

Enhancing Depression Detection with Multimodal Cross-Attention: A Robust Framework Using the D-Vlog Dataset

Taufeeq Ahmed¹, Ramesh Chandra Sahoo^{2*}

¹ Research Scholar, Manav Rachna International Institute of Research and Studies, Faridabad, India.
Email: ahmed.taufeeq@gmail.com ORCID: 0009-0002-9353-6789

² Manav Rachna International Institute of Research and Studies, Faridabad, India.
Email: rcsahoo.set@mriu.edu.in ORCID: 0000-0003-0296-8106
Corresponding Author: Ramesh Chandra Sahoo

Abstract: Worldwide, depression is a serious public health issue, and struggles around early detection are evident in countless clinical practices. Nearly all traditional methods of depression detection involve interviewing patients and having patients fill out screening questionnaires, a process that is inefficient and tough to reproduce. As depression detection methods become more automated, the ability to identify clues that may suggest depression through behavioral changes in people is capturing more attention. This is especially true as more people share their personal lives through vlogs via social media. In this paper, we present the various techniques that we have built to find depression in vlog videos. In the proposed method, we employ a cross-attention mechanism to integrate acoustic and visual behavioral cues (or “modalities”) that reflect the alignment (or relationship) between the two. Then, we conduct cross-modal fusion of the processed individual streams from the specific encoders (or, ‘modality encoders’). We empirically assessed our proposed model using the D-Vlog dataset, consisting of 961 vlogs from 816 patients with depression annotations. Our proposed model is able to achieve 82% accuracy, 0.79 F1-score, and 0.89 Area Under Curve (AUC) score. Compared with unimodal/depression detection systems and traditional (i.e., early & late) cross fusion models, our proposed model outperformed all these models. The results suggest that explicitly modelling cross-modal relationships improves the ability to capture behavioural patterns associated with depressive states. These findings highlight the potential of multimodal learning for developing automated tools that can assist in large-scale depression screening using real-world video data.

Keywords: Multimodal Learning, Depression Detection, Psychological Assessment, Cross-Attention Mechanism, Multimodal Dataset

1. INTRODUCTION

The World Health Organization estimates the global burden of disease caused by Major Depressive Disorder (MDD). Depression, as a mental illness, is equally complicated, multifaceted, and difficult to treat as a social disease. The importance of early recognition of the severity of the illness cannot be overstated. The lack of recognition and appropriate diagnosis of depression has serious long-term impacts on the individual and society as a whole, affecting the psychosocial and economic domains. In a clinical context, a diagnosis typically means a structured clinical interview and case history analysis, along with a set of standard questionnaires completed by the patient, who is usually an expert. The advantage of this methodology is reliability. The disadvantage is the time it takes and the clinical skills required to implement it. In this instance, a clinical depression diagnosis is the best justification for the research interest in the automated estimation of behavioral phenomena to facilitate the early clinical evaluation of behavioral symptoms of depression in order to facilitate further research in this field.

Depression is associated with a variety of behavioral features, including reduced expressivity, slowed and reduced volume of speech, and a greater number of long speech pauses. Many of the features associated with



depression that have been listed in previous research can be documented and analyzed for their feasibility for use in advanced (machine learning) based behavioral analysis systems [1][2]. Since depressive symptoms may appear across several behavioural channels simultaneously, multimodal learning approaches that combine information from speech and facial expressions have been found to produce more robust predictions than models relying on a single modality [3].

Automated depression detection in the past has focused on studies in which participants are biased due to the controlled environment where the data was collected. For instance, DAIC-WOZ participants manually answer questions in interviews while being recorded in a quiet room. These studies are great for developing algorithms and benchmarking depression detection systems. However, detection systems fail to process data when they are moved to less structured environments. For example, the recording environment of social media videos is often dimly lit and has background noise, and the participants might behave erratically. These unstructured environments are a challenge, but also a chance to analyze depression detection systems in soft behavioral signals [4].

In the example of the D-Vlog dataset, researchers began to analyze depression detection on social media videos. The dataset is a collection of vlogs, some of which are tagged as depressed and some as not depressed [5]. Vlogs are still interviews, but they have a loose format where a person narrates a story in a free and natural way. As a result, they offer a useful setting for exploring multimodal models that aim to operate in real-world conditions.

Despite this progress, several challenges remain. Data collected in natural environments tend to contain greater variability and noise, which can affect both feature extraction and model stability. In addition, although many studies combine multiple modalities, the strategies used to fuse audio and visual information vary considerably across approaches. The relative effectiveness of different fusion mechanisms therefore remains an active area of investigation. Ethical considerations also play an important role when working with publicly available social media data, particularly with respect to privacy, consent, and potential demographic bias in the collected samples [6].

Motivated by these challenges, this work investigates a multimodal framework designed to analyse speech and facial cues from vlog recordings in order to identify indicators of depression. The proposed approach employs modality-specific encoders to capture temporal patterns in both audio and visual signals, followed by a cross-attention mechanism that models interactions between the two modalities. By learning relationships between speech dynamics and facial behaviour, the model seeks to improve the reliability of automated depression detection in naturalistic video data.

