CINEFORUM
ISSN: 0009-7039
Vol. 66. No. S9, 2026 (Special Issue S9)

A Study of "Multimodal Depression Detection Using Machine Learning

Mukundam Uduta[1]*, Dr. K. Srujan Raju[2]

[1]Research Scholar, Bharatiya Engineering Science & Technology Innovation University (BESTIU) Andhra Pradesh, India.
[2]Dean, Research & Development, CMR Technical Campus, Hyderabad, Telangana, India
***Corresponding Author**
Mukundam Uduta
Research Scholar, Bharatiya Engineering Science & Technology Innovation University (BESTIU) Andhra Pradesh, India.

ABSTRACT
Depression is a major mental-health concern characterized by persistent disturbances in mood, interest, cognition, behavior, and daily functioning. The increasing availability of digital behavioural data has encouraged researchers to investigate artificial intelligence (AI) and machine-learning (ML) approaches for supporting depression screening and assessment. While conventional approaches frequently depend on questionnaires and clinical interviews, machine-learning research has explored potentially informative signals from text, speech, facial behaviour, video, social-media activity, physiological measurements, and other behavioural sources. Multimodal machine learning is particularly important because depression is expressed through interconnected linguistic, vocal, visual, and behavioural characteristics that may not be adequately represented by a single data source. This article descriptively reviews published research on multimodal machine-learning approaches for depression detection using secondary data. The analysis covers conventional ML, deep learning, recurrent neural networks, attention mechanisms, Transformers, graph-based approaches, contrastive learning, and emerging large-language-model-based systems. The reviewed evidence indicates a gradual movement from single-modality approaches toward multimodal representation and fusion, particularly involving text, audio, and video. However, differences in datasets, clinical labels, populations, preprocessing procedures, evaluation measures, and fusion strategies restrict direct comparison of reported performance. Privacy, bias, interpretability, dataset imbalance, and external validation remain important challenges. The study concludes that multimodal ML provides a promising research direction for depression screening support, but automated detection should complement rather than replace professional clinical assessment.
Keywords: Depression Detection, Multimodal Machine Learning, Artificial Intelligence, Deep Learning, Mental Health

1. INTRODUCTION

Depression is a common mental disorder with substantial consequences for individuals, families, workplaces, and health systems. The World Health Organization reports that approximately 322 million people were living with depression in 2023 and emphasizes that depression is distinct from ordinary short-term fluctuations in mood. The scale and complexity of the condition have encouraged research into technologies that can support earlier identification and continuous monitoring. At the same time, conventional assessment remains dependent on clinical interviews, questionnaires, observation, and professional interpretation. Such approaches remain central to clinical practice, but researchers have explored computational methods that can identify behavioural patterns associated with depressive symptoms. The limitations of individual modalities provide a major rationale for multimodal machine learning. Text can represent linguistic and semantic information; audio can capture prosody, speaking rate, energy, pitch, and other acoustic characteristics; video can provide information about facial movements, gaze, head position, and other non-verbal behaviors. Combining these sources may therefore provide a broader representation of observable behaviour. Reviews of multimodal mental-health detection have found increasing use of neural architectures for model-level fusion and for learning relationships within and between modalities.

Nevertheless, multimodal systems face important limitations. Depression datasets are often relatively small, heterogeneous, and collected under specific experimental conditions. Different studies may use different diagnostic instruments, thresholds, populations, modalities, preprocessing procedures, fusion strategies, and evaluation metrics. Consequently, a high score on one dataset cannot automatically establish universal superiority or clinical effectiveness. The research problem addressed in this article is therefore not the development of another

OPEN ACCESS © CINEFORUM

ML classifier but the descriptive synthesis of existing evidence concerning how multimodal machine learning has developed, what modalities and approaches dominate the field, and what technical and ethical barriers remain.

2. RESEARCH OBJECTIVES

To examine the development of machine-learning approaches used for depression detection.
To analyse the major modalities used in machine-learning-based depression detection.

3. REVIEW OF LITERATURE

Chen et al. (2025) proposed text-guided multimodal depression detection using cross-modal feature reconstruction, decomposition, bidirectional cross-attention, and Transformer-based fusion. The study used DAIC-WOZ and E-DAIC and specifically addressed information redundancy and heterogeneity between modalities.

Cummins et al. (2015) reviewed automatic depression and suicide-risk assessment using speech analysis. Their work established speech as an important behavioural signal and highlighted problems associated with speaker variability, recording conditions, standardization, and generalizability. The study remains relevant because these limitations continue to influence multimodal research.

