Detecting the Impact of Social Media on Users' Mental Health Using Machine Learning and XAI

Ara Bela Zulfa Laila[1]
Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia
Sekaran, Gunungpati, Semarang, 50229, Central Java, Indonesia
Email: [1] arabelazulfalaila08@students.unnes.ac.id

Abstrak. *Penelitian ini mengembangkan sistem prediktif berbasis machine learning untuk mendeteksi potensi depresi akibat penggunaan media sosial, serta membandingkan kinerja algoritma seperti Random Forest, XGBoost, dan Naïve Bayes. Data survei yang meliputi usia, jenis kelamin, status hubungan, durasi penggunaan harian, dan platform media sosial digunakan untuk membangun model dengan evaluasi akurasi, precision, recall, dan F1-score. XGBoost menunjukkan kinerja terbaik dengan akurasi 90% dan F1-score tinggi. Fitur utama yang memengaruhi prediksi depresi meliputi durasi penggunaan media sosial, usia, dan platform. Teknik Explainable AI (XAI) dengan LIME meningkatkan transparansi model, memberikan penjelasan yang relevan untuk individu, dan memperkuat kepercayaan terhadap prediksi. Penelitian ini menekankan pentingnya transparansi dalam penerapan model di bidang kesehatan mental dan menawarkan solusi fleksibel yang dapat diadopsi untuk aplikasi digital seperti chatbot atau dashboard pemantauan kesehatan mental real-time.*
Kata Kunci: *Machine Learning, Media Sosial, Kesehatan Mental, Explainable AI, XGBoost.*

Abstract. *This research develops a machine learning-based predictive system to detect potential depression due to social media use, and compares the performance of algorithms such as Random Forest, XGBoost, and Naïve Bayes. Survey data, including age, gender, relationship status, daily usage duration, and social media platform, were used to build the model, with accuracy, precision, recall, and F1-score evaluated. XGBoost showed the best performance with 90% accuracy and a high F1-score. The main features that affect depression prediction include duration of social media use, age, and platforms. Explainable AI (XAI) techniques with LIME increase the transparency of the model, provide relevant explanations for individuals, and strengthen confidence in the predictions. This research emphasizes the importance of transparency in model implementation in the mental health field and offers a flexible solution that can be adopted for digital applications such as chatbots or real-time mental health monitoring dashboards.*
Keywords: *Machine Learning, Social Media, Mental Health, Explainable AI, XGBoost.*

1. Introduction

Social media has now become an integral part of daily life, serving as the primary platform for individuals to communicate, share information, express themselves, and build global social networks. With the growing popularity of platforms such as Instagram, Twitter, TikTok, and Facebook, individuals can interact without the constraints of time and space. In addition to being communication tools, these platforms are also widely used for seeking information and entertainment. However, the rapid increase in social media usage over the past few years has heightened the risk of addiction among its users [1].

Social media addiction can be defined as psychological dependence on social media that causes behavioral addiction symptoms [2]. Excessive use of social media can also amplify the potential impact of daytime media exposure on nighttime sleep quality. The higher the level of social media use during the day, the greater the extent to which it affects a person's sleep architecture and dominates their nighttime dreams [3]. Nightmares are closely associated with psychological issues such as depression and anxiety, and can be risk factors for anxiety disorders

and suicidal thoughts [4]. Montag and Hegelich [5] found that excessive social media use is associated with increased anxiety and depression. In addition, Woods and Scott [6], through a self-report survey, also found that the intensity of social media use influences feelings of loneliness and low self-esteem.

Machine learning classification models can be used in analyzing this impact because machine learning uses historical data that can be used to predict the future, allowing computers to learn from data without having to do explicit programming [7], [8]. In addition, Explainable AI (XAI) approaches are starting to be developed to increase transparency in machine learning models, so that prediction results can be more easily understood and interpreted by users, practitioners, and policymakers.

Several previous studies have applied machine learning methods to detect the impact of social media on mental health. For example, research conducted by Agarwal *et al*. [9] uses a social media data-based diagnosis approach with an ensemble deep learning approach designed to monitor users' linguistic expressions and recognize patterns related to mental health disorders more accurately through enhanced features. This can improve understanding of the relationship between social media activity and mental health disorders in more depth. Another research conducted by Yu *et al*. [10] emphasized that sleep disturbance can be a mediator between the intensity of social media use and the emergence of symptoms of psychological disorders such as anxiety and depression, which is why early detection based on digital data, such as social media, is relevant as a mental health intervention effort.

Although many studies have been successful in building Machine Learning-based detection models, the majority of such approaches still lack transparency in explaining how the model makes decisions. Most available research tends to focus on achieving high prediction accuracy, without adequately explaining the most significant factors that influence the mental health of social media users. Additionally, there is a lack of studies that use Explainable Artificial Intelligence (XAI) to contextualize social media-based mental health detection, despite the importance of model transparency in fostering trust and ease of use in analysis outcomes [11].

This study employs a methodology that involves the use of secondary datasets containing survey responses regarding social media usage behavior, the preprocessing of structured (tabular) data, and the application of several classification machine learning models, including Random Forest, XGBoost, and Naïve Bayes, to predict the level of impact on mental health. Additionally, Explainable AI techniques such as LIME will be applied to explain how these models generate their predictions, providing more transparent and practical interpretations for mental health professionals or practitioners [12], [13].

