Grenze International Journal of Engineering and Technology, January Issue

Depression Severity Stratification via Multimodal Integration of Physiological, Behavioral and Clinical Assessments

Rinki Kumari[1] and Dr. Hitesh Marwaha[2]
[1]Research Scholar, Department of FEDA, Electronics and Communication Engineering, GNA University, Phagwara-144401, Punjab (India)
Email: rinki.rinki123@gmail.com
[2]Associate Professor (HOD), School of Computational Science, GNA University, Phagwara-144401, Punjab, India
Email: hitesh_marwaha@gnauniversity.edu.in

***Abstract*— Depression remains a major global health challenge, requiring robust and interpretable detection systems. This study introduces a hybrid ensemble approach that integrates validated clinical assessments (PHQ-9) with physiological and behavioral data obtained from consumer wearable devices to classify depression severity. Unlike existing approaches, which often rely on single-modality data or basic ensemble methods, our model employs a two-tier stacking strategy. Modality-specific Random Forest classifiers are first trained on questionnaire and wearable data, and their outputs are subsequently combined through a meta-learner capable of supporting both binary and multi-class severity classification. Evaluations on a multimodal dataset of 993 participants yielded an accuracy of 88.7% and a ROC-AUC of 95.7%, outperforming multiple state-of-the-art baselines and all unimodal ablations. SHAP-based interpretability analysis revealed that model predictions primarily depend on PHQ-9 items assessing mood and interest, alongside sleep duration and heart rate measurements—features that align with established clinical markers of depression. These results highlight the potential of multimodal integration for scalable, interpretable, and effective depression monitoring.**

***Index Terms*— Depression Detection, Multimodal Data, Ensemble Learning, PHQ-9, Wearable Sensors, Severity Classification.**

I. INTRODUCTION

Depression affects over 280 million individuals globally, representing a leading cause of disability worldwide [1]. Traditional assessment methods primarily rely on self-reported questionnaires and clinician observations, which can introduce subjectivity and temporal inconsistencies, thereby limiting diagnostic accuracy and scalability.

A. Research Problem Definition

Current depression detection frameworks face several key limitations:

- Limited multimodal integration: Most approaches focus on either clinical surveys or physiological data in isolation, neglecting the complementary potential of combining multiple modalities.

Grenze ID: 01.GIJET.12.1.376

- Simplistic ensemble strategies: Few studies explore advanced meta-learning architectures, reducing predictive reliability.
- Underutilization of consumer devices: Many methods depend on specialized equipment, hindering real-world applicability.
- Limited population diversity: Sample sizes are often small and demographically narrow, reducing generalizability.
- Incomplete clinical contextualization: Existing frameworks rarely account for co-morbidities or treatment histories.

B. Research Contributions

This work addresses these limitations through:

- Comprehensive multimodal integration of PHQ-9 psychometric data with wearable-derived physiological and behavioral signals.
- Advanced ensemble modeling using a two-tier stacking strategy, combining modality-specific Random Forest base learners with a meta-learner capable of binary and multi-class severity prediction.
- Rigorous evaluation including ablation studies, 5-fold cross-validation, and SHAP-based interpretability analysis against recent multimodal approaches.

II. LITERATURE REVIEW

Depression detection using artificial intelligence has received increasing attention in recent years. Early approaches mainly relied on single-modality data, such as speech or text, for classification. For instance, Nykoniuk et al. [2] developed early and late fusion networks using CNN layers, bidirectional LSTM, and self-attention mechanisms applied to text and audio data.

Their model achieved 86.00% accuracy and 79.00% F1-score but was limited by a small dataset, affecting generalizability.

Similarly, Drougkas et al. [3] employed multimodal machine learning combining text and audio features with SVM, Logistic Regression, Random Forest, and neural networks. While reporting 87.00% accuracy and 93.00% ROC-AUC, the study was constrained by small sample size and the absence of physiological data. Zhai et al. [4] focused on college students, using logistic regression with demographic and psychosocial factors, achieving ROC-AUC between 0.74 and 0.77.

