Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026233243

Machine Learning Research on Depression Assessment: A Narrative Review of Physiological, Behavioral, and Multimodal Approaches

Xin Yang*

Shanghai Maritime University, Shanghai, China

**Corresponding author: Xin Yang.*

Abstract

Depression has become an important target of machine learning research because conventional assessment often depends on intensive clinician-patient interaction and may be influenced by subjectivity. This narrative review examines recent studies on depression assessment from three major perspectives: physiological signals, behavioral signals, and multimodal fusion. After outlining the conceptual background of depression detection and prediction—including common datasets, methodological workflows, and evaluation issues—it compares representative studies using electroencephalography, heart-rate-related signals, speech, facial behavior, text, and combined modalities, highlighting how different data modalities capture complementary aspects of depressive states. Across these domains, both traditional machine learning and deep learning methods have shown promising performance, with reported within-dataset accuracies often exceeding 90% and AUCs above 0.9 in some EEG and ECG studies. However, performance frequently drops sharply in cross-site or multi-cohort settings—falling to around 52% balanced accuracy after harmonization in one large multi-site MRI benchmark—and major challenges remain in sample size, label quality, external validation, interpretability, fairness, and clinical deployment. Multimodal strategies appear especially promising, though they also introduce greater complexity in data collection and integration. Overall, future progress will depend not only on improving predictive accuracy but also on building reliable, explainable, and clinically usable systems for real-world mental health care.

Keywords

depression assessment, machine learning, physiological signals, behavioral data, multimodal fusion

1. Introduction

Depression is one of the most important mental health challenges worldwide and continues to impose a substantial disease burden across populations [1]. Because of its impact on functioning, quality of life, and long-term well-being, early identification of depression remains a major clinical and public health priority [1]. In practice, however, conventional assessment still relies largely on clinical interviews and self-report rating scales. Although these tools are essential in psychiatric settings, they may be limited by subjectivity, time cost, and unequal access to mental health professionals [2]. These limitations have spurred growing interest in machine learning as a means to support more objective, scalable, and data-driven approaches to depression assessment [2,3].

Recent studies have applied machine learning to a wide range of data modalities related to depression. These include physiological signals such as electroencephalography and electrocardiography, as well as behavioral signals derived from speech, facial expressions, body movements, and text [3,4]. This growing literature suggests that depression is not only reflected in clinical symptoms, but is also associated with measurable biological and behavioral patterns that can potentially be captured computationally [2,3]. At the same time, machine learning studies in this field address different goals. Some focus on detection, namely identifying whether an individual currently shows depression-related characteristics, whereas others aim at prediction, such as estimating future risk, symptom trajectory, or treatment response [3,5]. However, the field is still dominated by detection-oriented classification studies, while prediction research remains less mature and often requires longitudinal data and stronger validation [3,5].

Methodologically, this body of research draws on both traditional machine learning algorithms—such as logistic regression, support vector machines, random forests, and gradient boosting—and deep learning approaches that learn representations directly from raw signals, images, audio, or text [2,3,10]. Performance is typically reported using accuracy, precision, recall, F1 score, and area under the curve. Still, these numbers must be interpreted cautiously, given the small, imbalanced, and site-specific nature of many depression datasets [3,8,10]. Notably, although many studies are labeled as "prediction," they often in fact perform cross-sectional detection, while genuinely prospective work using longitudinal data to identify depression risk before clinical manifestation remains rare [3,11].

A related but often under-discussed issue concerns where the "depression" labels used to train these models actually come from. Across studies, depression has been operationalized in highly heterogeneous ways—formal DSM-5 diagnosis, structured clinical interviews, self-report scales such as the PHQ-9 or BDI, hospital records, or even self-identification on social media—and these sources differ substantially in reliability. A high PHQ-9 score, for example, is not equivalent to a clinician-confirmed diagnosis of major depressive disorder, yet both are often used interchangeably as ground truth in the literature. This matters because no matter how sophisticated a model is, its outputs can only be as reliable as the labels it is trained on. The ambiguity in what counts as "depression" across datasets fundamentally limits how confidently reported performance can be compared or generalized [3,8].

Another notable trend is the growing use of deep learning and multimodal fusion. Reviews of recent studies indicate that integrating multiple sources of information may improve robustness and better reflect the complexity of depression than relying on a single modality alone [2,3]. In addition, systematic reviews on speech-based and text-based approaches show that behavioral data can provide useful and relatively scalable signals for depression-related assessment [6,7]. Nevertheless, promising results in research settings do not automatically translate into clinical usefulness. Persistent issues such as small sample sizes, dataset bias, limited external validation, insufficient interpretability, and uncertainty in real-world deployment continue to restrict progress toward clinical application [5,8].

