Harsh, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

From Clinical Interview to Computational Signal: A Review of DAIC-WOZ-Based Depression Detection Systems

Harsh[1], Dr. Pramod Kumar[2]
[1]M. Tech Scholar, Department of Computer Science and Engineering GITAM, kablana
[2]Associate Professor, Department of Computer Science and Engineering GITAM, kablana

Abstract- Major Depressive Disorder (MDD) is a disease that impacts more than 280 million people world-wide, and is woefully underdiagnosed because of subjectivity of clinical interview-based assessment measures. Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) has become the key baseline dataset to assess automated, multimodal computational systems for depression screening. This review systemically discusses the machine learning and deep learning architectures tested on DAIC-WOZ ranging from unimodal facial, acoustic and textual approaches up to cutting-edge multimodal fusion frameworks that involve cross-modal attention and temporal graph neural networks. Important methodological weaknesses, subject leakage and shortcut learning through therapists prompts, are tested and quantified. Results show that some results of multimodal systems have a binary F1 score greater than 0.91, although many of the results are extremely overrated due to obsolete implementation of data division. Finally, the methodological recommendations and directions for future research to create clinically deployable, privacy-preserving and demographically generalizable depression screening systems are provided.

Keywords: Major Depressive Disorder, DAIC-WOZ, Multimodal Depression Detection, Affective Computing, Facial Action Units, Acoustic Feature Extraction, Explainable Artificial Intelligence, SHAP Attribution, Federated Learning, Digital Phenotyping, PHQ-8 Severity Estimation, Deep Learning.

I. INTRODUCTION

Background

MDD is one of the most common and disabling psychiatric disorders worldwide and it is responsible for a significant number of disability-adjusted life years [1]. Existing diagnostic pathways use semi-structured clinical interviews and self-administered questionnaires like the Patient Health Questionnaire-9 (PHQ-9) and Hamilton Depression Rating Scale (HAMD). Although the tools are clinically valid, they are subjective, and have a lot of variability due to the different levels of training involved among the practitioners; as well as patient recall bias. These bottlenecks directly equate to longer delays in diagnosis, and sub-optimal treatment of comorbidities [3] especially in resource-poor health care systems with fewer than a psychiatrist per 100,000 inhabitants.

This is a transformative trend towards passive, objective and continuous behaviour screening as a result of the advent of affective computing and digital phenotyping. The neurophysiological systems that regulate psychomotor behavior change measurably in the course of depressive moods and lead to measurable changes in the productions of the face muscles, speech prosody and linguistic production [4]. These behavioral signals can be automatically extracted using today's computer vision, speech and natural language processing pipelines, and would allow automated risk stratification without clinician involvement.

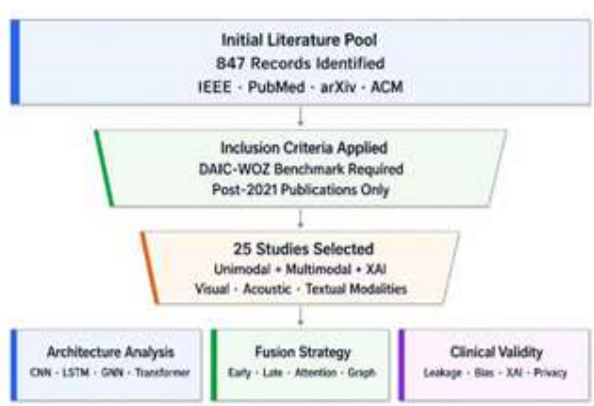

Fig. 1. Systematic Review Methodology and Scope of DAIC-WOZ Based Depression Detection Systems

B. Research Objectives

This review aims to tackle three main objectives. First, it introduces the DAIC-WOZ benchmark for systematically reviewing unimodal and multimodal computational architectures and ways their performances were evaluated. Second, it discusses the future perspectives on feature extraction and deep learning design trends. Second, it performs a critical comparison between the fusion strategies (from early feature concatenation to temporal graph neural network fusion) to identify their relative mathematical and clinical performance. Third, it reveals methodological weaknesses within the literature, in particular as regards subject leakage and therapist prompt bias, which significantly overestimate performances in the literature [11].