The study adds genuine value in three ways: first, it builds sharper machine-learning tools that spot how social media affects people's minds; second, it uses explainable-AI techniques so researchers and lay users can actually understand the results; and third, it hands practical findings to policymakers, health workers, and platform designers who want to craft smarter, digital-age support programs.

2. Literature Review

As mentioned earlier, research by Yu *et al*. [10] found that excessive use of social media has a negative impact on sleep quality and mental well-being among adolescents, primarily due to social pressure and device use during sleep. This finding is also supported by research by Meynadier *et al*. [14], which shows that social media addiction has a strong correlation with factors such as depression, anxiety, loneliness, and fear of missing out on information, often referred to as FOMO. Additionally, Xiao *et al*. [15] also noted that mobile phones have a mediating effect in the relationship between physical activity and sleep disorders, as well as addiction among adolescents. This addiction can even be considered severe as it is associated with high levels of anxiety, which is a common predictor of sleep disorders and poor mental health.

In the development of technology, machine learning approaches have been utilized by many previous researchers to identify indicators of mental disorders. Geetha *et al*. [8] developed a Multi-Layer Perceptron (MLP) model capable of detecting stress levels based on sleep patterns,

demonstrating the great potential of machine learning in detecting psychological issues through physiological and behavioral data. Similar research conducted by Alshammari [16] also shows algorithms such as Artificial Neural Network (ANN) that can classify sleep disorders with high enough accuracy and open up opportunities in detecting health problems related to sleep patterns.

In terms of social media data processing, ensemble-based approaches such as the one developed by Agarwal *et al.* In [9] showed effectiveness in identifying mental disorders based on the content of posts and interactions between users on social media. Thus, this research is at the intersection of the need for early detection and the interpretability of analysis results. By integrating Machine Learning and Explainable AI approaches to analyze the impact of social media on mental health, it is hoped that this research can make a meaningful contribution to the development of decision support systems in the field of mental health, especially in the context of today's digital society.

3. Research Method

3.1 Research Dataset

This study uses a structured secondary dataset obtained from the public Kaggle repository titled "Social Media and Mental Health." According to its original description, this dataset was collected through an online survey to identify the relationship between time spent on social media and its impact on mental health. The sampling method for this dataset is convenience sampling, in which participants voluntarily completed the provided questionnaire.

In total, the data used in this study were collected from 481 respondents and consist of 21 feature attributes. The data consists of tabular data comprising both numerical and categorical values. An analysis of the class distribution for the target variable (divided into 5 categories: 0, 1, 2, 3, and 4) reveals the following sample distribution: Class 2 has more than 117 samples, followed by class 3 (116 samples), class 4 (105 samples), class 1 (83 samples), and class 0 has only 60 samples. This indicates a balanced dataset for multi-class classification modeling. This presentation of the class distribution is provided to offer transparency regarding the characteristics of the data to be trained using machine learning algorithms.

3.2 Research Phases

This study was designed with several research phases, including data exploration, data preprocessing and dimensionality reduction (PCA), model training and performance evaluation, and result interpretation using Explainable AI. The flow of these research phases is shown in Figure 1, which illustrates the sequence of processes involved in building the system for this study:

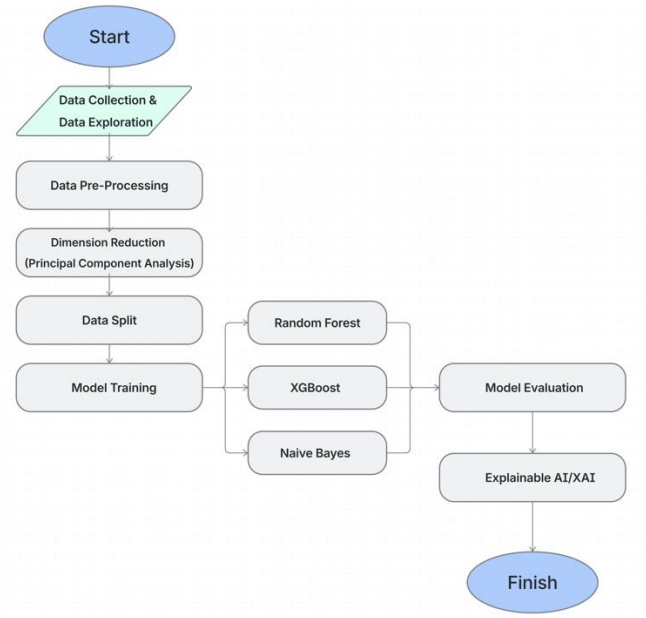

Figure 1. System flowchart

The first stage of this study involves data collection and exploratory data analysis (EDA). This stage includes a brief description of the dataset obtained, which contains features such as age, gender, relationship status, daily social media usage time, and social media usage intensity. Next is understanding the relationships between features to identify which features have the greatest influence on depression.

This is followed by data preprocessing, which involves removing null or empty values and replacing them with zeros. Then, encoding is performed by converting categories into numerical values using label encoding and one-hot encoding. The target feature to be classified is depression (multi-class label classification). After that, the next stage is Principal Component Analysis (PCA). PCA is used for dimensionality reduction and to determine the extent to which each feature contributes to the data variance.

The most important stage in this research is modeling or making machine learning models. Before training the model, several research variables must be defined so that the model performance evaluation process has clear parameters. The operational definitions of these variables are shown in Table 1 below.

