International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Multimodal Sentiment Analysis for Mental Health Detection

Vanetta P. Silveira
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Reva L. Gaonkar
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Riya N. Padwalkar
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Sanisha Faria
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Louella M. Colaco
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Supriya Patil
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Meghana Pai Kane
Department of Computer Engineering
Padre Conceição College of Engineering
Verna, Goa, India

Abstract - **Mental health is a critical component of overall well-being, yet conditions such as depression are frequently underdiagnosed due to the subjective and time-intensive nature of traditional assessment. This paper proposes a multimodal approach for automated depression detection that combines textual and acoustic representations extracted from patient interviews. A BERT-based encoder, enriched with auxiliary sentiment features, is used to model linguistic patterns, while a WavLM-based encoder captures acoustic and prosodic cues from speech. Text and audio embeddings are combined using a learned gated fusion mechanism at the utterance level, and a Bidirectional Gated Recurrent Unit further models temporal dependencies at the session level. The system was evaluated on the E-DAIC WOZ dataset, where the text-only model achieved 73.21% accuracy for depression detection, and the fusion model obtained the highest precision (72.73%) for identifying individuals with depression. The auxiliary sentiment classification head further attained an accuracy of 87.34% at the utterance level, demonstrating the benefit of jointly modelling sentiment and depression cues through complementary textual and acoustic modalities.**

Keywords *- multimodal sentiment analysis; depression detection; mental health; BERT; WavLM; multimodal fusion*

I. INTRODUCTION

Mental health is a state of well-being that allows one to deal with life's stresses, to realize their potential, to perform duties normally and to be an asset to the society. Chisholm et al. [1] stated that without mental well-being there is no true physical health, highlighting the strong interdependence between physical and mental well-being. Mental health is often overlooked in comparison to physical health. An imbalance between the two contributes to many physical and mental disorders.

According to a survey conducted by the World Health Organization (WHO) in 2021, 1 in 7 individuals globally are affected by mental disorders, with anxiety and depression being the most prevalent. Anxiety and depression alone affect over 359 and 280 million people respectively, with prolonged untreated depression linked to increased suicide risk [2– Despite the severity, these disorders are often overlooked due to various factors, particularly stigma and lack of awareness [1].

Mental disorders are diagnosed through various approaches, primarily relying on clinical interviews and self-reporting questionnaires. Some of the detection methods include the Patient Health Questionnaire (PHQ-8/9), the Beck Depression Inventory (BDI-II), the Hamilton Depression Scale, and the Center for Epidemiological Studies Depression Scale (CES-D) [5]. Nevertheless, there is a lack of focus on early detection of men-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

tal disorders; hence, depression and suicidal tendencies often remain undetected at the initial stages. Affected individuals often show unclear clinical characteristics, making traditional manual diagnosis complex, time-consuming, and potentially biased [5]. The shortage of professionals in the field of psychiatry makes the situation even more critical [6–8].

Over the years, researchers have identified w ays t o mitigate mental health issues using Artificial Intelligence (AI) techniques [9–11]. Natural Language Processing (NLP) plays a crucial role in sentiment analysis (SA), gaining popularity as a powerful diagnostic tool with the capability of early detection of emotions and disorder symptoms.

Despite significant progress in SA, most existing systems remain unimodal, predominantly using text. Consequently, they miss important characteristics that can be obtained using a multimodal approach. Recent studies explored systems that utilize social media data, consisting of tweets, texts, and images [12, 13]. These approaches proved effective in monitoring mental health, particularly among the youth. However, these methods may not be practical, since not all users are active on social media or have access to it. Hence, there is a need for an automated system that integrates multiple modalities to detect the sentiments of individuals.

This paper proposes a multimodal system that integrates textual and audio data to address the mental health issue of depression. The proposed system is trained on the Extended Distress Analysis Interview Corpus-Wizard of Oz (E-DAIC WOZ) dataset, using the audio data and their corresponding text. The system aims to detect depression based on audio and text and classifies individuals as depressed or non-depressed.