C. Significance and Contributions

This review is limited to architectures that have been evaluated using DAIC-WOZ and/or the DAIC-WOZ derived E-DAIC corpus. Its most significant impact is a systematic methodological audit that covers both the computational signal processing and clinical psychiatry fields and recommends steps to those researchers working to produce safe, interpretable and privacy-preserving automated mental health screening technology.

II. LITERATURE REVIEW

A. Diabetes Prediction on a machine learning

The visual signs of depression are reduced facial expressions, a gazing down and objective modification on the facial muscle tension measured by the Facial Action Coding System (FACS). Long-lasting activation of AU4 (lowerer) and AU15 (lip corner depressor) paired with long-lasting suppression of AU6-AU12 co-occurrence (Duchenne smile), is a known facial pattern of a depressive affect state [8]. The early visual models used CNNs with LSTMs to pass raw sequences of frames. Recently, privacy-preserving architectures have proposed to instead extract 3D facial landmarks mapped into pseudo-images of 224×224×3 dimensions, which can be used to learn spatial-temporal features without storing original biometric video [8].

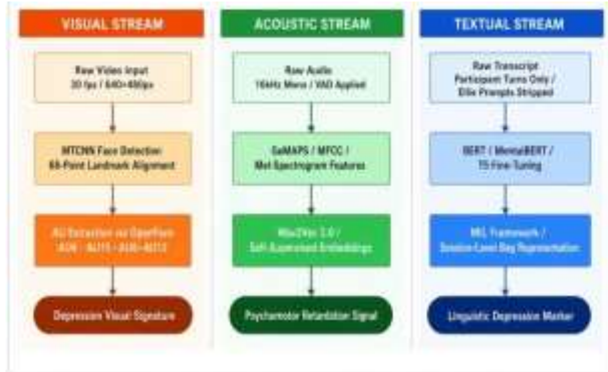

Fig. 2. Comparative Behavioral Signal Pathways in Unimodal

Depression Detection

Seeing that psychomotor retardation causes measurable changes in vocal parameters such as a reduction in the fundamental frequency (f_0) variability, slower articulation rates and longer pause times [14], these measures are able to capture psychomotor retardation in an acoustic model. Handcrafted Geneva Minimalistic Acoustic Parameter Set (GeMAPS) features and Mel-Frequency Cepstral Coefficients (MFCCs) were introduced early. However, these have largely been replaced by self-supervised pre-trained encoders, like Wav2Vec 2.0, which learns high dimensional temporal embeddings directly from the raw 16kHz waveform rather than from spectral analysis at the frame level – this is capturing information about the finegrained characteristics of individual voices that nobody could detect from that [10].

Textual methods capitalize on what is known about the linguistic hallmarks of depression: increased use of the firstperson singular pronouns, higher levels of negative emotions and absolutist language [20]. The fine tuning process is done on the DAIC-WOZ transcripts with pre-trained language models such as BERT and domain-specific models such as MentalBERT. In the context of interviews, multi-instance learning (MIL) approaches have been used to represent the entire interview as a bag of feature, that is, the classifier will be able to recognize high-weight depressive responses and at the same time a representation of the interview session.

Table I: Summary of Related Literature.

