Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

Explainable Ensemble Learning for Depression Risk Classification Using Multidomain Behavioral Features

Erfian Junianto[1], Siti Nurkhodijah*[2]

[1]Informatics, Universitas Adhirajasa Reswara Sanjaya, Indonesia
[2]Information Systems, Universitas Adhirajasa Reswara Sanjaya, Indonesia

Email: [2]sitinurkhodijah16@gmail.com

Received : Jul 2, 2025; Revised : Aug 12, 2025; Accepted : Aug 13, 2025; Published : Apr 15, 2026

Abstract

Depression is a growing global health concern, particularly among adolescents and university students. Despite the availability of standardized assessments, delays in early detection remain a major barrier to effective treatment. Digital behavioral data holds considerable potential for mental health assessment, but its utilization remains limited due to the absence of integrated and interpretable computational models. This study presents an interpretable machine learning framework for classifying depression risk using multi-domain *behavioral features* extracted from simulated digital life datasets. Three public datasets were integrated and mapped to five psychological clusters based on *DSM-5* criteria: self-regulation, negative affect, cognitive strain, comparison and avoidance, and sleep disturbance. Two ensemble classifiers, Random Forest and XGBoost, were applied and evaluated using 10-fold stratified cross-validation. Depression risk was categorized into three levels: Low, Medium, and High. The Random Forest model achieved the highest accuracy (81%) and macro-averaged F1-score (0.81), showing strong performance especially in identifying transitional Medium-risk users. To enhance transparency, both global and local model interpretations were performed using *SHapley Additive exPlanations (SHAP)*. Results revealed that digital stressors such as excessive screen time and disrupted sleep patterns were prominent in high-risk classifications, while mood stability and mindfulness were protective factors in low-risk groups. The proposed framework offers a scalable and explainable for early depression screening by integrating psychological theory with artificial intelligence methods. The findings contribute to the field of behavioral informatics by demonstrating the practical value of interpretable models in enhancing the reliability, transparency, and applicability of digital mental health systems and personalized behavioral monitoring.

Keywords: *Behavioral Features, Depression Screening, Explainable Machine Learning, Mental Health, SHAP, XGBoost*

1. INTRODUCTION

Depression has emerged as a major global mental health concern, affecting over 280 million people and ranking among the leading causes of disability worldwide [1]. In Indonesia, mental health issues are increasingly prevalent, with over 500,000 depression cases reported in 2022, predominantly among adolescents and university students [2]. Delays in screening and diagnosis often exacerbate symptoms and reduce treatment efficacy. This situation underscores the need for early, proactive, and scalable mental health risk detection strategies.

With the growing integration of digital technologies in daily life, digital behavioral patterns such as excessive screen time, social media usage, sleep disruption, physical inactivity, and impulsive browsing are increasingly linked to psychological distress [3], [4]. These patterns can be captured passively through device usage logs or behavioral surveys, making them attractive data sources for machine learning (ML) models aimed at predicting mental health risk levels. However, most prior

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

research either focused on single-domain features or applied black-box models lacking transparency, raising challenges in real-world clinical or public health applications [5].

Depression is a multidimensional condition influenced by overlapping factors such as cognitive fatigue, negative affect, low self-control, and disrupted circadian rhythms [6], [7]. This study draws from the Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (DSM-5) classification to group symptoms into five psychological clusters: self-regulation, negative affect, cognitive strain, comparison and avoidance, and sleep disturbance [8]. Accurately modeling this complexity requires not only integrating features across domains but also ensuring interpretability to enable trust and actionable insights from the predictions, as outlined in several recent works summarized in Table 1 Related Studies.

Table 1. Comparison of Related Studies

No	Authors (Year)	Data & Modality	Method or Model	Findings & Limitations
1	Asare et al. (2021) [9]	Smartphone sensors (sleep, screen time)	RF, SHAP, hyperparameter tuning	Accuracy 85%; sleep duration & social interaction most influential.
2	Imans et al. (2024) [10]	Behavioral & demographic data	Gradient Boosting, RF, SHAP	Accuracy >90%; interpretable via SHAP; focused on severity.
3	Amirhosseini et al. (2024) [11]	Text, metadata, user activity (multimodal)	RF, SVM	RF outperformed; emphasized personalization of prediction.
4	Liu & Shi (2022) [12]	Social media posts	Chi-Square, Info Gain, RF	Affective & temporal features are key; ensemble performed best.
5	Li & Xiao (2025) [13]	Text + behavioral features from social media	Transformer w/ Cross-Attention	Accuracy 94.95%, F1-score 94.69%; integrated multi-source signals.

While previous research in digital mental health has applied machine learning techniques, many have been limited to single-domain features or relied on black-box models with limited interpretability. This study addresses these limitations by proposing a multi-domain behavioral fusion framework integrated with explainable ensemble learning. Leveraging DSM-5-based clustering and SHAP interpretation, the research advances behavioral informatics through the development of transparent and Interpretable AI systems for mental health risk screening.

This study proposes a classification framework that fuses multi-domain digital behavioral features and maps them into psychological clusters aligned with DSM-5 depression indicators. Using three public-simulated datasets, features are engineered, merged, and labeled into three mental health categories: non-depressed, mildly depressed, and severely depressed. Random Forest and XGBoost are employed for classification, with SHAP used to generate global and local explanations of feature contributions.

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

The contributions of this research are threefold: (1) constructing a structured digital behavior fusion dataset aligned with DSM-5-based clusters, (2) building interpretable ensemble classifiers for depression risk classification, and (3) visualizing feature impact using SHAP to support transparent model understanding. This framework presents a scalable, explainable, and domain-aware approach to mental health screening using behavioral data, bridging psychological theory with AI for real-world impact.

2. METHOD

This study adopts a structured experimental framework to classify depression risk based on behavioral indicators captured from digital life. The approach integrates multiple simulated datasets, applies feature-level fusion and psychological clustering based on DSM-5, and employs explainable machine learning (Random Forest and XGBoost with SHAP) for both prediction and interpretation.

