TYPE Original Research
PUBLISHED 28 January 2025
DOI 10.3389/fpsyt.2025.1508772

Check for updates

OPEN ACCESS

EDITED BY
Steven Fernandes,
Creighton University, United States

REVIEWED BY
Alwin Poulose,
Indian Institute of Science Education and Research, India
Kuldeep Singh,
Guru Nanak Dev University, India

*CORRESPONDENCE
Peng Li
3308260670@qq.com

RECEIVED 13 October 2024
ACCEPTED 02 January 2025
PUBLISHED 28 January 2025

CITATION
Jin N, Ye R and Li P (2025) Diagnosis of depression based on facial multimodal data. *Front. Psychiatry* 16:1508772.
doi: 10.3389/fpsyt.2025.1508772

COPYRIGHT
© 2025 Jin, Ye and Li. 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.

Diagnosis of depression based on facial multimodal data

Nani Jin[1], Renjia Ye[1] and Peng Li[2]*

[1]Materdicine Lab, School of Life Sciences, Shanghai University, Shanghai, China, [2]Research Department, Third Xiangya Hospital of Central South University, Changsha, China

Introduction: Depression is a serious mental health disease. Traditional scale-based depression diagnosis methods often have problems of strong subjectivity and high misdiagnosis rate, so it is particularly important to develop automatic diagnostic tools based on objective indicators.

Methods: This study proposes a deep learning method that fuses multimodal data to automatically diagnose depression using facial video and audio data. We use spatiotemporal attention module to enhance the extraction of visual features and combine the Graph Convolutional Network (GCN) and the Long and Short Term Memory (LSTM) to analyze the audio features. Through the multi-modal feature fusion, the model can effectively capture different feature patterns related to depression.

Results: We conduct extensive experiments on the publicly available clinical dataset, the Extended Distress Analysis Interview Corpus (E-DAIC). The experimental results show that we achieve robust accuracy on the E-DAIC dataset, with a Mean Absolute Error (MAE) of 3.51 in estimating PHQ-8 scores from recorded interviews.

Discussion: Compared with existing methods, our model shows excellent performance in multi-modal information fusion, which is suitable for early evaluation of depression.

KEYWORDS

depression, multi-modal data, feature fusion, spatial-temporal attention, artificial intelligence

1 Introduction

Depression, also known as clinical depression or depressive disorder, is a prevalent and serious mental health condition that manifests through persistent low mood, lack of energy, and other symptoms that significantly impact an individual's thoughts, emotions, behaviors, and overall health (1). According to the World Health Organization, approximately 280 million people worldwide suffer from depression, with 15% of those affected eventually dying by suicide (2). The multifaceted nature of depression, influenced by social, psychological, and biological factors, underscores the necessity for a comprehensive approach to its treatment (3). Long-term stress, genetic predispositions, and adverse social environments are key contributors to the onset of depression, necessitating multifaceted treatment strategies to help patients regain a healthy life.

Traditional methods for diagnosing depression often rely on clinical evaluations by doctors and self-reports from patients. These scale-based methods are fraught with challenges such as high subjectivity, potential misdiagnosis, regional disparities, and a general lack of medical awareness. Moreover, the subtle nature of depressive symptoms means many individuals fail to recognize their condition promptly, leading to delayed treatment and worsening symptoms. Therefore, developing auxiliary diagnostic tools based on objective indicators is crucial for improving early diagnosis and treatment outcomes.

Recent advancements in artificial intelligence (AI) and deep learning have introduced new possibilities for assisting in the diagnosis of depression. These technologies have shown promise in identifying patterns and features indicative of depression through various data modalities (4). However, there are still some limitations in the research aimed at automatic diagnosis of depression. Some studies only consider global features and ignore local features in facial video data, which may lead to insufficient capture of subtle facial changes related to depression. Other studies only consider video data without combining audio information, ignoring the importance of multimodal information. In addition, the design of some models is too complex, which leads to the poor interpretability of the model and the difficulty in understanding its inner mechanism. We address the limitations of previous related work and propose a novel multimodal deep convolutional network, aiming to overcome these problems and provide a more efficient solution for the automatic diagnosis of depression.

In this study, we propose a novel deep learning approach that leverages multimodal data fusion to automatically diagnose depression using facial video and audio data. Our method enhances the extraction of visual features through a spatiotemporal attention module and combines Graph Convolutional Networks (GCN) and Long Short-Term Memory (LSTM) networks to analyze audio features. By integrating these multimodal features, our model effectively captures diverse patterns associated with depression. Our experimental results demonstrate that the proposed method outperforms existing approaches in terms of performance metrics, making it a promising tool for the early evaluation and diagnosis of depression. The main contributions of our study are as follows:

1. We introduce a novel multimodal network architecture that comprehensively integrates video and audio information, significantly enhancing the reliability of depression diagnosis.
2. We design a feature fusion model that effectively combines temporal and spatial features, providing a more comprehensive representation of video data and facilitating a deeper analysis of the patient's psychological state.
3. We employ a combined GCN and LSTM model to process audio data, constructing a graph structure to analyze Mel-Frequency Cepstral Coefficients (MFCC), thereby improving the interpretability and accuracy of the diagnostic process.

This paper is organized into five sections. The first section presents the research background, discusses the status and challenges of depression identification, and introduces the objectives and significance of the study. The second section reviews recent methods for depression evaluation using both single-modal and multi-modal data. The third section details the proposed method, including the overall network architecture and its components. Section four includes the experimental environment, training process, dataset details, results and discussions. Finally, section five summarizes the contributions of this study, evaluates the advantages and limitations of the proposed method, and outlines directions for future research.

2 Related work

Studies have shown that depression state is closely related to patients' head and face activities (5). Currently, some studies have tried multi-modal fusion of facial video information and other types of data, such as voice features and text information, to improve the accuracy of depression diagnosis. By utilizing multiple sources of information, the emotional state and psychological characteristics of patients can be captured more comprehensively, leading to more accurate assessment.

Al Jazaery and Guo (6) used 3D convolutional neural networks to extract deep spatio-temporal features of closely cropped aligned facial regions and relatively large head regions respectively, and then used recurrent neural networks to continue learning spatio-temporal information for final prediction. It is the first application of 3D convolutional neural networks to depression level analysis and shows great promise. But it focuses more on visual-based non-verbal data and does not take audio into account. Sun Haohao et al. (7) performed face detection, alignment and cropping on video frames in AVEC2013 (8) and AVEC2014 (9) depression databases to obtain the whole face image and the local eye and mouth region. Then, the deep convolutional neural network that fuses the attention mechanism of the channel layer is used to extract the corresponding global features and local features. The multiple visual features learned are more discriminative than the global features alone. But this study does not consider the influence of the audio. Yuchen Pan et al. (10) proposed the Spatio-Temporal Attention Depression Recognition Network (STA-DRN), which mainly uses the spatio-temporal attention (STA) mechanism to generate spatial and temporal attention vectors, so as to capture the global and local spatio-temporal relationships of features. In the STA module, there is also an attention vector fusion strategy that fuses spatial and temporal domain information. This model can capture the dynamic change process of facial expression and enhance the feature correlation in the process of depression recognition. JH Kim et al.

(11) introduces the customized VGG-19 (CVGG-19) architecture, which integrates designs from VGG, Inception-v1, ResNet, and Xception to enhance facial emotion recognition (FER). The CVGG-19 significantly improves performance by 59.29% and reduces computational cost by 89.5% compared to the original VGG-19. Additionally, CVGG-19 achieves an average F1-score that is 3.86% higher than Inception-V1, ResNet50, and Xception architectures. Constantino Álvarez Casado et al. (12) extracted remote photoplethysmography (rPPG) signals directly from facial videos and computed a variety of statistical, geometric and physiological features including heart rate. These features were fed into machine learning regression models to identify different levels of depression. The results of this approach are comparable to other audiovisual models based on voice or facial expression.

