Project Title: Multimodal Depression Detection through Mutual Transformer

Group Number: 8

Group Members

Surname, First Name	Student ID	Department
Haque, Md Rezwanul	21108752	ECE
Lu, Daniel	21113256	SYDE

Main Contributions

1. Deep Learning Framework: Developed a transformer-based framework for depression detection using multimodal data.
2. Acoustic Feature Extraction Module: Introduced a global self-attention mechanism to model content and positional relationships in acoustic features.
3. Video Feature Extraction Module: Designed to capture local and global patterns using patch embeddings and hierarchical attention.
4. Mutual Transformer: Calculated interdependencies between audio and visual embeddings for a joint multimodal representation.

Tools and Technologies Used

We have used the following programming languages, platforms, and tools (e.g., Python, PyTorch, TensorFlow, etc.):

- **Programming Language:** Python
- **Deep Learning Framework:** PyTorch
- **Code Editor:** Visual Studio Code
- **Operating System:** Ubuntu 24.04 LTS

Project Implementation Details

Check the boxes that best describe your project:

☐ We used an existing implementation of the algorithm(s).

☑ We implemented the algorithm(s) ourselves.

☑ We used an existing dataset.

☐ We created a new dataset.

Abstract

Depression is a major mental health condition that severely affects the emotional and physical well-being of individuals. The simple nature of data collection from social media platforms has attracted significant interest in properly utilizing this information for mental health research. In this paper, we present a multimodal depression detection network based on acoustic and visual data generated from social media networks. We propose to exploit mutual transformers to efficiently extract and fuse multimodal features for efficient depression detection. The proposed system consists of four core modules: acoustic feature extraction module to retrieve relevant acoustic attributes, visual feature extraction module to extract significant high-level patterns, mutual transformer to compute the correlations among the generated features, and multi-fusion detection module to fuse these features from multiple modalities and detect the depression. The extensive experiments are performed using the multimodal D-Vlog dataset, and the findings reveal that the developed multimodal depression detection network surpasses the state-of-the-art, with recall of 0.8065 and F1 score of 0.7707.

1 Introduction

Depression is a mental health condition that affects person's mood, thoughts, and behavior. It often causes feelings of sadness, hopelessness, and a lack of interest in daily activities [1]. Currently, depression is primarily diagnosed through questionnaire surveys and professional assessments, but these methods can be affected by participant cooperation and clinician expertise [2]. Early detection and treatment are crucial, yet traditional methods are often time-consuming and error-prone. Therefore, developing an automated and intelligent system is essential for faster, more accurate, and accessible depression detection.

Most previous studies often concentrate on single-modality text information or treat each modality with equal importance when developing fusion methods. In contrast, using multi-modal data provides more complete and complementary information for diagnosing depression. For example, Seneviratne et al. [3] introduced a multi-modal depression detection model by combining audio and text data, addressing overfitting in session-level classifiers and utilizing segment-level classifiers. Yuan et al. [4] proposed a multimodal multiorder factor fusion approach to take advantage of high-order interactions between different modalities. Shen et al. [5] presented an automated depression detection model using bidirectional long-short-term memory (BiLSTM) and fused audio and text data with attention mechanisms. Additionally, Zhou et al. [6] designed the Time-Aware Attention Multi-modal Fusion Network to handle Vlog data by overcoming issues with unimodal representations and fusion processes.

To address the limitations of existing datasets and models for depression detection, we propose a robust multimodal depression detection framework based on mutual transformer networks that efficiently capture and fuse acoustic and visual features. Our approach utilizes mutual transformers to compute correlations across modalities, providing a better representation of depression-related behavioral patterns using the D-Vlog dataset collected from social media platforms. Our contributions are as follows:

- We develop a deep learning framework through transformer architecture for detecting depression from multimodal data.
- We propose acoustic feature extraction module by integrating a global self-attention mechanism to design content-based and positional relationships within the feature space.

- We introduce video extraction module is designed to extract both local and global patterns through patch embeddings and hierarchical attention mechanisms.
- We design the mutual transformer to calculate interdependencies between audio and visual embeddings to capture multimodal information among the modalities through joint representation.
- We exploit the D-Vlog dataset to achieve state-of-the-art performance, demonstrating the proposed model's effectiveness in distinguishing between depressed and non-depressed individuals based on non-verbal behavior.

The rest of the article is structured as follows. Section 2 provides an overview of previous related studies, including advancements in depression detection through unimodal and multimodal approaches. Section 3 provides the details of the proposed mutual transformer-based architecture for multimodal depression detection, highlighting its innovative fusion strategies. Section 4 discusses results and performance comparisons. Finally, Section 5 concludes the article with a summary of findings and directions for future research.