The main contributions of this study can be summarized as follows:

1. A multimodal framework is developed that integrates speech and facial information from vlog recordings for automated depression detection.
2. A cross-attention fusion mechanism is introduced to model interactions between audio and visual behavioural cues.
3. The proposed approach is evaluated against unimodal and conventional fusion baselines using the D-Vlog dataset.
4. Additional experiments on the DAIC-WOZ dataset provide insight into the ability of the model to generalize across different recording environments.

Through these contributions, the study aims to further advance research in multimodal mental health analysis and provide additional evidence that combining speech and visual information can improve automated depression screening in real-world video settings.

2. RELATED WORK

2.1 Behavioral and Speech-Based Depression Detection

Automated detection of depression has centred around diverse behavioral signals that indicate emotional and cognitive changes. In the past, researchers primarily analyzed facial expressions and identified reduced smiles and head movements as behaviors correlated with depression. Computer vision techniques have been developed to evaluate facial action units and expression changes to identify depression from video recordings [7], [8].

The analysis of speech has been another major area of research, as the vocal profile of depressed individuals shows important changes. Depressed speakers tend to speak more slowly, have less pitch variation, and take longer pauses. These characteristics of speech have been incorporated into many depression detection algorithms, including fundamental frequency, energy, and speech rate [9], [10]. Although visual information is not available, audio recordings have been shown to be sufficient to capture actionable behavioral information.

Beyond facial and speech signals, physiological measurements have also been investigated in the context of mental health analysis. Signals such as electroencephalography and heart rate variability can reveal biological responses related to emotional processing and stress. Deep learning models trained on physiological data have reported promising results in controlled experimental settings [11]. However, collecting such signals typically requires specialized sensors and laboratory conditions, which makes large scale data collection difficult. For this reason, recent work has increasingly focused on behavioral signals that can be extracted from widely available audiovisual recordings.

2.2 Multimodal Learning Approaches

Because depression affects multiple aspects of behavior simultaneously, combining information from several modalities has become an important research direction. Multimodal learning approaches attempt to integrate cues from speech, facial expressions, and sometimes textual information in order to produce more reliable predictions.

Several fusion strategies have been proposed to combine multimodal information. Early fusion approaches merge features from different modalities before model training, whereas late fusion methods combine predictions produced by separate unimodal models. Intermediate fusion strategies attempt to learn modality specific representations and then integrate them through shared neural layers [12], [13].

In recent years, attention-based models have been introduced to better capture relationships between modalities. These models allow the network to focus on the most informative signals at different points in time. Transformer architectures, in particular, have shown strong capability in modeling temporal dependencies and cross modal interactions. Studies using attention-based fusion have reported improved performance for depression detection tasks, especially when analyzing long conversational recordings or personal video narratives [14], [15].

Despite these advances, the effectiveness of different fusion strategies can vary depending on the dataset and recording conditions. Understanding how multimodal representations interact across different environments remains an active area of research.

2.3 Datasets and Research Challenges

The creation of models for the detection of depression has greatly relied on the existence of open-access datasets. Most depression datasets consist of data collected during closed interviews where a participant speaks with a virtual interviewer or responds to a set of predetermined questions. One of the most popular datasets used in this field is the DAIC WOZ dataset, which contains multiple audiovisual recordings directed at assessing clinical levels of depression [16].

Clearly, these datasets contain multiple important reference points. However, they often lack the diversity and variability found in most of the communications an average individual interfaces with. As a result, when data from interviews is used to train models, the models will often underperform when interfaced with non-structured interviews, where the data has a disparate range of recording situations and speaking styles.

To tackle this deficiency, a number of studies have focused on detecting depression via social media video blogs. The D Vlog dataset was created to facilitate the study of detecting depression through natural vlog recordings. The dataset contains hundreds of videos marked for the presence or absence of depressive states [17]. Unlike structured interviews, vlogs contain natural speech and face behavior as they are usually self-directed videos and involve a personal story of the vlogger. Therefore, the recorded material is much closer to the natural behavior of a person than the recorded material used in structured interviews.

More recent datasets have attempted to expand this approach by increasing both the scale and diversity of video samples. For example, the LMVD dataset provides a larger collection of annotated vlog recordings drawn from multiple online sources [18]. These resources allow researchers to investigate depression detection models under more realistic conditions where factors such as lighting variation, background noise, and recording quality are less controlled.

Despite the growing availability of datasets, several challenges remain. Many datasets provide only binary labels rather than clinically validated severity scores, which limits the ability to model fine grained depression levels. In addition, domain differences between datasets often lead to performance degradation when models trained on one dataset are evaluated on another [19], [20]. Finally, ethical issues related to privacy, consent, and demographic bias require careful consideration when using publicly available online data for mental health research [21].

3. DATASET DESCRIPTION

This study utilizes the D-Vlog dataset, which is the only dataset created for the study of depression detection in natural spontaneous audio-visual recordings. The dataset contains Vlogs from Youtube that were published from January 2020 to January 2021. Videos were collected using searches for personal, emotional, mental health, and narrative videos. Each video was screened to verify that it contained a first-person narrative and was of sufficient audio-visual quality to permit behavioral analysis. This video selection process was intended to avoid scripted videos to maintain the authentic quality typical of spontaneous online vlog recordings [22].