Table 1. Table of Operational Definitions

No	Variable	Operational Definition	Parameters	Scale
1	Classification Algorithms	Machine learning models are used for model prediction.	Use of Random Forest, XGBoost, and Naive Bayes.	Nominal
2	Accuracy	The proportion of correct predictions, both true positives (TP) and true negatives (TN), out of the total number of predictions	$\frac{TP + TN}{TP + TN + FP + FN}$	Ratio
3	Precision	The ratio of true positive predictions to the total number of positive predictions.	$\frac{TP}{TP + FP}$	Ratio
4	Recall	The proportion of actual positive cases correctly predicted out of the total number of actual positive cases	$\frac{TP}{TP + FN}$	Ratio
5	F1-Score	The harmonic mean of precision and recall is used to measure the balance between the two.	$2 \times \frac{Precision \times Recall}{Precision \times Recall}$	Rasio
6	ROC-AUC	A probabilistic metric representing the area under the ROC curve, used to evaluate a model's discriminative ability to distinguish between classes using the One-vs-Rest (OvR) approach.	$AUC_{OvR} = \frac{1}{C}\sum_{i=1}^{C} AUC_i$	Ratio

Modeling starts by dividing the data into training data and test data (x_train, x_test, y_train, y_test). The model used is to compare the performance of Random Forest, XGBoost, and Naïve Bayes. Random Forest is a bagging-based ensemble learning method that builds many decision trees at random to reduce variance. This makes the model more resistant to overfitting when handling complex data.

While Naïve Bayes is a method for predicting target classes based on the principle of conditional probability, assuming strong feature independence. Mathematically, the probability of class C given a set of features X is formulated as $P(C|X) = \frac{P(X|C)P(C)}{P(X)}$. On the other hand, XGBoost is a boosting-based ensemble algorithm that aims to minimize the loss function iteratively. The basic objective function in XGBoost involves regularization to control model complexity, namely: $obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$, where the l component calculates the difference between the prediction and the actual value, and Ω is a penalty to prevent overly complex tree structures [17].

Once the variables have been defined, the process continues with hyperparameter tuning to find the best combination of parameters and optimize the model using RandomizedSearchCV. This process involves exploring parameters such as n_estimators, max_depth, learning_rate, subsample, gamma, and others. Empirical results show that this process effectively balances bias and variance and significantly improves classification performance metrics compared to models without optimization [18].

Subsequently, the model training process continued with a comprehensive evaluation using the metrics Accuracy, Precision, Recall, and F1-Score to assess the initial performance of each algorithm. To ensure the validity and generalizability of the model and to avoid bias in data splitting, additional testing was conducted using 5-fold cross-validation. Additionally, the ROC-AUC (Receiver Operating Characteristic – Area Under Curve) metric was used to measure the model's discriminative ability in distinguishing between classes probabilistically. All evaluation results were then analyzed in greater depth using a Confusion Matrix to identify patterns of prediction errors (False Positives and False Negatives), thereby determining the model with the best and most stable performance.

The last stage in this research is the final explanation using Explainable AI. The type of explainable AI that will be used in this research is LIME. LIME (Local Interpretable Model-agnostic Explanations) is used to explain the reasons for the model's predictions locally (per individual) and to display important features that influence the model's decisions.

4. Result and Discussion

4.1 Data Analysis and Visualization

Data analysis and visualization conducted to understand the patterns, distribution, and relationships between the datasets showed several results. The visualization of the analysis of the average time of social media use per day, shown in Figure 2, shows that the majority of respondents use social media for around three to six hours per day. In fact, the number of respondents who use social media for more than five hours is the highest, and only a few respondents use social media between four and five hours. This high duration of use can have an effect on increasing the risk of mental health disorders in its users.

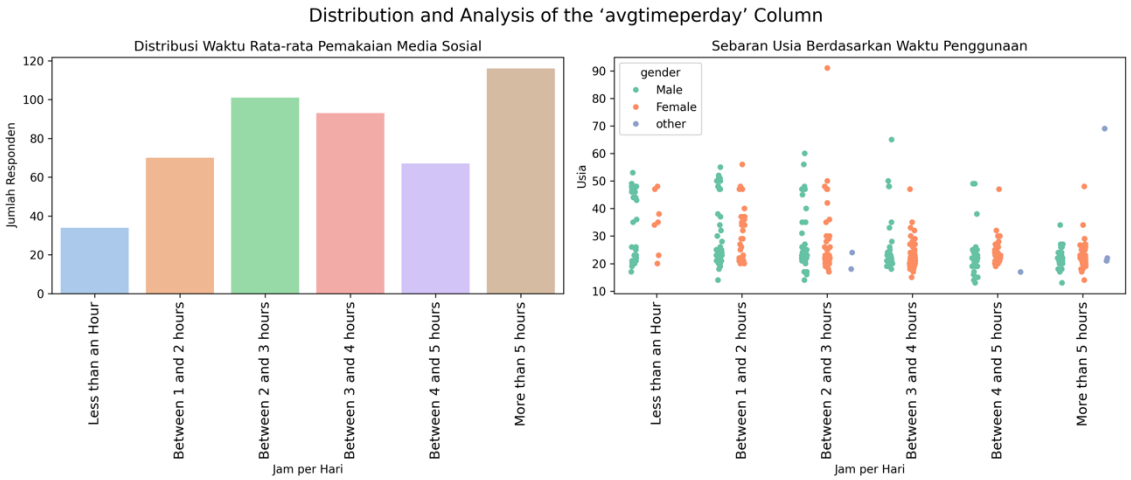

Figure 2. Average Time of Social Media Use Per Day.