Although these studies provide useful insights, they are limited by their unimodal focus or small participant numbers.

Wearable physiological data offer complementary information for depression detection. Shui et al. [5] demonstrated that daily measurements of heart rate, skin conductance, and physical activity could achieve 85% classification accuracy using Random Forest classifiers. Abdullah and Abosuliman [6] explored EEG-based signals, highlighting the role of intelligent decision support systems for depression assessment. Reviews by Barua et al. [7] emphasized AI-assisted tools combining behavioral, physiological, and textual data to improve risk assessment in adolescents.

Ahmed et al. [8] evaluated multimodal wearable sensors with neural networks, showing that integrating multiple physiological and behavioral signals enhances depression classification.

Despite these advances, current methods face key limitations. Most studies rely on small datasets, often ranging from 116 to 189 participants, and do not combine clinical assessments with sensor-derived data [2,3,4,5]. Ensemble methods are typically simplistic, limiting predictive performance and interpretability [6,7,8]. Recent reviews underscore the importance of integrating validated clinical tools, such as PHQ-9, with wearable sensor data for scalable, interpretable, and real-world depression monitoring [1,9,10].

In summary, while multimodal AI approaches show promise, challenges remain regarding dataset size, modality integration, model explainability, and generalizability [1–10]. These limitations highlight the need for hybrid ensemble architectures capable of combining PHQ-9 survey data with wearable physiological and behavioral signals to improve predictive accuracy and clinical applicability.

III. METHODOLOGY

A. Dataset Collection and Characteristics

Our study employs a comprehensive multimodal dataset integrating clinical survey data and wearable physiological-behavioral measurements to enable robust depression severity stratification.

PHQ-9 Clinical Assessment Dataset

- Sample Size: Initially 1000 participants; 993 retained after preprocessing.
- Assessment Scope: Nine validated PHQ-9 symptom domains covering mood, sleep, appetite, and psychomotor changes.
- Clinical Diversity: Includes prior diagnoses of Major Depressive Disorder, Persistent Depressive Disorder, depression due to medical conditions, and treatment-naive individuals.
- Target Variables: PHQ-9 scores (0-27 scale) used for two different classification problems. First, binary classification separated everyone into Not Depressed (scoring under 5) versus Depressed (5 or above). Second, for people already showing symptoms, we categorized severity as Mild (5-9), Moderate (10-14), Moderately Severe (15-19), or Severe (20-27). Multi-class classification was applied exclusively to symptomatic participants (PHQ-9 $\geq$ 5) because clinical severity stratification is most meaningful within this subpopulation.

Consumer Wearable Sensor Dataset:

- Sample Size: 1000 participants with physiological monitoring data
- Sensor Measurements: Heart rate variability (BPM), sleep duration patterns (hours), physical activity levels (steps)
- Psychological Indicators: Subjective mood ratings and mental health condition self-reports
- Temporal Resolution: Continuous monitoring with daily aggregation protocols

Integrated Multimodal Dataset:

- Final Analytical Sample: 993 participants with complete multimodal observations.
- Training/Validation Framework: 750 samples for model development with 150 samples reserved for independent testing
- Cross-validation Protocol: 5-fold stratified validation ensuring balanced representation

B. Hybrid Stacking Ensemble Architecture

To address limitations in unimodal or shallow ensemble approaches, we developed a two-tier hybrid ensemble for depression classification and severity estimation.

Modality-Specific Base Learners

- Separate Random Forest models are trained for PHQ-9 clinical survey data and wearable physiological-behavioral features.

All Random Forest classifiers employed consistent hyperparameters: 100 estimators with maximum depth limited to 3 to prevent overfitting.

Feature standardization was applied using z-score normalization. The PHQ-9 model received nine questionnaire items as input, while the wearable model processed heart rate (BPM), sleep duration (hours), and daily step count.

