DATA SCIENCE AND APPLICATIONS SESSION

# A REVIEW OF AI METHODS USED IN DIGITAL TWINS FOR MENTAL HEALTH MONITORING

Nevena Radešić*,
[0009-0009-0107-5518]

Nemanja Jeličić
[0009-0004-2302-204X]

Singidunum University,
Belgrade, Serbia

Abstract:

Digital twins represent a novel approach to the continuous monitoring of mental health. This paper provides a review and critical analysis of four representative papers that apply various AI methods within digital twins for mental health monitoring: TwinMind (passive smartphone sensing with XGBoost and SHAP), Quamer et al. (multimodal fusion of facial, vocal, and physiological signals using a Swin Transformer-TGNN architecture), Wang et al. (EEG signals with a genetic algorithm for accelerating explainable AI methods), and Abilkaiyrkyzy et al. (a BERT based chatbot for conversational depression screening). For each work we described: the data, the AI/ML model, the explainability mechanism and the achieved results. The works are analysed from multiple perspectives: accuracy, explainability and practical applicability. The results show that passive sensing offers the best trade off between practicality and accuracy, whereas multimodal fusion achieves the highest accuracy at the cost of expensive equipment. Genetic algorithms significantly accelerate XAI (from ~30s to <10s), which is of great importance for real time operation. The chatbot exhibits high usability but low accuracy. The conclusion is that there is no universally optimal solution. Instead, the choice depends on the context. Future directions include federated learning for privacy preservation and hybrid architectures that combine sensors and conversational agents.

Keywords:

Digital Twin, Artificial Intelligence, Mental Health, Explainable AI.

## INTRODUCTION

Depression and stress are among the most common global mental health challenges. They are affecting people across all age groups and professions. A particularly vulnerable population is university students: meta-analyses show that 34-45% of students experience moderate to severe stress or anxiety - roughly double the rate of their non-student peers [1]. Traditional assessment methods, such as annual PHQ-9 questionnaires or occasional counselling sessions, provide only snapshot views and cannot capture the dynamic fluctuations of mental states over time.

Correspondence:
Nevena Radešić

e-mail:
nradesic@singidunum.ac.rs



DOI: 10.15308/Sinteza-2026-275-282

In contrast, digital twins have migrated from industrial applications to medicine. A digital twin is a virtual representation of a physical system that combines real-time data coupled with sophisticated computational models [2]. In the context of mental health, a digital twin can integrate data from smartphones, wearables, EEG devices, or chatbot dialogues to predict the risk of depression or anxiety.

Although a growing number of studies apply artificial intelligence to digital twins for mental health, a systematic comparison of different AI methods is lacking. It remains unclear which data sources they use, how accurate they are, whether their predictions are explainable, and how practical they would be in everyday use.

To address this gap, this paper analyses four representative approaches:

- TwinMind [1]- passive smartphone sensing with XGBoost and SHAP explanations;
- Quamer et al. [3] - multimodal fusion (face, voice, physiological signals) with a Swin Transformer-TGNN architecture;
- Wang et al. [4] - EEG signals with a genetic algorithm to accelerate XAI; and
- Abilkaiyrkyzy et al. [5]- a conversational BERT chatbot for depression screening.

For each approach, we describe the data used, the AI/ML model, the explainability (XAI) method, and the reported results. In the discussion, we compare these approaches across four criteria: accuracy, explainability, invasiveness, and practicality. We conclude with recommendations for future research, including federated learning [6] and hybrid architectures.

## 2. LITERATURE SELECTION

For the purposes of this paper, a survey of existing work was conducted in the fields of digital twins, artificial intelligence, and mental health. The papers were selected based on the following criteria:

- the work explicitly employs the concept of a digital twin for mental health;
- the work implements an AI/ML model for classification, prediction, or assessment of mental state;
- the work was published in a peer reviewed journal or as a preprint in the period 2021-2026; and
- the work provides sufficient detail on architecture, data, and results to enable comparative analysis.

Based on these criteria, four representative papers were selected, covering different types of input data (passive sensing, multimodal signals, EEG, text) and different AI paradigms (XGBoost, Transformer-GNN, genetic algorithms, BERT).

