

Multimodal Depression Detection from Audio, Visual, and Textual Cues: Evaluation of Adaptive Gating versus Naive Concatenation

Azhar A. Hadi

azhar@uobabylon.edu.iq

Jawaharlal Nehru Technological University, Hyderabad

Supreethi K.P.

Jawaharlal Nehru Technological University, Hyderabad

Article

Keywords: Multimodal models, WavLM model, RoBERTa, Trustworthy, Deep learning

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11111831/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multimodal Depression Detection from Audio, Visual, and Textual Cues: Evaluation of Adaptive Gating versus Naive Concatenation

Azhar A. Hadi*, Supreethi K.P.

University College of Engineering, Science & Technology (UCEST)

Jawaharlal Nehru Technological University Hyderabad, India

Email: azhar@uobabylon.edu.iq, supreethi.pujari@jntuh.ac.in

*Corresponding author: azhar@uobabylon.edu.iq

Abstract—Depression detection on DAIC-WOZ often relies on interviewer-driven shortcuts rather than genuine participant behaviour, and fusion-architecture gains are often validated only via cross-validation, never confirmed on unseen participants. This study addresses both issues directly. First, we propose a multimodal deep learning classifier for depression that processes audio input using the WavLM model, facial behaviour using descriptors derived from OpenFace, and speech transcripts using RoBERTa, ensuring all inputs rely only on what participants say to remove interviewer-induced bias. This paper compared three fusion approaches: naive feature concatenation, a learned attention-based gating mechanism, and cross-modal attention, first using stratified 5-fold cross-validation and then on a strictly held-out test set of participants never seen during either model development or selection. During cross-validation, the gating mechanism performs better than simple concatenation, but this advantage disappears on the held-out data, where naive concatenation instead shows the best and most stable performance (74.5% accuracy, F1-macro 0.681 ± 0.046 across 5 random seeds). A diagnostic analysis shows why: the gate settles on nearly constant weights across modalities with no meaningful connection to input quality, meaning the apparent ability to adapt during cross-validation was due to overfitting rather than genuine learning. A complementary interpretability analysis that combines occlusion and corrected gradient-based attribution confirms that acoustic features have the strongest influence on the model’s predictions, followed by facial behaviour and then text. Although the model’s held-out sensitivity (0.50) shows that it is not yet suitable for clinical use, our results prove that rigorous generalization testing, rather than architectural complexity, is the key factor in developing trustworthy multimodal models based on small clinical datasets.

Keywords: Multimodal models, WavLM model, RoBERTa, Trustworthy, Deep learning.

I. INTRODUCTION

Around the world, depression is estimated to affect 280 million people. It remains a major cause of disability, and more than 75% of those affected in low and middle income countries do not receive any treatment, mainly because of the subjective nature of diagnosis and the lack of adequate clinical resources [21]. Consequently, researchers have shown strong interest in using automated methods to screen for depression by analysing multiple behavioural indicators gathered during clinical interviews, with the DAIC-WOZ corpus [10], [7] serv-

ing as the standard reference in this area. In response, recent studies have developed increasingly advanced fusion models, including gated and cross-modal attention mechanisms [25], [14], contrastive feature decomposition [5], and hierarchical transformer fusion [11], often achieving significant improvements over unimodal baselines. However, two particular risks associated with this dataset have so far been ignored: the structured format of the DAIC-WOZ interview allows models to use the position of the interviewer’s prompts as a kind of shortcut instead of interpreting them as a true behavioral signal, a point clearly shown by Burdisso et al. [4], and the recent research on the Extended DAIC-WOZ corpus, which achieved near-ceiling performance (F1 is 0.992), has itself pointed out that this result is indicative of leakage rather than reflecting clinical validity [22].

Moreover, because this body of literature seldom directly examines the more general risk, the limited sample size of 189 to 275 participants means that architecturally complex, instance-adaptive fusion mechanisms might show strong cross-validation performance without generalizing to a truly unseen, held-out set of participants, and the assertion that these mechanisms learn the true per-instance modality reliability is rarely checked against ground-truth signal-quality indicators. The present paper addresses both shortcomings by means of a participant-only, bias-controlled pipeline based on modern pretrained encoders (WavLM [6], OpenFace [2], RoBERTa [18]), and it systematically compares naive concatenation, learned gating, and cross-modal attention fusion using both stratified cross-validation and a strictly held-out test set.

This paper makes the following contributions:

- A controlled multimodal benchmark on DAIC-WOZ comparing unimodal, naive-concatenation, gated, and cross-attention fusion strategies using modern pretrained encoders under a common, participant-level evaluation protocol.
- A rigorous held-out evaluation demonstrating that cross-validation superiority of adaptive fusion does not reliably translate into improved generalization on unseen participants, with results reported across multiple random seeds and a parameter-matched control ruling out model capacity as a confound.

- A gate-collapse analysis showing that the learned gating mechanism converges to near-static modality weighting with negligible association with measured video-signal quality, quantitatively contrasted against the naive fusion model’s emergent quality sensitivity.
- A quantified bias-control ablation testing dependence on the known interviewer-prompt shortcut, reporting effect size and confidence intervals rather than a qualitative judgment.
- A complementary interpretability analysis combining modality-level occlusion and gradient-based (Integrated Gradients) attribution, including explicit identification and correction of a known early-timestep artifact in recurrent-network attribution, validated via a non-circular held-out threshold check.

