Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Refinement Feature Fusion for Multi-model Depression Detection

Yongwei Liao

weijie0018@163.com

Shenzhen City Polytechnic

Yinbin Zhang

Shenzhen City Polytechnic

Weiwei Chen

Shenzhen City Polytechnic

Zhenjun Li

Shenzhen City Polytechnic

Bing Zhou

Shenzhen City Polytechnic

Research Article

Keywords: Depression Detection, Multimodal, Data Fusion, Large Model, Artificial Intelligence

Posted Date: May 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9000115/v1

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

Additional Declarations: No competing interests reported.

Refinement Feature Fusion for Multi-model Depression Detection

Yongwei Liao[1,2*], Yibin Zhang[1], Weiwei Chen[1], Zhenjun Li[1], Bing Zhou[1]

[1*]School of Information and Communications Technology, Shenzhen City Polytechnic, Longgang, Shenzhen, 518000, Guangdong, China.
[2]Shenzhen Institute of Technology, Shenzhen, 518000, Guangdong, China.

*Corresponding author(s). E-mail(s): weijie0018@163.com;
Contributing authors: 37975964@qq.com; chenwwayne@gmail.com;
lizhenjun@szcp.edu.cnm; bingzhou6@gmail.com;

Abstract

This paper proposes UniMDD (Unified Multi-modal Depression Detection), a novel framework designed for multimodal depression detection based on video interviews. It employs an efficient multi-model data fusion strategy, comprising four sequential modules: feature extraction, feature fusion, feature refinement, and depression detection. Specifically, the feature extraction module leverages OpenFace for facial features, VGGish for audio features, and a Large Language Model for textual features. The core innovation lies in the fusion and refinement stages. The feature fusion module designs a Mixed Attention Network(MAN) to capture intricate cross-modal interactions and effectively integrate complementary information from different data streams. Subsequently, the feature refinement module constructs an EHA-Attention mechanism within decoder blocks. This serves a dual purpose: it generates higher-level semantic and contextual representations while mitigating potential inter-modal conflicts and enhancing the comprehension of each individual modality. Finally, the refined features are fed into the depression detection module for prediction. Experimental validation on the LMVD dataset demonstrates that UniMDD achieves an accuracy of 76.68%, surpassing the current state-of-the-art method, DepMamba, by a margin of 4.55 percentage points. These results underscore the advancement and robustness of the proposed UniMDD framework.

Keywords: Depression Detection, Multimodal, Data Fusion, Large Model, Artificial Intelligence

1 Introduction

Traditional methods for diagnosing depression primarily rely on patients' subjective self-reports and clinical interviews conducted by medical professionals, often formalized through scale-based assessments [1]. These methods employ standardized assessment tools such as HAMD [2], BDI forms [3], and PHQ-8/PHQ-9 [4, 5] to assign quantitative scores for evaluation. However, inherent limitations, including patient reluctance to disclose symptoms and variability in clinical experience among professionals, contribute to diagnostic inaccuracies. Studies indicate that misdiagnosis rates using these traditional approaches can be as high as 20% in certain regions. Recent research has identified promising objective behavioral and physiological correlates for depression. Data from

modalities such as speech patterns, facial expressions, behavior, eye movements, skin response, blood pressure, and electroencephalogram (EEG) signals have demonstrated significant potential to aid in diagnosis [6, 7]. Consequently, with the rapid advancement of artificial intelligence, the use of multimodal data [8] has emerged as a prominent research focus for automated depression detection. Multimodal data fusion [8] integrates complementary information from diverse sources including images, text, audio, video, and physiological signals to construct a more comprehensive and enriched profile for analysis. This integrated approach can provide nuanced insights into a patient's affective state, physical gestures, expressed emotions, and thought processes, which are often critical indicators in depression assessment [7, 9].

Currently, the most popular multimodal depression detection models predominantly leverage deep learning architectures based on Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks [10]. Representative examples include models such as [11], TFN [12], TAMFN [13], DepMamba [14], and CANAMRF [15]. Notably, the leading method, DepMamba, employs convolutional neural networks alongside bidirectional Mamba modules [16] for feature extraction and utilizes a multimodal collaborative state space model for modality fusion. Despite these advances, existing models exhibit several persistent limitations for depression detection. Their inherent structural constraints often lead to inefficient processing of complex, high-dimensional multimodal data. Moreover, contemporary feature fusion techniques struggle to adequately capture intricate cross-modal interactions and dependencies, which ultimately impairs overall detection performance. Consequently, the accuracy of current multimodal detection systems generally plateaus between 70% and 80%, a range still considered insufficient for reliable clinical deployment. Furthermore, the dominant paradigm remains centered on multimodal data fusion using CNN, LSTM, or their derivatives. Therefore, we argue that a unified detection framework built upon transformer-based architectures [17] presents a promising avenue for advancement. Such a framework could enable more effective integration of multimodal features, generate more comprehensive depression-related representations, and fundamentally facilitate deeper cross-modal feature interaction.

To address these challenges, this paper proposes UniMDD, a novel depression detection framework built upon a large language model. The core objective of UniMDD is to establish a unified architecture that effectively facilitates information sharing and feature fusion across different modalities (e.g., text, audio, visual). It is designed to independently process the distinctive features of each modality while preserving their unique characteristics. Crucially, the framework incorporates mechanisms to prevent conflicts between heterogeneous features during fusion, thereby simultaneously enhancing the comprehensive understanding of each individual modality and their synergistic integration. By mining deeper, more contextual information from multimodal data, UniMDD aims to significantly improve model sensitivity and the accurate recognition of depression-related indicators.

The main contributions of this paper are summarized as follows:

- A refinement feature fusion model for Multimodel Depression Detection, named UniMDD, is proposed based on the large model, which consists of four core components: feature extraction, feature fusion, feature refinement, and depression detection.
- A mixed attention network(MAN) is utilized to fuse multimodel information, which can capture complementary information across modalities and improve feature representation for depression recognition.
- The feature refinement module is composed of decoder blocks with EHA-Attention, which facilitates cross-modal information interaction and generates higher-level semantic representations.