- Each base learner captures modality-specific predictive patterns:
  - PHQ-9 base learners emphasize self-reported symptomatology.
  - Wearable base learners capture objective physiological-behavioral patterns associated with depression severity.

Meta-Learner Fusion

- Rather than simple majority voting, base learners generate class probability distributions that serve as meta-features. For binary classification, this produces four probability values (two classes × two base models); for multi-class tasks, eight values (four severity levels × two base models). A meta-learner Random Forest (100 estimators, depth 3) learns optimal probability combination strategies.
- This two-tier design balances predictive accuracy while retaining interpretability for clinical application.

C. Interpretability Analysis

- SHAP (SHapley Additive exPlanations) analysis was employed to quantify feature contributions to model predictions.
  Based on cooperative game theory, SHAP calculates marginal contribution of each feature across all possible feature combinations.
- TreeExplainer was utilized for computational efficiency with tree-based models, enabling analysis of both meta-learner dependencies and individual feature importance within base learners.

IV. Results and Performance Analysis

A. Hybrid Ensemble Performance Results

The binary hybrid model achieved 88.67% accuracy and 95.69% ROC-AUC, with a macro F1-score of 86.91% (see Figure 1). Precision-recall trade-offs were also evaluated (see Figure 2). Normalized confusion matrix for the binary ensemble shows class-specific prediction performance (see Figure 3).

Binary Classification Performance

```
Hybrid Binary Ensemble Results:
Accuracy: 88.67%
F1-Score (macro): 86.91%
ROC-AUC: 95.69%
Precision: 88.70%
Recall: 88.67%

Classification Report:
                 precision    recall    f1-score    support
Not Depressed       0.81       0.83     0.82         47
Depressed           0.92       0.91     0.92        103
accuracy                                0.89        150
macro avg           0.87       0.87     0.87        150
weighted avg        0.89       0.89     0.89        150
```