Some studies only focus on audio information for depression recognition. Momoko Ishimaru et al. (13) input the feature vector converted from audio data into graph convolution layer and dense layer in turn, and finally obtain the prediction score. This new regression model uses the generated graph-structured data to express correlations between audio features, which can be exploited to assess the severity of depression. Li et al. (14) built speech signals into speech graphs based on feature similarity to input Graph-LSTM neural network for classification. The network is a new graph neural network structure combining LSTM aggregator and weighted pool, which enhances the interpretability of the model and can effectively identify speech emotional features. However, the model also has the shortcomings of redundant speech graph features and lack of visual features.

Some advancements in bimodal speech emotion recognition (SER) using both acoustic and text data, focusing on the significance of attention mechanisms and fusion strategies in combination with traditional deep learning techniques. Also there are some challenges such as limited datasets and difficulties in data acquisition (15).

Uddin et al. (16) input the preprocessed audio clips and video clips into the spatio-temporal network based on audio and video. The dynamic feature descriptor Volume Local Directional Structural Pattern is introduced to encode the structure, so as to extract the dynamic facial features. Then, Temporal Attentive Pooling is used to summarize the segmentation features, and Multi-modal factorized bilinear pooling is used to fuse the multimodal features. Finally, the corresponding BDI-II scores were obtained by regression to estimate the severity of depression. This method has strong feature extraction ability and considers multi-modal data but ignores the association between high-level semantic features and channels. Ming Fang et al. (17) comprehensively considered video, audio and text information, and designed a multi-modal fusion model with multi-level attention mechanism (MFM-Att) for depression detection. The model uses two LSTMs to learn video and audio features, and a Bi-LSTM with attention mechanism to learn text features, and then inputs these three outputs into the MFM-Att for feature fusion. This design can make information complementary between different modalities. However, the complexity of the model needs to be improved.

Improving the interpretability of diagnostic models for depression is crucial for clinical practice. David Gimeno-Gomez et al. (18) present a simple and flexible multimodal temporal model capable of recognizing nonverbal cues to depression from noisy real-world videos. They visualize the level to which these features contribute to the results through integrated gradients (19) based on audio-speech embeddings, facial emotion embeddings, facial, body and hand signatures, as well as gaze and blink information.

3 Methods

Figure 1 shows the framework of the proposed method for diagnosing depression based on multimodal data. Firstly, visual information and audio information are extracted from the recorded videos of the participants, and the two kinds of information are pretreated separately. Then, the feature extraction is performed on the preprocessed data and the multimodal feature set is obtained by feature fusion. After that, the processed features are classified and the respective classification results are output. This framework allows the model to synthesize visual and audio information, which helps to deeply mine the hidden information in the data. In the process of facial behavior feature extraction, we use the spatio-temporal attention module to strengthen the correlation between features and video frames. For audio features, GCN and LSTM are mainly used.

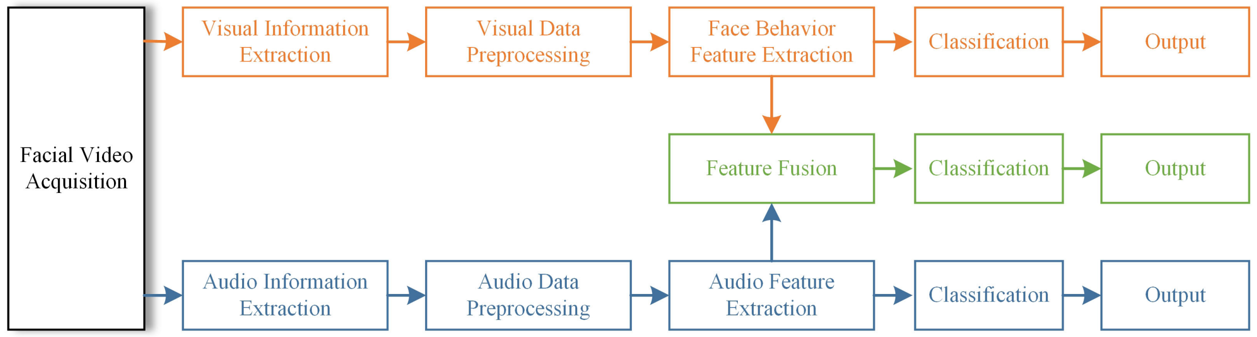

FIGURE 1
Framework for depression diagnosis.

3.1 Visual feature extraction

In order to effectively extract information from the facial behavior features, we propose Temporal-Spatial Network for Depression Diagnosis (TSNet-DD). The proposed model incorporates a temporal attention module and a spatial attention module to capture global and local features at the temporal and spatial levels from video frames. The core of TSNet-DD is that it can use the Temporal-Spatial Attention Module (TSAM) to enhance the correlation between pixels and frames.

The overall architecture of TSNet-DD network is shown in Figure 2. The initial layer of the network uses a 7×7×7 convolution kernel to perform downsampling with a step size of 1×2×2 to extract low-level features of the input image. Next, a 3×3×3 pooling operation with a step size of 1×2×2 is performed in the second layer, and the resulting features are denoted as F_v. The subsequent module1, module2, module3 and module4 correspond to different convolutional layers in the ResNet, and each module consists of a different number of residual blocks. TSNet32-DD corresponds to the ResNet18, that is, each module contains two residual blocks, and two sub-modules are also contained within each residual block. In ResNet18, these submodules are 3×3 convolutional layers, while in our network, we introduce TSAM. Therefore, a total of 32 TSAMs are used in the ResNet18-based network, and we refer to this network as TSNet32-DD.

Similarly, in the ResNet34-based network, there are 3,4,6 and 3 residual blocks in each module (sections marked red in Figure 2). Each residual block still contains two TSAMs, resulting in a total of 64 TSAMs in the final network, thus this network is called TSnet64-DD. After all residual modules, a pooling operation is performed on the feature map to resample the features into fixed shapes, and finally a fully connected layer is used to classify the subjects.

The feature extraction module TSAM in TSNet-DD contains a temporal attention module and a spatial attention module. These two modules are used to generate the temporal attention weight vector W_t and spatial attention weight vector W_s of the input F_v, respectively, so as to obtain the corresponding temporal feature F_t and spatial feature F_s. Then, these two kinds of features are fused to capture the intrinsic relationship between spatial-temporal features, assigning adaptive weights to the features with spatio-temporal information. The structure of TSAM is shown in Figure 3.

The fusion process of temporal attention module and spatial attention module could be expressed by the following formula:

$$F_{ts} = F_t + F_s \tag{1}$$

3.1.1 Temporal attention module

For video data of patients with depression, intra-frame temporal changes are crucial for depression recognition. Such temporal changes can be short term or long term dynamics spanning several seconds. Although short-term features could capture dynamic information between several frames, their ability to extract long-term dynamic features is limited. To address this problem, we introduce a temporal Attention module (TAM) for