The remainder of this paper is organized as follows. Section II reviews recent works that motivate this study. Section III presents the proposed system and methodology. Sections IV and V present and discuss the experiments conducted along with the results. Finally, Section VI concludes the paper.

II. RELATED WORKS

SA is an NLP application that has been extensively researched and widely applied in various domains. It has significantly contributed to the field of mental health, where sentiments are used in diagnosing an individual's mental state. However, existing SA systems are largely unimodal, with text being the most common input.

Initially, traditional machine learning (ML) methods such as Logistic Regression (LR), k-Nearest Neighbors (k-NN), and Naive Bayes (NB) formed the foundation of most existing approaches. Alvarez et al. [14] employed six ML methods on Electronic Health Records (EHRs), where Random Forest (RF) classifier gave the best performance (Table I). In parallel, social media emerged as a valuable source for mental health detection and analysis [12]. Detection of depression and Post-Traumatic Stress Disorder (PTSD) among Twitter users was explored through sentiment analysis of tweets. Multiple ML algorithms were implemented, where Decision Tree (DT) achieved the best performance, while RF performed least effectively [15]. Selva et al. [13] proposed a deep learning framework integrating word2vec and BERT embeddings with a Bi-LSTM classifier and knowledge distillation for depression and anxiety detection from social media data.

Advanced methods such as Deep Learning (DL), Large Language Models (LLMs), and transformer-based architectures were widely adopted [16–18]. Social media data sourced from Reddit and Twitter was used to predict multiple emotional states such as happiness, anger, surprise, sadness, and fear. A range of techniques, including LR, NB, RF, and Support Vector Machine (SVM), along with DL approaches such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), were employed, achieving a maximum accuracy of 93.28% [19]. Using 30-second speech segments, Zhang et al. [20] trained a deep learning classifier to distinguish depressed from non-depressed individuals, with evaluation spanning three corpora: Multimodal Open Dataset for Mental Disorder Analysis (MODMA), DAIC-WOZ, and the Chinese Outpatient Audio Depression Corpus (COAD).

Hassan et al. [21] identified depression using assessment scales such as the Hamilton Depression Rating Scale (HAM-D) and Personal Health Questionnaire (PHQ-8). ML methods, including SVM, Gaussian Mixture Model (GMM), and LR, along with deep learning models such as CNN, RNN, and LSTM, were applied. DL methods excelled in speech emotion recognition, and CNNs effectively captured acoustic features for speech classification. Jadhav et al. [22] utilized six audio and two video features, adopting a voting-based ensemble classifier comprising LR, Support Vector Classifier (SVC), RF, and Gradient Boosting (GB). Textual data was handled via a Mamdani fuzzy approach for textual analysis , achieving up to 99.54% accuracy through late fusion (Table I).

Consequently, researchers recognized that unimodal systems failed to capture important information available in modalities such as speech and visual data, leading to the development of multimodal systems that integrate multiple data sources to improve analysis and performance. Gandhi et al. [23] explored Multimodal Sentiment Analysis (MSA) using speech, visual, and textual data through fusion techniques, and presented various MSA architectures along with interdisciplinary applications and future research directions. Wang et al. [24] conducted sentiment analysis on the Weibo platform, classifying the data into positive and negative categories, using the SMP2020 Weibo Emotion Classification Evaluation dataset and weibo_senti_100k, comprising both textual and visual data.