II. RELATED WORK

DAIC-WOZ and its extended version E-DAIC [10], [24] are still the leading benchmarks for the automatic detection of depression from clinical interviews, using a semi-autonomous virtual interviewer to combine audio, video, and the transcribed text with clinician-generated PHQ-8 labels. Even though it has been widely adopted, recent surveys have pointed out that the corpus’s small size (ranging from 186 to 275 participants) and the strong class imbalance remain ongoing limitations which restrict the reliability of the results reported in the literature [28].

Although the potential of multimodal fusion is clear, most studies report only cross-validation or single-split results, so it remains unclear whether the improvements generalize to completely new participants. This paper directly addresses that question through a held-out, multi-seed evaluation. The earlier literature began with fusing speech and text representations using gated CNN-LSTM networks [25], then moved to more expressive fusion methods: cross-modal attention networks combining audio and text [14], contrastive frameworks for modality decomposition [5], a dual-level attention fusion method that achieved an F1 of 0.83 based on five-fold cross-validation [20], a hierarchical transformer-based fusion approach [11], and text-audio fusion using a canonical-correlation method that reported accuracies exceeding 92% [19].

The key takeaway is that the question-based, structured format used in DAIC-WOZ allows shortcut learning. Burdisso et al. [4] showed that models that leverage the position of the interviewer’s prompts rather than participants’ actual behaviour achieved inflated performance (F1 of 0.90 when using only text). This issue has also come up in more recent studies, as a hierarchical DistilBERT-BiLSTM fusion study on Extended DAIC-WOZ found near-ceiling F1 scores (0.992) in the text-enhanced setups. They clearly attributed this to shortcut learning or leakage rather than clinical validity [22]. Likewise, speech-based approaches have been suggested that focus on low-complexity models and feature selection guided by the need for explainability to reduce overfitting and spurious, artefact-driven correlations on this dataset [8].

An increasing amount of research makes use of post-hoc explainability in depression models, including SHAP and LLM-

based feature extraction from transcripts [9], SHAP/LIME-based interpretable procedures for behavioural and activity data [1], attention-guided sentence-level explanations together with sentence-transformer classifiers [17], causal graph-based interpretability modules [29], and density-adaptive attention mechanisms, which are said to achieve an F1-macro of 0.70 to 0.72 on DAIC-WOZ speech alone while also improving interpretability [12]. However, the key gap is that attribution techniques for recurrent architectures are seldom validated against known artefacts, such as recency bias in gradient-based attribution; to address this gap, we explicitly identify and correct this artefact using a non-circular, externally derived threshold, rather than relying on the same attribution methods for comparison.

III. MATERIALS AND METHODS

A. Dataset

The DAIC-WOZ corpus [10] comprises 188 separate interviews with individual participants, divided into 107 for training, 34 for development, and 47 for the held-out test, with no participant appearing in more than one split (verified by a program). It consists of semi-structured clinical interviews conducted by an animated virtual interviewer called "Ellie" and includes audio recordings, video recordings obtained through automatic extraction of facial features, and transcripts. Each interview is assigned a binary depression status based on the participant’s PHQ-8 score; therefore, a score of 10 or above is considered depressed, in accordance with the official AVEC 2017 train/dev/test division [27]. We excluded one of the 189 sessions originally distributed (participant 440) because we confirmed byte-level file corruption: the ZIP header was invalid. Consequently, class distribution is imbalanced in all the splits (approximately 70% are non-depressed), as summarized in Table I.

TABLE I: Dataset split composition.

Split	N	Non-Depressed	Depressed
Train	107	77	30
Development	34	23	11
Test (held-out)	47	33	14
Total	188	133	55

B. Bias-Controlled Preprocessing

To avoid the interviewer-prompt shortcut documented in this corpus [4], we obtain every modality entirely from segments of participant speech, identified by speaker-diarization labels and their associated timestamps in the transcript. As a result, before feature extraction, we discard all interviewer (“Ellie’s”) statements. We also exclude segments shorter than 0.3 s ($n = 607$, mostly transcription artefacts). This leaves 31,495 segments at the individual-utterance level across the corpus, with all three modalities temporally aligned.

C. Feature Extraction

For each participant’s speech segment, this paper used frozen, pretrained encoders to obtain modality-specific embeddings, following standard transfer-learning procedures. Specifically, we sampled the audio at 16 kHz and mean-pooled it

using WavLM. In addition, we used the layer-9 hidden states, which appear to capture paralinguistic information, to produce 1024-dimensional vectors. The video descriptors (facial Action Units, head pose, gaze) are obtained through OpenFace 2.0 and are mean-pooled together with the standard deviation for each segment to form 76-dimensional vectors; any frames with a tracking confidence below 0.7 are excluded, and segments that have no valid frames are clearly marked as missing (valid_frac) rather than being silently omitted, so that a data-quality signal can be used in the subsequent analyses. The text is encoded with RoBERTa-base and is mean-pooled across the non-padding tokens to yield 768-dimensional vectors. Throughout the main fusion comparisons, all encoders remain frozen.