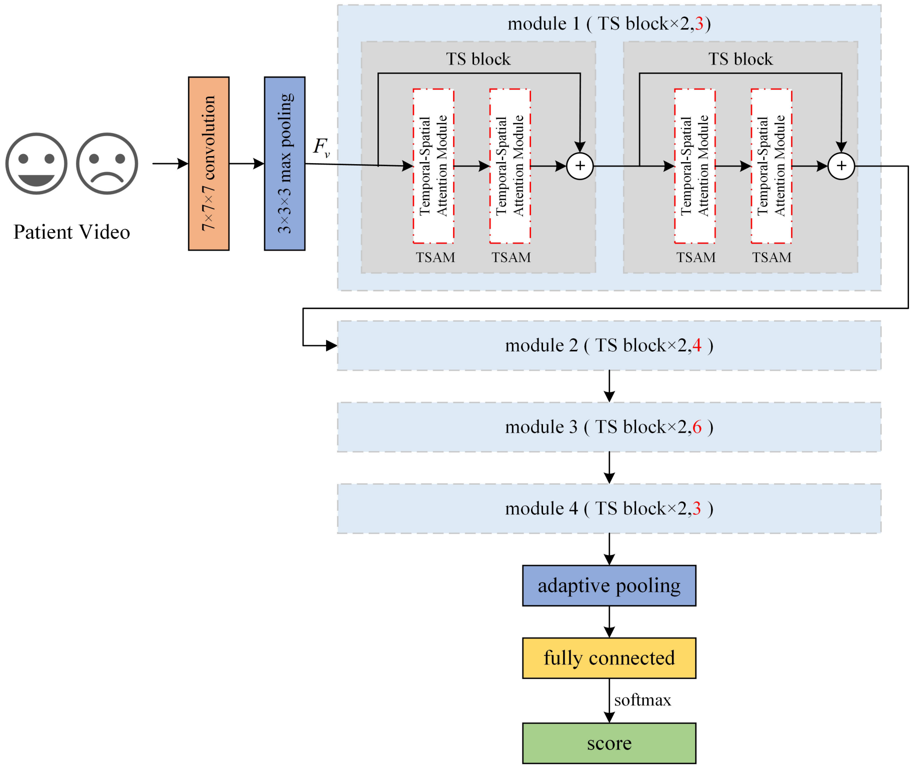
module 1 (TS block×2,3)
TS block
TS block
Temporal-Spatial Attention Module
Temporal-Spatial Attention Module
Temporal-Spatial Attention Module
Temporal-Spatial Attention Module
TSAM
TSAM
TSAM
TSAM
Patient Video
7×7×7 convolution
3×3×3 max pooling
F_v
module 2 (TS block×2,4)
module 3 (TS block×2,6)
module 4 (TS block×2,3)
adaptive pooling
fully connected
softmax
score

FIGURE 2
Architecture of TSNet-DD.

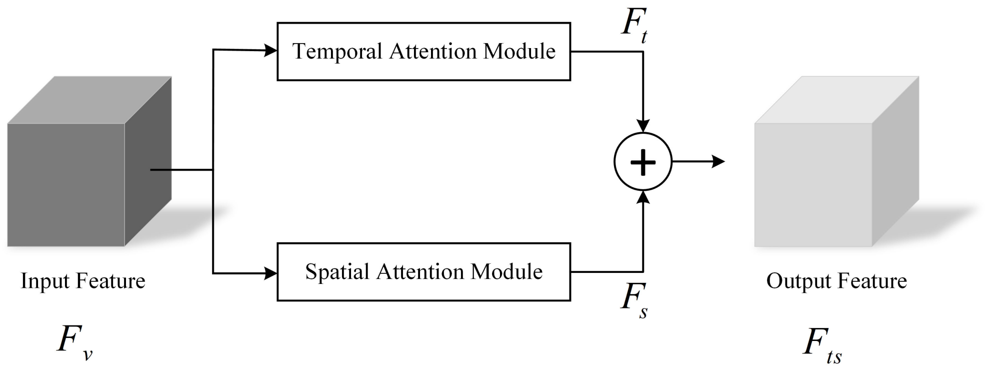

FIGURE 3
Temporal-spatial attention mechanism.

enhancing temporal information. The specific structure of this module is shown in Figure 4.

In the TSnet-DD model, after the second layer of 3×3×3 pooling operation, the feature F_v is obtained, and its size is assumed to be H×W×C. The feature F_v is sent to the TAM. Firstly, the global average pooling and max pooling operations are performed respectively to obtain two 1×1×C channel descriptions. Subsequently, these two descriptions are fed into a two-layer neural network with shared weights for processing. Then the resulting two features are added and the weight coefficient W_t is obtained through the sigmoid activation function. Finally, W_t is multiplied with the original input feature F_v to obtain the new scaled feature F_t. The process of TAM can be summarized as follows:

$$W_t = \sigma\Big(Conv\big(AvgPool(F_v)\big) + Conv\big(MaxPool(F_v)\big)\Big) \quad (2)$$

$$F_t = F_v * W_t \quad (3)$$

To further visualize the architecture and data transfer process of TAM, we show its pseudo-code in Algorithm 1.

```
Step 1 Input: Feature map $F_v \epsilon R^{H\times w\times c}$

Step 2 Global Average Pooling: $F_{avg} = AvgPool(F_v)$ where $F_{avg}\ \epsilon R^{1\times 1\times c}$.

Step 3 Global Max Pooling: $F_{max} = MaxPool(F_v)$ where $F_{max}\ \epsilon R^{1\times 1\times c}$.

Step 4 Shared Network: $W_t = \sigma\big(Conv(F_{avg}) + Conv(F_{max})\big)$ where $W_t\ \epsilon R^{1\times 1\times c}$, and $\sigma$ is the sigmoid function.

Step 5 Attention Scaling: $F_t = F_v * W_t$ where $*$ denotes element-wise multiplication.
```

Algorithm 1. Pseudocode of temporal attention module.

3.1.2 Spatial attention module

In our collection of videos about people with depression, some useful features usually appear in a sequence of consecutive video frames. Therefore, whether features can identify spatial order information is crucial in depression diagnosis. In addition, different locations of the face have their own unique features, and there are subtle relationships between these location features that cannot be captured by our naked eyes (20). With this in mind, we employ a Spatial Attention Module (SAM) to generate spatial vectors to capture the spatial information. The structure of SAM is shown in Figure 5.

In SAM, the input feature F_v can determine where the features are meaningful. Firstly, the average pooling and maximum pooling of the channel dimension are performed on F_v respectively to obtain two channel descriptions of size H×W×1, then the two descriptions are concatenated together in the channel dimension. Next, through a 7×7 convolutional layer and the activation function sigmoid, the weight coefficient W_s are obtained. Finally, the W_s is multiplied with the input feature F_v to obtain the final spatial attention vector F_s. This process can be expressed as follows:

$$W_s = \sigma\big(Conv([AvgPool(F_v), MaxPool(F_v)])\big) \quad (4)$$

$$F_s = F_v * W_s \quad (5)$$

The flow of the Spatial Attention Module is shown in Algorithm 2.

```
Input: Feature map $F_v \epsilon R^{H\times w\times c}$Channel-wise Average Pooling: $F_{avg} = AvgPool(F_v,\ axis = C)$ where $F_{avg}\ \epsilon R^{H\times W\times 1}$.

Channel-wise Max Pooling: $F_{max} = MaxPool(F_v,\ axis = C)$ where $F_{max}\ \epsilon R^{H\times W\times 1}$.

Concatenation: $F_{concat} = [F_{avg};\ F_{max}]$ where $F_{concat}\ \epsilon R^{H\times W\times 2}$.
```

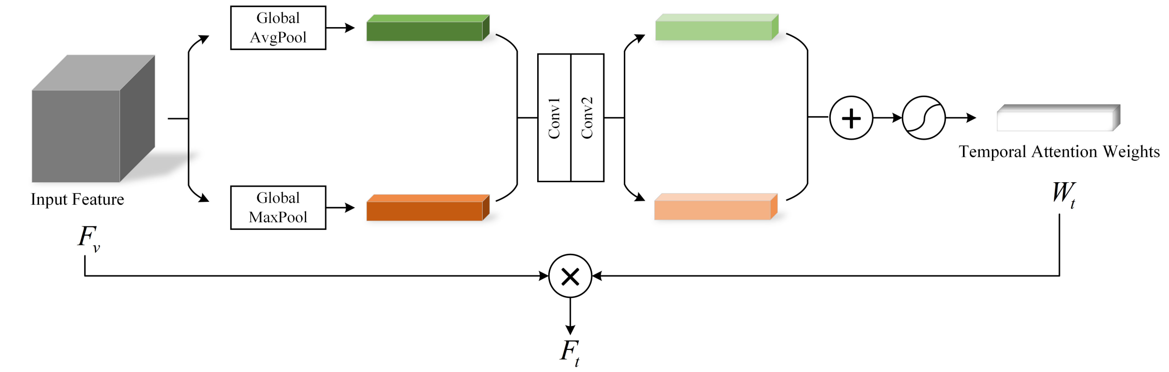

FIGURE 4
Temporal attention module.

```
Convolution: $W_s = \sigma(Conv(F_{concat}))$ where $W_s \epsilon R^{H\times W\times 1}$, and $\sigma$ is the sigmoid function.

Attention Scaling: $F_s = F_v * W_s$ where $*$ denotes element-wise multiplication.
```

Algorithm 2. Pseudocode of spatial attention module.