In addition to the four representative works analysed in detail, several other studies employ similar AI techniques either outside the digital twin paradigm or for different applications within mental health. Mujiyanto et al. (2024) [7] apply a Swin Transformer with enhanced dropout for facial expression recognition on the FER2013 and CK+ datasets, achieving 78.65% accuracy - confirming the validity of this architecture for emotion feature extraction outside the digital twin context. A 2025 study in Cognitive Computation [8] uses a Random Forest classifier on resting-state EEG data with dry electrodes for mental health triage (86% file-level accuracy), a technique comparable to Wang et al. but without the genetic algorithm acceleration or digital twin framing. Marrelli et al. (2025) [9] propose a digital twin chatbot with a self-attention mechanism within the Rasa framework, reaching 74% accuracy on the E-DAIC dataset - a different architectural choice compared to Abilkaiyrkyzy et al., though still lacking explainability. A systematic review and meta-analysis of 20 explainable ML studies for depression detection (2025) [10] confirms that XGBoost achieves the best average F1-score (0.86) and that SHAP is used in 70% of studies. This fact supports the methodological choices in TwinMind. Finally, a conceptual study on digital twins in forensic mental health (2026) [11] explores dynamic risk assessment for violence and self-harm. Study demonstrates the broader applicability of digital twin technology beyond depression monitoring in academic settings.

## 3. ANALYSIS OF SELECTED PAPERS

### 3.1. ABILKAIYRKYZY ET AL. (2024) - BERT-BASED CHATBOT FOR DEPRESSION SCREENING

In this paper, the authors identify three primary barriers to mental healthcare: stigma (reluctance to seek help due to shame), limited accessibility (shortage of psychiatrists, particularly outside major urban centres), and cost [5]. Although standardized methods such as the Patient Health Questionnaire-9 (PHQ-9) are effective, users are not always able or willing to complete them. A chatbot capable of conducting natural, anonymous conversations, available 24/7 without appointment delays, can address these challenges. The objective is not to replace clinicians, but to enable early detection and facilitate referral to appropriate care [5].

The authors propose a chatbot named Emi, implemented within the Rasa framework. Rasa comprises two core components: NLU (Natural Language Understanding) and Core (dialogue management). User input undergoes the following steps:

1. A BERT tokenizer maps words to numerical identifiers;
2. Featurization - A LanguageModelFeaturizer leverages a pre-trained BERT model to extract a contextualized feature vector (embedding) from the utterance;
3. Intent Classification - Instead of Rasa's default classifiers, the authors fine-tuned BERT for the target task. A fully connected layer with three neurons (corresponding to three depression severity classes) was appended to BERT. Fine-tuning was conducted on the E-DAIC dataset, which comprises transcripts of clinical interviews with 219 individuals. Optimization was performed using AdamW;
4. Dialogue Management (Core) - Rasa employs three policies: Memoization (for exact sequence recall), TED (Transformer Embedding Dialogue, for contextual understanding), and RulePolicy (for routine conversational segments, e.g., greetings) [5].

Additionally, a Fallback classifier was implemented. With that, when BERT's confidence falls below an acceptable threshold, the user is prompted to rephrase their input.

On the E-DAIC test set, the model achieves 69% accuracy and 0.77 F1-score for the severe depression class. Evaluation on 20 previously unseen students gave lower accuracy - 40%, or 65% when considering the top two predicted classes. This was expected, as student speech patterns differ from those observed in clinical interviews with diagnosed patients. Usability was rated highly, with a System Usability Scale (SUS) score of 84.75%. Notably exceeding the industry average of 68%.

Comparative analysis with analogous systems (Ada Health, ECAs) indicates that Emi is competitive, particularly with respect to ease of use [5].

### 3.2. WANG ET AL. (2025) - GENETIC ALGORITHM FOR ACCELERATING XAI IN EEG-BASED SYSTEMS

Wang et al. (2025) focus on addressing the response time problem in the application of explainable artificial intelligence (XAI) in digital twins. Traditional XAI methods, such as LIME [12] and SHAP [13], are computationally expensive. They generate hundreds of perturbed examples to explain a single prediction. That process can take up to 30 seconds [4]. For a digital twin expected to provide interpretable alarms in real time, this latency is unacceptable.

The authors introduce E-XAI (Evolving XAI), a framework that employs a genetic algorithm (GA) to first identify a small, yet informative subset of input features. The GA operates by evolving a population of binary vectors (indicating which features are selected), where the fitness of each candidate is measured by the accuracy of the model on that subset. Following GA-based feature selection, standard XAI methods (SHAP, LIME) are applied only to this reduced feature subset, enabling them to operate substantially faster.