The relation between the amount of time spent on social media and negative impact scores is shown in Figure 3. The dotted trend line connecting the mean values clearly indicates an upward trend. This suggests that the level of negative social media impact generally tends to increase as the average time spent on social media increases.

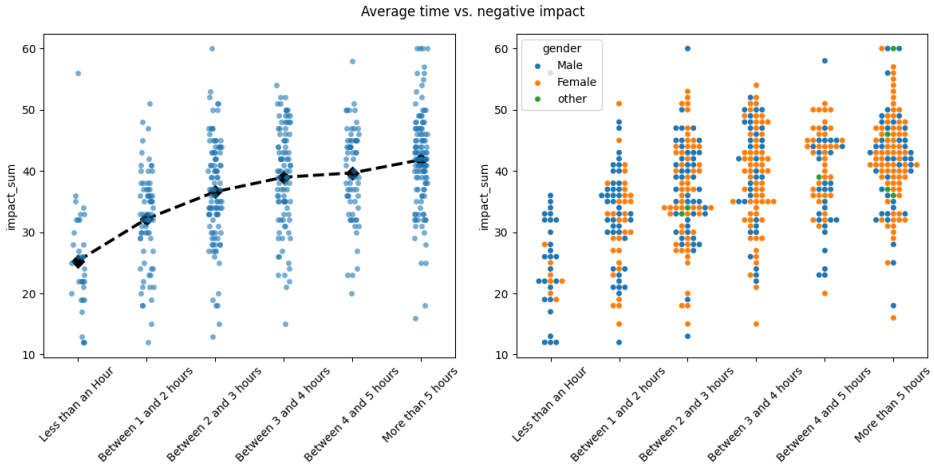

Figure 3. The relation between the average time spent on Social Media and the negative impact score.

Figure 4 is a heatmap of the correlation between features in the dataset, which shows that the feature 'avg_time_spent,' or the average time of social media usage, has a good and significant correlation with the level of depression. In addition, the 'platform_tiktok' feature also shows a good correlation and indicates that high TikTok social media usage is also associated with an increased risk of depression.

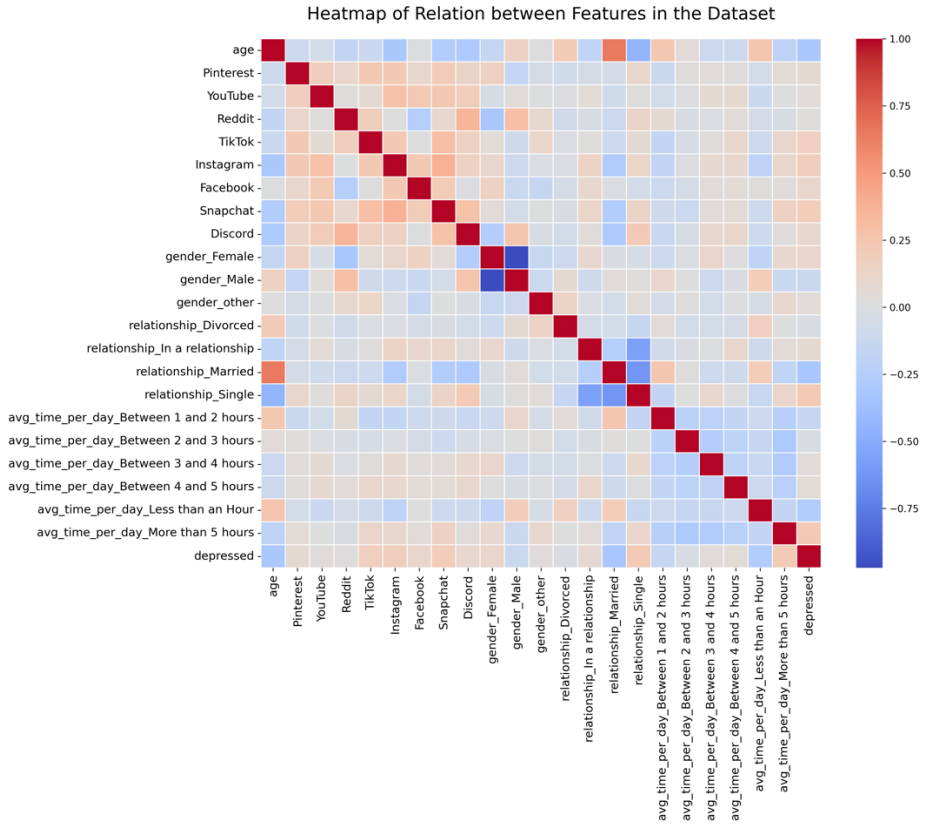

Figure 4. Heatmap of the Relation between Features in the Dataset.

4.2 Results of Dimension Reduction Using Principal Component Analysis (PCA)

The data that has undergone preprocessing is standardized using StandardScaler before proceeding to the machine learning model training stage. StandardScaler is used to transform the data so that it falls within the same scale range. This process scales the data so that it has a mean

of 0 and a standard deviation of 1, which is an optimal prerequisite before performing dimensionality reduction. Next, dimensionality reduction is applied using the Principal Component Analysis (PCA) algorithm.

Through this transformation process, the original data is condensed into 6 principal components. An analysis of variance explained by the components shows that the first through sixth components account for 15.06%, 11.25%, 7.90%, 6.53%, 6.27%, and 5.97% of the variance, respectively. Cumulatively, these six principal components are able to explain and represent 53.01% of the total variance in the original data. Although the data dimensions have been significantly reduced, this 53% variance is proven to sufficiently represent the dominant and most relevant features for the classification process while successfully filtering out irrelevant features. This is evidenced by the model's consistently high final evaluation metrics.