3.2 Audio feature extraction

The researchers found that people with depression tended to speak in a monotonous and lower tone than healthy controls (21). Therefore, in addition to the analysis of visual features, it is particularly important to mine the key features hidden in speech signals for the diagnosis of depression. See Figure 6 for the audio feature processing method used in this paper.

We use Mel-frequency Cepstral coefficients (MFCC) of audio data as effective features for depression recognition (22). MFCC takes into account the auditory properties of the human ears and can well capture the features in speech. The calculation process of MFCC is as follows: Firstly, the input audio signal is pre-weighted to enhance the high-frequency components. Then, the pre-weighted signal is divided into multiple short-time frames, and a window is applied to each frame to reduce the spectral leakage. Next, the fast fourier transform is performed on each windowed frame to convert the time domain signal to the frequency domain. Then the power spectrum of each frame is calculated. Finally, the power spectrum is passed through a bank of Mel filters, and its output is log transformed and discrete cosine transformed. At this point, we have the MFCC feature vector for each frame of the input audio. After that, we consider the MFCC feature vector of each frame as a node, and calculate the feature similarity of each node based on the Euclidean distance between its feature vectors, so as to construct the

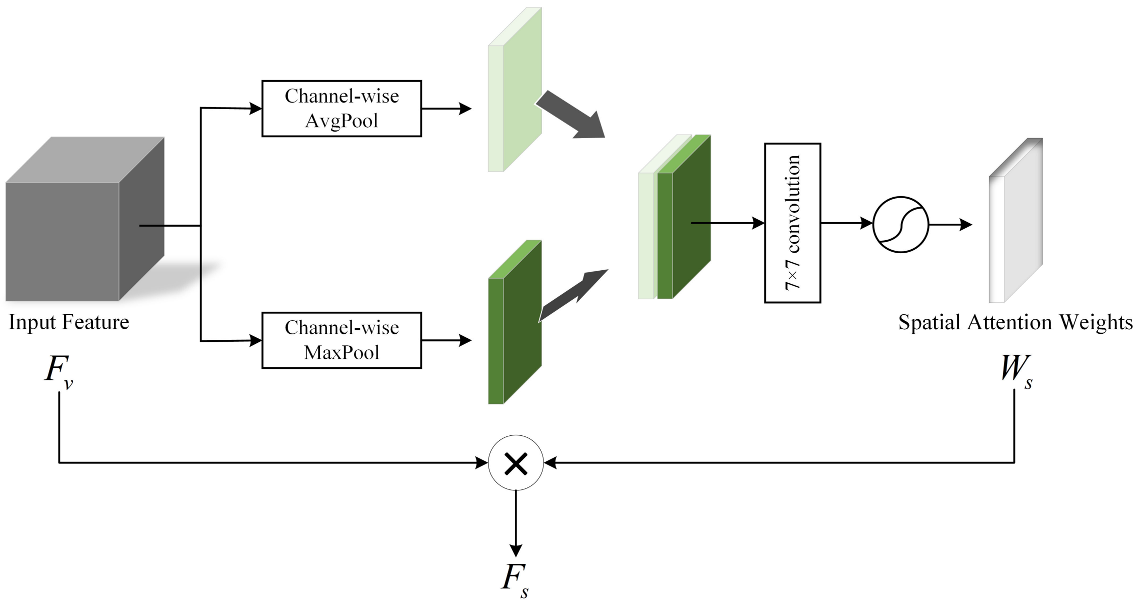

FIGURE 5
Spatial Attention Module.

edges between nodes. In this process, we set a threshold of 0.5 to limit the addition of edges, that is, only adding edges between nodes with high enough feature similarity and small enough distance. Finally, we assign a corresponding weight to each edge based on the inverse of the distance to better capture the local and global features in the audio signal. The specific algorithmic architecture of MFCC Calculation is shown in Algorithm 3.

```
Step 1 : Apply a pre-emphasis filter to the audio signal $X_t$ .
Step 2 : Divide the signal into overlapping frames .
Step 3 : Windowing : Apply a Hamming window to each frame .
Step 4 : Compute the Fast Fourier Transform of each frame .
Step 5 : Apply a Mel filter bank to the power spectra .
Step 6 : Take the logarithm of the Mel-filtered spectra .
Step 7 : Apply the Discrete Cosine Transform to obtain MFCCs .
```

Algorithm 3. Pseudocode of MFCC calculation.

To explore the complex patterns and temporal features of audio data, we input the constructed graph structure data into GCN and LSTM for processing. GCN aggregates the features of nodes and their neighbors through convolution operations. LSTM could capture long and short-term dependencies in the sequence. Algorithm 4 further details the process of combining GCN and LSTM. By combining GCN and LSTM, we can capture the high-level graph features of each node and the dynamics and dependencies in the time series. Finally, by subsampling and classifying the complex features, we can obtain a diagnosis of whether the speaker in the audio has depression or not.

```
Step 1 : Construct a graph (G = (V, E)) where each node ($V_i$) represents an MFCC feature vector of a frame .
Step 2 : Compute edge weights based on feature similarity (Euclidean distance) .
Step 3 : Aggregate features from neighboring nodes : $H^{l+1} = \sigma(D^{-1/2}AD^{-1/2}H^l W^l)$, where A is the adjacency matrix, D is the degree matrix, $H^l$ is the feature matrix at layer l, and $W^l$ is the weight matrix .
Step 4 : LSTM Layer : Capture temporal dependencies : $H_t = LSTM(x_t, h_{t-1})$, where $x_t$ is the input at time t and $h_t$ is the hidden state .
```

Algorithm 4. Pseudocode of GCN-LSTM model.

3.3 Video-audio fusion

Assuming that the final extracted visual feature is F_V and the obtained audio feature is F_A, we now discuss how to fuse these two features. In view of the fact that not all modality features play a positive role in the severity assessment of depression, we propose a Video-Audio Fusion Network (VAFN) to fuse the feature information of the two modalities. The structure of VAFN is shown in Figure 7. The input of VAFN is the multi-modal feature $F_{MM} = \{F_V, F_A\}$, and the output feature is the fused F_{VA}. The nature of video and audio data are different, which leads to different feature vector dimensions. Therefore, in the actual processing, we first perform zero-padding on the side with smaller size in F_V and F_A to ensure that the resulting dimensions of F_{VP} and F_{AP} are consistent. Then, F_{VP} and F_{AP} are superimposed in the horizontal and vertical directions respectively to obtain H_{VA} and V_{VA}. A fully connected layer is used to reduce the dimension of V_{VA}, and the attention weight vector V_{VAF} is obtained. Finally, H_{VA} and V_{VAF} are multiplied to obtain the final multi-modal fusion feature F_{VA}. The obtained fusion features are max-pooling and classified to obtain the depression prediction results. The entire fusion process described above is summarized in Algorithm 5.

```
Step 1 : Zero-pad the smaller feature vector $F_V$ or $F_A$ to match dimensions .
Step 2 : Concatenate the padded features horizontally and vertically : $H_{VA}$=[$F_{VP}$;$F_{AP}$] , $V_{VA}$=[$F_{VP}$;$F_{AP}$] .
Step 3 : Apply a fully connected layer to $V_{VA}$ to obtain an attention weight vector $V_{VAF}$=FC ($V_{VA}$)
Step 4 : Multiply $H_{VA}$ with $V_{VAF}$ to obtain the fused feature $F_{VA} = H_{VA} * V_{VAF}$ .
```

Algorithm 5. Pseudocode of VAFN.

4 Experiments

4.1 Experimental settings

The GPU used in this paper is NVIDIA RTX3090. The development and testing are carried out in the Python3.9 environment, and the integrated development tool is Pycharm.