E-XAI reduces explanation time from approximately 30 seconds to under 10 seconds, with a negligible loss in accuracy (less than 1.5%). The authors evaluated five machine learning models (Random Forest, XGBoost, MLP, AdaBoost, and KNN). The obtained results are presented in Table 1.

Three XAI techniques are compared: SHAP (theoretically the most rigorous), LIME (faster but less stable), and GA-evolved decision trees - the latter being the most intuitive for end users, as they yield simple rules of the form: "if feature A > X and feature B < Y, then the risk is high".

**Table 1.** Comparison between XAI and E-XAI [4]

| ML model | Accuracy | | Mean Fitness | Mean Training Time / s |
|---|---|---|---|---|
| | XAI | E-XAI | | |
| Random Forest | 0.8231 | 0.8125 | 0.9281 | 8.574 |
| Ada Boost | 0.8017 | 0.7960 | 0.9132 | 8.698 |
| MLP | 0.7876 | 0.7794 | 0.9025 | 8.934 |
| XGBoost | 0.8122 | 0.8055 | 0.9190 | 8.643 |
| KNN | 0.7654 | 0.7580 | 0.8918 | 8.785 |

The two main limitations of this study are the small sample size (only 17 participants) and the fact that the data are laboratory acquired. Nevertheless, the work makes an important contribution - it demonstrates that accelerating XAI is feasible through the application of genetic algorithms. Future work includes the integration of multimodal signals and closed-loop feedback with a clinician.

### 3.3. CHITIKELA (2025) - PASSIVE SENSING WITH XGBOOST (TWINMIND)

Reference [1] addresses the increasingly prevalent problem of depression among university students. The author proposes a Mental Health Digital Twin (MHDT) whose main objectives are to:

- integrate multimodal behavioural, physiological, and self-reported data;
- predict short-term risk for clinically relevant outcomes; and
- provide interpretable explanations for its predictions.

For the implementation of the MHDT, the publicly available StudentLife dataset was used. It contains data from 214 students at Dartmouth College. The data originate from three sources:

- three-daily ecological momentary assessments measuring mood and PHQ-4 responses;
- static demographic data (age, sex, major); and
- continuous passive smartphone sensing, including GPS entropy, lock/unlock frequency, call/SMS volume, and step counts.

The collected data are then organized into 7-day rolling (sliding) windows with a 1-day sliding increment. For each window, three statistical descriptors are computed: the mean (behavioural level), the linear trend (difference between the first and last values), and the slope estimated using ordinary least squares (OLS) regression. This process yields a feature vector of approximately 2800 dimensions per student per day. Following this aggregation of the data, the authors proceeded to label the data. Since the dataset contains no true clinical diagnoses for each day, the authors applied k-means clustering with k = 3 to the 2800 extracted features. In this manner, all seven-day windows were automatically grouped into one of three clusters: low risk, moderate risk, and high risk.

Upon examining the cluster centroids, it was observed that the categorization is meaningful: the high-risk cluster exhibits elevated PHQ-4 scores, irregular sleep patterns, and signs of social withdrawal.

A family of horizon-specific XGBoost ensembles then predicts a student's risk level at t+1, t+3, and t+7 days. Five-fold cross-validation was used in order to get more reliable results. For a model to be usable, the reasoning behind the prediction also must be explained. Therefore, the author computed SHAP (SHapley Additive exPlanations) values for each prediction. SHAP analysis was conducted at two levels:

- Global explanations: Three features were found to account for 67% of the predictive power: sleep regularity, mobility entropy (diversity of movement), and the trend of PHQ-4 scores.
- Local explanations: For each individual student, a SHAP force plot can be generated, showing which specific features of that student pushed the prediction toward high risk and which toward low risk.

The predictive performance of the MHDT model was evaluated across three time horizons: 1 day, 3 days, and 7 days ahead. The results are presented in Table 2.

A more detailed analysis by risk class reveals that the model best identifies students with low risk (precision 0.89). The greatest challenge lies in detecting high risk at the 7-day horizon, where recall is 0.55 - meaning that the model misses nearly every second student who will be in a high-risk state one week later.

The author also performed an ablation study. It was determined that EMA questionnaires contribute the most to prediction, while demographic data have the weakest impact.