2 Related Work

The detection of depression has been an active research area in the field of psychology, artificial intelligence (AI), and healthcare. Depression diagnosis traditionally relied on clinician assessments and questionnaires[7]. However, machine learning (ML) and multimodal data analysis have played a important role in enhancing detection accuracy and scalability. Multimodal methods combine various data formats, including text, audio, visual and physiology signals, to capture the complexity of depression symptoms[8]. Text-based methods use nature language processing (NLP) to analyze linguistic patterns indicative of depression. By leveraging the vast amount of digital information on social media posts, electronic health records, this technology can detect subtle linguistic patterns of depressive symptoms[9]. Audio-based depression detection analyzes speech characteristics, including pitch, rate and intensity. Audio data have shown to be an effective modality for detecting depression[10]. Miao et al. designed a fusion feature combining high-order spectral feature and traditional speech features to classify deprssion and this model reached 0.85 accuracy[11]. While visual signals, such as facial expressions and body movements, also offer valuable insights into depression. Lang He et al.[12] proposed a pattern recognition based method to effectively capture facial dynamic expressions as non-verbal measurements for assessing the severity of depression. Multimodal approaches have shown promising performance compared to unimodal methods. Despite these advances, issues of data privacy, interpretability, and generalization across populations remain significant challenges for this field.

3 Methodology

The proposed model is designed to solve a binary classification problem, where the goal is to classify a vlog as either depression or not. Let, a set of vlogs $G = \{G_n\}_{n=1}^{|G|}$, and each vlog g_n can be represented as like in the equation (1).

$$g_n = \left(X_A^n \in \mathbb{R}^{t \times d_a}, X_V^n \in \mathbb{R}^{t \times d_v}\right) \tag{1}$$

Here, X_A^n refers to the acoustic features, X_V^n refers to the visual features, and d_a, d_v, and t are the dimensions of the acoustic features, visual features, and the sequence length, respectively.

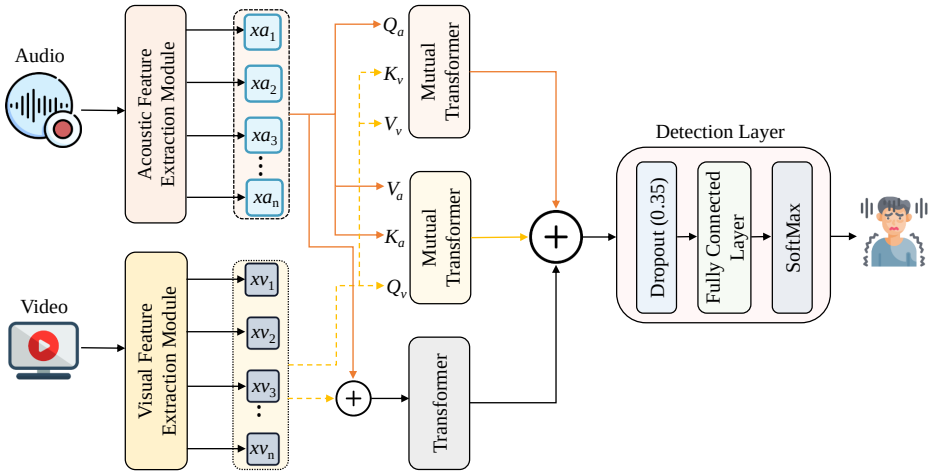


Figure 1: The overall architecture of the proposed model: A mult-imodal fusion approach for audio-visual feature integration using mutual transformers and detection Layer.

Both the acoustic and visual sequences are aligned to have the same length. The objective is to classify each vlog in G as either "Depression" or "Normal" by learning important patterns from the vlog's acoustic and visual features. The overall architecture is plotted in Fig. 1.

3.1 Acoustic Feature Extraction Module

In the Acoustic Feature Extraction Module (AFEM), we use a combination of a global self-attention network [13] and a bottleneck layer [14] as the backbone for extracting relevant features from acoustic data. The output feature map is obtained by aggregating information from the input map using both content-based and positional attention layers. We define the matrices of keys, queries, and values as $K_a = [k_{ij}] \in \mathbb{R}^{t \times d_a}$, $Q_a = [q_{ij}] \in \mathbb{R}^{t \times d_a}$, and $V_a = [v_{ij}] \in \mathbb{R}^{t \times d_o}$. The new feature map $X_A^c \in \mathbb{R}^{t \times d_a}$ is then computed using the following content-based dot-product attention [15] equation (2).

$$X_A^c = Q_a \cdot \rho(K_a^\top) \cdot V_a \tag{2}$$

where ρ represents softmax normalization. Normalizing the queries constrains the output features to be convex combinations of the global context vectors. As these constraints could restrict the expressive power of the attention mechanism, we remove the softmax normalization on queries. This allows the output features to span the entire subspace of the context vectors d_a.