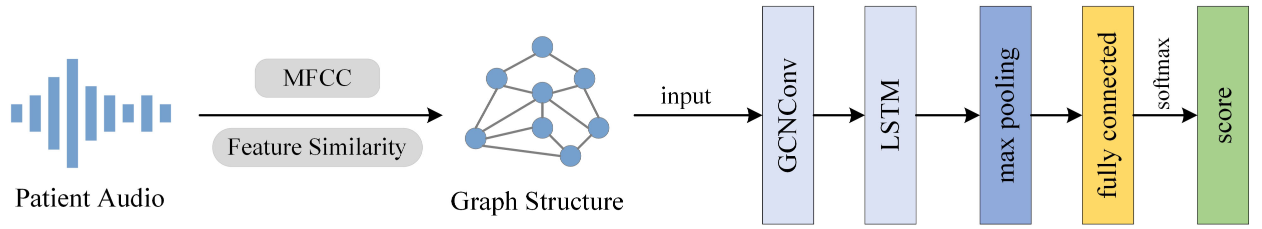

FIGURE 6
Audio feature extraction.

We use PyTorch v1.12.0 as the deep learning framework and use CUDA 11.6 in the model training process. The operating system is Windows10. In order to alleviate the over-fitting problem, we use the AdamW optimizer for training, and add the Dropout layer to the network backbone. The dropout rate is set between [0.4, 0.6]. As for optimization, the learning rate is set to be 10e−4 for modality feature extraction and 5×10e−5 for modality fusion with linear schedule strategy. During training, the number of iterations is consistently set to 5000.

In this study, we used the Extended Distress Analysis Interview Corpus (E-DAIC) dataset (23, 24) to validate our proposed method. The E-DAIC dataset is an extended version of the WOZ-DAIC dataset (23) and consists of semi-structured clinical interviews designed to identify psychological distress conditions such as anxiety, depression, and PTSD. In E-DAIC dataset, OpenSMILE (25) was used to extract the acoustic features of subjects, including Mel-frequency cepstral coefficients (MFCC) (26), and OpenFace (27) was used to extract the corresponding visual features. Facial features, eye fixations, head poses, and motor units are included. To protect the privacy of participants, the dataset provides these extracted features directly instead of raw video recordings. The E-DAIC dataset consists of clinical interview transcripts from 219 participants, along with corresponding assessments of depression and PTSD severity. To ensure a representative distribution of the data, the training set contains 163 samples, the validation set contains 56 samples, and the test set contains 10 samples. Each participant in the E-DAIC dataset was annotated according to their Patient Health Questionny-8 (PHQ-8) score (28), with scores higher than 10 classified as 1 (indicating the presence of depression) and scores lower than 10 labeled as 0 (indicating the absence of depression).

4.2 Evaluation metrics

Evaluation measures to evaluate the depression diagnostic model included F1 score (Equation 7) (29), root-mean-square error (RMSE, Equation 8) and mean absolute error (MAE, Equation 9) (30). The F1 score is the harmonic average of precision and recall (Equation 6) and is used to comprehensively

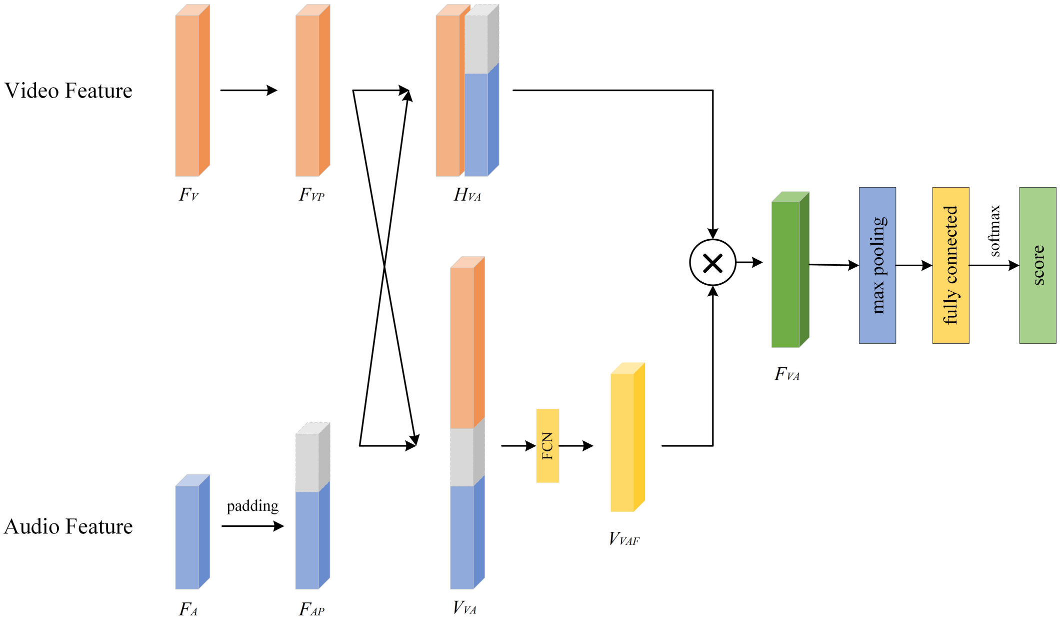

FIGURE 7
Video-audio fusion network.

measure the performance of the depression diagnostic model. RMSE can reveal how the model performs in extreme cases, such as severely overestimating or underestimating a patient"s depression, which can have a significant impact on clinical decision making. MAE gives the average difference between the model prediction and the actual value.

$$Precision = \frac{TP}{TP + FP},\ Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{8}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{9}$$

4.3 Results

4.3.1 Comparison with other methods

In this study, we preliminarily use video data and audio data separately for depression recognition based on single-modal features, and the results are shown in Table 1. Specifically, for video features, we compare the proposed TSNet-DD with references (6, 10, 12). It shows that TSNet-DD consistently outperforms the other three models, and TSNet64-DD outperforms TSNet32-DD. The TSNet64-DD model achieved the highest F1 score of 0.853, demonstrating its ability to capture both spatial and temporal features effectively. This improvement over previous models (6, 10, 12) suggests that our temporal-spatial attention mechanism significantly enhances feature extraction. Although the RMSE value of TSNet64-DD is slightly higher than that of TSNet32-DD at 5.11, the small difference in RMSE here is negligible compared to the advantages of its F1 value and MAE value. For audio data, the GCN-LSTM model achieved an F1 score of 0.827, outperforming previous models (13) and (14). This indicates that combining GCN and LSTM can effectively capture the complex patterns in audio features related to depression. Although the MAE value of GCN-LSTM is not the lowest, considering the characteristic that MAE is insensitive to outliers and its excellent performance on F1, we believe that GCN-LSTM has a unique advantage in processing audio features of depression.

TABLE 1 Results of depression recognition under single-modal features.

feature	model	F1	RMSE	MAE
video	RNN-C3D (6)	0.723	8.07	5.78
	STA-DRN (10)	0.702	8.94	6.77
	RFR (12)	0.710	8.49	6.57
	TSNet32-DD	0.800	5.11	5.03
	TSNet64-DD	0.853	5.23	4.45
audio	GCNN (13)	0.690	9.28	6.65
	GLNN (14)	0.788	8.43	5.04
	GCN-LSTM	0.827	6.67	6.28

Subsequently, we fuse facial video features and audio features to evaluate the performance of multimodal data in depression diagnosis. The results are shown in Table 2. We find that the model based on multi-modal data consistently outperforms the performance using only single-modal data in terms of F1 value. The F1 value of our proposed method finally reaches 0.922, which is not only better than the performance of all single-modal data, but also the highest in all experiments based on multi-modal data. This may be due to the diversity of the input data, and it indicates that there are clear differences in facial visual features and voice features between patients with depression and healthy participants. It turns out that the multimodality-based assistive method has its unique significance in depression diagnosis when the privacy of the participants is protected as much as possible.

