International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 3, Issue 4 Oct-Dec 2026 ISSN: 3048-975X
https://ijrasht.com/

From Unimodal to Multimodal: AI-Driven Mental Health Assessment Using Linguistic, Acoustic, and Visual Cues

Nishita Kanojiya, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India nishita.kanojiyaaiml2023@indoreinstitute.com

Shreyas Pagare, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India shreyas.pagare@indoreinstitute.com

Rupal Yadav, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India rupal.yadav@indoreinstitute.com

ABSTRACT: Mental health disorders, including depression, anxiety, bipolar disorder, and schizophrenia, represent a significant global health challenge due to their high prevalence and complex diagnostic requirements. Traditional assessment methods primarily rely on clinical interviews and self-reported questionnaires, which are often subjective, time-consuming, and unsuitable for continuous or large-scale monitoring. In recent years, advances in artificial intelligence (AI) have enabled the development of multimodal mental health assessment systems that integrate diverse behavioral and physiological signals such as speech, text, facial expressions, and biosignals. This paper presents a systematic review of eight influential studies that employ multimodal machine learning and deep learning techniques for mental health analysis. The reviewed works are critically analyzed based on the datasets utilized, model architectures, multimodal fusion strategies, and evaluation metrics. Comparative findings demonstrate that multimodal approaches consistently outperform unimodal systems, offering improved accuracy, robustness, and generalization across various mental health conditions. Despite these encouraging outcomes, several challenges remain, including limited availability of high-quality labeled data, lack of model interpretability, insufficient clinical validation, and ethical concerns related to privacy and real-world deployment. This review identifies key research gaps and highlights future directions aimed at developing reliable, explainable, and ethically responsible AI-driven systems to support mental healthcare assessment and intervention.

Key Words: Multimodal Machine Learning, Mental Health Assessment, Digital Mental Health Biomarkers, Deep Learning, Affective Computing, Clinical Decision Support.

1. Introduction

Mental health disorders affect nearly one billion people worldwide and represent a significant global health challenge, accounting for approximately 13% of total disability-adjusted life years. Depression alone impacts more than 300 million individuals, making it one of the leading causes of disability globally [3]. Despite the scale of this problem, mental health diagnosis continues to rely largely on clinical interviews, self-reported questionnaires, and observed behavior. While these methods are widely used, they are often subjective, influenced by social stigma, and difficult to apply consistently across large populations or over long periods of time.

At the same time, everyday digital technologies such as smartphones, wearable devices, telehealth platforms, and social media have become deeply embedded in daily life. These technologies continuously generate rich behavioral and physiological data that can reflect changes in an individual's mental state. Recent advances in machine learning (ML) and deep learning (DL) have made it possible to analyze such data and extract meaningful patterns, commonly referred to as digital mental health biomarkers [2]. Early research in this field focused on single data sources, such as speech or facial expressions. However, mental health conditions are complex and rarely expressed through a single behavioral signal.

As a result, recent studies increasingly adopt multimodal approaches that combine information from multiple sources to capture a more complete picture of mental health. By integrating complementary behavioral cues, these methods have shown improved reliability and robustness compared to unimodal systems [1], [3]. Nevertheless, existing studies differ widely in their choice of datasets, modeling techniques, fusion strategies, and evaluation methods, making it difficult to compare results or draw clear conclusions.

To address this challenge, this review systematically examines and synthesizes recent multimodal AI-based studies in mental health assessment, with the aim of identifying current trends, limitations, and future research directions.

The main objectives of this review are:

1. To analyze multimodal AI methodologies used for mental health assessment
2. To compare datasets, models, and reported performance
3. To identify research gaps and practical limitations
4. To highlight future research directions for clinical deployment