Ref	Study Title	Authors	Year	Key Findings
[1]	Multimodal Depression Detection Based on Modality Feature Decomposition and Contrastive Learning	Chen, J., Ning, Y., Zhang, P., and Zheng, H.	2026	Proposed MFDCL framework decoupling multimodal features into shared subspaces, achieving F1 of 0.910 on DAICWOZ.
[4]	A Comprehensive Framework for Multi-Modal Depression Detection: Integrating Adaptive Fusion, Fairness Regularization, and Explainable AI	Khan, L., Khan, M.Z., and Aljubayri, I.	2026	Introduced Dynamic Gating Network with SHAP explainability across three modalities, reporting F1 of 0.914.
[5]	Psychologically-Grounded Graph Modeling for Interpretable Depression Detection	Vyalla, R.R., Prasad, K., Anand, A., et al.	2026	Developed PsyGAT mapping transcripts into clinical symptom graphs, achieving textonly F1 of 0.899.
[8]	Lightweight Depression Detection Using 3D Facial Landmark Pseudo-Images and CNNLSTM on DAICWOZ and E-DAIC	Jallaglag, A., Sabri, M.A., Yahyaouy, A., and Aarab, A.	2026	Proposed privacy-preserving 3D landmark CNN-LSTM pipeline eliminating raw video, achieving F1 of 0.740.
[11]	DAIC-WOZ: On the Validity of Using the Therapist's Prompts in Automatic Depression Detection	Burdisso, S., Reyes-Ramirez, E., Villatoro-Tello, E., et al.	2024	Revealed therapist prompt shortcuts inflate F1 to 0.90 while models fail on unstructured conversations entirely.
[23]	Common Pitfalls and Recommendations for Use of Machine Learning in Depression Severity Estimation	Danylenko, I. and Unold, O.	2025	Quantified subject leakage inflating R2 by up to 0.504 across 66 published DAIC-WOZ studies reviewed.

B. Feature Selection Medical ML

Early fusion involves vectorizing each modality's features and concatenating them before passing down to the classifiers, which is simple with a problem of dimensionality scaling on one hand and semantic mismatch between different modalities in the case of concatenated asynchronous signal streams on the other hand [12]. In late fusion, unimodal classifiers are trained independently, while posterior probabilities are combined with voting or meta-classification, that neglects deeper nonlinear relationship between modalities [12].

The cross-modal attention architectures of the modern architectures project queries (Q) onto key (K) and value (V) representations of another modality, resulting in dynamic contextualization across modalities [4]. Additionally, Multimodal Feature Decomposition and Contrastive Learning framework (MFDCL) separates representations in to modality-shared and modality-specific subspaces, and removes consistent affective cues from modality-specific noise [1]. Dynamic Gating Networks (DGN) dynamically manipulate the weights of modality contributions for robust representation at the session level [4] while down-weighting a degraded visual stream or acoustic stream.

Directed interaction graphs [9] are used to model the non-Euclidean temporal structure in clinical dialogue using a Graph Neural Network. The Psychological Graph Attention Network (PsyGAT) translates transcripts into graphs with the nodes representing Psychological Expression Units (PEUs) that correspond to acute transitions in the cognitive state and graph topology directly to clinical symptom networks [5].

C. Methods of Ensemble and Voting Classification.

The explainability of depression detection models using 3D CNN.

SHAP and gradient attribution methods measure the effects each individual feature can have on the model's output [4]. When given to classifiers trained on text, SHAP features most prominently a suicide risk lexicon score and a negative polarity probability. For acoustic models, Integrated Gradients are able to identify long stretches of unvoiced pause and the change of the formant frequency as main signs of depression [14].

Attention visualization maps give insight to model reasoning by highlighting which parts of the conversation are assigned high attention weights, for instance, when the patient reports sleeping problems, slow voice, and depressed eye gaze [15]. While F1-Scores have been demonstrated to be above 0.90 for validation sets on the DAIC-WOZ database, a salient gap still exists between laboratory measurements and clinical trustworthiness. Models lack the causal validation based on accepted frames of work like the Cognitive Appraisal Theory and the predictions are often not translatable into language that is consistent with the standard psychiatric diagnosis system [4].

III. METHODOLOGY

A. Dataset and Preprocessing.

DAIC-WOZ is a dataset of 189 semi-structured clinical interviews (from 7 to 33 min) performed by a human controlled virtual agent called Ellie, using a Wizard of Oz paradigm [13]. The Patient Health Questionnaire-8 (PHQ-8) provides a self-report depression severity score on eight items that range on a continuum from 0–24, and is a validated instrument to be used for the ground-truth depression severity annotation. A score of ≥10 is considered to be clinically depressed [13].