The content attention layer doesn't consider spatial positions, so it remains unchanged if elements are shuffled. Following the work of [16], [17], we tackle this issue using a positional attention layer that accounts for both content and spatial relationships between neighboring elements. Inspired by axial attention formulations [18], this layer processes the feature space in specific directions, ensuring each element attends to its neighbors. This approach effectively spreads information across the entire $N \times N$ neighborhood. Let $\Delta = \left\{-\frac{N-1}{2}, \ldots, 0, \ldots, \frac{N-1}{2}\right\}$ be a set of N offsets, and $R = [r_\delta] \in \mathbb{R}^{N \times d_a}$ denote the matrix of learnable relative position embeddings corresponding to N spatial offsets $\delta \in \Delta$. Let $V_{a(xy)} = [v_{x+\delta,y}] \in \mathbb{R}^{N \times d_o}$ be the matrix consisting of the values at the N neighboring elements of position (x, y). Let $X_{A(xy)}^p$ denote the output of the positional attention mechanism at

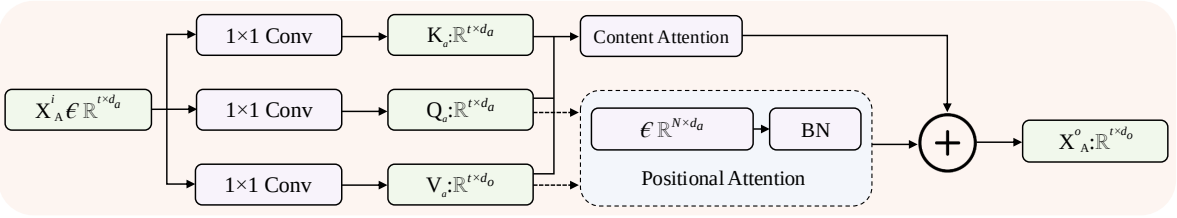


Figure 2: Acoustic feature extraction module: the module processes input audio features using 1×1 convolutions, applies both content and positional attention mechanisms, and performs batch normalization before aggregating the results for the final output.

position (x, y). Then, our attention mechanism, which uses the relative position embeddings R as keys, can be described as like equation (3).

$$X^p_{A(xy)} = Q_{a(xy)} \cdot R^\top \cdot V_{a(xy)} \tag{3}$$

where $Q_{a(xy)}$ is the query at position (x, y). The design audio feature extraction module is represented in the Fig. 2.

The final output feature map, which is a combination of these two attention mechanisms, can be written as like equation (4).

$$X^o_A = X^c_A + f_{bn}(X^p_A) \tag{4}$$

Here, f_{bn} is the 1×1 convolution with batch normalization (BN) layer. The final feature map X^o_A is the sum of these two components.

We utilize a bottleneck layer to perform binary classification by efficiently extracting features from the combined outputs X^o_A of the content-based and positional attention mechanisms. The operation of the bottleneck layer is represented by equation (16).

$$\hat{y}_a = \beta_o \left(\frac{1}{t} \sum_{i=1}^{t} (f_{bn,i}(f_{cn,i}(X^o_A))) + f_{bn}(f_{cn}(X^o_A)) \right) \tag{5}$$

Where, f_{cn} refers to the 1×1 convolutional layers, β_o is the weight matrix used for final classification, and t is the sequence length, where $t = 3$.

3.2 Visual Feature Extraction Module

For visual feature extraction module (VFEM), the network uses a patch embedding layer followed by a transformer block. Assume the visual input feature map is $X^i_V \in \mathbb{R}^{t \times d_v}$. This input is processed through a 1×1 convolution layer, followed by group normalization, an activation function such as SiLU, and another 1×1 convolution. This acts as the patch embedding mechanism, projecting the input data (X^i_V) into a suitable dimension for the transformer block.

Once the embedding is complete, the data is passed through transformer blocks. In the transformer block, query (Q_v), key (K_v), and value (V_v) are calculated using the equation (6).

$$Q_v = X^i_V \beta^q, \quad K_v = X^i_V \beta^k, \quad V_v = X^i_V \beta^v \tag{6}$$

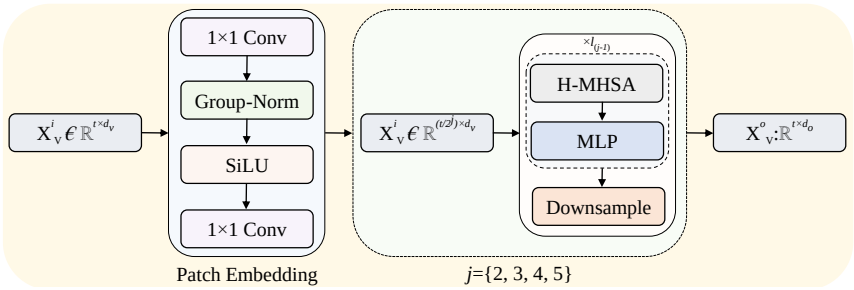


Figure 3: Visual feature extraction module: the module applies 1×1 convolutions, group normalization, and SiLU activation to extract patch embeddings from visual data. These embeddings are passed through a series of hierarchical multi-head self-attention layers (H-MHSA), a multi-layer perceptron (MLP), and a downsampling operation to generate refined visual features as output.