The remainder of this paper is organized as follows. Section 2 reviews related work on depression detection. Section 3 details the proposed UniMDD framework. Section 4 presents the experimental evaluation and analysis. Finally, Section 5 concludes the paper.

2 Related Work

In this section, we investigate the representative works on the feature fusion, feature refinement and depression detection as follows.

2.1 Feature Fusion

Feature alignment [18] addresses the inherent heterogeneity between modalities such as the divergent original distributions of text, image, and audio data by projecting different modal features into a shared semantic space, thereby making them comparable. Building upon aligned features, multimodal feature fusion [19] is the process of combining or integrating these features to generate more robust and discriminative joint representations that enhance downstream tasks like classification. In the specific domain of multimodal depression detection [7, 9], effectively fusing features from video, audio, and text modalities is a critical pathway to improving diagnostic performance. Common fusion strategies include early, intermediate, and late fusion, as well as more sophisticated paradigms such as transformer-based, mixture-of-experts (MoE)-based, and contrastive learning-based fusion [20]. Transformer-based fusion leverages self-attention mechanisms to capture long-range dependencies across modalities. For instance, VITA [21] projects text, video, and audio features into a unified space before employing a transformer for cross-modal fusion. The Mixture-of-Experts (MoE) paradigm, exemplified by Uni-MoE [22], trains multiple specialized expert networks and integrates them within a unified framework, often mapping diverse modalities into a common representation space via MoE layers. Contrastive learning-based approaches, like CLIP [23], project data from different modalities into a shared embedding space and optimize these representations through contrastive objectives, effectively adopting a variant of late fusion. Other models, such as ViLT [24] and ALBEF [25], explore early fusion and unified multitask training paradigms, respectively. While these methods provide a strong foundation, effective depression detection requires a fusion mechanism that not only integrates information but also prioritizes the most diagnostically relevant cues across modalities. Therefore, this paper employs a mixed attention mechanism to integrate multimodal features. This approach is chosen for its ability to dynamically model inter-modal relationships, allowing the model to capture complementary information inherent in the multimodal data and, crucially, to selectively focus on the modal features most beneficial for depression recognition. The resulting fused representation can more comprehensively reflect a subject's emotional and behavioral state, thereby improving the accuracy and reliability of the detection system.

2.2 Feature Refinement

In multimodal depression detection, the feature refinement module is a critical component for achieving efficient and accurate diagnosis by transforming fused features into higher-level, more discriminative representations. Existing work has developed various refinement strategies tailored to specific modalities or fusion paradigms. For the audio modality, TFCAV [26] integrates squeeze-and-excitation components with time-frequency channel attention blocks to extract depression-related features from spectrograms. AHDD [27] introduces a feature variation coordination measurement (FVCM) algorithm to model correlations in time-lagged multichannel variations, fusing speaker identification and emotion recognition features. In the video modality, DepressNet [28] utilizes pre-trained deep learning models to extract depression representations from facial images, while the SRBP [29] framework learns patterns separately from RGB and encoded video frames, fusing them via average pooling. Recent advanced models like DepMamba [14] and CANAMRF [15] have made significant progress in cross-modal modeling. DepMamba addresses long-sequence inefficiencies through hierarchical context modeling and progressive multimodal fusion. CANAMRF employs a cross-attention mechanism to adaptively fuse data from different modalities. However, while effective, these models often prioritize learning either modality-shared characteristics or modality-specific information, struggling to achieve an optimal balance that captures comprehensive cross-modal interactions while preserving the unique discriminative power of each modality. To bridge this gap and further improve detection accuracy, we argue that an explicit feature refinement stage is necessary. This stage should process the initially fused features to generate

richer, higher-level semantic information and contextual relationships. Therefore, in this work, we propose to employ a transformer decoder module for feature refinement. Leveraging its multi-head self-attention mechanism and feed-forward networks, this module can effectively integrate information across the fused multimodal sequence, model complex dependencies, and ultimately generate advanced, task-specific semantic representations that are more sensitive to depression-related indicators.

2.3 Depression Detection

Depression detection is generally divided into manual detection based on physician subjective judgment (MDD) and automated detection based on artificial intelligence (ADD). Manual detection often employs scales and questionnaires, whereas ADD utilizes physiological signals such as EEG, ECG, body temperature, MEG, eye movements, behavioral cues including facial expressions, actions, and postures. Detection methods include text-based methods such as BDI [3] and PHQ [4, 5]; speech-based methods like MFCC, SVM and DepAudioNet [30]; video-based methods such as DepressNet [28] and DLGA-CNN [31]; multimodal methods such as Bi-LSTM [11], MFM-attention [32], DCGAN [33], and DepMamba [14]. Specifically, Bi-LSTM can process time-series data from text (e.g., dialogue) and audio to extract emotional and semantic features. TFN [12] represents features from different modalities as tensors and fuses them through tensor products, capturing cross-modal interactions and generating a unified feature vector for classification or regression tasks. TAMFN [13] consists of a Global Temporal Convolution Network (GTCN), an Inter-modal Feature Extraction module (IFE), and a Temporal-Aware Attention-based Multimodal Fusion module (TAMF). GTCN integrates local and global temporal information, IFE extracts early inter-modal interactions, and TAMF dynamically fuses multimodal features using temporal-aware attention. CANAMRF [15] employs an adaptive mechanism to dynamically adjust modality weights and uses a recurrent structure for iterative multimodal fusion. It combines adaptive fusion with attention mechanisms to more effectively capture cross-modal interactions. DepMamba [14] is an efficient multimodal depression detection model that adopts hierarchical context modeling and progressive multimodal fusion. It extracts features using convolutional neural networks and bidirectional Mamba modules [16], and employs a multimodal collaborative state-space model (SSM) for modality fusion. MMIM [34] employs hierarchical mutual information maximization within a noise-contrastive estimation (NCE) framework to the maximize mutual information between fused outputs and unimodal inputs. Although these multimodal models demonstrate promise for early diagnosis, they face several persistent challenges that hinder clinical application: inefficient modeling of long-sequence data, suboptimal fusion strategies that fail to fully capture cross-modal dynamics, high model complexity, and strong dependency on large, high-quality datasets. To address these limitations particularly concerning data dependency and fusion efficacy this paper proposes the Unified Multi-modal Depression Detection (UniMDD) framework. UniMDD is designed to enhance performance and reliability by providing a more robust architecture for integrating multimodal information.

