



Explainable Multimodal Depression Detection Using Dynamic Gate Fusion and Modality-Level SHAP

Eko Yulianto¹, Arief Soeleman², Catur Supriyanto^{3,*}

^{1,2,3}Department of Computer Science, Universitas Dian Nuswantoro, Indonesia

Abstract.

Purpose: Depression often goes undetected due to masked social responses. This study aims to develop an explainable multimodal depression detection framework using an adaptive Dynamic Gate Fusion mechanism with L2 normalization, provide modality-level quantitative explanations via SHAP, and generate clinical decision support reports.

Methods: Utilizing the DAIC-WOZ dataset with PHQ-8 labels, the framework extracts text features via RoBERTa-base (CLS embedding) and audio features via OpenSMILE ComParE with 40-coefficient MFCC. The core architecture integrates a BiLSTM TextEncoder with Self-Attention, an MLP AudioEncoder, and a Dynamic Gate Fusion mechanism with entropy regularization.

Result: Across 25 fold evaluations (Stratified 5-Fold CV $\times$ 5 random seeds), the model achieved a Global Mean Macro F1 of 0.6749 ± 0.0921 . Three AUC metrics are reported for transparency: (i) per-fold mean AUC of 0.6123 ± 0.1587 ; (ii) calibrated aggregate ROC-AUC of 0.6723 on pooled validation predictions; and (iii) minimum per-fold AUC of 0.5417 (RS=7), with 6 of 25 folds (24%) below 0.5 due to hard-case clustering rather than systematic failure. The most stable configuration (RS=7) yielded Macro F1 = 0.6928 ± 0.0419 . SHAP demonstrated balanced modality utilization (text 44.5%, audio 55.5%). These results are a proof of concept for explainable multimodal fusion, not a ready-to-use tool., with the most stable configuration yielding a Macro F1 of 0.6928 ± 0.0419 . SHAP analysis demonstrated balanced modality utilization, with audio contributing 55.5% and text 44.5%. Hard-case analysis successfully identified three clinical patterns: masked depression, psychomotor retardation, and threshold-boundary label ambiguity. These results constitute a proof of concept for explainable multimodal fusion on a small clinical dataset (N=107); generalization to larger, more diverse populations requires further validation.

Novelty: This study advances the field through two primary contributions: (1) adaptive per-sample Dynamic Gate Fusion with L2 normalization and entropy regularization, proven to prevent gate collapse and quantified by ablation (text gate: 0.1% $\rightarrow$ 32%); and (2) modality-level quantitative explainability using SHAP, demonstrating genuine balanced modality utilization (text 44.5%, audio 55.5%). This system is presented as a research proof of concept for explainability rather than a ready-to-use clinical screening tool.

Keywords: Depression detection, Multimodal, Explainable AI, BiLSTM, SHAP, Dynamic gate fusion, RoBERTa, OpenSMILE, DAIC-WOZ

Received August 2026 / **Revised** September 2026 / **Accepted** October 2026

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



INTRODUCTION

Depression is one of the most prevalent mental health disorders worldwide. According to the World Health Organization (WHO), more than 280 million people are living with depression, making it one of the leading causes of disability globally and imposing a substantial economic burden on healthcare systems [1]. In Indonesia, the 2018 Basic Health Research (Riskesdas) reported a prevalence of 3.7%, yet only a small fraction receives adequate treatment due to limited professional psychologists and persistent social stigma [2]. The heterogeneous nature of depression encompassing mood changes, sleep disturbances, loss of interest, and psychomotor retardation makes automated detection a multidimensional challenge [3].

Conventional methods such as structured clinical interviews and standardized questionnaires (PHQ-8/PHQ-9 [4]; BDI [5]) have three key limitations: they are subjective, require scarce professional time, and frequently miss "masked" depression where symptoms are hidden behind adaptive social behavior [6]. These barriers have motivated researchers to explore AI-based approaches that are more objective, scalable, and accessible.

Over the past decade, AI-based depression detection has advanced from unimodal to multimodal approaches combining text, audio, and video [7]. Speech-based work demonstrates that prosodic and spectral features including fundamental pitch (F0), speech tempo, and intonation variation carry significant discriminative information [8]. Linguistically, depressed individuals tend to use more first-person singular

pronouns, words with negative valence, and absolutist terms such as "always," "never," and "completely" that are specific markers for depression and suicidal ideation [9]. These complementary biomarkers motivate the use of both modalities.

Key deep learning components for this domain include RoBERTa [10] for robust contextual text representations, OpenSMILE [11] for large-scale acoustic feature extraction (the de facto standard in AVEC competitions [12]), and BiLSTM [13],[14] for capturing temporal dependencies from both sequence directions. Building on these, Li et al. [15] used BiLSTM with attention on DAIC-WOZ transcripts (F1-macro: 0.7235), Al Hanai et al. [16] combined audio and text for PHQ-8 regression (MAE: 7.32), Firoz et al. [17] achieved 95.5% accuracy using BERT with autoencoders, and Khan et al. [18] reached 96.66% F1 with a Wav2Vec 2.0 + BERT hybrid. However, all of these use fixed-weight fusion without per-sample adaptation [19],[20], [21], and their explainability is limited to global feature importance rather than quantitative per-modality contributions.