Video annotation was performed by a team of trained annotators. Each Vlog was classified as either depression or non-depression, depending on behavioral and linguistic criteria that were defined in the annotation guidelines. To resolve inconsistencies, multiple annotators reviewed each video, and discussions were held to resolve disagreements. Inter-annotator agreement was measured to establish confidence in the labelling. Such procedures are commonly adopted in behavioral annotation tasks to reduce subjective bias and improve dataset quality [23].

The final dataset contains 961 video blogs from 816 unique speakers, representing approximately 160 hours of recorded material. Instead of distributing raw video files, the dataset provides pre-extracted acoustic and visual feature representations. This design choice simplifies the preprocessing stage and allows researchers to focus on model development and multimodal analysis. The acoustic features include prosodic and spectral descriptors derived from speech signals, while the visual features consist of frame-level embeddings obtained from convolutional neural networks that capture facial expressions and head movements. Table 1 summarizes the main components of the dataset.

Table 1. D-Vlog dataset statistics

Component	Description	Size
Subjects	Unique vloggers	816
Vlogs	Total annotated videos	961
Acoustic features	Prosodic and spectral descriptors (.npy)	112 MB
Visual features	Frame level CNN embeddings (.npy)	597 MB
Labels	Depression and non-depression annotations	70 KB

The availability of synchronized audio and visual features makes the dataset suitable for evaluating multimodal learning methods. Because the recordings originate from natural online environments, the dataset also contains realistic variations in lighting, background noise, and recording conditions. These characteristics provide a useful testbed for studying depression detection models intended for real world deployment [24].

4. PROPOSED METHODOLOGY

This section describes the methodology adopted for detecting depressive indicators from vlog recordings. The proposed framework integrates acoustic and visual behavioral cues through a multimodal learning architecture. The overall pipeline begins with preprocessing of the audio and visual feature representations provided by the dataset and proceeds with a neural architecture that models temporal patterns within each modality and captures their interactions through a cross-attention mechanism.

The objective of this design is to allow the model to learn relationships between speech characteristics and facial behavior that may reflect depressive states. By jointly analyzing both modalities, the model can identify patterns that might not be observable when each modality is considered independently [25][26].

4.1 Feature preparation

The experiments utilize the pre extracted acoustic and visual features distributed with the D Vlog dataset. Acoustic features are derived from short segments of speech and include prosodic and spectral descriptors such as pitch, energy, speaking rate, and spectral coefficients. These features capture variations in vocal expression that are often associated with emotional and psychological conditions [27][28].

Visual features are represented as frame level embeddings extracted from convolutional neural networks trained on facial analysis tasks. These embeddings encode facial expressions, head movements, and other visual cues that may indicate behavioral changes related to depression [29].

Before training the model, several preprocessing steps are applied to ensure that the data from both modalities are suitable for multimodal learning. First, the feature values are normalized using z score standardization so that each feature dimension has zero mean and unit variance. This normalization improves training stability and ensures that the two modalities operate on comparable scales.

Next, the continuous feature sequences are divided into fixed length temporal segments. Each segment represents a short window of the original recording and is extracted with partial overlap to preserve temporal continuity. This segmentation strategy allows the model to process long recordings more efficiently while still capturing meaningful temporal patterns.

Minor inconsistencies such as missing visual frames or silent audio segments are addressed through simple imputation or masking. These procedures enable the model to learn from incomplete observations without introducing significant bias. After preprocessing, aligned audio and visual segments are grouped into batches so that the model can jointly process both modalities during training.

4.2 Multimodal model architecture

The proposed model consists of modality specific encoders followed by a cross-modal attention module that integrates the learned representations. An overview of the architecture is illustrated in Figure 1.

Each modality is first processed by a dedicated encoder designed to capture temporal patterns within the feature sequences. For the acoustic modality, a temporal neural network such as a bidirectional long short term memory network or a temporal convolutional layer is used to learn contextual representations from speech features. Similarly, the visual modality is processed through a temporal encoder that analyzes sequences of facial feature embeddings extracted from video frames. These encoders transform the raw input features into higher level representations that summarize temporal dynamics within each modality [30].

To formally describe the interaction between modalities, let

$$A = \{a_1, a_2, \dots, a_T\} \quad (1)$$

denote the sequence of acoustic feature vectors and

$$V = \{v_1, v_2, \dots, v_T\} \quad (2)$$

represent the sequence of visual feature vectors.

The encoded representations produced by the modality specific encoders can be expressed as

$$H_a = Enc_a(A) \quad (3)$$

$$H_v = Enc_v(V) \quad (4)$$

where H_a and H_v denote the contextual representations learned from audio and visual inputs respectively.

To capture dependencies between modalities, a cross-attention mechanism is applied. The general attention operation can be written as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (5)$$

where Q , K , and V represent query, key, and value matrices, and d denotes the feature dimension used for scaling.

In the proposed framework, the audio representation attends to the visual representation and the visual representation attends to the audio representation. This bidirectional interaction allows the model to associate speech characteristics such as pauses or pitch variation with corresponding facial expressions occurring at the same moment in the video. Attention based multimodal learning has been shown to be effective for modeling such cross-modal relationships in audiovisual analysis tasks [31], [32].

The outputs of the cross-attention module are then combined to form a unified multimodal representation. This fused representation is passed through a feedforward network with dropout regularization to reduce overfitting. The final layer produces a probability distribution over two classes representing depression and non-depression.