Fang et al. (2023) proposed a multimodal fusion model with a multi-level attention mechanism integrating audio, visual, and textual information. The model used recurrent networks to learn modality-specific representations before attention-based fusion. The study highlights the role of attention in reducing redundant information and learning relationships between modalities.

Flores et al. (2022) introduced AudiFace, combining temporal facial features, audio, and transcripts through transfer learning, bidirectional LSTM, and self-attention. The researchers reported strong performance across their evaluated datasets and identified eye gaze as an important temporal facial feature. The study nevertheless emphasized the challenge created by limited participant numbers and heterogeneous modalities.

Fu et al. (2025) introduced facial-action-unit-guided graph representation learning for multimodal depression detection. The approach integrated visual and audio information while attempting to reduce the effects of interpersonal variability through graph-based representation learning. The study evaluated the approach on AVEC datasets, demonstrating continued development of graph-based multimodal methods.

He et al. (2024) proposed DPD-Net, a graph-neural-network-enhanced Transformer framework using textual, audio, and visual features. The model was evaluated across multiple benchmark datasets and was designed to address the limited robustness associated with dependence on a single dataset. Its cross-dataset design is particularly relevant to the generalizability problem.

Hussain et al. (2026) reviewed recent deep-learning and large-language-model approaches involving multimodal depression detection. Their synthesis indicates increasing integration of speech, facial expression, EEG, text, and other behavioural or physiological signals, while emphasizing continuing problems in dataset diversity, standardization, interpretability, and clinical validation.

Li et al. (2024) proposed an intra- and inter-sample contrastive-learning approach for multimodal depression detection. The framework decomposed audio, visual, and textual information into modality-common and modality-specific representations and incorporated contrastive objectives before multimodal fusion. This represents a shift from simple fusion toward learning more discriminative cross-modal representations.

Lim et al. (2025) proposed a lightweight cross-modality model integrating audio and text. The model was evaluated using English, Chinese, and Korean depression datasets, with reported F1 scores varying across datasets. The multilingual evaluation is important because it highlights both the potential and the difficulty of transferring depression-related representations across linguistic environments.

Yadav et al. (2023) reviewed automated depression detection using social posts, audio, and video and considered more than 140 related publications, with a comparative analysis of a subset of studies. Their review identified diverse datasets and methodological approaches and highlighted unresolved issues surrounding data quality, modality differences, and future research directions.

Ye et al. (2021) examined multimodal depression detection using emotional audio and evaluation text. The study used 160 Chinese participants and combined low-level acoustic features, DeepSpectrum representations, and word-vector features. The reported multimodal model achieved an accuracy of 0.912 and an F1 score of 0.906 within that study's experimental context.

Zhang et al. (2024) developed a multimodal framework using text, audio, and video for depression detection and assessment. Using the DAIC-WOZ dataset, the work addressed the challenge of extracting useful information from long textual inputs and high-dimensional multimodal data. The study further illustrates the continuing importance of the DAIC-WOZ benchmark in multimodal research.

OPEN ACCESS © CINEFORUM

4. RESEARCH METHODOLOGY

The study adopts a **descriptive** research design based exclusively on secondary data. No primary participants were recruited, no questionnaire was administered, and no new machine-learning model was trained. The purpose is to organize, describe, compare, and critically synthesize existing published evidence concerning multimodal machine-learning approaches for depression detection.

Secondary evidence was obtained from peer-reviewed journal articles, conference publications, systematic reviews, scholarly databases, official dataset documentation, and authoritative institutional sources. Particular emphasis was placed on publications indexed or discoverable through major scholarly databases and on studies published between 2019 and 2026, while seminal earlier work was retained where it established important methodological foundations. The evidence was examined using descriptive, comparative, thematic, trend, and content analysis. Studies were classified according to modality, algorithmic approach, fusion strategy, dataset context, and major limitations. Published findings were interpreted comparatively rather than statistically pooled. Information was cross-checked across publisher pages, PubMed, scholarly repositories, and recognized institutional sources. Reported performance values are attributed to the original studies rather than presented as results generated by the present article. The study uses published evidence and therefore does not directly process identifiable patient information. Nevertheless, ethical issues reported by previous research—including privacy, informed consent, fairness, explain ability, responsible AI, and avoidance of clinical overclaiming—are treated as central analytical concerns.

