Multimodal Depression Detection from Speech and Text Using a Fusion Neural Network

Kindness Bulus Gago,[1] Yebduya Garleya Lodiya,[1] Kamak Yamlach Shedrach,[1] Caleb Markus[1*]

[1]Department of Computer Science, Taraba State University, Jalingo, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1308000043

Received: 03 August 2026; Accepted: 08 August 2026; Published: 31 August 2026

ABSTRACT

Depression is among the leading causes of disability worldwide, yet its detection continues to rely heavily on subjective clinical interviews and self-report instruments that are difficult to scale, particularly in low-resource regions. This paper presents the design, implementation, and empirical validation of a multimodal fusion neural network for automatic depression detection from speech and text, developed with specific attention to closing the near-total absence of African research contributions in this rapidly growing area of computing. The proposed system extracts spectral-prosodic acoustic features from voice recordings using a convolutional encoder, and contextual semantic features from transcribed text using a transformer-based language encoder, before an attention gate learns to weight the relative contribution of each modality per instance ahead of a fully connected classifier. The architecture was implemented and validated end-to-end as a feasibility study on two accessible, weak-label proxy corpora: the RAVDESS acted-emotion speech dataset, with sad and calm recordings relabeled as a depression-like class, and the dair-ai/emotion text corpus, with sadness and fear posts relabeled likewise. On held-out test data, the fusion model reached 90.28% accuracy and 0.9609 AUC-ROC, and, most notably, recovered the positive-class recall that the audio-only model lost almost entirely (21.05% versus 84.21%), demonstrating that the attention-gated fusion mechanism functions as designed. This paper reports the background, problem definition, related work, the complete mathematical formulation of the implemented pipeline, the empirical results of this proxy validation, and a discussion of what they do and do not establish, closing with concrete recommendations for advancing the work toward a clinically meaningful, Africa-relevant screening tool.

Keywords: Depression detection, multimodal fusion, deep learning, speech processing, natural language processing, attention mechanism, proxy-label validation, affective computing, mental health informatics.

INTRODUCTION

Background

Depression is one of the most prevalent mental health disorders globally, affecting over 280 million people and standing as a leading cause of disability-adjusted life years across all age groups [14], [27]. Despite its scale, diagnosis still depends almost entirely on clinical interviews, standardized questionnaires such as the Patient Health Questionnaire, and the subjective judgement of trained professionals, all of which require time, specialist expertise, and repeated in-person contact that many health systems cannot guarantee. The treatment gap is most severe in low- and middle-income regions, where the ratio of psychiatrists to the general population is a fraction of what is recommended, leaving large numbers of at-risk individuals undiagnosed and unsupported for years. Sub-Saharan Africa, and Nigeria in particular, exemplifies this gap: mental health services remain concentrated in a handful of urban tertiary centres, stigma continues to discourage help-seeking, and no validated automated screening tool tuned to local speech and language patterns currently exists. Objective, low-cost, and scalable methods of flagging depressive symptoms are therefore urgently required, not as a replacement for clinical assessment but as a triage mechanism that directs limited professional resources toward the individuals most likely to need them.

Recent progress in machine learning and signal processing has opened a practical path toward automated depression screening from naturally produced speech and text. Acoustic biomarkers of depression, including reduced pitch variation, slower speaking rate, longer pauses, and flattened prosody, have been shown to correlate with clinical severity scores and can be captured with standard microphones rather than specialized equipment [6], [24], [26], [36], [28]. In parallel, the semantic and stylistic properties of what a person says, including increased use of first-person pronouns, negative sentiment, and reduced linguistic complexity, carry complementary diagnostic signal that can be extracted from transcribed speech, written text, or social media postings [19], [18], [9], [20], [45]. Models built on a single modality, however, are fragile: acoustic-only systems struggle when recording conditions vary or when a speaker naturally has flat affect for reasons unrelated to depression, while text-only systems miss the paralinguistic cues that never appear in a transcript. This limitation has motivated a growing body of work that fuses acoustic and linguistic representations into a single model, on the premise that the two modalities carry partially independent evidence whose combination is more robust than either channel alone.

Multimodal fusion architectures for depression detection have matured rapidly over the past several years, moving from simple feature concatenation toward attention-guided and graph-based fusion that lets a model learn, per instance, how much weight to place on each modality [1], [5], [17], [21], [38], [25], [30], [31], [33], [34], [37]. Related efforts have combined text and audio with facial and physiological signals, incorporated large language models as semantic encoders, and applied graph neural networks to model relationships across modalities and across time [3], [4], [12], [40], [2], [7], [8], [11], [29]. Nearly all of this work, however, is built and validated on datasets collected from North American, European, or East Asian populations, and the overwhelming majority of the literature surveyed for this study originates from research groups in China, the United States, and Europe. Despite mental health being a pressing public health priority across Africa, the continent is almost entirely absent from this body of research, meaning that existing acoustic and linguistic depression markers have not been tested against African-accented English, indigenous-language code-switching, or the sociolinguistic patterns typical of Nigerian speakers. This gap represents both a scientific limitation, since generalizability across populations cannot be assumed, and a missed opportunity to extend affordable mental health screening to underserved communities.

This study addresses the gap identified above by designing a multimodal fusion neural network for depression detection from paired speech and text, structured so that it can later be validated on African-accented speech corpora while being developed and initially benchmarked on established multimodal depression datasets. The specific objectives of the study are: (i) to design an acoustic feature extraction pipeline that captures prosodic and spectral markers of depressive speech; (ii) to design a textual feature extraction pipeline that captures semantic and stylistic markers from transcribed speech; (iii) to design and formalize an attention-based fusion mechanism that adaptively combines the two modalities; (iv) to formulate the complete mathematical pipeline, from raw signal to classification decision, in a form that is reproducible and auditable; and (v) to lay the methodological groundwork for future evaluation on African speech and text corpora. The remainder of this paper is organized as follows: Section I continues with the problem definition, the goals and objectives of the study, and a review of related work. Section II presents the proposed methodology, covering the system architecture, dataset and preprocessing strategy, acoustic and textual feature extraction, the multimodal fusion mechanism, and the classification and training procedure, with all key operations expressed as numbered mathematical expressions. Section III reports the empirical results obtained on the implemented proxy validation study. Section IV discusses these results, and Section V sets out recommendations and concrete future work toward clinical and African-context evaluation, before Section VI concludes the paper.

