



## OPEN ACCESS

## EDITED BY

Bing Liu,  
Beijing Normal University, China

## REVIEWED BY

Yong Li,  
Nanjing University of Science and Technology,  
China  
Alwin Poulse,  
Indian Institute of Science Education and  
Research, India  
Kuldeep Singh,  
Guru Nanak Dev University, India  
Çağlar Uyulan,  
Izmir Katip Celebi University, Türkiye

## \*CORRESPONDENCE

Shouliang Qi  
✉ qisl@bmie.neu.edu.cn

RECEIVED 30 March 2025

ACCEPTED 26 August 2025

PUBLISHED 24 September 2025

## CITATION

Li Y, Yang X, Zhao M, Wang J, Yao Y, Qian W  
and Qi S (2025) Predicting depression by  
using a novel deep learning model and  
video-audio-text multimodal data.  
*Front. Psychiatry* 16:1602650.  
doi: 10.3389/fpsyt.2025.1602650

## COPYRIGHT

© 2025 Li, Yang, Zhao, Wang, Yao, Qian and  
Qi. This is an open-access article distributed  
under the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other forums  
is permitted, provided the original author(s)  
and the copyright owner(s) are credited and  
that the original publication in this journal is  
cited, in accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Predicting depression by using a novel deep learning model and video-audio-text multimodal data

Yifu Li <sup>1,2</sup>, Xueping Yang <sup>3</sup>, Meng Zhao <sup>1,2</sup>,  
Jiangtao Wang <sup>1,2</sup>, Yudong Yao <sup>4</sup>, Wei Qian <sup>1</sup>  
and Shouliang Qi <sup>1,2\*</sup>

<sup>1</sup>College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, <sup>2</sup>Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China, <sup>3</sup>Department of Psychology, The People's Hospital of Liaoning Province, Shenyang, China, <sup>4</sup>Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, United States

**Objective:** Depression is a prevalent mental health disorder affecting millions of people. Traditional diagnostic methods primarily rely on self-reported questionnaires and clinical interviews, which can be subjective and vary significantly between individuals. This paper introduces the Integrative Multimodal Depression Detection Network (IMDD-Net), a novel deep-learning framework designed to enhance the accuracy of depression evaluation by leveraging both local and global features from video, audio, and text cues.

**Methods:** The IMDD-Net integrates these multimodal data streams using the Kronecker product for multimodal fusion, facilitating deep interactions between modalities. Within the audio modality, Mel Frequency Cepstrum Coefficient (MFCC) and extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features capture local and global acoustic properties, respectively. For video data, the TimeSformer network extracts both fine-grained and broad temporal features, while the text modality utilizes a pre-trained BERT model to obtain comprehensive contextual information. The IMDD-Net's architecture effectively combines these diverse data types to provide a holistic analysis of depressive symptoms.

**Results:** Experimental results on the AVEC 2014 dataset demonstrate that the IMDD-Net achieves state-of-the-art performance in predicting Beck Depression Inventory-II (BDI-II) scores, with a Root Mean Square Error (RMSE) of 7.55 and a Mean Absolute Error (MAE) of 5.75. A classification to identify potential depression subjects can achieve an accuracy of 0.79.

**Conclusion:** These results underscore the robustness and precision of the IMDD-Net, highlighting the importance of integrating local and global features across multiple modalities for accurate depression prediction.

## KEYWORDS

deep learning, depression, multimedia, information fusion, local and global features

## 1 Introduction

Depression is the primary mental health disorder contributing to the disease burden and it impacts roughly 300 million individuals worldwide (1). Depression is a widespread, costly, and debilitating condition that significantly increases the risk of suicide (2). More than 80% of individuals fail to receive appropriate treatment due to the lack of early intervention services and treatments for depression and researchers estimate that approximately one in five individuals will experience depression at some point in their lifetime (3). Consequently, it is evident that the diagnosis and screening of depression are essential.

Current approaches to evaluating depression rely predominantly on the verbal accounts provided by patients, their families, or caregivers, whether through clinical interviews or questionnaires (4). However, these traditional methods have certain limitations because the subjectivity of individuals can affect responses to questions, and symptoms of depression may manifest differently across individuals (5). Traditionally, accurate diagnosis of depression severity requires comprehensive information and extensive clinical training (6). Fortunately, advanced computing methods, such as machine learning, deep learning, and artificial intelligence, are ideally suited to enhance the assessment of mental health outcomes for individuals (7).

Utilizing audio and video methods for detecting depression offers distinct advantages, including the ability to capture direct cues and subtle behavioral changes that may not be evident in traditional assessments, while numerous indicators are used to detect depression, such as hormonal imbalances (8), changes in sleep patterns (9), cognitive performance assessments (10), resting-state functional magnetic resonance imaging (fMRI) data (11), EEG data (12, 13) and other physiological data (14). In recent years, a variety of automatic depression estimation (ADE) systems have emerged (7, 15, 16). These systems automatically assess the severity of depression using audiovisual cues, employing advanced

techniques from machine learning and deep learning (17). Research indicates that speech exhibits numerous unique characteristics that can be used to identify an individual's mental state (18–20). With the aid of various gestures involving the eyes, mouth, nose, and hands, emotions such as anger, happiness, sadness, and neutrality can be identified through depression detection systems that utilize image and video processing (21). Similarly, textual information can also be analyzed to extract features relevant to depression (22). Similarly, textual information can also be analyzed to extract features relevant to depression (22, 23). And many researchers also have explored diagnosing depression through social networks by textual information (24–26). While unimodal approaches can be effective to some extent in detecting depression, multimodal or hybrid modalities often exhibit superior performance. However, the selection of different modalities, the choice of network architectures, and the methods of fusion all significantly impact the effectiveness of depression detection (17). Many depression detection networks utilize multimodal architectures, yet few of these systems effectively incorporate both local and global features within various modalities.

Consequently, this study introduces a novel deep learning network architecture for multimodal depression detection (Figure 1), termed the Integrative Multimodal Depression Detection Network (IMDD-Net). This advanced framework not only integrates data from video, text, and audio modalities but also considers both local and global information within each modality. By doing so, the IMDD-Net enhances the estimation efficacy by capturing a more comprehensive representation of depressive symptoms. Specifically, within the audio modality, we preprocessed the audio signals and extracted both time-frame level and global statistical features. Both sets of features were then fed into the IMDD-Net through separate channels. For video data, we sampled the video at regular intervals and processed each frame. Then we utilized the specific network to extract both global and local features. In the text modality, we employed a pre-trained Bert-

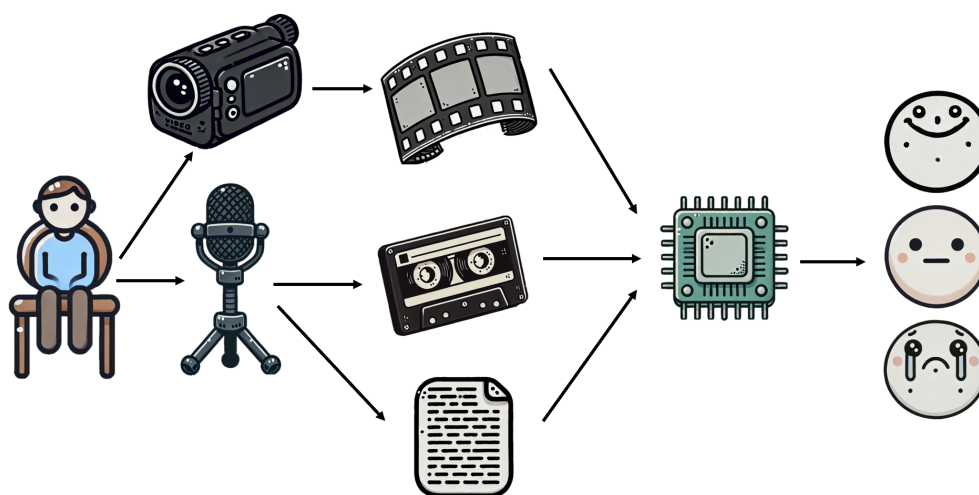


FIGURE 1  
Brief diagram of multimodal depression detection by IMDD.

base-german-cased (27) network. The high-dimensional features obtained from these three modalities were fused. This IMDD-Net was evaluated by using the AVEC2014 dataset (28), demonstrating the effectiveness of our approach in the practical assessment of depression.

The primary innovations and contributions of this paper are summarized as follows:

1. A novel multimodal network architecture has been proposed for the identification of depression;
2. This network effectively integrates and analyzes both global and local features across multiple modalities, including text, video, and audio;
3. Using the Kronecker product for multimodal fusion explores deep interactions between modalities and enhances the detection accuracy of depression.
4. The IMDD-Net achieves state-of-the-art performance in depression assessment areas.

The remainder of this paper is organized as follows. Section 2 reviews the related work on automatic depression estimation. Section 3 describes the dataset used in this study, details the data preparation and preprocessing steps, and provides a comprehensive explanation of the proposed IMDD-Net architecture. Section 4 presents the experimental results obtained using IMDD-Net and provides an analysis of the model's performance across regression and classification tasks. Section 5 discusses the implications of our findings, outlines current limitations, and considers directions for future improvement and clinical applicability. Section 6 concludes the paper and summarizes the main contributions of this work.

## 2 Related works

Recent years have witnessed significant advancements in the field of automatic depression estimation through deep learning. In 2021, Dong et al. (19) developed an automatic depression estimation method using speech signals. This approach combines deep speaker recognition and speech emotion recognition features from pre-trained models to utilize complementary vocal and emotional data. In 2020, Li et al. (29) introduced DRR\_DepressionNet to predict depression severity from facial expressions. This method enhances facial images to enlarge the training dataset and uses a modified ResNet divided into C\_M block, Resblock, and global average pooling. It employs Euclidean loss instead of traditional cross-entropy loss for training. Compared to static images, videos often contain more information and Uddin et al. (30) introduced a two-stream deep spatiotemporal network to assess depression levels from video data. This framework employs the Inception-ResNet-v2 for spatial data and a volume local directional number descriptor for facial motion analysis, enhanced by convolutional neural network (CNN) processing. It also features a multilayer bidirectional long short-term memory (Bi-LSTM) with temporal median pooling to

integrate spatial and temporal features effectively. Then, He et al. (31) introduced an end-to-end trainable system for depression detection, utilizing a 3D CNN with a spatiotemporal feature aggregation module. This system utilizes a 3D DEP-NetVLAD aggregation method to effectively identify depression. In 2023, Rajawat et al. (32) introduced a fusion fuzzy logic model combined with deep learning to identify depression through facial expressions from image and video files. Their model uses a fuzzy algorithm and unordered fuzzy rule initiation for depression recognition, transforming facial expressions into detectable indicators of depression.