5. SECONDARY DATA ANALYSIS AND FINDINGS

Table 1. Major Modalities Used in Depression Detection

Modality	Data Type	Major Features	Potential Contribution	Key Limitation
Text	Linguistic/semantic data	Words, syntax, sentiment, contextual embeddings	Captures linguistic and semantic indicators	Language and cultural dependence
Audio	Speech/acoustic data	Pitch, prosody, energy, speaking rate, spectral features	Captures vocal and paralinguistic behaviour	Sensitive to recording conditions and speaker variability
Video	Facial/visual data	Facial action units, gaze, head movement, expressions	Captures non-verbal behavioural cues	Lighting, camera position, interpersonal variability
Social media	Text, images, behaviour	Posts, interactions, linguistic patterns	Enables naturalistic longitudinal information	Privacy, self-selection, noisy labels
Physiological	EEG, rPPG and related signals	Neural/cardiovascular features	Provides physiological information	Equipment, privacy, limited datasets
Passive sensing	Smartphone/wearable data	Activity, mobility, communication patterns	Supports longitudinal behavioural monitoring	Data governance and contextual ambiguity

OPEN ACCESS © CINEFORUM

Interpretation: Text, audio, and video dominate clinical-interview-based multimodal research, while passive sensing and physiological modalities expand the potential scope of automated assessment. Systematic evidence indicates that multimodal approaches can represent complementary behavioural information, but modality-specific associations vary according to individuals, demographic characteristics, and context.

Table 2. Selected Multimodal Machine-Learning Approaches Reported in Literature

Author/Year	Modalities	ML/DL Approach	Research Context	Main Finding
Williamson et al. (2016)	Vocal, facial, semantic	Feature-based multimodal analysis	Clinical communication	Demonstrated complementary communication cues
Yin et al. (2019)	Text, audio, video	Hierarchical Bi-LSTM	AVEC depression challenge	Modelled multimodal temporal structure
Ray et al. (2019)	Text, audio, video	Multi-level attention	AVEC depression challenge	Learned intra- and inter-modal relevance
Fan et al. (2019)	Text, audio, other temporal features	Dilated CNN	AVEC 2019	Captured temporal dependencies
Ye et al. (2021)	Audio, text	Deep learning fusion	Chinese dataset	Reported improved multimodal classification
Muzammel et al. (2021)	Audio, visual, text	Deep neural networks	Clinical depression recognition	Compared multimodal learning/fusion strategies
Flores et al. (2022)	Facial, audio, transcripts	Transfer learning + Bi-LSTM attention	Clinical interviews	Demonstrated value of combining modalities
Fang et al. (2023)	Audio, visual, text	LSTM + multi-level attention	DAIC-WOZ	Used intra- and inter-modal attention
Zhang et al. (2024)	Text, audio, video	Multimodal fusion	DAIC-WOZ	Addressed high-dimensional multimodal information
He et al. (2024)	Text, audio, visual	GNN + Transformer	Multiple datasets	Emphasized cross-dataset robustness
Li et al. (2024)	Text, audio, visual	Contrastive learning + fusion	Public datasets	Learned common and modality-specific representations
Fu et al. (2025)	Audio, visual	Graph representation learning	AVEC datasets	Addressed interpersonal variability

Interpretation: The table demonstrates the transition from manually engineered multimodal features toward learned representations. Recurrent models remain useful for sequential information, whereas attention and Transformer mechanisms increasingly dominate studies concerned with long-range cross-modal relationships. More recent work adds graph learning, contrastive objectives, and foundation models.

Table 2. Major Challenges in Multimodal Depression Detection

Challenge	Description	Research Implication
Data heterogeneity	Modalities differ in structure, quality, scale, and temporal resolution	Requires robust alignment and fusion
Dataset limitations	Many datasets are small or context-specific	External validation is necessary

OPEN ACCESS © CINEFORUM

Class imbalance	Depressed and non-depressed samples may not be equally represented	Evaluation should consider appropriate metrics
Privacy	Audio, video, text and physiological data may contain sensitive information	Strong data governance is essential
Bias	Demographic, linguistic and cultural differences can influence predictions	Fairness evaluation is required
Interpretability	Deep multimodal architectures can be difficult to explain	Explainable AI should accompany model development
Generalizability	Results may not transfer across populations or datasets	Cross-dataset and cross-cultural validation is needed
Missing modalities	Real-world systems may not always have every data source	Models should tolerate incomplete inputs

Interpretation: Technical performance alone is insufficient for evaluating multimodal depression detection. The literature increasingly recognizes that trustworthy systems require reliable data, appropriate validation, transparency, privacy safeguards, and evidence of applicability beyond the dataset on which a model was developed.

