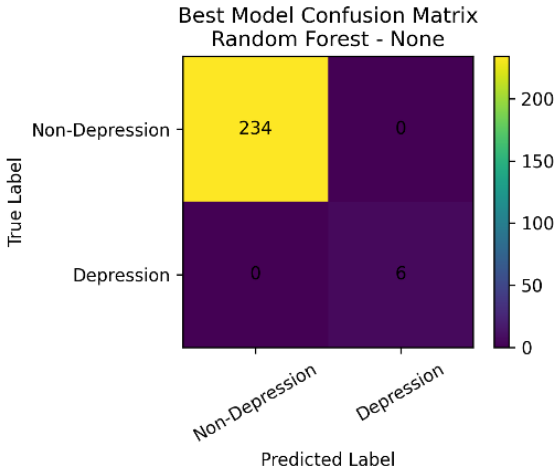

Figure 3. Confusion matrix of the best model.

The ROC curve and precision-recall curve further confirmed the best model's discriminative ability. As shown in **Figure 4**, the PR-AUC value reached 1.0000, indicating perfect precision-recall performance for the minority depression-risk class. Similarly, Figure 5 shows an ROC-AUC of 1.0000, indicating complete separation between the two classes in the test data.

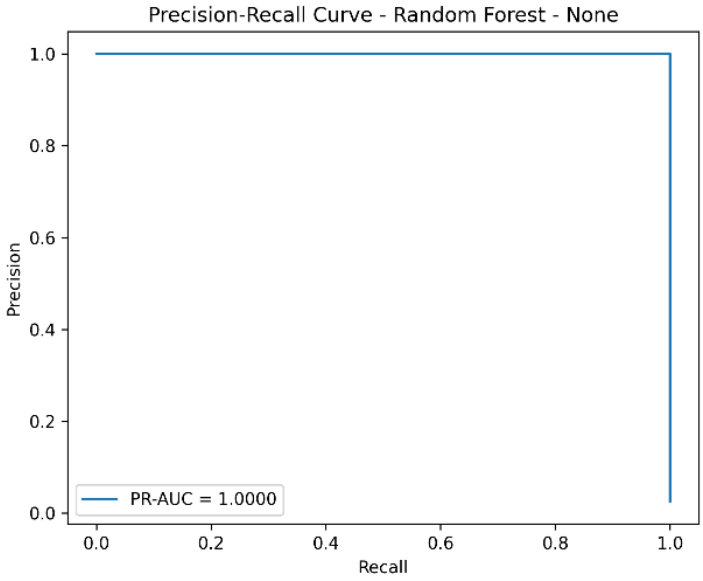

Figure 4. Precision-recall curve of the best model.

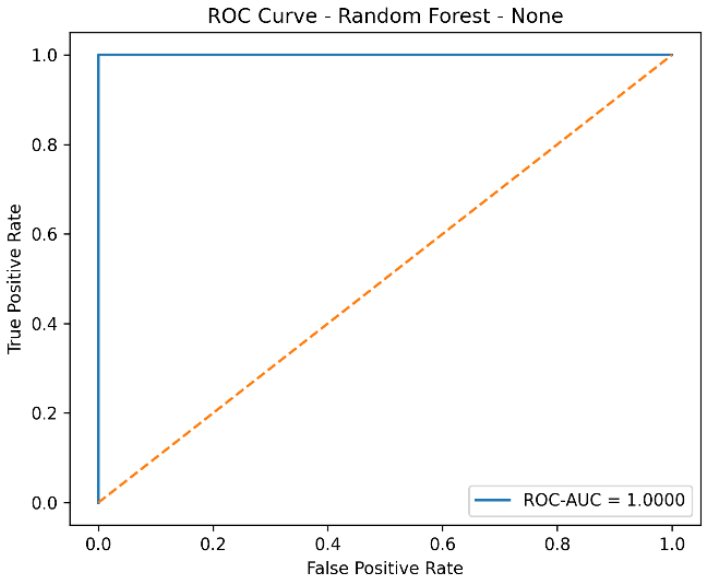

Figure 5. ROC curve of the best model.

However, this perfect performance should be interpreted carefully. The result suggests that the depression label in this dataset may be highly separable based on the available features, particularly sleep hours, stress level, anxiety level, and daily social media hours. Since the test set contained only 6 depression-risk samples, perfect performance may also be influenced by the small number of positive test cases. Therefore, although the model achieved excellent performance in this experiment, further validation using larger, real-world, and clinically verified datasets is necessary before the model can be considered reliable for practical mental health screening.

