Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 2, 2026

A MULTIMODAL EXPLAINABLE AI FRAMEWORK FOR REAL-TIME DEPRESSION AND ANXIETY DETECTION

Azhan Bhayo

Department of Computer Science, NED University of Engineering and Technology Karachi, Pakistan

bhayo4506859@cloud.neduet.edu.pk

DOI: https://doi.org/10.5281/zenodo.18657323

Keywords
Depression detection, Anxiety analysis, Multimodal AI, Explainable AI, Digital phenotyping, Eye-tracking, Facial expression analysis, Voice biomarkers, Mental health monitoring

Article History
Received: 17 December 2025
Accepted: 01 February 2026
Published: 16 February 2026

Corresponding Author: *
Azhan Bhayo

Abstract
Depression and anxiety are among the leading causes of disability worldwide, yet current diagnostic methods rely mainly on self-reports and clinical interviews, which suffer from subjectivity and limited ecological validity. Advances in artificial intelligence (AI) and digital phenotyping offer novel opportunities to improve detection by leveraging multimodal behavioral and physiological signals. This paper presents the design of an explainable multimodal AI system that integrates three primary modalities: eye-tracking, facial expressions, and voice biomarkers, with future extensions to GPS mobility, typing dynamics, and gesture analysis. The proposed system activates three models at regular intervals of 2 hours, each recording data for 5 minutes to ensure unobtrusive, energy-efficient, and ecologically valid monitoring. The framework is designed to be lightweight, interpretable, and optimized for mobile deployment. Expected outcomes include improved accuracy in early detection of depression and anxiety, interpretable outputs for clinician trust, and a scalable framework extendable to additional modalities in the future.

I. INTRODUCTION

DEPRESSION and anxiety are among the most prevalent mental health disorders globally, together accounting for a substantial proportion of disability-adjusted life years (DALYs) and healthcare costs [1]. Traditional diagnostic methods rely heavily on self-reported questionnaires and clinical interviews, which are subject to recall bias, underreporting, and lack of ecological validity [2]. As a result, there is a growing interest in digital phenotyping the use of data from personal devices and sensors to provide objective, continuous, and real-time insights into mental health [3].

Recent advances demonstrate that multimodal approaches which integrate multiple behavioral and physiological signals offer superior performance in detecting and monitoring depression compared to unimodal systems [4]. Each modality contributes unique information about cognitive, affective, and behavioral states. Eye-tracking reveals attentional and emotional biases, with depressed individuals often showing reduced fixation on positive stimuli and prolonged attention to negative or threatening cues [5], [6]. Bekler and Yılmaz [7] further demonstrated the potential of eye-tracking within virtual reality for detecting depressive states through explainable AI. Similarly, facial expressions serve as non-verbal indicators of affective state, where micro-expressions and reduced emotional expressivity have been linked to both depression and anxiety [8]. Deep learning-based facial analysis models, such as convolutional neural networks (CNNs), have shown promise in automated emotion recognition [9].

Voice features, including prosody, pitch, intensity, jitter, and shimmer, are also sensitive biomarkers of depression, as individuals often display reduced pitch variability and slower speech rates. Prior research by Cao et al. [10] and Quang-Anh et al. [11] has highlighted the value of voice-based emotion detection using neural networks. Beyond these core modalities, research has also demonstrated the potential of smartphone GPS data for identifying social withdrawal and mobility patterns [12], typing dynamics and keystroke metadata for capturing psychomotor slowing and cognitive impairments [10], [13], and body/hand gestures for quantifying reduced movement, slumped posture, and lower gesture frequency in depressed individuals [14], [15].

While these modalities provide valuable insights, most existing research remains unimodal, limiting accuracy and generalizability [16]. Furthermore, although deep learning approaches have achieved strong classification performance, they often lack interpretability, which is essential for clinical adoption [17]. In the present work, this study focuses specifically on three core modalities facial expressions, eye-tracking, and voice features owing to their strong empirical foundation and feasibility for implementation. At the same time, the broader range of modalities such as GPS mobility, typing dynamics, and gestures are acknowledged as promising avenues for future work.

Therefore, this paper proposes a multimodal explainable AI system for detecting and monitoring depression and anxiety. By integrating complementary modalities, the system aims to enhance predictive accuracy, reduce false positives, and generate interpretable explanations that support clinical trust and decision-making.