Ikeuchi et al. [17] focused on depression detection from conversational data through fine-tuning. After converting speech to text via Whisper-large-v3, the authors benchmarked several LLMs, including Llama, Gemma, Mistral, and Qwen for the detection task. QLoRA-based fine-tuning was employed to reduce model parameters and optimize memory usage. Among these, Llama 3.1-8B Instruct achieved the strongest results. Liu et al. [25] proposed an AI-based mobile application, MoodEcho, for automatic speech-based depression detection. The system relied on Mel-spectrogram representations to train a CNN augmented with Squeeze-and-Excitation (SE) blocks. Iyortsuun et al. [26] introduced an Additive Cross-Modal Attention Network (ACMA), a Bi-LSTM architecture with cross-modal attention that fuses text and audio cues for depression classification, evaluated on the DAIC-WOZ dataset. It should be noted that accuracies reported on social-media-derived datasets [13, 19, 22] are

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

not directly comparable to those obtained on clinical-interview datasets such as DAIC-WOZ/E-DAIC, since social media labels are typically self-reported or keyword-derived, whereas E-DAIC WOZ relies on clinically validated PHQ-8 assessments [27], representing a more challenging and realistic detection task. Table I summarises the datasets, methods, and reported performance of the works discussed above.

III. METHODOLOGY

This work presents a multimodal approach for detecting depression through audio signals and their corresponding transcripts, along with extracted sentiments, as depicted in Fig. 1. The audio input is decomposed into two modalities, speech signals and textual transcripts. Both modalities undergo preprocessing to remove noise, eliminate irrelevant information, and improve data quality. Subsequently, feature extraction is performed to derive acoustic features from speech signals and contextual embeddings from transcripts using a pretrained BERT model.

The extracted features from both modalities are combined and fused to leverage complementary information, enabling a more effective understanding of the individual's mental and emotional state. Additionally, sentiment features are extracted from the textual data to capture emotional information, which further helps improve the overall performance of detecting depression.

A. Dataset

This work uses the E-DAIC WOZ dataset, a semi-clinical multimodal dataset introduced by DeVault et al. [28, 29], designed for automated mental health assessment. It comprises 275 semi-structured human-computer interviews conducted with a virtual assistant, operating in both Wizard-of-Oz (human controlled) and fully autonomous interaction modes. The dataset is split into training, development, and test sets while preserving speaker diversity in terms of age, gender, and depression severity based on the eight-item PHQ-8 [27]. The training and development sets contain a mix of human-controlled and AI-driven sessions, whereas the test set consists exclusively of fully automated interactions [28]. Each interview includes synchronized audio recordings, textual transcripts, and pre-trained multimodal features, with clinically validated labels such as PHQ-8 scores and PTSD severity measures.

B. Data Pre-processing

As the proposed system utilises multiple modalities as input, preprocessing was carried out separately for each modality.

1) Text Pre-processing: Textual transcripts were preprocessed to improve data quality while preserving semantically important linguistic information. The preprocessing pipeline involved merging the PHQ transcripts with train, development, and test splits using a common participant identifier to ensure proper alignment with labels.

2) Sentiment Feature Extraction: To enrich textual representations, sentiment-based features were extracted using a locally stored RoBERTa-based model. Long transcripts were segmented into overlapping chunks of up to 510 tokens with a stride of 256 tokens using a sliding window approach. Each chunk was processed independently to obtain sentiment probabilities such as negative, neutral, and positive.

3) Audio Pre-processing: Audio preprocessing aims to enhance signal quality, remove noise, and isolate the target speaker for effective feature extraction and model training. All audio recordings were resampled to 16 kHz for consistency, and non-relevant introductory and ending segments were removed. Speech regions were identified using WebRTC Voice Activity Detection (VAD) [30]. Speaker separation was performed by extracting embeddings using Resemblyzer [31], followed by K-means clustering [22]. Noise reduction included high-pass filtering, declipping, spike removal, and soft compression, followed by peak normalization.

C. Training

The proposed multimodal framework was trained separately on textual and audio modalities to effectively learn contextual, acoustic, and emotional patterns associated with depression.

1) Text-based Training: The text-based framework was trained using a multi-task learning strategy involving depression classification and auxiliary sentiment prediction. Cross-entropy loss was employed for both tasks, with lower weight assigned to the auxiliary sentiment loss. To address class imbalance, weighted random sampling with softened class weighting was adopted. The model was optimized using the AdamW optimizer with learning rate scheduling, warm-up, gradient clipping, dropout, and layer normalization. Since predictions were generated at the utterance level, session-level depression predictions were obtained using aggregation strategies such as mean probability, confidence-weighted mean, and median probability.