To provide a comprehensive evaluation of our model's performance, we also included the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) values. The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC provides an aggregate measure of performance across all possible classification thresholds. Our model's ROC curves for both single-modal and multi-modal data are shown in Figure 8. The AUC values for TSNet64-DD, GCN-LSTM, and VAFN(TSNet64+GL) are summarized in Table 3. The results indicate that our multi-modal fusion model achieves the highest AUC value, further confirming its superior performance in distinguishing between depressed and non-depressed individuals.

In addition, our proposed models strike a balance between computational complexity and performance. The computational complexity of TSNet-DD is primarily determined by the number of convolutional layers and the attention modules. The attention modules add a small overhead due to the additional operations for attention weight calculation. Despite the added complexity of the attention modules, TSNet-DD is designed to be non-redundant, ensuring efficient processing without unnecessary computational overhead. This balance between complexity and efficiency allows TSNet-DD to achieve high performance while maintaining reasonable computational requirements. GCN-LSTM combines the strengths of GCN and LSTM to process audio features. GCN is used to aggregate features from neighboring nodes, while LSTM captures temporal dependencies. GCN-LSTM is designed to handle

TABLE 2 Results of depression recognition under multi-modal features.

model	F1	RMSE	MAE
MFM-Att (17)	0.895	7.29	4.03
GCN	0.918	6.24	3.88
VAFN(TSNet32+GL)	0.903	5.77	4.00
VAFN(TSNet64+GL)	0.922	6.06	3.51

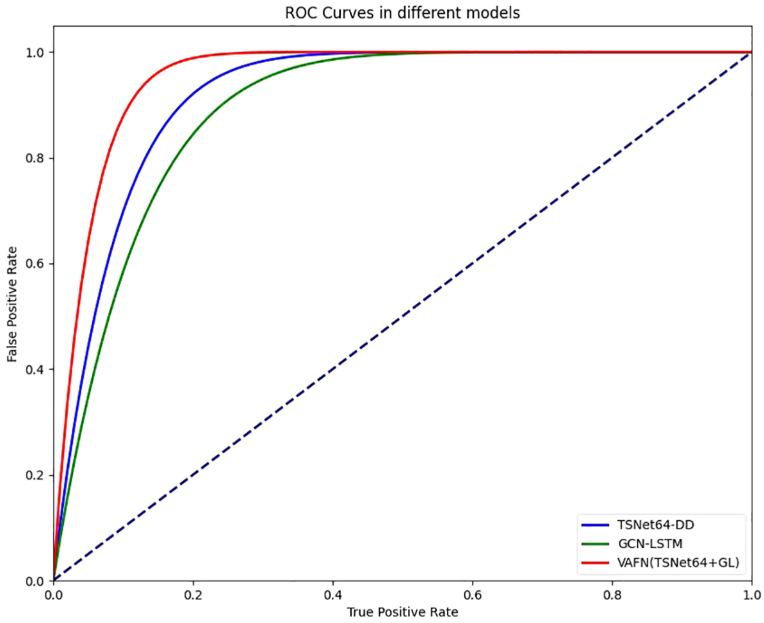

FIGURE 8
Comparison of different ROCs.

the variability and noise in audio data effectively. While the combination of GCN and LSTM increases the computational complexity, the model's ability to capture complex patterns and temporal features justifies the additional computational cost. VAFN fuses the feature information from both video and audio modalities. It uses zero-padding to handle different feature vector dimensions and attention mechanisms to assign adaptive weights to the features. The overall complexity of VAFN depends on the individual complexities of TSNet-DD and GCN-LSTM, along with the additional operations for feature fusion. The fusion process involves concatenation, fully connected layers, and attention weight calculation, which add to the computational load. VAFN is designed to leverage the complementary nature of visual and audio features, resulting in improved diagnostic performance. The fusion process, while adding some computational overhead, is optimized to ensure that the model remains efficient and scalable.

4.3.2 Ablation study

To better understand the contributions of various components in our proposed model, we conducted an ablation study. This study evaluates the impact of the Temporal-Spatial Attention Module (TSAM), the combination of Graph Convolutional Network (GCN) and Long Short-Term Memory (LSTM) for audio features, and the Video-Audio Fusion Network (VAFN). We performed experiments by systematically removing or modifying these components and observing the resulting changes in performance metrics.

We first assess the effect of the TSAM by comparing the full TSNet64-DD model with a variant that does not include the TSAM. The results are shown in Table 4. It demonstrates that the TSAM significantly improves the performance of the model, highlighting its importance in capturing temporal and spatial features. Next, we evaluate the impact of using GCN and LSTM for audio feature extraction. We compare the full GCN-LSTM model with variants that use only GCN or only LSTM. The results are shown in Table 5. The results indicate that the combination of GCN and LSTM outperforms the individual models, demonstrating the effectiveness of integrating both graph-based and temporal features for audio data. Finally, we assess the impact of the VAFN by comparing the full VAFN(TSNet64+GL) model with variants that use only video features (TSNet64-DD) or only audio features (GCN-LSTM). The results are shown in Table 6. The results clearly show that the fusion of video and audio features significantly enhances the performance, confirming the complementary nature of these modalities. The ablation study confirms that each component of our proposed model contributes to its overall performance. The TSAM enhances the extraction of temporal and spatial features from video data, the combination of GCN and LSTM effectively captures complex audio patterns, and the VAFN successfully integrates multimodal features to improve diagnostic accuracy. These findings validate the design choices and highlight the importance of multimodal data fusion in the automatic diagnosis of depression.

TABLE 3 Results of different model's AUC values.

model	AUC	FLOPs
TSNet64-DD	0.912	453,787,648
GCN-LSTM	0.880	4,915,200
VAFN(TSNet64+GL)	0.950	463,424,512

4.3.3 Effects of different subject groupings

We divide the dataset into three categories by sex: male group, female group, and mixed gender group. For each data set, we

TABLE 4 Impact of Temporal-Spatial Attention Module (TSAM) on video feature extraction.

Model	F1	RMSE	MAE
TSNet64-DD	0.853	5.23	4.45
TSNet64-DD w/o TSAM	0.789	6.12	5.37

TABLE 5 Impact of GCN and LSTM on audio feature extraction.

Model	F1	RMSE	MAE
GCN-LSTM	0.827	6.67	6.28
GCN only	0.742	7.89	7.12
LSTM only	0.756	7.65	6.87

TABLE 6 Impact of Video-Audio Fusion Network (VAFN) on multimodal feature fusion.

Model	F1	RMSE	MAE
VAFN(TSNet64+GL)	0.922	6.06	3.51
TSNet64-DD	0.853	5.23	4.45
GCN-LSTM	0.827	6.67	6.28

conducted experiments based on single mode and multi-mode respectively. In the video mode experiment, we adopt TSNet64, which has better performance. The experimental results are shown in Figure 9, Figure 10 and Figure 11. We find that for each modal, the F1 values of the mixed gender group are consistently lower than those assessed on either the male or female group alone. In addition, the female group almost always outperformed the male group, which may be due to a category imbalance between the samples. It also suggests that men and women differ in the information conveyed in facial behavior and speech during the diagnosis of depression. In addition, for each subject group, the experimental results based on the multimodal feature set are generally better than those based on the single-modal feature set, which is mainly due to the diversity of training data which brings more abundant features.