3 Methods

In this section, we introduce the architecture of UniMDD.

3.1 The structure of UniMDD

As illustrated in Figure 1, the proposed UniMDD framework comprises four sequential stages: unimodal feature extraction, multimodal feature fusion, hierarchical feature refinement, and final depression detection. The overall pipeline first extracts features from video, audio, and text inputs independently, then fuses them into a unified representation. This representation is subsequently refined by a large model to learn deep, discriminative patterns before final classification. The model processes each modality with a dedicated backbone:

- Video Features: The OpenFace toolkit is employed to extract 68 facial landmark points per frame. By integrating these points across multiple frames, the model captures dynamic

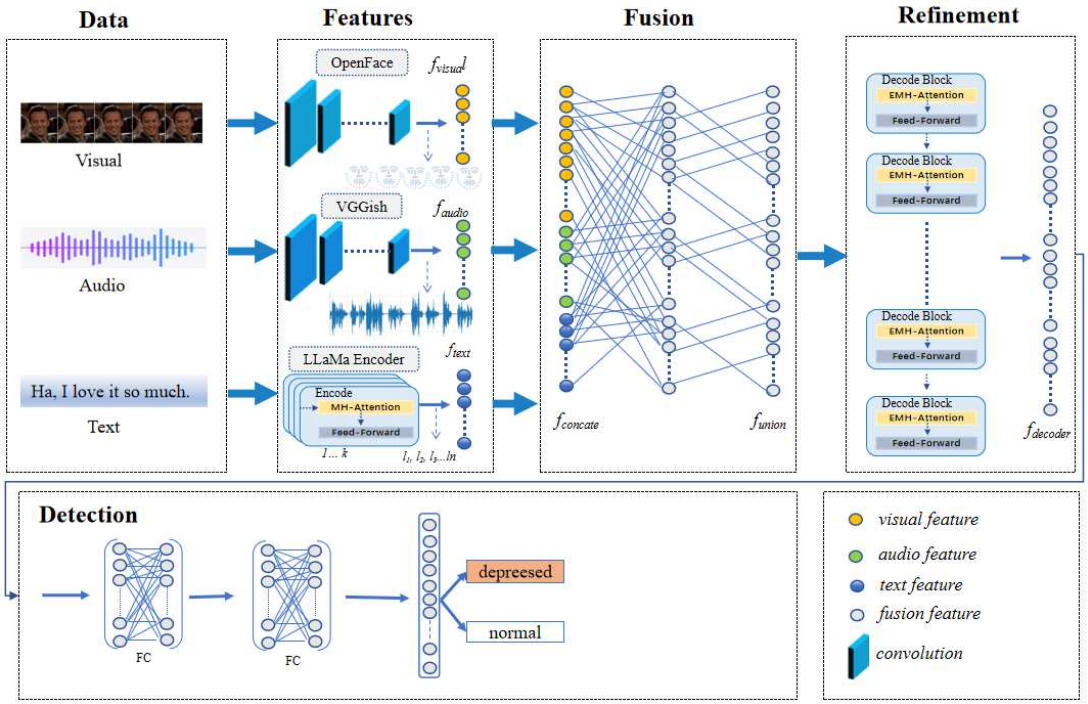


Fig. 1: The structure of UniMDD.

facial expressions while filtering out redundant background scene information. This process yields a 136-dimensional contextual feature vector for each segmented video clip.

- Audio Features: For each utterance, log-mel spectrogram features are extracted using the VGGish model, initially producing a 6,373-dimensional vector. This high-dimensional representation is then projected into a more compact 128-dimensional feature space via a fully connected layer.
- Text Features: Textual input is processed by the encoder stack of the LLaMA large language model. After passing through four encoder blocks, a 256-dimensional semantic feature representation is obtained.

Given that the primary experimental dataset contains aligned video and audio streams, the fusion stage focuses on these two modalities. First, temporal alignment is performed by synchronizing the video and audio features along the time axis. The aligned 136-dimensional video and 128-dimensional audio features are then concatenated, resulting in a unified 264-dimensional spatio-temporal feature vector for each time step.

The fused sequence serves as input to the feature refinement module, which is built upon a Transformer decoder architecture. This module consists of four stacked decoder blocks, each equipped with multi-head self-attention and feedforward neural network sub-layers. It processes the temporal sequence to model long-range dependencies, suppress inter-modal noise or conflicts, and generate a refined, higher-level semantic representation. Finally, this refined representation is passed to the depression detection head, which consists of two fully connected layers, to produce the final binary classification outcome (depressed or non-depressed).

3.2 Feature Extraction

The UniMDD framework extracts features from video, audio, and text modalities using dedicated models. The following sections detail the process for each modality.

3.2.1 Video Feature Extraction via OpenFace

For video input, the OpenFace toolkit is employed to extract facial action features. Given an

input image (e.g., 96 × 96 pixels), the process begins with face detection and alignment. Multi-task Cascaded Convolutional Networks (MTCNN) detect 68 facial landmarks. Based on these points, an affine transformation is calculated to correct for rotation and scale, aligning the face to a standard position. The aligned image is then resized to the required input dimensions for the neural network.

The core of OpenFace uses a Constrained Local Neural Field (CLNF) model to precisely localize facial landmarks. The response map for the j-th landmark is calculated as:

$$R_j(x) = \sum_{k=1}^{n} w_k \cdot \phi_k(I(N(x)); \theta_k) \tag{1}$$

where ϕ_k is the k-th filter, $I(N(x))$ is the image patch at location x, θ_k are the filter parameters, and w_k is a combination weight. The landmark positions are constrained by a point distribution model:

$$S = S_0 + \sum_{i=1}^{m} p_i \cdot S_i \tag{2}$$

where S_0 is the mean shape, S_i are shape bases, and p_i are the shape parameters. Finally, Facial Action Units (AUs) are recognized using a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. The classification function is:

$$f_{visual} = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}i, \mathbf{x}) + b\right) \tag{3}$$

where $K(\cdot, \cdot)$ is the RBF kernel. The resulting 136-dimensional landmark data and AU activations are combined and projected to form the final video feature vector f_{visual}.