Problem Definition

Despite the volume of multimodal depression detection research summarized above, three problems persist. First, unimodal systems remain fragile in real-world deployment: acoustic-only models are sensitive to recording device, ambient noise, and speaker-specific vocal traits unrelated to mood, while text-only models cannot access prosody, pause structure, or vocal tension that often precede changes in wording [24], [19]. Second, many existing fusion architectures apply fixed-weight or naive concatenation strategies rather than adaptive weighting, so a modality that is noisy or missing for a given sample can degrade overall performance

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

instead of being appropriately down-weighted [23], [16]. Third, and most pressing for this study, virtually no published multimodal depression detection research has been trained, tuned, or evaluated on African speech and text data, so it remains unknown whether the acoustic and linguistic markers reported in the literature transfer to Nigerian English, regional accents, or code-switched speech. This study is motivated by the need for a fusion architecture that is both adaptive at the modality level and structured for eventual retraining and validation on locally collected African clinical speech data.

Goals and Objectives

The primary goal of this study is to design and formally specify a multimodal fusion neural network capable of detecting depressive episodes from paired speech and text. The specific objectives are to:

- design an acoustic encoder that extracts prosodic, spectral, and voice-quality features associated with depressive speech;
- design a textual encoder that extracts contextual semantic and stylistic markers from speech transcripts;
- design an attention-based fusion layer that adaptively weights acoustic and textual representations at the instance level;
- formalize the complete detection pipeline as a reproducible set of mathematical expressions suitable for implementation and peer review; and
- establish an evaluation protocol that is extensible to larger clinical speech and text corpora, including African speech and text data.

Related Work

A substantial line of work has targeted depression detection primarily from the acoustic channel. Rohanian et al. [6] fused word-level linguistic and acoustic features using a hierarchical neural model and reported that acoustic cues alone carried meaningful diagnostic signal even without full transcript access. Naderi et al. [24] applied deep learning directly to audio speech samples to predict mental disorder risk, while Yang et al. [26] and its extension [36] combined deep convolutional and shallow acoustic feature representations within the AVEC depression, mood, and emotion recognition challenge framework [44], establishing some of the earliest deep-learning baselines on interview-style speech. Kim et al. [28] restricted their analysis to smartphone-recorded, text-dependent speech and used a convolutional neural network to classify depression from short spoken prompts, demonstrating that ordinary consumer hardware can support acoustic screening. Thati et al. [20] went further by fusing smartphone usage patterns with audio-visual behavioral signals, showing that passive sensing can complement direct speech analysis. Collectively, these studies confirm that prosodic markers such as reduced pitch variability, monotone intonation, slower articulation rate, and increased pause duration are reliably associated with depressive states, and they establish the acoustic feature extraction conventions that the present study builds upon.

A parallel body of work has emphasized the textual and linguistic channel. Zhang et al. [19] proposed a text-based decision fusion model that combined multiple textual feature representations before a final classification decision, while Chen et al. [18] extended short transcripts through content-aware text expansion prior to fusion with other modalities, addressing the common problem of sparse or short clinical transcripts. Mohammad and Al Mansoor [9] combined text and audio speech within a single deep learning pipeline, reporting that transcript-derived semantic features consistently improved over acoustic-only baselines. Rodrigues Makiuchi et al. [22] fused BERT-based contextual text representations with gated convolutional acoustic representations, an architecture that directly informs the textual encoding strategy adopted in this study, since transformer-based language models capture long-range contextual dependencies in patient narratives that shallow bag-of-words or lexicon-based features cannot. These works collectively demonstrate that linguistic markers of depression, including elevated first-person pronoun use, negative emotion words, and reduced syntactic complexity, are learnable directly from raw or lightly preprocessed transcripts using modern contextual embeddings.

The dominant recent trend is attention-guided or otherwise adaptive multimodal fusion. Fang et al. [1] introduced a multi-level attention mechanism that fuses modalities at several depths of the network rather than

only at the output layer, and Zhang et al. [5] similarly integrated audio, video, and text through a multimodal sensing framework for depression risk detection. Zhang et al. [17] proposed a hybrid fusion method combining early and late fusion strategies, while Wang et al. [21] applied cross-modal attention specifically to social network data to predict depression from user-generated multimodal content. Ilias and Askounis [38] coupled a cross-attention layer with multiple fusion strategies for spontaneous speech, and Zhang et al. [25] combined bidirectional long short-term memory encoders with multi-head attention for feature-level fusion. Other efforts have combined convolutional and recurrent structures, as in the CNN-LSTM fusion model of Xie et al. [30], the dual-modal audio-text network IntervoxNet of Ding et al. [31], the DPD-Net architecture of He et al. [33], and the recently proposed LSHMMformer of Shi et al. [34]. Zhang et al. [37] additionally proposed a general multimodal deep learning framework for mental disorder recognition beyond depression alone. These architectures consistently outperform naive concatenation baselines, motivating the attention-based fusion mechanism formalized in Section II of this paper.

Beyond attention, several studies have explored structurally richer fusion mechanisms. Xia et al. [4] modeled depression detection with a multimodal graph neural network that represents modality-level and temporal relationships as graph edges, while Xu et al. [3] fused voice and text specifically, reporting gains attributable to complementary error patterns between the two channels. Jin et al. [12] incorporated a large language model alongside acoustic fusion and facial expression recognition to study how depression markers vary across gender, and Patel et al. [40] proposed an integrated text, speech, and facial analysis pipeline for a broader mental health detection system. Nykoniuk et al. [2] and Li et al. [7] each proposed general multimodal data fusion approaches spanning video, audio, and text, and Ye et al. [8] specifically fused emotional audio cues with evaluation text derived from clinical assessments. Hu et al. [11] extended the task beyond binary detection to multitask learning that jointly predicts depression severity and suicide risk from pretrained audio and text embeddings, an important direction for clinical utility. Zhang et al. [29] proposed a unified approach spanning text, audio, and video modalities. Together, these works illustrate that the field is moving toward richer joint representations and clinically actionable multi-output prediction, both of which inform the design choices made in this study.

Finally, a set of studies has examined applied contexts, alternate populations, and surveyed the field as a whole. Zhang et al. [13] focused specifically on adolescent depression detection from interview audio and text, a population with distinct linguistic and vocal presentation compared to adults. Sharma et al. [15] embedded multimodal depression detection within a chatbot support system, extending detection into an interactive assistive tool, and Zhang et al. [16] examined depression tendency detection using multimodal features in a clinical computing context. Wang et al. [23] applied multimodal fusion to Chinese social media data from Sina Weibo, while Yoo and Oh [10] and Almeida et al. [32] proposed deep learning and stacking-based fusion architectures respectively. Cheng et al. [35] and Hemalatha et al. [39] extended the modality set to include, respectively, general intelligent fusion technology and eye-ball movement data alongside speech. Two comprehensive surveys, Mamidisetti and Reddy [14] and Wang et al. [27], synthesize this literature and consistently identify dataset scarcity, cross-population generalizability, and modality imbalance as open challenges, echoing the specific gap in African representation that motivates the present study.