Feature Importance Analysis:

Permutation importance was used to identify the most influential variables in the best-performing model. The results showed that sleep hours had the highest contribution to model performance, followed by stress level, anxiety level, and daily social media hours.

Table 10. Feature Importance Based on Permutation Importance

Rank	Feature	Importance Mean	Importance Std
1	Sleep hours	7,526	1,016
2	Stress level	7,035	1,165
3	Anxiety level	5,772	1,601
4	Daily social media hours	4,998	810
5	Platform usage	0	0
6	Gender	0	0
7	Age	0	0

Rank	Feature	Importance Mean	Importance Std
8	Screen time before sleep	0	0
9	Physical activity	0	0
10	Academic performance	0	0
11	Social interaction level	0	0
12	Addiction level	0	0

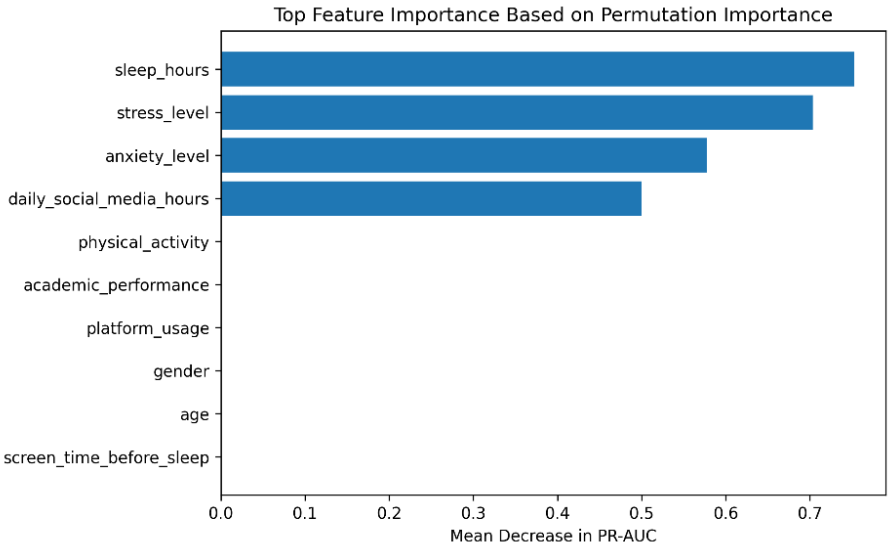

Figure 6. Top feature importance based on permutation importance.

The feature importance results are consistent with the descriptive statistics. Teenagers labeled as depression risk had lower sleep duration and higher stress and anxiety levels than those in the non-depression group. Daily social media hours also contributed substantially to prediction, indicating that prolonged social media use was an important behavioral indicator in the model.

Interestingly, several variables such as age, gender, platform usage, academic performance, physical activity, screen time before sleep, social interaction level, and addiction level had zero importance in the best model. This does not necessarily mean that these variables are irrelevant in all mental health contexts. Instead, it means that, within this dataset and model configuration, these variables did not provide additional predictive value after the model had already used sleep hours, stress level, anxiety level, and daily social media hours.

Discussion:

The findings of this study indicate that machine learning can effectively predict depression risk among teenagers when behavioral and psychosocial indicators are used together. The comparison between behavioral-only and full-

feature settings showed that social media and lifestyle variables alone produced limited predictive performance. However, when stress level, anxiety level, and addiction level were added, model performance improved substantially.

This suggests that depression-risk prediction should not rely solely on social media usage duration. Although teenagers in the depression-risk group had higher average daily social media use, the model's strongest predictors were sleep hours, stress level, anxiety level, and daily social media hours. Therefore, social media use should be interpreted as part of a broader behavioral and psychosocial pattern rather than as a single causal factor.

The results also highlight the importance of using appropriate evaluation metrics in imbalanced mental health prediction. The depression-risk class represented only 2.58% of the dataset. In such a condition, accuracy alone is insufficient because it can hide poor detection of the minority class. PR-AUC, recall, F1-score, and balanced accuracy provide more meaningful insights into the model's ability to identify teenagers at risk of depression. The use of threshold tuning based on F2-score was also useful because it prioritized recall, which is important in early risk screening.

The role of imbalance handling was more visible in the behavioral-only setting. Class weighting and SMOTE improved recall and F1-score compared with the baseline setting. However, in the full-feature setting, the difference between imbalance handling methods became smaller because the added psychosocial variables made the classes easier to distinguish. This indicates that feature quality had a stronger effect than resampling strategy in this dataset.