3.2.2 Audio Feature Extraction via VGGish

Audio features are extracted using the VGGish model, which maps audio signals to a compact 128-dimensional embedding space. The audio signal is first transformed into a time-frequency representation: a Short-Time Fourier Transform (STFT) is applied, followed by a Mel-filter bank and logarithmic compression to obtain a Log-Mel spectrogram (e.g., of size 96 × 64 frames × frequency bands). This spectrogram is processed by a VGG-16-like convolutional network. The convolution operation for a given layer is defined as:

$$Conv_{i,j,k} = Relu(\sum_{l,m,n} x_{i+l-1,j+m-1} \cdot w_{l,m,n,k} + b_k) \tag{4}$$

where w denotes the convolutional kernel weights and b the bias. The final audio feature can be summarized as:

$$f_{audio} = \text{VGGish}(S_{mel}) \tag{5}$$

where S_{mel} is the input Log-Mel spectrogram.

3.2.3 Text Feature Extraction via LLaMA Encoder

Textual input is processed by the LLaMA large language model encoder. The text is first tokenized using LLaMA's Byte-Pair Encoding (BPE) tokenizer and mapped into a high-dimensional embedding space (e.g., dimension 4096 for LLaMA-2 7B). The token sequence is then passed through the encoder stack. Within each transformer block, self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, V are the query, key, and value matrices, and d_k is the key dimension. This is followed by a feed-forward network (FFN) with Gaussian Error Linear Unit (GELU) activation and layer normalization. The operation of a single FFN can be expressed as:

$$\text{FFN}(x) = \text{LayerNorm}\left(\text{GELU}(xW_1 + b_1)W_2 + b_2\right) \tag{7}$$

where W_1, W_2 and b_1, b_2 are learnable parameters. After processing through four encoder blocks, the resulting contextualized token representations (shape: [batch_size, sequence_length, hidden_size]) are aggregated via pooling to produce a fixed-length text feature vector f_{text} of 256 dimensions.

3.3 Feature Fusion(MAN)

Following independent feature extraction, the core challenge is to effectively integrate information

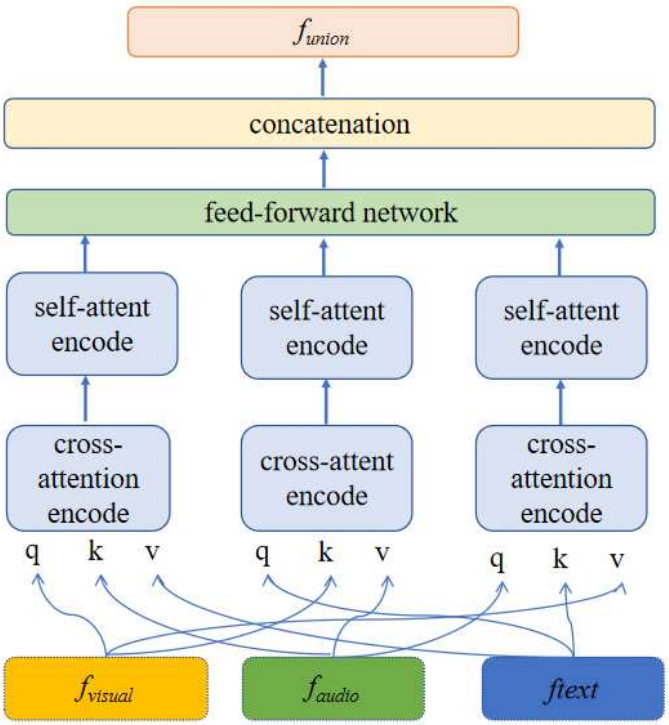


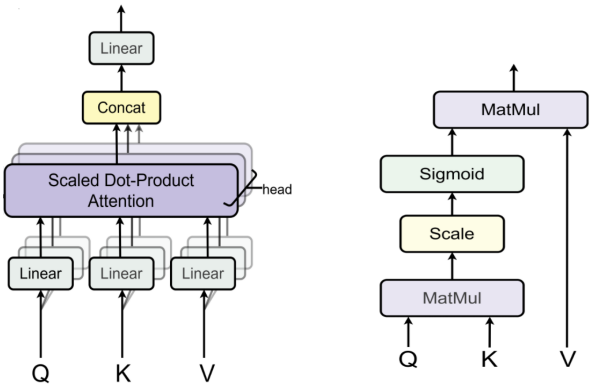
Fig. 2: The structure of mixed attention network(MAN).

from video, audio, and text modalities. As illustrated in Figure 2, we implement a feature-level fusion strategy centered on a Mixed Attention Network. The primary goal of this stage is to capture the complementary information inherent in multimodal data and dynamically prioritize modality-specific features that are most discriminative for depression recognition. The initial step involves the concatenation of the extracted video, audio, and text feature vectors along the feature dimension, forming a high-dimensional joint representation.

The primary goal of this stage is to capture the complementary information inherent in multimodal data and dynamically prioritize modality-specific features that are most discriminative for depression recognition. The initial step involves the concatenation of the extracted video, audio, and text feature vectors along the feature dimension, forming a high-dimensional joint representation. However, simple concatenation alone is insufficient as it cannot model the complex, non-linear interactions between modalities. To address this, we introduce a mixed attention transformation block. This block sequentially applies:

1. Cross-Attention: To explicitly model the relationships and dependencies between different modalities, allowing the model to contextualize information from one modality with respect to others.
2. Self-Attention: To refine the representation within the newly formed joint feature space, capturing long-range dependencies and enhancing the integration of the cross-modality cues.

The output of this mixed attention process is a new, refined fusion feature representation that comprehensively encodes emotional and behavioral information, thereby improving the model's accuracy and robustness. Prior to fusion, temporal alignment is performed: the feature sequences from each modality are synchronized along the time axis based on their timestamps. For model training, we employ a data augmentation technique to improve generalization and handle variable-length sequences. From the original aligned sequence, we uniformly and randomly sample multiple subsequences, each guaranteed to contain at least 400 frames. This process yields the final training instances: temporal data with



(a) Multi-Head Attention Layer (b) Enhanced Multi-Head Layer

Fig. 3: The structure of EHA-attention

a shape of (sequence length, feature dimension), paired with their corresponding depression labels.