2) Audio-based Training: The audio-based framework was trained using acoustic and prosodic representations extracted using a WavLM-based encoder from the preprocessed speech signals. The extracted audio representations were used to train a depression classification model. Optimization was performed using the AdamW optimizer along with learning rate scheduling and gradient clipping. Utterance-level or segment-level predictions were aggregated to obtain session-level predictions.

D. Fusion

The proposed system adopts a two-stage fusion strategy that combines textual, sentiment, and acoustic representations to obtain a unified session-level depression prediction.

1) Utterance-Level Gated Fusion: At the utterance level, text and audio embeddings are combined using a learned gated fusion mechanism. The text branch produces a 128-dimensional embedding by concatenating four BERT pooling strategies, forming a 4×768-dimensional vector projected to 128 dimensions using a linear layer with LayerNorm and GELU. The audio branch produces a 128-dimensional embedding from WavLM frame-level hidden states pooled using an attention pooling layer. The two embeddings are concatenated and passed through a gating network:

$$g = \sigma(W_g[e_t; e_a] + b_g) \tag{1}$$

where e_t and e_a denote the text and audio embeddings respectively, W_g is a learned weight matrix, and σ is the sigmoid activation. The gate g is split into text-specific (g_t) and audio-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

TABLE I. Comparison of existing approaches for mental health detection.

Sr. No.	Author	Dataset	Method	Accuracy / F1-Score	Classification	SA Present
1	Alvarez et al. [14]	EHRs	RF, LR, k-NN, NB, SVM	0.72 / –	Binary	Yes
2	Tiwari et al. [15]	Twitter	DT, RF, ML methods	– / –	Binary	Yes
3	Selva et al. [13]	Reddit, Twitter	word2vec + BERT + Bi-LSTM	98.5% / –	Binary	Yes
4	Borah et al. [19]	Reddit, Twitter	LR, NB, RF, SVM, LSTM, BERT	93.28% / –	Multi-class	Yes
5	Hassan et al. [21]	DAIC-WOZ	SVM, GMM, CNN, RNN, LSTM	– / –	Binary	No
6	Zhang et al. [20]	MODMA, DAIC-WOZ, COAD	DL (speech, 30s samples)	– / 0.94, 0.67, 0.93	Binary	No
7	Jadhav et al. [22]	E-DAIC, PHQ-8	Ensemble (LR, SVC, RF, GB) + Mamdani	99.54% / –	Binary	No
8	Gandhi et al. [23]	Multiple	Multimodal Fusion (Speech + Visual + Text)	– / –	Multi-class	Yes
9	Wang et al. [24]	SMP2020, weibo_senti_100k	Multimodal (Text + Visual)	99.10% / –	Binary	Yes
10	Ikeuchi et al. [17]	DAIC-WOZ, PROMPT	Llama, Gemma, Mistral (QLoRA)	– / 0.84	Binary	No
11	Liu et al. [25]	DAIC-WOZ	CNN + SE blocks (Mel-spectrogram)	– / 0.86	Binary	No
12	Iyortsuun et al. [26]	DAIC-WOZ	ACMA + Bi-LSTM (Multimodal)	75.8% / –	Binary	No

specific (g_a) components. An element-wise interaction term is also computed, and all three are concatenated to form the fused representation:

$$f = [g_t \odot e_t \; ; \; g_a \odot e_a \; ; \; e_t \odot e_a] \tag{2}$$

where $\odot$ denotes element-wise multiplication. The resulting vector f is linearly projected to 256 dimensions and then passed through a residual block, wherein the MLP consists of two blocks of LayerNorm, GELU activation, and dropout, with an intermediate 256-to-256 linear transformation.

2) Attention Pooling: Since audio features are extracted frame-by-frame from WavLM, an attention pooling mechanism is employed to aggregate frame-level representations into a single utterance embedding:

$$\tilde{e}_a = \sum_i \text{softmax}(s_i) \cdot h_i \tag{3}$$

where h_i are the WavLM hidden states and s_i are the learned attention scores.