Although the data demonstrate high performance, several limitations are encountered in this work:

- Single campus - Only Dartmouth College was studied;
- EMA dependence - Students must respond three times daily; those who fail so reduce the model's accuracy;

**Table 2.** Model Performance Metrics Across Prediction Horizons

| Metric | 1-day | 3-days | 7-days |
|---|---|---|---|
| Accuracy | 85.3% | 81.0% | 76.0% |
| AUC | 0.958 | 0.9304 | 0.8970 |

- Pseudo-labels - k-means clusters are not equivalent to clinical diagnoses;
- Assumption of sensor consistency - Phones vary; different Android/iOS models produce different signals; and
- No federated learning - All data are transmitted to a central server, posing a privacy risk.

### 3.4. QUAMER ET AL. (2026) - MULTIMODAL SWIN TRANSFORMER-TGNN ARCHITECTURE

The authors Amanullah Quamer et al. propose a framework for continuous, real-time monitoring of students' mental states using digital twins and artificial intelligence.

For training and model evaluation, two publicly available datasets were used: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) for emotion recognition from facial and vocal expressions, and WESAD (Wearable Stress and Affect Detection) for physiological stress signals (heart rate - HR, electrodermal activity - EDA, body temperature - TEMP). In addition, the authors synthetically generated behavioural data (screen time, typing speed) to emulate an academic environment. Preprocessing includes image augmentation (rotation, brightness adjustment), conversion of audio signals to Mel-spectrograms, and filtering of physiological signals using a Butterworth filter. These data represent the foundation of a four-layer system architecture:

1. Sensing Layer: IoT-enabled devices (e.g., webcam, smartwatch, microphone);
2. Processing Layer: The Swin Transformer-TGNN hybrid model processes raw signals into high-level emotional and stress features;
3. Digital Twin Layer: A virtual twin of each student is kept up to date with streaming data and evolving internal psychological states; and
4. Decision Layer: Computes the Mental Health Index (MHI) and generates intervention alerts when deviations exceed defined thresholds [3].

The Processing Layer is the core of the system and consists of two parts: Swin Transformer and TGNN.

The Swin Transformer extracts localized emotional features from facial and acoustic modalities using window-based self-attention. This allows for efficient feature extraction at different hierarchical levels, while dealing with occlusions and illumination variations effectively [3].

A Temporal Graph Neural Network (TGNN) is used to model physiological and behavioural signals. Each signal (HR, EDA, temperature, screen time) represents a node in the graph G_t=(V_t,E_t), where edges encode both multimodal and temporal relationships. This captures long-term dependencies such as increasing EDA combined with reduced activity, indicating stress buildup [3].

Fusion of the outputs from the Swin Transformer and the TGNN is achieved through an attention-based fusion mechanism. This mechanism enables the system to adapt when one of the sensors fails - the weight of the unreliable modality is automatically reduced, while the remaining modalities assume a greater role.

The Mental Health Index is a composite measure of a student's overall well-being on a scale from 0 to 1, where a higher value indicates a better state. It is calculated according to Equation 1:

$$MHI = \gamma_1 E_{AI} + \gamma 2 (1 - S_{stress}) + \gamma_3 (1 - \Delta B)$$

**Equation 1.** The Mental Health Index

where:

- $E_{AI}$ is the probability of a positive emotion (0-1) provided by the Swin Transformer;
- $S_{stress}$ is the normalized stress intensity (0-1) from the TGNN; and
- $\Delta B$ is the behavioral deviation (e.g., a drastic change in screen time).

Coefficients ($\gamma_1$, $\gamma_2$, $\gamma_3$) are set to (0.4, 0.4, 0.2) empirically. If the MHI falls below 0.4, the system generates an early intervention alert. Note that the MHI is fully transparent: each component is interpretable, which is a form of explainability, although the authors do not employ standard XAI methods such as SHAP or LIME.

To evaluate the proposed system, a comparison was performed using 5-fold cross-validation. The CNN-BiLSTM model was selected as the baseline. The proposed model achieved better performance in accuracy, F1 score, Pearson correlation and average inference latency. The results are shown in Table 3.