where $\beta^q, \beta^k, \beta^v \in \mathbb{R}^{d_v \times d_v}$ are the trainable weight matrices of linear transformations. Assuming the input and output have the same dimensions, the Hierarchical Multi-Head Self-Attention (H-MHSA) [19] can be formulated as like equation (7).

$$M_V = \rho\left(\frac{Q_v K_v^\top}{\sqrt{d}}\right) V_v \tag{7}$$

where $\sqrt{d}$ means an approximate normalization, and the Softmax normalization function (ρ) is applied to the rows of the feature matrix.

After computing MHSA, a residual connection is added to improve optimization, as shown in the equation (8).

$$M'_V = M_V \beta^p + X_V^i \tag{8}$$

where $\beta^p \in \mathbb{R}^{d_v \times d_v}$ is a trainable weight matrix for feature projection. Next, a multi-layer perceptron (MLP) is applied to enhance the representation, which is done by the equation (9).

$$X_V^o = f_{\text{mlp}}(M'_V) + M'_V \tag{9}$$

where X_V^o represents the output of the transformer block. The transformer block is repeated $l_{(j-1)}$ times where $j = \{2, 3, 4, 5\}$, and f_{mlp} is the MLP layer function. The design video feature extraction module is represented in th Fig. 3.

3.3 Mutual Transformer

The proposed model computes multimodal mutual correlations and speaker-related depression features with the MT module. From the acoustic feature extraction module, we get the acoustic feature X_A^o (as seen in equation (4)), which is an audio embedding represented as $X_A^o = \{x_{a1}, x_{a2}, \ldots, x_{an}\}$, where $n = 256$ is the sequence length and $d_a = 71$ is the size of each embedding. In a similar way, the visual feature extraction module provides the visual feature X_V^o (as shown in equation (9)), which is a video embedding represented as $X_V^o = \{x_{v1}, x_{v2}, \ldots, x_{vn}\}$, where $n = 256$ is the sequence length and $d_v = 139$ is the size of each embedding.

There are two mutual transformer layers to calculate the mutual correlations (MC), $MC_{\text{AV}} \in \mathbb{R}^{n \times d}$ and $MC_{\text{VA}} \in \mathbb{R}^{n \times d}$ from $X_A^o \in \mathbb{R}^{n \times d_a}$ and $X_V^o \in \mathbb{R}^{n \times d_v}$, and there is another transformer layer by fusing both $X_A^o \in \mathbb{R}^{n \times d_a}$ and $X_V^o \in \mathbb{R}^{n \times d_v}$ to calculate the mutual

correlations $MC_{f_{\text{AV}}} \in \mathbb{R}^{(2n)\times(d_a+d_v)}$. The mutual transformers receives query (Q), key (K), and value (V) from different modalities [20]. In particular, the MC_{AV} mutual transformer receives the audio embedding X_A^o as Q_a and video embedding X_V^o as K_v and V_v. Similarly, MC_{AV} receives X_V^o as Q_v and X_A^o as K_a and V_a. Meanwhile, a particular transformer is applied to model the correlations $MC_{f_{\text{AV}}} \in \mathbb{R}^{(2n)\times(d_a+d_v)}$ of the concatenation of X_A^o and X_V^o.

The corresponding formulations are presented in equations (10), (11), and (12).

$$\begin{aligned} f_{\text{attn}(X_A^o, X_V^o)} &= \rho\left(\frac{Q_a K_v^\top}{\sqrt{d_{K_v}}}\right) V_v, \\ f_{\text{attn}(X_V^o, X_A^o)} &= \rho\left(\frac{Q_v K_a^\top}{\sqrt{d_{K_a}}}\right) V_a. \end{aligned} \tag{10}$$

$$\begin{aligned} MC_{\text{AV}} &= L\left(X_A^o + \sigma\left(\beta^\top\left(L(X_A^o + f_{\text{attn}(X_A^o, X_V^o)})\right)\right) + b\right), \\ MC_{\text{VA}} &= L\left(X_V^o + \sigma\left(\beta^\top\left(L(X_V^o + f_{\text{attn}(X_V^o, X_A^o)})\right)\right) + b\right). \end{aligned} \tag{11}$$

$$MC_{f_{\text{AV}}} = f_{\text{transformer}}(f_{\text{concat}}(X_A^o, X_V^o)) \tag{12}$$

where, L denotes the LayerNorm [21]. f_{attn} is the abbreviation of the self-attention function. MC_{AV}, MC_{VA}, and $MC_{f_{\text{AV}}}$ denote the output of mutual transformer. $f_{\text{transformer}}$ and f_{concat} are the transformer and concatenation layer function of X_A^o and X_V^o. β and b are learnable parameters.