3) Session-Level Fusion via BiGRU: To capture temporal dependency across a clinical interview, a Bidirectional Gated Recurrent Unit (BiGRU) session encoder is applied on top of the utterance-level embeddings. For each utterance i, the text embedding, audio embedding, and their element-wise interaction are concatenated and projected:

$$x^{(i)} = \text{proj}([e_t^{(i)}; e_a^{(i)}; e_t^{(i)} \odot e_a^{(i)}]), \quad \text{proj} : \mathbb{R}^{384} \to \mathbb{R}^{256} \tag{4}$$

The projected sequence is fed into the BiGRU:

$$H = \text{BiGRU}(x^{(1)}, \ldots, x^{(n)}) \tag{5}$$

The BiGRU produces hidden states $H \in \mathbb{R}^{n \times 128}$ (64 units per direction). Attention pooling is applied to obtain a session-level vector, passed through a linear classification head.

4) Calibration and Threshold Tuning: Temperature scaling is applied as a post-hoc calibration step. A single learnable temperature parameter T is optimized on the development set by minimizing cross-entropy loss on the scaled logits $\hat{z}/T$. The decision threshold is tuned on the development set by maximising balanced recall, constrained to the range $[0.30, 0.70]$.

IV. RESULTS

The proposed system was evaluated on the E-DAIC WOZ test split (56 sessions, 4,614 utterances) across four depression-detection configurations and two sentiment-classification granularities.

A. Depression Detection

Table II outlines the test-set performance across all four configurations. The text-only BERT model achieved the highest overall accuracy (73.21%) and macro F1-score (0.7214). The audio-only WavLM model reached 67.86% accuracy but showed a pronounced recall imbalance (91.43% Non-Depressed vs. 28.57% Depressed). The Fusion MLP achieved the highest precision for the Depressed class (72.73%), reducing false-positive predictions relative to the text-only model (62.50%). The Session BiGRU obtained the highest Depressed-class recall (71.43%), at the cost of lower overall accuracy (58.93%). Fig. 2 depicts the precision–recall trade-off.

TABLE II. Depression Detection Performance on the E-DAIC WOZ Test Set

Model	Acc.	Prec.	Rec.	Macro F1
Text-only (BERT)	0.7321	0.7188	0.7286	0.7214
Audio-only (WavLM)	0.6786	0.6738	0.6000	0.5902
Fusion MLP (BERT+WavLM)	0.7143	0.7192	0.6476	0.6500
Session BiGRU	0.5893	0.6094	0.6143	0.5881