Against this background, this paper presents a narrative review of recent machine learning research on depression. It first introduces the basic conceptual background of depression detection and prediction, then reviews the literature in three areas: physiological-signal-based studies, behavioral-signal-based studies, and multimodal fusion. Finally, it discusses the key methodological and clinical challenges that currently shape future research in this field [3,5,8].

2. Literature Review

This section reviews recent machine learning studies on depression assessment from three main perspectives [3,9]. First, it examines physiological-signal-based approaches, including EEG, ECG, MRI, and fNIRS, which aim to capture biological correlates of depression. Second, it discusses behavioral-signal-based studies, such as those using speech, facial expression, body movement, and text, which are generally more accessible and scalable for large-scale screening. Third, it reviews multimodal fusion approaches that combine multiple information sources to improve robustness and better reflect the complexity of depressive symptoms [33]. Across these three strands, the review compares common modeling strategies, reported performance, and the major methodological challenges that continue to limit clinical translation [3,9,33]. Although some studies address prediction and longitudinal outcomes, most of the existing literature still focuses primarily on detection and assessment in cross-sectional settings [3,5].

2.1 Physiological-Signal-Based Approaches to Depression Assessment

Physiological signals have become an important data source in machine-learning research on depression because they provide relatively direct information about neural and autonomic processes associated with affective dysfunction [13]. Among these modalities, electroencephalography (EEG) has received particular attention due to its low cost, portability, and high temporal resolution [13]. Recent EEG studies also show a clear methodological shift from handcrafted features toward deep neural architectures trained on raw signals or transformed time–frequency representations [14].

Recent EEG studies illustrate both methodological evolution and a recurring problem of generalization. Algorithmically, the field has shifted from signal-processing pipelines toward biologically motivated architectures and richer input designs: Bagherzadeh et al. paired the synchrosqueezed wavelet transform with transfer learning to obtain very high within-dataset accuracy, though their cross-database performance dropped considerably [15]; Shen et al.'s HEMAsNet instead embeds inter-hemispheric asymmetry directly into the network architecture [14]; and Noda et al. expanded the input space by combining resting-state EEG with TMS-evoked EEG, reaching a mean AUC of 0.922 when multiple feature sets were integrated [13]. Acquisition has likewise diversified: Suzuki et al. showed that even consumer-grade EEG could support depression detection (Macro-F1 91.59%), suggesting feasibility outside specialized labs [16]. Together, these advances illustrate that EEG research has become both methodologically diverse and more portable—but Bagherzadeh's cross-database drop foreshadows a broader concern that within-dataset performance does not necessarily translate to robust generalization.

Compared with EEG, electrocardiography (ECG) has been studied less frequently in depression research. Still, it remains clinically attractive because autonomic dysregulation is closely related to depressive states, and ECG acquisition is relatively simple [17]. Habib et al. proposed MDDBranchNet, a deep learning model for detecting major depressive disorder from ECG signals, achieving an accuracy of up to 95.4% [17]. This suggests that cardiac signals serve as useful complementary biomarkers, especially in wearable or low-burden monitoring scenarios [17]. However, the current ECG literature is still much smaller than the EEG literature, which means its clinical value remains less well established at present [17].

Neuroimaging-based approaches, especially those using MRI, aim to capture distributed brain-network abnormalities associated with depression [18]. Li et al. conducted a two-center radiomic study based on multisequence MRI features and reported AUC values ranging from 0.794 to 0.866 across models [18]. Pilmeyer et al. extended this line of work by integrating T1-weighted MRI, resting-state fMRI, and diffusion tensor imaging, and their multimodal ensemble achieved AUCs of 0.746 for diagnosis and 0.932 for six-month outcome prediction [19]. However, performance drops sharply when models are tested on larger, more heterogeneous samples. In a large multi-site benchmark study including 5,365 participants, Belov et al. reported balanced accuracy of only about 62%, which further declined to about 52% after ComBat harmonization [20]. This pattern reflects well-documented sources of multi-site MRI variability—differences in scanner vendors, magnetic field strength, acquisition sequences (such as TR, TE, and voxel size), and preprocessing pipelines across centers. That further drop suggests that some of the apparent "signal" within-site analyses may have reflected site-specific confounds rather than depression-related neurobiology, and that aggressive harmonization may strip away biologically meaningful variance along with the noise [20]. These results indicate that MRI-based machine-learning models may detect biologically meaningful patterns, but cross-site generalizability remains a major challenge [20].