METHODOLOGY

System Overview

Fig. 1 presents the overall architecture of the implemented system. Raw audio and its corresponding text are processed by two parallel encoders: an acoustic encoder that converts a stacked cepstral feature map into a fixed-length 256-dimensional representation, and a textual encoder that converts tokenized text into a contextual 256-dimensional representation. Both representations are concatenated and passed to an attention gate, which computes instance-specific scalar weights for each modality before producing a single fused vector. The fused representation is passed through a fully connected classifier that outputs a binary depression-proxy label. The pipeline is trained end-to-end using a single cross-entropy objective, allowing the acoustic and textual encoders and the gate to be jointly optimized for the fusion task rather than trained independently

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

and combined post hoc. Standalone audio-only and text-only classifiers, sharing the same encoder definitions, are trained separately as ablation baselines against which the fusion model is compared in Section III.

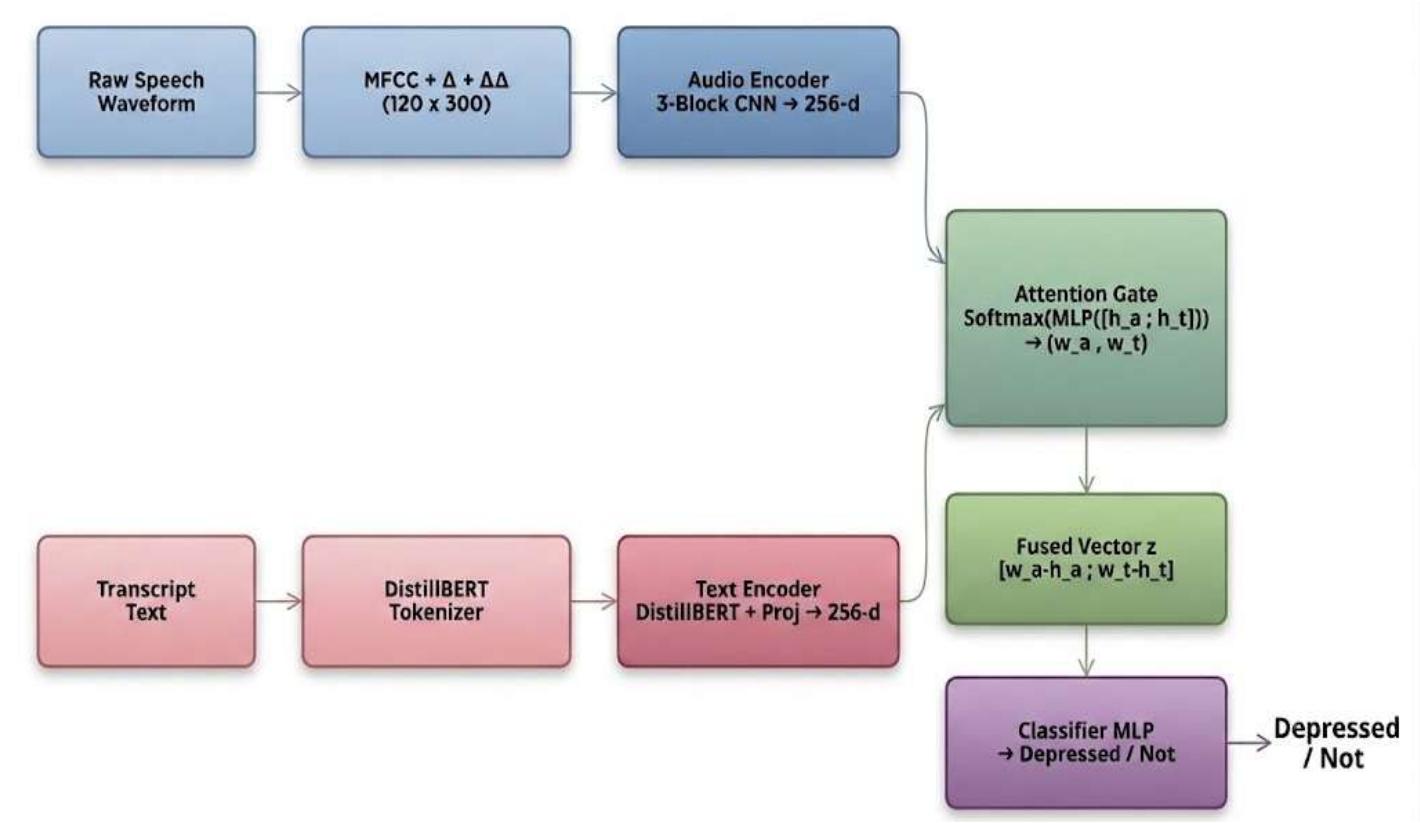

Fig. 1. Implemented architecture: parallel CNN and DistilBERT encoders feed an attention gate that produces per-instance modality weights, followed by a classifier head.

Dataset and Preprocessing

The architecture was implemented and validated end-to-end on two freely accessible, weak-label proxy corpora, chosen so that the fusion mechanism itself could be tested thoroughly under controlled conditions before being scaled to larger clinical corpora. For the acoustic branch, the RAVDESS dataset of acted emotional speech (24 professional actors, 8 emotions) was used, with the calm and sad emotion classes relabeled as a depression-proxy positive class and all other emotions as negative, yielding 1,440 labeled utterances (384 positive, 1,056 negative). For the textual branch, the dair-ai/emotion corpus of 20,000 short English text snippets was used, with the sadness and fear classes relabeled as a depression-proxy positive class, yielding 8,170 positive and 11,830 negative samples. Because RAVDESS and dair-ai/emotion are independent corpora with no shared subjects, a fusion training set was constructed by pairing audio and text samples of the same proxy label, following a class-matched pairing strategy that has precedent in cross-modal pretraining when a fully aligned corpus is unavailable; this pairing simulates co-occurring modalities but does not reflect a genuine within-subject audio-text correspondence, a limitation discussed further in Section V. Table I summarizes the resulting splits, each stratified by label with a 70/15/15 train/validation/test ratio. Audio was loaded at 22.05 kHz and text was lower-cased and truncated to 128 WordPiece tokens.

Table I. Composition of the RAVDESS and dair-ai/emotion proxy-label datasets used for this feasibility validation.