II. Literature Review

Depression detection using artificial intelligence (AI) and digital phenotyping has become a rapidly growing research domain. Prior studies have investigated multiple modalities text, facial expressions, voice, smartphone usage, and eye-tracking. While these approaches achieve promising accuracy in experimental settings, their limitations in scalability, interpretability, mobile applicability, and clinical validation hinder practical adoption. In this study, the focus is primarily on three modalities—voice, face, and eyes—while other modalities are considered as future work.

A. Text-Based Approaches (Future Work)

Abid [18] applied LSTM, SVM, and Random Forest on anonymized patient health records, achieving up to 87% accuracy. The work demonstrated the potential of text mining in capturing contextual cues for diagnosis. However, the approach was restricted to structured and unstructured clinical notes, raising issues of privacy, interpretability, and limited real-world deployment. Since such systems depend on sensitive medical records and lack multimodal integration, their applicability in mobile or outpatient settings remains minimal. For these reasons, text-based methods are treated as a potential future extension of this system.

B. Facial Expression-Based Approaches

Singh and Srivastava [19] and Li et al. [20] applied CNN models on the FER-2013 dataset, reporting around 77% accuracy. These studies highlight the feasibility of detecting depression from facial cues in real time. However, CNNs are computationally intensive, limiting mobile deployment. Additionally, facial recognition accuracy is highly sensitive to environmental conditions such as lighting and camera resolution, and cultural differences in emotional expression reduce generalizability. Importantly, these models have not been validated in clinical practice, which questions their robustness outside controlled experiments. In the proposed system, facial analysis is one of the primary modalities, with emphasis on lightweight models optimized for mobile deployment.

C. Voice-Based Approaches

Sharma et al. [21] used a hybrid SVM+KNN classifier on DAIC-WOZ and MODMA datasets, achieving 82–87% accuracy. Similarly, Karlin et al. [22] employed wav2vec2 embeddings with a Longformer-based semantic model on 2007 real-world calls, achieving 81% accuracy. While voice-based methods are attractive due to their non-invasive nature and suitability for mobile applications, they remain highly sensitive to background noise and recording quality. Proprietary components also hinder transparency, while the absence of clinical trials reduces medical acceptance. Furthermore, computational demands of advanced speech models pose challenges for mobile devices with limited resources. Despite these challenges, voice analysis is central to the proposed application due to its natural and continuous ability to capture depressive cues.

D. Multilingual and Attention Mechanisms (Future Work)

Quang-Anh et al. [23] proposed Dynamic-CBAM + Attention-BiGRU on Vietnamese emotional speech, achieving 87% accuracy. Attention mechanisms improved feature selection, but the approach relied on a small, language-specific dataset. This limits cross-lingual generalizability and reduces real-world feasibility, as models would need retraining for different cultural contexts. The study also did not evaluate deployment in noisy, real-world conditions such as phone calls or mobile applications. While promising, multilingual and attention-driven methods are considered for future extensions of the system.

E. Mobile and Behavioral Data (Future Work)

Cao et al. [24] introduced DeepMood, which used RNN variants on typing dynamics and mobile sensor data. While it captured temporal dependencies effectively, the method required constant smartphone usage, raising compliance issues and bias against users with low digital literacy. Carr et al. [25] monitored circadian activity using accelerometers, but data from only 28 patients limited generalizability. Similarly, Nguyen et al. [26] and Ajilore et al. [27] demonstrated the utility of keystroke dynamics, but these studies lacked standardized protocols and clinical evaluation. Although smartphone-based systems suggest potential for mobile deployment, their success is constrained by small datasets, user compliance challenges, and the absence of multimodal fusion. For the scope of this study, these methods are highlighted as future work.

F. Eye-Tracking Approaches

Bekler and Yılmaz [28] used explainable AI (XAI) to analyze eye movements in VR environments, improving interpretability a major limitation in prior black-box models. However, their system requires VR hardware, restricting accessibility and scalability. Noyes et al. [29] highlighted through a systematic review that eye-tracking studies are heterogeneous and lack standard biomarkers for depression. Skaramagkas et al. [30] presented the large-scale eSEE-d dataset, enabling neural network applications. Nevertheless, classification success rates remain below 80%, and deployment in portable/mobile devices has not been tested. In the proposed system, eye-tracking is integrated as one of the three core modalities, with emphasis on adapting techniques for mobile-compatible platforms rather than VR environments.