From an explainability perspective, the feature importance results provide practical insight into the main indicators associated with depression-risk prediction. Sleep hours emerged as the most influential feature, followed by stress level, anxiety level, and daily social media hours. This finding supports the idea that adolescent mental health screening should consider sleep behavior and psychosocial conditions alongside digital behavior.

Nevertheless, several limitations should be acknowledged. First, the number of depression-risk samples was very small, with only 31 positive cases in the entire dataset and 6 positive cases in the test set. Second, the perfect performance of the best model suggests that the dataset may contain highly separable or rule-based label patterns. Third, the dataset should not be interpreted as a substitute for clinical diagnosis because depression is a complex mental health condition that requires professional assessment. Therefore, the model developed in this study should be positioned as an early risk screening framework rather than a clinical diagnostic tool.

Overall, the results demonstrate that explainable machine learning can provide a useful framework for identifying depression risk among teenagers using social media behavior, lifestyle patterns, and psychosocial indicators. The study also shows that handling class imbalance and applying interpretable analysis are essential in developing responsible AI-based mental health prediction systems.

4. Conclusion

This study developed an explainable machine learning framework for predicting depression risk among teenagers using social media usage, lifestyle behavior, and psychosocial indicators. The dataset consisted of 1,200 records with a highly imbalanced depression label, where only 31 samples, or 2.58%, were categorized as depression risk. This

imbalance made the prediction task methodologically challenging and required evaluation metrics beyond accuracy, such as recall, F1-score, balanced accuracy, ROC-AUC, and PR-AUC.

The experimental results showed that the full-feature setting, which combined behavioral, lifestyle, and psychosocial variables, produced substantially better performance than the behavioral-only setting. The best-performing model was Random Forest using full features without imbalance handling, which achieved perfect performance on the test set, with accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC all reaching 1.0000. The confusion matrix showed that all 234 non-depression samples and 6 depression-risk samples in the test set were correctly classified.

Feature importance analysis revealed that sleep hours, stress level, anxiety level, and daily social media hours were the most influential predictors of depression risk. Among these variables, sleep hours had the highest contribution, followed by stress level and anxiety level. These findings indicate that depression risk among teenagers in this dataset was not only associated with social media use but also strongly related to sleep behavior and psychosocial conditions. Therefore, early mental health risk screening should consider multiple behavioral and psychological factors rather than relying only on social media usage duration.

Although the proposed model achieved excellent performance, the results should be interpreted with caution. The number of depression-risk samples was very small, and the perfect classification result may indicate that the dataset contains highly separable or rule-based label patterns. Therefore, the model should not be considered a clinical diagnostic tool. Instead, it should be positioned as an AI-based early risk screening framework that may support preliminary identification of teenagers who require further attention or professional assessment.

Future research should validate the proposed approach using larger, more diverse, and clinically verified datasets. Further studies may also explore deep learning models, longitudinal behavioral data, multimodal data sources, and more advanced explainable AI techniques such as SHAP to improve interpretability and reliability. In addition, collaboration with mental health professionals is necessary to ensure that AI-based depression risk prediction can be developed responsibly and ethically for adolescent mental health support.

References:

[1] B. Lu, L. Lin, and X. Su, “Global burden of depression or depressive symptoms in children and adolescents: A systematic review and meta-analysis,” *J. Affect. Disord.*, vol. 354, pp. 553–562, Jun. 2024, doi: 10.1016/j.jad.2024.03.074.

[2] N. Saleem, P. Young, and S. Yousuf, “Exploring the Relationship Between Social Media Use and Symptoms of Depression and Anxiety Among Children and Adolescents: A Systematic Narrative Review,” *Cyberpsychology Behav. Soc. Netw.*, vol. 27, no. 11, pp. 771–797, Nov. 2024, doi: 10.1089/cyber.2023.0456.

[3] D. J. Korczak, C. Westwell-Roper, and R. Sassi, “Diagnosis and management of depression in adolescents,” *Can. Med. Assoc. J.*, vol. 195, no. 21, pp. E739–E746, May 2023, doi: 10.1503/cmaj.220966.

[4] N. Agyapong-Opoku, F. Agyapong-Opoku, and A. J. Greenshaw, “Effects of Social Media Use on Youth and Adolescent Mental Health: A Scoping Review of Reviews,” *Behav. Sci.*, vol. 15, no. 5, p. 574, Apr. 2025, doi: 10.3390/bs15050574.

[5] H. Shannon, K. Bush, P. J. Villeneuve, K. G. Hellemans, and S. Guimond, “Problematic Social Media Use in Adolescents and Young Adults: Systematic Review and Meta-analysis,” *JMIR Ment. Health*, vol. 9, no. 4, p. e33450, Apr. 2022, doi: 10.2196/33450.