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

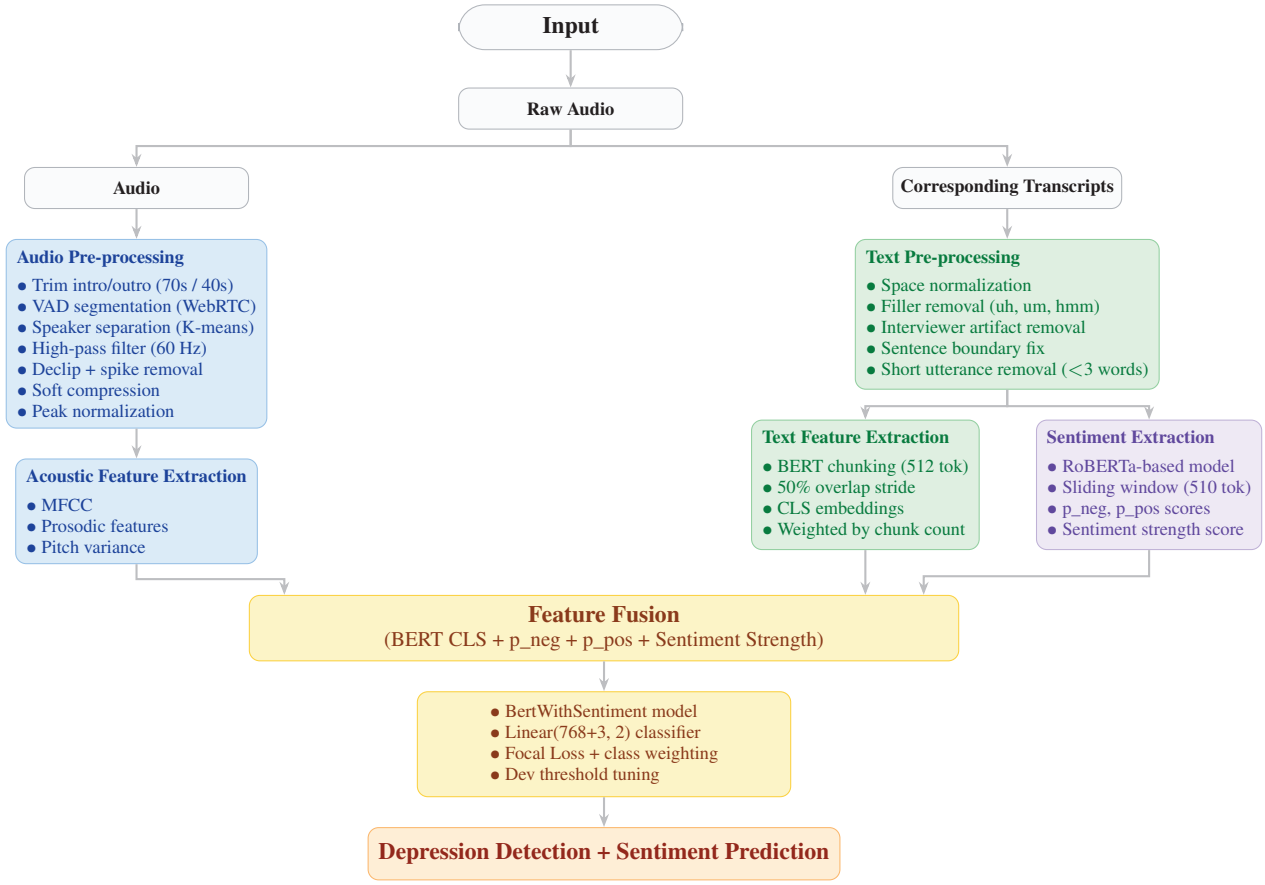

Fig. 1. Proposed multimodal pipeline for depression detection using the E-DAIC-WOZ dataset.

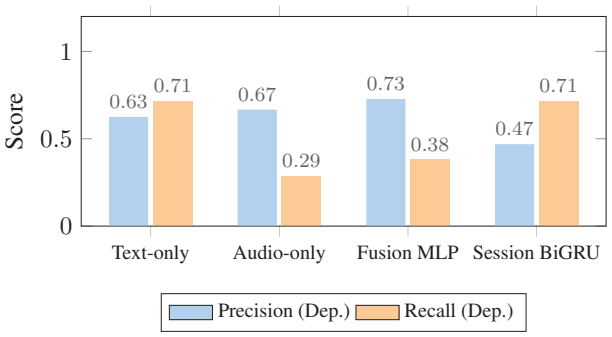

Fig. 2. Precision–recall trade-off for the Depressed class across the four model configurations.

B. Sentiment Classification

The auxiliary sentiment head, trained jointly with the text branch, achieved 87.34% accuracy (macro F1 = 0.8734) at the utterance level across 4,614 test utterances. Aggregating predictions to the session level using a tuned threshold of 0.5206 yielded 85.71% accuracy (macro F1 = 0.8564) across the 56 test sessions.