G. Critical Analysis

The reviewed works show significant progress in AI-driven depression detection. However, common limitations are evident across modalities: limited clinical validation with most studies remaining at pilot or experimental stage; mobile applicability challenges as heavy models (CNNs, transformers) often lack optimization for deployment in real-time or smartphone-based systems; environmental sensitivity with performance varying with background noise, lighting, device quality, and user compliance; cultural and linguistic constraints as many approaches rely on language- or region-specific datasets, limiting global applicability; and interpretability and trust issues as deep models often function as black boxes, reducing clinician acceptance.

Given these gaps, this study focuses on three modalities—voice, face, and eyes—which offer the most practical potential for mobile deployment. Other modalities such as text, multilingual analysis, and behavioral data will be explored as part of future work to extend the system toward a more comprehensive, multimodal framework.

TABLE I
COMPARATIVE ANALYSIS OF DEPRESSION DETECTION APPROACHES

Study	Modality	Method	Accuracy	Strengths	Limitations
Abid [18]	Text	LSTM, SVM, RF	87%	Good contextual analysis	Privacy concerns, limited mobile applicability
Singh & Srivastava [19]	Facial	CNN	77%	Real-time detection feasible	Computationally intensive, environmental sensitivity
Sharma et al. [21]	Voice	SVM+KNN	82-87%	Non-invasive, mobile-friendly	Noise sensitivity, lack of clinical validation
Bekler & Yılmaz [28]	Eye-tracking	XAI in VR	<80%	Explainable, interpretable	Requires VR hardware, not scalable
Proposed System	**Multimodal**	**CNN+LSTM, MLP, SVM+KNN**	**TBD**	**Multimodal fusion, mobile-optimized, explainable**	**Requires clinical validation**

III. Proposed Methodology

This study proposes a multimodal depression detection framework that integrates facial behavior, eye movement, and vocal biomarkers into a unified mobile-based platform. By leveraging deep learning, machine learning, and smartphone sensors, the framework is designed to ensure robust and clinically relevant predictions of depression severity. The pipeline consists of data acquisition, preprocessing, feature extraction, modality-specific modeling, multimodal fusion, and evaluation, as illustrated in Fig. 1.

A. Data Acquisition

Data collection will be conducted through both publicly available datasets and real-time smartphone inputs. For facial and eye data, the DAIC-WOZ dataset (clinical interview recordings with PHQ-8 scores), MoodCapture (facial images captured in real-world settings), and Smartphone Eye-Tracking Dataset (2025) for gaze patterns will be utilized. Auxiliary datasets including OpenEDS and TEyeD will be used for pretraining eye-tracking models. For voice data, DAIC-WOZ, MODMA, and VNEMOS datasets will be employed, along with live audio recording via the smartphone microphone.

B. Preprocessing

Each modality undergoes customized preprocessing. Facial video processing includes frame extraction, face detection and alignment using Mediapipe or DLib, grayscale conversion, and histogram equalization. Eye data preprocessing involves noise reduction, fixation and saccade detection, and blink rate normalization. Voice preprocessing includes noise filtering through spectral gating, resampling, framing, MFCC extraction, and jitter/shimmer normalization.

C. Feature Extraction

Relevant features are extracted across modalities. Facial features include CNN embeddings using ResNet-50 or Inception-v3, temporal dynamics via LSTM, and micro-expressions such as lip-corner depression and gaze aversion. Eye features comprise fixation duration, saccade velocity, gaze stability, and blink frequency. Voice features include MFCCs, jitter, shimmer, harmonic-to-noise ratio (HNR), and wav2vec2 embeddings.

D. Modality-Specific Modeling

Each modality is processed using optimized models. Facial behavior analysis employs CNN combined with Attention-LSTM for spatiotemporal expression recognition. Eye movement analysis utilizes Deep MLP and Extra Trees Classifier with SHAP-based interpretability. Voice analysis incorporates Wav2vec2 embeddings with LightGBM for efficiency, Bi-GRU and Attention-BiGRU for temporal acoustic modeling, and hybrid SVM+KNN for interpretable classification.