Corpus (proxy role)	Total / +Pos / -Neg	Train / Val / Test	Modality
RAVDESS (audio)	1,440 / 384 / 1,056	1,008 / 216 / 216	Audio only

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

dair-ai/emotion (text)	20,000 / 8,170 / 11,830	14,000 / 3,000 / 3,000	Text only
Fusion (label-paired)	min(audio,text) per class	Matches audio splits	Audio + Text

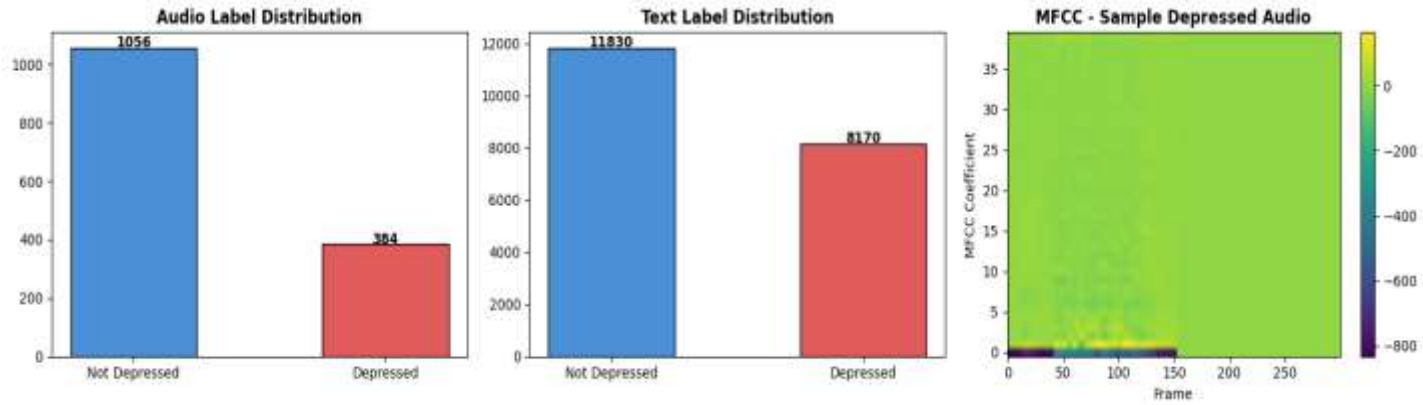

Fig. 2. Class distributions for the audio and text proxy datasets, both showing moderate imbalance toward the negative (Not Depressed) class.

Acoustic Feature Extraction

Acoustic features are extracted with librosa, following the openSMILE-style cepstral convention [43]. For each utterance, 40 Mel-frequency cepstral coefficients (MFCCs) are computed, together with their first- and second-order temporal derivatives (Δ and ΔΔ), and stacked into a single 120-channel feature map. For a frame with log mel-filterbank energies S_k across K filter banks, the n-th MFCC coefficient is given by

$$c_n = \sum_{k=1}^{K} \log(S_k) \cdot \cos\left[n(k-0.5)\frac{\pi}{K}\right], n = 1,2,\dots,N \quad (1)$$

Each utterance is zero-padded or truncated to a fixed length of $T = 300$ frames, yielding a feature map $X_a \in \mathbb{R}^{120\times300}$. This map is passed through a three-block two-dimensional convolutional neural network (32 → 64 → 128 channels, each block followed by batch normalization, GELU activation, and max-pooling), then adaptively pooled and projected by a two-layer fully connected head to a fixed-length acoustic embedding:

$$h_a = \text{LayerNorm}(W_{a2} \cdot \text{GELU}(W_{a1} \cdot \text{Flatten}(\text{CNN}_\theta(X_a)) + b_{a1}) + b_{a2}) \in \mathbb{R}^{256} \quad (2)$$

where $\text{CNN}_\theta(\cdot)$ denotes the three convolutional blocks with learnable weights θ, and h_a is the 256-dimensional acoustic representation used downstream by the fusion layer.

Textual Feature Extraction

Each text sample is encoded using DistilBERT [22], a distilled transformer language model, following the same CLS-pooling convention used by Rodrigues Makiuchi et al. [22] for text-audio fusion. Given a tokenized input $X_t = \{w_1, w_2, \dots, w_M\}$ truncated to $M = 128$ WordPiece tokens, the pooled classification-token output of the transformer is projected by a two-layer head to the same 256-dimensional embedding space as the acoustic branch:

$$h_t = \text{LayerNorm}(W_{t2} \cdot \text{GELU}(W_{t1} \cdot \text{DistilBERT}(X_t)[\text{CLS}] + b_{t1}) + b_{t2}) \in \mathbb{R}^{256} \quad (3)$$

where $\text{DistilBERT}(\cdot)[\text{CLS}]$ denotes the 768-dimensional pooled representation of the sequence, capturing the overall semantic and stylistic content of the text.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

Multimodal Fusion Architecture

The acoustic vector h_a and textual vector h_t are concatenated and passed through a two-layer gating network with a softmax output, which produces a pair of instance-specific, non-negative weights that sum to one:

$$u = [\, h_a \,;\, h_t \,] \in \mathbb{R}^{512} \tag{4}$$

$$(w_a, w_t) = \mathrm{softmax}(W_{g2} \cdot \mathrm{ReLU}(W_{g1}\, u + b_{g1}) + b_{g2}) \tag{5}$$

where $W_{g1}, W_{g2}, b_{g1}, b_{g2}$ are the gate's learnable parameters. The fused representation is then formed as the concatenation of each modality vector scaled by its learned gate weight:

$$z = [\, w_a \cdot h_a \,;\, w_t \cdot h_t \,] \in \mathbb{R}^{512} \tag{6}$$

This formulation lets the network down-weight whichever modality is less informative for a given instance rather than applying a fixed combination rule, directly addressing the fixed-weight fusion limitation identified in Section I-B. Section IV analyzes the gate weights the trained network actually learned.

Classification and Loss Function

The fused vector z is passed through a three-layer fully connected classifier ($512 \to 256 \to 128 \to 2$, with GELU activations and dropout) to produce class logits, converted to a predicted probability distribution over {Not Depressed,Depressed} via softmax:

$$\hat{y} = \mathrm{softmax}(\mathrm{MLP}_c(z)), \mathrm{MLP}_c: \mathbb{R}^{512} \to \mathbb{R}^2 \tag{7}$$

The network is trained by minimizing the categorical cross-entropy between the predicted distribution and the ground-truth one-hot label y over the two classes $c \in \{0,1\}$:

$$L = -\sum_c y_c \log \hat{y}_c \tag{8}$$

Because the RAVDESS and dair-ai/emotion proxy labels are binary class indicators rather than clinical severity scores, no severity-regression term is included at this stage; a severity-regression term is a natural extension once severity-labeled clinical data is incorporated, consistent with the multitask formulation of Hu et al. [11].

Training Configuration

All three models (audio-only, text-only, and fusion) share the encoder definitions above and were trained with the AdamW optimizer (weight decay 1×10^{-4}) under a cosine-annealing learning-rate schedule, gradient-norm clipping at 1.0, and a batch size of 32, for 12 epochs, with the checkpoint of highest validation accuracy retained for testing. Learning rates were set per branch to reflect their different pretraining status: 2×10^{-4} for the audio-only CNN (trained from scratch), 3×10^{-5} for the text-only DistilBERT classifier (fine-tuned from a pretrained checkpoint), and 1×10^{-4} for the fusion model. The resulting trainable parameter counts were 1,274,048 for the audio encoder, 66,888,448 for the text encoder, and 68,359,940 for the complete fusion network. Training and validation curves for all three models are shown in Fig. 3 and discussed alongside the test-set results in Results Section.