V. DISCUSSION

The results indicate that text and audio modalities contribute complementary information rather than redundant signal. The BERT text encoder captures linguistic depression markers effectively but on its own is prone to false positives, while the WavLM audio encoder, despite lower standalone accuracy, supplies acoustic cues that sharpen precision when combined through the gated fusion mechanism. This is reflected in the Fusion MLP's improved Depressed-class precision (72.73% vs. 62.50% for text-only), which is clinically relevant since it reduces unnecessary referrals. Compared to prior DAIC-WOZ fusion models such as the ACMA framework by Iyortsuun et al. [26] (75.8% accuracy), the proposed Fusion MLP achieves a comparable 71.43% accuracy while additionally performing joint sentiment classification, a capability absent from existing DAIC-WOZ depression-detection systems.

The gap between utterance-level and session-level performance suggests that temporal dynamics are harder to learn from the small labelled set (21 Depressed, 35 Non-Depressed sessions in the test split), though the BiGRU's higher Depression-class recall makes it a useful complementary high-sensitivity signal alongside the higher-precision Fusion MLP.

The consistently strong sentiment classification performance at both utterance (87.34%) and session (85.71%) levels confirms that the shared BERT encoder generalises well across the two related tasks, supporting the multi-task design choice.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

VI. CONCLUSION

This paper presented a multimodal approach for depression detection that combines textual and acoustic representations extracted from patient interviews in the E-DAIC WOZ dataset. Four configurations were evaluated: a text-only model, an audio-only model, an utterance-level gated fusion model, and a session-level BiGRU model. The text-only model achieved the highest overall accuracy (73.21%), while the gated fusion model achieved the highest precision for the Depressed class (72.73%). The session-level BiGRU achieved the highest Depressed-class recall (71.43%), making it a useful high-sensitivity signal for flagging at-risk patients. The auxiliary sentiment classification head achieved strong performance at both the utterance level (87.34% accuracy) and the session level (85.71% accuracy). These findings show that the proposed system achieves performance comparable to existing DAIC-WOZ-based approaches while uniquely combining depression detection with sentiment classification, offering added clinical interpretability beyond single-task depression screening systems. Future work will explore incorporating additional modalities such as facial expressions, expanding the model to detect related conditions such as anxiety and stress, and validating the system on larger and more diverse clinical datasets.

REFERENCES

[1] K. Srivastava, K. Chatterjee, and P. S. Bhat, "Mental health awareness: The indian scenario," *Industrial Psychiatry Journal*, vol. 25, no. 2, pp. 131–134, 2016, editorial.

[2] World Health Organization, "Mental disorders," Sep. 2025, accessed: 2026-03-26. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[3] ——, "Depressive disorder (depression)," Aug. 2025, accessed: 2026-03-26. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[4] Institute for Health Metrics and Evaluation (IHME), "Global burden of disease study 2021 results," 2024, accessed: 2026-03-26. [Online]. Available: https://vizhub.healthdata.org/gbd-results/

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[6] Y. Li, S. Kumbale, Y. Chen, T. Surana, E. S. Chng, and C. Guan, "Automated depression detection from text and audio: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[7] M. Kowalewski, M. Stroinski, K. Kwarciak, V. Laptiev, and D. Hemmerling, "End-to-end multimodal system for depression detection from online recordings," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023, pp. 1–4.

[8] S. Teng, S. Chai, J. Liu, T. Tateyama, L. Lin, and Y.-W. Chen, "Multi-modal and multi-task depression detection with sentiment assistance," in *Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE)*, 2024, pp. 1–5.

[9] C. Madhumitha, "Prediction of mental health (depression) using data science and machine learning techniques," *B. Eng. thesis, Dept. Comput. Sci. Eng.*, 2022.

[10] M. Maulyanda and S. A. Nazhifah, "Sentiment analysis of mental health using support vector machine (svm) with fastapi implementation," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, pp. 568–575, 2025.

[11] P. Nedungadi, G. Veena, K.-Y. Tang, R. R. Menon, and R. Raman, "Ai techniques and applications for online social networks and media: Insights from bertopic modeling," *Ieee Access*, 2025.

[12] F. Arias, M. Z. Nunez, A. Guerra-Adames, N. Tejedor-Flores, and M. Vargas-Lombardo, "Sentiment analysis of public social media as a tool for health-related topics," *Ieee Access*, vol. 10, pp. 74 850–74 872, 2022.