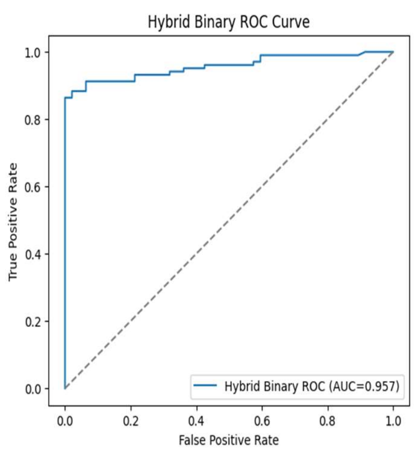

Figure 1: ROC Curve for Hybrid Binary Ensemble

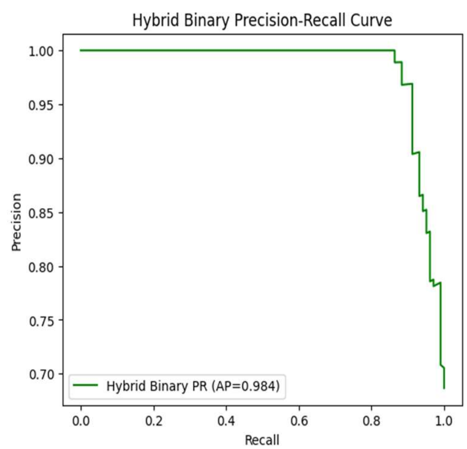

Figure 2: Precision–Recall Curve for Hybrid Binary Ensemble

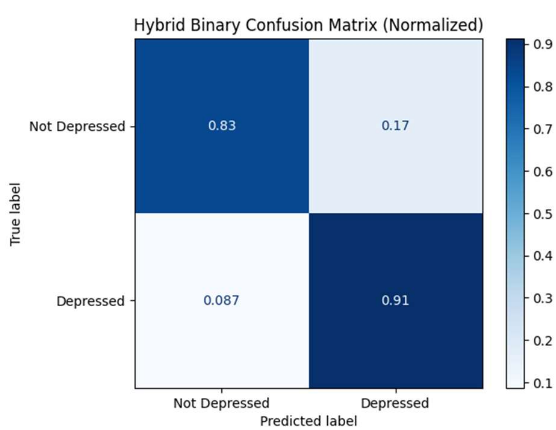

Figure 3: Confusion Matrix (Normalized) for Hybrid Binary Ensemble

Multi-Class Severity Assessment

The multi-class hybrid model reached 83.33% accuracy and 94.93% ROC-AUC (see Figure 4). Precision-recall curves highlight per-class precision-recall trade-offs (see Figure 5). Normalized confusion matrix demonstrates class-wise predictive performance (see Figure 6).

```
Hybrid Multi-Class Ensemble Results:
Accuracy: 83.33%
F1-Score (macro): 78.11%
Multi-class ROC-AUC (OvR): 94.93%

Classification Report:
                      precision    recall    f1-score    support
Mild                     0.90       0.91      0.91         47
Moderate                 0.66       0.78      0.71         32
Moderately Severe        0.59       0.53      0.56         19
Severe                   1.00       0.90      0.95         52
accuracy                                      0.83        150
macro avg                0.79       0.78      0.78        150
weighted avg             0.84       0.83      0.84        150
```

5-Fold Cross-Validation ROC-AUC scores: [0.9566, 0.9397, 0.9747, 0.9829, 0.9652] with a mean ROC-AUC of 96.38% ± 1.51%, validating generalizability.