D. Fusion Architectures

This paper looked at four fusion strategies, as shown in Figure 1. The unimodal baselines involve using an independent BiLSTM classifier for each modality’s embedding sequence. In the case of naive (early) fusion, the three embeddings at each time step (1868-d) are concatenated and then encoded using a single BiLSTM classifier. Gated fusion encodes each modality separately using three BiLSTM branches, each followed by a lightweight reliability head that produces a per-modality logit. the logit from the video modality is also penalized by (1 - valid_frac), and the three logits are SoftMax normalized into instance-specific weights which are then applied to a weighted sum of the embeddings. Cross-modal attention fusion treats the three projected embeddings as a token sequence of length 3, passes it through multi-head self-attention, and then classifies via mean pooling. With respect to the gated and attention architectures, we also assess a variant in which the BiLSTM encoders are pretrained once on the entire training pool and are then either kept frozen or fine-tuned at a greatly reduced learning rate (1×10^{-5}) as compared to the fusion head (1×10^{-3}) to reduce overfitting due to the small size of the corpus.

Above all, the number of trainable parameters in Naive and full Gated Fusion differs by only 7.4% (see Table II), which shows that the two main architectures are roughly equal in capacity. It is therefore reasonable to attribute the generalization differences to the fusion mechanism itself rather than model size.

TABLE II: Model parameter counts (trainable).

Model	Trainable Parameters
Text-only (BiLSTM, hidden=64)	431,000
Video-only (BiLSTM, hidden=64)	120,000
Naive Fusion (BiLSTM, hidden=64)	994,369
Gated Fusion (full, jointly trained)	1,068,197
Gated Fusion (lightweight, frozen encoders)	38,165

E. Training, Evaluation, Bias Control, and Interpretability

This study trained our classifiers using Adam [16] and a class-weighted cross-entropy loss function, with gradient clipping and early stopping based on the validation macro-F1 score. Hyperparameters were chosen entirely through stratified

5-fold cross-validation on the combined training and development dataset (n = 141 participants). This paper then retrained the final model once and assessed it on the strictly separated 47-participant test set using five different random splits. Results are reported as mean $\pm$ standard deviation. To examine dependence on the content of the interviewer’s prompts [4], we ran a paired t-test with bootstrap confidence intervals comparing the text model that included Ellie with the baseline using only participant data. This paper assessed interpretability using modality occlusion and Integrated Gradients [26]. To avoid attribution bias in a non-circular way, we defined the threshold for excluding artefacts at early time steps using data from held-out sessions independent of the case study being evaluated.

F. Evaluation Metrics

Because of class imbalance, we report accuracy, sensitivity, specificity, balanced accuracy, precision, F1 (per-class and macro), and ROC-AUC, computed using Equations (1 to 5). F1-macro and balanced accuracy are selected as the primary metrics, consistent with current recommendations for imbalanced clinical classification.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (3)$$

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (4)$$

$$F1_{\text{macro}} = \frac{1}{2} (F1_{\text{depressed}} + F1_{\text{non-depressed}}) \quad (5)$$

IV. RESULTS

A. Unimodal and Cross-Validation Performance

Table III gives the 5-fold cross-validation results (mean F1-macro $\pm$ SD) for the unimodal baselines and all four fusion methods on the combined training/development dataset (n = 141). Among the unimodal models, text achieved the highest F1-macro score (0.718 ± 0.079), followed by audio (0.707 ± 0.047) and finally video (0.638 ± 0.044), consistent with previous findings that facial or visual descriptors alone are generally the weakest single modality for this task. The naive (early) fusion method (0.693 ± 0.103) did not outperform the best unimodal baseline in cross-validation, while gated fusion achieved the highest F1-macro score in cross-validation among all the models examined (0.725 ± 0.071); cross-modal attention fusion performed worse than both (0.658 ± 0.026 , for the best-tuned version).