Functional near-infrared spectroscopy (fNIRS) has drawn growing interest as a more portable, lower-cost alternative to MRI for capturing cortical hemodynamic responses [21]. Early fNIRS-based depression work focused on extracting handcrafted biomarkers from cortical signals—Li et al. showed the feasibility of this approach [21], and Mao et al. extended it to a verbal-fluency task [23]—while more recent work designs deep models around known neurobiology: Lee et al., for example, built a channel-embedding layer that encodes inter-hemispheric asymmetry in prefrontal responses [22]. A separate line of work has begun to look beyond diagnosis: Ho et al. used baseline fNIRS with clinical variables to predict six-month treatment response, an early example of fNIRS applied to longitudinal clinical outcomes rather than cross-sectional classification [24].

Overall, recent studies of physiological signals show that deep learning has become increasingly dominant across EEG, ECG, MRI, and fNIRS research [14]. At the same time, the strongest reported results still tend to come from relatively small or internally validated datasets [15]. By contrast, large multi-site evidence shows that robust generalization remains difficult to achieve [20]. For this reason, the future clinical value of physiological-signal-based models will depend not only on higher accuracy but also on standardized preprocessing, external validation, and improved interpretability [20].

2.2 Behavioral Signals for Depression Assessment

Behavioral-signal research examines whether depression-related information can be inferred from how people speak, write, and behave in observable settings. Compared with physiological modalities, behavioral data are often easier to collect using microphones, cameras, smartphones, and online platforms, making them attractive for scalable assessment outside highly specialized clinical environments [6,29].

Speech has received sustained attention among behavioral modalities. A recent meta-analysis (Maran et al.) found that speech-based methods are promising but vary substantially across studies in recording protocols, feature sets, and modeling pipelines [6]. Two directions extend this work: broadening the population—Wei et al. showed that acoustic-based machine learning supports detection in Chinese university students, a younger and non-Western sample [25]—and broadening what speech is taken to measure, with Wiseman et al. reporting that objective speech measures capture not only depressive symptoms but also associated cognitive difficulties, suggesting speech may reflect clinically meaningful variation rather than a binary distinction [26]. Speech is, therefore, a practical, low-burden signal source, although performance depends on language, speaking task, and recording quality [6].

Visual and other nonverbal behavioral cues form another important branch of the literature. Ozturk et al. showed that digitally assessed nonverbal behaviors could forecast the first onset of depression, which is especially notable because it moves beyond simple cross-sectional recognition [27]. In facial analysis research, Liu et al. proposed a stimulus-response pattern framework for robust cross-stimulus facial depression recognition, directly addressing the fact that facial markers may shift across elicitation contexts [28]. Together, these studies suggest that facial expressions, head movements, posture, and related nonverbal signals contain useful information. Still, they are highly context-sensitive and can be influenced by camera conditions, task design, and individual expressive style [27,28].

Text-based approaches form another major direction, especially using social media or open-ended self-report data. Recent work has extended the basic paradigm of binary classification on social media in several directions. Kerasiotis et al. enriched the input side by combining transformer models with auxiliary metadata and linguistic markers, substantially improving weighted F1 over a transformer-only baseline for depression detection [29]. Qasim et al. enriched the output side, targeting depression severity rather than binary presence and indicating that NLP work is moving toward fine-grained assessment [30]. Chung et al. shifted the data source itself, applying BERT and BERTopic to open-ended mobile-app messages from older adults and extending text-based analysis from anonymous online posts into a more clinically oriented, age-specific context [32].

At the same time, behavioral text models raise important concerns about fairness and transferability. Rai et al. demonstrated that key language markers of depression on social media depended on race. Language-based models were considerably more predictive for White participants than for Black participants [31]. This finding is important for the wider behavioral-signal literature because it warns against assuming that a model trained on one population or platform will generalize cleanly to another. Overall, behavioral approaches offer clear advantages in cost, accessibility, and repeated measurement, but their future clinical value will depend on stronger external validation, better bias analysis, and more interpretable modeling strategies [6,31].

2.3 Multimodal Fusion Approaches

Multimodal fusion has become one of the most active directions in machine-learning-based depression assessment, since depressive symptoms are often expressed across several channels simultaneously: language, speech prosody, facial behavior, and sometimes physiological signals. A recent meta-analysis confirmed that multimodal approaches generally outperform unimodal systems for screening, though with substantial heterogeneity across datasets, labels, and evaluation protocols [33].

Recent work has focused on combining text, audio, and video collected from clinical interviews or structured assessments. The target task itself has broadened: Zhang et al. integrated text, audio, and video for binary detection, showing that complementary cues from different modalities can be combined within a single pipeline [34], while Sadeghi et al. extended this approach to severity estimation, suggesting that different modalities may contribute differently depending on the target task [35].