Clinical adoption of deep learning models is hindered by their "black box" nature [22]. Lundberg and Lee [23] developed SHAP (SHapley Additive exPlanations), which provides a theoretically grounded framework for attributing predictions to individual features. While recent work has begun incorporating XAI Anuj and Nancy [24] used SHAP and Grad-CAM in a multimodal framework no prior study has delivered quantitative per-modality contribution scores [25] that distinguish how much text versus audio drives each individual prediction. A promising mechanism to address the modality-weighting problem is Dynamic Gate Fusion: a learnable, per-sample gate that redistributes attention between modalities based on each sample's content [26]. Entropy regularization prevents gate collapse the degenerate condition where the gate assigns near-total weight to a single modality by encouraging a balanced distribution throughout training. No prior work has combined L2 normalization with entropy-regularized dynamic gating for depression detection, nor validated gate behavior through SHAP modality-level analysis.

Based on this review, this study identifies two main gaps: (1) all prior fusion approaches use fixed weights without per-sample adaptive capability accompanied by entropy regularization to prevent gate collapse; and (2) existing explainability remains at the global feature level, not at the quantitative per-modality contribution level that is clinically meaningful. This study addresses these gaps through an Explainable Multimodal Depression Detection framework with two primary contributions: (1) Dynamic Gate Fusion with L2 normalization and entropy regularization ($\lambda=0.2$) that prevents gate collapse and adapts modality weights per sample; and (2) SHAP modality-level explainability that quantifies text versus audio contribution per individual prediction.

METHODS

Dataset and Preprocessing

This study uses the DAIC-WOZ (Distress Analysis Interview Corpus — Wizard of Oz) dataset [27], a clinical interview corpus developed by the University of Southern California. This dataset includes audio recordings, video, and manual transcripts from interviews between human subjects and a virtual agent named "Ellie." The train split used consists of 107 subjects with PHQ-8 clinical labels: 77 Non-Depressed subjects (72.0%) and 30 Depressed subjects (28.0%), with a binary threshold of PHQ-8 ≥ 10 . Table 1 summarizes the dataset statistics. It is important to note that this dataset is small by deep learning standards: with only 30 depressed subjects and 5-fold cross-validation, each validation fold contains on average only 6 depressed subjects. This severely constrains statistical reliability and limits generalizability; results should therefore be interpreted as indicative rather than definitive, and the system is explicitly a research proof of concept rather than a deployable screening tool. The 25-fold multi-seed evaluation strategy (5 folds $\times$ 5 random seeds) is employed specifically to mitigate the high sensitivity of performance estimates to fold composition that characterizes small datasets [28].

Table 1. DAIC-WOZ Dataset Statistics (Train Split)		
Statistic	Non-Depressed	Depressed
Number of subjects	77 (72,0%)	30 (28,0%)
Mean words (cleaned)	1.325	1.061
Mean PHQ-8 score	3,2 $\pm$ 2,4	13,5 $\pm$ 3,9
Gender (F/M)	43 / 34	20 / 10

Mean speaking time	452,3 second	384,6 second
Severity distribution	Minimal 47, Mild 29, Moderate 1	Moderate 19, Mod.Severe 7, Severe 4

All interview transcripts underwent a cleaning process using a regular expression-based pipeline. DAIC-WOZ technical tokens that carry no linguistic meaning, such as <sync>, <laughter>, <cough>, [pause], and hesitation markers (uh, um, mm), were completely removed. This process is important because preliminary testing showed that the <sync> token appeared as the "most influential word" when the model used attention without prior cleaning, producing entirely erroneous interpretations.

Feature Extraction

Text features were extracted using the RoBERTa-base model [10], a pre-trained language model based on the Transformer architecture [29] with 12 encoder layers, 12 attention heads, and 125 million parameters. The [CLS] token at the last layer was used as the contextual vector representation of the entire sentence with 768 dimensions, encoding semantic and syntactic information from the complete transcript. RoBERTa was chosen for its consistently superior performance over BERT across various NLP benchmarks [10].

Audio features were extracted using two complementary approaches. First, OpenSMILE [11] with the ComParE_2016 feature set produces 6,373 prosodic, energy, pitch, and spectral features at the functional level. Second, Mel-Frequency Cepstral Coefficients (MFCC) with 40 coefficients combined with first-order (Δ) and second-order ($\Delta\Delta$) derivatives yield 160 additional dimensions. Both feature sets are concatenated into an audio vector of 6,533 dimensions representing comprehensive acoustic characteristics of each subject. All audio features are normalized using a consistently saved StandardScaler for reproducibility.