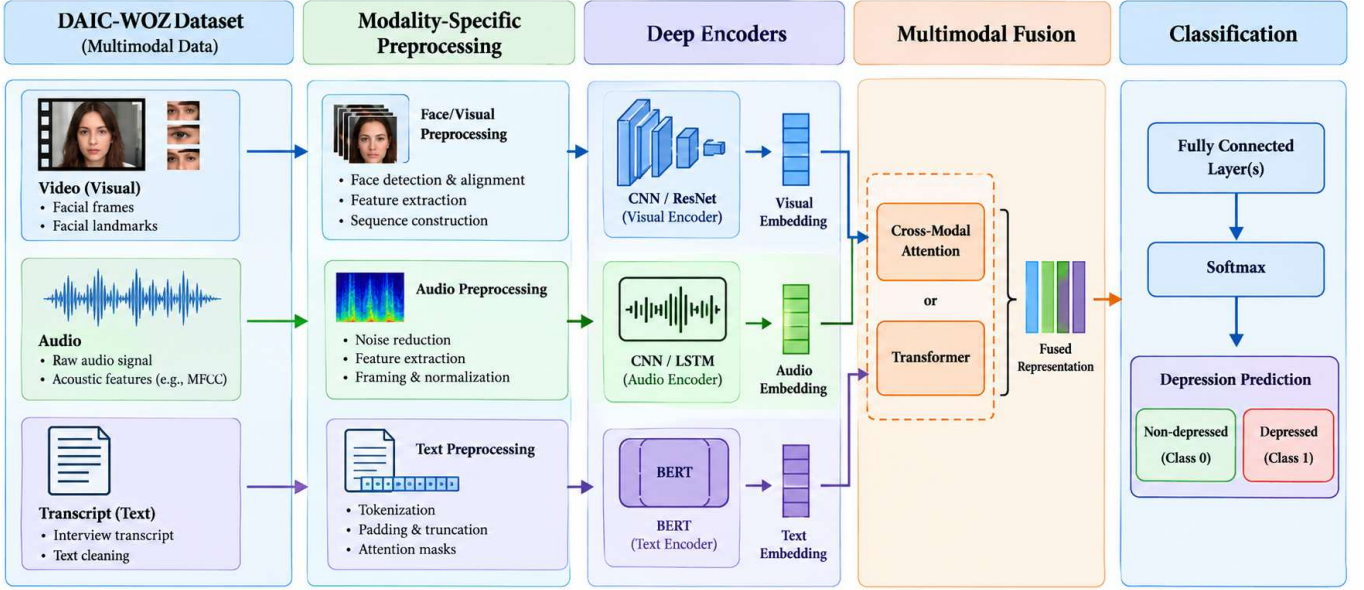


Fig. 1: Full pipeline of the proposed multimodal architecture.

TABLE III: Five-fold cross-validation performance (train+dev, $n = 141$).

Model	F1-macro	Recall	Precision
Text	0.718 ± 0.079	0.611 ± 0.140	0.605 ± 0.111
Audio	0.707 ± 0.047	0.492 ± 0.165	0.779 ± 0.187
Video	0.638 ± 0.044	0.492 ± 0.093	0.498 ± 0.067
Naive Fusion	0.693 ± 0.103	0.489 ± 0.238	0.733 ± 0.165
Late Fusion (avg.)	0.655 ± 0.065	0.417 ± 0.132	0.591 ± 0.111
Gated Fusion	0.725 ± 0.071	0.614 ± 0.126	0.609 ± 0.089
Cross-Attention Fusion	0.658 ± 0.026		

B. Held-Out Test Performance: A Generalization Reversal

The results in Table IV and Figure 2 reflect performance on the strictly held-out, previously unexamined 47-participant test set, with figures averaged over five random seeds. The ranking that was obtained during cross-validation did not persist: naive fusion achieved the highest F1-macro on the held-out data (0.689 ± 0.046 , ranging from 0.644 to 0.776 for each seed), outperforming all of the unimodal baselines as well as both adaptive fusion methods, whereas gated fusion, although it had been the leader in the cross-validation, dropped to 0.613 ± 0.013 on the held-out data. This is a clear reversal, since the fusion strategy that performed best in cross-validation is not the one which generalizes most effectively to new participants.

TABLE IV: Held-out test performance (five-seed mean $\pm$ SD; $n = 47$).

Model	F1-macro (Test)	F1-macro (CV)	CV $\rightarrow$ Test Gap
Text	0.599 ± 0.015	0.718 ± 0.079	-0.119
Audio	0.645 ± 0.031	0.707 ± 0.047	-0.062
Video	0.608 ± 0.047	0.638 ± 0.044	-0.030
Naive Fusion	0.689 ± 0.046	0.693 ± 0.103	~ 0.000
Gated Fusion	0.613 ± 0.013	0.725 ± 0.071	-0.112

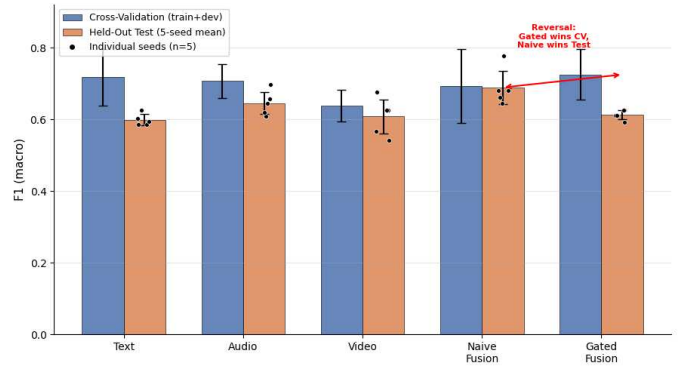


Fig. 2: CV vs. Held-Out Test Performance.

Figure 2 illustrates the near-zero generalization gap associated with naive fusion as opposed to the considerable gap seen in gated fusion, also showing the individual results for each seed in order to provide transparency. The comparison in Table II, which is parameter-matched, reveals that Naive Fusion (994,369 parameters) and full Gated Fusion (1,068,197 parameters) differ by just 7.4%, so the number of parameters alone is not a likely explanation for this generalization gap.

C. Gate-Collapse Analysis

To examine the mechanism behind the generalization gap associated with gated fusion, we looked at the weight behaviour for individual instances on the held-out test set (Figure 3A). The gate weights settled down to nearly constant values in all 47 of the test sessions (audio: 0.984 ± 0.002 , video: 0.006 ± 0.001 , text: 0.010 ± 0.002), indicating that the gate was performing nearly constant modality weighting rather than adapting in a way that was meaningful at the instance level. Consistently, when we correlated the gate’s video weight with an independent, objectively known measure of video quality (the mean facial-tracking valid-frame fraction), we found a negligible relationship ($r = 0.012$). In contrast, a similar post-hoc analysis of the naive fusion model’s video importance based on occlusion revealed a weak positive association with the same quality indicator $r = 0.175$, (see Figure 3B). Taken together, these results support the idea that the gate collapsed and overfit to patterns specific to each fold during cross-validation, rather than developing a generalizable, quality-sensitive fusion rule.