Recent multimodal transformer-based architectures have demonstrated that attention mechanisms can effectively capture interactions between behavioral modalities and improve prediction performance in mental health analysis tasks [33], [34], [35]. The proposed architecture follows a similar principle by allowing dynamic information exchange between audio and visual streams.

Figure 1 illustrates the overall architecture of the proposed model, including the modality specific encoders, the cross attention fusion module, and the final classification layer.

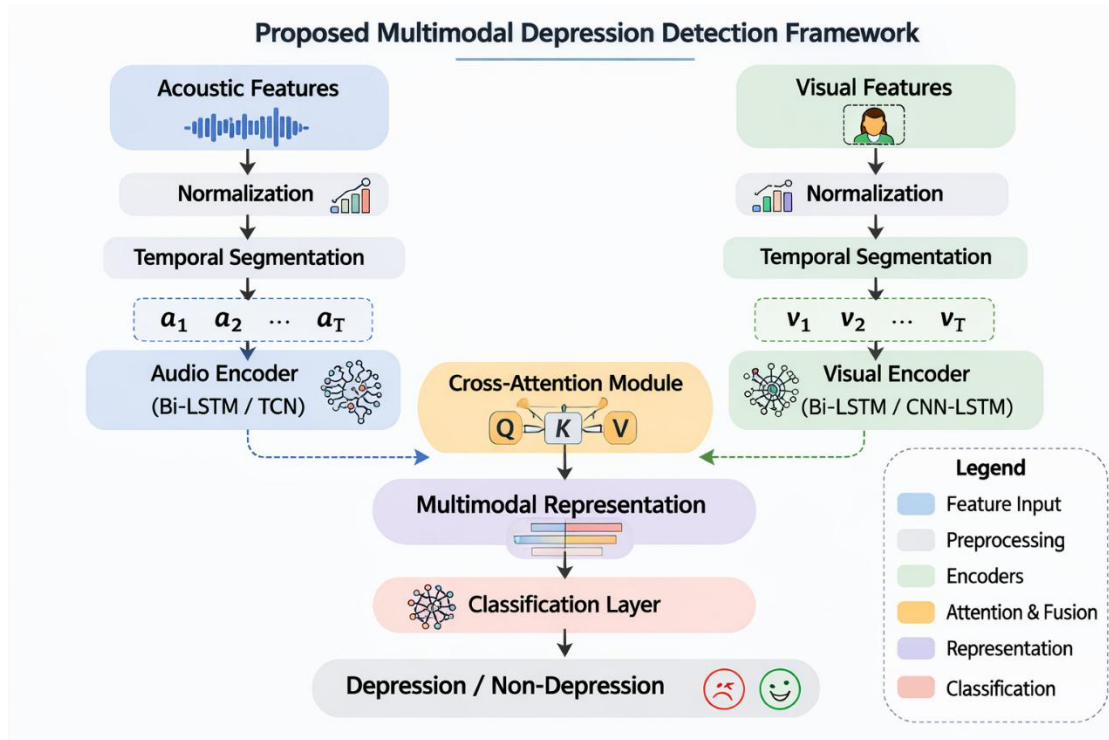


Figure 1. Proposed multimodal architecture for depression detection

Algorithm 1: Training procedure of the proposed multimodal model

Input:

Acoustic feature sequence A

Visual feature sequence V

Labels Y

Output:

Trained multimodal depression detection model

1. Load acoustic and visual feature sequences from dataset
2. Normalize features using z score standardization
3. Segment sequences into fixed length temporal windows
4. Handle missing segments using masking or simple imputation
5. Pass acoustic segments through the audio encoder to obtain H_a
6. Pass visual segments through the visual encoder to obtain H_v
7. Apply cross attention to model interactions between H_a and H_v
8. Fuse the attended representations into a multimodal feature vector
9. Feed fused representation into a fully connected network
10. Compute prediction probabilities using softmax layer
11. Calculate classification loss using cross entropy
12. Update model parameters using Adam optimizer

13. Repeat steps 5 to 12 for each training epoch
--

Algorithm 1 summarizes the training procedure of the proposed framework. The process begins with loading and preprocessing the audio and visual feature sequences obtained from the dataset. Each modality is processed through its respective encoder to extract temporal representations that capture behavioral dynamics. The cross-attention module then models the interactions between the two modalities, enabling the network to identify relationships between speech patterns and facial expressions. The fused representation produced by this module is finally passed to the classification layer, where the model predicts whether the input recording indicates depressive behavior. During training, the network parameters are optimized using cross entropy loss and the Adam optimizer until convergence.

5. EXPERIMENTAL DETAILS

This section presents the experimental methodology used to evaluate the proposed multimodal depression detection framework. The experiments are designed to assess classification performance, compare the proposed model with baseline approaches, analyze the contribution of individual architectural components, and evaluate the robustness and generalization capability of the framework.

5.1 Evaluation Metrics

To evaluate the effectiveness of the proposed depression detection framework, several widely used classification metrics are employed. These metrics provide complementary perspectives on model performance and are commonly adopted in machine learning studies related to behavioral analysis and healthcare applications.

Accuracy measures the proportion of correctly classified instances among all predictions and provides an overall indication of model performance. Precision represents the proportion of correctly predicted positive instances among all predicted positive samples, reflecting the model’s reliability in identifying depression cases.