A second methodological trend is the shift from simple feature concatenation to more structured fusion strategies. For example, Chen et al.'s IIFDD framework separates intra-modal fusion (within each modality) from inter-modal fusion (across modalities) using attention-based mechanisms, reflecting the principle that each modality must be adequately represented before cross-modal interaction begins [36]. Yang et al. pursue a related goal through representation learning and knowledge transfer, arguing that better latent alignment across heterogeneous modalities improves diagnostic robustness [37].

More recent work draws on pretrained foundation models and increasingly knowledge-informed designs. Xu et al. built a voice–text fusion system on Wav2Vec 2.0 and BERT, outperforming unimodal baselines on both classification and severity estimation [38]. Yang et al. go further with a hierarchical prompt-fusion model that combines psychological knowledge with text, audio, and visual data, moving multimodal fusion beyond purely technical integration toward knowledge-informed modeling [39].

An additional clinical limitation of multimodal fusion is the absence of modalities at the time of use. Most current models implicitly assume that all input channels are available for every patient, but this assumption is difficult to sustain in real settings. A patient may refuse video recording, audio may be degraded by background noise, or a particular sensor may be unavailable. Missingness may also be clinically meaningful rather than random, since patients who avoid certain forms of data collection may differ in important ways from those who do not. However, this issue has received limited attention in depression-focused multimodal research. Although the broader multimodal-learning literature has explored approaches such as modality-dropout training and adaptive fusion, these strategies are still rarely discussed in studies of depression assessment.

Despite these advances, the clinical value of multimodal fusion remains unestablished. Weber et al. showed that depression diagnosis from patient interviews using multimodal machine learning is feasible in naturalistic psychiatric settings, but also highlight a central limitation: current models often rely on highly specific interview formats, small samples, and institution-bound data distributions [40]. Multimodal fusion remains promising because it better matches the multifaceted nature of depression than single-modality modeling, but larger real-world datasets, external validation, interpretability, and clinically realistic collection protocols are still needed before such systems can be deployed as dependable screening tools [33,40].

3. Discussion

Recent studies suggest that physiological, behavioral, and multimodal approaches should not be treated as mutually exclusive options, but rather as different points on a continuum of trade-offs [17,33]. Physiological signals are often closer to the neurobiological correlates of depression, while behavioral signals are cheaper, less invasive, and easier to collect at scale [13,29]. Multimodal fusion is attractive because it may combine biological specificity with ecological validity [33,35]. From this perspective, the most important question is no longer which single modality is “best,” but which kinds of models can remain reliable across heterogeneous datasets, devices, and clinical environments [20,40].

A second synthesis point concerns the interpretation of model performance. In this field, high accuracy on a curated dataset does not automatically imply clinical usefulness [20,40]. Recent work on explainable machine learning for depression detection shows that interpretability is becoming a major concern, but it is still an emerging rather than a mature solution [12,41]. In practice, explainability is valuable only if it helps clinicians understand whether a model is using psychologically and medically plausible signals, whether its errors are systematic, and whether its outputs can support individual-level judgment instead of functioning as a black box [12,41].

A related translational concern is that many of the most technically impressive models rely on data modalities such as EEG, fNIRS, or MRI that are still far from routine use in ordinary depression assessment workflows. In everyday clinical practice, depression assessment continues to depend largely on symptom-based evaluation, clinical interviews, and validated questionnaires rather than physiological or neuroimaging markers. In many cases, the entire diagnostic encounter consists of a brief clinical interview and a self-report scale. One of them is the PHQ-9—a process that bears little resemblance to the multimodal data pipelines proposed in the literature. This means that even when models achieve strong performance on research datasets, their relevance to real clinical decision-making may remain limited. From this perspective, the central challenge is not simply to report higher AUCs or F1 scores, but to explain why these methods have not yet reshaped routine mental health care and what kinds of evidence would be required for them to do so.

Another key issue is the instability introduced by heterogeneous labels and data-collection protocols. Across the literature, depression may be operationalized through formal diagnosis, scale-based thresholds, psychiatric samples, or self-reported symptoms, and these targets are not fully equivalent [3,8]. At the same time, differences in sensors, preprocessing pipelines, feature engineering, and recording conditions make direct comparison across studies difficult [3,8,20]. This helps explain why strong results in one dataset often do not transfer smoothly to another, especially when models are moved from research settings to broader clinical populations [20].