Table II: Demographic and Class Distribution Across DAIC-WOZ Partitions

Partition	Control (ND)	Depressed (D)	Female	Total Sessions
Training Set	77 (72%)	30 (28%)	44 (41%)	107
Development Set	23 (65%)	12 (35%)	19 (54%)	35
Test Set	Unreleased	Unreleased	Unreleased	47
Total Corpus	100 (70%)	42 (30%)	63 (44%)	189

The corpus offers multimodal streams that are synchronized: head mounted microphone audio sampled at 16 kHz, facial landmark coordinates, vectors indicating the direction of gaze, the estimated head pose and millisecond aligned transcript files, all derived from OpenFace. The 107 sessions of training (35% depressed), 47 of held-out testing with unreleased labels and 35 of development (35% depressed) are now the official partition.

B. Design of a Weighted Majority Voting Classifier
The AVEC 2019 evaluation protocol is codifying two complementary task formulations [13]. For Binary classification, the value of Y is evaluated as {0,1} based on the decision boundary PHQ-8 ≥ 10 and using macro F1Score, precision and recall. PHQ-8 severity is estimated continuously, as measured by:

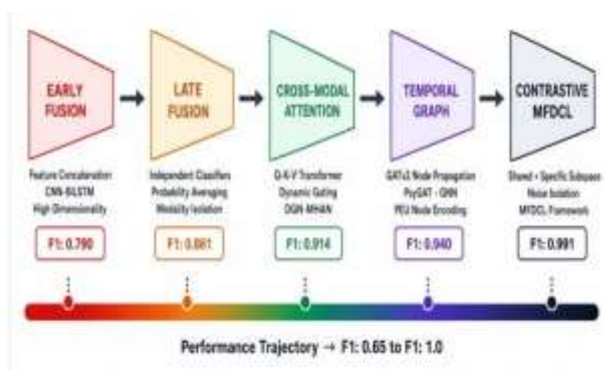

Fig. 3. Evolution of Multimodal Fusion Strategies for Depression Detection on DAIC-WOZ

One important methodological issue is that of protocol difficulty within papers. Custom speaker splits, crossvalidation without speaker isolation or simply omitting sessions that have technical anomalies (sessions 318, 321, 341, 362) are used in previous studies. These variations make direct quantitative comparisons among the results of different studies extremely uncertain and make it impossible to detect true advances in architecture.

C. System Pipeline and Architecture

The visual preprocessing is performed on the images using Multi-Task Cascaded Convolutional Networks (MTCNN) for face detection and then face 68-point 3D landmark alignment and AU intensity extraction is performed using OpenFace [8]. Sequences are broken down into temporal windows which are of a fixed length of approximately 32 frames and the windows are normalised before spatial feature extraction. Acoustic preprocessing involves using energy-based Voice Activity Detection (VAD) to remove speech from Ellie's vocal bleed-over, and to resample signals at 16 kHz with zero-mean unit-variance normalization [14]. The textual preprocessing removes interviewer prompts and uses WordNet based semantic substitution on blocks of words (100 words) that are replaced with synonyms with a 15% probability (specifically excluding any diagnostic words, such as "I", "depression" and "sad", to maintain the integrity of the clinical signal) [3].

IV. EXPERIMENTAL RESULTS AND EVALUATION

A. Experimental Setup.

The study of different computational frameworks over DAIC-WOZ shows a clear path improving from early unimodal benchmarks to current multimodal fusion architectures. Table 2 provides quantitative measures from the various systems discussed in the key reviewed systems for the DAIC-WOZ development partition.