D. Bias-Control Ablation

Table V shows the results of the ablation tests that include Ellie with respect to the content of the interviewer prompts. Including the interviewer’s statements led to a mean increase in the F1-macro score of $+0.022$ (95% CI: $[-0.051, +0.095]$, bootstrap CI: $[-0.037, +0.091]$, paired t-test, $p = 0.587$; Cohen’s $d = 0.26$), a value that is not statistically different from no effect when $n = 5$ folds. Therefore, while this does not exclude the possibility of interviewer-prompt leakage, it does not provide evidence that our participant-only model’s performance depends in any meaningful way on the interviewer’s content, which aligns with our bias-control design.

E. Interpretability

On the held-out test set, modality occlusion (as shown in Figure 4) revealed that audio made the greatest contribution to the model’s predictions (mean $|\Delta p| = 0.155 \pm 0.086$, followed by video (0.060 ± 0.052) and then text (0.008 ± 0.009). All pairwise differences were statistically significant (Wilcoxon signed-rank test, $p < 0.0001$ for all three comparisons, $n = 47$).

The Integrated Gradients analysis showed a consistent, content-independent attribution pattern during the first 15 to 20 timesteps in all sessions studied. The magnitude then decreased by about two orders of magnitude before stabilizing (as shown in Figure 5), consistent with known recency-bias effects in gradient-based attribution methods applied to recurrent architectures. When this initial period is excluded, one case study involving a participant who was correctly classified as not depressed showed relatively strong late-sequence attribution, reaching a magnitude of 0.854 at timestep 86. This attribution corresponded to the utterance "more tired and", a phrase relevant to symptoms. It was more than two orders of magnitude above the threshold independently derived from the

other two sessions examined, indicating that it represents a real late-sequence attribution by the model rather than the early-timestep artefact identified earlier, and that it is a clinically plausible phrase relating to symptoms rather than evidence of confirmed clinical reasoning.

F. Full Classification Performance and Clinical Considerations

Table VI and Figure 6 show that the final naive-fusion model had competitive overall discrimination on the held-out test set. However, sensitivity remained limited (0.50), so its performance should be seen as a methodological benchmark rather than evidence that it is ready for clinical screening. Accuracy (0.745) exceeded balanced accuracy (0.674) by 0.070, indicating class imbalance (70% non-depressed, $n = 47$). Performance was asymmetric because specificity (0.849) was much higher than sensitivity (0.500).

G. Comparison with Prior Work

Table VII provides a methodological comparison of four earlier DAIC-WOZ studies that reflect the field’s standard evaluation methods. Bhatt et al. [3] achieve 78.69% accuracy and an F1-macro of 0.7636 using a voting-ensemble audio classifier, along with speaker diarization and SMOTE balancing. However, they do not state whether the train/test split is participant-disjoint. Ismail et al. [13] assess an audio-only transformer under participant-independent 5-fold cross-validation, citing an F1-macro of 0.599, but they neither report multi-seed stability nor test for known dataset-specific biases. The lightweight video-only CNN-LSTM model [15] obtains an F1-macro of 0.74 on DAIC-WOZ. The audio spectrogram transformer with imbalance-aware negative mining [23] reports an F1-macro of 0.7353. Still, neither study verifies a participant-disjoint split, reports results across multiple seeds, or provides a parameter-matched comparison against an alternative fusion strategy. In all four studies, no one meets all four requirements simultaneously. The current work fulfils all four criteria at the same time, and its results (F1-macro = 0.681 ± 0.046 , accuracy = 74.5%) lie within the range covered by these four baseline studies, showing that our trimodal fusion result is consistent with, not anomalously below, the level of performance that can be achieved on this corpus when evaluation rigour is kept constant.

V. CONCLUSION

The study compared three multimodal fusion methods: naive concatenation, learned reliability gating, and cross-modal attention for detecting depression using the DAIC-WOZ corpus, with audio embeddings from WavLM, facial descriptors from OpenFace, and text embeddings from RoBERTa. In stratified 5-fold cross-validation, the gated fusion model achieved the highest F1-macro score (0.725), ahead of naive concatenation (0.693) and all unimodal baselines. However, on a strictly held-out, participant-disjoint test set, the ranking changed: naive concatenation achieved the best and most stable performance (74.5% accuracy, F1-macro = 0.681 ± 0.046 across