Fairness becomes even more important once models move closer to implementation. A recent study across four populations showed that standard machine-learning models for depression prediction can exhibit bias, and that debiasing methods can reduce these inequalities but not eliminate them [42]. This is especially important in mental health, where training data may already reflect unequal access to care, diagnostic bias, and differences in help-seeking behavior [42]. As a result, a model that appears accurate overall may still perform inadequately for specific demographic or social groups unless fairness is evaluated explicitly [42].

Clinical translation also depends on whether patients and professionals are willing to use AI-assisted systems in practice. Survey evidence indicates that patients often recognize potential benefits of AI in mental health care, but concerns about privacy, confidentiality, bias, and loss of human connection balance this interest [43]. Likewise, health professionals have expressed cautious optimism toward passive sensing, AI, and machine learning, while also raising ethical, practical, and therapeutic concerns [44]. These findings suggest that adoption will require more than technical accuracy; it will also depend on transparency, human oversight, and careful integration into clinical workflows [43,44].

This point is particularly relevant for digital phenotyping, which is often presented as a promising approach to continuous, ecologically valid depression assessment. Recent reviews indeed emphasize the broad clinical potential of digital phenotyping, but they also stress that meaningful impact depends on standardized sensor-based data collection, interoperable protocols, and stronger evidence from real-world deployment [45,46]. In other words, the field is moving from proof-of-concept research toward implementation science [45,46]. Future work should therefore focus not only on improving algorithms, but also on multicenter validation, prospective evaluation, fairness auditing, and interpretable outputs that genuinely fit mental health care pathways [41,42,45].

These limitations also suggest several technical directions for future work. Small sample sizes and strict privacy constraints, especially in clinical-interview and physiological-signal datasets, make federated learning worth exploring, since it may allow collaboration across hospitals or research sites without requiring sensitive patient data to be centrally pooled. The distribution shifts observed in multi-site MRI studies and cross-population speech data also indicate a need for methods such as domain adaptation or domain generalization, which are designed to improve robustness when training and deployment conditions differ. For the heterogeneous, sometimes imperfect labels discussed above, weak supervision and noisy-label learning may help mitigate the impact of annotation inconsistency. These approaches are not complete solutions, but they point to a more methodologically grounded research agenda than simply pursuing higher accuracy on curated benchmarks.

Overall, the literature reviewed in this paper suggests that machine learning for depression assessment has entered a more demanding stage of development [47]. The central challenge is no longer to demonstrate that computational models can detect depression-related patterns, but to determine whether such systems can become robust, fair, interpretable, and clinically acceptable in real settings [42,47]. This shift from technical feasibility to responsible implementation should shape the next phase of research in the field [47].

4. Conclusion

In conclusion, machine learning has become an important direction in depression assessment because it offers new ways to analyze complex physiological, behavioral, and multimodal data. The literature reviewed in this paper suggests that these approaches can support more objective and scalable assessments than conventional methods alone, especially when they are used to complement rather than replace clinical judgment.

This review has shown that different data modalities provide different kinds of value. Physiological signals may be more closely related to neurobiological processes, whereas behavioral signals are often easier to collect and more suitable for large-scale or repeated assessment. Multimodal fusion appears especially promising because it can integrate complementary information from different sources, but its practical use is still limited by data quality, external validation, interpretability, and deployment complexity.

Future progress in this field will depend less on reporting high performance on isolated datasets and more on developing robust, explainable, fair, and clinically usable systems. In particular, stronger longitudinal designs, more standardized data collection, better cross-population validation, and closer integration with real mental health care settings will be essential if machine learning is to make a meaningful contribution to depression assessment in practice.

References

[1] Liu, J., Ning, W., Zhang, N., Zhu, B., & Mao, Y. (2024). Estimation of the global disease burden of depression and anxiety between 1990 and 2044: An analysis of the Global Burden of Disease Study 2019. Healthcare, 12(17), 1721. https://doi.org/10.3390/healthcare12171721

[2] Song, S., Huo, Y., Tang, S., Cheong, J., Gao, R., Valstar, M. F., & Gunes, H. (2025). Automatic Depression Assessment Using Machine Learning: A Comprehensive Survey [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.18915

[3] Li, Q., Liu, X., Hu, X., Ahad, M. A. R., Ren, M., Yao, L., & Huang, Y. (2025). Machine learning-based prediction of depressive disorders via various data modalities: A survey. IEEE/CAA Journal of Automatica Sinica, 12(7), 1320–1349. https://doi.org/10.1109/JAS.2025.125393

[4] Chen, Y., Zhao, W., Yi, S., & Liu, J. (2023). The diagnostic performance of machine learning based on resting-state functional magnetic resonance imaging data for major depressive disorders: A systematic review and meta-analysis. Frontiers in Neuroscience, 17, Article 1174080. https://doi.org/10.3389/fnins.2023.1174080