4.3 Model Training and Performance Evaluation

The three Machine Learning models used in this research are Random Forest, XGBoost, and Naïve Bayes. Before testing the models, the data is divided into 80% training data and 20% test data. Based on the evaluation results using accuracy, precision, recall, and f1score metrics, the following results were obtained:

Table 2. Matrix Evaluation Results

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.67	0.67	0.67	0.67
XGBoost	0.90	0.91	0.90	0.90
Naïve Bayes	0.36	0.37	0.36	0.36

Based on the evaluation results, XGBoost shows the best performance results and is far superior to the other two models with 90% accuracy, 91% precision, 90% recall, and 90% F1-score. This shows that XGBoost is considered effective as a boosting algorithm that can handle dataset complexity and avoid overfitting through regularization techniques. The Random Forest model followed with an accuracy rate of 67%, while Naïve Bayes had the lowest performance among the three models used, with an accuracy value of only about 36%. The significant difference in model performance between XGBoost, Random Forest, and Naïve Bayes can be caused by several factors related to the characteristics of each model and how it processes the data.

The superiority of the XGBoost model in this study is not only due to the tuning process alone, but also the result of a combination of the robustness of the XGBoost architecture and hyperparameter optimization. In general, XGBoost has an architectural advantage because it uses a gradient boosting approach that builds sequential decision trees, where each new model focuses on correcting the residuals or prediction errors from the previous model. The XGBoost algorithm is also highly sensitive to changes in hyperparameters, which allows this optimal performance to be achieved. The model successfully found a balance for the features of this dataset through this random combination search. This indicates that XGBoost is considered an effective boosting algorithm capable of handling the complexity of the dataset and avoiding overfitting through regularization techniques.

To ensure the model's validity and minimize data split bias, the testing was expanded using K-fold cross-validation (k=5) and the ROC-AUC metric with a one-vs-rest (OvR) approach. Cross-validation yielded an average accuracy of 35.97% for Random Forest, 32.04% for XGBoost, and 29.97% for Naive Bayes. In addition, evaluation using the ROC-AUC metric showed ROC-AUC scores of 0.9558 for XGBoost and 0.9446 for Random Forest, which were significantly superior to the Naive Bayes model, which only achieved a value of 0.6741.

There is a significant gap between accuracy and the ROC-AUC score. This is to be expected because the accuracy metric requires absolute matching; if a prediction misses the mark on adjacent ordinal classes, it is immediately counted as an error. On the other hand, the ROC-AUC score achieves a very high value because this metric evaluates the model's prediction probabilities

comprehensively, not just the final decision label. This difference between accuracy and ROC-AUC is a common occurrence in modeling ordinal-scale questionnaire data. However, even though the model struggles to predict class boundaries absolutely due to feature overlap, ensemble models such as XGBoost and Random Forest still possess superior probabilistic capabilities in mapping the severity of mental health impacts.

The evaluation results of the XGBoost model visualized using the confusion matrix are shown in Figure 5. From the visualization, it can be seen that the XGBoost model shows a very good prediction performance, with most of the model's prediction results in accordance with the actual results. From the figure, it can be seen that as many as 40 samples predicted as class 0 are in accordance with the actual label. The same can also be seen in other classes, such as second, third, and fourth classes, each of which has the same good prediction accuracy level and even very high, namely 84, 90, and 78. Looking at the results of the confusion matrix shows that the XGBoost model is able to classify with a high level of accuracy and minimal error, which can make this model appropriate for predicting depression in this study.

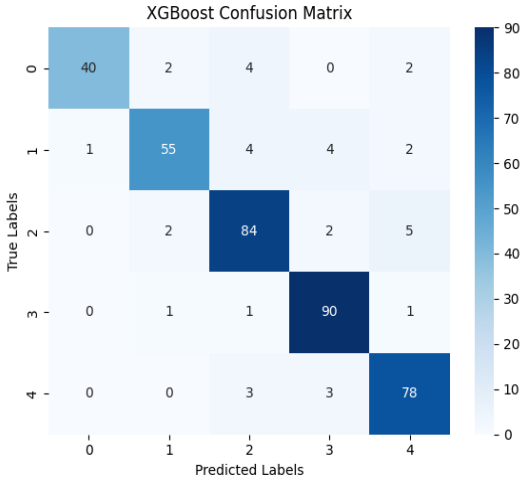

Figure 5. Visualization of Random Forest Model Evaluation Using Confusion Matrix.

The next model used in this study is Random Forest. A comparison of the accuracy of the Naïve Bayes model and the Random Forest method in diagnosing breast cancer shows that Random Forest achieves an accuracy of 94.91%, slightly better than Naïve Bayes, which achieves approximately 93% [19]. The results of the Random Forest model evaluation, visualized using the confusion matrix, are shown in Figure 6. From the visualization, it can be seen that the results of the random forest model evaluation show lower accuracy and a fairly large number of prediction errors. The prediction error is quite spread in each class, such as the second class, which is wrongly predicted as the third class 14 times, and the fourth class, which is wrongly predicted as the second class 15 times. After seeing the confusion matrix results, it shows that the Random Forest model is quite capable of predicting the level of depression in each class, but the results are not as good as the XGBoost model, so it still needs to be improved.