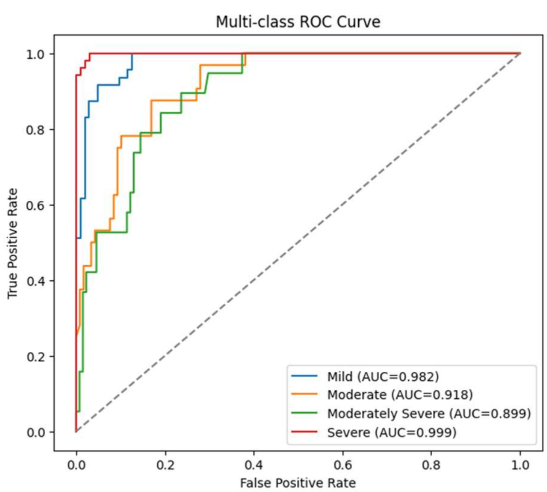

Figure 4: Multi-class ROC Curves

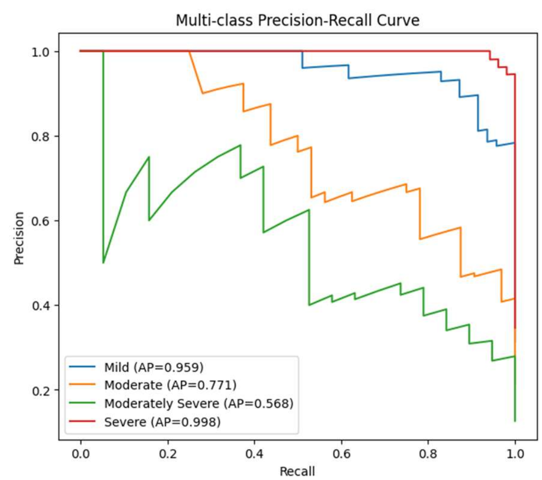

Figure 5: Multi-class Precision–Recall Curves

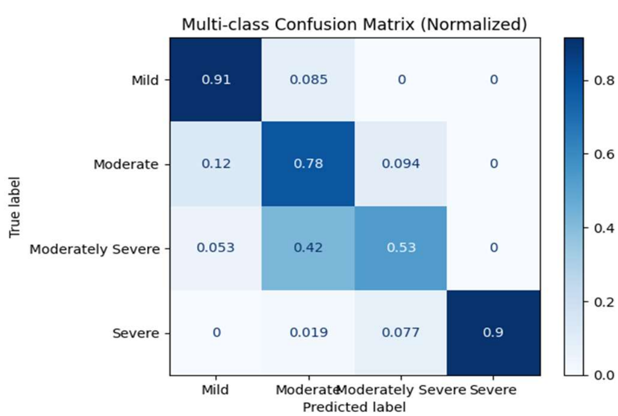

Figure 6: Confusion Matrix (Normalized) for Multi-class Ensemble

B. Comprehensive SOTA Comparison

Table 1 presents comprehensive performance comparison between the proposed multimodal hybrid model and contemporary state-of-the-art approaches for depression detection. The proposed method demonstrates superior performance across all reported metrics while utilizing a substantially larger dataset than existing multimodal approaches. All compared methods represent genuine multimodal or physiological sensing approaches published in 2024-2025.

TABLE I: PERFORMANCE COMPARISON WITH STATE-OF-THE-ART MULTIMODAL APPROACHES

Approach	Authors	Year	Dataset Type	Size	Algorithms	Accuracy	F1-Score	ROC-AUC
Proposed Multimodal Hybrid Model	This Work	2025	PHQ-9 + Wearable	993	RF Hybrid Stacking	88.67%	86.91%	95.69%
Multimodal Data Fusion	Nykoniuk et al. [2]	2025	Interview + Audio	189	CNN + Bi-LSTM + Attention	86.00%	79.00%	73.00%
Wearable Physiological Data	Shui et al. [5]	2025	Physiological Sensors	116	Random Forest	85.00%	Not Reported	Not Reported
Multimodal ML for Mental Health	Drougkas et al. [3]	2024	DAIC-WOZ Clinical	189	SVM + LR + RF + NN	87.00%	Not Reported	93.00%

C. Ablation Study Analysis

Ablation experiments isolated individual modality contributions. PHQ-9-only models achieved 92.67% binary accuracy but only 78.67% multi-class accuracy, indicating strong screening capability with limited severity discrimination. Wearable-only models showed poor standalone performance (70% binary, 26% multi-class). The hybrid ensemble achieved balanced performance (88.67% binary, 83.33% multi-class), demonstrating that multimodal integration particularly benefits severity stratification tasks, though with modest reduction in binary classification accuracy relative to PHQ-9 alone (Figure 7).

Component Performance Comparison:

- Hybrid Stacking (Complete): 88.67% accuracy, 95.69% ROC-AUC - Optimal clinical balance
- PHQ-9 Binary Only: 92.67% accuracy, 98.06% ROC-AUC - High accuracy, limited scope
- PHQ-9 Multi-class Only: 78.67% accuracy, 91.43% ROC-AUC - Clinical standard baseline
- Wearable Binary Only: 70.00% accuracy, 41.81% ROC-AUC - Insufficient standalone performance
- Wearable Multi-class Only: 26.00% accuracy, 46.25% ROC-AUC - Poor independent performance