The end-to-end process consists of five primary phases: data collection and merging, pre-processing and labelling, model training with 10-fold stratified validation, performance evaluation, and SHAP-based feature attribution. The complete framework which is illustrated in the research workflow shown in Figure 1.

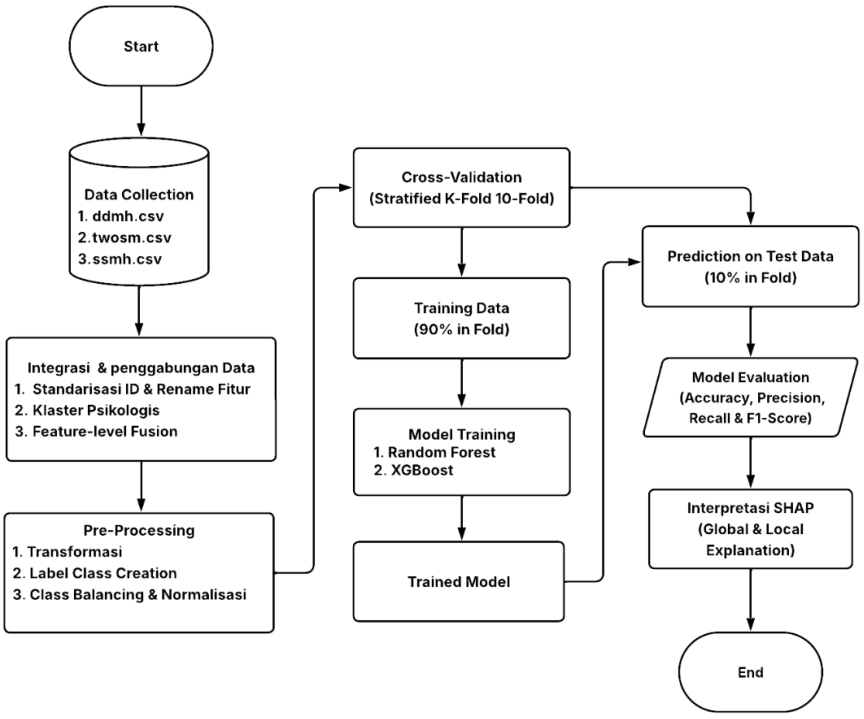

Figure 1. Research framework

1.1. Data Collection

Three publicly available simulated datasets were sourced from Kaggle to represent diverse dimensions of digital behavior associated with mental health. These datasets were selected to support the goal of building a multi-domain behavioral classification framework for depression risk.

The datasets are:

a. Impact of Screen Time on Mental Health: Captures behavioral variables such as daily screen exposure, sleep quality, stress, and mood ratings [14].

b. Social Media Menace: Focuses on impulsive social media use, self-control, and digital addiction behaviors [15].

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

c. Social Media and Mental Health: Includes survey-based data containing demographic information, PHQ-9 depression scores, and behavioral self-assessments [16].

Combined, these datasets provide complementary insights into individual behavioral tendencies that may indicate mental health risk.

1.2. Data Integration

To construct a unified behavioral matrix, a feature-level fusion strategy was carried out in four main stages:

a. Data Standardization: Standardized user identifiers and harmonized feature names across all datasets to ensure consistency and clarity.
b. Psychological Feature Mapping: Categorized behavioral indicators into five psychological clusters based on DSM-5 such as self-regulation, negative affect, cognitive strain, comparison and avoidance, and sleep disturbance [8].
c. Feature-Level Fusion: Integrated features across datasets using the standardized user ID as a key to align individual records.

Generated a final behavioral dataset comprising 24 attributes that represent multidomain patterns associated with depression risk. This integrated matrix served as the foundation for interpretable and comprehensive machine learning modeling. This fused dataset enabled comprehensive and interpretable modeling of depression risk using machine learning.

1.3. Preprocessing

The preprocessing phase involved several key steps to prepare the dataset for model training. First, categorical features were transformed into numerical format using label encoding to ensure compatibility with tree-based algorithms. Then, class labels for depression risk were constructed based on PHQ-9 scores and categorized into Low, Medium, and High risk levels [17], [18]. Due to initial class imbalance, an oversampling technique was applied to equalize the distribution across all three classes. Finally, all numerical features were normalized using Z-score standardization to stabilize variance and eliminate scale bias during model training.

1.4. Model Development and Validation

The selection of Random Forest and XGBoost was based on their reliability in handling multidimensional structured data and strong performance in previous studies [19], [20], which aligns with the objectives of this study.

Two ensemble-based classifiers were selected:

a. Random Forest (RF): Uses bagging to construct multiple decision trees; robust to overfitting [21], [22].
b. XGBoost (Extreme Gradient Boosting): A boosting-based model that introduces regularization and weighted updates for performance improvement [23], [24].

Model development was implemented in Python using Scikit-Learn and XGBoost within Google Colaboratory. The dataset was evaluated using Stratified K-Fold Cross Validation (K=10), which maintains label proportions across folds [25]. In each fold, 90% of data was used for training and 10% for testing.

1.5. Performance Metrics

To evaluate the effectiveness of the classification models, this study employed four standard performance metrics:

a. Accuracy: The ratio of correct predictions to the total number of predictions.

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

b. Precision: The proportion of true positives among all predicted positives, reflecting model exactness.
c. Recall: The proportion of true positives among all actual positives, indicating model sensitivity.
d. F1-Score: The harmonic mean of precision and recall, providing a balanced measure between the two.

All metrics were calculated using macro-averaging to ensure fair evaluation across the three depression risk classes (Low, Medium, High). Additionally, confusion matrices were constructed for each fold to capture detailed misclassification patterns and support class-level performance analysis [26].