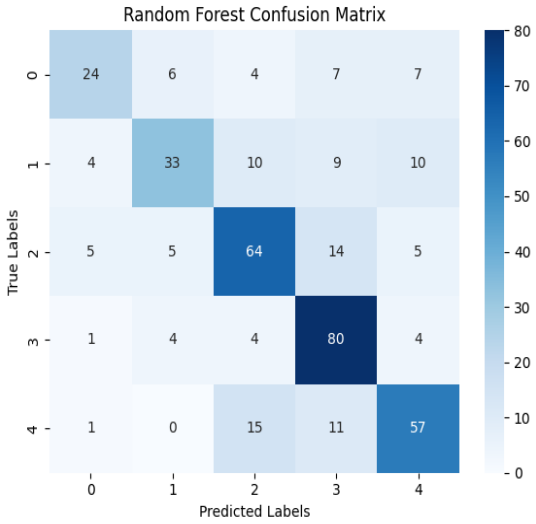

Figure 6. Visualization of Random Forest Model Evaluation Using Confusion Matrix.

The next model used in this research is Naïve Bayes. Naïve Naïve Bayes is not an ensemble algorithm but a probabilistic algorithm that assumes independence between features. A comparative study in [20] shows that although Naïve Bayes performed well with an accuracy of 97.07%, Random Forest performed even better, achieving an accuracy of 99.38%. The Naïve Bayes model evaluation results, visualized using a confusion matrix, are shown in Figure 7. From the visualization, it can be seen that the Naïve Bayes model shows low prediction performance compared to the previous two models, XGBoost and Random Forest. From the figure, it can be seen that out of 48 samples that should be in the zero class, only 22 can be predicted correctly, while the rest are spread across all classes. The same error pattern also occurs in other classes, indicating that this Naïve Bayes model has low accuracy and inaccurate predictions, so this model is less suitable for predicting depression levels in this study.

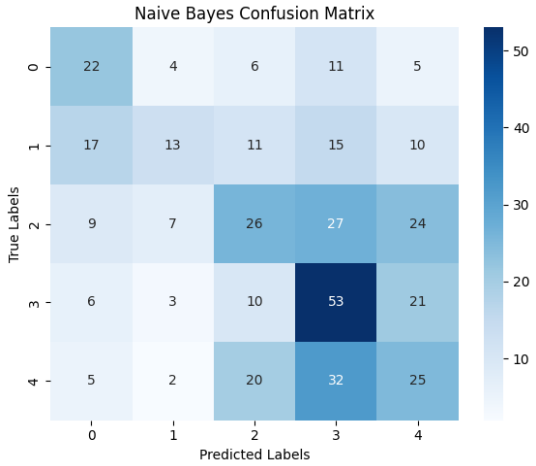

Figure 7. Visualization of Naïve Bayes Model Evaluation Using a Confusion Matrix.

Based on the confusion matrix visualizations of the three models above, a comprehensive analysis of the False Positive (FP) and False Negative (FN) distributions was conducted to analyze the characteristics of prediction errors. Overall, the best model, XGBoost, followed by Random Forest, showed an excellent concentration of correct predictions along the main diagonal, particularly for classes 2 and 3.

Further analysis revealed that critical prediction errors are extremely rare. For example, in the empirical XGBoost metrics, not a single respondent from actual class 4, which has the highest

impact level, was projected to be extremely misclassified into class 0 or class 1, with the extreme FN value being 0. Instead, actual class 0 respondents have a very low FP rate compared to class 4. There are only 3 samples from actual class 0 respondents. This indicates that the model can accurately and logically distinguish between opposing class boundaries.

Despite this, the distribution of FPs and FNs around the main diagonal of the matrix indicates that the model still faces challenges in distinguishing between adjacent ordinal classes. For example, in the case of False Negatives in the XGBoost model, 19 samples from class 4 were incorrectly predicted as class 2 (10 samples) and class 3 (9 samples). On the other hand, False Positive cases also occur, where most Class 1 samples are typically predicted to be Class 2 or 3. In ordinal-scale psychological survey data, where the decision boundaries between adjacent severity levels are very narrow, this error distribution pattern clustering in adjacent class areas indicates the presence of natural feature overlaps.

The comparison of the results of these three algorithms shows that ensemble-based and boosting models are considered more capable of handling the complexity of data patterns related to mental health, especially in detecting symptoms of depression influenced by social media. In addition, these results also show that the model chosen does not only depend on accuracy alone, but also on the model's ability to handle real data that is not always linear and simple.

4.3 Model Interpretation Results with Explainable AI (XAI)

The Explainable AI method using LIME in this research is used to evaluate how the model makes decisions on each individual in the dataset. LIME allows visualization of the influence of features locally, so that the factors that make up each prediction for an individual can be analyzed. Several individual case studies using LIME in this research show that depression predictions are strongly influenced by several factors, such as duration of use, age, and platform preference.

This research attempts to use LIME interpretation in several decision-making models of several case studies. In the case study, a man who is active on social media for more than 5 hours per day has a greater chance of being classified as 'Depressed', but because there are other features that can also compensate, such as the use of social media platforms TikTok, Reddit, and so on, the final result is more inclined to 'Other'. Visualization of how the model makes this decision can be seen in Figure 8.