[5] Ntam, V. A., Huebner, T., Steffens, M., & Scholl, C. (2025). Machine learning approaches in the therapeutic outcome prediction in major depressive disorder: A systematic review. Frontiers in Psychiatry, 16, Article 1588963. https://doi.org/10.3389/fpsyt.2025.1588963

[6] Maran, P. L., Braquehais, M. D., Vlaic, A., Alonzo-Castillo, M. T., Vendrell-Serres, J., Ramos-Quiroga, J. A., & Rodríguez-Urrutia, A. (2025). Performance of automatic speech analysis in detecting depression: Systematic review and meta-analysis. JMIR Mental Health, 12, e67802. https://doi.org/10.2196/67802

[7] Phiri, D., Makowa, F., Amelia, V. L., Phiri, Y. V. A., Dlamini, L. P., & Chung, M.-H. (2025). Text-based depression prediction on social media using machine learning: Systematic review and meta-analysis. Journal of Medical Internet Research, 27, e59002. https://doi.org/10.2196/59002

[8] Cao, Y., Dai, J., Wang, Z., Zhang, Y., Shen, X., Liu, Y., & Tian, Y. (2025). Machine learning approaches for depression detection on social media: A systematic review of biases and methodological challenges. Journal of Behavioral Data Science, 5(1), 67–102. https://doi.org/10.35566/jbds/caoyc

[9] Ren, W., Xue, X., Liu, L., & Huang, J. (2025). AI applications in depression detection and diagnosis: Bibliometric and visual analysis of trends and future directions. JMIR Mental Health, 12, e79293. https://doi.org/10.2196/79293

[10] Ghorbankhani, M., & Safara, M. (2025). Artificial intelligence in depression diagnostics: A systematic review of methodologies and clinical applications. Artificial Intelligence in Medicine, 172, 103320. https://doi.org/10.1016/j.artmed.2025.103320

[11] Zhang, A., & Zhang, H. (2025). Predicting depression in healthy young adults: A machine learning approach using longitudinal neuroimaging data. NeuroImage, 315, 121285. https://doi.org/10.1016/j.neuroimage.2025.121285

[12] Ahmed, I., Brahmacharimayum, A., Ali, R. H., Khan, T. A., & Ahmad, M. O. (2025). Explainable AI for depression detection and severity classification from activity data: Development and evaluation study of an interpretable framework. JMIR Mental Health, 12, e72038. https://doi.org/10.2196/72038

[13] Noda, Y., Sakaue, K., Wada, M., Takano, M., & Nakajima, S. (2024). Development of artificial intelligence for determining major depressive disorder based on resting-state EEG and single-pulse transcranial magnetic stimulation-evoked EEG indices. Journal of Personalized Medicine, 14(1), 101. https://doi.org/10.3390/jpm14010101

[14] Shen, J., Li, K., Liang, H., Zhao, Z., Ma, Y., Wu, J., Zhang, J., Zhang, Y., & Hu, B. (2024). HEMAsNet: A hemisphere asymmetry network inspired by the brain for depression recognition from electroencephalogram signals. IEEE Journal of Biomedical and Health Informatics, 28(9), 5247–5259. https://doi.org/10.1109/JBHI.2024.3404664

[15] Bagherzadeh, S., Norouzi, M. R., Ghasri, A., Tolou Kouroshi, P., Bahri Hampa, S., Farokhshad, F., & Shalbaf, A. (2025). Automated depression detection via cloud-based EEG analysis with transfer learning and synchrosqueezed wavelet transform. Scientific Reports, 15, 18008. https://doi.org/10.1038/s41598-025-02452-7

[16] Suzuki, K., Laohakangvalvit, T., & Sugaya, M. (2024). Machine-Learning-Based Depression Detection Model from Electroencephalograph (EEG) Data Obtained by Consumer-Grade EEG Device. Brain Sciences, 14(11), 1107. https://doi.org/10.3390/brainsci14111107

[17] Habib, M. A., Vaniya, S. N., Khandoker, A. H., & Karmakar, C. (2024). MDDBranchNet: A Deep Learning Model for Detecting Major Depressive Disorder Using ECG Signal. IEEE Journal of Biomedical and Health Informatics, 28(7), 3798–3809. https://doi.org/10.1109/JBHI.2024.3390847