1.6. Explainability Using SHAP

To enhance interpretability, SHAP (SHapley Additive exPlanations) was employed [5], [27]. This method assigns a contribution value to each feature in a prediction based on cooperative game theory (Shapley values). It provides both:

a. Global explanation: Summary plots to visualize average importance per feature across the dataset.
b. Local explanation: Individual bar plots to identify dominant factors influencing single-user predictions.

This dual-level interpretability is particularly essential in mental health applications, where trust, accountability, and actionable insights are critical for ethically deploying AI-driven risk assessments (Explainable artificial intelligence for mental health through transparency and interpretability for understandability).

3. RESULT

This section presents the outcomes of the experimental workflow, including classification performance metrics, confusion matrix analysis, and model explainability using SHAP. Two ensemble models, Random Forest and XGBoost, were trained and evaluated to categorize users into three levels of depression risk. The performance of each model is quantitatively assessed through accuracy, precision, recall, and F1-score. Additionally, to ensure interpretability and transparency in prediction, both global and local SHAP analyses were conducted to examine how different behavioral features influence model decisions at population and individual levels.

3.1. Data Collection

This study utilized three publicly available simulated datasets from Kaggle, each representing distinct dimensions of digital behavior relevant to mental health analysis:

a. Impact of Screen Time on Mental Health (ddmh.csv): Comprising 2,000 entries and 25 attributes, this dataset captures behavioral patterns related to screen time, sleep quality, productivity, and stress. It explores correlations between screen exposure and psychological symptoms such as anxiety and stress.
b. Social Media Menace (twosm.csv): Containing 1,000 entries and 31 attributes, this dataset focuses on maladaptive social media behaviors, including impulsivity, low self-control, and digital addiction. It was designed to evaluate the psychological impact of excessive social media use.
c. Social Media and Mental Health (ssmh.csv): This survey-based dataset consists of 481 entries and 21 attributes, including demographic data, social media usage patterns, and psychological

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

self-assessment metrics. It includes PHQ-9 scores, serving as a standardized indicator of depression and anxiety symptoms.

3.2. Data Integration

A feature-level fusion strategy was employed to integrate the three source datasets into a unified behavioral matrix consisting of 24 attributes associated with depression risk. This fused dataset serves as the foundation for interpretable and multidomain machine learning modeling.

The features span key behavioral domains including screen time, self-control, emotional regulation, sleep quality, attention span, and social comparison capturing the complex relationship between digital behavior and mental health.

All variables were standardized, renamed, and harmonized across datasets, resulting in a structured mix of numerical features (e.g., scroll_rate, mood_rating, sleep_duration_hours) and categorical indicators (e.g., self_control, sleep_quality, validation_seeking) suitable for tree-based ensemble training. A detailed list of features, including their data types and original sources, is presented in Table 2.

Table 2. List of Features in the Combined Dataset

Dataset Source	Features	Data Type
Impact of Screen Time on Mental Health	user_id	Nominal (Unique ID)
	daily_screen_time_hours	Numerical (hours/day)
	social_media_hours	Numerical (hours/day)
	sleep_quality	Categorical (good/poor/average)
	stress_level	Numerical (scale 1–10)
	weekly_depression_score	Numerical (PHQ-9 scale)
	mood_rating	Numerical (scale 1–10)
	sleep_duration_hours	Numerical (hours/night)
	mindfulness_minutes_per_day	Numerical (minutes/day)
	physical_activity_hours_per_week	Numerical (hours/week)
Social Media Menace	total_screen_time	Numerical (hours/day)
	scroll_rate	Numerical (scrolls/minute)
	addiction	Categorical (low/medium/high)
	self_control	Categorical (low/medium/high)
	productivity_loss	Numerical (hours/day)
Social Media and Mental Health	distracted	Categorical (yes/no)
	restless	Categorical (yes/no)
	easily_distracted	Categorical (yes/no)
	worries	Categorical (yes/no)
	concentration_difficulty	Categorical (yes/no)
	comparison_freq	Numerical (frequency scale)
	comparison_feeling	Categorical (positive/neutral/negative)
	validation_seeking	Categorical (yes/no)
	feeling_down	Categorical (yes/no)
	sleep_issues	Categorical (yes/no)
	interest_fluctuation	Categorical (yes/no)

3.3. Preprocessing

With the dataset fully preprocessed, including categorical encoding, class creation, class balancing, and normalization, the data was ready for model training. These preprocessing steps ensured

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

that the input matrix was both statistically stable and structurally compatible with ensemble learning algorithms. The next phase focused on evaluating the classification performance of the selected models using stratified cross-validation and standard evaluation metrics.

3.3.1. Feature Transformation

Categorical features were transformed into numerical format using label encoding, ensuring model compatibility while preserving ordinal semantics. Table 3 shows the mapping from original categories to encoded values.

Table 3. Categorical Feature Transformation Results

Features	Category Values (Before Encoding)	Encoded Values (After Encoding)
sleep_quality	poor, average, good	0,1,2
addiction	low, medium, high	0,1,2
self_control	low, medium, high	0,1,2
distracted	no, yes	0, 1
restless	no, yes	0, 1
easily_distracted	no, yes	0, 1
worries	no, yes	0, 1
concentration_difficulty	no, yes	0, 1
comparison_feeling	negative, neutral, positive	0,1,2
validation_seeking	no, yes	0, 1
feeling_down	no, yes	0, 1
sleep_issues	no, yes	0, 1
interest_fluctuation	no, yes	0, 1

3.3.2. Class Labeling and Balancing

The target variable (depression_risk) was derived from PHQ-9 scores using APA clinical thresholds and categorized into Low, Medium, and High risk classes. However, initial label distribution was imbalanced, potentially introducing bias in model training. To address this, oversampling was employed to equalize the number of observations across all classes. The distribution before and after balancing is presented in Table 4.