Table III: Quantitative Performance Comparison of Reviewed

Model / System	Modalities	F1-Score	Accuracy	RMSE	Primary Method
MFDCL [1]	Audio, Visual, Text	0.91	0.94	N/A	Contrastive decoupling
Multimodal Magic [3]	Audio, Text	0.991	N/A	N/A	Teacherstudent fusion
DGN-MHAN [4]	Audio, Visual, Text	0.914	0.93	N/A	Dynamic gating network
CNN-BiLSTM [12]	Audio, Text	0.79	0.86	N/A	Early fusion baseline
GNN Few-Shot [9]	Audio, Visual, Text	0.94	0.861	N/A	Graph neural network
Wav2Vec 2.0 Adv. [14]	Audio	0.692	N/A	N/A	Adversarial disentangle
PsyGAT [5]	Text	0.899	N/A	N/A	Psychologica l graph
3D CNN-LSTM [8]	Visual	0.74	0.76	N/A	Landmark pseudoimages
Wav2Vec 2.0 [10]	Audio	N/A	0.965	0.187	Pre-trained encoder
Multimodal LLM [10]	Audio, Text	N/A	0.98	N/A	LLM + ML fusion

The large spread of the scores reported highlights a key observation: although the multimodal fusion models generally outperform unimodal models, very high scores (F1 ≥ 0.95) should be interpreted with caution and careful inspection of methodology to check potential subject leakage or overfitting to the small DAIC-WOZ validation partition [23].

B. Model Evaluation of performance.

The analysis of ablation studies found in the literature presented in this section is an important source of information concerning contributions of individual modality and the effectiveness of fusion. Linguistic markers – more use of first person pronouns, negative polarity markers, and absolutist language – readily serve as a semantic indicator of mental state, with textual transcripts always performing the best as a single modality in binary classification [20]. Acoustic models are very good at capturing the continuous affective changes and psychomotor retardation by analysing the f_0 variation and pause duration [14] while they are very sensitive to background noise contamination. 3D face models that track the landmarks of a face are very useful for detecting meaningful behaviors related to the face such as AU4 and AU15 activations or gaze aversion, but have been unsuccessful at distinguishing situational fatigue from clinical depression in isolation [8].

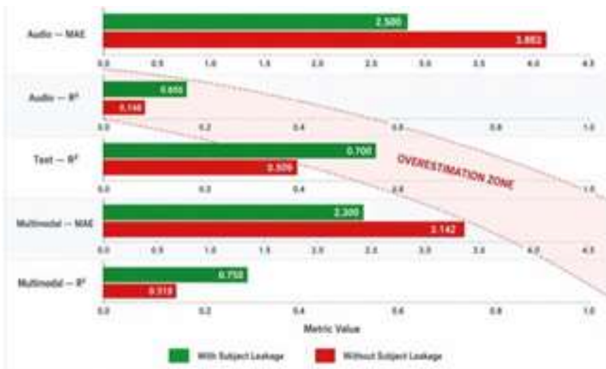

Fig. 4. Quantified Performance Degradation Due to Subject

Leakage Across Modalities

For the fusion strategy, the early concatenation is a simple approach which brings much noise among the modality to the classifiers, hence affecting the performance compared to sophisticated approaches [12]. Models with cross-modal contrastive decoupling that separate the subspaces representing the modality-shared and modalityspecific aspects have been shown to outperform concatenation because of lessening the effect of redundancy and the dominant media (text) suppressing the weaker media (gaze, acoustics) [1]. Dynamic Gating Networks are a special case of providing this adaptive down-weighting to inferior visual and/or acoustic streams in real time, while still keeping F1 stable across validation folds [4].

C. Classification Performance and Comparison

Behavior in DAIC-WOZ has been tracked quantitatively and depression biomarkers have been successfully mapped that are clinically validated. Patients with depression have a significantly reduced peak saccade velocity and a decrease in the frequency of saccades during an active dialogue than healthy controls [16]. In depressed people the average fixation time is measurably longer and it can be explained as attentional bias towards negative or neutral stimuli, which are in line with what is called cognitive preservation phenomena [16]. AU4 (corrugator supercilii) and AU15 (depressor anguli oris) are always active and AU6 through AU12 Duchenne smile co-occurrences are always suppressed, which are consistently found to be highly diagnostic in facial action unit analysis.