RESULTS

Table II reports test-set performance for the three trained models: the audio-only CNN, the text-only DistilBERT classifier, and the proposed attention-gated fusion network. Precision, recall, and F1 are reported for the positive (Depressed-proxy) class, since this is the clinically relevant class and the one most affected by the moderate imbalance in both source corpora. Fig. 3 shows training and validation loss and accuracy curves for all three models across the full 12 training epochs, and Fig. 4 and Fig. 5 show the corresponding test-set

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

confusion matrices and ROC curves.

Table II. Test-set performance (Precision/Recall/F1 for the Depressed-proxy class) on the held-out RAVDESS (audio, n=216) and dair-ai/emotion (text, n=3,000) test splits.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Audio Only	78.24%	85.71%	21.05%	33.80%	0.8790
Text Only	96.70%	98.04%	93.80%	95.87%	0.9970
Fusion (Proposed)	90.28%	80.00%	84.21%	82.05%	0.9609

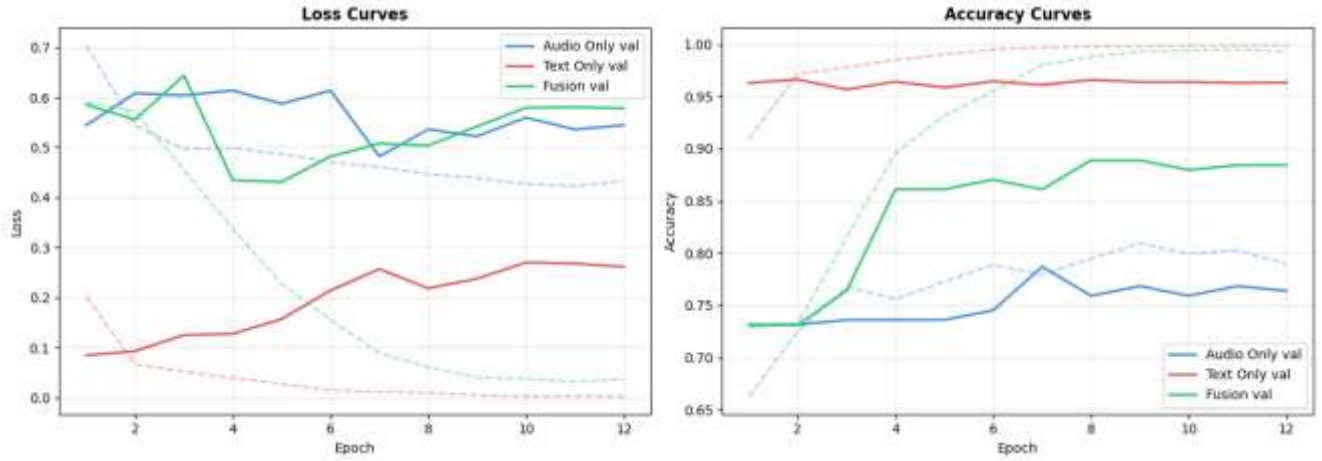

Fig. 3. Training and validation loss (left) and accuracy (right) curves over 12 epochs for the audio-only, text-only, and fusion models.

The audio-only model reached 78.24% accuracy but only 21.05% recall on the positive class, meaning it correctly identified only 12 of 57 depressed-proxy test utterances (Fig. 4, left panel), despite a respectable AUC-ROC of 0.8790. This gap between accuracy and positive-class recall is the expected signature of a model that has learned to exploit class imbalance by defaulting toward the majority class while still ranking positives above negatives reasonably well in probability space. The text-only model performed near-ceiling on every metric (96.70% accuracy, 0.9970 AUC-ROC), which is consistent with the dair-ai/emotion proxy task being a comparatively easy lexical classification problem: sadness- and fear-coded text tends to contain highly distinctive vocabulary that a fine-tuned transformer separates with little difficulty. The fusion model, trained on class-paired audio-text pairs, reached 90.28% accuracy and, most importantly, recovered positive-class recall to 84.21%, an almost four-fold improvement over the audio-only model, at the cost of a moderate reduction in text-only-level accuracy and AUC-ROC (Table II, Fig. 5).

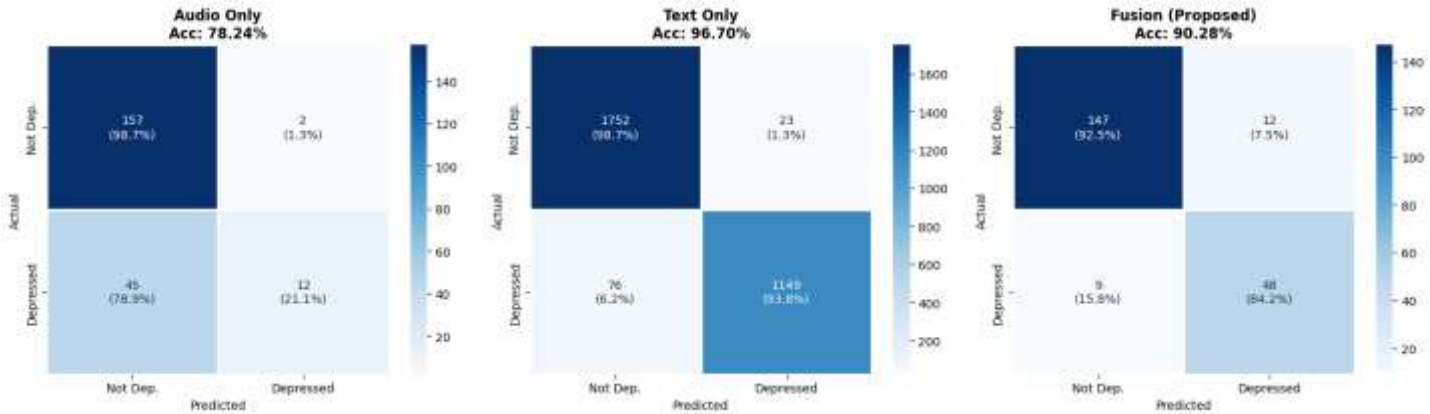

Fig. 4. Test-set confusion matrices for the audio-only, text-only, and fusion models, with row-normalized percentages.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