Furthermore, many multimodal networks based on audio and video have also achieved excellent results in the field of depression detection. Niu et al. (33) introduced a novel Spatio-Temporal Attention network combined with a Multimodal Attention Feature Fusion strategy for predicting depression levels by capturing multimodal cues in 2020. In addition, Sun et al. (34) developed a multi-modal adaptive fusion transformer network and it is tailored to extract long-term context from uni-modal data. This network employs an adaptive fusion technique to integrate multimodal features effectively. Bucur et al. (35) introduced a unique time-enriched multimodal transformer architecture that leverages pre-trained models to extract image and text embeddings from social media posts in 2023. Operating at the user level, their model integrates time2vec positional embeddings to account for the timing of posts. Furthermore, they developed a variant designed to handle randomly sampled, unordered post sets, thereby increasing robustness against dataset noise. Li et al. (36) proposed a Decoupled Multimodal Distillation (DMD) framework to address modality heterogeneity in emotion recognition by separating each modality's representation into modality-exclusive and modality-irrelevant components. A graph-based distillation unit (GD-Unit) enables dynamic, adaptive knowledge transfer between modalities via learned edge weights. This flexible structure improves feature discrimination and crossmodal alignment, achieving superior performance on standard multimodal emotion recognition benchmarks.

In summary, recent advancements in the field of automatic depression estimation have showcased the potential of deep learning and multimodal approaches in improving the accuracy and reliability of depression detection. Various methods utilizing speech signals, facial expressions, and video data have been developed, each contributing unique strengths and innovations. As the field continues to evolve, future research should focus on refining these models, addressing their limitations, and improve the effectiveness of depression detection.

## 3 Materials and methods

### 3.1 Dataset and preprocessing methods

This chapter provides an overview of the datasets utilized in this study, as well as the preprocessing methods applied to the raw data, encompassing audio, video, and text modalities.

### 3.1.1 AVEC 2014 dataset

In this study, we utilize the Audio-Visual Emotion Recognition Challenge (AVEC) 2014 dataset (28). This dataset is among the few that offer unprocessed audio and video data, which are critical for analyzing nuanced behavioral cues and expressions associated with depression. The AVEC 2014 dataset includes 150 German participants (96 female and 54 male) and the subjects had a mean age of 31.5 years, with a standard deviation of 12.3 years, ranging from 18 to 63 years. Each of them completed two tasks to generate differentiated audiovisual data. The tasks selected are as follows: (1) Northwind dataset: Participants recite a passage from the fable “Die Sonne und der Wind” (The North Wind and the Sun) in German. (2) Freeform dataset: Participants express themselves spontaneously by answering various prompts, such as: “What is your favorite dish?”; “What was your best gift, and why?”; or “Discuss a sad childhood memory.” All responses are answered in German. Therefore, these tasks have been specifically selected to enable a comprehensive analysis of both verbal and nonverbal cues associated with depression. Additionally, each participant’s depression severity is quantified using the Beck Depression Inventory-II (BDI-II) (37) scores, which serve as labels for the dataset. According to standard interpretation guidelines, BDI-II scores can be categorized as follows: 0–13 indicates minimal depression, 14–19 mild depression, 20–28 moderate depression, and 29–63 severe depression.

### 3.1.2 Video preprocessing

In this section, we detail the preprocessing steps applied to the video data of the AVEC 2014 dataset, and the process is depicted in Figure 2.

First, we perform sampling on the raw videos. Since the original videos vary in length, we employ an adaptive sampling interval approach. This method calculates the sampling interval based on the length of the video and the number of frames, ensuring that each video yields exactly one hundred frames after sampling.

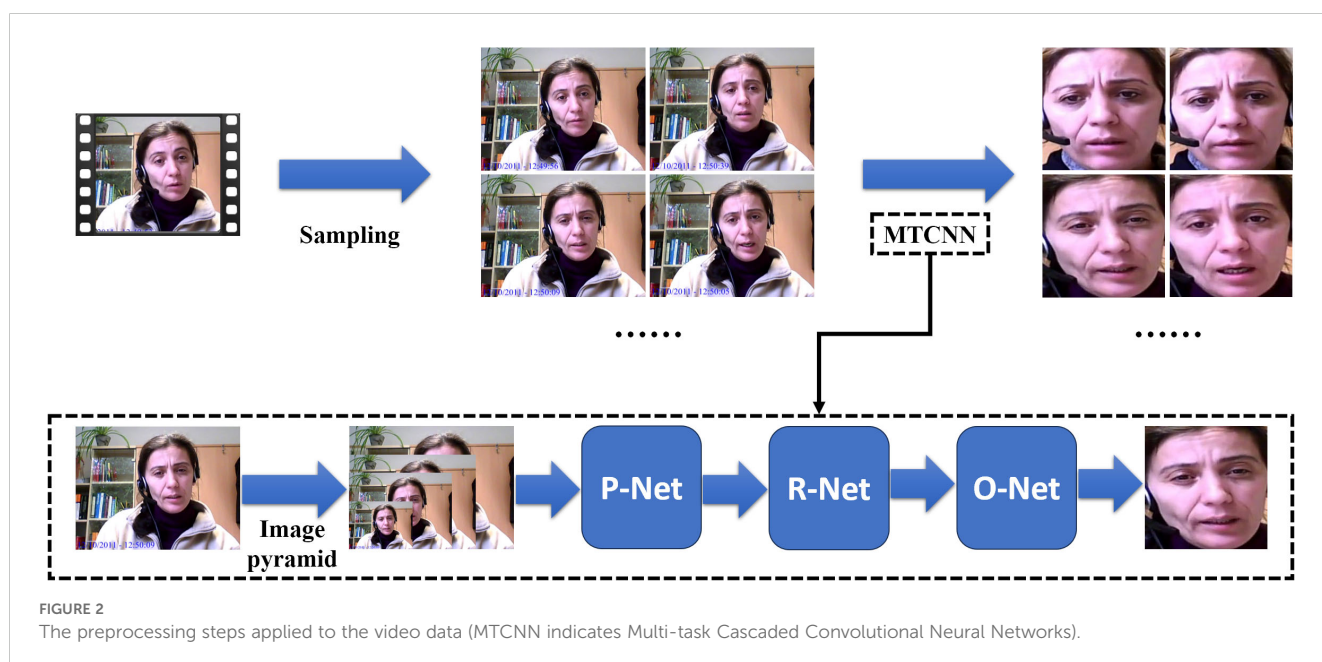
To minimize the interference from the background and other irrelevant information, each frame is processed through Multi-task Cascaded Convolutional Neural Networks (MTCNN) (38). First, MTCNN employs an image pyramid (39) to handle various face sizes within the frame, allowing it to perform detection across multiple scales. Then, the processed images are sent to Proposal Network (P-Net). It generates candidate facial regions by rapidly scanning the resized images, proposing potential areas that likely contain faces. After that, the outputs of the P-Net are sent to the Refine Network (R-Net) and this stage refines these candidates, filtering out false positives. Finally, the outputs of R-Net are sent to the Output Network and this stage also provides final bounding boxes and associated confidence scores to confirm the presence of facial features and output the final processed frames.

Following this preprocessing procedure, we have successfully processed the video data for 150 participants, with each set consisting of 100 frames.

### 3.1.3 Audio preprocessing

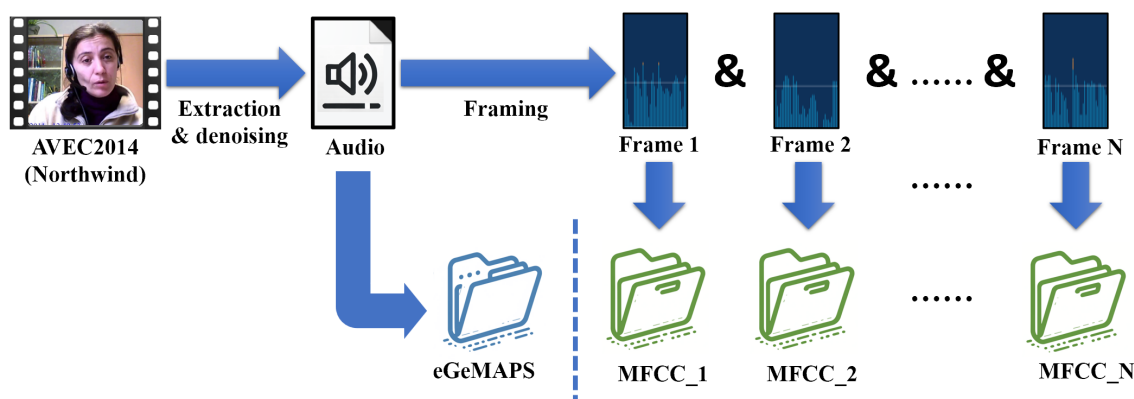
The processing of audio data in this study focuses on both global and local features, utilizing eGeMAPS (40) and MFCC (41) respectively. This section will detail the preprocessing steps used to extract these features, ensuring a comprehensive analysis for depression detection.

To control variables that might affect experimental outcomes, the extraction of audio features employed the Northwind dataset from AVEC2014. This ensures consistency in the content spoken by all participants. After extracting audio from original video files, the audio undergoes a noise reduction process. This involves spectral gating and wavelet denoising to remove noise while preserving essential aspects of the voice. Following the denoising, MFCC and eGeMAPS are extracted and the overall process is shown in the upper part of Figure 3.