[18] Li, Q., Dong, F., Gai, Q., Che, K., Ma, H., Zhao, F., Chu, T., Mao, N., & Wang, P. (2023). Diagnosis of major depressive disorder using machine learning based on multisequence MRI neuroimaging features. Journal of Magnetic Resonance Imaging, 58(5), 1420–1430. https://doi.org/10.1002/jmri.28650

[19] Pilmeyer, J., Lamerichs, R., Schielen, S., Ramsaransing, F., van Kranen-Mastenbroek, V., Jansen, J. F. A., Breeuwer, M., & Zinger, S. (2024). Multi-modal MRI for objective diagnosis and outcome prediction in depression. NeuroImage: Clinical, 44, 103682. https://doi.org/10.1016/j.nicl.2024.103682

[20] Belov, V., Erwin-Grabner, T., Aghajani, M., et al. (2024). Multi-site benchmark classification of major depressive disorder using machine learning on cortical and subcortical measures. Scientific Reports, 14(1), 1084. https://doi.org/10.1038/s41598-023-47934-8

[21] Li, Z., McIntyre, R. S., Husain, S. F., Ho, R., Tran, B. X., Nguyen, H. T., Soo, S.-C., Ho, C. S., & Chen, N. (2022). Identifying neuroimaging biomarkers of major depressive disorder from cortical hemodynamic responses using machine learning approaches. EBioMedicine, 79, 104027. https://doi.org/10.1016/j.ebiom.2022.104027

[22] Lee, K., Kwon, J., Chun, M., Choi, J., Lee, S.-H., & Im, C.-H. (2024). Functional Near-Infrared Spectroscopy-Based Computer-Aided Diagnosis of Major Depressive Disorder Using Convolutional Neural Network with a New Channel Embedding Layer Considering Inter-Hemispheric Asymmetry in Prefrontal Hemodynamic Responses. Depression and Anxiety, 2024, 4459867. https://doi.org/10.1155/2024/4459867

[23] Mao, L., Hong, X., & Hu, M. (2024). Identifying neuroimaging biomarkers in major depressive disorder using machine learning algorithms and functional near-infrared spectroscopy during verbal fluency task. Journal of Affective Disorders, 365, 9–20. https://doi.org/10.1016/j.jad.2024.08.082

[24] Ho, C. S. H., Wang, J., Ho, R., Lin, H., Li, Z., & Chen, N. (2025). Application of functional near-infrared spectroscopy and machine learning to predict treatment response after six months in major depressive disorder. Translational Psychiatry, 15(1), 7. https://doi.org/10.1038/s41398-025-03224-7

[25] Wei, Y., Qin, S., Liu, F., Liu, R., Zhou, Y., Chen, Y., Xiong, X., Zheng, W., Ji, G., Meng, Y., Wang, F., & Zhang, R. (2025). Acoustic-based machine learning approaches for depression detection in Chinese university students. Frontiers in Public Health, 13, 1561332. https://doi.org/10.3389/fpubh.2025.1561332

[26] Wiseman, M., Yep, R., Wood Alexander, M., Pople, C. B., Perri, L., Gopinath, G., Vasileiadi, M., Robin, J., Spilka, M. J., Simpson, W., Yunusova, Y., Munoz, D. P., Coe, B. C., Brien, D. C., Nestor, S. M., Lipsman, N., Giacobbe, P., & Rabin, J. S. (2025). Objective speech measures capture depressive symptoms and associated cognitive difficulties. Translational Psychiatry, 15, 525. https://doi.org/10.1038/s41398-025-03728-2

[27] Ozturk, S., Feltman, S., Klein, D. N., Kotov, R., & Mohanty, A. (2024). Digital assessment of nonverbal behaviors forecasts first onset of depression. Psychological Medicine, 54(12), 3507–3518. https://doi.org/10.1017/S0033291724002010

[28] Liu, Z., Zhang, S., Chen, B., Li, G., Chen, Q., Ding, Z., Zhang, X., & Hu, B. (2025). Stimulus-response pattern: The core of robust cross-stimulus facial depression recognition. IEEE Transactions on Affective Computing, 16(2), 1146–1158. https://doi.org/10.1109/TAFFC.2024.3496524

[29] Kerasiotis, M., Ilias, L., & Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. Social Network Analysis and Mining, 14, 196. https://doi.org/10.1007/s13278-024-01360-4

[30] Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of depression severity in social media text using transformer-based models. Information, 16(2), 114. https://doi.org/10.3390/info16020114

[31] Rai, S., Stade, E. C., Giorgi, S., Francisco, A., Ungar, L. H., Curtis, B., & Guntuku, S. C. (2024). Key language markers of depression on social media depend on race. Proceedings of the National Academy of Sciences of the United States of America, 121(14), e2319837121. https://doi.org/10.1073/pnas.2319837121