4.4 Interpretability analysis

Deep learning techniques are usually 'black-box', but in clinical practice we need more transparent models to increase their credibility and interpretability. Therefore, we perform an interpretability analysis of our model. We show the attribution scores for audio, gaze, action unit (AU), and pose, where higher values indicate strong attribution to positive predictions. The E-DAIC dataset has more than 19,000 frames of facial images for each

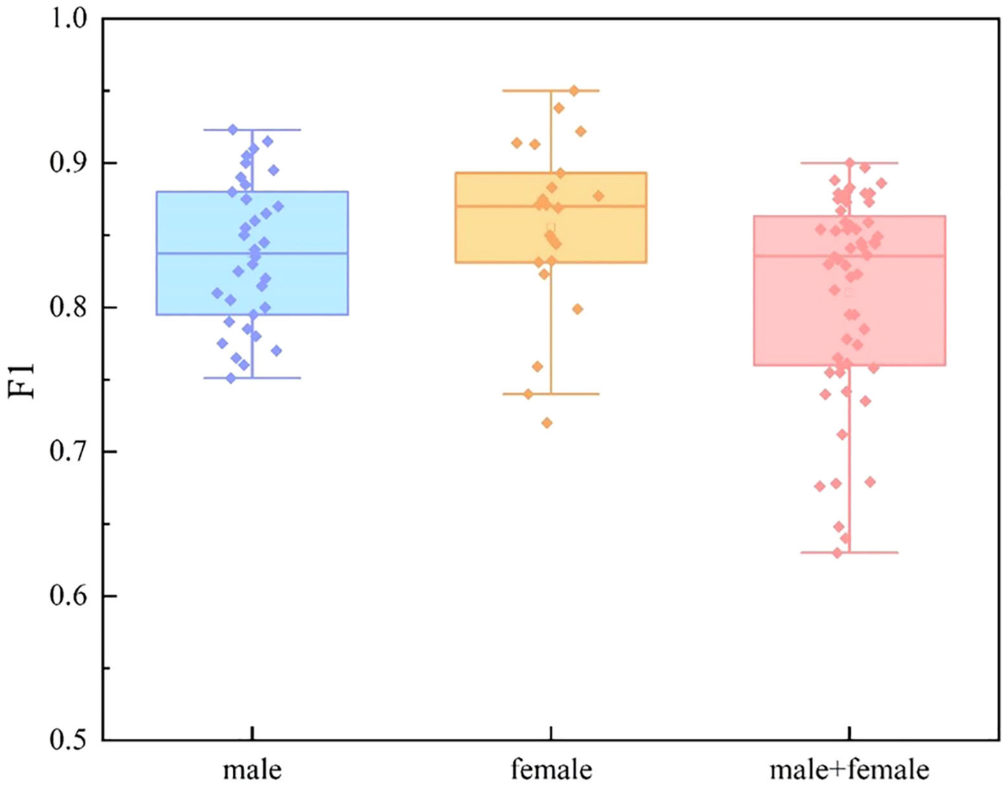

FIGURE 9
Comparison of different subject groups on video modal features.

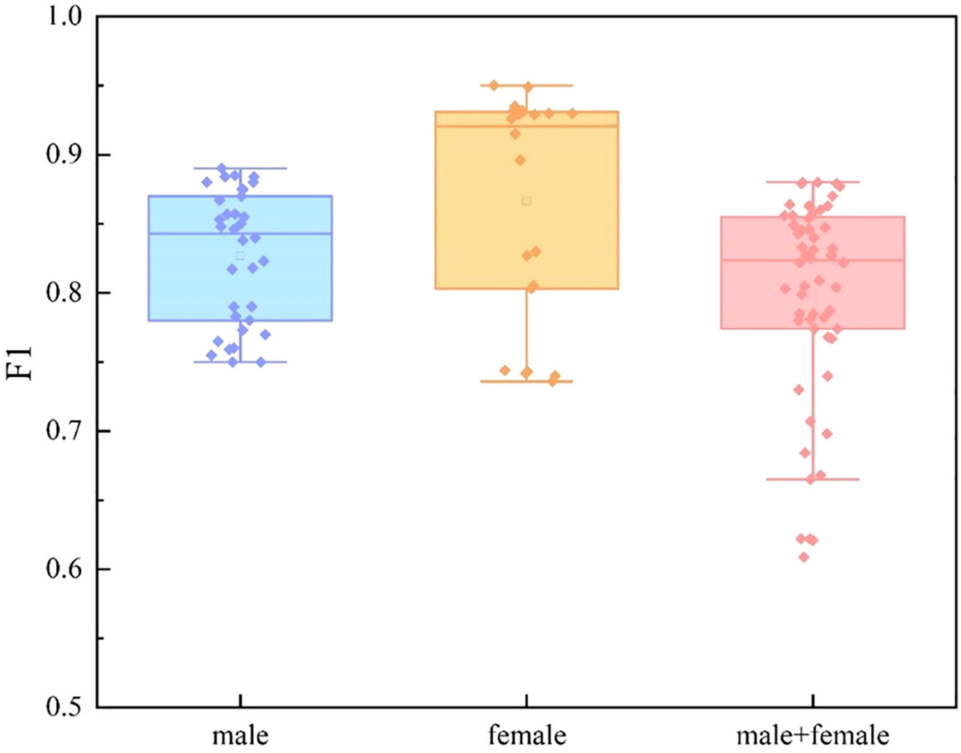

FIGURE 10
Comparison of different subject groups on audio modal features.

sample, and we aggregate every 100 frames into a whole to explain depression detection. As shown in Figure 12, AU contributes to the model diagnostic results to the highest degree, followed by pose and audio, and the smallest contribution is gaze. We further visualize the degree of influence of each AU feature on the diagnostic results of different frames through the contribution matrix. As can be seen in Figure 13, AU04, AU05, AU14, AU15, AU17, AU23, AU26, and AU45 have a higher degree of influence, indicating a stronger correlation with depression. These units thus play a more significant role in the diagnostic process for depression.

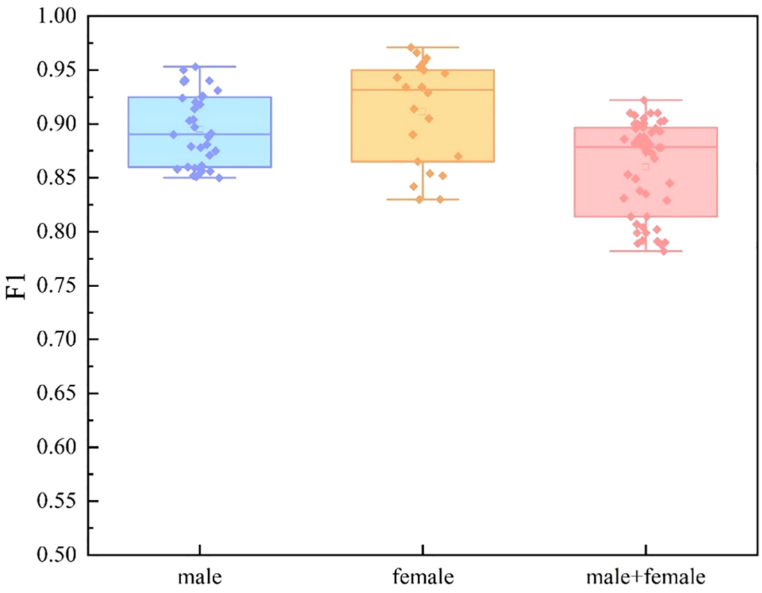

FIGURE 11
Comparison of different subject groups on multimodal features.