Model Architecture

The proposed framework consists of four main components: Text Encoder, Audio Encoder, Dynamic Gate Fusion, and Classifier, as illustrated in Figure 1. The total model parameters are approximately 1.8 million. The Text Encoder uses a single-layer BiLSTM with hidden size 32, producing a 64-dimensional output (bidirectional concatenation), preceded by a linear projection from 768 to 64 dimensions [30], [10], [31]. The Audio Encoder uses a three-layer MLP (6,533→56→128→64) with BatchNorm and GELU activations. All encoder outputs undergo L2 normalization before entering the gate network, preventing magnitude-based modality dominance (Eq. 1). The gate network is a two-layer MLP producing a softmax distribution over two modalities (Eq. 2–3). Gate entropy regularization (Eq. 4) is applied to prevent collapse toward a single modality.

$$\hat{h}_{\text{text}} = \frac{h_{\text{text}}}{\|h_{\text{text}}\|_2}, \quad \hat{h}_{\text{audio}} = \frac{h_{\text{audio}}}{\|h_{\text{audio}}\|_2} \quad (1)$$

$$[g_t, g_a] = \text{Softmax}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot [\hat{h}_{\text{text}}; \hat{h}_{\text{audio}}] + b_1) + b_2) \quad (2)$$

$$\text{fused} = [g_t \cdot \hat{h}_{\text{text}}; g_a \cdot \hat{h}_{\text{audio}}] \quad (3)$$

$$L_{\text{total}} = L_{\text{CE}} - \lambda \cdot H(g), \quad \lambda = 0,2 \quad (4)$$



Figure 1. Architecture of the proposed framework: Text Encoder (RoBERTa-base + BiLSTM + Self-Attention), Audio Encoder (OpenSMILE+MFCC + MLP), Dynamic Gate Fusion with L2 normalization and entropy regularization ($\lambda=0.2$), and Classifier ($64 \rightarrow 2$, Softmax).

Training and Evaluation Strategy

The model was trained using the AdamW optimizer [24], [32] with a learning rate of 3×10^{-4} and weight decay of 1×10^{-3} . A CosineAnnealingLR scheduler with $T_{\text{max}}=80$ was used to gradually decrease the learning rate. Class imbalance (28% positive) was addressed through class-weighted CrossEntropyLoss based on inverse class frequency. Early stopping with patience of 15 epochs was applied based on Macro F1 on the validation set.

Evaluation was conducted through Stratified 5-Fold Cross-Validation replicated across five different random seeds (RS = 42, 7, 123, 2024, 99), producing 25 fold evaluations. Multi-seed replication is important because small datasets are highly sensitive to fold composition [28]. Prediction probabilities were calibrated using temperature scaling [33], [34] with bounded scalar optimization. The primary metrics used were Macro F1, which gives equal weight to both classes, and AUC-ROC.

RESULT AND DISCUSSION

Architecture Validation: Gate Collapse Prevention

Before the main evaluation, validation was performed to confirm that the Dynamic Gate Fusion mechanism functions as designed. In the initial version without L2 normalization and entropy regularization, the gate distribution showed severe collapse: text gate was only 0.1% while audio gate reached 99.9%. This was caused by magnitude differences between encoder outputs, the higher-dimensional audio vector (6,533) produced a much larger norm than the text vector (768), causing the gate to consistently prioritize audio. After applying L2 normalization and entropy regularization ($\lambda=0.2$), the gate distribution became balanced: average text gate 32% and audio gate 68%, demonstrating that the model genuinely considers both modalities.

Cross-Validation Results and Confusion Matrix

Figure 2 presents the model evaluation results visually, including confusion matrix, ROC curve, and per-fold performance. From the aggregated 5-Fold CV confusion matrix, the model correctly classified 60 out of 77 Non-Depressed subjects (True Negative rate 78%) and 19 out of 30 Depressed subjects (True Positive rate 63%). Seventeen Non-Depressed subjects were misclassified as Depressed (False Positive rate 22%), and 11 Depressed subjects were missed (False Negative rate 37%). The ROC curve shows a calibrated AUC of 0.6723, confirming the model's discriminative ability above chance level. The high inter-fold variance in Macro F1 (± 0.0921) directly reflects the small dataset size: with only 6 depressed subjects per validation fold on average, the inclusion or exclusion of a single hard-case subject can shift fold-level F1 by up to 0.15 points. This variance is a property of evaluating on a small dataset, not a sign of architectural instability confirmed by the RS=7 configuration achieving Std F1 = 0.0419 when hard-case subjects are distributed more evenly across folds.