In our general framework, the text, audio, and visual features denoted as f_{text}, f_{audio}, and f_{visual} respectively are processed through a mixed attention module to obtain a unified feature representation f_{union}. This module sequentially applies cross-attention and self-attention mechanisms, which can be formulated as:

$$f_{union} = \text{MixAtt}(f_{text}, f_{audio}, f_{visual}; Att_{cross}, Att_{self}) \tag{8}$$

where Att_{cross} denotes the cross-attention module for inter-modal interaction, Att_{self} denotes the self-attention module for intra-modal refinement, and MixAtt$(\cdot)$ represents the combined function of these operations. The final unified feature f_{union} is subsequently used for downstream tasks.

In the current study, experiments are conducted on the LMVD dataset, which provides synchronized video and audio modalities but does not include textual transcripts. Consequently, the implemented UniMDD model adapts by fusing only the audio and visual modalities. In this adaptation, the final fused representation is obtained through feature concatenation:

$$f_{concat} = [f_{audio}; f_{visual}] \tag{9}$$

This concatenated feature vector f_{concat} serves as the direct input to the subsequent feature refinement module. The design readily generalizes to the three-modal case when textual data is available, as formalized in Equation (8).

3.4 Feature Refinement(EHA)

The Transformer decoder, composed of multiple stacked, identically structured layers, has demonstrated remarkable effectiveness in capturing sequential features and long-range dependencies. Each layer typically includes core components such as Attention mechanisms (e.g., Multi-Head Attention or Grouped-Query Attention), Feed-Forward Networks (FFN), and Layer Normalization.

To adapt this powerful architecture for the specific demands of depression detection from sequential multimodal data, we develop Enhanced Hybrid Attention (EHA-Attention) to construct our decoder blocks, forming the core of the feature refinement module. As illustrated in Figure 3, we implement the following key modifications to the standard decoder layer:

1. Enhanced Attention with Sigmoid Activation: In the self-attention sub-layer, we replace the standard softmax activation function with a sigmoid function (see Fig.3-b). This modification allows the model to capture more subtle and graded relationships between different positions in the sequence, moving beyond a strict probability distribution to a weighted gating mechanism, which can improve feature sensitivity.

2. Enhanced Feed-Forward Network: The standard FFN is replaced with a more expressive Multi-Layer Perceptron (MLP). This MLP consists of two linear layers: the first expands the feature dimension by a tunable factor, and the second projects it back to the original dimension. GELU activation and dropout layers are incorporated within this block to enhance nonlinear representation capacity and prevent overfitting, enabling the learning of more complex patterns while maintaining efficiency.

The foundational Transformer decoder operation for one layer can be described as:

$$\begin{aligned} f' &= \text{LayerNorm}\big(f_{in} + \text{MHA}(f_{in})\big) \\ f_{decoder} &= \text{LayerNorm}\big(f' + \text{FFN}(f')\big) \end{aligned} \tag{10}$$

where f_{in} is the input to the layer, MHA is the Multi-Head Attention mechanism, and FFN is the Feed-Forward Network.

Integrating our proposed enhancements, the operation of a single EHA-Attention decoder block in our feature refinement module becomes:

$$\begin{aligned} f' &= \text{LayerNorm}\big(f_{in} + \text{SigmoidMHA}(f_{in})\big) \\ f_{decoder} &= \text{LayerNorm}\big(f' + \text{MLP}(f')\big) \end{aligned} \tag{11}$$

where SigmoidMHA denotes the multi-head attention with sigmoid activation, and MLP represents the enhanced perceptron block described above. The final output of the stacked refinement module provides a contextually enriched, high-level representation for the subsequent depression detection head.

3.5 Depression Detection

The final component of the UniMDD framework is the classification head, which processes the output of the feature refinement module, denoted as $f_{decoder}$. This refined feature has a shape of [B, L, N], where B is the batch size, L is the temporal sequence length, and N is the feature dimension. First, an adaptive average pooling operation is applied along the temporal dimension (L). This compresses the sequential representation into a fixed-dimensional, global feature vector with shape [B, N], effectively aggregating the contextual information across the entire sequence into a single summary vector. Subsequently, this condensed high-level representation is fed into a linear layer (a fully connected layer). This layer serves as the classifier, mapping the pooled feature vector to the final prediction scores specifically, a logit for the binary classification task (depressed vs. non-depressed). Through this architecture from temporal pooling to linear projection the model effectively distills the learned, depression-relevant representations from the refined multimodal sequence into a robust and accurate classification decision.

4 Experiments

In this section, we evaluate the performance of UniMDD on for depression detection.

4.1 Datasets and Metrics

To validate the effectiveness of UniMDD, we conducted extensive experiments on the large-scale multimodal depression dataset LMVD(Large-Scale Multimodal Vlog Dataset) [35]. LMVD is a real-world audio-visual depression detection dataset collected from vlog videos published on social media platforms such as YouTube and TikTok. It contains over 1,200 samples from vloggers clinically diagnosed with depression and more than 2,500 samples from a control group (non-depressed). Each sample consists of complete vlogs, with an average duration of 10 to 30 minutes.The dataset includes visual features such as facial expressions and behaviors, audio features such as emotional intonation and phonetic features. LMVD is specifically designed for depression detection, where depression symptoms can be identified by analyzing creators video content, vocal attributes and behavioral patterns. D-Vlog [35] dataset was partitioned into training, validation and test sets in a ratio of 7:1:2, while LMVD dataset was split in a ratio of 8:1:1. To mitigate the impact of randomness, all reported experimental results are the average of three experiments conducted. For the binary classification task of depression detection, we use accuracy, precision, recall and F1-score to evaluate the model. In contrast, for he regression

Table 1: Comparison between UniMDD and popular algorithms.