Table 4. Class Distribution Before and After Balancing

Depression Risk	Number of Observations (Before Balancing)	Number of Observations (After Balancing)
Low	92	92
Medium	51	92
High	34	92

3.3.3. Normalization

The The final preprocessing step involved normalizing all numerical features using Z-score standardization, which transforms each variable to have a mean of 0 and a standard deviation of 1. Although tree-based algorithms are generally robust to feature scaling, normalization was applied to reduce the influence of features with large numerical ranges and ensure consistent model behavior. An example of raw vs normalized data is shown in Table 5.

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

Table 5. Sample Data Before and After Normalization

user_id	Feature	Raw Value	Normalized Value
user_305	total_screen_time	5.3	0.19
user_305	scroll_rate	21.7	-1.35
user_305	self_control	2	0.49
user_305	sleep_quality	2	1.24

3.4. Classification Performance

To evaluate the effectiveness of the proposed framework, two ensemble-based classifiers, Random Forest (RF) and XGBoost (XGB), were trained and validated using 10-fold Stratified Cross-Validation, ensuring consistent class distribution across each fold. Model performance was assessed using four standard evaluation metrics: Accuracy, Precision, Recall, and F1-Score, all computed using macro-averaging to accommodate multi-class classification. The classification results are presented in Table 6 across three depression risk levels Low, Medium, and High.

Table 6. Classification Metrics for Random Forest and XGBoost

Class	Metric	Random Forest	XGBoost
Low	Precision	0.77	0.77
	Recall	0.80	0.78
	F1-Score	0.78	0.78
Medium	Precision	0.86	0.84
	Recall	0.80	0.78
	F1-Score	0.83	0.81
High	Precision	0.81	0.79
	Recall	0.83	0.83
	F1-Score	0.82	0.81
Overall	Accuracy	**0.81**	0.80
	Macro Avg	**0.81**	0.80

The Random Forest model slightly outperformed XGBoost, particularly in the Medium risk class, which is often more difficult to distinguish due to its overlap with both Low and High categories. This performance distinction suggests RF's higher robustness in identifying transitional mental states.

3.4.1. Confusion Matrix Analysis

To further analyze misclassification patterns, a global confusion matrix was constructed by aggregating predictions from all 10 folds. The results are shown in Table 7.

Table 7. Confusion Matrix for Random Forest and XGBoost

Model	Class	Correct	Misclassified (To)
Random Forest	Low	144	Medium (16), High (21)
	Medium	145	Low (21), High (15)
	High	151	Low (22), Medium (8)
XGBoost	Low	142	Medium (15), High (24)
	Medium	142	Low (23), High (16)
	High	150	Low (19), Medium (12)

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

3.4.2. Global Feature Attribution with SHAP

To understand the impact of each feature on the model's decision-making process, SHAP summary plots were generated for each risk class (Low, Medium, High). These plots illustrate the mean absolute SHAP value of each feature, ranked by their overall contribution to predictions.

In the Low Risk class, the SHAP summary plot (Figure 2) the most impactful features were mood_rating, mindfulness_minutes_per_day, and self_control, all contributing negatively to the model's output score as shown in Figure 2. This indicates that users who maintain a positive emotional state, engage regularly in mindfulness practices, and exhibit strong self-regulation are consistently classified with low depression risk. These findings align with established protective psychological factors such as emotional stability, self-awareness, and behavioral discipline [5], [28].

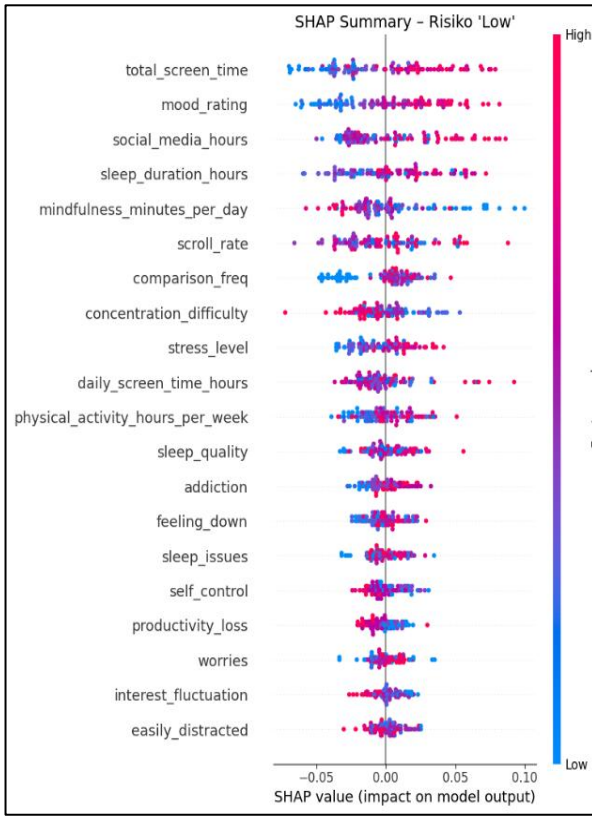

Figure 2. Global Summary Plot – Low Risk Depression Class (Random Forest)

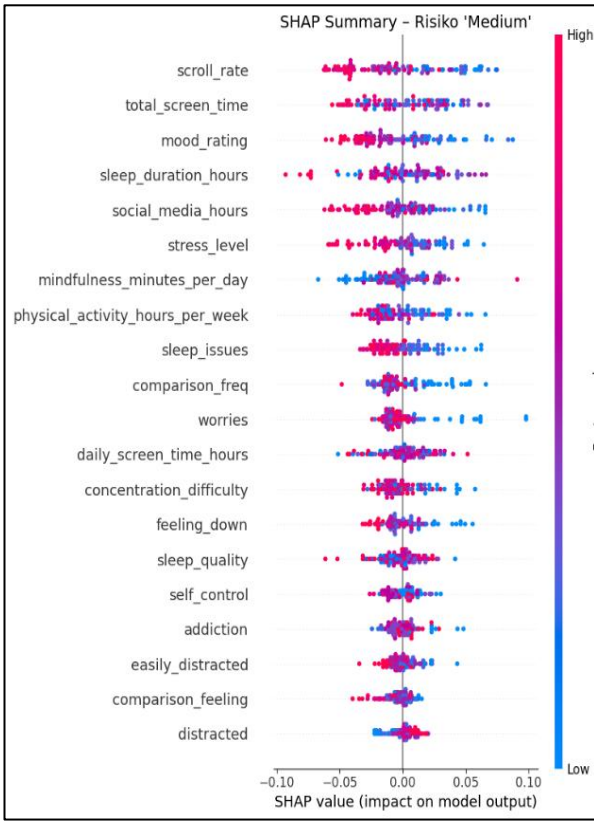

Figure 3. Global Summary Plot – Medium Risk Depression Class (Random Forest)