6. DISCUSSION

The secondary evidence indicates that multimodal machine learning has developed in response to the limitations of single-modality depression detection. Text captures semantic and linguistic information, speech provides acoustic and paralinguistic information, and video captures visible behavioural cues. Their combination potentially allows algorithms to represent different aspects of the same behavioural episode. Traditional ML methods remain relevant because they can work effectively with carefully selected features and relatively modest datasets. However, deep learning has reduced dependence on manually designed features by learning representations directly from raw or minimally processed inputs. Recurrent networks have been particularly relevant to sequential interview data, while attention mechanisms provide a mechanism for emphasizing informative segments and relationships. More recent Transformer and graph-based approaches attempt to model cross-modal interactions more explicitly. Contrastive-learning approaches further attempt to distinguish modality-common information from modality-specific information. Recent LLM-oriented research suggests another methodological transition toward contextual and semantically rich representations.

7. THEORETICAL AND CONCEPTUAL IMPLICATIONS

Multimodal machine learning provides a conceptual framework for representing depression-related information across interconnected data sources. Its central concepts include feature fusion, representation learning, cross-modal attention, multimodal embeddings, and temporal modelling. The progression from early feature concatenation toward learned cross-modal representations suggests that the field increasingly views modalities as complementary rather than merely additive. Explainable AI is also conceptually important because clinical users require information about why an algorithm generates a particular screening signal. Recent reviews similarly emphasize the need for transparency and reproducibility in automated clinical depression assessment.

8. PRACTICAL AND ETHICAL IMPLICATIONS

For mental-health researchers, multimodal ML offers opportunities to investigate behavioural indicators that may be difficult to quantify manually. For healthcare professionals, such systems may potentially provide supplementary screening information, particularly where repeated monitoring is difficult. For AI developers, the evidence emphasizes the importance of representative datasets, robust validation, transparent reporting, and uncertainty-aware modelling.

For digital-health organizations, privacy-preserving data practices are essential because multimodal information can include highly sensitive speech, facial images, written content, physiological information, and behavioural records. For policymakers, responsible deployment requires attention to consent, data protection, fairness, accountability, and appropriate clinical oversight.

OPEN ACCESS © CINEFORUM

9. LIMITATIONS

This descriptive study has several limitations. First, it depends on published secondary evidence and therefore reflects the methodological choices and reporting quality of previous researchers. Second, multimodal datasets vary substantially in size, population, modality availability, clinical assessment instruments, and data-collection environments. Third, reported performance metrics cannot always be directly compared because studies employ different validation protocols and evaluation measures. Fourth, many multimodal systems rely on benchmark datasets that may not adequately represent demographic, linguistic, cultural, or socioeconomic diversity. Fifth, privacy restrictions can limit access to realistic clinical datasets. Finally, the present study does not perform a statistical meta-analysis and therefore does not estimate a pooled effect or universally valid performance level.

10. FUTURE RESEARCH

Future research should prioritize larger and more diverse multimodal datasets representing different languages, cultures, ages, and demographic groups. Longitudinal data may help researchers examine changes in depressive symptoms rather than relying exclusively on single assessment episodes. Cross-dataset and cross-cultural validation should become routine before claims of generalizability are made. Explainable AI, fairness assessment, privacy-preserving machine learning, and federated learning deserve further attention. Research should also investigate models capable of functioning when one or more modalities are unavailable. Finally, collaboration between AI researchers, mental-health professionals, and affected communities is essential for determining whether computational outputs are clinically meaningful and ethically appropriate.