### Audio preprocessing



### Text preprocessing

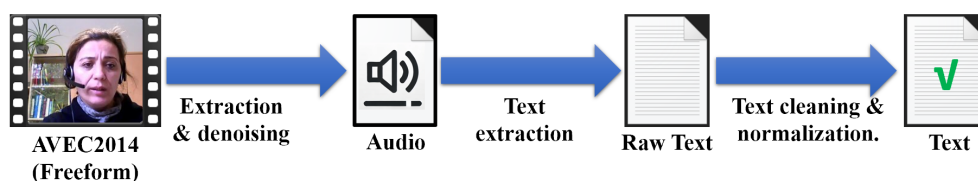


FIGURE 3  
The preprocessing steps applied to the audio data and text data.

MFCCs are widely used in speech and audio processing and they play a pivotal role in capturing the local features of audio for our analysis. MFCCs are derived from the Fourier transform of a signal, with a focus on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the standard cepstral analysis. The calculation of MFCCs is as Equation 1:

$$f_{mel} = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \quad (1)$$

Initially, each audio track is segmented into 1,000 frames and for each frame, we extract the first 20 MFCCs. To further enhance the descriptive power of our features, we compute both the first and second derivatives of these coefficients, effectively capturing the dynamic changes in the cepstral features over time. As a result, each frame is represented by 60 MFCC features (20 coefficients plus their first and second derivatives), culminating in a 1000 \* 60 feature matrix for each audio sample. This feature set allows us to analyze the local characteristics of the speech.

Meanwhile, we utilize the eGeMAPS (40) to represent the global features of speech. Unlike the MFCCs which focus on capturing fine-grained local properties of audio, eGeMAPS is designed to encapsulate the overall statistical characteristics of speech. This feature set is particularly effective for its comprehensive coverage of voice attributes that are commonly implicated in emotional states and psychological conditions.

eGeMAPS is an acoustic parameter set built on the Geneva Minimalistic Acoustic Parameter Set (GeMAPS). eGeMAPS

consists of 88 features derived from basic acoustic descriptors (Low-Level Descriptors, LLDs) through various statistical methods, and these statistical features are known as High-Level Statistics (HSFs). The features cover multiple acoustic aspects including frequency, energy, and spectral properties. Each audio sample in our dataset is processed to extract a complete set of these eGeMAPS features, ultimately transforming the entire speech signal into a single 1 \* 88 vector. This representation captures the global attributes of the speech.

#### 3.1.4 Text preprocessing

To capture the diverse semantic information expressed by participants during interviews, our study begins by extracting textual data from the AVEC 2014 Freeform dataset, and the process is shown in the lower part of Figure 3. The first step involves extracting audio from the video recordings, followed by a noise reduction process to ensure the clarity of human voices. Several preprocessing techniques are employed here, including spectral gating to reduce background noise and dynamic range compression to maintain a consistent audio level.

Once the audio is cleaned, it is converted into German text. Following transcription, the German text undergoes text cleaning and normalization. This involves transforming all text to lowercase, stripping away punctuation and special characters, and discarding stopwords that do not add significant value to the depression detection.

Thus, we have thoroughly and meticulously preprocessed the data for 150 participants, obtaining refined final datasets across

three modalities: audio, video, and text. This comprehensive preparation ensures data well-suited for the subsequent stages of our analysis.

## 3.2 Framework of IMDD-Net

This chapter is dedicated to detailing the architecture of the IMDD-Net, a multimodal network designed for assessing depression. The IMDD-Net integrates audio, text, and video modalities, using both local and global information within each modality to enhance estimation accuracy. The network processes these modalities through four specialized channels and a multimodal fusion and inference process, culminating in the output of BDI-II scores to evaluate the severity of depressive disorders. The architecture of the network is illustrated in Figure 4. In the following sections, the composition and functionalities of the IMDD-Net's architecture will be introduced.

### 3.2.1 Video feature extraction subnetwork

In order to capture both local and global features from video data effectively, the TimeSformer (42) network has been selected for the video modality channel of the IMDD-Net. The TimeSformer builds upon the Vision Transformer (ViT) (43) by integrating temporal processing capabilities, allowing it to effectively address the dynamic nature of video data. Specifically, the TimeSformer extends the Vision Transformer architecture by incorporating temporal dimensions into the self-attention mechanism. This enables the model to process sequences of video frames rather than static images.

Initially, each frame ( $64 \times 64$  pixels) of the video is divided into patches of size  $8 \times 8$  pixels. These patches are then linearly embedded into 512-dimensional vectors. After patch embedding, each patch is concatenated with learned spatial (position within a frame) and temporal (position within the sequence) embeddings. Space and time embeddings enhance the model's ability to interpret the positional context of each patch both within individual frames and across the sequence. The embedded patches are processed through multiple layers of spatial and temporal self-attention mechanisms. Each layer consists of 12 blocks (depth=12), where both local interactions within frames and global interactions across frames are captured (Figure 5). Each attention block is followed by a multilayer perceptron (MLP) and the output from the final layer of the TimeSformer is passed through a linear layer that shapes the output into a high-dimensional feature vector of size  $5 \times 1$  per video. This vector represents the key features extracted and processed from the video which contain both local and global features.

### 3.2.2 Textual feature extraction subnetwork

In the text modality channel of the IMDD-Net, we leverage the capabilities of the Bidirectional Encoder Representations from Transformers (BERT) (27) to analyze textual data. BERT is particularly well-suited for this task due to its ability to capture both local and global contextual information from text. Unlike directional models, which read the text input sequentially (left-to-right or right-to-left), the BERT model reads the entire sequence of words at once. This characteristic allows BERT to capture the meaning of a specific word in different contexts, providing a deeper understanding of the text as a whole.

For our study, we utilize the German variant of BERT, specifically the Bert-base-german-cased model, which has been

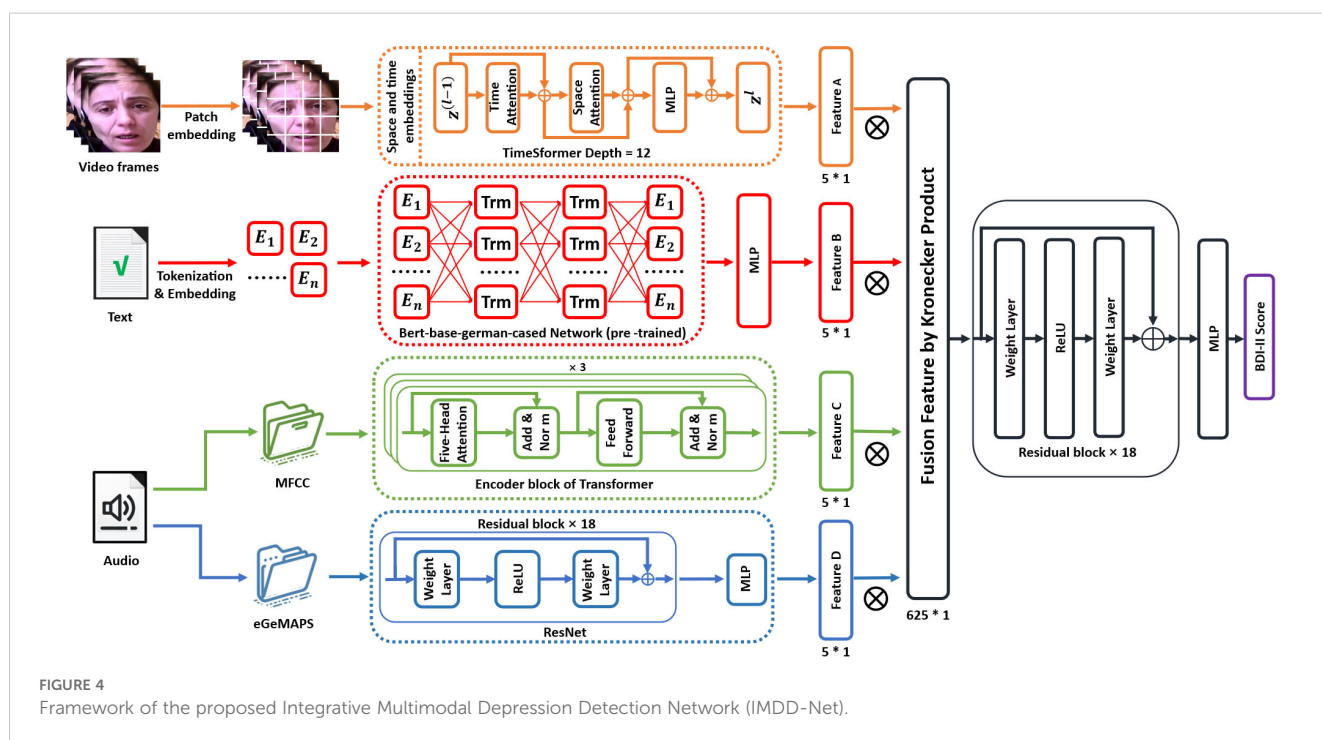




FIGURE 5  
Local & global features extracted by space & time attention of TimeSformer.

pre-trained on a large corpus of German text. The model features an embedding layer that converts tokenized text into numerical vectors. Text data is initially processed using the BertTokenizer, and adjusts the length of these sentences to a uniform 350 tokens by padding. The subsequent transformer blocks process the input using self-attention mechanisms, followed by an MLP layer at the end to transform the token embeddings into a single vector.

The final output is a high-dimensional feature vector sized  $5 \times 1$  for each piece of text, encapsulating both local and global contextual information.

### 3.2.3 Audio feature extraction subnetwork

The audio modality channel in the IMDD-Net processes audio data to capture both local and global features.

**Local Feature Extraction with MFCC:** The audio files are first processed to extract MFCC, which focus on local information at the frame level. Each audio file is segmented into frames, and for each frame, 60 MFCC features are computed, resulting in a feature matrix of size  $60 \times 1000$ . These MFCC features are then processed through a Transformer model with 8 attention heads and a hidden dimension of 64. The Transformer encoder is composed of 3 layers and the  $5 \times 1$  feature vector representing local speech information is outputted.