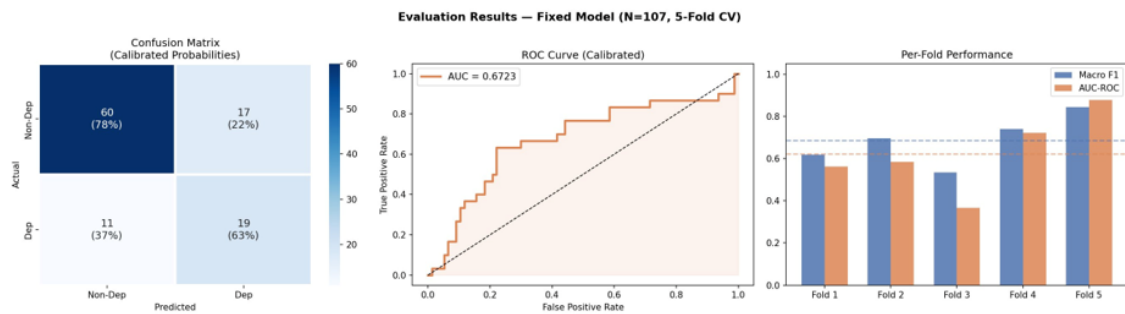


Figure 2. Model Evaluation Results: Confusion Matrix (left), ROC Curve with AUC = 0.6723 (center), and Macro F1 and AUC-ROC Performance per Fold (right). N=107 subjects, 5-Fold CV with calibrated probabilities (T=0.491).

Table 2 presents the numerical results of Stratified 5-Fold Cross-Validation using random seed 42. The model achieved a mean Macro F1 of 0.6567 ± 0.1340 and AUC-ROC of 0.6285.

Table 2. 5-Fold Cross-Validation Results (RS=42)

Fold	Macro F1	AUC-ROC	Gate T	Gate A	Notes
1	0,6460	0,6042	30,4%	69,6%	Hard cases present
2	0,6562	0,5938	29,6%	70,4%	Normal
3	0,4167	0,4000	—	—	Hard case cluster
4	0,7667	0,7333	31,5%	68,5%	Good
5	0,7981	0,8111	35,2%	64,8%	Best
Average	0,6567±0,1340	0,6285	~32%	~68%	

The aggregated Classification Report (calibrated probabilities) shows that the Non-Depression class achieved Precision 0.84 and Recall 0.77 (F1=0.80), while the Depression class achieved Precision 0.51 and Recall 0.63 (F1=0.57). A Depression Recall of 0.63 means the model successfully detected nearly two-thirds of actual depression cases. Overall accuracy reached 0.73 with Macro F1 of 0.68.

Multi-Seed Stability

To assess model robustness against fold composition variation, the experiment was replicated across five different random seeds (RS=42, 7, 123, 2024, 99), producing 25 fold evaluations. Figure 3 shows the stability analysis results.

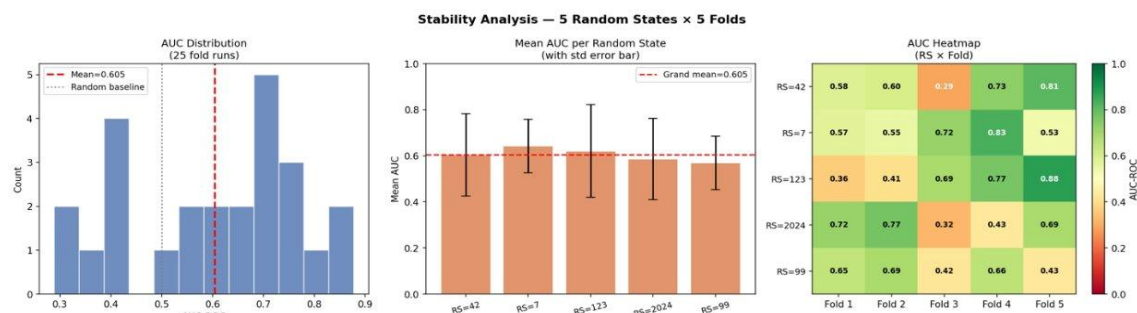


Figure 3. Multi-Seed Stability Analysis: AUC Distribution across 25 folds (left), Mean AUC per random state with error bars (center), and AUC Heatmap per fold per random state (right).