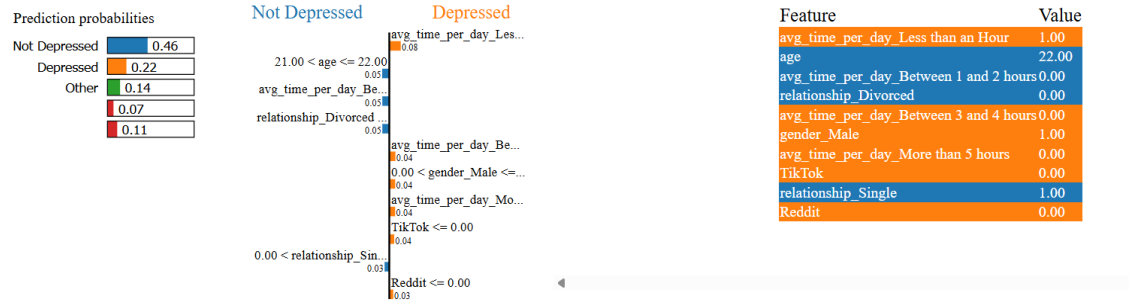

Figure 8. Interpretation Using LIME for the First Case Study.

The next case study is a woman who uses social media for about 1-2 hours per day and never uses social media for less than one hour or more than five hours. The model predicts that the individual is classified as 'Other' with a probability of 0.48. Other probabilities include 'Not Depressed' with 0.05 and 'Depressed' with 0.11. Although this individual only uses social media for about 1-2 hours per day, the presence of other time categories, such as 'Less than an Hour' and 'More than five hours,' in the essential features shows that the similarity of patterns with high-risk groups can still affect the prediction results. The LIME visualization for this individual can be seen in Figure 9.

Figure 9. Interpretation using LIME for the Second Case Study.

The above case studies show that LIME not only improves the transparency of the model's prediction results, but also provides more value in trusting the model in real-life implementations in the field of mental health. Practitioners such as psychologists and counselors can also use these visual explanations for things such as understanding the personal context behind classification results, verifying model results, and tailoring intervention or counseling approaches to each individual's unique circumstances. Explainable AI is thus an important bridge between predictive computing and clinical ethics that makes the results of the model not only accurate but also socially accountable.

4.4 Implications of the Findings

The results in this study reinforce several previous studies regarding the relationship between excessive social media use and psychological disorders. Research by Keles *et al.* [21] showed that intensive social media use has a correlation with an increased risk of depression, anxiety, and psychological stress. Another study in [22] found that limiting the use of social media can significantly reduce symptoms of depression and loneliness. This suggests that interventions that target the reduction of time spent using social media can have a positive impact on individuals' mental health.

Alternatively, this study also emphasizes the importance of integrity and ethical standards in the use of predictive models. It is important to adhere to the principles of privacy, accountability, and transparency, as these systems interact with personal data and psychological aspects of individuals. Therefore, the application of Explainable AI techniques such as LIME is crucial for building trust in the model's prediction process and reassuring both users and experts [23].

Overall, this study highlights the importance of collaboration between the fields of technology, psychology, and public policy. Predictive models using machine learning can not only identify risks with high accuracy and efficiency, but also serve as a foundation for developing decision support systems that can improve mental well-being in society.

4.5 Limitations and Future Research Directions

Even though the study produced encouraging findings and relied on a solid predictive framework, a few clear weaknesses still deserve attention if later projects want to improve. For one thing, almost all the input data came straight from participants' own accounts, so it could easily pick up biases tied to memory lapses or the urge to answer in socially acceptable ways. Because of that, the labels and feature values that feed the model might not reflect reality as closely as hoped.

Looking ahead, the team advises adding time-series or sequential-learning algorithms that can track moods as they ebb and flow over hours or days. Mixing in text, audio, video, and physiological streams will make the resulting models sharper and more tailored to each user. Growing the dataset by including people from varied backgrounds and settings will further test and hopefully confirm that the findings hold up outside the lab. Finally, a careful audit for bias across age, gender, and ethnicity is essential so any tool that reaches the public is honest, just, and answerable to the communities it serves.

5. Conclusions

This study shows that machine-learning tools like XGBoost can spot how social-media habits (daily time spent, age, gender, and chosen platforms) affect mental health. When the team

tested the models, XGBoost reached 90% accuracy, beating both Random Forest and Naive Bayes at flagging depression risk. Adding Explainable AI methods like LIME gave clear, step-by-step reasons for each prediction, so users and clinicians can trust and follow the logic. Together, these findings pave the way for smart, user-friendly apps that monitor mental health in real time while respecting each person's unique pattern.

References

[1] C. Cheng, Y. C. Lau, L. Chan, and J. W. Luk, "Prevalence of social media addiction across 32 nations: Meta-analysis with subgroup analysis of classification schemes and cultural values," *Addict. Behav.,* vol. 117, Art. no. 106845, 2021, doi: 10.1016/j.addbeh.2021.106845.

[2] Y. Sun and Y. Zhang, "A review of theories and models applied in studies of social media addiction and implications for future research," *Addict. Behav.,* vol. 114, Art. no. 106699, 2021, doi: 10.1016/j.addbeh.2020.106699.

[3] M. Schredl, "Continuity Between Waking Life and Dreaming: A Research Note and Study in Adolescents," *Imagination, Cognition and Personality,* vol. 44, no. 1, pp. 104–116, 2024, doi: 10.1177/02762366241254818.

[4] B. Sheaves, S. Rek, and D. Freeman, "Nightmares and psychiatric symptoms: a systematic review of longitudinal, experimental, and clinical trial studies," *Clin. Psychol. Rev.,* vol. 100, Art. no. 102241, 2023, doi: 10.1016/j.cpr.2022.102241.