E. Multimodal Fusion and Decision Layer

Outputs from the three modalities are fused using decision-level weighted averaging, where weights are learned based on validation performance. The final system outputs a depression severity score mapped to four levels according to PHQ-8: Normal, Mild, Moderate, and Severe. This multimodal approach ensures robustness, such that poor lighting in facial analysis can be compensated by voice or eye-based signals.

F. Evaluation Protocol

The system will be evaluated using multiple datasets including DAIC-WOZ, MODMA, VNEMOS, MoodCapture, Smartphone Eye-Tracking Dataset, OpenEDS, and TEyeD. Performance metrics will include Accuracy, Precision, Recall, F1-score, AUROC, MAE, and CCC. Validation approaches will comprise stratified k-fold cross-validation, blind test set evaluation, and ablation studies across modalities.

G. Future Extensions

Future extensions include typing pattern analysis incorporating keystroke features such as hold time, latency, and typing speed from datasets like KeyRecs, EmoSurv, and BiAffect. Additionally, mobility and behavioral features from GPS and smartphone usage patterns will provide richer depression detection capabilities. Privacy and security measures including federated learning and on-device inference will ensure user data confidentiality.

H. Ethical Considerations

The framework adheres to strict ethical guidelines. All datasets are anonymized and securely stored. The system acts as a decision-support tool, not a replacement for clinical diagnosis. User consent and transparency are mandatory in real-world deployment.

ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

Flow Diagram

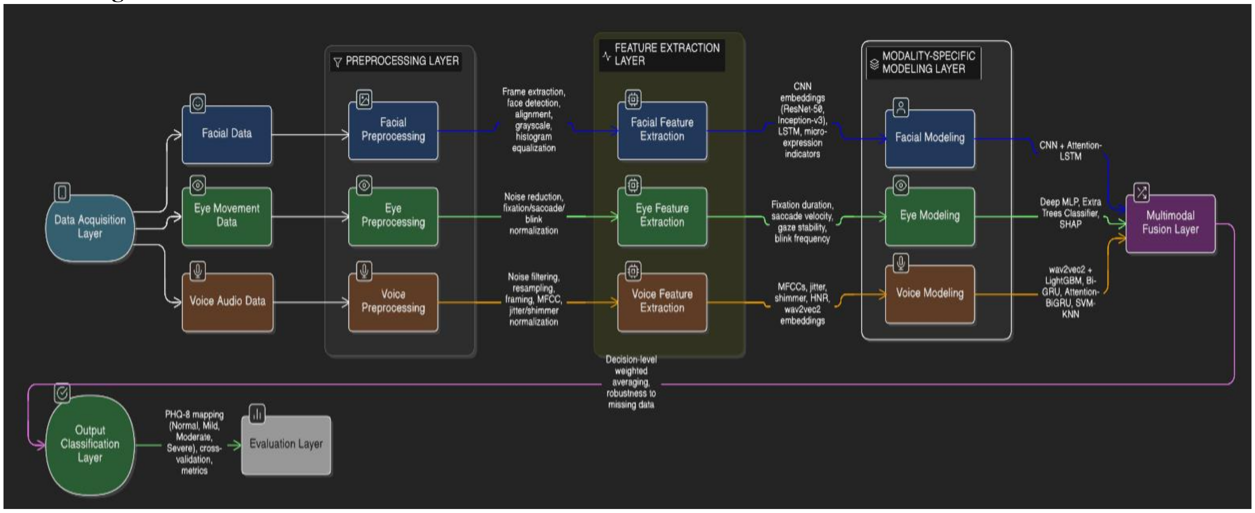

Fig. 1. Proposed multimodal framework architecture showing data flow from acquisition through preprocessing, feature extraction, modality-specific modeling, fusion, and final classification into depression severity levels.