**Global Feature Extraction with eGeMAPS:** In parallel, eGeMAPS features are extracted to capture the global acoustic properties of the audio file and the size of input is  $88 \times 1$ . The eGeMAPS features are processed using a network structured similar to ResNet, which is adapted here for one-dimensional audio signal processing. This network is composed of 18 residual blocks and each block includes layers of linear transformations with skip connections. After processing through an MLP, the  $5 \times 1$  high-dimensional feature vector representing global audio information is outputted.

At this point, we have obtained two vectors that respectively represent the global and local features of the audio information.

### 3.2.4 Multimodal feature fusion and inference

In this final phase of the IMDD-Net, we converge the extracted features from different modalities to form a unified representation, leveraging the power of multimodal data fusion to enhance estimation precision.

The use of the Kronecker product in our fusion process allows for a detailed interaction between features from different modalities. Unlike simpler fusion techniques such as concatenation or averaging, the Kronecker product facilitates a richer and more expressive combination by mathematically intertwining the feature sets, thus capturing both inter-modal and intra-modal dependencies.

The Kronecker product, denoted by  $\otimes$ , is a mathematical operation on two matrices. For matrices  $A$  of size  $m \times n$  and  $B$  of size  $p \times q$ , the Kronecker product  $A \otimes B$  is a block matrix of size  $m \times n$ . The operation is defined as Equation 2:

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \dots & \dots & \dots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix} \quad (2)$$

where each element  $a_{ij}$  of matrix  $A$  is multiplied by matrix  $B$ .

In the context of our IMDD-Net, the  $5 \times 1$  feature vectors from each modality (representing both global and local information) undergo a transformation through the Kronecker product, resulting in a combined feature matrix. When these  $5 \times 1$  vectors from the three modalities (audio, video, text) are subjected to the Kronecker product sequentially, the dimensionality of the resulting feature matrix expands to  $625 \times 1$ . This expansion not only increases the feature space but also preserves the unique characteristics of each modality.

The  $625 \times 1$  feature vector is then processed through a deep residual network consisting of 18 layers and the output from the ResNet is a predictive value that correlates with the Beck Depression Inventory-II (BDI-II) scores.

In summary, the architecture of the IMDD-Net is designed for multimodal data. By integrating and processing local and global features from video, audio, and text inputs through advanced models, the system improves the accuracy of depression estimation and provides new insights for detecting depression.

### 3.3 Training methodology

The training and validation strategy of the IMDD-Net is designed to ensure a robust and comprehensive assessment of the model's performance.

The dataset consists of data from 150 participants. To validate the effectiveness and stability of the model, five-fold cross-validation is employed. The dataset is divided into five equal parts randomly. In each fold, one part is held out as a validation set while the other four parts are used for training. This process is repeated five times, with each of the five parts used exactly once as the validation set.

The Huber loss function is selected because it is an error metric that combines the best aspects of L1 norm (mean absolute error) and L2 norm (mean squared error) loss functions, making it particularly effective for regression problems.

The Huber loss is defined by the Equation 3:

$$L_{\delta}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{for } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & \text{for } |y - \hat{y}| > \delta \end{cases} \quad (3)$$

where  $y$  represents the true value,  $\hat{y}$  is the predicted value and  $\delta$  is a threshold parameter that defines the boundary between the quadratic loss and the linear loss. This threshold is adjusted as 1 ( $\delta = 1$ ) in our experiment.

The optimizer used is AdamW (Initial learning rate = 0.001), which combines the advantages of AdaGrad and RMSProp optimization methods and includes weight decay (weight\_decay=0.01) regularization to prevent overfitting.

To evaluate the performance of our model, we adopt Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as

primary evaluation metrics. These two measures are widely used in the field of automatic depression estimation, particularly on the AVEC 2014 dataset, allowing for direct and fair comparisons with prior state-of-the-art methods. MAE provides a straightforward interpretation of average error, while RMSE penalizes larger deviations more heavily, offering insight into prediction stability. MAE and RMSE which are defined as Equations 4 and 5:

$$RMSE = \sqrt{\frac{1}{N} \sum (r_i - r'_i)^2} \quad (4)$$

$$MAE = \frac{1}{N} \sum |r_i - r'_i| \quad (5)$$

where  $N$  is the total number of observations,  $r_i$  is the prediction from the model, and  $r'_i$  is the actual observed value.

The model was trained for 300 epochs, and Figure 6 presents the Huber loss, MAE, and RMSE curves averaged across the five cross-validation folds.

## 4 Results

### 4.1 Experiments result of IMDD-Net

In this section, we present the experimental results of the IMDD-Net, comparing its performance against state-of-the-art (SOTA) methods and the baseline from the AVEC 2014 challenge. The evaluation metrics used to assess the performance are RMSE and MAE.

As shown in Table 1, our model has achieved state-of-the-art performance with an RMSE of 7.55 and an MAE of 5.75. The AVEC 2014 baseline has an RMSE of 9.89 and an MAE of 7.89 (28). Cai et al. (44) proposed an end-to-end time-domain channel attention network (TDCA-Net) for depression detection and on the AVEC 2014 dataset, their model achieved an RMSE of 8.90 and an MAE of 7.08. Additionally, Dong et al. (19) proposed a hierarchical depression detection model combining deep speaker recognition and speech

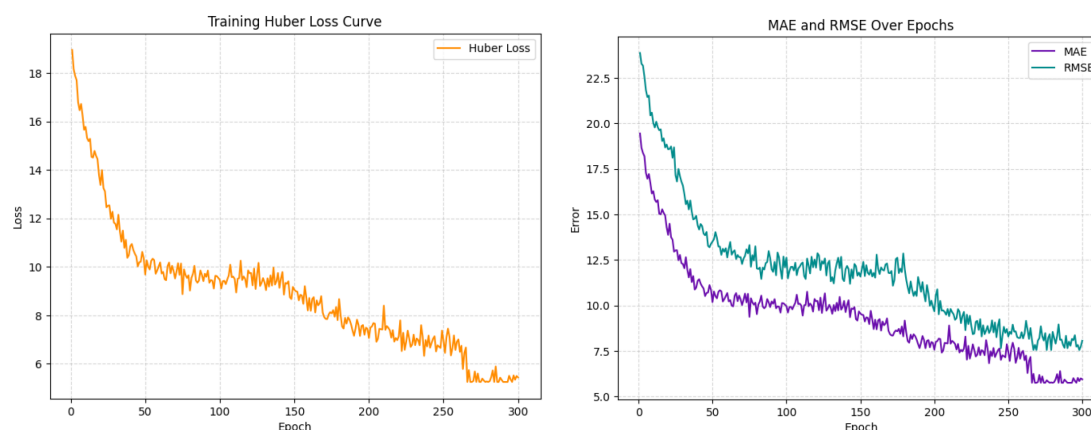


FIGURE 6  
Averaged training curves of Huber loss, MAE, and RMSE over 300 epochs across five cross-validation folds.



TABLE 1 The comparison of different methods and their structures on the AVEC 2014 dataset.

| Method                  | Network structure   | Modality         | RMSE↓       | MAE↓        |
|-------------------------|---|------------------|-------------|-------------|
| AVEC 2014 baseline (28) | Support Vector Regression   | Audio-Video      | 9.89        | 7.89        |
| (44)                    | An end-to-end time-domain channel attention network (TDCA-Net)  | Audio            | 8.90        | 7.08        |
| (19)                    | A hierarchical model combining deep speaker recognition and speech emotion recognition features                             | Audio            | 8.73        | 7.32        |
| (30)                    | A two-stream deep spatiotemporal network  | Video            | 8.78        | 6.96        |
| (31)                    | An end-to-end trainable system utilizing a 3D convolutional neural network with a spatiotemporal feature aggregation module | Audio-Video      | 8.42        | 6.78        |
| (45)                    | Local Quaternion and Global Deep Network  | Video            | 7.84        | 6.08        |
| (46)                    | Maximization and Differentiation Network  | Video            | 7.90        | 6.19        |
| (47)                    | A multi-scale and multi-region facial dynamic representation method   | Video            | 7.98        | 6.14        |
| (48)                    | A two-stream model with a novel temporal pooling method for capturing spatio-temporal dynamics in video clips               | Video            | 7.94        | 6.20        |
| (49)                    | Spatial-Temporal Attention Depression Recognition Network   | Video            | 7.75        | 6.00        |
| (50)                    | Using multi-scale video-level features and novel spectral representations   | Video            | <b>7.15</b> | 5.95        |
| <b>Ours</b>             | Integrative Multimodal Depression Detection Network by using local and global multimodality features                        | Audio-Video-Text | 7.55        | <b>5.75</b> |

The symbol "↓" indicates that lower values represent better performance (as for MSE and RMSE). Bold values denote the best performance across methods.

emotion recognition features, achieving an RMSE of 8.73 and an MAE of 7.32 on the AVEC 2014 dataset. Moreover, Uddin et al. (30) developed a two-stream deep spatiotemporal network for depression level estimation, resulting in an RMSE of 8.78 and an MAE of 6.86. Furthermore, He et al. (31) introduced an end-to-end trainable intelligent system utilizing a 3D convolutional neural network with a spatiotemporal feature aggregation module achieving an RMSE of 8.42 and an MAE of 6.78 on the AVEC 2014 dataset and Shang et al. (45) proposed a method called Local Quaternion and Global Deep Network, which integrates local quaternion and global deep features for facial depression recognition, achieving an RMSE of 7.84 and an MAE of 6.08. Moreover, Melo et al. (46) introduced the Maximization and Differentiation Network to represent facial expression variations relevant for depression assessment, achieving an RMSE of 7.90 and an MAE of 6.19 on the AVEC 2014 dataset. Additionally, Niu et al. (47) proposed a multi-scale and multi-region facial dynamic representation method for depression prediction, achieving an RMSE of 7.98 and an MAE of 6.14 and Melo et al. (48) proposed a two-stream model with a novel temporal pooling method for capturing spatio-temporal dynamics in video clips, achieving an RMSE of 7.94 and an MAE of 6.20 on the AVEC 2014 dataset. Pan et al. (49) proposed the Spatial-Temporal Attention Depression Recognition Network, which enhances feature extraction by capturing global and local spatial-temporal information, achieving an RMSE of 7.75 and an MAE of 6.00 and Song et al. (50) proposed a method for video-based automatic depression analysis using multi-scale video-level features and novel spectral representations, achieving an RMSE of 7.15 and an MAE of 5.95 on the AVEC 2014 dataset.