3.4 Mult-Fusion Depression Detection Layer

In the Mult-Fusion Depression Detection Layer (MFDDL), the multimodal transformer representation Z can be obtained using mutual transformer fusion strategy. The general form of Z is shown in equation (13).

$$Z = f_{\text{concat}}\left(f_{\text{AvgPool}}\left(MC_{\text{AV}}, MC_{\text{VA}}, MC_{f_{\text{AV}}}\right)\right) \tag{13}$$

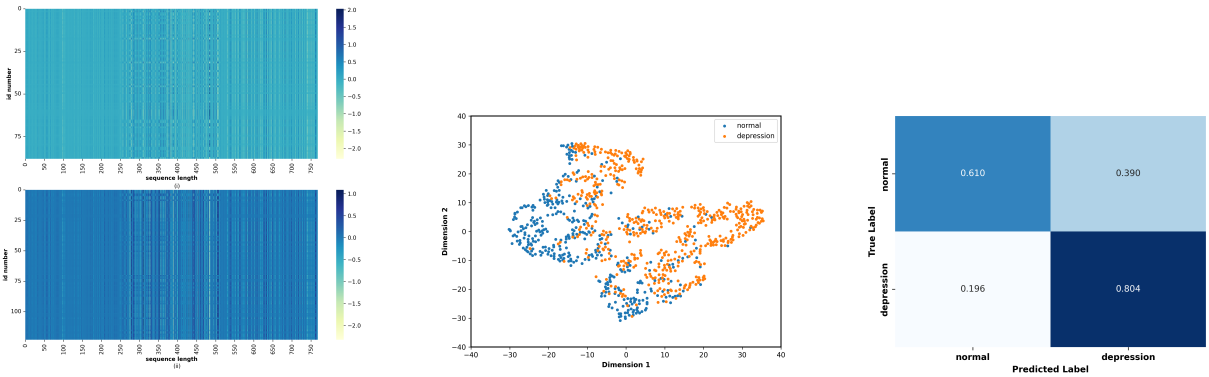
where $Z \in \mathbb{R}^{(4n)\times 2(d_a+d_v)}$. The final logits for multimodal depression detection can then be derived as shown in equation (14).

$$\begin{aligned} \alpha &= F(\beta^\top \cdot Z), \\ P &= \rho\left(\beta^\top(\alpha^\top \cdot Z) + b\right) \end{aligned} \tag{14}$$

Here, $F(\cdot)$, α, and ρ denote the fully connected layer, attention scores, and softmax activation function, respectively. β and b are learnable parameters, and $P \in \mathbb{R}^2$ is the depression probability, classifying into either "Depression" or "Non-depression." The right side in the Fig. 1 is represent Depression Detection layer.

For depression detection, a customized loss function is utilized by combining various types of losses to address the challenges of noisy and imbalanced datasets [22]. This loss function combines the Smoothed Binary Cross-Entropy (BCE) with Logits, Focal Loss, and L2 Regularization, and can be written as in equation (15) (See the Appendix for detailed explanation).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{L2}} \tag{15}$$



(a) (b) (c)

Figure 4: Visualization of (a) CAM weight based on attention weight of the final embedding obtained by fusing encoding audio-visual features: (i) people with normal (ii) people with depression, (b) t-SNE feature of the Proposed Model (MFDDL) for multimodal depression detection, and (c) Confusion matrix of the Proposed Model (MFDDL).

4 Experiments and Results

The D-Vlog dataset [23] includes 961 YouTube vlog videos from 816 individuals, labeled as depressed (555) or non-depressed (406) based on keywords in the video titles. It provides 25 acoustic features from OpenSmile [24] and 68 facial landmarks from Dlib [25], with the data split into training, validation, and test sets in a 7:1:2 ratio. The model was trained for 200 epochs with the Adam optimizer, a cosine annealing scheduler, and early stopping to avoid overfitting. Metrics such as accuracy, precision, recall, and F1 score were evaluated through 10-fold cross-validation.

4.1 Visual Results

In this part, we present CAM attention weights, t-SNE feature visualization, and the confusion matrix to show the results clearly.