Recall, also referred to as sensitivity, measures the proportion of actual depression cases that are correctly identified by the model. This metric is particularly important in mental health applications, where failing to detect individuals with depressive symptoms can have serious consequences.

The F1-score is defined as the harmonic mean of precision and recall and provides a balanced measure when both false positives and false negatives must be considered.

In addition, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used to evaluate the model’s ability to distinguish between depressed and non-depressed samples across different decision thresholds. A higher AUC value indicates stronger discriminative capability and improved classification reliability [38], [39].

Together, these evaluation metrics provide a comprehensive assessment of the performance of the proposed multimodal depression detection model.

5.2 Experimental Setup

Dataset Splitting Strategy

Estimating the performance of the models requires the dataset to be split first, for training purposes, validation, and testing. For training, an estimated 70% of the samples will be used, while 15% will be used for validation and the other 15% will be used for testing.

To ensure that there is no leakage of information, the dataset is split at the speaker level so that videos from the same speaker are contained in only one partition. This speaker-disjoint splitting method guarantees that the model is tested on previously unexamined speakers and gives an accurate assessment of how generalizable the model is.

Along with the predetermined split of the dataset, a cross-validation method of five folds is employed. For this method, speakers are assigned to the five folds in a way that maintains speaker exclusivity. The model is trained and

tested a total of five times, with each fold used as the testing fold at least once. The performance is evaluated as a mean and standard deviation from the five trials.

Training Configuration

The deep learning optimization procedure works the best with the Adam optimizing algorithm and so this was the one used to train the models. Early stopping was employed to prevent the model from overfitting, and this was done by stopping once the validation AUC was reached.

Dropout regularization with a probability of 0.5 is used in the classification layers to improve model generalization. Each training run is performed for a maximum of 50 epochs, and the model parameters corresponding to the highest validation AUC are selected for final evaluation.

5.3 Comparative Performance Analysis

To evaluate the effectiveness of the proposed multimodal architecture, three categories of models are examined: unimodal models, baseline multimodal fusion models, and the proposed cross-attention-based framework.

Unimodal models rely on a single modality. In the audio-only configuration, acoustic features are processed using the audio encoder followed by the classification layer. In the visual-only configuration, only visual features are used for prediction.

Baseline multimodal models combine the two modalities using conventional fusion strategies. In early fusion, acoustic and visual features are concatenated at the feature level before being processed by a shared encoder. In late fusion, separate unimodal models generate predictions that are later combined through averaging or logistic regression.

The proposed model integrates the two modalities through a bidirectional cross-attention mechanism that dynamically models interactions between speech characteristics and facial expressions. The comparative performance of these models is summarized in Table 2.

Table 2. Comparative Performance on Test Set (mean $\pm$ std over 5 folds)

Model	Accuracy	Precision	Recall	F1-score	AUC
Audio-only	0.72 $\pm$ 0.03	0.70 $\pm$ 0.04	0.68 $\pm$ 0.05	0.69 $\pm$ 0.04	0.78 $\pm$ 0.03
Visual-only	0.70 $\pm$ 0.02	0.68 $\pm$ 0.03	0.66 $\pm$ 0.04	0.67 $\pm$ 0.03	0.75 $\pm$ 0.02
Early fusion	0.76 $\pm$ 0.03	0.74 $\pm$ 0.04	0.73 $\pm$ 0.03	0.73 $\pm$ 0.03	0.82 $\pm$ 0.02
Late fusion	0.78 $\pm$ 0.02	0.76 $\pm$ 0.03	0.75 $\pm$ 0.04	0.75 $\pm$ 0.03	0.84 $\pm$ 0.02
Cross-attention	0.82 $\pm$ 0.02	0.80 $\pm$ 0.03	0.79 $\pm$ 0.03	0.79 $\pm$ 0.03	0.89 $\pm$ 0.01

The results demonstrate that multimodal models consistently outperform unimodal baselines. While audio-only and visual-only models capture meaningful behavioral cues, integrating both modalities significantly improves classification performance.

Among the multimodal approaches, the proposed cross-attention architecture achieves the highest performance across all metrics. The model improves AUC by approximately 5–7 percentage points compared with late fusion and nearly 12 percentage points compared with unimodal models. These improvements indicate that explicitly modeling interactions between speech and facial behavior enables the model to capture richer multimodal representations.

5.4 Comparison with Existing Studies

To further evaluate the effectiveness of the proposed approach, the performance of the model is compared with representative methods reported in previous depression detection studies.

Table 3. Comparison with Existing Depression Detection Methods

Study	Dataset	Method	Accuracy	F1-score	AUC
Study A	DAIC-WOZ	CNN-based model	0.74	0.71	0.80
Study B	AVEC	LSTM-based model	0.76	0.73	0.83
Study C	D-Vlog	Transformer fusion	0.79	0.76	0.86
Proposed Method	D-Vlog	Cross-attention multimodal model	0.82	0.79	0.89

The results indicate that the proposed cross-attention architecture achieves superior performance compared with previously reported models. The improvement can be attributed to the ability of the attention mechanism to dynamically capture interactions between speech patterns and facial expressions rather than relying on static feature concatenation. This allows the model to learn more informative multimodal representations and improves overall classification performance.