In summary, our model has achieved state-of-the-art performance, demonstrating its effectiveness in depression detection. Notably, our model achieved the lowest MAE among

all compared models, with a score of 5.75, indicating its superior accuracy in predicting depression severity. Furthermore, our model attained the second-lowest RMSE, with a score of 7.55, closely following the RMSE of 7.15 achieved by Song et al. [38]. These results underscore the robustness and precision of our IMDD-Net in integrating multimodal data and global and local features to enhance depression detection.

In our study, we generated a BDI-II value comparison bar plot and an error histogram to evaluate the prediction performance of the IMDD-Net visually (Figure 7). Figure 7A presents a bar plot comparing the real and predicted BDI-II values for each sample. The x-axis corresponds to the sample indices, and the y-axis represents the BDI-II values. Each sample is represented by two bars: one for the real value and one for the predicted value. The red dashed line in the figure represents the threshold at a BDI-II score of 13. Typically, BDI-II scores below 13 are considered indicative of no depression, while scores above 13 suggest the presence of depressive symptoms. Analysis of Figure 7A reveals that the IMDD-Net tends to underestimate the actual BDI-II scores of individuals with depression (BDI-II scores greater than 13, 43 of 73 are underestimated), while overestimating the BDI-II scores of those without depression (BDI-II scores less than 13, 45 of 77 are overestimated).

As shown in Figure 7B, the error histogram illustrates the distribution of the prediction errors, which are calculated as the difference between the real and predicted BDI-II values. The histogram provides a visual representation of how closely the predicted values align with the actual values. The x-axis represents the prediction error, while the y-axis represents the frequency of each error value. The majority of the errors are concentrated around zero and 57.33% sample is located in the

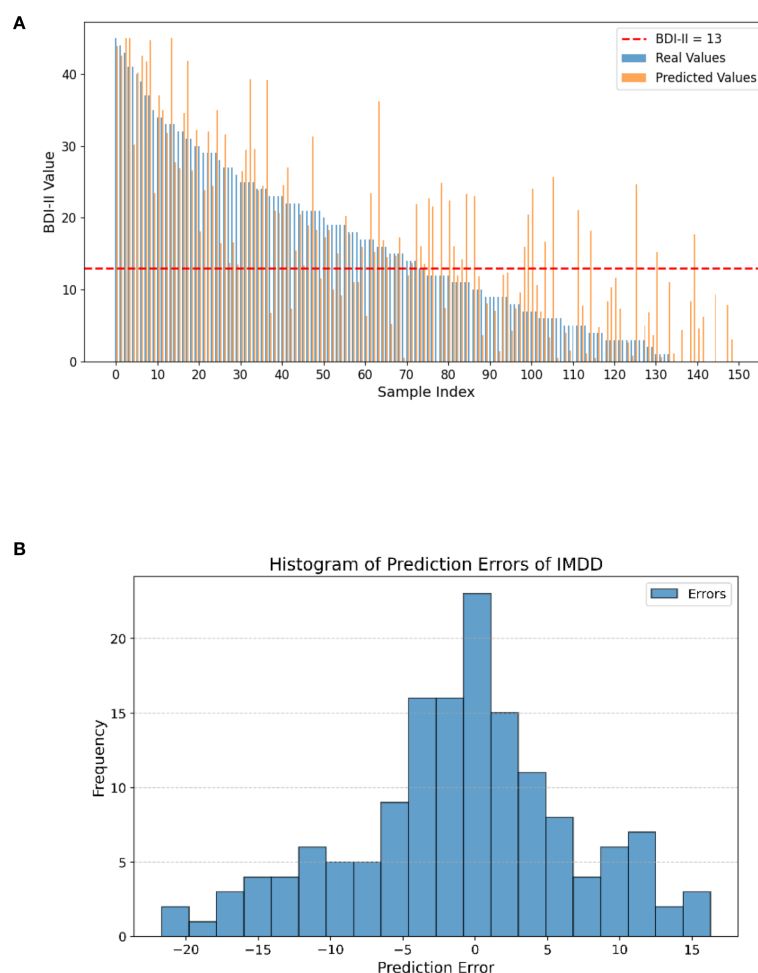


FIGURE 7

Differences between real and predicted BDI-II values and histogram of prediction errors of IMDD-Net. (A) The comparison of real and predicted BDI-II values; (B) Histogram of prediction errors.

error range of  $\pm 5$ , indicating that the IMDD-Net's predictions are generally accurate.

To further illustrate the experimental results, we conducted a Brand-Altman analysis and performed a regression analysis. The Brand-Altman scatter plot and the regression plot are presented in

Figure 8 to illustrate the agreement and predictive accuracy of the IMDD-Net.

The Brand-Altman plot (the left part of Figure 8) compares the differences between the predicted BDI-II scores and the actual BDI-II scores against their averages. The mean difference (bias) is

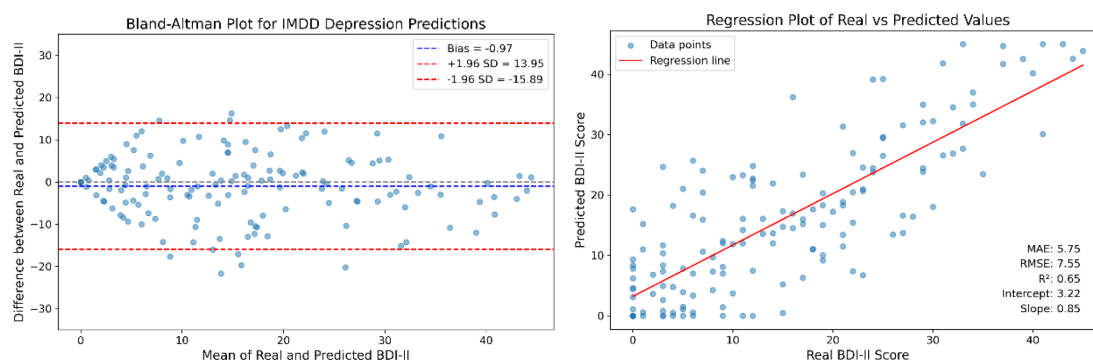


FIGURE 8

The Brand-Altman plot and regression analysis for IMDD-Net predictions.

calculated to be  $-0.97$  with a standard deviation of  $7.61$ . The limits of agreement, defined as the mean difference  $\pm 1.96$  times the standard deviation, range from  $-15.89$  to  $13.95$ .

The regression analysis (the right part of Figure 8) shows the relationship between the predicted BDI-II scores and the actual BDI-II scores. The regression equation is given by: Predicted BDI-II =  $0.85 \times$  Actual BDI-II +  $3.22$ . The coefficient of determination ( $R^2$ ) is calculated to be  $0.65$ , indicating that approximately 65% of the variance in the actual BDI-II scores can be explained by the IMDD-Net's predictions.

In summary, the Bland-Altman analysis shows good agreement between the predicted and actual BDI-II scores, while the regression analysis confirms the predictive accuracy of the IMDD-Net. These results collectively highlight the effectiveness of the IMDD-Net in assessing depression severity.

## 4.2 Classification performance analysis

To further validate the effectiveness of IMDD-Net, we performed a classification analysis using a threshold of 13 on the BDI-II scores. Participants with scores above 13 were classified as depressed, while those with scores of 13 or below were classified as non-depressed. The confusion matrix (The left part of Figure 9) provides a detailed breakdown of the classification results, showcasing the true positive (TP) is 62, false positive (FP) is 20, true negative (TN) is 57, and false negative (FN) is 11. Therefore, the performance metrics are calculated as follows: Accuracy (ACC) is approximately 79.3%, Sensitivity (SEN) is 84.9%, Specificity (SPE) is 74.0%, Positive Predictive Value (PPV) is 75.6%, and Negative Predictive Value (NPV) is 83.8% (The right part of Figure 9).

## 4.3 Ablation experiment

To further emphasize the importance of multimodal data in the detection of depression, we conducted an ablation study. In this study, we evaluated the performance of the IMDD-Net using

individual modalities and combinations of two modalities and compared the results with the complete multimodal configuration.

For the ablation study, we considered the following configurations: (a) Audio Only (MFCC and eGeMAPS); (b) Video Only; (c) Text Only; (d) Audio + Video; (e) Audio + Text; (f) Video + Text; (g) IMDD (Audio + Video + Text). Each configuration was used to train and test the model independently, following the same training methodology as described in the previous sections.

The performance of each configuration was evaluated by RMSE and MAE. The results are summarized in Table 2. Among the single modalities, the text modality achieves the best performance with the lowest RMSE ( $7.81$ ) and MAE ( $6.36$ ). For dual-modality combinations, the Video + Text configuration yields the best results with an RMSE of  $7.66$  and an MAE of  $5.84$ . Previous clinical psychology studies have indicated that the relationship between language users (e.g., speakers or writers) and their texts is meaningful and shows considerable promise for the depression detection (51). A study (51) also suggests that it may be possible to identify individuals at risk for depression through text-based analysis, which aligns with the results of our ablation experiments.

The ablation experiments and corresponding statistical analysis results are also shown in Figure 10, where the standard deviation is also included from five-fold cross-validation and paired t-tests are conducted on the RMSE and MAE metrics. The RMSE of the Audio Only and Video Only configuration is significantly higher than that of other configurations ( $p < 0.05$ ). The Text Only configuration presents significantly lower RMSE than other two single modality configurations ( $p < 0.05$ ), indicating that the information underlying the text might be more valuable for the depression detection or easier to be extracted by our IMDD-Net. The MAE of the IMDD (Audio + Video + Text) is significantly lower than that of single modality and Audio + Video configurations ( $p < 0.05$ ).