Fig. 4a shows the attention weights for both groups based on the fused acoustic-visual features. We visualize the CAM attention weights on the final embeddings, which combine acoustic and visual features. Using the trained model, we process the test set data, visualize attention weights, and analyze differences in sequence length between depressed and non-depressed individuals. The horizontal axis represents sequence length, and the vertical axis represents subject numbers. In the figure, darker blue colors indicate higher attention weights for the depressed category, while lighter blue colors indicate lower attention for the non-depressed category.

Fig. 4b shows the distribution of fused acoustic-visual features from the D-Vlog dataset using t-SNE visualization. Some normal samples overlap with the depressed category, leading to minor misclassifications. These misclassifications occur because the model struggles to distinguish these categories but can still identify the clusters.

Fig. 4c displays the confusion matrix for the normal and depressed categories in the dataset. Overall, the proposed model (MFDDL) outperforms state-of-the-art methods in predicting depression classes on the D-Vlog dataset.

4.2 Results Comparison with the State-of-the-Art Models

This subsection will compare the performances of our proposed model with other methods as shown in Table 1.

Our proposed model achieved higher metrics, with a precision of 0.7392, recall of 0.8065, and F1 score of 0.7707, showing strong performance in detecting depression or healthy group. This is a significant improvement over single-feature modules (like acoustic or visual alone) in Table 2. This result highlights the value of using mutual transformers to fuse features before passing them to the detection layer. This unique fusion approach in our model improves accuracy and interpretability, while reducing overfitting, making it more reliable in distinguishing between normal and depression cases across different data types.

Table 1: Performances Comparison in Different Methods.

Method	**Precision**	**Recall**	**F1-Score**
BLSTM [26]	0.6081	0.6179	0.5970
TFN [27]	0.6139	0.6226	0.6100
Depression Detector [23]	0.6540	0.6557	0.6350
TAMFN [6]	0.6602	0.6650	0.6582
CAIINET [28]	0.6656	0.6698	0.6655
STST [29]	0.7250	0.7767	0.7500
MDAVIF [30]	**0.7425**	0.7600	0.7525
Proposed Model	0.7392	**0.8065**	**0.7707**

Table 2: Performances of Proposed Model.

Method	**Feature**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
AFEM	Acoustic	0.4770	0.5552	0.4667	0.4106
VFEM	Visual	0.4604	0.5551	0.4577	0.4239
MFDDL	Both	0.7260	0.7392	0.8065	0.7707

5 Discussion and Conclusion

In this paper, we introduced a multimodal depression detection system based on mutual transformers. The mutual transformer is designed to fuse the extracted features from the acoustic and visual modules for consolidating the correlation among multiple modalities. We performed experiments on the latest social medial based multimodal D-Vlog dataset, and the proposed network obtained comparatively better performances compared to the existing methods, exhibiting the effectiveness of multimodal fusion in mental health assessment. However, depression datasets often include limited amounts of data and imbalanced data distributions, posing difficulties for real-world applications. Given that the D-Vlog dataset is subjectively labeled by humans, mislabeling is inevitable, increasing the chance of model overfitting and limiting the model's ability to learn robust features. In the future, we will develop more robust and reliable multimodal depression detection architectures across various datasets by minimizing the impact of mislabeled samples, ensuring broader applicability.

References

[1] R. H. Belmaker and G. Agam, "Major depressive disorder," *New England Journal of Medicine*, vol. 358, no. 1, pp. 55–68, 2008.

[2] U. Yadav, A. K. Sharma, and D. Patil, "Review of automated depression detection: Social posts, audio and video, open challenges and future direction," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 1, p. e7407, 2023.

[3] N. Seneviratne and C. Espy-Wilson, "Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6252–6256.

[4] C. Yuan, X. Liu, Q. Xu, Y. Li, Y. Luo, and X. Zhou, "Depression diagnosis and analysis via multimodal multi-order factor fusion," in *International Conference on Artificial Neural Networks*. Springer, 2024, pp. 56–70.

[5] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6247–6251.

[6] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "Tamfn: time-aware attention multimodal fusion network for depression detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 669–679, 2022.

[7] M. Squires, X. Tao, S. Elangovan, R. Gururajan, X. Zhou, U. R. Acharya, and Y. Li, "Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment," *Brain Informatics*, vol. 10, no. 1, p. 10, 2023.

[8] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *A practical guide to sentiment analysis*, pp. 1–10, 2017.

[9] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: a systematic review and research framework," *Translational Psychiatry*, vol. 13, no. 1, p. 309, 2023.

[10] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 1568–1586, 2023.

[11] X. Miao, Y. Li, M. Wen, Y. Liu, I. N. Julian, and H. Guo, "Fusing features of speech for depression classification based on higher-order spectral analysis," *Speech Communication*, vol. 143, pp. 46–56, 2022.

[12] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *International journal of intelligent systems*, vol. 37, no. 12, pp. 10 140–10 156, 2022.