11. CONCLUSION

The descriptive synthesis demonstrates that depression-detection research has progressed from individual behavioural signals and conventional machine-learning techniques toward increasingly sophisticated multimodal architectures. Text, audio, and video remain the principal modalities in clinical-interview research, while physiological and passive-sensing information are expanding the scope of multimodal assessment. Recurrent neural networks, attention mechanisms, Transformers, graph neural networks, contrastive learning, and large language models represent successive developments in the effort to learn richer relationships among heterogeneous data sources.The literature also demonstrates that multimodal integration should not be interpreted as automatically producing reliable clinical diagnosis. Dataset limitations, demographic and cultural variation, class imbalance, privacy, interpretability, missing modalities, and external validation remain substantial challenges. Systematic reviews consistently emphasize heterogeneity in datasets, features, clinical labels, and evaluation procedures, limiting direct comparison of published performance. The major contribution of multimodal machine learning is therefore its potential to integrate complementary behavioural and physiological information into a broader computational representation. The field is moving toward increasingly contextual, adaptive, and representation-rich systems, particularly through attention, Transformer, graph, contrastive, and LLM-based architectures

REFERENCES

1. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10–49.
2. https://doi.org/10.1016/j.specom.2015.03.004
3. Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., & Yu, J. (2024). Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion, 104, 102161.
4. https://doi.org/10.1016/j.inffus.2023.102161
5. Fan, W., He, Z., Xing, X., Cai, B., & Lu, W. (2019). Multi-modality depression detection via multi-scale temporal dilated CNNs. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (pp. 73–80). ACM. https://doi.org/10.1145/3347320.3357695
6. Fang, M., Peng, S., Liang, Y., Hung, C.-C., & Liu, S. (2023). A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control, 82, 104561. https://doi.org/10.1016/j.bspc.2022.104561
7. Flores, R., Tlachac, M. L., Toto, E., & Rundensteiner, E. (2022). AudiFace: Multimodal deep learning for depression screening. In Proceedings of the 7th Machine Learning for Healthcare Conference (Vol. 182, pp. 609–630). PMLR.
8. Fu, C., Qian, F., Su, Y., Su, K., Song, S., Niu, M., Shi, J., Liu, Z., Liu, C., Ishi, C. T., & Ishiguro, H. (2025).

OPEN ACCESS © CINEFORUM

9. Facial action units guided graph representation learning for multimodal depression detection. Neurocomputing, 619, 129106. https://doi.org/10.1016/j.neucom.2024.129106
10. He, M., Bakker, E. M., & Lew, M. S. (2024). DPD (DePression Detection) Net: A deep neural network for multimodal depression detection. International Journal of Medical Informatics, 191, 105593.
11. https://doi.org/10.1016/j.ijmedinf.2024.105593
12. Hossain, M. S., & Muhammad, G. (2018). Emotion recognition using deep learning approach from audio-visual signals. IEEE Access, 6, 10871–10882.
13. Hou, R., Teng, S., Liu, J., Chai, S., Li, Y., Lin, L., & Chen, Y.-W. (2025). Retrieval-augmented multimodal depression detection. In 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 1–5). IEEE. https://doi.org/10.1109/EMBC58623.2025.11254056
14. Hussain, Y., Zaheer, M. A., Khan, A. M., & Malik, A. S. (2026). Depression detection using deep learning and large language models from multimodalities. Frontiers in Digital Health, 8, 1759857.
15. https://doi.org/10.3389/fdgth.2026.1759857
16. Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. Sensors, 24(2), 348.
17. https://doi.org/10.3390/s24020348
18. Li, M., Wei, Y., Zhu, Y., Wei, S., & Wu, B. (2024). Enhancing multimodal depression detection with intra- and inter-sample contrastive learning. Information Sciences, 684, 121282.
19. https://doi.org/10.1016/j.ins.2024.121282
20. Lim, E., Jhon, M., Kim, J.-W., Kim, S.-H., Kim, S., & Yang, H.-J. (2025). A lightweight approach based on cross-modality for depression detection. Computers in Biology and Medicine, 186, 109618.
21. https://doi.org/10.1016/j.compbiomed.2024.109618
22. Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. Computer Methods and Programs in Biomedicine, 211, 106433. https://doi.org/10.1016/j.cmpb.2021.106433
23. Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019). Multi-level attention network using text, audio and video for depression prediction. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (pp. 9–16). ACM. https://doi.org/10.1145/3347320.3357697
24. Ye, J., Yu, Y., Wang, Q., Li, W., Liang, H., Zheng, Y., & Fu, G. (2021). Multi-modal depression detection based on emotional audio and evaluation text. Journal of Affective Disorders, 295, 904–913.
25. https://doi.org/10.1016/j.jad.2021.08.090
26. Yin, S., Liang, C., Ding, H., & Wang, S. (2019). A multi-modal hierarchical recurrent neural network for depression detection. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (pp. 65–71). ACM. https://doi.org/10.1145/3347320.3357696
27. 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

OPEN ACCESS © CINEFORUM