Methods	Accuracy	Precision	Recall	F1 Score	Avg
KNN	56.83±1.64	57.43±1.52	50.92±4.44	53.92±2.97	54.77
Bi-LSTM	66.85±1.92	65.81±3.58	70.33±4.49	67.83±0.70	67.71
TBN	67.94±1.76	67.06±2.92	70.33±3.96	68.56±1.35	68.47
STFT	67.76±1.09	69.20±2.65	64.01±7.20	66.23±3.03	66.80
TFN	63.93±1.89	64.08±2.25	62.64±1.10	63.34±1.59	63.50
DepTran	61.93±3.01	60.36±4.19	72.16±13.43	65.08±3.52	64.88
TAMFN	70.49±0.55	71.15±2.73	68.86±5.19	69.84±1.30	70.09
DepMamba	72.13±1.89	70.18±1.66	**76.56±2.54**	73.20±1.01	73.02
UniMDD	**76.68±0.83**	**78.05±2.11**	73.99±1.68	**75.94±0.57**	**76.17**

task of estimating depression severity, the metric are root mean square error(RMSE) and mean absolute error(MAE).

4.2 Experiment settings

During the experiment, the learning rate was set to 1e-4, batch size was set to 4, and the AdamW optimizer was used to train the model. Two 24 GB NIVDIA RTX 3090 GPUs are used for parallel training. We use the cross-entropy function as the objective function, set the epoch to 100, and use L2 regularization with weight coefficients set to 0.1. The head in the multi-head attention layer of each transformer block is set to 8, the dropout of MLP is set 0.4. The overall trainable parameters reached 8.54 million by stacking four transformer blocks, where the average GPU usage is about 17 GB of video memory and the average training time is approximately 1.5 hours.

4.3 Comparison between UniMDD and popular algorithms

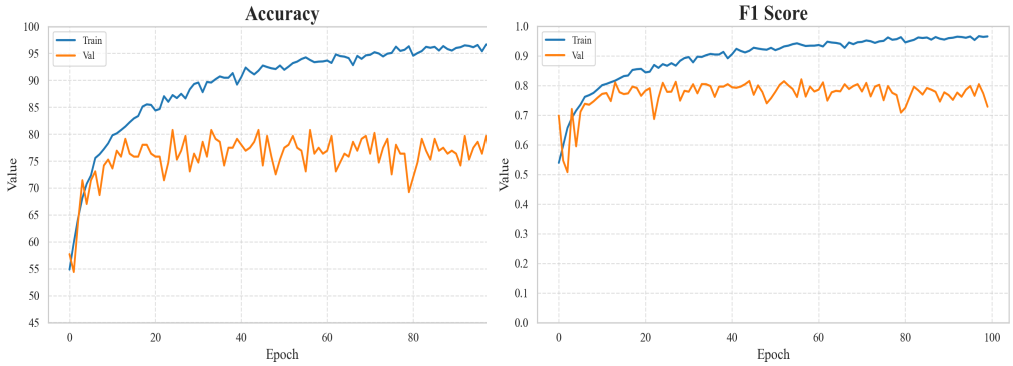
Table 1 presents a comparison between UniMDD and other popular algorithms conducted on LMVD dataset, showing that UniMDD achieves state-of-the-art performance. From Table 1 , we can make the following observations: 1)UniMDD outperforms other methods on LMVD dataset. Compare to the latest method DepMamba, UniMDD improves accuracy by approximately 4.55 and precision by 7.87. This improvement can be attributed to the transformer-based large model architecture, which has strong capabilities in representing and identifying multimodal features. 2)The increase in accuracy and precision is evident, which indicates an improvement in the accuracy of overall and positive predictions. Moreover, the recall decreased by 2.57, but the F1-score still increased by 2.74, which reflects a more balanced performance between precision and recall. 3)On average, UniMDD outperforms DepMamba by nearly 3.15, indicating the effectiveness of the transformer-based large model framework in depression detection compared to Mamba-based approaches. However, it is noted that increasing the model parameter count can also result in longer training time and higher GPU resource consumption.

4.4 The training process of UniMDD

Figure 4 illustrates the training process of UniMDD on the LMVD dataset. From the figure, we can make the following observations: 1)The large transformer-based model for depression detection exhibits noticeable oscillations in the later stages of training, indicating potential data imbalance in the distribution of samples. 2)The oscillations in accuracy, precision, recall, and F1-score during training remain largely consistent, suggesting strong correlations among these evaluation metrics. 3)In the later phases of training, all metrics accuracy, precision, recall, and F1-score consistently exceed those of DepMamba, demonstrating that the proposed large-scale depression detection model significantly outperforms the current SOTA algorithm.

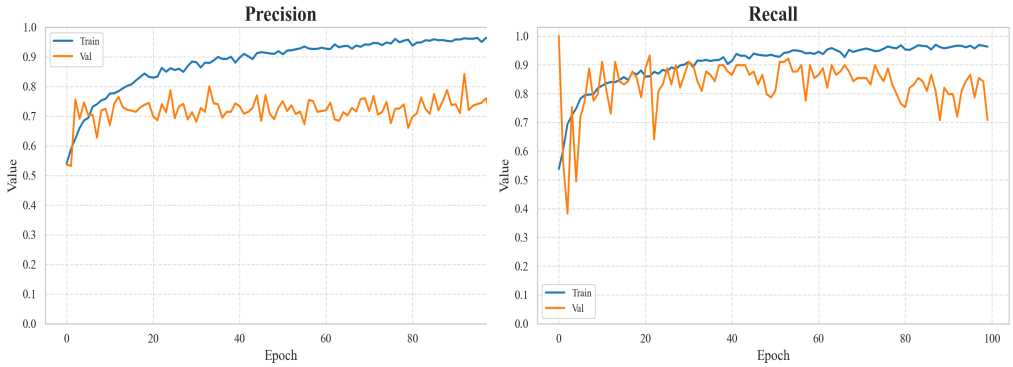
4.5 The Loss in training process

Figure 5 shows the loss of UniMDD during the training process on the LMVD dataset. From



(a) Accuray

(b) F1 Score



(c) Precision

(d) Recall

Fig. 4: The training process of UniMDD.