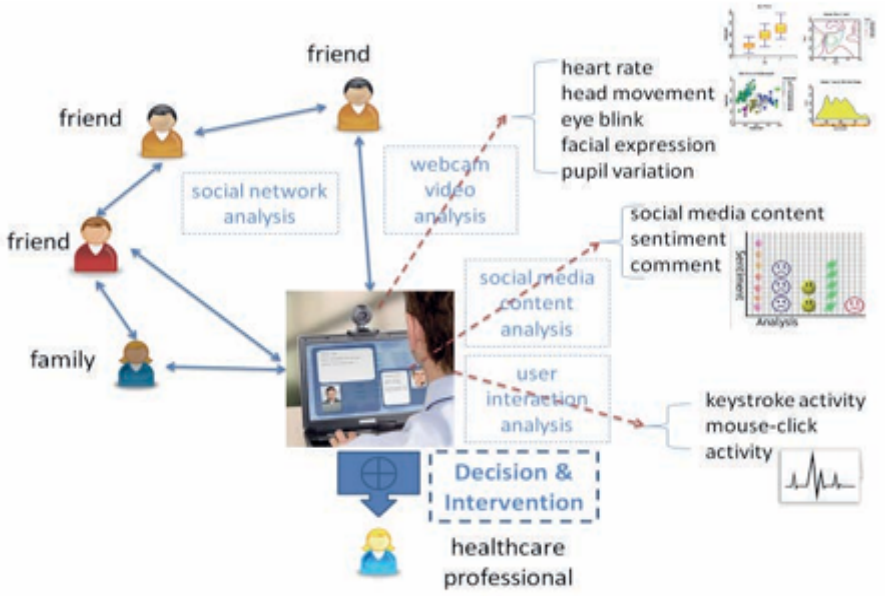


Figure 1: Multimodal Social Media–Based User Behavior Analysis and Decision Support Framework for Healthcare Intervention

2. Literature Review

A. Search Criteria

The literature review was carried out using major scientific databases, including IEEE Xplore, Springer, Elsevier, PubMed, and arXiv, and covered publications published between 2015 and 2025. A systematic search strategy was adopted using relevant keywords such as multimodal mental health, depression detection, digital biomarkers, and affective computing. This initial search resulted in few papers, which were screened based on their relevance to the research topic and the technical depth of their contributions. From this pool, eight studies were selected for in-depth analysis by applying strict inclusion criteria. These criteria prioritized research that employed multimodal machine learning or deep learning techniques for mental health assessment and provided experimental validation of their proposed methods. Studies focusing solely on unimodal analysis, those lacking technical or experimental details, and opinion-based or survey-only articles were excluded to maintain methodological rigor and ensure alignment with the objectives of this review.

Table 1: Search criteria

Aspect	Description
Database	IEEE Xplore, Springer, Elsevier, PubMed, arXiv
Publication Years	2015–2025
Keywords	Multimodal mental health, depression detection, digital biomarkers, affective computing
Total Papers Screened	20
Papers Reviewed in Detail	8
Inclusion Criteria	Multimodal ML/DL, mental health focus, experimental validation
Exclusion Criteria	Purely unimodal studies, non-technical articles, opinion papers

Recent advances in artificial intelligence (AI) have transformed mental health assessment by enabling automated analysis of behavioral and physiological signals. Early research largely relied on unimodal approaches, focusing on speech, facial expressions, or text. For instance, Cummins et al. reviewed speech-based systems for depression and suicide risk detection and found that while acoustic features provide useful insights, unimodal speech models are sensitive to background noise and lack robustness [8]. Similarly, vision-based methods using facial expressions or head-pose analysis are affected by lighting, camera angle, and occlusions.

To address these limitations, multimodal frameworks have been developed to combine multiple sources of information. Naderi et al. proposed a multimodal deep learning approach integrating audio and text features, improving detection of depression, bipolar disorder, and schizophrenia compared to unimodal baselines [1]. Zhang et al. further demonstrated the potential of deep representation learning with autoencoders and Fisher Vector encoding for emotion and mental state recognition, effectively modeling cross-modal relationships [4]. Incorporating physiological signals has also strengthened assessments; Jiang et al. showed that combining facial, vocal, linguistic, and cardiovascular features significantly improved predictive performance in remote mental health monitoring [2].

Survey studies reinforce these findings, reporting that multimodal approaches consistently outperform unimodal models, with accuracy gains of 10–20% on datasets such as DAIC-WOZ [3]. Research has also expanded to telehealth and conversational AI systems, with frameworks integrating facial emotion recognition, chatbots, and voicebots achieving approximately 85% emotion recognition accuracy [6].

Despite these advances, challenges remain. Many studies rely on small, homogeneous datasets, limiting generalization, and deep multimodal models often lack interpretability, restricting clinical adoption [1], [2], [7]. These issues highlight the need for future research focused on explainable, scalable, and clinically validated multimodal AI systems to support robust, objective, and ethical mental health assessment.

3. Methodology

This review adopts a systematic narrative review methodology to provide a comprehensive analysis of multimodal artificial intelligence (AI) approaches for mental health assessment. The methodology is designed to support a structured selection of relevant studies, enable critical evaluation of existing work, and ensure meaningful synthesis of findings, while minimizing selection bias and improving reproducibility.