[Note: System architecture diagram to be inserted here showing: (1) Data Acquisition Layer with three input streams (Facial Video, Eye Movement, Voice Audio), (2) Preprocessing Layer, (3) Feature Extraction Layer, (4) Modality-Specific Models Layer (CNN+LSTM, MLP+SHAP, Wav2vec2+LightGBM), (5) Fusion Layer (Weighted Averaging), and (6) Output Classification (Normal/Mild/Moderate/Severe)]

IV. Expected Outcomes

The proposed multimodal framework is expected to yield several significant outcomes. First, the development of a lightweight multimodal AI framework specifically designed for depression and anxiety detection that can operate efficiently on mobile devices. Second, improved accuracy and robustness compared to unimodal approaches through the integration of complementary data sources. Third, enhanced interpretability of predictions to support clinical trust and acceptance, addressing a critical barrier to AI adoption in healthcare settings. Fourth, the creation of a smartphone-based application capable of real-time monitoring that can be deployed in naturalistic settings. Finally, a significant contribution to scalable, cost-effective digital mental health solutions that can reach underserved populations and provide continuous monitoring capabilities beyond traditional clinical settings.

V. Conclusion

This paper presents a comprehensive multimodal AI framework for real-time depression and anxiety detection that addresses several critical gaps in existing research. By integrating facial expressions, eye-tracking, and voice biomarkers, the proposed system leverages complementary information sources to enhance detection accuracy while maintaining computational efficiency suitable for mobile deployment. The framework's emphasis on explainability through techniques such as SHAP-based interpretability addresses the critical need for clinical trust and adoption of AI-based mental health tools.

The proposed methodology incorporates rigorous preprocessing, feature extraction, and modality-specific modeling techniques drawn from state-of-the-art research, while acknowledging the limitations of existing approaches in terms of clinical validation, mobile applicability, and environmental sensitivity. The system's architecture is designed with future extensibility in mind, allowing for the incorporation of additional modalities such as typing dynamics, GPS mobility patterns, and gesture analysis as the research evolves.

While the framework presents a promising approach to automated mental health monitoring, it is important to emphasize that the system is intended as a decision-support tool to complement, rather than replace, professional clinical judgment. Future work will focus on clinical validation studies, optimization for real-world deployment conditions, and expansion to include additional behavioral and physiological modalities. Through continued development and validation, this framework has the potential to significantly contribute to accessible, continuous, and objective mental health monitoring, ultimately improving early detection and intervention for individuals experiencing depression and anxiety.

Acknowledgment

The author would like to thank the Department of Computer Science at NED University of Engineering and Technology for providing the necessary resources and support for this research.

REFERENCES

World Health Organization, "Depression and Other Common Mental Disorders: Global Health Estimates," WHO, Geneva, 2017.

O. Carr, F. Andreotti, K. E. A. Saunders, N. Palmius, G. M. Goodwin, and M. De Vos, "Monitoring depression in bipolar disorder using circadian measures from smartphone accelerometers," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 251–257, Jan. 2020.

J. Onnela and S. Rauch, "Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health," *Neuropsychopharmacology*, vol. 41, pp. 1691–1696, 2016.

S. K. Sharma, A. I. Alutaibi, and A. R. Al-Dhamari, "Early detection of mental health disorders using machine learning models based on behavioral and voice data analysis," *Healthcare Analytics*, vol. 5, 100244, 2025.

D. Armstrong and C. Olatunji, "Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis," *Clin. Psychol. Rev.*, vol. 32, no. 8, pp. 704–723, 2012.

B. Noyes, A. Biorac, G. Vazquez, S. Khalid-Khan, D. Munoz, and L. Booij, "Eye-tracking in adult depression: protocol for a systematic review and meta-analysis," *BMJ Open*, vol. 13, no. 6, e069256, Jun. 2023.

M. Bekler and M. Yılmaz, "Eye-tracking as an indicator of depressive emotional states in virtual reality: An explainable artificial intelligence approach," Gazi Univ., Nov. 2024.

J. Cohn and F. De la Torre, "Automated face analysis for affective computing," in *Handbook of Affective Computing*, Oxford Univ. Press, 2014.

Z. Zhang, P. Luo, C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3631–3639.

B. Cao, L. Zheng, C. Zhang, P. S. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. D. Leow, "DeepMood: Modeling mobile phone typing dynamics for mood detection," *Proc. 8th ACM Int. Conf. Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, Boston, MA, USA, 2017, pp. 329–336.

P. Quang-Anh, H. N. Binh, N. H. Hieu, and N. H. Phuong, "Emotional Vietnamese speech-based depression diagnosis using dynamic attention mechanism," *IEEE Access*, vol. 12, pp. 11583–11595, 2024.

N. Saeb, M. Zhang, C. Karr, S. Schueller, M. Corden, K. Kording, and D. Mohr, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," *J. Med. Internet Res.*, vol. 17, no. 7, e175, Jul. 2015.