[6] Z. Yue and M. Rich, “Social Media and Adolescent Mental Health,” *Curr. Pediatr. Rep.*, vol. 11, no. 4, pp. 157–166, Sep. 2023, doi: 10.1007/s40124-023-00298-z.

[7] D. J. Yu, Y. K. Wing, T. M. H. Li, and N. Y. Chan, “The Impact of Social Media Use on Sleep and Mental Health in Youth: a Scoping Review,” *Curr. Psychiatry Rep.*, vol. 26, no. 3, pp. 104–119, Mar. 2024, doi: 10.1007/s11920-024-01481-9.

[8] J. M. Nagata *et al.*, “Social Media Use and Depressive Symptoms During Early Adolescence,” *JAMA Netw. Open*, vol. 8, no. 5, p. e2511704, May 2025, doi: 10.1001/jamanetworkopen.2025.11704.

[9] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, and J. Guo, “Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review,” *JMIR Ment. Health*, vol. 9, no. 3, p. e27244, Mar. 2022, doi: 10.2196/27244.

[10] D. Phiri, F. Makowa, V. L. Amelia, Y. V. A. Phiri, L. P. Dlamini, and M.-H. Chung, “Text-Based Depression Prediction on Social Media Using Machine Learning: Systematic Review and Meta-Analysis,” *J. Med. Internet Res.*, vol. 27, p. e59002, Apr. 2025, doi: 10.2196/59002.

[11] O. Ahmed, E. I. Walsh, A. Dawel, K. Alateeq, D. A. Espinoza Oyarce, and N. Cherbuin, “Social media use, mental health and sleep: A systematic review with meta-analyses,” *J. Affect. Disord.*, vol. 367, pp. 701–712, Dec. 2024, doi: 10.1016/j.jad.2024.08.193.

[12] B. Brosnan, J. J. Haszard, K. A. Meredith-Jones, S.-R. Wickham, B. C. Galland, and R. W. Taylor, “Screen Use at Bedtime and Sleep Duration and Quality Among Youths,” *JAMA Pediatr.*, vol. 178, no. 11, p. 1147, Nov. 2024, doi: 10.1001/jamapediatrics.2024.2914.

[13] W. B. Tahir, S. Khalid, S. Almutairi, M. Abohashrh, S. A. Memon, and J. Khan, “Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques,” *IEEE Access*, vol. 13, pp. 12789–12818, 2025, doi: 10.1109/ACCESS.2025.3530862.

[14] L. As. Brautsch, L. Lund, M. M. Andersen, P. J. Jennum, A. P. Folker, and S. Andersen, “Digital media use and sleep in late adolescence and young adulthood: A systematic review,” *Sleep Med. Rev.*, vol. 68, p. 101742, Apr. 2023, doi: 10.1016/j.smrv.2022.101742.

[15] U. M. Haque, E. Kabir, and R. Khanam, “Detection of child depression using machine learning methods,” *PLOS ONE*, vol. 16, no. 12, p. e0261131, Dec. 2021, doi: 10.1371/journal.pone.0261131.

[16] A. Yoo *et al.*, “Prediction of adolescent depression from prenatal and childhood data from ALSPAC using machine learning,” *Sci. Rep.*, vol. 14, no. 1, p. 23282, Oct. 2024, doi: 10.1038/s41598-024-72158-9.

[17] M. T. Hawes, H. A. Schwartz, Y. Son, and D. N. Klein, “Predicting adolescent depression and anxiety from multi-wave longitudinal data using machine learning,” *Psychol. Med.*, vol. 53, no. 13, pp. 6205–6211, Oct. 2023, doi: 10.1017/S0033291722003452.

[18] X. Liu *et al.*, “Identification of depressive symptoms in adolescents using machine learning combining childhood and adolescence features,” *BMC Public Health*, vol. 25, no. 1, p. 264, Jan. 2025, doi: 10.1186/s12889-025-21506-z.

[19] Md. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, “An in-depth analysis of machine learning approaches to predict depression,” *Curr. Res. Behav. Sci.*, vol. 2, p. 100044, Nov. 2021, doi: 10.1016/j.crbeha.2021.100044.

[20] Y. Xin *et al.*, “Machine learning-based analysis and prediction of factors influencing mental health among children and adolescents in Jiangsu Province,” *Child Adolesc. Psychiatry Ment. Health*, vol. 19, no. 1, p. 100, Aug. 2025, doi: 10.1186/s13034-025-00959-5.