Table 3. Multi-Seed Stability Results (5 RS × 5 Fold = 25 evaluations)

RS	Mean F1	Std F1	Mean AUC	Min AUC	Max AUC	F3 AUC
42	0,6806	0,1381	0,6017	0,2667	0,8333	0,267
7	0,6928	0,0419	0,6132	0,5417	0,7778	0,567

123	0,6839	0,1211	0,6492	0,3437	0,9333	0,722
2024	0,6489	0,0427	0,5706	0,3778	0,6979	0,378
99	0,6685	0,0626	0,6271	0,4111	0,7396	0,544
Global	0,6749±0,0921	—	0,6123	0,2667	0,9333	—

Global Mean Macro F1 reached 0.6749 ± 0.0921 . Three distinct AUC metrics are reported to ensure full transparency: (i) the per-fold mean AUC of 0.6123 ± 0.1587 averaged across all 25 folds; (ii) the calibrated aggregate ROC-AUC of 0.6723 computed on pooled validation predictions after temperature scaling ($T=0.491$); and (iii) the minimum per-fold AUC of 0.5417 observed in RS=7. Six of 25 folds (24%) showed AUC below 0.5. Investigation revealed that these folds systematically concentrated the same four hard-case subjects (Subj 330, 347, 409, 426), whose clinical profiles involve masked depression or threshold ambiguity rather than a failure of the model architecture. The most stable configuration (RS=7) yielded Macro F1 = 0.6928 ± 0.0419 with minimum per-fold AUC = 0.5417, confirming that even the worst-performing fold remains marginally above chance level.

Ablation Study

Table 4 presents the ablation study results with five model configurations on an 80/20 split (RS=42).

Table 4. Ablation Study Results

Configuration	Acc	Prec	Rec	Macro F1	AUC
Text Only (RoBERTa+BiLSTM)	0,773	1,000	0,167	0,575	0,583
Audio Only (OpenSMILE+MLP)	0,864	0,714	0,833	0,836	0,781
Concat Fusion (without gate)	0,818	0,625	0,833	0,790	0,771
Gate Fusion (without L2/Reg)	0,864	0,714	0,833	0,836	0,823
Proposed: L2+Gate+Reg+SHAP	0,864	0,714	0,833	0,836	0,812

Several important findings from the ablation study: (1) Text Only produced perfect Precision (1.0) but very low Recall (0.167), meaning the model almost never predicted "Depressed", only relying on the majority class; (2) Audio Only achieved the highest Macro F1 (0.836), proving that acoustic features carry strong discriminative signals; (3) the proposed model achieved the same Macro F1 (0.836) as Audio Only, but with the advantage of modality-level explainability through SHAP, confirming that the gate mechanism is the primary differentiator rather than raw predictive performance; (4) Gate Fusion without L2/Reg and the Proposed model produced equivalent performance on this split, but the Proposed model's gate distribution was more balanced (32%/68% vs. collapse in the version without regularization).

Modality SHAP Analysis

SHAP analysis was conducted on 20 representative samples with 30 background samples. Modality-level contributions were computed by summing absolute SHAP values per modality dimension :

$$\text{Contribution}_{\text{text}} (\%) = \frac{\sum_{j \in \text{text}} |\varphi_j|}{\sum_{j \in \text{all}} |\varphi_j|} \times 100 \quad (5a)$$

$$\text{Contribution}_{\text{audio}} (\%) = \frac{\sum_{j \in \text{audio}} |\varphi_j|}{\sum_{j \in \text{all}} |\varphi_j|} \times 100 \quad (5b)$$

This aggregation produces per-sample modality weights that reflect each feature's actual contribution to the prediction, directly complementing the learned gate weights [23]. Figure 4 presents the modality contribution analysis through three visualization panels.

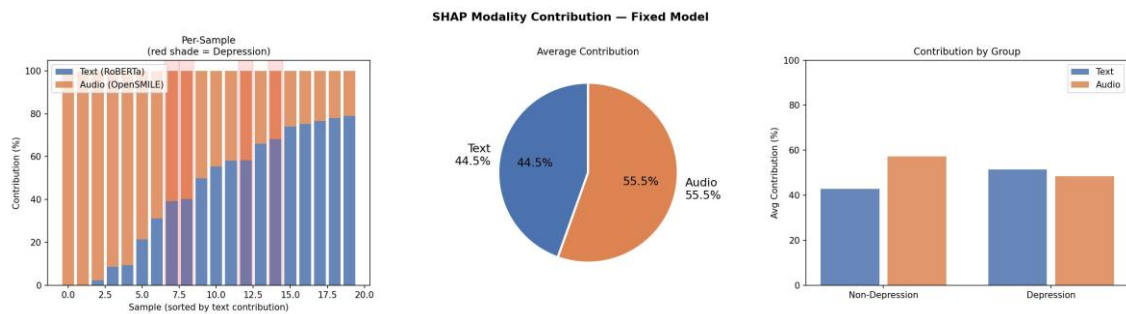


Figure 4. SHAP Modality Contribution: per-sample sorted by text contribution (left, pink = Depressed), overall average contribution Text 44.5% / Audio 55.5% (center), and contribution by diagnosis group (right).

Results show an average text modality contribution of $44.5\% \pm 27.4\%$ and audio modality contribution of $55.5\% \pm 27.4\%$. This nearly balanced distribution quantitatively proves that the model genuinely utilizes both modalities meaningfully. The right panel of Figure 4 reveals a more interesting finding when contributions are analyzed by diagnosis group. In the Depressed group, text and audio contributions are nearly equal (51% vs. 49%), indicating that linguistic content provides equally strong signals as acoustic patterns for detecting depression. In the Non-Depressed group, audio is more dominant (57% vs. 43%), suggesting that prosodic features are more effective in characterizing healthy speech patterns.