5.5 Ablation Study

To analyze the contribution of different components of the proposed architecture, several ablation experiments are conducted.

First, the cross-attention module is replaced with simple feature concatenation. This modification reduces AUC by approximately 3–4 percentage points, indicating that attention-based fusion plays an important role in modeling cross-modal relationships.

Second, the bidirectional attention mechanism is replaced with a single-direction attention configuration. In this case, performance decreases by approximately two percentage points, suggesting that bidirectional interaction provides richer multimodal representations.

Third, the temporal segmentation step is removed, and the model is trained on entire feature sequences. Although this configuration achieves moderate performance (AUC around 0.85), it exhibits reduced stability when handling recordings of varying duration. The results of these experiments are illustrated in Figure 2.

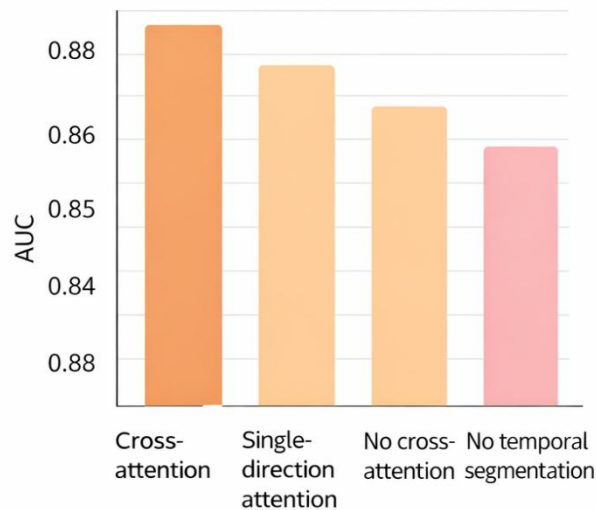


Figure 2. AUC comparison for ablation variants of the proposed architecture.

5.6 Visualization of Model Performance

To provide a clearer comparison of model performance across different evaluation metrics, the results are visualized using performance graphs.

Figure 3 presents a comparison of accuracy, precision, recall, and F1-score across unimodal models, baseline multimodal models, and the proposed cross-attention framework.

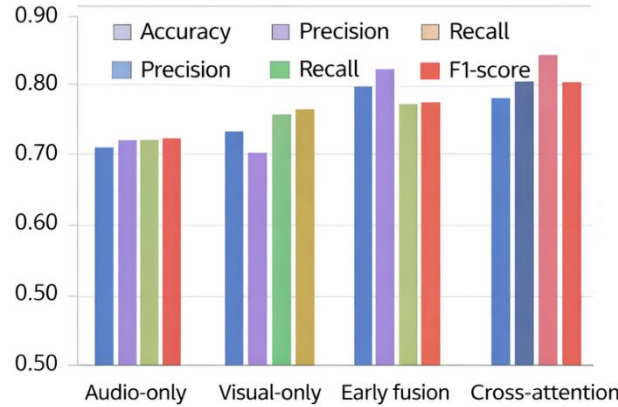


Figure 3. Performance comparison across different models and evaluation metrics.

The graph illustrates that the proposed cross-attention model consistently achieves higher performance across all evaluation metrics. The improvement is particularly evident in F1-score and AUC, indicating that the model effectively balances precision and recall while maintaining strong classification separability.

5.7 Robustness and Generalization Analysis

To further assess the reliability of the proposed framework, additional experiments are conducted to evaluate robustness under different conditions.

First, subgroup analysis based on inferred gender is performed. The model demonstrates consistent performance across male and female speakers, with AUC differences remaining within one to two percentage points.

Second, cross-corpus evaluation is performed by applying the model trained on the D-Vlog dataset to samples from the DAIC-WOZ dataset. Although the performance decreases due to domain differences, the model still achieves an AUC of approximately 0.78. After fine-tuning on a small subset of DAIC-WOZ samples, the AUC improves to approximately 0.84.

Finally, noise robustness experiments are conducted by introducing artificial degradation to the audio and visual inputs. Visual frames are down sampled and blurred, while white noise is added to the audio signal. The cross-attention model shows only moderate performance degradation (approximately 5%), whereas unimodal models experience larger declines. These results indicate that multimodal fusion improves resilience when one modality becomes unreliable.

6. RESULTS

6.1 Quantitative Results

The experimental evaluation demonstrates that the proposed multimodal cross-attention model consistently outperforms all baseline configurations across the evaluation metrics. Table 4 summarizes the mean performance obtained over five cross-validation folds.

Table 4. Quantitative Performance Metrics (Mean $\pm$ Std over 5 Folds)

Model	Accuracy	Precision	Recall	F1-score	AUC
Audio-only	0.72 ± 0.03	0.70 ± 0.04	0.68 ± 0.05	0.69 ± 0.04	0.78 ± 0.03
Visual-only	0.70 ± 0.02	0.68 ± 0.03	0.66 ± 0.04	0.67 ± 0.03	0.75 ± 0.02
Early Fusion	0.76 ± 0.03	0.74 ± 0.04	0.73 ± 0.03	0.73 ± 0.03	0.82 ± 0.02
Late Fusion	0.78 ± 0.02	0.76 ± 0.03	0.75 ± 0.04	0.75 ± 0.03	0.84 ± 0.02
Cross-Attention	0.82 ± 0.02	0.80 ± 0.03	0.79 ± 0.03	0.79 ± 0.03	0.89 ± 0.01