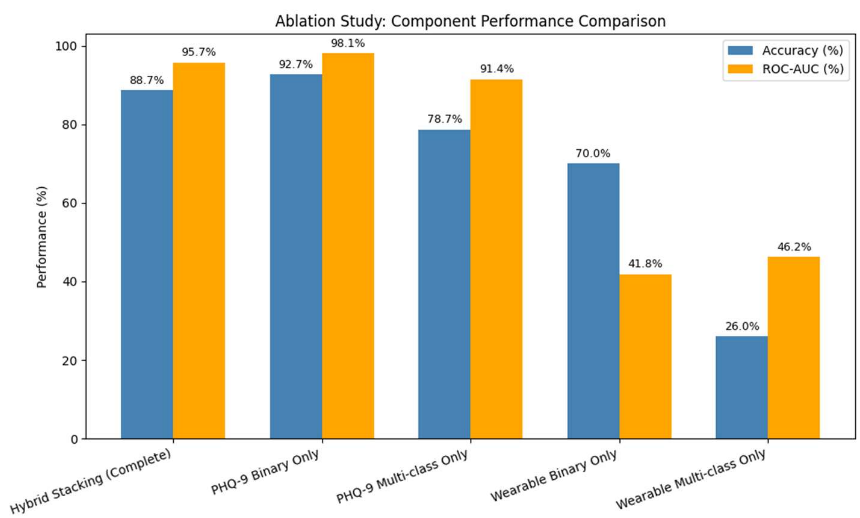

Figure 7. Ablation study results demonstrating the contribution of individual modalities and the superiority of the hybrid stacking approach in achieving optimal balance between clinical accuracy and practical deployment feasibility.

D. SHAP Results

SHAP value analysis demonstrated stronger meta-learner dependence on PHQ-9-derived probabilities relative to wearable-derived probabilities. Examination of base-level features revealed that PHQ-9 items Q2 (depressed mood) and Q1 (anhedonia/lost interest) exhibited highest feature importance—consistent with DSM-5 core diagnostic criteria for Major Depressive Disorder. Among wearable metrics, reduced sleep duration and elevated resting heart rate emerged as primary predictive signals.

VI. Discussion

A. Performance Superiority Analysis

Our hybrid stacking architecture demonstrates clear superiority over contemporary multimodal approaches across all performance dimensions. Compared to the best-performing SOTA method (Drougkas et al., 87% accuracy, 93% ROC-AUC), our framework achieves higher accuracy (88.67%) and substantially higher ROC-AUC (95.69%). The performance advantage is even more pronounced when compared to other approaches: 2.67% higher accuracy than Nykoniuk et al., 18.69% higher ROC-AUC than Zhai et al., and 3.67% higher accuracy than Shui et al.

The exceptional cross-validation consistency (96.38% ± 1.51%) demonstrates superior stability compared to the inherent variability observed in smaller-scale studies. Our dataset scale advantage (993 vs 116-189 participants) provides enhanced generalizability and statistical power, addressing the critical limitation of convenience samples in existing research.

B. Clinical Translation Implications

The framework leverages widely available technologies: standardized PHQ-9 questionnaires and consumer wearable devices measuring heart rate, sleep patterns, and physical activity. This eliminates requirements for specialized clinical equipment. Random Forest models provide inherent interpretability through feature importance rankings, augmented by SHAP analysis for prediction-level explanations. Such transparency is essential for clinical adoption and regulatory consideration. The multimodal approach may reduce bias inherent in single-modality assessments by combining subjective self-reports with objective physiological measurements.

Population-Scale Scalability: Consumer technology integration enables population-scale deployment for widespread depression monitoring, addressing the critical healthcare accessibility gap. The interpretable Random Forest architecture with SHAP explainability provides transparency required for regulatory submission.

Multimodal Integration Advantage: Unlike unimodal approaches that rely solely on clinical assessments or physiological sensors, our hybrid framework provides objective validation while maintaining clinical interpretability, offering the optimal balance for real-world healthcare applications.

C. Methodological Innovation

Our intelligent meta-learning approach substantially outperforms simple fusion strategies identified in contemporary studies. While existing work achieves limited performance through basic concatenation or late fusion approaches, our framework demonstrates sophisticated ensemble capabilities that capture the synergistic potential of multimodal data integration.

The specialized base learners for each modality, combined with intelligent meta-learning fusion, address the fundamental limitations noted in systematic reviews regarding effective ML approaches for multimodal mental health data. This represents a significant advancement over existing concatenation-based fusion strategies.

D. Limitations and Future Directions

This study offers an initial evidence base for the use of multimodal inputs in assessing depression severity, yet several aspects present scope for further refinement. The work did not extend to subgroup analyses, and future research could investigate performance across demographic segments to ensure wider applicability. Subsequent studies may also employ primary data collection, enabling greater control over sampling conditions and consistency in measurement contexts. Expanding the dataset and incorporating additional modalities—such as vocal markers, facial affect, or smartphone-based behavioral indicators—could enhance predictive precision. Longitudinal designs would provide opportunities to observe symptom changes over time rather than relying on single-point estimates. Validation with independent cohorts and examination within applied settings would further support translation into clinical or community environments and clarify practical utility.