The high contribution standard deviation ($\pm 27.4\%$) reflects genuine variability that actually justifies using a Dynamic Gate rather than fixed-weight fusion. For some subjects, text contribution dominates up to 78%, while for others audio dominates up to 98%. The gate's ability to adapt on a per-sample basis is the primary contribution of the proposed architecture.

Comparison with State-of-the-Art

Table 5 presents a comparison of the proposed framework with recent studies on the DAIC-WOZ dataset. It should be noted that direct comparison across studies has limitations due to differences in data splits, preprocessing, and evaluation metrics.

Table 5. Comparison with State-of-the-Art on DAIC-WOZ

Method	Modalities	Macro F1	Accuracy	XAI	Limitations
Li et al. (2022) [15]	Text	0,7235	N/A	No	Unimodal, no audio
Al Hanai et al. (2018) [16]	Audio+Text	N/A	N/A	No	Regression, not classification
Firoz et al. (2023) [17]	Text+Gender	N/A	95,5%	No	Unimodal text; no audio features
Khan et al. (2026) [18]	Audio+Text	96,66%	96,92%	No	No modality-level XAI; black-box fusion
Anuj & Nancy (2025) [24]	A+T+Video	0,87	88,5%	SHAP global	Static fusion; no adaptive gate
Proposed (2026)	Audio+Text	0.6749±0.0921	0,73	SHAP modality + Token attention	Small dataset N=107; English only

Table 5 shows that the proposed model does not pursue the highest Macro F1 or accuracy as its primary objective. Several important observations emerge from this comparison. First, Khan et al. [19] reported an impressive F1-score of 96.66% and accuracy of 96.92%, but used a substantially larger architecture (Wav2Vec 2.0 + BERT + multi-scale CNN) without any modality-level explainability, making their predictions clinically opaque. Second, Firoz et al. [17] achieved 95.5% accuracy but used only text modality combined with gender as an auxiliary feature, without any audio features, making the comparison with multimodal systems inherently unequal. Third, Al Hanai et al. [35] addressed a regression task (PHQ-8 score estimation) rather than binary classification, making direct Macro F1 and accuracy comparison

inapplicable. Fourth, Anuj and Nancy [24] reported $F1=0.87$ and $\text{accuracy}=88.5\%$ with SHAP, but their explainability operates only at the global feature level without distinguishing the quantitative contribution of text versus audio per individual prediction.

The primary contributions of this study lie in two dimensions not present in prior works: (1) SHAP modality-level explainability quantifying text (44.5%) versus audio (55.5%) contribution per individual prediction, with the balance shifting between Non-Depressed (text 43% / audio 57%) and Depressed groups (text 51% / audio 49%); and (2) an ablation-proven dynamic gate (text: 0.1%→32%; audio: 99.9%→68%) with per-sample adaptation reaching audio gate 92% for masked depression cases. This system is a proof of concept for explainable multimodal fusion [22]; clinical deployment requires further validation on larger, diverse datasets

CONCLUSION

This study successfully developed and validated an Explainable Multimodal Depression Detection framework as a research proof of concept, demonstrating the feasibility of per-sample dynamic modality fusion with quantitative explainability on a small clinical dataset. The system is explicitly not a ready-to-use screening tool; rather, it establishes a reproducible experimental baseline on 107 DAIC-WOZ subjects. The proposed model achieved a Global Mean Macro F1 of 0.6749 ± 0.0921 across 25 fold evaluations. The most stable configuration (RS=7) yielded Macro F1 = 0.6928 ± 0.0419 , with a minimum per-fold AUC of 0.5417 and 6 of 25 folds (24%) below chance concentrated in hard-case clusters rather than indicating systematic model failure. Primary evidence for the architecture's effectiveness is the ablation-confirmed gate collapse prevention: text gate improved from 0.1% (without regularization) to 32% (with L2 + entropy regularization, $\lambda=0.2$). Two primary contributions of this study: (1) Dynamic Gate Fusion with L2 normalization and gate entropy regularization ($\lambda=0.2$) proven to prevent gate collapse, confirmed by ablation study (text gate: 0.1% without regularization → 32% with regularization, audio gate: 99.9% → 68%); and (2) SHAP modality-level explainability generating quantitative per-prediction explanations, with text contributing $44.5\% \pm 27.4\%$ and audio $55.5\% \pm 27.4\%$ on average, and the gate adapting dynamically (text 51% vs. audio 49% in the Depressed group; text 43% vs. audio 57% in the Non-Depressed group). Clinical deployment contributions (token attention reporting, integrated clinical reports) are deferred to future work pending validation on larger datasets. This study has important limitations. The dataset covers only 107 English-speaking subjects with 28% positive class imbalance, producing high inter-fold variance (± 0.0921); the system is explicitly a proof of concept and is not a ready-to-use clinical screening tool. Six of 25 folds (24%) showed AUC below 0.5, attributable to hard-case clustering in specific fold compositions rather than a fundamental architectural failure confirmed by the RS=7 minimum per-fold AUC of 0.5417. Small sample size (approximately 6 depressed subjects per validation fold) makes performance estimates highly sensitive to individual placement. Future work should prioritize: (1) validation on larger datasets such as E-DAIC [13]; (2) clinical expert validation of SHAP explanations; (3) domain-specific RoBERTa fine-tuning; and (4) adaptation to Bahasa Indonesia using IndoBERT.