Fig. 5: The training loss of UniMDD on the LMVD dataset.

the figure, we can observe that: 1)The loss value decreases rapidly in the early stages of training, indicating the model can learn discriminative features quickly. 2)In the middle and later stages, the decline in loss gradually stabilizes, but occasional fluctuations are still observable. This could be attributed to the complexity of the data distribution and the challenges of optimizing a large-scale model. 3)Compared to the baseline method DepMamba, UniMDD achieves a lower final loss value, indicating superior convergence and stronger representation. 4)These loss dynamics reflect the effectiveness of the proposed architecture and training strategy, and highlight the challenges associated with large-scale multimodal learning.

4.6 Ablation experiment

Table 2 shows the ablation studies of UniMDD, and the accuracy of individual modality. From the table, we can observe that: 1)The effectiveness of our module optimizations and strategies is evident in the results. This is because information from a single modality (such as text, audio, or vision) is often incomplete and noisy.

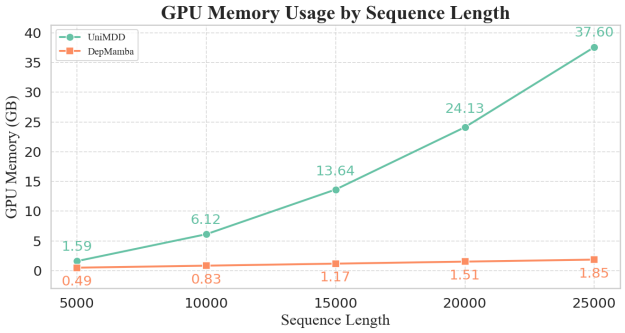
Table 2: Ablation experiment of UniMDD.

Methods	Accuracy
only audio	59.02
only visual	67.77
softmax	72.03
attention	66.45
mlp	74.40
UniMDD	**76.17**

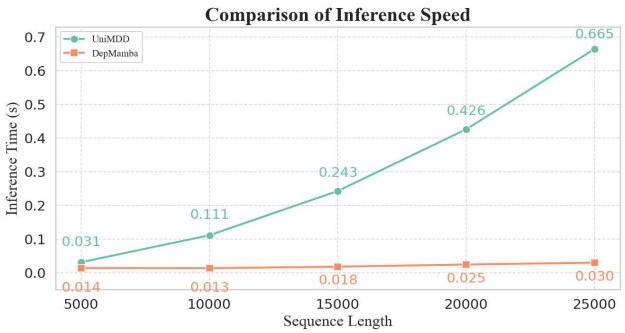
The advantage of multimodal fusion lies in its ability to simulate the cognitive process of comprehensive judgment in humans. When information from one modality is ambiguous or noisy, other modalities can provide crucial evidence for correction and supplementation. By learning these cross-modal correspondences and complementary relationships, the model makes more robust and accurate judgments. 2)The accuracy improves significantly after incorporating the attention and MLP modules. This is because the attention mechanism allows the model to directly capture relationships between any two elements in a sequence, regardless of their distance. For example, the model can detect long-range dependencies between "a forced smile in the opening remarks" and "negative statements at the end." The MLP further maps the fused features into a new feature space that is more conducive to the final task, such as classification.

4.7 Analysis of Performance and Effectiveness

Figure 6 compares the efficiency of UniMDD and DepMamba in various dimensions, including computational resource consumption (e.g. GPU memory, parameter counting), training/inference speed and model performance. From an architectural perspective, DepMamba may have an edge in terms of computational efficiency and memory usage when dealing with long sequence data, due to its linear computational complexity, compared to UniMDD which is based on the Transformer architecture. These advantages are particularly evident in terms of training speed and inference latency, potentially reducing hardware requirements. However, it is important that Transformer-based architecture of UniMDD has



(a) GPU Memory Usage



(b) Infence Speed

Fig. 6: Efficiency Comparison: UniMDD vs. DepMamba.

undergone extensive optimization and benefits from a mature ecosystem and powerful hardware support, which cannot be ignored.

5 Conclusion

This paper builds a framework for multimodal depression detection in video interviews, called UniMDD, which uses an efficient multi-model data fusion. This framework consists of four modules: feature extraction, feature fusion, feature refinement, and depression detection. The feature extraction module utilizes OpenFace to extract facial features, VGGish to extract audio features, and the large language model to extract text features. In particular, the feature fusion module develops a mixed attention network to capture cross-modal interactions and enable the model to leverage complementary information from multimodal data. The feature refinement module constructs EHA-Attention to form decoder blocks

for generating higher-level semantic and contextual information, which can prevent the confict from different modalities and strengthen the comprehension of each modality. The depression detection module predicts the presence of depression. Experimental results on the LMVD dataset demonstrate that UniMDD achieves an accuracy of 76.68%, outperforms current state-of-the-art method DepMamba by 4.55, and proves its advancement and effectiveness.

Declarations

Funding

This work was supported by (i)Natural Science Foundation of Shenzhen Municipality(Grant No.JCYJ20250604184310015) (ii)Guangdong Province Ordinary University Project(Grant No.2024KTSCX260) (iii)Shenzhen Institute of Technology School-level Project(Grant No.2611010).

Conflicts of Interest

The authors declare no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

This study is a secondary analysis of data collected in an earlier study. The dataset is publicly available at: https://github.com/helang818/LMVD/ [35].

References

[1] He, L., Guo, C., Tiwari, P., Pandey, H.M., Dang, W.: Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. International Journal of Intelligent Systems **37**(12), 10140–10156 (2021) https://doi.org/10.1002/int.22426

[2] Wang, L., Chignell, M., Jiang, H., Lokuge, S., Mason, G., Fotinos, K., , Katzman, M.: Discovering the causal structure of the hamilton rating scale for depression using causal discovery. IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 1–4 (2021) https://doi.org/10.1109/BHI50953.2021.9508538

[3] Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F.: Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. Journal of personality assessment **67**(3), 588–597 (1996)

[4] Kroenke, K., Spitzer, R.L.: The phq-9: A new depression diagnostic and severity measure. Psychiatric Annals **32**(9), 509–515 (2002)

[5] Williams, L.S., Brizendine, E.J., Plue, L., Bakas, T., W. Tu, H.H., Kroenke, K.: Performance of the phq-9 as a screening tool for depression after stroke. stroke **36**(3), 635–638 (2005)

[6] Suto, T., Fukuda, M., Ito, M., Uehara, T., Mikuni, M.: Multichannel near-infrared spectroscopy in depression and schizophrenia: cognitive brain activation study. Biological psychiatry **55**(5), 501–511 (2004)