[13] G. Selva Mary, J. Blesswin, M. Venkatesan, S. Vairagar, S. Adagale, C. Shravage, and J. Barpute, "Original research article enhancing conversational sentimental analysis for psychological depression prediction with bi-lstm," *Journal of Autonomous Intelligence*, vol. 7, no. 1, 2023.

[14] E. Alvarez-Mellado, E. Holderness, N. Miller, F. Dhang, P. Cawkwell, K. Bolton, J. Pustejovsky, and M.-H. Hall, "Assessing the efficacy of clinical sentiment analysis and topic extraction in psychiatric readmission risk prediction," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 2019, pp. 81–86.

[15] P. K. Tiwari, M. Sharma, P. Garg, T. Jain, V. K. Verma, and A. Hussain, "A study on sentiment analysis of mental illness using machine learning techniques," in *IOP conference series: materials science and engineering*, vol. 1099, no. 1. IOP Publishing, 2021, p. 012043.

[16] R. V. Sharan, C. Mascolo, and B. W. Schuller, "Emotion recognition from speech signals by mel-spectrogram and a cnn-rnn," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

[17] K. Ikeuchi, T. Kishimoto, F. Nakai, T. Horigome, M. Kitazawa, and T. Ohtsuki, "Efficient and effective fine-tuning method for depression detection from conversation," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2025, pp. 1–5.

[18] J. He and H. Hu, "Mf-bert: Multimodal fusion in pre-trained bert for sentiment analysis," *IEEE Signal Processing Letters*, vol. 29, pp. 454–458, 2021.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

[19] T. Borah and S. Ganesh Kumar, "Application of nlp and machine learning for mental health improvement," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3*. Springer, 2022, pp. 219–228.

[20] S. Zhang, Z. Zheng, D. He, H. Zhu, Y. Gu, Y. Hu, X. Lv, T. Zhu, and W. Zhang, "Speech-based depression detection: Enhancing emotional support via clinical data," *IEEE Transactions on Affective Computing*, 2025.

[21] A. Hassan and S. Bernadin, "A comprehensive analysis of speech depression recognition systems," in *SoutheastCon 2024*. IEEE, 2024, pp. 1509–1518.

[22] G. D. Jadhav, S. D. Babar, and P. N. Mahalle, "Ensemble-based multimodal analysis for depression detection," *Engineered Science*, vol. 35, p. 1604, 2025.

[23] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[24] C. Wang, J. Konpang, A. Sirikham, and S. Tian, "Enhancing weibo sentiment analysis with multi-modal learning: Integrating text and synthesized images with contrastive learning," *IEEE Access*, 2025.

[25] L. Liu, F. Tydeman, W. Xie, and Y. Wang, "Development of an ai-based mobile app for automatic depression screening using speech in english and chinese," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2025, pp. 1–5.

[26] N. K. Iyortsuun, S.-H. Kim, H.-J. Yang, S.-W. Kim, and M. Jhon, "Additive cross-modal attention network (acma) for depression detection based on audio and textual features," *IEEE Access*, vol. 12, pp. 20 479–20 489, 2024.

[27] S. S. Dhingra, K. Kroenke, M. M. Zack, T. W. Strine, and L. S. Balluz, "Phq-8 days: a measurement option for dsm-5 major depressive disorder (mdd) severity," *Population health metrics*, vol. 9, no. 1, p. 11, 2011.

[28] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *Lrec*, vol. 14. Reykjavik, 2014, pp. 3123–3128.

[29] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[30] J. Wiseman and I. Y. Bondarenko, "Python interface to the webrtc voice activity detector," *URL: https://github.com/wiseman/py-webrtcvad (: 15.04. 2025)*, 2016.

[31] C. Jemine, "Resemblyzer: Voice encoder for speaker verification," https://github.com/resemble-ai/Resemblyzer, 2019, accessed: 2026.