M. J. Hoptman et al., "Assessment of cognitive function in bipolar disorder with passive smartphone keystroke metadata: a BiAffect digital phenotyping study," *Front. Psychiatry*, vol. 16, 2025, doi: 10.3389/fpsyt.2025.1430303.

R. Joshi, A. Patel, and M. Goyal, "Automated detection of depression using body movement and gesture analysis," in *Proc. Int. Conf. Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2021.

A. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, Feb. 2013.

H. Baird et al., "Multimodal mental health assessment: Current advances and future directions," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 25–40, 2024.

A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, Jul. 2019.

N. Abid, "Intelligent Depression Pattern Identification Using Artificial Intelligence in the Patient Health Records," *AlgoVista: Journal of AI & Computer Science*, vol. 2, no. 1, pp. 37–47, 2025.

S. Singh and N. Srivastava, "Face Emotion Recognition System for Depression Detection Using AI Techniques," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 12, no. 1, pp. 175–182, Jan. 2024. doi: 10.22214/ijraset.2024.57890.

Y. Li, J. Zhang, Y. Zhang, and H. Li, "Automatic Identification of Depression Using Facial Images with Deep Convolutional Neural Network," *Medical Science Monitor*, vol. 28, e937004, pp. 1–10, Jul. 2022. doi: 10.12659/MSM.937004.

S. K. Sharma, A. I. Alutaibi, A. R. Khan, G. G. Tejani, F. Ahmad, and S. J. Mousavirad, "Early detection of mental health disorders using machine learning models using behavioral and voice data analysis," *Scientific Reports*, vol. 15, Art. no. 16518, May 2025. doi: 10.1038/s41598-025-00386-8.

B. Karlin, D. Henry, R. Anderson, S. Cieri, M. Aratow, E. Shriberg, and M. Hoy, "Digital Phenotyping for Detecting Depression Severity in a Large Payor-Provider System: Retrospective Study of Speech and Language Model Performance," *JMIR AI*, vol. 4, Art. no. e69149, Jun. 19 2025, doi: 10.2196/69149.

Q.-A. N. D., M.-H. H., T. K. D., M.-D. P., and N. N. V., "Emotional Vietnamese Speech-Based Depression Diagnosis Using Dynamic Attention Mechanism," *arXiv preprint arXiv:2412.08683*, Dec. 2024.

B. Cao, L. Zheng, C. Zhang, P. S. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. D. Leow, "DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection," *arXiv preprint arXiv:1803.08986*, Mar. 2018.

O. Carr, D. D. Zulueta, K. Ryan, O. Ajilore, and A. D. Leow, "Monitoring depression in bipolar disorder using circadian measures from smartphone accelerometers," *JMIR Mental Health*, vol. 5, no. 1, p. e15, 2018.

T. M. Nguyen, A. D. Leow, and O. Ajilore, "A review on smartphone keystroke dynamics as a digital biomarker for understanding neurocognitive functioning," *Brain Sciences*, vol. 13, no. 6, p. 959, Jun. 2023.

O. Ajilore et al., "Assessment of cognitive function in bipolar disorder with passive smartphone keystroke metadata: a BiAffect digital phenotyping study," *Frontiers in Psychiatry*, vol. 16, p. 1430303, Mar. 2025. doi: 10.3389/fpsyt.2025.1430303.

M. Bekler and M. Yılmaz, "Eye-Tracking as an Indicator of Depressive Emotional States in Virtual Reality: An Explainable Artificial Intelligence Approach," Gazi University, Ankara, Türkiye, received Nov. 11, 2024, accepted Nov. 16, 2024.

B. Noyes, A. Biorac, G. Vazquez, S. Khalid-Khan, D. Munoz, and L. Booij, "Eye-tracking in adult depression: protocol for a systematic review and meta-analysis," *BMJ Open*, vol. 13, no. 6, p. e069256, Jun. 2023, doi: 10.1136/bmjopen-2022-069256.

V. Skaramagkas, E. Ktistakis, D. Manousos, E. Kazantzaki, N. S. Tachos, E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "eSEE-d: Emotional state estimation based on eye-tracking dataset," *Brain Sciences*, vol. 13, no. 4, art. 589, Mar. 2023.