In the Medium Risk class, the SHAP summary plot highlights sleep_issues, social_media_hours, and total_screen_time, all of which contributed positively to the predicted risk. These attributes reflect the early signs of digital fatigue and mild cognitive strain that characterize moderate depressive tendencies. Conversely, features like physical_activity_hours_per_week exerted a negative SHAP influence, suggesting that regular physical activity may help buffer stress-related impacts [29]. These findings are illustrated in Figure 3.

In the High Risk category, the SHAP summary plot identifies scroll_rate, total_ screen_time, and sleep_duration_hours displayed strong positive SHAP values, identifying them as critical contributors to high-risk classification. Other significant drivers included comparison_freq and worries, reinforcing the psychological patterns of emotional instability, disrupted sleep, and compulsive digital engagement that are often associated with severe depression [30]. These findings are illustrated in Figure 4.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

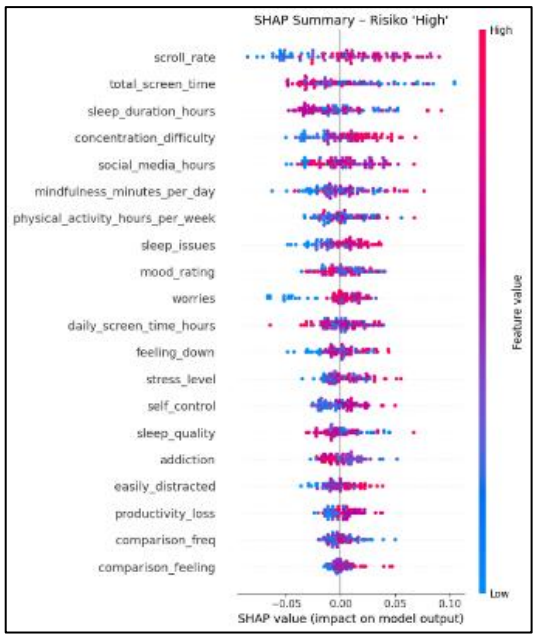

Figure 4. Global Summary Plot – High Risk Depression Class (Random Forest)

3.5. Local SHAP Analysis (Individual-Level)

In addition to global attribution, individual SHAP bar plots were generated to examine personalized feature influence for one representative user per class. These plots show how individual feature values positively or negatively affect the predicted outcome.

In the Low Risk case, the SHAP bar plot shows that mood_rating, mindfulness_ minutes_per_day, and self_control had the most significant negative SHAP values, meaning they reduced the model's output probability toward the depressive classes. These insights align with recent evidence showing that trait mindfulness and self-regulation enhance emotional well-being and reduce susceptibility to depressive symptoms, thereby reinforcing the observed negative SHAP contributions of these features in low-risk subjects [31]. These findings are illustrated in Figure 5.

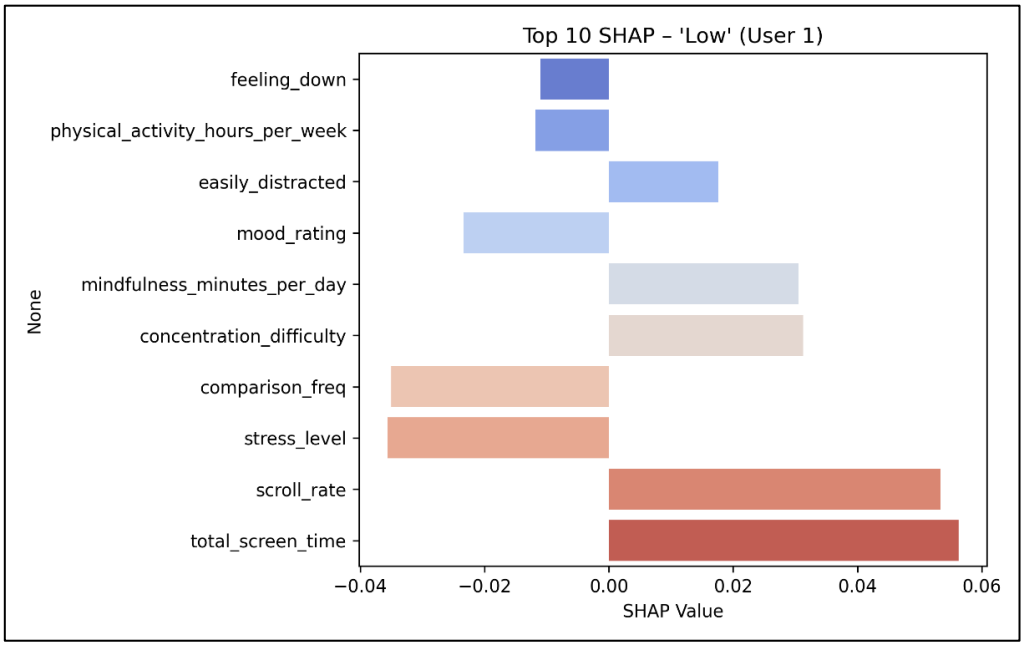

Figure 5. Local Explanation Bar Plot - Low Risk Depression Class (Random Forest)

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

For the Medium Risk user, the SHAP bar plot reveals that sleep_issues, social_media_hours, and total_screen_time contributed positively to the predicted risk, while mindfulness_minutes_per_day and physical_activity_hours_per_week exerted modest negative influence. This mixed contribution reflects a transitional risk profile, where protective and risk features co-occur, resulting in a moderate classification. Prior work using SHAP shows similar patterns: sleep disruption and diminished physical activity are among the strongest contributors to elevated mental health risk, supporting the role of these behavioral features in intermediate risk stratification [32]. These findings are illustrated in Figure 6.