The ablation study clearly illustrates the critical role of multimodal data integration in the IMDD-Net. Each modality contributes unique and valuable information and enhances the network's ability to detect depression accurately. The superior performance of the complete multimodal configuration

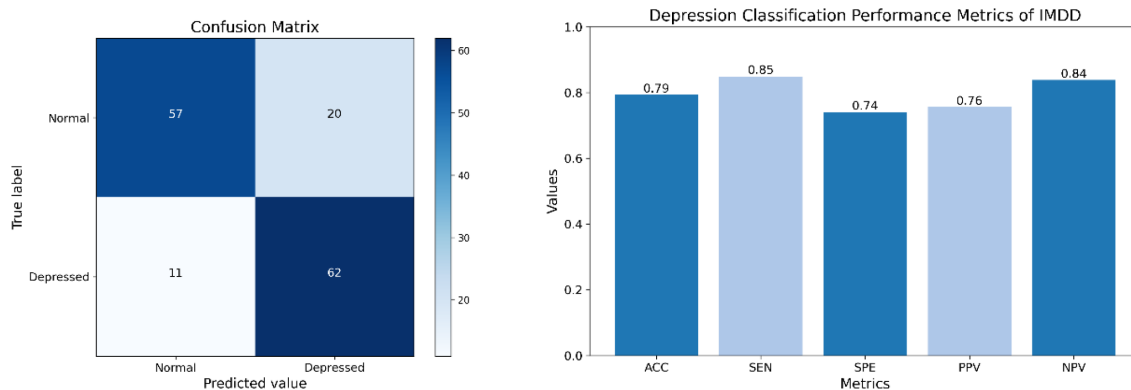


FIGURE 9  
The confusion matrix and classification performance of IMDD-Net.

TABLE 2 The result of the ablation experiment of IMDD-Net.

| Model variant               | RMSE↓       | MAE↓        |
|-----------------------------|-------------|-------------|
| Audio Only                  | 8.49        | 6.88        |
| Video Only                  | 8.27        | 6.63        |
| Text Only                   | 7.81        | 6.36        |
| Audio + Video               | 7.72        | 6.05        |
| Audio + Text                | 7.68        | 5.99        |
| Video + Text                | 7.66        | 5.84        |
| IMDD (Audio + Video + Text) | <b>7.55</b> | <b>5.75</b> |

The symbol “↓” indicates that lower values represent better performance (as for MSE and RMSE).  
Bold values denote the best performance across methods.

underscores the necessity of leveraging diverse data sources in the field of depression detection.

An important experimental discovery in this study is the identification of the critical role played by multimodal data integration, specifically emphasizing the complementary nature of text, audio, and video modalities. Through ablation experiments, we observed that integrating text modality consistently improved performance more than audio or video alone, indicating that linguistic patterns provide highly discriminative signals for depression. This finding is consistent with clinical research, which suggests that linguistic expressions offer reliable and sensitive indicators of depressive states, capturing nuanced cognitive and emotional disturbances that might be less prominently reflected through facial expressions or vocal characteristics alone.

Furthermore, the comparative analysis of prediction errors and Bland-Altman agreement plots revealed specific behavioral patterns in our model’s predictions: the IMDD-Net tended to underestimate depressive severity in participants with higher BDI-II scores, possibly due to subtle or suppressed emotional cues in severely depressed individuals that are difficult to capture comprehensively from video or audio data alone. Conversely, it slightly overestimated scores in non-depressed individuals, suggesting a potential sensitivity toward ambiguous or transient emotional cues. These nuanced error patterns highlight opportunities for further refining modality-specific feature extraction and fusion techniques.

Overall, these experimental insights underscore the necessity of multimodal fusion in depression detection frameworks, particularly emphasizing the distinct contribution and sensitivity of linguistic information. This emphasizes the need for future research to prioritize advanced linguistic feature extraction methods and adaptive multimodal fusion strategies, ultimately improving the clinical utility and interpretability of automated depression detection systems.

5 Discussion

This study introduced IMDD-Net, a multimodal deep learning model designed to enhance depression detection by integrating

video, audio, and text data. Our findings demonstrate that incorporating multiple modalities provides a more comprehensive and robust assessment of depressive symptoms. By leveraging both local and global feature extraction, IMDD-Net effectively captures short-term behavioral cues as well as long-term patterns in facial expressions, vocal tone, and linguistic structures. The model achieved an RMSE of 7.55 and an MAE of 5.75 for depression severity estimation, along with a classification accuracy of 79.3%, demonstrating high sensitivity (84.9%) and specificity (74.0%).

This study underscores the potential of multimodal deep learning in advancing objective, scalable, and accessible mental health assessments. By addressing current challenges, models like IMDD-Net could significantly contribute to the development of AI-driven tools for depression screening and monitoring, bridging the gap between computational psychiatry and clinical practice.

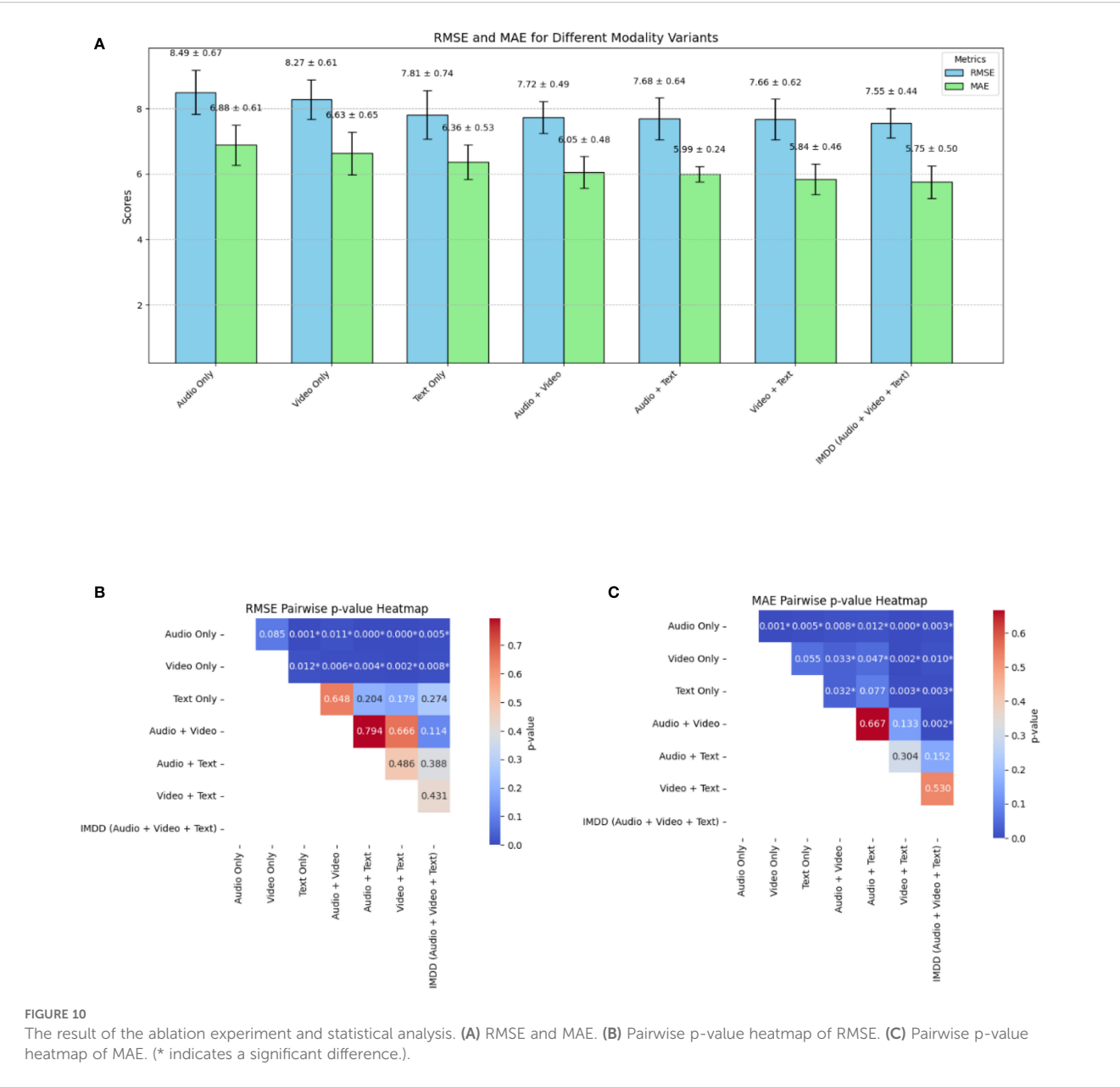
5.1 Effectiveness of multimodal data in depression prediction

The use of multimodal data, particularly the combination of audio, video, and text, has proven to be highly effective in depression prediction. Depression is a disorder that manifests through various behavioral, cognitive, and linguistic patterns (52). Studies have consistently shown that individuals with depression exhibit distinct facial expressions, speech patterns, and language use, which together provide a more comprehensive and reliable means of assessment (53, 54).

Facial expression analysis plays a crucial role in depression detection, as individuals suffering from depression often exhibit reduced facial variability, diminished smiling, and increased negative affect (55). Similarly, speech-based features such as prosody, pitch variation, articulation rate, and speech fluency offer valuable indicators of depressive states (56). Depressed individuals often demonstrate monotonous speech, slower articulation, increased hesitations, and longer response latencies (57), all of which can be extracted using acoustic analysis techniques such as MFCC and eGeMAPS. In addition to audio-visual cues, the linguistic patterns of individuals with depression often reflect negative sentiment, cognitive distortions, self-referential focus, and reduced syntactic complexity (51). By analyzing transcribed speech or written text, natural language processing (NLP) models can identify markers such as higher usage of first-person pronouns, excessive expressions of sadness or hopelessness, and reduced use of complex sentence structures (58).