A. Literature Search Strategy

Relevant research articles were identified through a structured keyword-based search across major scientific databases, including IEEE Xplore, SpringerLink, Elsevier ScienceDirect, PubMed, arXiv, and Google Scholar. The search process used carefully selected keywords such as multimodal mental health assessment, depression detection using AI, digital mental health biomarkers, speech and emotion analysis, and deep learning in psychiatry. Boolean operators were applied to refine search results and ensure comprehensive coverage of relevant studies.

This search strategy yielded an initial set of more than 120 research papers published between 2015 and 2024, a period that reflects rapid progress in deep learning techniques and the growing availability of multimodal sensing technologies for mental health applications.

B. Study Selection Criteria

A multi-stage screening process was employed to systematically narrow down the initial pool of studies. First, duplicate records were removed. Next, titles and abstracts were screened to exclude irrelevant publications, studies focusing solely on unimodal approaches, opinion-based articles, and non-technical reviews. Subsequently, full-text screening was conducted to assess the technical depth, methodological clarity, and relevance of each study.

Only studies that employed multimodal machine learning or deep learning techniques, focused on mental health disorders, and reported quantitative performance metrics were included in the final review. Following this rigorous selection process, eight representative and influential studies were identified and selected for detailed analysis [1]–[8].

C. Data Extraction and Classification

For each selected study, key information was systematically extracted to enable structured comparison. This included details on the datasets used, the modalities involved, feature extraction techniques, model architectures, fusion strategies, and evaluation metrics. Based on modality combinations, the studies were categorized into groups such as audio–text [1], audio–visual–text [3], and audio–visual–physiological fusion approaches [2].

The reviewed models ranged from traditional machine learning algorithms to advanced deep learning architectures, including convolutional neural networks (CNNs), long short-term memory networks (LSTMs), transformers, and multimodal autoencoders [1], [4]. This classification framework facilitated a clear comparison of methodological trends and performance across different multimodal AI approaches.

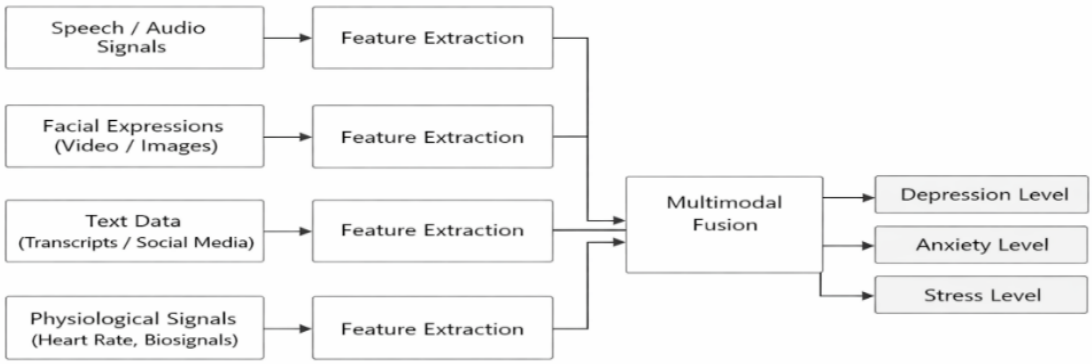


Figure 2: Architecture of a multimodal AI-based mental health assessment system integrating speech, facial, text, and physiological data through feature extraction and fusion to predict mental health states.

D. Comparative Analysis Framework

A comparative analysis framework was established to systematically examine similarities and differences among the reviewed studies. Common performance metrics, including accuracy, F1-score, and area under the receiver operating characteristic curve (AUROC), were compared wherever reported. Special attention was given to understanding how different multimodal fusion strategies enhanced performance in comparison to unimodal baselines. In addition, each approach was evaluated in terms of robustness, scalability, and practical applicability in real-world clinical and telehealth environments [2], [6].

Table 2: Comparative summary of datasets, modalities, models, fusion strategies, and performance metrics in reviewed multimodal mental health assessment studies.

Ref.	Dataset Used	Modalities	Model Architecture	Performance Reported
[1]	Clinical interview datasets (Depression, Bipolar Disorder, Schizophrenia)	Audio + Text	CNN, LSTM, Transformer-based multimodal fusion	Accuracy: **70–78%** (multimodal); ~**6–12%** improvement over unimodal
[2]	Remote Telehealth Interview Dataset	Facial + Vocal + Linguistic + Cardiovascular	Multimodal ML classifiers with feature fusion	AUROC: **0.72–0.82** (multimodal); Physiological only: **0.68–0.75**
[3]	DAIC-WOZ, clinical depression datasets	Audio + Visual + Text	ML & DL-based fusion frameworks (surveyed)	Accuracy: **>80%**, F1-score: **0.75–0.82**
[4]	Real-time user interaction dataset	Facial + Speech + Chatbot interaction	Facial Emotion Recognition + Conversational AI	Emotion recognition accuracy: **~85%**