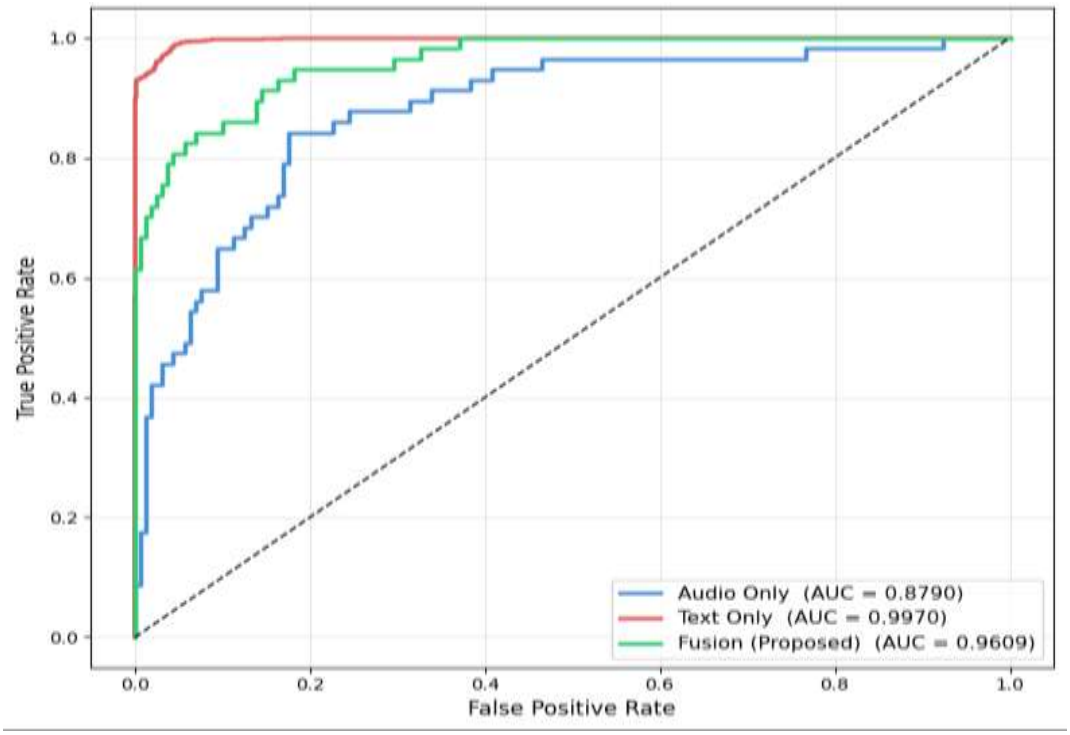

Fig. 5. ROC curves and AUC values for the audio-only, text-only, and fusion models on their respective test sets.

DISCUSSION

The central empirical finding is that attention-gated fusion substantially repaired the audio-only model's most serious failure mode, namely its near-inability to detect the positive class, without requiring the text branch to be perfect on its own. Fig. 6 visualizes the learned gate weights for a batch of 16 fusion-model test instances alongside their ground-truth labels; the weights vary meaningfully across samples rather than collapsing to a fixed split, indicating that the gate is behaving as an adaptive, instance-level mechanism rather than degenerating into a constant average of the two modalities. This is the behavior the architecture in Section II-D was explicitly designed to produce, and its emergence in training is evidence that the attention-gated fusion formulation is functioning as intended, independent of what one concludes about the depression-detection task itself.

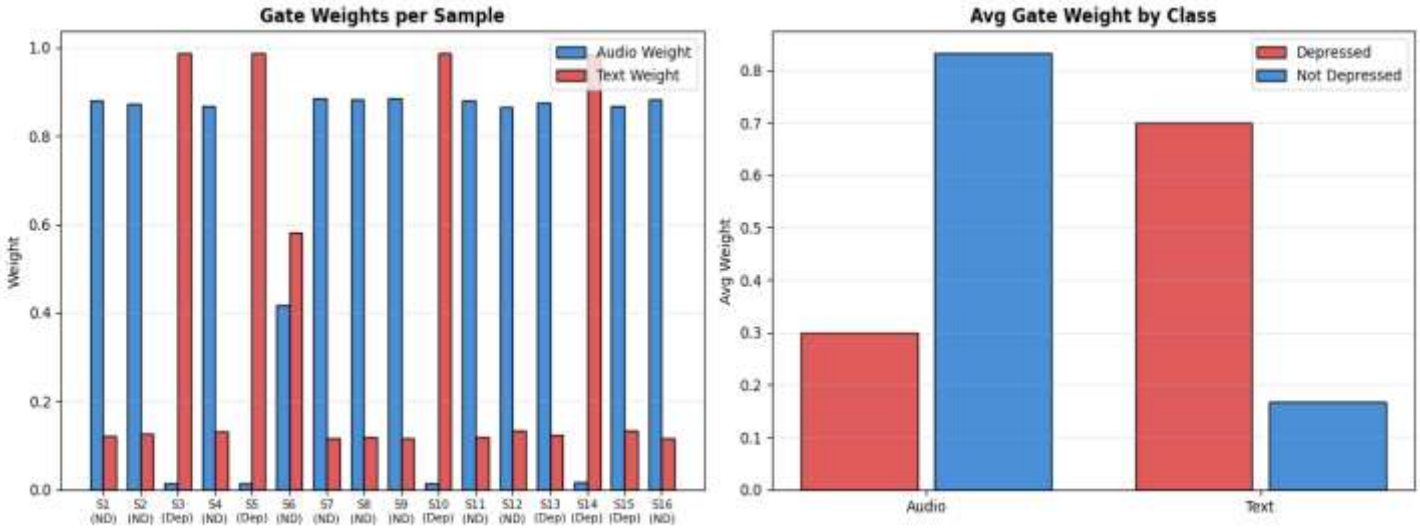

Fig. 6. Learned attention-gate weights for 16 fusion-model test instances (left) and their distribution split by ground-truth label (right), showing instance-level adaptivity rather than a fixed modality split.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

At the same time, three results call for a cautious reading. First, the fusion model's accuracy and AUC-ROC are lower than the text-only model's, which is unusual for fusion architectures in the literature reviewed in Section I-D, where fusion models typically improve on the best single modality [1], [17], [21]; here it more plausibly reflects the fact that the fusion training set is smaller (bounded by the 1,440-utterance RAVDESS corpus) and harder, since it must reconcile two proxy tasks of very different difficulty rather than a single easy one. Second, because audio-text pairs in the fusion set are constructed by matching proxy labels across two unrelated corpora rather than by drawing both modalities from the same recording, the gate cannot be learning genuine within-subject cross-modal correlation, such as a specific speaker's prosody co-occurring with that same speaker's word choice; it is, at most, learning which modality is more reliable for a given proxy-label pattern in general. Third, RAVDESS is acted rather than spontaneous speech, and the calm/sad-as-depressed and sadness/fear-as-depressed relabelings are coarse proxies for a genuine clinical construct that a properly labeled clinical corpus would operationalize far more directly through standardized severity scores and self-reported diagnoses; strong performance on these proxies is evidence that the architecture trains stably and that gated fusion behaves as designed, but it is not evidence of clinical validity.