**Table 3.** Performance Comparison Between Baseline and Proposed Models [3]

| Metric | CNN-BiLSTM (Baseline) | Proposed Swin-TGNN Model | Improvement (%) |
|---|---|---|---|
| Emotion recognition accuracy | 84.6 | 93.2 | +8.6 |
| Stress detection F1 score | 0.79 | 0.88 | +11.3 |
| MHI correlation (r) | 0.82 | 0.93 | +13.4 |
| Inference latencu (ms) | 310 | 176 | -43.2 |

Qualitative inspection of attention maps revealed that the Swin Transformer emphasized critical facial regions (eyes, eyebrows, mouth corners), while the TGNN identified temporal dependencies between physiological and behavioural cues, such as heart-rate spikes linked with prolonged screen time. The model's adaptive fusion mechanism prioritized reliable modalities in real time, maintaining over 87% accuracy even when one modality was missing. This validates its generality and accuracy and its potential use for application in university environments where the sensors may not always be consistent [3].

Despite the impressive figures, the work has several significant limitations:

- Behavioural data were synthetically generated, not collected from real students in an academic environment;
- Evaluation was conducted exclusively on laboratory datasets (RAVDESS, WESAD) that contain acted performances rather than authentic emotional reactions of students under stress;
- Although the authors refer to the system as a "digital twin", it actually implements only one-way synchronization (from student to model) without a closed feedback loop in which the model learns from intervention outcomes - an essential characteristic of a true digital twin;
- Explainability is limited to attention maps and the transparent MHI formula; and
- The work does not include validation on a real student population, remaining at the conceptual level.

Nevertheless, this work makes an important contribution to the field. The hybrid Swin Transformer-TGNN architecture represents an advancement over simpler models (e.g., XGBoost in TwinMind). Furthermore, the explicit introduction of federated learning is a big step toward better data privacy preservation.

## 4. DISCUSSION

The four analyzed works demonstrate different philosophies for applying AI to digital twins in mental health.

TwinMind relies on passive sensing - data that the phone already collects (GPS, unlock events, step counts). Its advantage is that it requires no additional equipment. The drawback is that EMA questionnaires are critical for accuracy - when removed, AUC drops by 0.075. If a student does not respond to them (which is common), the model loses reliability. Also, the use of k-means pseudo-labels instead of clinical diagnoses further limits validity.

Quamer et al.'s approach is the most demanding in terms of equipment - a camera, microphone, and smartwatch are required. However, it also achieves the highest accuracy (93.2% for emotion recognition). Its hybrid architecture (Swin Transformer for spatial features, TGNN for temporal features) is technically impressive, but its feasibility in a real academic environment is questionable. Students would likely not want to be recorded by a camera while studying. Furthermore, the synthetic behavioural data diminish validity.

Wang et al. address a key problem of digital twins - explanation speed. They demonstrate that a genetic algorithm can reduce the runtime of XAI methods from ~30s to <10s, with negligible loss in accuracy (1.5%). This is a significant contribution - clinicians will not wait half a minute to understand why a student has been flagged as at-risk. However, experiment was conducted with only 17 participants in a laboratory setting. It is unclear how the system would perform on real students in a classroom environment.

Abilkaiyrkyzy et al. approach the problem from a conversational perspective - the chatbot Emi conducts a dialogue and classifies depression severity. This is the most accessible approach. However, the price is low accuracy - only 40% on real users, or 65% when considering the top two predicted classes. Additionally, the model is a black box - there are no explanations for why a particular class was assigned to an individual.

A comparison according to the mentioned criteria is provided in Table 4.

**Table 4.** Comparison of the four reviewed works according to the defined criteria

| Criterion | TwinMind | Quamer et al. | Wang et al. | Emi chatbot |
|---|---|---|---|---|
| Data | GPS, unlock, EMA, steps | face, voice, HR, EDA, synthetic | EEG (5 electrodes) | text (dialogue) |
| AI model | XGBoost | Swin-TGNN | RF/XGBoost + GA | fine-tuned BERT |
| XAI | SHAP | attention maps | SHAP/LIME/trees + GA | none |
| Accuracy (best) | AUC 0.958 (1 day) | 93.2% for emotions | 81.25% (RF) | 69% (dataset) / 40% (real-world) |
| Explanation time | ~42 ms | not measured | <10s (vs. 30s for LIME) | not applicable |
| Sample | 214 real students | datasets + synthetic | 17 laboratory participants | 20 students |
| Practicality | high (phone only) | low (expensive equipment) | medium (EEG device) | high (phone only) |

## 5. CONCLUSION

This review analysed four representative approaches to integrating artificial intelligence into digital twins for mental health monitoring: passive smartphone sensing, multimodal signal fusion, EEG based classification with accelerated explainability, and a conversational chatbot.