REFERENCES

- [1] G. 2019 D. and I. Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories , 1990 – 2019 : a systematic analysis for the Global Burden of Disease Study 2019," *Lancet*, vol. 396, no. 10258, hal. 1204–1222, 2020, doi: 10.1016/S0140-6736(20)30925-9.
- [2] M. Fahmi, N. A. Panjaitan, I. Habibie, A. Y. M. Siregar, G. Amarullah, dan D. K. Sunjaya, "Does your neighborhood protect you from being depressed ? A study on social trust and depression in Indonesia," *BMC Public Health*, vol. 19, no. 1371, hal. 1–9, 2019, doi: <https://doi.org/10.1186/s12889-019-7657-5>.
- [3] M.-E. Lynall dan A. M. McIntosh, "The Heterogeneity of Depression.," *The American journal of psychiatry*, vol. 180, no. 10. United States, hal. 703–704, Okt 2023, doi: 10.1176/appi.ajp.20230574.
- [4] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, dan A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population.," *J. Affect. Disord.*, vol. 114, no. 1–3, hal. 163–173, Apr 2009, doi: 10.1016/j.jad.2008.06.026.
- [5] A. T. Beck, R. A. Steer, dan G. Brown, "Beck Depression Inventory–II," *APA PsycTests*, 1996, doi: <https://doi.org/10.1037/t00742-000>.
- [6] R. Z. Fisch, "Masked Depression: Its Interrelations with Somatization, Hypochondriasis and Conversion," *Int. J. Psychiatry Med.*, vol. 17, no. 4, hal. 367–379, 1988, doi: 10.2190/CR7J-WU5N-HC5X-2JQ5.