[13] Z. Shen, I. Bello, R. Vemulapalli, X. Jia, and C.-H. Chen, "Global self-attention networks for image recognition," *arXiv preprint arXiv:2010.03019*, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3531–3539.

[16] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[17] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[18] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.

[19] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, and L. Van Gool, "Vision transformers with hierarchical attention," *Machine Intelligence Research*, pp. 1–14, 2024.

[20] Z. Zhao, Y. Wang, G. Shen, Y. Xu, and J. Zhang, "Tdfnet: Transformer-based deep-scale fusion network for multimodal emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3771–3782, 2023.

[21] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv e-prints*, pp. arXiv–1607, 2016.

[22] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[23] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 226–12 234.

[24] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[25] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[26] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 65–71.

[27] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[28] L. Zhou, Z. Liu, X. Yuan, Z. Shangguan, Y. Li, and B. Hu, "Caiinet: Neural network based on contextual attention and information interaction mechanism for depression detection," *Digital Signal Processing*, vol. 137, p. 103986, 2023.

[29] Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, and B. Hu, "Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer," *Digital Communications and Networks*, vol. 10, no. 3, pp. 577–585, 2024.

[30] T. Ling, D. Chen, and B. Li, "Mdavif: A multi-domain acoustical-visual information fusion model for depression recognition from vlog data," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8115–8119.

[31] I. Goodfellow, "Deep learning," 2016.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[33] T. Lin, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[34] A. Y. Ng, "On feature selection: learning with exponentially many irreverent features as training examples," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[35] E. Stevens, L. Antiga, and T. Viehmann, *Deep learning with PyTorch*. Manning Publications, 2020.

[36] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Appendix

Extension from AFEM

This layer reduces the dimensionality of the input, preserving crucial information while simplifying the data for further processing. It uses several convolutional layers, followed by batch normalization, to capture both local and global information. The operation of the bottleneck layer is represented by equation (16).

$$\hat{y}_a = \beta_o \left(\frac{1}{t} \sum_{i=1}^{t} (f_{bn,i}(f_{cn,i}(X_A^o))) + f_{bn}(f_{cn}(X_A^o)) \right) \tag{16}$$

Where, f_{cn} refers to the 1×1 convolutional layers, β_o is the weight matrix used for final classification, and t is the sequence length, where $t = 3$. The output, $\hat{y}_a$, which is the predicted probability, is then sent to the final classification layer to produce logits for binary classification using only the acoustic features.

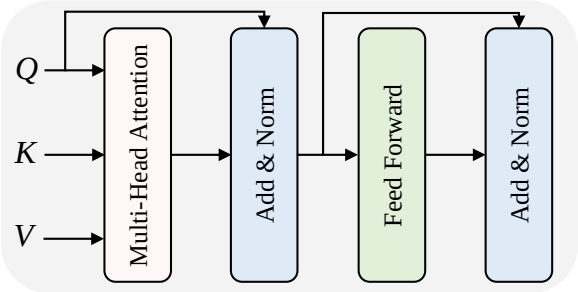


Figure 5: The architecture of transformer encoder: Q, K and V represent the query, key and value of inputs, respectively.

Extension from VFEM

After obtaining the output $X_V^o \in \mathbb{R}^{t \times d_o}$from the transformer block from equation (9), the model applies Global Average Pooling (GAP) to reduce the spatial dimensions by averaging the features across the feature map, resulting in a pooled feature vector. The pooled features are then passed through a fully connected (FC) layer.

$$\hat{y}_v = F\left(f_{gap}(X_V^o)\right) \tag{17}$$

where $F(\cdot)$ and f_{gap} denote fully connected layer and the function for GAP layer, respectively. The output, $\hat{y}_v$ from equation (17), which is the predicted probability, is then sent to the final classification layer to produce logits for binary classification using only the visual features.

Transformer Encoder

In the MT, the transformer encoder calculates the attention from multimodal downscaled features. As illustrated in Fig. 5, the transformer calculates the self-attention by a multihead attention mechanism.

Customized Loss Function (for equation (15)):

Smoothed Binary Cross-Entropy (BCE) with Logits forms the foundational component of this combined loss, serving as the primary loss for binary classification. This function directly optimizes binary classification tasks by minimizing the logarithmic loss between the true and predicted probabilities [31]. Label smoothing further regularizes the model by preventing it from becoming overly confident in its predictions. Instead of assigning a hard probability of 1 to the correct class and 0 to the others, label smoothing assigns a slightly lower probability (e.g., 0.9) to the correct class and distributes the remaining probability among the incorrect classes [32]. This is expressed mathematically by equations (18).

$$\begin{aligned} y_{\text{smooth}} &= y \cdot (1 - \epsilon_{\text{smooth}}) + 0.5 \cdot \epsilon_{\text{smooth}}, \\ \mathcal{L}_{\text{BCE}} &= -\left(y_{\text{smooth}} \log(p) + (1 - y_{\text{smooth}}) \log(1 - p)\right). \end{aligned} \tag{18}$$

where p represents the predicted probability, y is the true label, and ϵ_{smooth} is the smoothing factor.