While unimodal approaches relying on either facial expressions, speech, or text have demonstrated promising results, they often suffer from limited accuracy due to the heterogeneous nature of depressive symptoms (59). For instance, an individual may suppress facial expressions while still exhibiting changes in speech tone and linguistic patterns, or their speech may remain neutral while textual markers indicate distress. By integrating these three modalities, deep learning models can compensate for the weaknesses of each individual modality, leading to more robust and reliable depression prediction.





5.2 The role of global and local features in multimodal fusion

The integration of global and local features in multimodal learning plays a crucial role in capturing the complex manifestations of depression. Depression affects individuals in both momentary expressions and long-term behavioral patterns (60), making it essential for predictive models to analyze information across multiple temporal scales. Local features focus on fine-grained, short-term behavioral signals such as micro-expressions, brief tonal fluctuations in speech, and word-level linguistic markers, which provide immediate but transient indicators of depressive symptoms (61). On the other hand, global features capture long-term dependencies, trends in speech fluency, sustained emotional states in facial expressions, and discourse-level linguistic patterns, which

reflect the broader psychological state of an individual over an extended period (62, 63).

Depression manifests at multiple temporal and behavioral levels, necessitating an approach that jointly models local variations and global trends. Future research should continue to explore adaptive multimodal fusion strategies that dynamically balance local and global information, ensuring that automated depression detection models remain both effective and clinically interpretable.

5.3 Limitations and future directions

In this paper, the IMDD-Net is introduced as a novel deep learning model designed to enhance the accuracy of depression

detection by leveraging local and global features from video, audio, and text cues. This model effectively integrates multimodal data to provide a comprehensive analysis of depressive symptoms. By using the Kronecker product for multimodal fusion, our network explores deep interactions between modalities, further enhancing assessment accuracy. Experimental results demonstrate that the IMDD-Net achieves state-of-the-art performance. The robustness and precision of IMDD-Net are proved in identifying depression by extra experiment. In summary, via deep learning approach and video-audio-text multimodal data, IMDD-Net might be a useful tool of estimating depression risk in an undisturbed and convenient manner.

While the IMDD-Net has demonstrated state-of-the-art performance, there are several limitations that need further investigation. First, the IMDD-Net does not effectively address the issue of missing modalities. If any one of the input modalities (audio, video, or text) is absent, the model's performance may be impacted. But it also means that depression can manifest through various modalities, encompassing diverse behavioral, emotional, and physiological indicators. Second, the integration of multimodal data and the use of advanced feature extraction techniques increase the computational complexity and resource requirements. This may hinder the deployment of the IMDD-Net in real-world, resource-constrained environments. Third, the current implementation of the IMDD-Net focuses on accuracy rather than real-time processing capabilities. For practical applications in clinical settings, the model needs to be optimized for faster inference without compromising accuracy. Furthermore, the current version of IMDD-Net has not been optimized for real-time deployment. Due to the model's reliance on high-dimensional features, multiple deep architectures and Kronecker-based multimodal fusion, the computational load is relatively high. This may pose challenges for real-time inference, particularly in resource-constrained environments such as mobile platforms or telehealth applications. Finally, each modality (audio, video, text) comes with its own set of challenges, such as background noise in audio data, varying lighting conditions in video data, and the need for accurate transcription in text data. Addressing these issues comprehensively remains a challenge.

The ethical considerations are of crucial importance. While the IMDD-Net shows potential for use in clinical or diagnostic settings, its deployment raises important ethical considerations. First, the collection and analysis of audio, video, and textual data involve sensitive personal information, necessitating strict data privacy protection and informed consent protocols. Second, false positives may lead to unnecessary anxiety or stigmatization, while false negatives could result in missed opportunities for early intervention. Therefore, the model should not function as a standalone diagnostic tool but rather as an assistive system to support clinical decision-making. Additionally, bias in training data—such as demographic imbalance—could propagate disparities in predictions. Ensuring fairness, transparency, and human oversight will be essential in any future clinical deployment of the IMDD-Net.

Future work should focus on training and validating the IMDD-Net on larger and more diverse datasets, including data from various cultural backgrounds and different age groups. This would enhance the model's generalizability and robustness. As a future direction, we also plan to collaborate with clinical partners to conduct prospective validation studies or pilot trials, aiming to assess the model's reliability and generalizability in diverse patient populations. More efforts should be made to optimize the IMDD-Net for real-time processing through reducing computational complexity and improving the efficiency of the feature extraction and inference processes. Developing robust methods to handle missing modalities is crucial. This could involve creating imputation techniques or designing the model to be more resilient to incomplete input data, ensuring consistent performance regardless of data availability. Recovering missing modalities also represents a viable approach (64). By addressing these limitations and exploring these future directions, the IMDD-Net can provide a more reliable and efficient tool for the detection of depression.

## 6 Conclusion

In this study, we introduced the Integrative Multimodal Depression Detection Network (IMDD-Net), a novel deep-learning framework that effectively integrates video, audio, and textual modalities to enhance depression detection accuracy. By systematically capturing both local and global features within each modality and employing an advanced multimodal fusion strategy based on the Kronecker product, IMDD-Net provides a robust representation of depressive symptoms.

Experimental results conducted on the AVEC 2014 dataset demonstrate the superior performance of IMDD-Net, achieving state-of-the-art outcomes with a Root Mean Square Error (RMSE) of 7.55 and a Mean Absolute Error (MAE) of 5.75 in predicting BDI-II scores. The classification analysis further validates the model's practical utility, yielding an accuracy of approximately 79.3% in distinguishing depressed individuals from non-depressed ones. Ablation studies underscore the critical contribution of each modality and reinforce the necessity of incorporating multimodal data.

Despite these promising outcomes, IMDD-Net faces several limitations that merit attention. The computational complexity associated with multimodal fusion and high-dimensional feature extraction poses challenges for real-time and resource-constrained applications. Moreover, the model currently lacks comprehensive clinical validation, limiting immediate deployment in clinical practice. Future studies will address these limitations by optimizing the computational efficiency, developing methods to handle incomplete modality inputs, and rigorously evaluating clinical validity through prospective trials.

Ultimately, by bridging advanced computational methods with multimodal behavioral indicators, IMDD-Net represents a significant

step toward objective, accurate, and accessible depression screening, paving the way for enhanced mental health assessment practices.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://doi.org/10.1145/2647868.2647869>.

## Ethics statement

The studies involving humans were approved by the Ethics Committee of Northeastern University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

## Author contributions

YL: Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. XY: Conceptualization, Data curation, Investigation, Validation, Visualization, Writing – review & editing. MZ: Data curation, Investigation, Software, Writing – review & editing. JW: Data curation, Software, Writing – review & editing. YY: Conceptualization, Validation, Visualization, Writing – review & editing. WQ: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing. SQ: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

## References

- Herrman H, Kieling C, McGorry P, Horton R, Sargent J, Patel V. Reducing the global burden of depression: a Lancet-World Psychiatric Association Commission. *Lancet*. (2019) 393:e42–e3. doi: 10.1016/S0140-6736(18)32408-5
- Marwaha S, Palmer E, Suppes T, Cons E, Young AH, Upthegrove R. Novel and emerging treatments for major depression. *Lancet*. (2023) 401:141–53. doi: 10.1016/S0140-6736(22)02080-3
- Vandana, Marriwala N, Chaudhary D. A hybrid model for depression detection using deep learning. *Measur. Sens.* (2023) 25:100587. doi: 10.1016/j.measen.2022.100587
- Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, et al eds. "Detecting depression from facial actions and vocal prosody, " In: 2009 3rd international conference on affective computing and intelligent interaction and workshops, Piscataway, NJ, USA: IEEE, (2009) pp. 1–7.
- Beck AT, Alford BA. *Depression: Causes and treatment*. Philadelphia, USA, Pennsylvania: University of Pennsylvania Press (2009).
- Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguistics*. (2007) 20:50–64. doi: 10.1016/j.jneuroling.2006.04.001
- Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cognit Neurosci Neuroimaging*. (2018) 3:223–30. doi: 10.1016/j.bpsc.2017.11.007
- Nunes SO, Reiche EM, Morimoto HK, Matsuo T, Itano EN, Xavier EC, et al. Immune and hormonal activity in adults suffering from depression. *Braz J Med Biol Res*. (2002) 35:581–7. doi: 10.1590/S0100-879X2002000500011
- Kupfer DJ, Foster FG. Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression. *Lancet*. (1972) 2:684–6. doi: 10.1016/S0140-6736(72)92090-9
- McDermott LM, Ebmeier KP. A meta-analysis of depression severity and cognitive function. *J Affect Disord*. (2009) 119:1–8. doi: 10.1016/j.jad.2009.04.022
- Hu J, Luo J, Xu Z, Liao B, Dong S, Peng B, et al. Spatio-temporal learning and explaining for dynamic functional connectivity analysis: Application to depression. *J Affect Disord*. (2024) 364:266–73. doi: 10.1016/j.jad.2024.08.014
- Tian X, Zhu L, Zhang M, Wang S, Lu Y, Xu X, et al. Social anxiety prediction based on ERP features: A deep learning approach. *J Affect Disord*. (2024) 367:545–53. doi: 10.1016/j.jad.2024.09.006
- Xiao W, Moncy JC, Ghazi-Noori A-R, Woodham RD, Rezaei H, Bramon E, et al. Enhanced network synchronization connectivity following transcranial direct current stimulation (tDCS) in bipolar depression: effects on EEG oscillations and deep learning-based predictors of clinical remission. *J Affect Disord*. (2025) 369:576–87. doi: 10.1016/j.jad.2024.09.054
- Jin H, Nath SS, Schneider S, Junghaenel D, Wu S, Kaplan C. An informatics approach to examine decision-making impairments in the daily life of individuals with depression. *J BioMed Inform*. (2021) 122:103913. doi: 10.1016/j.jbi.2021.103913

## Funding

The author(s) declare financial support was received for the research and/or publication of this article. This study was supported by the National Natural Science Foundation of China (62271131) and the Fundamental Research Funds for the Central Universities (N25BJD013).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

15. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital Phenotyping of Mental Health using multimodal sensing of multiple situations of interest: A Systematic Literature Review. *J BioMed Inform.* (2023) 138:104278. doi: 10.1016/j.jbi.2022.104278
16. Thai M, Nair AU, Klimes-Dougan B, Albott CS, Silamongkol T, Corkrum M, et al. Deep transcranial magnetic stimulation for adolescents with treatment-resistant depression: A preliminary dose-finding study exploring safety and clinical effectiveness. *J Affect Disord.* (2024) 354:589–600. doi: 10.1016/j.jad.2024.03.061
17. He L, Niu M, Tiwari P, Marttinen P, Su R, Jiang J, et al. Deep learning for depression recognition with audiovisual cues: A review. *Inf Fusion.* (2022) 80:56–86. doi: 10.1016/j.inffus.2021.10.012
18. Churi H, Keshri P, Khamkar S, Sankhe A. *A deep learning approach for depression classification using audio features.* Tamilnadu: Fast Track Publications (2021).
19. Dong Y, Yang X. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing.* (2021) 441:279–90. doi: 10.1016/j.neucom.2021.02.019
20. He L, Cao C. Automated depression analysis using convolutional neural networks from speech. *J BioMed Inform.* (2018) 83:103–11. doi: 10.1016/j.jbi.2018.05.007
21. Joshi ML, Kanoongo N. Depression detection using emotional artificial intelligence and machine learning: A closer review. *Mater Today: Proc.* (2022) 58:217–26. doi: 10.1016/j.matpr.2022.01.467
22. Trotszek M, Koitka S, Friedrich CM. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans Knowledge Data Eng.* (2020) 32:588–601. doi: 10.1109/TKDE.69
23. Saffar AH, Mann TK, Ofoghi B. Textual emotion detection in health: Advances and applications. *J BioMed Inform.* (2023) 137:104258. doi: 10.1016/j.jbi.2022.104258
24. Dalal S, Jain S, Dave M. DepressionFeature: Underlying ontology for user-specific depression analysis. *J Supercomput.* (2025) 81:1–21. doi: 10.1007/s11227-024-06585-w
25. Liu Y. Depression detection via a Chinese social media platform: a novel causal relation-aware deep learning approach. *J Supercomput.* (2024) 80:10327–56. doi: 10.1007/s11227-023-05830-y
26. Liu Y. Depression clinical detection model based on social media: a federated deep learning approach. *J Supercomput.* (2024) 80:7931–54. doi: 10.1007/s11227-023-05754-7
27. Devlin J, Chang M-W, Lee K, Toutanova K. *Bert: Pre-training of deep bidirectional transformers for language understanding.* Minneapolis, Minnesota: Association for Computational Linguistics (2018).
28. Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. *AVEC '14: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* New York NY United States: Association for Computing Machinery, (2014).
29. Li X, Guo W, Yang H, eds. "Depression severity prediction from facial expression based on the DRR\_DepressionNet network." In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Piscataway, NJ, USA: IEEE, (2020), pp. 2757–64.
30. Uddin MA, Joolee JB, Lee Y-K. Depression level prediction using deep spatiotemporal features and multilayer bi-LSTM. *IEEE Trans Affect Comput.* (2022) 13:864–70. doi: 10.1109/TAFFC.2020.2970418
31. He L, Guo C, Tiwari P, Pandey HM, Dang W. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *Int J Intelligent Syst.* (2021) 37:10140–56. doi: 10.1002/int.22426
32. Rajawat AS, Bedi P, Goyal SB, Bhaladhare P, Aggarwal A, Singhal RS. Fusion fuzzy logic and deep learning for depression detection using facial expressions. *Proc Comput Sci.* (2023) 218:2795–805. doi: 10.1016/j.procs.2023.01.251
33. Niu M, Tao J, Liu B, Huang J, Lian Z. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Trans Affect Comput.* (2023) 14:294–307. doi: 10.1109/TAFFC.2020.3031345
34. Sun H, Liu J, Chai S, Qiu Z, Lin L, Huang X, et al. Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sens (Basel).* (2021) 21:4764. doi: 10.3390/s21144764
35. Bucur A-M, Cosma A, Rosso P, Dinu LP, eds. "It's just a matter of time: Detecting depression with time-enriched multimodal transformers." In: *European Conference on Information Retrieval*, Cham, Switzerland: Springer, (2023), pp. 200–15.
36. Li Y, Wang Y, Cui Z, eds. "Decoupled multimodal distilling for emotion recognition." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Los Alamitos: IEEE Computer Society, (2023), pp. 6631–40.
37. Beck AT, Steer RA, Brown GK. *BDI-II: Beck depression inventory.* Springer, NY, USA: Pearson (1996).
38. Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process Lett.* (2016) 23:1499–503. doi: 10.1109/LSP.2016.2603342
39. Adelson EH, Anderson CH, Bergen JR, Burt PJ, Ogden J. Pyramid methods in image processing. *RCA engineer.* (1984) 29:33–41.
40. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Comput.* (2016) 7:190–202. doi: 10.1109/TAFFC.2015.2457417
41. Tiwari V. MFCC and its applications in speaker recognition. *Int J Emerg Technol.* (2010) 1:19–22.
42. Bertasius G, Wang H, Torresani L, eds. *Is space-time attention all you need for video understanding?* In: *IICML Red Hook, NY: Curran Associates, Inc.* (2021) (2):3–4.
43. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
44. Cai C, Niu M, Liu B, Tao J, Liu X, eds. *TDCA-Net: Time-Domain Channel Attention Network for Depression Detection.* Red Hook, NY, USA: Interspeech Curran Associates, Inc. (2021).
45. Shang Y, Pan Y, Jiang X, Shao Z, Guo G, Liu T, et al. LQGDNet: A local quaternion and global deep network for facial depression recognition. *IEEE Trans Affect Comput.* (2023) 14:2557–63. doi: 10.1109/TAFFC.2021.3139651
46. de Melo WC, Granger E, López MB. MDN: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Trans Affect Comput.* (2023) 14:578–90. doi: 10.1109/TAFFC.2021.3072579
47. Niu M, Tao J, Liu B, eds. "Multi-scale and multi-region facial discriminative representation for automatic depression level prediction." In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Piscataway, NJ, USA: IEEE, (2021), pp. 1325–9.
48. De Melo WC, Granger E, Lopez MB, eds. "Encoding temporal information for automatic depression recognition from facial analysis." In: *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Piscataway, NJ, USA: IEEE, (2020), pp. 1080–4.
49. Pan Y, Shang Y, Liu T, Shao Z, Guo G, Ding H, et al. Spatial–Temporal Attention Network for Depression Recognition from facial videos. *Expert Syst Appl.* (2024) 237:121410. doi: 10.1016/j.eswa.2023.121410
50. Song S, Jaiswal S, Shen L, Valstar M. Spectral representation of behaviour primitives for depression analysis. *IEEE Trans Affect Comput.* (2022) 13:829–44. doi: 10.1109/TAFFC.2020.2970712
51. Havigerová JM, Haviger J, Kučera D, Hoffmannová P. Text-based detection of the risk of depression. *Front Psychol.* (2019) 10:513. doi: 10.3389/fpsyg.2019.00513
52. Buchwald AM, Rudick-Davis D. The symptoms of major depression. *J Abnormal Psychol.* (1993) 102:197. doi: 10.1037/0021-843X.102.2.197
53. Bourke C, Douglas K, Porter R. Processing of facial emotion expression in major depression: a review. *Aust New Z J Psychiatry.* (2010) 44:681–96. doi: 10.3109/00048674.2010.496359
54. Balsters M JH, Krahmer E J, Swerts M GJ, Vingerhoets A JJM. Verbal and nonverbal correlates for depression: a review. *Curr Psychiatry Rev.* (2012) 8:227–34. doi: 10.2174/157340012800792966
55. Hale WW III. Judgment of facial expressions and depression persistence. *Psychiatry Res.* (1998) 80:265–74. doi: 10.1016/S0165-1781(98)00070-5
56. Koops S, Brederoo SG, de Boer JN, Nadema FG, Voppel AE, Sommer IE. Speech as a biomarker for depression. *CNS Neurol Disorders Drug Targets CNS Neurol Disorders.* (2023) 22:152–60. doi: 10.2174/1871527320666211213125847
57. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* (2015) 71:10–49. doi: 10.1016/j.specom.2015.03.004
58. Wolohan J, Hiraga M, Mukherjee A, Sayeed ZA, Millard M, eds. "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP." In: *Proceedings of the first international workshop on language cognition and computational models*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, (2018), pp. 11–21.
59. Angst J, Merikangas KR. Multi-dimensional criteria for the diagnosis of depression. *J Affect Disord.* (2001) 62:7–15. doi: 10.1016/S0165-0327(00)00346-3
60. Wichers M. The dynamic nature of depression: a new micro-level perspective of mental disorder that meets current challenges. *psychol Med.* (2014) 44:1349–60. doi: 10.1017/S0033291713001979
61. Pampouchidou A, Simantiraki O, Fazlollahi A, Padiaditis M, Manousos D, Roniotis A, et al. "Depression assessment by fusing high and low level features from audio, video, and text." In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA: Association for Computing Machinery, (2016), pp. 27–34.
62. de Melo WC, Granger E, Hadid A, eds. "Combining global and local convolutional 3d networks for detecting depression from facial expressions." In: *2019 14th IEEE international conference on automatic face & gesture recognition (fg 2019)*, Piscataway, NJ, USA: IEEE, (2019), pp. 1–8.
63. Shang Y, Pan Y, Jiang X, Shao Z, Guo G, Liu T, et al. LQGDNet: A local quaternion and global deep network for facial depression recognition. *IEEE Trans Affect Comput.* (2021) 14:2557–63. doi: 10.1109/TAFFC.2021.3139651
64. Wang Y, Cui Z, Li Y, eds. "Distribution-consistent model recovering for incomplete multimodal learning." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Los Alamitos: IEEE Computer Society, (2023), pp. 22025–34.