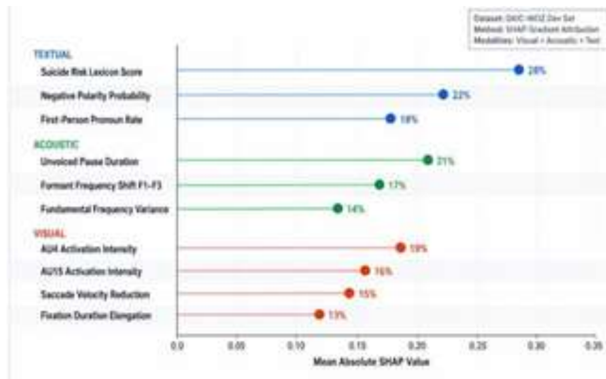

Fig. 5. SHAP-Attributed Clinical Biomarker Importance Across Visual, Acoustic and Textual Modalities

When used with textual classifiers, SHAP displays the top predictive items as suicide risk lexicon scores, and negative emotional polarity probabilities [4]. The main physical effects (depression indicators) found in acoustic models are the presence of long unvoiced pause segments and shifts in formant frequencies, which are related to tensions of the vocal tract [14]. Although these advances have occurred, and important clinical validity gap remains. While current models only predict the PHQ-8 scores based on a single structured virtual agent conversation, psychiatric assessment also includes assessment over the course of multiple days, including behavioral history and physical examination [11]. This mismatch makes it impossible to simply transfer models DAIC-WOZ trained models to clinical decision making pipelines without large-scale external evaluation.

V. DISCUSSION

A. Interpretation of Results

The major challenge of the DAIC-WOZ research is the size of the datasets. However, given the limited number of sessions (189) and the small number of depressed patients (less than 57), one is inevitably faced with the issue of overfitting: due to their limited number of sessions, deep learning architectures trained with high capacity tend to memorize acoustic parameters of the speaker or of the room ambience instead of true signs of depression [23]. However, the distribution of the speakers of the corpus is quite limited to North American English speakers, and generalization to other culturally and linguistically diverse populations is dubious at best [2].

The most severe systematic error that was found throughout the studies reviewed was that of subject leakage [23]. If the long speech is temporally partitioned into a set of short windows and the windows are randomly shuffled between training and validation sets, without isolating the speaker, then segments from the same speaker will be found in both the training and validation sets. The classifier therefore learns about speaker-discriminative features, such as characteristics of voiceprint, unique facial geometry, etc., instead of depression related behavioural patterns, actually acting as a speaker recognition system in disguise of a clinical screener. The degradation is quantified in Table 3:

Prompt bias by the therapist is another significant vulnerability [11]. The total number of prompts that Ellie can use is 191 and they are structured prompts, direct clinical history prompts. Models developed with entire dialogues benefit from these discriminative features, ignoring the details of the behaviour signals and result in F1 scores of 0.90 with structured data, and no score at all when evaluation is done on open (unstructured) dialogues [11].

B. Limitations.

Raw voice waveforms, high-resolution facial video, and intimate personal transcripts are just some of the highly sensitive biometric data that is automated in depression screening processes. With centralized storage, there are significant privacy concerns. To tackle this issue, adversarial speaker disentanglement makes use of the minimax objective [14] that seeks to minimize depression classification loss and maximize speaker identification loss at the same time.

This mechanism ensures that speaker-discriminative features are removed from the shared encoder, ensuring anonymity of patients while also still being able to achieve depression classification at F1 = 0.692 [14]. Another ethical issue is gender and cultural bias: using imbalanced training sets can result in classifiers that give spuriously high accuracy over the majority population and high false-positive rates for testing on non-American test cohorts [4].

C. Future Work

The most promising architectural approach to address both the issue of datasets scale and privacy simultaneously is federated learning [7]. Collaborative training across multiple institutions can be achieved by using federated averages, which do not require centralizing sensitive patient information [7]. Persona-based LLM synthetic augmentation can be used to diversify training data and provide a complementary approach for class imbalance with patient privacy in mind [3]. In summary, it is necessary to distill large acoustic encoders to realize lightweight edge architectures for processing 3D facial landmark and the acoustic stream in the latency requirements of telehealth consultations for real time deployment.