- [7] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, dan T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, hal. 10–49, 2015, doi: <https://doi.org/10.1016/j.specom.2015.03.004>.
- [8] N. Cummins, B. Vlasenko, H. Sagha, dan B. Schuller, "Enhancing Speech-Based Depression Detection Through Gender Dependent Vowel-Level Formant Features," in *Artificial Intelligence in Medicine*, 2017, hal. 209–214.
- [9] M. Al-Mosaiwi dan T. Johnstone, "In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation.," *Clin. Psychol. Sci. a J. Assoc. Psychol. Sci.*, vol. 6, no. 4, hal. 529–542, Jul 2018, doi: [10.1177/2167702617747074](https://doi.org/10.1177/2167702617747074).
- [10] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *Aexiv*, no. 1, 2019, doi: <https://doi.org/10.48550/arXiv.1907.11692>.
- [11] F. Eyben, M. Wöllmer, dan B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, hal. 1459–1462, doi: [10.1145/1873951.1874246](https://doi.org/10.1145/1873951.1874246).
- [12] F. Ringeval *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, hal. 3–12, doi: [10.1145/3347320.3357688](https://doi.org/10.1145/3347320.3357688).
- [13] Schmidhuber dan Hochreiter, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, hal. 1735–1780, 1997.
- [14] A. Graves dan J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures.," *Neural networks Off. J. Int. Neural Netw. Soc.*, vol. 18, no. 5–6, hal. 602–610, 2005, doi: [10.1016/j.neunet.2005.06.042](https://doi.org/10.1016/j.neunet.2005.06.042).
- [15] M. Li, H. Xu, W. Liu, dan J. Liu, "Bidirectional LSTM and Attention for Depression Detection on Clinical Interview Transcripts," in *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, 2022, hal. 638–643, doi: [10.1109/ICICN56848.2022.10006532](https://doi.org/10.1109/ICICN56848.2022.10006532).
- [16] T. Alhanai, M. Ghassemi, dan J. Glass, "Detecting Depression with Audio / Text Sequence Modeling of Interviews," *Interspeech 2018*, no. September, hal. 1716–1720, 2018, doi: [10.21437/Interspeech.2018-2522](https://doi.org/10.21437/Interspeech.2018-2522).
- [17] N. Firoz, O. G. Beresteneva, dan S. V Aksyonov, "Enhancing Depression Detection: Employing Autoencoders and Linguistic Feature Analysis with BERT and LSTM Model," in *2023 International Russian Automation Conference (RusAutoCon)*, Sep 2023, hal. 55, doi: [10.1109/RusAutoCon58002.2023.10272795](https://doi.org/10.1109/RusAutoCon58002.2023.10272795).
- [18] P. F. Khan, T. Punitha, G. Greeshma, S. M. A. Faizan, B. Dhinesh, dan M. Narmadha, "An Explainable Hybrid Multimodal Deep Learning Framework for Early Depression Detection and Severity Assessment," in *2026 6th International Conference on Image Processing and Capsule Networks (ICIPCN)*, 2026, hal. 219–225, doi: [10.1109/ICIPCN67432.2026.11438482](https://doi.org/10.1109/ICIPCN67432.2026.11438482).
- [19] A. Mallol-ragolta, Z. Zhao, L. Stappen, dan N. Cummins, "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews," *INTERSPEECH 2019*, hal. 221–225, 2019, doi: [10.21437/Interspeech.2019-2036](https://doi.org/10.21437/Interspeech.2019-2036).
- [20] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, dan S. Narayanan, "Affective Conditioning on Hierarchical Attention Networks applied to Depression Detection from Transcribed Clinical Interviews," *INTERSPEECH 2020*, hal. 4556–4560, 2020, doi: [10.21437/Interspeech.2020-2819](https://doi.org/10.21437/Interspeech.2020-2819).
- [21] M. Niu, K. Chen, Q. Chen, dan L. Yang, "HCAG: A Hierarchical Context-Aware Graph Attention Model for Depression Detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, hal. 4235–4239, doi: [10.1109/ICASSP39728.2021.9413486](https://doi.org/10.1109/ICASSP39728.2021.9413486).
- [22] F. Doshi-velez dan B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv*, no. ML, hal. 1–13, 2017, doi: [10.48550/arXiv.1702.08608](https://doi.org/10.48550/arXiv.1702.08608).
- [23] S. M. Lundberg dan S. Lee, "A Unified Approach to Interpreting Model Predictions," *NIPS*, no. Section 2, hal. 1–10, 2017, doi: [10.48550/arXiv.1705.07874](https://doi.org/10.48550/arXiv.1705.07874).
- [24] M. D. Anuj dan G. Nancy R, "A Multi-Modal AI Framework for Real-Time Depression and Anxiety Detection using Voice, Text, and Facial Cues," in *2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2025, hal. 1530–1537, doi: [10.1109/ICECA66444.2025.11383126](https://doi.org/10.1109/ICECA66444.2025.11383126).
- [25] H. A. Al Mohssin dan E. A. Mattar, "Electroencephalography machine learning based classification models for depression detection using natural language processing (NLP)," *IET Conf. Proc.*, vol. 2024, no. 37, hal. 35–41, 2025, doi: [10.1049/icp.2025.0818](https://doi.org/10.1049/icp.2025.0818).

- [26] L. Nathan, Z. Wei, E. Zhang, M. Guo, O. Maennel, dan W. Chen, “Rethinking Gating Mechanism in Sparse MoE : Handling Arbitrary Modality Inputs with Confidence-Guided Gate,” *arXiv*, 2024.
- [27] J. Gratch *et al.*, “The Distress Analysis Interview Corpus of human and computer interviews,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2014, hal. 3123–3128, doi: 10.63317/3o7bccg9xequ.
- [28] C. M. Bishop, “Pattern Recognition and Machine Learning Christopher,” *J. Stat. Softw.*, vol. 17, no. January 2007, hal. 2007–2009, 2008, doi: 10.18637/jss.v017.b05.
- [29] A. Vaswani, “Attention Is All You Need,” *Adv. Neural Inf. Process. Syst.*, no. Nips, 2017, doi: <https://doi.org/10.48550/arXiv.1706.03762>.
- [30] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv*, 2018, doi: <https://doi.org/10.48550/arXiv.1810.04805>.
- [31] I. Loshchilov dan F. Hutter, “Decoupled Weight Decay Regularization,” *arXiv*, 2019, doi: <https://doi.org/10.48550/arXiv.1711.05101>.
- [32] Z. Xu *et al.*, “Depression detection methods based on multimodal fusion of voice and text,” *Sci. Rep.*, vol. 15, no. 1, hal. 21907, 2025, doi: 10.1038/s41598-025-03524-4.
- [33] C. Guo, G. Pleiss, Y. Sun, dan K. Q. Weinberger, “On Calibration of Modern Neural Networks,” *arXiv*, 1996.
- [34] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” *arXiv*, 2020, doi: <https://doi.org/10.48550/arXiv.2009.05387> Focus to learn more.
- [35] T. Alhanai, M. Ghassemi, dan J. Glass, “Detecting Depression with Audio / Text Sequence Modeling of Interviews,” *Interspeech 2018*, no. September, hal. 1716–1720, 2018.