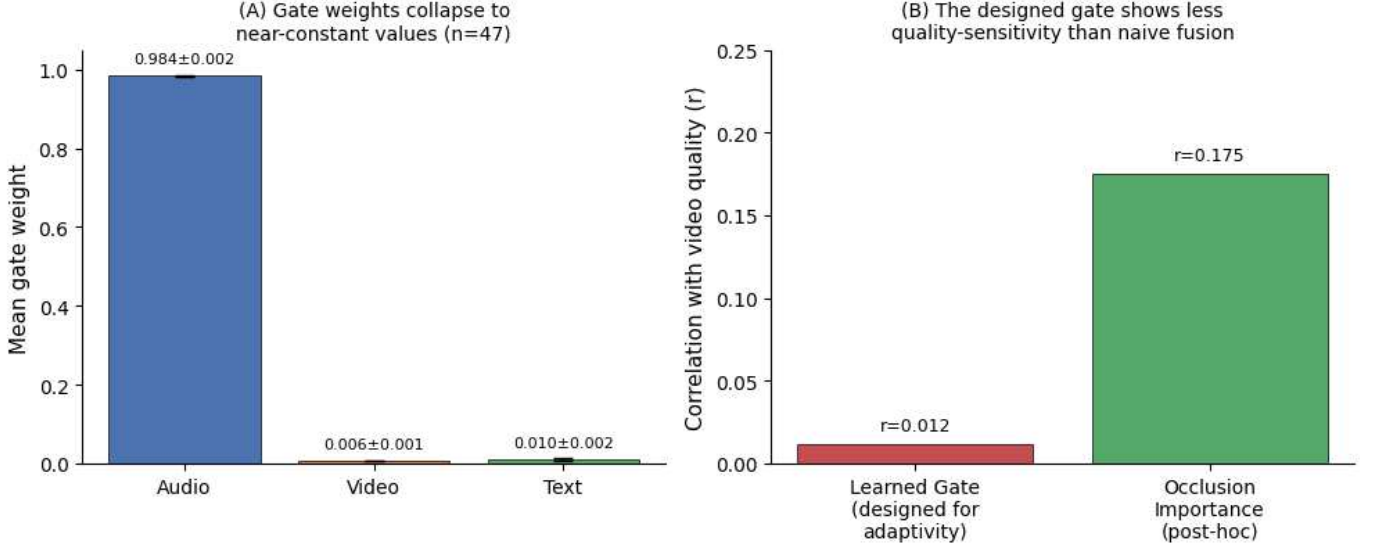


Fig. 3: Gate-weight collapse (A) and its weak association with modality quality relative to naive fusion (B).

TABLE V: Ellie inclusive ablation (5 fold CV).

Condition	F1-macro	Δ F1	95% CI	p-value
Participant-only (ours)	0.718 $\pm$ 0.079	-	-	-
Ellie inclusive	0.740 $\pm$ 0.063	+0.022	[-0.051, +0.095]	0.587

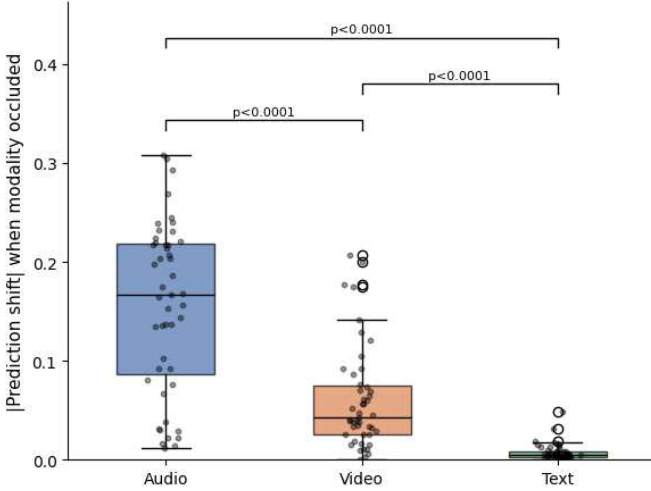


Fig. 4: Modality importance via input occlusion (n = 47 held-out sessions).

TABLE VI: Full classification metrics, held-out test (naive fusion, seed 42).

Metric	Value
Accuracy	0.745
Balanced Accuracy	0.674
Sensitivity (Depressed)	0.500
Specificity (Non-Depressed)	0.849
Precision (Depressed)	0.583
F1-macro	0.681
F1 (Depressed)	0.539
F1 (Non-Depressed)	0.824
ROC-AUC	0.671

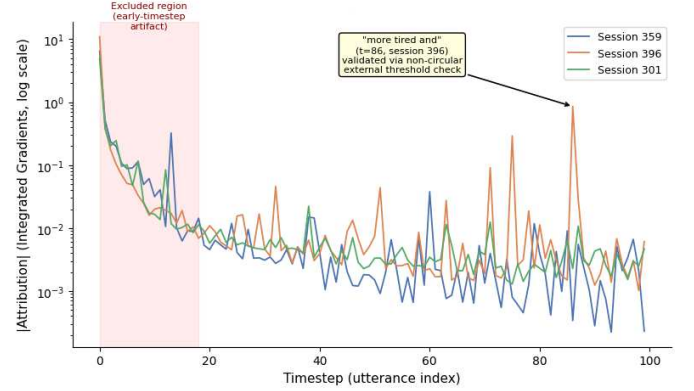


Fig. 5: Integrated Gradients attribution decay and a validated late-sequence case study.

5 seeds), whereas the gated fusion model fell to an F1-macro of 0.613 ± 0.013 , and cross-modal attention performed worse than both. The mechanistic analysis then pointed to gate collapse as the reason for the gating model's failure, describing a nearly constant weighting of the modalities (audio is 0.98) that was not related to the independently measured video quality ($r = 0.012$), showing that the advantage seen in cross-validation was due to overfitting and not to true instance-level adaptation. The interpretability analysis, which combined occlusion with artefact-corrected Integrated Gradients, provided statistically significant evidence (Wilcoxon, $p < 0.0001$) that audio contributed most to the model's predictions, followed by facial behaviour and then text, and a validated case study linked the model's attention to clinically reasonable