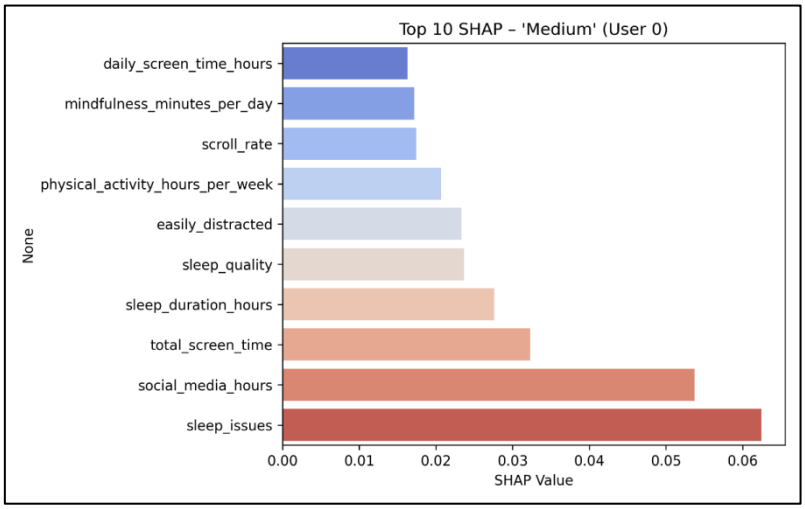

Figure 6. Local Explanation Bar Plot – Medium Risk Depression Class (Random Forest)

In the High Risk case, the SHAP bar plot highlights scroll_rate, sleep_duration_hours, and comparison_freq as strong positive contributors to risk prediction. These features align with previous findings that extended screen engagement and disrupted sleep patterns are primary behavioral signals associated with elevated depression risk, while social comparison further intensifies this signal. The minimal SHAP contribution from self-regulatory indicators corresponds to a diminished role of internal protective factors in high-risk individuals. This interpretation is supported by recent studies demonstrating that sleep duration and variability are dominant predictors for depression-related outcomes [33] and that screen-derived behavioral markers such as device usage patterns serve as critical risk indicators in mental health sensing models [34]. These findings are illustrated in Figure 7.

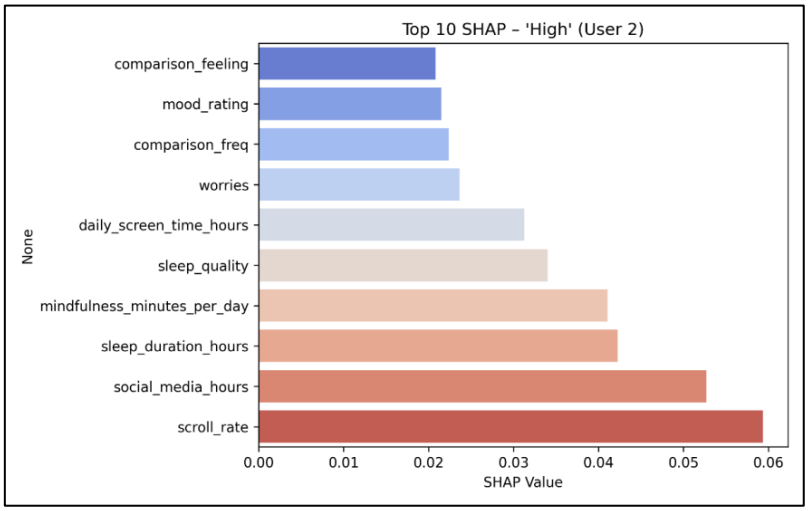

Figure 7. Local Explanation Bar Plot – High Risk Depression Class (Random Forest)

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

These localized SHAP explanations demonstrate the model's capacity to deliver not only accurate but also interpretable predictions tailored to individual users. This level of explainability is crucial for the deployment of AI in sensitive domains such as mental health monitoring, where trust, transparency, and actionable feedback are essential [35].

4. DISCUSSIONS

The results demonstrate the effectiveness of feature-level fusion in integrating multidomain digital behaviors such as screen time, emotional regulation, and sleep patterns into a cohesive framework for depression risk classification. Among the two ensemble models evaluated, Random Forest (RF) showed the best performance, achieving 81% accuracy and macro-averaged F1-score. It was particularly effective in classifying Medium-risk users, a group that is often difficult to distinguish due to overlapping behavioral traits.

These findings validate the hypothesis that passive digital indicators like scroll rate, sleep issues, and total screen time are highly predictive of mental health status. SHAP-based interpretation confirmed this, showing that mood rating, mindfulness duration, and self-control were dominant protective factors in the Low-risk group. Conversely, features such as sleep duration, comparison frequency, and compulsive scrolling were strong contributors to High-risk classifications. These patterns align with existing literature on digital fatigue, poor sleep quality, and emotional dysregulation.

The model's behavior is consistent with previous research by Asare et al. [9] and Imans et al. [10], who emphasized the importance of interpretable ensemble models in mental health prediction. SHAP not only enhances transparency but also provides actionable insights that can support personalized digital interventions.

This study contributes to behavioral informatics by demonstrating how explainable AI can be applied to psychological risk modeling. The integration of DSM-5-based psychological clustering, feature-level fusion, and interpretable ensemble methods offers a reproducible and scalable framework for ethical algorithmic decision-making in mental health contexts.