Passive sensing, as implemented in TwinMind, offers the best trade off between practicality and accuracy. The inclusion of SHAP based explanations, both global and local, further strengthens its clinical utility. However, its dependence on EMA questionnaires - which suffer from real world non compliance - and the use of k means pseudo labels instead of clinical diagnoses remain significant limitations.

Multimodal fusion, proposed by Quamer et al., achieves the highest raw accuracy among the reviewed works. The hybrid Swin Transformer–TGNN is technically impressive, but this approach requires expensive equipment, raises privacy concerns in real academic environments, and relies on synthetic behavioral data.

Wang et al. make an important technical contribution by demonstrating that genetic algorithms can reduce the runtime of explainable. Nevertheless, validation on only 17 laboratory participants leaves generalizability unproven.

The conversational chatbot developed by Abilkaiyrkyzy et al. is the most accessible approach. It requires a smartphone and text messages. Users rate its usability highly, with a System Usability Scale score of 84.75%. However, its accuracy on real users is only 40%, and the model operates as a black box without any explainability mechanism. In its current form, it is unsuitable for clinical diagnosis.

It is also worth noting that none of the four works implement a true closed loop digital twin that learns from intervention outcomes. Despite being labelled as digital twins, they are essentially predictive monitoring systems.

Looking forward, future research should prioritize federated learning for privacy preserving multi institution training, hybrid architectures that combine passive sensing with conversational agents, clinical validation with real patient populations rather than synthetic or laboratory data, and standardized evaluation of explainability. Until then, digital twins for mental health remain a promising but maturing field.

## REFERENCES

[1] Y. Chitikela, "TwinMind: An Explainable Digital Twin for Early Detection of Anxiety and Depression Using Passive Sensing," SSRN, 2025.

[2] A. B. Abdelnabi, A. Bany and G. G. Rabadi, "Real-Time Medical Aid Delivery: A Digital Twin Approach with Dynamic Vehicle Routing Problem," in *2025 IEEE 5th International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 2025.

[3] A. Quamer and K. H. Mir, "Digital Twin-Based AI System for Mental Health Monitoring in Academic Environments," 2026.

[4] Z. W. Zhang, A. Y. A. Hammadi, X. Huang, F. Guo, E. Damiani, C. Y. Yeun and L. Li, "Evolving Explainable Artificial Intelligence for electroencephalography-based mental health classification in digital twin systems," *Ad hoc networks*, p. 103964, 2025.

[5] A. Abilkaiyrkyzy, A. A. Laamarti, M. Hamdi and A. E. Saddik, "Dialogue system for early mental illness detection: toward a digital twin solution," *IEEE Access*, 2024.

[6] M. B, B. M. Moore, D. Ramage and S. Hampson, "Communication-efficient learning of deep networks from decentralized data," *InArtificial intelligence and statistics*, pp. 1273-1282, 2017.

[7] M. Mujiyanto, A. Setyanto, K. Kusrini and E. Utami, "Swin Transformer with Enhanced Dropout and Layer-wise Unfreezing for Facial Expression Recognition in Mental Health Detection," *Engineering, Technology & Applied Science Research*, vol. 14, pp. 19016-19023, 2024.

[8] J. Damian, "Non-Invasive Classification of Mental Health Disorders Using RestingState EEG with Dry Electrodes for Scalable Triage," *Cognitive Computation*, 2025.

[9] M. RS, R. S. M. Hasan, S. Yarram and B. S. Guttikonda, "Digital-Twin System based on Chatbot Framework with Self-Attention Mechanism for Early Mental Illness Detection," in *International Conference on Data Science and Information System (ICDSIS)*, 2025.

[10] A. Trelles, A. T. Ruiz and A. P. Rojo, "Systematic Review and Meta-Analysis of Explainable Machine Learning Models for Clinical Depression Detection," in *Behavioral Sciences*, 2025.

[11] J. O. Ogunleye, T. O. Ebo, J. O. Alabi, B. D. Makanjuola, E. Egbon and D. B. Olawade, "Digital twin technology in forensic mental health: a narrative review," *Journal of Forensic and Legal Medicine.*, 2026.

[12] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier.," in *ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[13] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions.," *Advances in neural information processing systems*, 2017.