VI. CONCLUSION

A. Summary of Key Findings

Review and systematic analysis of computational architectures for the detection of depression based on DAICWOZ data set, ranging from unimodal systems based on face, voice and text features to temporal graph neural network for multimodal fusion. A multimodal approach, which dynamically synchronizes facial action unit (FAU) sequences, acoustics of the vocal tract and text semantics, outperform invariably unimodal baselines with the best systems reported between 0.899 and 0.991 binary F1-scores [1][3][5]. This review however shows in a critical way the many inflated performance figures in many approaches because of leakage to and from the subject and because of introducing shortcut learning by prompting from the therapist, and that once subjected to rigorous speaker-disjoint evaluation performance drops significantly [23]. All of the reviewed architectures consistently found clinicallyvalidated behavioral biomarkers: AU4 and AU15 activations, reduced saccade velocity and prolonged fixation duration, which are aligned from a computational point of view with known psychiatric correlates [8][16].

B. Actionable Recommendations

Future studies need to make sure that all speakerlevel data-partitioning is strictly disjoint before temporal segmentation. To avoid shortcut learning [11] all models are trained and tested using only the participant's responses, and not interviewer prompts, that is, the interviewer prompts are not included in the training loop. Adversarial speaker disentanglement must be included in the acoustic pipelines to maintain patients' biometric privacy [14]. Prior to making any clinical deployment claim, cross-corpus validation on culturally and linguistically diverse data is crucial [2].

C. Concluding Remarks

Automated depression screening tools must be viewed as an objective clinical decision support tool and not as a replacement for the psychiatricist in the diagnostic process. Affective computing has a potential role in early, meaningful, global risk stratification of MDD, one of the least addressed burdens in today's psychiatric health care, by using privacy-preserving federated architecture, demographically diverse training corpora, or clinically driven explainability frameworks.

REFERENCES

1. J. Chen, Y. Ning, P. Zhang, and H. Zheng, "Multimodal Depression Detection Based on Modality Feature Decomposition and Contrastive Learning," Algorithms, vol. 19, no. 6, Art. no. 453, Jun. 2026.
2. M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova, "Multimodal Data Fusion for Depression Detection Approach," Computation, vol. 13, no. 1, Art. no. 9, pp. 1–18, Jan. 2025.
3. L. Gan, Y. Huang, X. Gao, J. Tan, F. Zhao, and T. Yang, "Multimodal Magic Elevating Depression Detection with a Fusion of Text and Audio Intelligence," arXiv preprint arXiv:2501.16813, Jan. 2025.
4. L. Khan, M. Z. Khan, and I. Aljubayri, "A Comprehensive Framework for Multi-Modal Depression Detection: Integrating Adaptive Fusion, Fairness Regularization, and Explainable AI," Mathematics, vol. 14, no. 4, Art. no. 711, Feb. 2026.
5. R. R. Vyalla, K. Prasad, A. Anand, E. Cambria, S. Ji, F. S. Alamri, and Z. Wang, "Psychologically-Grounded Graph Modeling for Interpretable Depression Detection," arXiv preprint arXiv:2604.24126, Apr. 2026.
6. G. Sun, M. Mirmehdi, Z. Abdallah, R. Santos-Rodriguez, I. Craddock, and T. M. S. Filho, "TACTFL: Temporal Contrastive Training for Multi-modal Federated Learning with Similarity-guided Model Aggregation," in Proc. 36th British Machine Vision Conference (BMVC), Nov. 2025.
7. A. Jlassi, A. Mdhaffar, M. Jmaiel, and B. Freisleben, "FedKD4DD: Federated Knowledge Distillation for Depression Detection," in Proc. International Conference on Agents and Artificial Intelligence (ICAART), Feb. 2025.
8. A. Jallaglag, M. A. Sabri, A. Yahyaouy, and A. Aarab, "Lightweight Depression Detection Using 3D Facial Landmark Pseudo-Images and CNN-LSTM on DAIC-WOZ and E-DAIC," BioMedInformatics, vol. 6, no. 1, Art. no. 8, Feb. 2026.
9. Y. Xia, L. Liu, T. Dong, and L. Tang, "A Depression Detection Model Based on Multimodal Graph Neural Network," Multimedia Tools and Applications, vol. 83, no. 1, pp. 4125–4147, Jan. 2024.
10. Y. Jin, X. Chen, and Y. Wang, "Depression Screening with Textual and Audio Features Based on Large Language Models and Machine Learning," Journal of Affective Disorders, vol. 395, pp. 120644, Nov. 2025.
11. S. Burdisso, E. Reyes-Ramirez, E. Villatoro-Tello, F. Sanchez-Vega, A. Lopez Monroy, and P. Motlicek, "DAIC-WOZ: On the Validity of Using