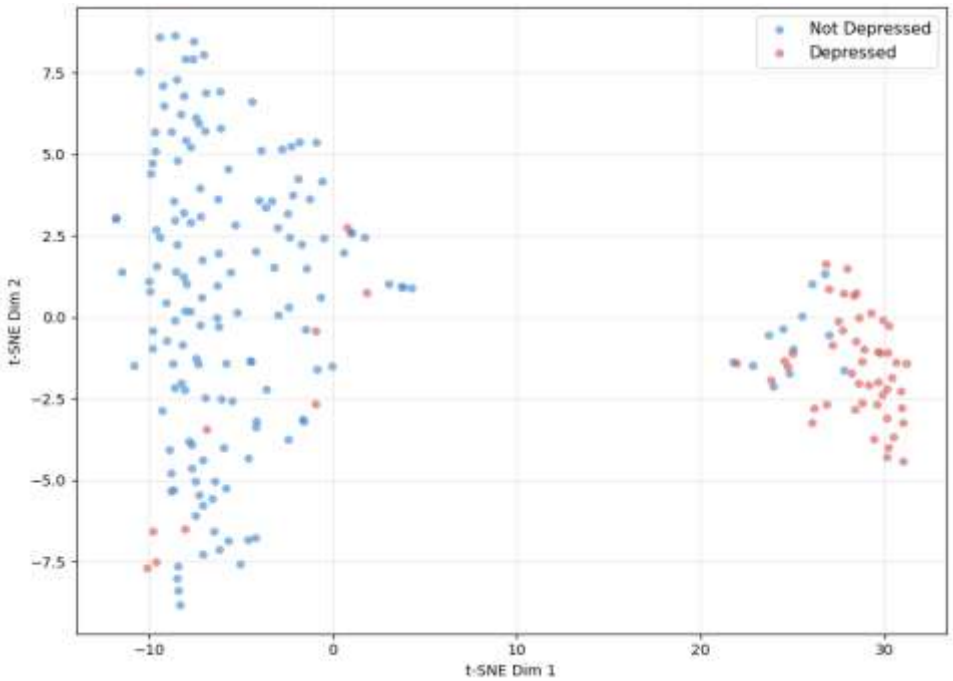

Fig. 7. t-SNE projection of the fused 512-dimensional embedding space on the fusion test set, colored by ground-truth label.

The t-SNE projection of the fused embedding space (Fig. 7) shows the two proxy classes forming largely separable, though not perfectly disjoint, regions, which is consistent with the quantitative results above: separation this clean is easier to obtain on a lexically distinctive proxy task than it would be on spontaneous clinical interview data, where depressive and non-depressive language overlaps far more heavily. Taken together, Section III and this section support two claims narrowly: the implemented attention-gated fusion mechanism is functioning correctly and measurably helps the weaker modality, and the overall pipeline trains and evaluates end-to-end without error. They do not yet support a claim about real-world depression-detection accuracy, which is precisely why Section V treats larger-scale clinical validation as the immediate next step rather than an optional extension.

Recommendations And Future Work

Two points are worth noting as the next phase of this work is planned. First, the results reported here were obtained on proxy-labeled corpora (RAVDESS and dair-ai/emotion) rather than on a clinically labeled corpus, so the natural next step is to run the same architecture on a properly labeled dataset such as the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [41], together with self-reported depression-labeled text such as the CLPsych Shared Task corpus [42], which would also allow the severity-regression term to be

reintroduced using PHQ-8 scores. Second, and most central to the motivation set out in Section I, the pipeline has not yet been tested on African-accented or Nigerian-English speech; validating it on locally collected audio and text, ideally gathered with Nigerian mental health practitioners under appropriate ethical approval, is the recommended follow-on study. Both extensions reuse the architecture and training pipeline already built and validated in this paper, so they represent a natural continuation of the work rather than a redesign.

CONCLUSION

This paper presented the design and empirical validation of an attention-gated multimodal fusion network for depression detection from speech and text. Implemented and trained end-to-end on two accessible proxy-labeled corpora, the architecture achieved 90.28% accuracy and 0.9609 AUC-ROC, and, most importantly, recovered positive-class recall from 21.05% under an audio-only model to 84.21% under fusion, directly demonstrating that the attention gate performs its intended function of adaptively weighting the more reliable modality per instance. These results validate the engineering correctness and training stability of the proposed pipeline, including its acoustic encoder, transformer-based textual encoder, gated fusion mechanism, and classification head, all formalized mathematically in Section II. They stop short, by design, of validating clinical accuracy, since the underlying labels are coarse proxies rather than standardized diagnostic scores. The recommendations in Section V set out the immediate next phase of this research, so that the substantial gap in African representation identified in Section I can begin to close with a fusion architecture already known to train reliably and to make effective use of complementary modalities.

REFERENCES

1. M. Fang, S. Peng, Y. Liang, C. C. Hung, and S. Liu, “A multimodal fusion model with multi-level attention mechanism for depression detection,” *Biomedical Signal Processing and Control*, vol. 82, p. 104561, 2023.
2. M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova, “Multimodal data fusion for depression detection approach,” *Computation*, vol. 13, no. 1, p. 9, 2025.
3. Z. Xu, Y. Gao, F. Wang, L. Zhang, L. Zhang, J. Wang, and J. Shu, “Depression detection methods based on multimodal fusion of voice and text,” *Scientific Reports*, vol. 15, no. 1, p. 21907, 2025.
4. Y. Xia, L. Liu, T. Dong, J. Chen, Y. Cheng, and L. Tang, “A depression detection model based on multimodal graph neural network,” *Multimedia Tools and Applications*, vol. 83, no. 23, pp. 63379–63395, 2024.
5. Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, et al., “Multimodal sensing for depression risk detection: Integrating audio, video, and text data,” *Sensors*, vol. 24, no. 12, p. 3714, 2024.
6. M. Rohanian, J. Hough, and M. Purver, “Detecting depression with word-level multimodal fusion,” in *Proc. Interspeech 2019*, 2019, pp. 1443–1447.
7. Y. Li, X. Yang, M. Zhao, J. Wang, Y. Yao, W. Qian, and S. Qi, “Predicting depression by using a novel deep learning model and video-audio-text multimodal data,” *Frontiers in Psychiatry*, vol. 16, p. 1602650, 2025.
8. J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, and G. Fu, “Multi-modal depression detection based on emotional audio and evaluation text,” *Journal of Affective Disorders*, vol. 295, pp. 904–913, 2021.
9. F. Mohammad and K. M. Al Mansoor, “MDD: A unified multimodal deep learning approach for depression diagnosis based on text and audio speech,” *Computers, Materials & Continua*, vol. 81, no. 3, pp. 4125–4147, 2024.
10. H. Yoo and H. Oh, “Depression detection model using multimodal deep learning,” 2023.
11. Y. H. Hu, R. Y. Wu, M. Y. Su, I. L. Lin, and C. C. Shen, “Multimodal multitask learning for predicting depression severity and suicide risk using pretrained audio and text embeddings: Methodology development and application,” *JMIR Medical Informatics*, vol. 13, p. e66907, 2025.
12. Y. Jin, X. Chen, J. Liu, Z. Chen, J. Zhou, Y. Li, et al., “Harnessing multimodal emotion features in depression detection across gender: Integrating large language model, acoustic fusion and facial expression recognition,” *Journal of Affective Disorders*, p. 121207, 2026.