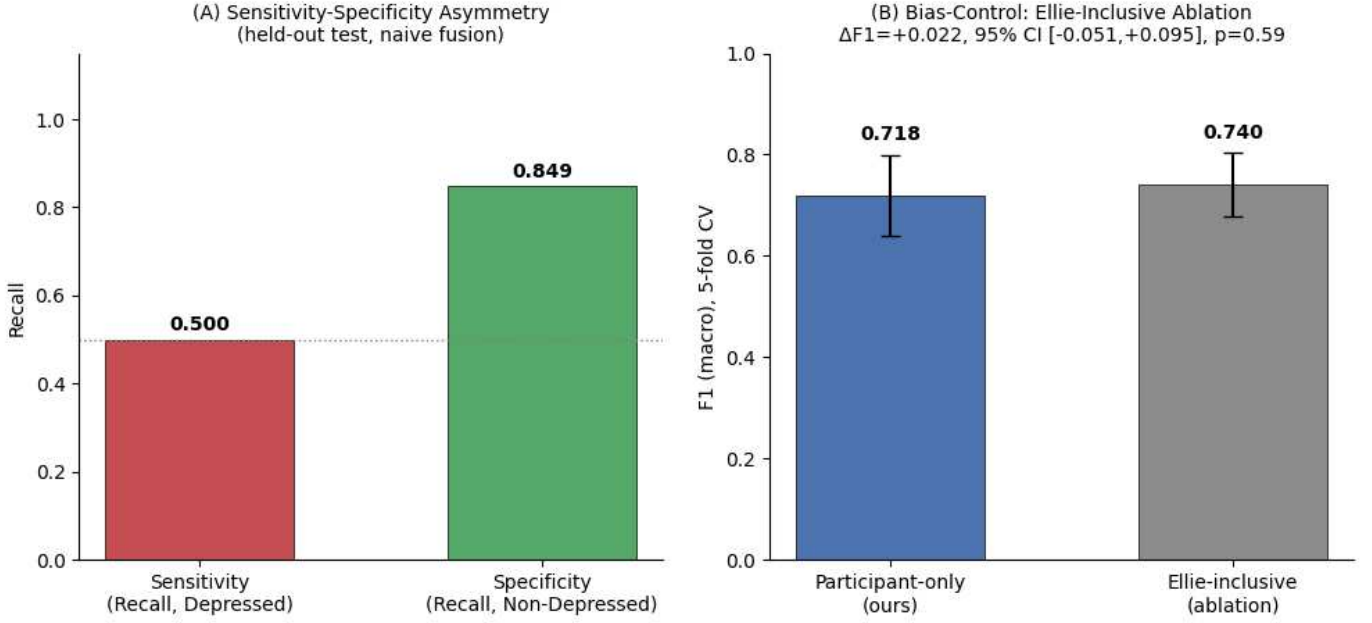


Fig. 6: Sensitivity–specificity asymmetry (A) and Ellie-inclusive bias-control ablation (B) on the held-out test set.

TABLE VII: Methodological comparison with prior work on DAIC-WOZ.

Study	Modalities	Held-Out Test	Participant-Disjoint	Multi-Seed	Bias/Shortcut Control	Param.-Matched	F1-macro
Bhatt et al. [3]	Audio	Yes	No	No	No	No	0.7636
Ismail et al. [13]	Audio	Yes	Yes	No	No	No	0.599
Jallaglag et al. [15]	Video	Yes	No	No	No	No	0.74
Ramrakhiyani et al. [23]	Audio	Yes	No	No	No	No	0.7353
This work	Audio+Video+Text	Yes	Yes	Yes	Yes	Yes	0.681 $\pm$ 0.046

symptom language. An ablation study removing interviewer inclusion confirmed that the results were not due to the known interviewer-prompt shortcut. Because the model had only limited sensitivity on the held-out data (0.50), these findings represent a solid methodological benchmark rather than indicating that the model is ready for clinical use. Future work should extend this evaluation to larger corpora such as Extended-DAIC, include parameter-matched architectural controls, and investigate regularisation strategies that maintain the flexibility of adaptive fusion while improving generalisation to new participants.

Author contributions

Conceptualization: A.H. and S.K.P.; Methodology: A.H.; Investigation: A.H.; Discussion of results: A.H.; Writing – Original Draft: A.H.; Writing – Review and Editing: A.H. and S.K.P.; Resources: A.H. and S.K.P.; Supervision: S.K.P.; Approval of the final text: S.K.P.

Acknowledgements

The authors gratefully acknowledge Jawaharlal Nehru Technological University Hyderabad, India, for providing the computing resources and technical support essential to this research. Artificial intelligence tools were utilised exclusively for language editing and grammatical correction, and were not used in the scientific content, experimental design, data analysis,

interpretation of results, or conclusions. All such contributions were developed and verified solely by the authors.

Funding

Not applicable.

Data availability

The dataset analyzed in this study is publicly available from the DAIC-WOZ corpus, distributed by the University of Southern California Institute for Creative Technologies. Access requires signing a data usage agreement, available at: <https://dcapswoz.ict.usc.edu/>

Declarations

Ethics approval

Not applicable. This article did not involve human participants, animal subjects, or the collection of primary data requiring ethical approval.

Consent to Participate

Not applicable. Human participants were not included in this article.