[5] C. Montag and S. Hegelich, "On the Psychology of TikTok Use: A First Glimpse From Empirical Findings," *Front. Public Health,* vol. 9, Art. no. 641673, 2021, doi: 10.3389/fpubh.2021.641673.

[6] H. C. Woods and H. Scott, "Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem," *J. Adolesc.,* vol. 51, pp. 41–49, 2016, doi: 10.1016/j.adolescence.2016.05.008.

[7] T. F. Kurnaz, C. Erden, U. Dağdeviren, *et al.,* "Comparison of machine learning algorithms for slope stability prediction using an automated machine learning approach," *Nat. Hazards,* vol. 120, no. 8, pp. 6991–7014, 2024, doi: 10.1007/s11069-024-06490-8.

[8] R. Geetha, S. Gunanandhini, G. Umarani Srikanth, and V. Sujatha, "Human Stress Detection in and Through Sleep Patterns Using Machine Learning Algorithms," *J. Inst. Eng. India Ser. B,* vol. 105, no. 6, pp. 1691–1713, 2024, doi: 10.1007/s40031-024-01079-y.

[9] D. Agarwal, V. Singh, A. K. Singh, and P. Madan, "Stacked ensemble model for analyzing mental health disorder from social media data," *Multimedia Tools Appl.,* vol. 83, pp. 53923–53948, 2023, doi: 10.1007/s11042-023-17395-2.

[10] 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, pp. 104–119, 2024, doi: 10.1007/s11920-024-01481-9.

[11] Y. Ibrahimov, T. Anwar, and T. Yuan, "Explainable AI for Mental Disorder Detection via Social Media: A Survey and Outlook," *arXiv preprint arXiv:2406.05984,* 2024.

[12] Z. Bao, A. Pérez, and J. Parapar, "Explainable depression symptom detection in social media," *Health Inf. Sci. Syst.,* vol. 12, Art. no. 47, 2024, doi: 10.1007/s13755-024-00303-9.

[13] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-modalities using a hybrid deep learning model on social media," *World Wide Web,* vol. 25, no. 1, pp. 1–25, 2022, doi: 10.1007/s11280-021-00992-2.

[14] J. Meynadier, J. M. Malouff, N. S. Schutte, N. M. Loi, and M. D. Griffiths, "Relationships Between Social Media Addiction, Social Media Use Metacognitions, Depression, Anxiety, Fear of Missing Out, Loneliness, and Mindfulness," *Int. J. Ment. Health Addict.,* 2025, doi: 10.1007/s11469-024-01440-8.

[15] T. Xiao, M. Pan, X. Xiao, and Y. Liu, "The Relationship Between Physical Activity and Sleep Disorders in Adolescents: A Chain-Mediated Model of Anxiety and Mobile Phone Dependence," *BMC Psychol.,* vol. 12, Art. no. 751, 2024, doi: 10.1186/s40359-024-02237-z.

[16] T. S. Alshammari, "Applying Machine Learning Algorithms for the Classification of Sleep Disorders," *IEEE Access,* vol. 12, pp. 36110–36125, 2024, doi: 10.1109/ACCESS.2024.3374408.

[17] N. C. Ramadhan, H. Hikmayanti, T. Rohana, and A. M. Siregar, "Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur XGBoost Untuk Klasifikasi Kanker Payudara," *TIN: Terapan Informatika Nusantara,* vol. 5, no. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408

[18] Vincent and N. Rachmat, "Penerapan Algoritma Gradient Boosting dalam Mendiagnosa Penyakit Kucing dan Anjing," *Jurnal Buana Informatika,* vol. 16, no. 2, pp. 134–143, Oct. 2025.

[19] N. Rohmah, E. A. Safitri, C. Alinta, Y. Oktalina, and W. Setiawan, "Perbandingan Akurasi Metode Naive Bayes dan Metode Random Forest dalam Mendiagnostik Penyakit Kanker Payudara," *DoubleClick: J. Comput. Inf. Technol.,* vol. 8, no. 2, pp. 109–118, 2025, doi: 10.25273/doubleclick.v8i2.20383.

[20] D. Kurniasari, R. N. Hidayah, N. Notiragayu, W. Warsono, and R. K. Nisa, "Classification Models for Academic Performance: A Comparative Study of Naïve Bayes and Random Forest Algorithms in Analyzing University of Lampung Student Grades," *Jurnal Teknik Informatika (JUTIF),* vol. 5, no. 5, pp. 1267–1276, Oct. 2024, doi: 10.52436/1.jutif.2024.5.5.2066.

[21] B. Keles, N. McCrae, and A. Grealish, "A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents," *Int. J. Adolesc. Youth,* vol. 25, no. 1, pp. 79–93, 2020, doi: 10.1080/02673843.2019.1590851.

[22] M. G. Hunt, R. Marx, C. Lipson, and J. Young, "No more FOMO: Limiting social media decreases loneliness and depression," *J. Soc. Clin. Psychol.,* vol. 37, no. 10, pp. 751-768, 2018, doi: 10.1521/jscp.2018.37.10.751

[23] J. M. Twenge, T. E. Joiner, M. L. Rogers, and G. N. Martin, "Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time," *Clin. Psychol. Sci.,* vol. 6, no. 1, pp. 3–17, 2018, doi: 10.1177/2167702617723376.