[7] He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., Pan, X., Dang, W.: Deep learning for depression recognition with audiovisual cues: A review. Information Fusion **88**, 56–86 (2022) https://doi.org/10.1016/j.inffus.2021.10.012

[8] Nykoniuk, M., Basystiuk, O., Shakhovska, N., Melnykova, N.: Multimodal data fusion for depression detection approach. Computation **13**(9), 1–18 (2024) https://doi.org/10.3390/computation13010009

[9] Tahir, W.B., Khalid, S., Almutairi, S.S.,

Abohashrh, M., Memon, S.A., Khan, J.: Depression detection in social media: A comprehensive review of machine learning and deep learning techniques. IEEE Access **13**, 12789–12818 (2025) https://doi.org/10.1109/ACCESS.2025.3530862

[10] Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of twitter users. CLPsych@NAACL-HTL, 88–97 (2018) https://doi.org/10.18653/v1/W18-0609

[11] Yin, S., Liang, C., Ding, H., Wang, S.: A multi-modal hierarchical recurrent neural network for depression detection. AVEC@MM, 65–71 (2019) https://doi.org/10.1145/3347320.3357696

[12] Zadeh, A., M. Chen, S.P., Cambria, E., Morency, L.: Tensor fusion network for multimodal sentiment analysis. EMNLP, 1103–1114 (2017)

[13] Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., L, Y., Hu, B.: Tamfn: Time-aware attention multimodal fusion network for depression detection. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 669–679 (2022)

[14] Ye, J., Zhang, J., Shan, H.: Depmamba: Progressive fusion mamba for multimodal depression detection. ICASSP, 1–5 (2025) https://doi.org/10.1109/ICASSP49660.2025.10889975

[15] Wei, Y., Zhang, Y., Zhang, S., Zhang, H.: Canamrf: An attention-based model for multimodal depression detection. PRICAI **2**, 111–116 (2023) https://doi.org/10.1007/978-981-99-7022-3_10

[16] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. ICML (2024)

[17] Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., Yu, J.: Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion **104**, 102161–102169 (2024) https://doi.org/10.1016/j.inffus.2023.102161

[18] Xiao, Z., Zhang, X., Sun, G., Shao, H.: Ftan: Feature transform and alignment network for cross-domain specific emitter identification. Signal Process **230**, 106800–109810 (2025) https://doi.org/10.1016/j.sigpro.2024.109800

[19] Wang, Y., Wang, Z., Li, C., Zhang, Y., Wang, H.: A multimodal feature fusion-based method for individual depression detection on sina weibo. IPCCC, 1–8 (2020) https://doi.org/10.1109/IPCCC50635.2020.9391501

[20] Khan, M., Tran, P., Pham, N., Saddik, A., Othmani, A.: Memocmt: multimodal emotion recognition using cross-modal transformer-based feature fusion. SCIENTIFIC REPORTS (2025) https://doi.org/10.1038/s41598-025-89202-x

[21] Fu, C., Lin, H., Wang, X., Zhang, Y., Shen, Y., Liu, X., Cao, H., Long, Z., Gao, H., Li, K., Ma, L., Zheng, X., Ji, R., Sun, X., Shan, C., , He, R.: Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. CoRR abs/2501.01957 (2025) https://doi.org/10.48550/arXiv.2501.01957

[22] Li, Y., Jiang, S., Hu, B., Wang, L., Zhong, W., Luo, W., Ma, L., Zhang, M.: Uni-moe: Scaling unified multimodal llms with mixture of experts. IEEE Trans. Pattern Anal. Mach. Intell. **47**(5), 3424–3439 (2025) https://doi.org/10.1109/TPAMI.2025.3532688

[23] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. ICML, 8748–8763 (2021)

[24] Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. ICML, 5583–5594 (2021)

[25] Li, J., Selvaraju, R.R., Gotmare, A., Joty,

S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS, 9694–9705 (2021)

[26] Niu, J., Liu, B., Tao, J., Li, Q.: A time-frequency channel attention and vectorization network for automatic depression level prediction. Neurocomputing **450**, 208–218 (2021) https://doi.org/10.1016/j.neucom.2021.04.056

[27] Dong, Y., Yang, X.: A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing **450**, 279–290 (2021) https://doi.org/10.1016/j.neucom.2021.02.019

[28] Zhou, X., Jin, K., Shang, Y., Guo, G.: Visually interpretable representation learning for depression recognition from facial images. IEEE Transactions on Affective Computing **11**(3), 542–552 (2020) https://doi.org/10.1109/TAFFC.2018.2828819

[29] Song, S., Jaiswal, S., Shen, L., Valstar, M.F.: Spectral representation of behaviour primitives for depression analysis. IEEE Transactions on Affective Computing **13**(2), 829–844 (2022) https://doi.org/10.1109/TAFFC.2020.2970712

[30] Ma, S., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: An efficient deep model for audio based depression classification. AVEC@ACM Multimedia, 35–42 (2016) https://doi.org/10.1145/2988257.2988267

[31] He, L., Chan, J.C., Wang, Z.: Automatic depression recognition using cnn with attention mechanism from videos. Neurocomputing, 165–175 (2021) https://doi.org/10.1016/j.neucom.2020.10.015

[32] Su, L., Hu, C., Li, G., Cao, D.: Msaf: Multimodal split attention fusion. CoRR abs/2012.07175 (2020)

[33] Yang, L., Jiang, D., Sahli, H.: Feature augmenting networks for improving depression severity estimation from speech signals. IEEE Access, 24033–24045 (2020) https://doi.org/10.1109/ACCESS.2020.2970496

[34] Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. EMNLP, 9180–9192 (2021) https://doi.org/10.18653/v1/2021.emnlp-main.723

[35] He, L., Chen, K., Zhao, J., Wang, Y., Pei, E., Chen, H., Jiang, J., Zhang, S., Zhang, J., Wang, Z., He, T., Tiwari, P.: Lmvd: A large-scale multimodal vlog dataset for depression detection in the wild. Information Fusion, 103632–103642 (2026) https://doi.org/10.1016/j.inffus.2025.103632