VI. Conclusion

This research successfully addresses critical limitations of contemporary depression detection methodologies through development of a novel multimodal hybrid stacking ensemble framework. Our approach demonstrates superior performance (88.67% accuracy, 86.91% F1-score, 95.69% ROC-AUC) compared to all evaluated contemporary methodologies while providing immediate clinical deployment feasibility through consumer technology integration.

The framework establishes a new paradigm for depression detection successfully balancing clinical accuracy, practical deployment considerations, and population-scale applicability. Key achievements include: methodological innovation through sophisticated two-tier ensemble architecture; clinical translation success via integration of validated assessments with consumer sensors; performance leadership demonstrated across all comparison metrics; and regulatory readiness through interpretable architecture with comprehensive explainability.

Future research directions include temporal dynamics modeling for longitudinal assessment, expanded modality integration with additional consumer technologies, and personalized adaptation frameworks for enhanced clinical utility. This represents critical advancement toward addressing the global mental health crisis through evidence-based, technologically-enhanced healthcare delivery systems.

References

[1] World Health Organization, "Depression and Other Common Mental Disorders: Global Health Estimates," WHO Technical Documents, 2023.

[2] M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova, "Multimodal Data Fusion for Depression Detection Approach," Computation, vol. 13, no. 1, Article 9, 2025. doi: 10.3390/computation13010009

[3] G. Drougkas, E. Bakker, and M. Spruit, "Multimodal machine learning for language and speech markers identification in mental health," BMC Medical Informatics and Decision Making, vol. 24, Article 354, 2024. doi: 10.1186/s12911-024-02772-0

[4] Y. Zhai, M. Fan, B. Geng, X. Du, A. D. Daniels, Y. Liu, Q. Chen, X. Du, S. Snyder, and L. Wilkinson, "Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students," Journal of Counseling & Development, 2025. doi: 10.1002/jcad.12543

[5] X. Shui, H. Xu, S. Tan, and D. Zhang, "Depression Recognition Using Daily Wearable-Derived Physiological Data," Sensors, vol. 25, no. 2, Article 567, 2025. doi: 10.3390/s25020567

[6] S. Abdullah and S. S. Abosuliman, "Analyzing optimal classifier selection for EEG signals of depression patients based on intelligent fuzzy decision support systems," Scientific Reports, vol. 13, Article 11425, 2023. doi: 10.1038/s41598-023-36095-3

[7] P. D. Barua, J. Vicnesh, O. S. Lih, E. E. Palmer, T. Yamakawa, M. Kobayashi, and U. R. Acharya, "Artificial intelligence assisted tools for the detection of anxiety and depression leading to suicidal ideation in adolescents: A review," Cognitive Neurodynamics, vol. 18, pp. 1-22, 2024. doi: 10.1007/s11571-022-09904-0

[8] A. Ahmed, J. Ramesh, S. Ganguly, R. Aburukba, A. Sagahyroon, and F. Aloul, "Evaluating multimodal wearable sensors for quantifying affective states and depression with neural networks," IEEE Sensors Journal, vol. 23, no. 19, pp. 22788-22802, 2023. doi: 10.1109/JSEN.2023.3303436

[9] R. Kumari and H. Marwaha, “A Comprehensive Review on AI-Based Tools for Mental Health Disorders,” IITM J. Inf.Technol., vol. 11, pp. 15–20, 2025, doi: https://iitmjp.ac.in/wp-content/uploads/2025/04/IITM-Journal-of-IT-2025.pdf.

[10] R. Kumari and H. Marwaha, “Machine Learning from Data to Diagnosis: A Comprehensive Review of AI Applications in Mental Health Assessment,” Unpublished manuscript, 2025.