Despite its strengths, the study has limitations. The use of simulated datasets may not fully capture the complexity of real-world behavior, and SHAP only reveals correlations, not causation. Future research should validate this approach using clinical or real-world data, with additional inputs such as biometrics, self-reported mood journals, and longitudinal tracking to improve robustness and generalizability.

5. CONCLUSION

This study presents an interpretable machine learning framework for classifying depression risk based on multidomain digital behavior. By fusing behavioral features related to screen time, sleep quality, social media usage, and emotional regulation, the model successfully categorized users into three levels of depression risk: Low, Medium, and High. Among the two classifiers evaluated, Random Forest achieved the highest performance, particularly in identifying Medium risk users. The integration of SHAP for global and local interpretability revealed that features such as mood_rating, self_control, and mindfulness_minutes_per_day were strong protective indicators, while scroll_rate, sleep_duration_hours, and comparison_freq were key contributors to elevated risk.

These findings underscore the feasibility of using digital behavioral data for early mental health screening and highlight the importance of explainability in sensitive domains like psychological diagnostics. From a computer science perspective, this research contributes to the advancement of interpretable artificial intelligence (XAI) by applying explainable ensemble learning to real-world psychological modeling. It demonstrates a practical and ethical framework for behavioral informatics, supporting decision transparency in AI-driven health technologies.

Future work should explore the integration of this framework into mobile applications or digital platforms for real-time mental health monitoring. In addition, investigating alternative explainability methods (e.g., LIME, counterfactuals), expanding the behavioral feature set, or incorporating temporal modeling with sequence-based architectures such as LSTM may further improve both prediction accuracy and interpretability in real-world deployment.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

Acknowledgement is only addressed to funders or donors and object of research. Acknowledgement can also be expressed to those who helped carry out the research.

REFERENCES

[1] World Health Organization, "Depressive Disorder (Depression)," 2023. Accessed: Apr. 10, 2025. [Online]. Available: Https://Www.Who.Int/News-Room/Fact-Sheets/Detail/Depression

[2] Kemenkes, "Profil Kesehatan Indonesia 2023." Accessed: Apr. 10, 2025. [Online]. Available: Https://Kemkes.Go.Id/Id/Profil-Kesehatan-Indonesia-2023

[3] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications And Research Directions," *Sn Comput. Sci.*, Vol. 2, No. 3, P. 160, 2021, Doi: 10.1007/S42979-021-00592-X.

[4] H. Byeon, "Advances In Machine Learning And Explainable Artificial Intelligence For Depression Prediction," *Int. J. Adv. Comput. Sci. Appl.*, Vol. 14, Pp. 520–526, Jul. 2023, Doi: 10.14569/Ijacsa.2023.0140656.

[5] S. M. Lundberg And S.-I. Lee, "A Unified Approach To Interpreting Model Predictions," In *Advances In Neural Information Processing Systems*, 2017, Pp. 4765–4774. Doi: 10.48550/Arxiv.1705.07874.

[6] D. C. Mohr, M. Zhang, And S. M. Schueller, "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors And Machine Learning," *Annu. Rev. Clin. Psychol.*, Vol. 13, No. Volume 13, 2017, Pp. 23–47, 2017, Doi: 10.1146/Annurev-Clinpsy-032816-044949.

[7] C. S. Andreassen, "Online Social Network Site Addiction: A Comprehensive Review," *Curr. Addict. Reports*, Vol. 2, No. 2, Pp. 175–184, 2015, Doi: 10.1007/S40429-015-0056-9.

[8] M. B. First, *Dsm-5-Tr®Handbook Of Differential Diagnosis*. Washington, Dc: American Psychiatric Publishing, 2024.

[9] K. Opoku Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, And D. Ferreira, "Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, And Feature Importance Analysis: Exploratory Study," *Jmir Mhealth Uhealth*, Vol. 9, No. 7, P. E26540, Jul. 2021, Doi: 10.2196/26540.

[10] 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, 2024, Doi: 10.3390/Diagnostics14212385.

[11] M. H. Amirhosseini, A. L. Ayodele, And A. Karami, "Prediction Of Depression Severity And Personalised Risk Factors Using Machine Learning On Multimodal Data," In *2024 Ieee 12th International Conference On Intelligent Systems (Is)*, 2024, Pp. 1–7. Doi: 10.1109/Is61756.2024.10705185.

[12] J. Liu And M. Shi, "A Hybrid Feature Selection And Ensemble Approach To Identify Depressed Users In Online Social Media," *Front. Psychol.*, Vol. Volume 12, 2022, Doi: 10.3389/Fpsyg.2021.802821.

[13] S. Li, Y. Xiao, And S. Hu, "A Depression Detection Method Based On Multi-Modal Feature Fusion Using Cross-Attention," In *2025 8th International Conference On Advanced Algorithms And Control Engineering (Icaace)*, 2025, Pp. 1825–1831. Doi: 10.1109/Icaace65325.2025.11019096.