REFERENCES

- [1] Iftikhar Ahmed, Anushree Brahmacharimayum, Raja Hashim Ali, Talha Ali Khan, and Muhammad Ovais Ahmad. Explainable ai for

- depression detection and severity classification from activity data: Development and evaluation study of an interpretable framework. *JMIR Mental Health*, 12(1):e72038, 2025.
- [2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Open-face: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.
 - [3] Nidhi Bhatt, Amita Jain, Minni Jain, and Shobha Bhatt. Depression detection from speech using a voting ensemble approach. In *2024 IEEE 8th International Conference on Information and Communication Technology (CICT)*, pages 1–6. IEEE, 2024.
 - [4] Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Adrian Lopez Monroy, and Petr Motlicek. Daic-woz: On the validity of using the therapist’s prompts in automatic depression detection from clinical interviews. In *Proceedings of the 6th clinical natural language processing workshop*, pages 82–90, 2024.
 - [5] Jingxia Chen, Yikun Ning, Pengwei Zhang, Haifeng Chen, and Huiru Zheng. Multimodal depression detection based on modality feature decomposition and contrastive learning. *Algorithms*, 19(6):452, 2026.
 - [6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
 - [7] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.
 - [8] Mariel Estevez, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. A fair and transparent framework for speech-based depression detection: Balancing interpretability and performance. *arXiv preprint arXiv:2606.31730*, 2026.
 - [9] María Teresa García-Ordás, Sergio Rubio-Martín, Alicia Merayo, Arturo Crespo-Álvarez, Oscar García-Olalla, and Isaías García-Rodríguez. Explainable depression prediction using shap and large language models to extract features from text transcriptions. *Applied Soft Computing*, page 116129, 2026.
 - [10] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik, 2014.
 - [11] Md Rezwanul Haque, Md Milon Islam, SM Taslim Uddin Raju, Hamdi Altaheri, Lobna Nassar, and Fakhri Karray. Mmformer: Multimodal fusion transformer network for depression detection. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 5912–5919. IEEE, 2025.
 - [12] Georgios Ioannides, Adrian Kieback, Aman Chadha, and Aaron Elkins. Density adaptive attention-based speech network: enhancing feature understanding for mental health disorders. *arXiv preprint arXiv:2409.00391*, 2024.
 - [13] Nur Nesa Nashuha Ismail, Nurulhuda Firdaus Mohd Azmi, Nilam Nur Amir Sjarif, and Nik Nur Wahidah Nik Hashim. Interpretable transformer-based depression detection in low-resource malay speech using shap-based feature analysis. *IEEE Access*, 2026.
 - [14] Ngumimi Karen Iyortsuun, Soo-Hyung Kim, Hyung-Jeong Yang, Seung-Won Kim, and Min Jhon. Additive cross-modal attention network (acma) for depression detection based on audio and textual features. *IEEE Access*, 12:20479–20489, 2024.
 - [15] Achraf Jallaglag, My Abdelouahed Sabri, Ali Yahyaouy, and Abdellah Aarab. Lightweight depression detection using 3d facial landmark pseudo-images and cnn-lstm on daic-woz and e-daic. *BioMedInformatics*, 6(1):8, 2026.
 - [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
 - [17] Arvind Kumar, Manyu Kumar, Abhay Raj, Agnik Banerjee, and Gopal Giri. Leveraging heuristic-driven explanations and sentence transformers for interpretable depression detection. In *International Conference on Intelligent Vision and Computing*, pages 268–281. Springer, 2025.
 - [18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
 - [19] Farah Mohammad and Khulood Mohammed Al Mansoor. Mdd: a unified multimodal deep learning approach for depression diagnosis based on text and audio speech. *Computers, Materials & Continua*, 81(3):4125–4147, 2024.
 - [20] Nguyen Nguyen, Minh Nguyen, Tri Doan, Hau Vo, Nha Tran, and Hung Nguyen. Vta-depressnet: a dual-level attention fusion for multimodal depression detection. *CCF Transactions on Pervasive Computing and Interaction*, pages 1–20, 2026.
 - [21] World Health Organization et al. Depressive disorder (depression). who fact sheet. 2023. URL <https://www.who.int/news-room/fact-sheets/detail/depression>. Accessed, pages 09–01, 2025.
 - [22] Francisco Ortiz-Vidal, Pedro Salcedo-Lagos, and Juan M Górriz Sáez. Multimodal depression detection using hierarchical fusion of distilbert and bi-lstm encoders: A deep learning approach on the daic-woz corpus. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 272–279. Springer, 2026.
 - [23] Hitesh Arjunbhai Ramrakhiyani, Himashri Deka, Sandeep Kumar Pandey, NS Sreenivasalu, Hanumant Singh Shekhawat, and Ravi Jasuja. Audio spectrogram transformer for automatic speech-based depression detection. In *International Conference on Intelligent Human Computer Interaction*, pages 484–492. Springer, 2025.
 - [24] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria 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*, pages 3–12, 2019.
 - [25] Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63, 2019.
 - [26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.
 - [27] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.
 - [28] Rishitej Reddy Vyalla, Kritarth Prasad, Avinash Anand, Erik Cambria, Shaoxiong Ji, Faten S Alamri, and Zhengkui Wang. Psychologically-grounded graph modeling for interpretable depression detection. *arXiv preprint arXiv:2604.24126*, 2026.
 - [29] Wenjie Zheng, Qiming Xie, Zengzhi Wang, Jianfei Yu, and Rui Xia. Towards explainable multimodal depression recognition for clinical interviews. *arXiv preprint arXiv:2501.16106*, 2025.