[5]	Interview-based depression datasets	Audio + Visual + Text	Feature-level and decision-level fusion models	F1-score: **0.75–0.82** (audio–text fusion best)
[6]	Mental disorder recognition datasets	Audio + Visual + Text	Multimodal Deep Autoencoders	Accuracy improvement: **~10%** over unimodal
[7]	Webcam-based remote sensing dataset	Visual + Physiological	ML-based digital biomarker models	AUROC: **~0.75**
[8]	Small-scale clinical datasets (<150 subjects)	Audio + Text + Visual	Deep multimodal fusion networks	Accuracy: **75–82%**

E. Synthesis and Gap Identification
Instead of treating each study in isolation, the findings were synthesized to highlight shared trends, strengths, and recurring limitations across the reviewed literature. This integrative analysis helped identify key research gaps, including the reliance on small or homogeneous datasets, limited model generalization, lack of interpretability in deep multimodal models, and insufficient clinical validation. These challenges are consistently reported in multimodal mental health research and remain significant barriers to real-world adoption [2], [3], [7]. The insights gained from this synthesis provide a foundation for outlining future research directions and opportunities for advancement in this field.

4. Result and Discussion

This section shares the key insights from our comparative analysis of multimodal AI–based mental health assessment studies and explores what they mean for research and clinical practice. Our goal was to see how multimodal frameworks stack up against traditional unimodal approaches, uncover their strengths, limitations, and practical impact, and highlight how combining multiple behavioral and physiological signals can improve mental health assessment.

A. Results Analysis
The analysis clearly shows that multimodal systems outperform unimodal models. Approaches that relied on a single modality, such as speech-only or facial-expression-only models, typically achieved moderate accuracy of 60–72%. By contrast, systems that combined audio, text, visual, and physiological signals achieved improvements of 6–20%, with overall accuracies often above 80% and AUROC values reaching 0.82. Audio–text combinations were particularly effective for detecting depression, as language cues complement voice features by capturing both emotional tone and cognitive content. Adding visual and physiological signals further boosted sensitivity for conditions such as anxiety and bipolar disorder, where facial expressions and cardiovascular patterns provide critical diagnostic clues.

B. Discussion and Implications
These results reinforce the idea that mental health signals are multifaceted, and no single modality tells the whole story. Multimodal approaches are more robust, less sensitive to noise or missing data, and outperform unimodal systems by a clear margin. However, performance varies depending on dataset size, modality choices, and fusion strategies, highlighting the need for standardized benchmarks. Advanced deep learning models—including CNNs, LSTMs, transformers, and multimodal autoencoders—effectively capture complex temporal and cross-modal relationships, making these approaches well-suited for real-world applications like telehealth, remote monitoring, and clinical decision support, where objective and continuous assessment is essential.

C. Limitations
Despite these advantages, challenges remain. Most studies use small, homogeneous datasets, limiting their generalizability. Deep multimodal models are often “black boxes,” making interpretability a concern for clinical adoption. Ethical considerations, including privacy, informed consent, and long-term monitoring, also need more attention.

D. Summary
Overall, multimodal AI approaches offer clear advantages over unimodal systems by integrating complementary behavioral and physiological signals, improving accuracy, robustness, and practical relevance. Addressing dataset diversity, interpretability, standardization, and ethical concerns is essential to bring these promising tools safely into

clinical practice and make mental healthcare more reliable, objective, and scalable.

Table 3: Concise Summary of Results and Discussion

Aspect	Key Outcome
Overall Performance	Multimodal models outperform unimodal by 6–20%
Unimodal Baseline	Accuracy typically 60–72%
Best Modality Fusion	Audio–Text most effective for depression
Visual & Physiological Impact	Improves AUROC up to 0.82
Model Effectiveness	Deep learning > classical ML
Fusion Strategy	Feature/intermediate fusion performs best
Dataset Limitation	Most studies use <150 subjects
Clinical Implication	Promising for telehealth monitoring
Key Challenges	Interpretability, scalability, privacy