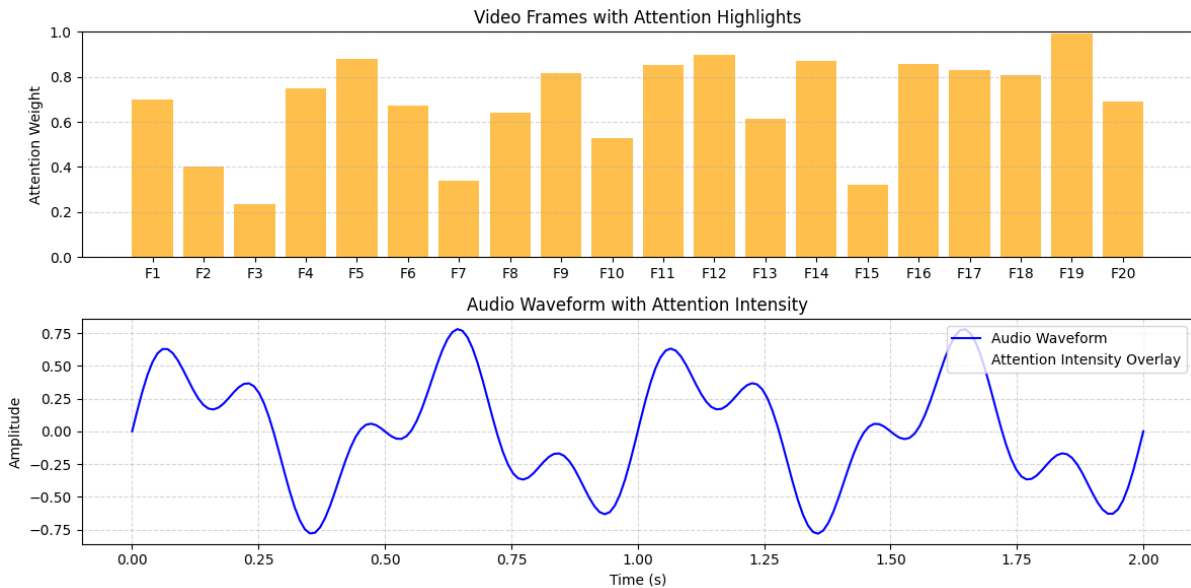
The results show that multimodal models outperform unimodal approaches, confirming that speech and facial cues provide complementary information for depression detection. Among the multimodal methods, the proposed cross-attention architecture achieves the best performance across all metrics. In particular, the model improves AUC by approximately 5–6 percentage points compared with late fusion and nearly 12 percentage points compared with unimodal models.

The performance comparison across different models and evaluation metrics is illustrated in Figure 3. The cross-attention model consistently achieves higher accuracy, recall, and F1-score, indicating improved classification reliability.

6.2 Qualitative Analysis

To better understand the behavior of the model, attention maps generated by the cross-attention module were examined. The analysis indicates that the model dynamically adjusts its focus depending on the available behavioral cues.

For example, in recordings where facial expressions were subtle but speech patterns exhibited longer pauses and monotonic tone, the model assigned greater attention to acoustic features. Conversely, when facial movements were more expressive, the model emphasized visual features. These observations suggest that the cross-attention mechanism effectively captures complementary multimodal signals. Example attention visualizations are presented in Figure 4.

**Figure 4.** Visualization of attention weights highlighting alignment between audio segments and facial regions.

6.3 Discussion

The improved performance of the proposed framework can be attributed to the ability of the cross-attention mechanism to model interactions between modalities. Unlike early or late fusion methods that combine modalities in a fixed manner, cross-attention dynamically adjusts the contribution of each modality based on the strength of the observed signals.

Despite the strong performance, certain limitations remain. In some cases, degraded visual quality or emotionally intense narratives without clinical depression can lead to misclassification. Addressing these challenges may require incorporating additional contextual information, such as textual sentiment analysis or more diverse training data.

Overall, the results demonstrate that multimodal cross-attention provides a more effective strategy for detecting depression-related behavioral patterns in real-world video recordings.

7. CONCLUSION

This study explored the potential of multimodal learning in identifying depression indicators within natural vlogs. It was found that the model that covertly fused audio and visual behavioural cues using cross-attention achieved better results than unimodal systems and traditional fusion methods. This demonstrates that automated systems for depression detection in videos can benefit from the targeted robust detection of the interactions in the patterns of speech and concomitant facial expressions. Additionally, the experiments demonstrated the potential of the D-Vlog dataset for the study of behavioural depression indicators within unstructured observation of the participants. Cross-attention fusion was shown to be superior to other integration methods in providing richer multimodal representation, and this was evident in the comparative experiments and the ablation study. However, several shortcomings were noted. There is an inherent subjectivity in the annotation of the dataset and the features provided an inherent limitation to the capturing of additional behavioural cues. Future investigations should include the transcribed texts of the speech in the dataset, the behavioural cues of participants from diverse demographics and cultures, and should expand cross-cultural inclusivity. Addressing these aspects will be important for developing reliable and ethically responsible tools for real-world mental health screening.