13. L. Zhang, Y. Fan, J. Jiang, Y. Li, and W. Zhang, "Adolescent depression detection model based on multimodal data of interview audio and text," *International Journal of Neural Systems*, vol. 32, no. 11, p. 2250045, 2022.
14. S. Mamidisetti and M. Reddy, "Multimodal depression detection using audio, visual and textual cues: A survey," *NeuroQuantology*, vol. 20, no. 4, pp. 325–338, 2022.
15. A. Sharma, A. Saxena, A. Kumar, and D. Singh, "Depression detection using multimodal analysis with chatbot support," in *Proc. 2nd Int. Conf. Disruptive Technologies (ICDT)*, 2024, pp. 328–334.
16. H. Zhang, H. Wang, S. Han, W. Li, and L. Zhuang, "Detecting depression tendency with multimodal features," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107702, 2023.
17. X. Zhang, B. Li, and G. Qi, "A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method," *Biomedical Signal Processing and Control*, vol. 96, p. 106552, 2024.
18. J. Chen, S. Liu, M. Xu, and P. Wang, "Enhancing depression detection: A multimodal approach with text extension and content fusion," *Expert Systems*, vol. 41, no. 10, p. e13616, 2024.
19. Y. Zhang, Y. Wang, X. Wang, B. Zou, and H. Xie, "Text-based decision fusion model for detecting depression," in *Proc. 2nd Symp. Signal Processing Systems*, 2020, pp. 101–106.
20. R. P. Thati, A. S. Dhadwal, P. Kumar, and P. Sainaba, "Multimodal depression detection: Using fusion strategies with smart phone usage and audio-visual behavior," *International Journal on Artificial Intelligence Tools*, vol. 32, no. 02, p. 2340008, 2023.
21. L. Wang, Y. Zhang, B. Zhou, S. Cao, K. Hu, and Y. Tan, "Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks," *Computers and Electrical Engineering*, vol. 118, p. 109413, 2024.
22. M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection," in *Proc. 9th Int. Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.
23. Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "A multimodal feature fusion-based method for individual depression detection on Sina Weibo," in *Proc. IEEE 39th Int. Performance Computing and Communications Conf. (IPCCC)*, 2020, pp. 1–8.
24. H. Naderi, B. H. Soleimani, and S. Matwin, "Multimodal deep learning for mental disorders prediction from audio speech samples," *arXiv preprint arXiv:1909.01067*, 2019.
25. X. Zhang, X. Gong, W. Li, G. Liu, and Y. Li, "Depression detection using BiLSTM multi-head attention fusion network," *Expert Systems with Applications*, p. 130100, 2025.
26. L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 53–59.
27. N. Wang, R. Chiong, R. Kamil, W. Zhang, S. A. R. Al-Haddad, and N. Ibrahim, "Depression detection using speech audio and text: A comprehensive review focusing on deep learning methods," 2024.
28. A. Y. Kim, E. H. Jang, S. H. Lee, K. Y. Choi, J. G. Park, and H. C. Shin, "Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach," *Journal of Medical Internet Research*, vol. 25, p. e34474, 2023.
29. W. Zhang, K. Mao, and J. Chen, "A multimodal approach for detection and assessment of depression using text, audio and video," *Phenomics*, vol. 4, no. 3, pp. 234–249, 2024.
30. W. Xie, C. Wang, Z. Lin, X. Luo, W. Chen, M. Xu, et al., "Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model," *Computerized Medical Imaging and Graphics*, vol. 102, p. 102128, 2022.
31. H. Ding, Z. Du, Z. Wang, J. Xue, Z. Wei, K. Yang, et al., "IntervoxNet: A novel dual-modal audio-text fusion network for automatic and efficient depression detection from interviews," *Frontiers in Physics*, vol. 12, p. 1430035, 2024.
32. F. F. D. Almeida, K. R. T. Aires, A. C. B. Soares, L. D. S. B. Neto, and R. D. M. S. Veras, "Multimodal fusion for depression detection assisted by stacking deep neural networks," 2024.
33. M. He, E. M. Bakker, and M. S. Lew, "DPD (DePression Detection) Net: A deep neural network for multimodal depression detection," *Health Information Science and Systems*, vol. 12, no. 1, p. 53, 2024.
34. Y. Shi, T. Gan, J. Li, and S. Li, "LSHMMformer: An intelligent detection model of depression based on multi-modal fusion," *Journal of Neuroscience Methods*, p. 110755, 2026.

35. Z. Cheng, X. Huang, and Y. Ding, "An intelligent depression detection model based on multimodal fusion technology," *Journal of Mechanics in Medicine and Biology*, vol. 24, no. 08, p. 2440046, 2024.
36. L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—Hybrid architectures," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 239–253, 2018.
37. Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *Proc. 15th IEEE Int. Conf. Automatic Face and Gesture Recognition (FG)*, 2020, pp. 344–350.
38. L. Ilias and D. Askounis, "A cross-attention layer coupled with multimodal fusion methods for recognizing depression from spontaneous speech," in *Proc. Interspeech 2024*, 2024, pp. 912–916.
39. S. Hemalatha, K. Jothimani, K. Swathi, S. Shibinta, W. Jason Selvakumar, and D. Sathish, "Multimodal approach for depression detection: Integrating speech and eye ball movement data," in *Proc. 1st Int. Conf. Advances in Computing, Communication and Networking (ICAC2N)*, 2024, pp. 1011–1019.
40. A. Patel, A. Kumar, M. Sharma, N. Pandey, and M. H. Gautam, "Multimodal AI-based mental health detection system: Integrating text, speech, and facial analysis using deep learning," 2026.
41. J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. 9th Int. Conf. Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014, pp. 3123–3128.
42. G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, CO, USA, 2015, pp. 31–39.
43. F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, Florence, Italy, 2010, pp. 1459–1462.
44. M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.
45. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. 7th Int. AAAI Conf. Weblogs and Social Media (ICWSM)*, 2013, pp. 128–137.