[32] Chung, M.-K., Lee, S. Y., Shin, T., Park, J. Y., Hwang, S., Kim, M.-H., et al. (2025). BERT and BERTopic for screening clinical depression on open-ended text messages collected through a mobile application from older adults. BMC Public Health, 25(1), 1. https://doi.org/10.1186/s12889-025-23337-4

[33] Wang, L., Wang, C., Li, C., et al. (2025). AI-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis. npj Digital Medicine, 8, 523. https://doi.org/10.1038/s41746-025-01933-3

[34] Zhang, W., Mao, K., & Chen, J. (2024). A multimodal approach for detection and assessment of depression using text, audio and video. Phenomics, 4(3), 234–249. https://doi.org/10.1007/s43657-023-00152-8

[35] Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., Berking, M., & Eskofier, B. M. (2024). Harnessing multimodal approaches for depression screening and diagnosis. npj Mental Health Research. https://doi.org/10.1038/s44184-024-00112-8

[36] Chen, J., Hu, Y., Lai, Q., Wang, W., Chen, J., Liu, H., Srivastava, G., Bashir, A. K., & Hu, X. (2024). IIFDD: Intra and inter-modal fusion for depression detection with multi-modal information from Internet of Medical Things. Information Fusion, 102, 102017. https://doi.org/10.1016/j.inffus.2023.102017

[37] Yang, S., Cui, L., Wang, L., Wang, T., & You, J. (2024). Enhancing multimodal depression diagnosis through representation learning and knowledge transfer. Heliyon, 10(4), e25959. https://doi.org/10.1016/j.heliyon.2024.e25959

[38] Xu, Z., Gao, Y., Wang, F., Zhang, L., Zhang, L., Wang, J., & Shu, J. (2025). Depression detection methods based on multimodal fusion of voice and text. Scientific Reports, 15, 21907. https://doi.org/10.1038/s41598-025-03524-4

[39] Yang, S., Liu, S., Nie, G., Wang, L., Wang, T., You, J., & Cambria, E. (2025). Fine-grained multimodal fusion for depression assisted recognition based on hierarchical knowledge-enhanced prompt learning. Expert Systems with Applications, 291, 128532. https://doi.org/10.1016/j.eswa.2025.128532

[40] Weber, J., Weber, M., & Lopez Alcaraz, J. M. (2025). Depression diagnosis from patient interviews using multimodal machine learning. Frontiers in Psychiatry, 16, 1694762. https://doi.org/10.3389/fpsyt.2025.1694762

[41] Trelles, A., Fontaines Ruiz, T. I., & Ponce Rojo, A. (2025). Systematic review and meta-analysis of explainable machine learning models for clinical depression detection. Behavioral Sciences, 15(11), 1476. https://doi.org/10.3390/bs15111476

[42] Dang, V. N., Cascarano, A., Mulder, R. H., Cecil, C., Zuluaga, M. A., Hernández-González, J., & Lekadir, K. (2024). Fairness and bias correction in machine learning for depression prediction across four study populations. Scientific Reports, 14, 7848. https://doi.org/10.1038/s41598-024-58427-7

[43] Benda, N., Desai, P., Reza, Z., Zheng, A., Kumar, S., Harkins, S., Hermann, A., Zhang, Y., Joly, R., Kim, J., Pathak, J., & Reading Turchioe, M. (2024). Patient perspectives on AI for mental health care: Cross-sectional survey study. JMIR Mental Health, 11, e58462. https://doi.org/10.2196/58462

[44] Rogan, J., Bucci, S., & Firth, J. (2024). Health care professionals' views on the use of passive sensing, AI, and machine learning in mental health care: Systematic review with meta-synthesis. JMIR Mental Health, 11, e49577. https://doi.org/10.2196/49577

[45] Alam, N. B., Surani, M., Das, C. K., Giacco, D., Singh, S. P., & Jilka, S. (2025). Challenges and standardisation strategies for sensor-based data collection for digital phenotyping. Communications Medicine, 5(1), 360. https://doi.org/10.1038/s43856-025-01013-3

[46] Zhang, Y., Wang, J., Zong, H., Singla, R. K., Ullah, A., Liu, X., Wu, R., Ren, S., & Shen, B. (2025). The comprehensive clinical benefits of digital phenotyping: from broad adoption to full impact. npj Digital Medicine, 8, 196. https://doi.org/10.1038/s41746-025-01602-5

[47] Ali, M., Ali, S., Abbas, Q., Abbas, Z., & Lee, S. W. (2025). Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. Digital Health, 11. https://doi.org/10.1177/20552076251395548

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.