Focal Loss is employed to address class imbalance and place greater emphasis on more difficult samples [33]. The focal loss can be defined as in equation (19).

$$\mathcal{L}_{\text{Focal}} = \alpha \cdot (1-p)^{\gamma} \cdot \mathcal{L}_{\text{BCE}} \tag{19}$$

Here, α is a scaling factor applied depending on whether the true label is 1 or 0, and γ is the focusing parameter, which determines how much the loss is concentrated on harder-to-classify examples.

Additionally, L2 regularization helps mitigate overfitting by penalizing large weights. This discourages the model from relying excessively on any single feature, thereby promoting better generalization to unseen data [34]. The L2 regularization loss is computed as shown in equation (20).

$$\mathcal{L}_{\text{L2}} = \lambda \sum \|\theta_i\|_2^2 \tag{20}$$

where θ_i represents the model parameters and λ is the strength of the L2 regularization.

Dataset

The D-Vlog dataset[23], collected from YouTube by Yoon et al. [23], consists of 961 vlog videos from 816 individuals (322 males and 639 females) and includes 555 depressed and 406 non-depressed labels. These labels were assigned based on specific keywords in the titles, with terms like "depression daily vlog", "depression journey", "depression vlog", "depression episode vlog", "depression video diary", "my depression diary", and "my depression story", indicating depressed vlogs and "daily vlog", "grwm (get ready with me) vlog", "haul vlog", "how to vlog", "day of vlog", "talking vlog", and so on, for non-depressed ones. To ensure the validity of the labels, Yoon et al. conducted two tasks: verifying that each video adhered to the vlog format, with individuals speaking directly to the camera, and employing annotators to analyze videos for signs of depression using auto-generated transcripts. The dataset is divided into training, validation, and test sets in a 7:1:2 ratio, and includes 25 low-level acoustic descriptors from OpenSmile [24] (with the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)) and 68 facial landmarks extracted by Dlib [25], to maintain privacy while providing relevant features. [**Dataset Requested: *D-Vlog: Multimodal Vlog Dataset for Depression Detection***]

Experimental Setup

In this paper, all experiments and code implemented based on PyTorch [35] framework based Python language. The training, validation and testing process of the model were all run on NVIDIA RTX A6000 graphics card with 48 GB memory. We employed the Adam optimizer [36], setting the learning rate to 1e-4, weight decay to 0.1, and epsilon to 1e-8. The model trained for 200 epochs with a batch size of 8, using a cosine annealing learning rate scheduler. To prevent overfitting, we implemented early stopping with a patience of 15, saving the best model based on the validation set performance. Additionally, 10-fold cross-validation was applied, and model performance was evaluated using metrics including accuracy, weighted precision, recall, and F1 score. [**Code: *Multimodal-Depression-Detection*, Private Repo: Please email (rezwanh001@gmail.com) me to get access.**]

Table 3: The Results of Ablation Studies with Our Model for Acoustic and Visual Feature.

Fusion	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Add	0.6927	0.7086	0.7909	0.7445
Multiply	0.6854	0.7125	0.7721	0.7338
Concat	0.6979	0.7214	0.7799	0.7469
MT	**0.7260**	**0.7392**	**0.8065**	**0.7707**

Ablation Study

To examine the effectiveness of each component in our model, we conduct ablation experiments in this section. In the previous subsection, our proposed model achieved the best performance on the D-Vlog dataset compared to the current models. To assess the impact of different fusion strategies on our model's performance, we conducted a series of ablation studies, as presented in Table 3. The results reveal that the mutual transformer (MT) fusion method outperforms the other fusion techniques, achieving the highest accuracy of 0.7260, precision of 0.7392, recall of 0.8065, and F1 score of 0.7707. In comparison, the additive fusion method yields an accuracy of 0.6927, with precision, recall, and F1 scores of 0.7086, 0.7909, and 0.7445, respectively. Similarly, the concatenation method shows promising results with an accuracy of 0.6979 and an F1 score of 0.7469, while the multiplicative method achieves slightly lower performance across the metrics, with an accuracy of 0.6854. These findings highlight the effectiveness of the MT module in capturing and integrating the complex interactions between acoustic and visual features, thus enhancing the model's overall performance in multimodal depression recognition. The ablation studies confirm that the design of the MT module plays a crucial role in leveraging diverse feature sets for improved classification outcomes.