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

[14] Khushi Yadav, "Impact Of Screen Time On Mental Health," Kaggle. Accessed: May 02, 2025. [Online]. Available: Https://Www.Kaggle.Com/Datasets/Khushikyad001/Impact-Of-Screen-Time-On-Mental-Health
[15] Shahzad Aslam, "Social Media Menace," Kaggle. Accessed: May 02, 2025. [Online]. Available: Https://Www.Kaggle.Com/Datasets/Zeesolver/Dark-Web
[16] Souvik Ahmed, "Social Media And Mental Health," Kaggle. Accessed: May 02, 2025. [Online]. Available: Https://Www.Kaggle.Com/Datasets/Souvikahmed071/Social-Media-And-Mental-Health
[17] K. Kroenke, R. L. Spitzer, And J. B. W. Williams, "The Phq-9," *J. Gen. Intern. Med.*, Vol. 16, No. 9, Pp. 606–613, 2001, Doi: Https://Doi.Org/10.1046/J.1525-1497.2001.016009606.X.
[18] D. Liu, Z. Chen, W. Marrero, N. Jacobson, And T. Thesen, "Explainable Machine Learning-Based Prediction Of Depression Severity In Medical Students," *Medrxiv*, 2023, Doi: 10.1101/2023.12.14.23299975.
[19] Z. I. Bimawan, T. Astuti, And P. Arsi, "Comparison Of Random Forest, K-Nearest Neighbor, Decision Tree, And Xgboost Algorithms For Detecting Stunting In Toddlers," *J. Tek. Inform.*, Vol. 5, No. 6, Pp. 1599–1607, 2024, Doi: 10.52436/1.Jutif.2024.5.6.2629.
[20] P. Elisa And A. R. Isnain, "Comparison Of Random Forest, Support Vector Machine And Naive Bayes Algorithms To Analyze Sentiment Towards Mental Health Stigma," *J. Tek. Inform.*, Vol. 5, No. 1, Pp. 321–329, 2024, Doi: 10.52436/1.Jutif.2024.5.1.1817.
[21] L. Breiman, "Random Forests," *Mach. Learn.*, Vol. 45, No. 1, Pp. 5–32, 2001, Doi: 10.1023/A:1010933404324.
[22] A. Salman, H. A., Kalakech, A., & Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, Vol. 2024, Pp. 69–79, 2024, Doi: 10.58496/Bjml/2024/007.
[23] M. Nalluri, M. Pentela, And N. R. Eluri, "A Scalable Tree Boosting System: Xg Boost," *Int. J. Res. Stud. Sci. Eng. Technol*, Vol. 7, No. 12, Pp. 36–51, 2020, Doi: 10.22259/2349-476x.0712005.
[24] T. Chen And C. Guestrin, "Xgboost: A Scalable Tree Boosting System," In *Proceedings Of The Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*, 2016. Doi: 10.1145/2939672.2939785.
[25] Y. Xu And R. Goodacre, "On Splitting Training And Validation Set: A Comparative Study Of Cross-Validation, Bootstrap And Systematic Sampling For Estimating The Generalization Performance Of Supervised Learning," *J. Anal. Test.*, Vol. 2, No. 3, Pp. 249–262, 2018, Doi: 10.1007/S41664-018-0068-2/Figures/9.
[26] S. Farhadpour, T. A. Warner, And A. E. Maxwell, "Selecting And Interpreting Multiclass Loss And Accuracy Assessment Metrics For Classifications With Class Imbalance: Guidance And Best Practices," *Remote Sens.*, Vol. 16, No. 3, 2024, Doi: 10.3390/Rs16030533.
[27] Arunraju Chinnaraju, "Explainable Ai (Xai) For Trustworthy And Transparent Decision-Making: A Theoretical Framework For Ai Interpretability," *World J. Adv. Eng. Technol. Sci.*, Vol. 14, No. 3, Pp. 170–207, Mar. 2025, Doi: 10.30574/Wjaets.2025.14.3.0106.
[28] J. M. Twenge And W. K. Campbell, "Associations Between Screen Time And Lower Psychological Well-Being Among Children And Adolescents: Evidence From A Population-Based Study," *Prev. Med. Reports*, Vol. 12, Pp. 271–283, 2018, Doi: Https://Doi.Org/10.1016/J.Pmedr.2018.10.003.
[29] B. Keles, M. Niall, And A. And 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, Dec. 2020, Doi: 10.1080/02673843.2019.1590851.
[30] H. Appel, A. L. Gerlach, And J. Crusius, "The Interplay Between Facebook Use, Social Comparison, Envy, And Depression," *Curr. Opin. Psychol.*, Vol. 9, Pp. 44–49, 2016, Doi: 10.1016/J.Copsyc.2015.10.006.
[31] A. Mamede, I. Merkelbach, G. Noordzij, And S. Denktas, "Mindfulness As A Protective Factor Against Depression, Anxiety And Psychological Distress During The Covid-19 Pandemic: Emotion Regulation And Insomnia Symptoms As Mediators," *Front. Psychol.*, Vol. Volume 13-2022, 2022, Doi: 10.3389/Fpsyg.2022.820959.
[32] C.-H. Tsai, M. Christian, Y.-Y. Kuo, C. C. Lu, F. Lai, And W.-L. Huang, "Sleep, Physical

Vol. 7, No. 2, April 2026, Page. 778-792
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5009

Activity And Panic Attacks: A Two-Year Prospective Cohort Study Using Smartwatches, Deep Learning And An Explainable Artificial Intelligence Model.," *Sleep Med.*, Vol. 114, Pp. 55–63, Feb. 2024, Doi: 10.1016/J.Sleep.2023.12.013.

[33] G. D. Price, M. V Heinz, S. H. Song, M. D. Nemesure, And N. C. Jacobson, "Using Digital Phenotyping To Capture Depression Symptom Variability: Detecting Naturalistic Variability In Depression Symptoms Across One Year Using Passively Collected Wearable Movement And Sleep Data," *Transl. Psychiatry*, Vol. 13, No. 1, P. 381, 2023, Doi: 10.1038/S41398-023-02669-Y.

[34] S. Jafarlou *Et Al.*, "Objective Monitoring Of Loneliness Levels Using Smart Devices: A Multi-Device Approach For Mental Health Applications," *Plos One*, Vol. 19, Jun. 2024, Doi: 10.1371/Journal.Pone.0298949.

[35] S. C. E. Nouis, V. Uren, And S. Jariwala, "Evaluating Accountability, Transparency, And Bias In Ai-Assisted Healthcare Decision- Making: A Qualitative Study Of Healthcare Professionals' Perspectives In The Uk," *Bmc Med. Ethics*, Vol. 26, No. 1, P. 89, 2025, Doi: 10.1186/S12910-025-01243-Z.