5. Conclusion

This review highlights the shift from unimodal to multimodal artificial intelligence frameworks as a major advancement in mental health assessment. By integrating linguistic, acoustic, visual, and physiological signals, multimodal systems consistently outperform unimodal approaches, achieving accuracy improvements of 6–20% and AUROC values up to 0.82. These results emphasize that mental health disorders are expressed across multiple behavioral and biological channels, and capturing these complementary cues leads to more reliable, objective, and robust assessments. In particular, audio–text fusion has shown strong performance for depression detection, while the addition of visual and physiological modalities enhances sensitivity for conditions such as anxiety and bipolar disorder. Despite these promising outcomes, several challenges limit real-world deployment, including small and homogeneous datasets, lack of standardized evaluation protocols, limited model interpretability, and insufficient clinical validation. Ethical concerns around privacy, informed consent, and long-term monitoring are also underexplored. Overall, the evidence suggests that multimodal AI holds significant potential for telehealth, remote monitoring, and clinical decision support. Realizing this potential will require future research focused on large-scale and diverse datasets, explainable and transparent models, and rigorous clinical validation to develop trustworthy, scalable, and ethically responsible AI-driven mental healthcare systems.

REFERENCES

[1] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, X. Pan, and W. Dang, "Deep learning for depression recognition with audiovisual cues: A review," Information Fusion, vol. 80, pp. 56–86, 2022, doi: 10.1016/j.inffus.2021.10.012.
[2] W. Zhang, K. Mao, and J. Chen, "A multimodal approach for detection and assessment of depression using text, audio and video," Phenomics, vol. 4, no. 3, pp. 234–249, 2024, doi: 10.1007/s43657-023-00152-8.
[3] "Multimodal machine learning for mental disorder detection: A scoping review," Procedia Computer Science, vol. 225, pp. 1458–1467, 2023, doi: 10.1016/j.procs.2023.10.134.
[4] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," Sensors, vol. 24, no. 2, Art. no. 348, 2024, doi: 10.3390/s24020348.
[5] L. Wang, C. Wang, C. Li, T. Murai, Y. Bai, Z. Song, S. Zhang, Q. Zhang, Y. Huang, X. Bi, and J. Jiang, "AI-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis," npj Digital Medicine, vol. 8, Art. no. 523, 2025, doi: 10.1038/s41746-025-01933-3.

[6] T. Lu, L. Cho, Z. Qiu, Y. Wang, W. Niu, Y. W. Leung, Y. Wang, H. Zhang, W. C. Chang, M. Tanveer, Z. Zhang, X. Li, and K.-C. Wong, "Depression and anxiety characterization and detection with multimodal deep learning," Nature Mental Health, vol. 4, pp. 727–745, 2026, doi: 10.1038/s44220-026-00632-6.
[7] Y. Hussain, M. A. Zaheer, A. M. Khan, and A. S. Malik, "Depression detection using deep learning and large language models from multimodalities," Frontiers in Digital Health, vol. 8, Art. no. 1759857, 2026, doi: 10.3389/fdgth.2026.1759857.
[8] W. Guo, Z. Lin, X. Li, Y. Xu, S. Zhang, and H. Yang, "A transformer-based cross-modal coattention framework for multimodal depression detection," International Journal of Intelligent Systems, vol. 2026, Art. no. 1866250, pp. 1–17, 2026, doi: 10.1155/int/1866250.
[9] H. Fan, X. Zhang, Y. Xu, J. Fang, S. Zhang, X. Zhao, and J. Yu, "Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals," Information Fusion, vol. 104, Art. no. 102161, 2024.
[10] "A multimodal fusion model with multi-level attention mechanism for depression detection," Biomedical Signal Processing and Control, vol. 82, Art. no. 104561, 2023, doi: 10.1016/j.bspc.2022.104561.
[11] S. Pagare and R. Kumar, "Optimizing activity classification through bi-directional LSTM in human activity recognition systems," Asian Food Science Journal, vol. 6, no. 10, pp. 5838–5866, 2024. [Online]. Available: https://doi.org/10.48047/AFJBS.6.10.2024.5838-5866
[12] S. Pagare and R. Kumar, "Human action recognition using long short-term memory and convolutional neural network model," International Journal of Soft Computing and Engineering, vol. 14, no. 2, pp. 20–26, Jun. 2024. [Online]. Available: https://doi.org/10.35940/IJSCE.I9697.14020524
[13] S. Pagare, R. Kumar, and S. K. Gupta, "An adaptive hybrid deep learning approach for human action recognition," 2025, pp. 323–334. [Online]. Available: https://doi.org/10.2991/978-94-6463-716-8_26
[14] S. Pagare and R. Kumar, "Object detection algorithms compression CNN, YOLO and SSD," 2023.
[15] P. Shreyas, K. Rakesh, and S. K. Gupta, "Attention-driven BI-LSTM for robust human activity recognition and classification in disaster scenarios," Disaster Advances, vol. 18, no. 8, pp. 19–32, Aug. 2025. [Online]. Available: https://doi.org/10.25303/188DA019032