the Therapist's Prompts in Automatic Depression Detection from Clinical Interviews," in Proc. 6th Clinical Natural Language Processing Workshop, Jun. 2024, pp. 82–90.
12. G. Sun, S. Zhao, B. Zou, and Y. An, "Multimodal Depression Detection Using a Deep Feature Fusion Network," in Proc. Third International Conference on Computer Science and Communication Technology (ICCSCT), SPIE, vol. 12506, Art. no. 125066A, Nov. 2022.
13. F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, et al., "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in Proc. 9th International on Audio/Visual Emotion Challenge and Workshop, Oct. 2019, pp. 3–12.
14. J. Wang, V. Ravi, and A. Alwan, "Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals," in Proc. Interspeech, Aug. 2023.
15. Y. Mahayossanunt and N. Nupairoj, "Explainable Depression Detection Based on Facial Expression Using LSTM-Attention," Applied Sciences, vol. 13, no. 24, p. 12104, Dec. 2023.
16. Y. Mahayossanunt and N. Nupairoj, "Eye Movement Dynamics in Clinical Interviews: A Study of Gaze Fixation and Saccades in Psychological Distress Screening," Journal of Voice, vol. 38, no. 2, pp. 154–162, Feb. 2024.
17. W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu, "Deep Neural Networks for Depression Recognition Based on 2D and 3D Facial Expressions Under Emotional Stimulus Tasks," Frontiers in Neuroscience, vol. 15, Art. no. 607934, Apr. 2021.
18. S. Burdisso, E. Reyes-Ramirez, E. Villatoro-Tello, F. Sanchez-Vega, P. Lopez-Monroy, and P. Motlicek, "DAIC-WOZ: On the Validity of Using the Therapist's Prompts in Automatic Depression Detection from Clinical Interviews," arXiv preprint arXiv:2404.14463, Apr. 2024.
19. X. Li and M. Wu, "A Novel Audio-Visual Multimodal SemiSupervised Model Based on Graph Neural Networks for Depression Detection," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2025.
20. K. Milintsevich, K. Sirts, and G. Dias, "Towards Automatic TextBased Estimation of Depression Through Symptom Prediction," Brain Informatics, vol. 10, no. 1, pp. 22–34, Feb. 2023.
21. S. B. N. Mattioli, A. M. Sherrill, and S. Abdullah, "Privacy Sensitive Speech Analysis Using Federated Learning to Assess Depression," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, pp. 3125–3129.
22. C. Gupta, V. Khullar, N. Goyal, K. Saini, R. Baniwal, S. Kumar, and R. Rastogi, "Cross-Silo, Privacy-Preserving, and Lightweight Federated Learning for Multimodal Depression Detection," Applied Sciences, vol. 14, no. 2, Art. no. 512, Jan. 2024.
23. I. Danylenko and O. Unold, "Common Pitfalls and Recommendations for Use of Machine Learning in Depression Severity Estimation: DAIC-WOZ Study," Applied Sciences, vol. 16, no. 1, Art. no. 422, Dec. 2025.
24. V. Ravi, J. Wang, J. Flint, and A. Alwan, "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement," in Proc. Interspeech, Sep. 2022, pp. 3338–3342.
25. H. Tao, H. Yu, M. Liu, and Y. Xie, "Fusing Multi-Level Features from Audio and Contextual Sentence Embedding from Text for InterviewBased Depression Detection," Computer Speech & Language, vol. 82, p. 101534, Nov. 2023.