REFERENCES

1. J. Yoon, C. Kang, S. Kim, and J. Han, "D-Vlog: Multimodal Vlog Dataset for Depression Detection," *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12462–12470, 2022.
2. K. Min, J. Yoon, M. Kang, D. Lee, E. Park, and J. Han, "Detecting depression on video logs using audiovisual features," *Humanities and Social Sciences Communications*, vol. 10, Article 788, 2023.
3. L. He, Z. Li, P. Tiwari, F. Zhu, and D. Wu, "LSCAformer: Long- and short-term cross-attention-aware transformer for depression recognition from video sequences," *Biomedical Signal Processing and Control*, vol. 98, p. 106767, 2024.
4. I. Ario, U. Smrke, N. Plohl, and I. Mlakar, "Scoping review on the multimodal classification of depression and experimental study on existing multimodal models," *Diagnostics*, vol. 12, no. 11, p. 2683, 2022.
5. M. M. Han et al., "Automatic recognition of depression based on audio and video: A review," *World Journal of Psychiatry*, vol. 14, no. 2, pp. 225–233, 2024.
6. L. Wang et al., "AI-assisted multimodal information for the screening of depression: A systematic review and meta-analysis," *npj Digital Medicine*, vol. 8, p. 523, 2025.
7. M. Nasir, A. Jati, P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 43–50.
8. M. Muzammel, H. Salam, and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106433, 2021.
9. Z. Xu et al., "Depression detection methods based on multimodal fusion of voice and text," *Scientific Reports*, vol. 15, p. 21907, 2025.
10. K. Daly and O. Olukoya, "Depression detection in read and spontaneous speech: A multimodal approach for lesser-resourced languages," *Biomedical Signal Processing and Control*, vol. 108, p. 107959, 2025.
11. A. Ksibi et al., "Electroencephalography-based depression detection using multiple machine learning techniques," *Diagnostics*, vol. 13, no. 10, p. 1779, 2023.

12. J. Gratch et al., "The Distress Analysis Interview Corpus of human and computer interviews," in *Proc. Language Resources and Evaluation Conference (LREC)*, 2014.
13. F. Ringeval et al., "AVEC 2019 workshop and challenge: Detecting depression with AI," in *Proc. Audio/Visual Emotion Challenge and Workshop (AVEC)*, 2019.
14. H. Cai et al., "A multimodal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, p. 178, 2022.
15. P. Cao et al., "A multimodal depression consultation dataset of speech and text with HAMD-17 assessments," *Scientific Data*, vol. 12, 2025.
16. H. Lian et al., "A survey of deep learning-based multimodal emotion recognition: speech, text, and face," *Entropy*, vol. 25, p. 1440, 2023.
17. Y. Guo, C. Zhu, S. Hao, and R. Hong, "A topic-attentive transformer-based model for multimodal depression detection," *IEEE Access*, vol. 10, pp. 73615–73627, 2022.
18. S. Yang et al., "Cross-modal contrastive learning for multimodal sentiment recognition," *Applied Intelligence*, vol. 54, pp. 4260–4276, 2024.
19. Z. Lifelo et al., "Uncertainty-aware domain incremental learning for cross-domain depression detection," *Scientific Reports*, vol. 15, p. 25344, 2025.
20. C. Lau, X. Zhu, and W.-Y. Chan, "Automatic depression severity assessment with deep learning using parameter-efficient tuning," *Frontiers in Psychiatry*, vol. 14, 2023.
21. S. A. Qureshi, M. Hasanuzzaman, S. Saha, and G. Dias, "Combining acoustic, text and visual features for depression estimation," in *Proc. ACM Multimedia Workshop*, 2019.
22. S. Li, Y. Yang, Y. Li, and J. Chen, "Multimodal depression detection using cross-attention fusion," *IEEE Access*, vol. 12, pp. 115223–115235, 2024.
23. H. Sun et al., "Multimodal sentiment analysis with disentangled representation learning," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 1606–1617, 2025.
24. X. Wang, X. Hu, Y. Li, and C. Jiang, "Multimodal human behavior recognition using deep neural networks," *Pattern Recognition Letters*, vol. 153, pp. 36–43, 2022.
25. J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.
26. A. Vaswani et al., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems*, 2017.
27. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
28. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
29. L. Christ et al., "The MuSe multimodal sentiment analysis challenge," in *Proc. ACM Multimedia Workshop*, 2023.
30. J. Tanner et al., "Approaches to research ethics in health research on YouTube," *Journal of Medical Internet Research*, vol. 25, e43060, 2023.
31. Y. Zhang et al., "Reporting of ethical considerations in social-media-based health research," *Journal of Medical Internet Research*, vol. 26, e51496, 2024.
32. M. Rahsepar-Meadi et al., "Ethical challenges of conversational AI in mental health care," *JMIR Mental Health*, vol. 12, e60432, 2025.
33. M. Wang et al., "Depression detection from speech signals using channel attention mechanisms," *Biomedical Signal Processing and Control*, vol. 113, p. 108873, 2025.
34. T. He and W. Huang, "Automatic identification of depressive symptoms using convolutional neural networks," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, 2024.
35. R. Sivanaiah, "Multimodal analysis of depression and PTSD using AI methods," *International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 2360–2368, 2024.
36. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
37. J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. International Conference on Machine Learning*, 2006.
38. G. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020.