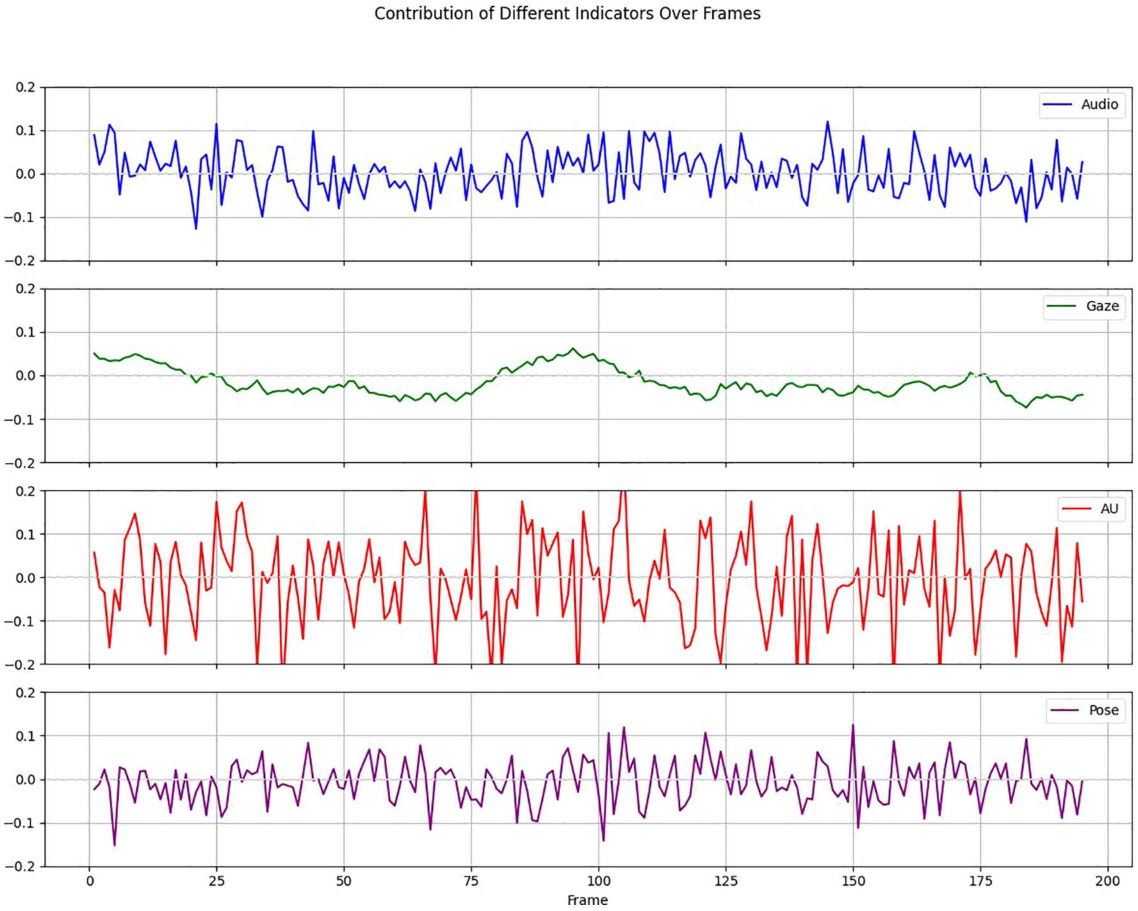

FIGURE 12
Contribution of different indicators over frames.

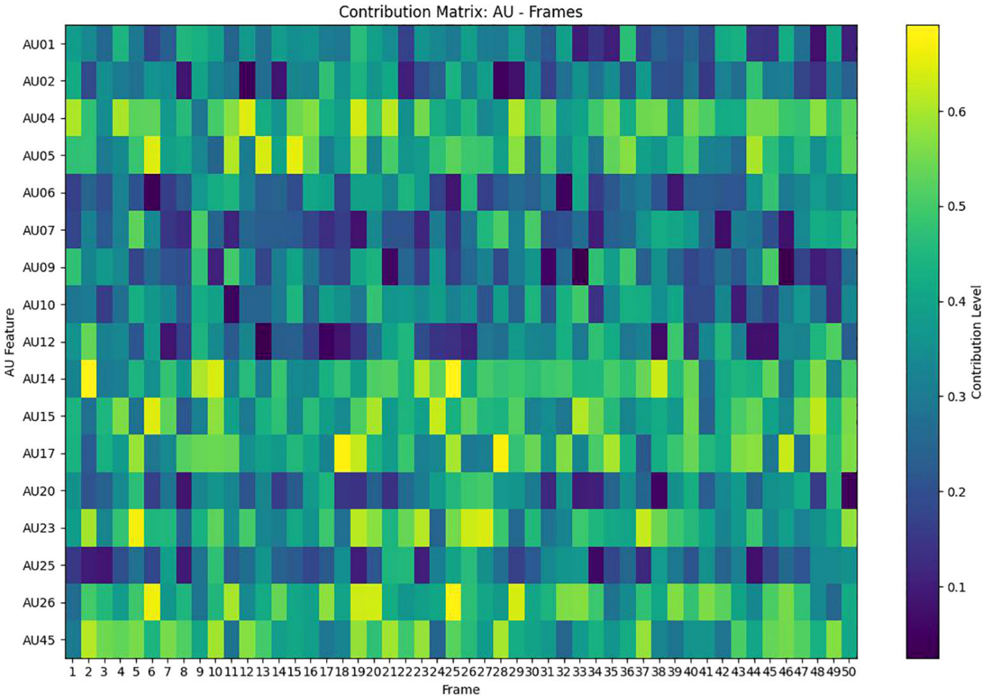

FIGURE 13
Contribution matrix of AU and frames.

4.5 Discussions

This study introduces a novel multi-modal deep convolutional network that leverages multi-source data fusion to provide a more effective solution for the automatic diagnosis of depression. We utilize a feature fusion module to effectively integrate temporal and spatial features, thereby extracting a more comprehensive representation that is conducive to analyzing the psychological state of patients. Our model's complexity is balanced by its non-redundant design, ensuring efficient processing without unnecessary computational overhead. From our comprehensive test results, several noteworthy insights can be gathered. The proposed TSNet-DD model for video data demonstrates significant advantages in capturing both spatial and temporal features. For audio data, the GCN-LSTM model effectively captures complex patterns related to depression. The fusion of video and audio features further improves diagnostic performance, demonstrating the complementary nature of visual and audio features.

However, it is important to acknowledge some limitations inherent in our study. First, due to the lack of publicly available high-quality datasets in this field, our research focuses specifically on E-DAIC datasets. Future research should aim to extend this approach to include different datasets for various psychiatric disorder diagnoses. Additionally, it is important to categorize different levels of depression, which we plan to address in future studies. Furthermore, inspired by ACFun (31) and LMAC-ZS (32), future research could explore the integration of additional data types, such as textual information, into the model to enhance classification performance. This integration could provide the model with a more comprehensive contextual understanding, thereby improving its ability to recognize emotional states. Considering the inherent limitations of deep learning—black-box—we can draw from the methodologies proposed in LMAC-ZS to enhance our model's interpretability. This kind of interpretability mechanism not only contributes to transparency in clinical applications but also provides significant directions for future research.

Overall, our results underscore the importance of multi-modal data in improving the accuracy and reliability of depression diagnosis. Future work should focus on expanding the dataset to include more diverse populations and exploring additional modalities such as text or physiological signals to further enhance diagnostic capabilities.

5 Conclusion

On the premise of protecting the privacy of patients, this paper discusses the method of realizing the high precision diagnosis of depression. We design TSNet-DD architecture for video data to comprehensively consider the temporal and spatial features of video frames through the spatial-temporal attention mechanism module. For audio data, we use a combination of GCN and LSTM to capture high-level graph features and dynamic changes in timing. Finally, the multi-modal feature fusion is realized through the video-audio fusion network. The experimental results show that our method has certain potential in the automatic diagnosis of depression. In the future, researchers can further explore the sensitive features of automatic recognition of depression through larger data sets and more diverse modalities, so as to improve the recognition accuracy and provide more powerful diagnostic and treatment support for clinicians. In addition, it is crucial to classify different levels of depression, which we plan to address in future studies.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://dcapswoz.ict.usc.edu/extended-daic-database-download/ (USC).

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants or their legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

NJ: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. RY: Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. PL: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

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.

References

1. Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med*. (2015) 13:72. doi: 10.1186/s12916-015-0325-4

2. WHO. Depressive disorder fact sheet (2023). World Health Organization. Available online at: https://www.who.int/news-room/fact-sheets/detail/depression/ (Accessed 22 February 2024).

3. Schaakxs R, Comijs HC, van der Mast RC, Schoevers RA, Beekman ATF, Penninx BWJH. Risk factors for depression: differential across age? *Am J Geriatric Psychiatry*. (2017) 25:966–77. doi: 10.1016/j.jagp.2017.04.004

4. Gao S, Calhoun VD, Sui J. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci Ther*. (2018) 24:1037–52. doi: 10.1111/cns.2018.24.issue-11

5. Jan A, Meng H, Gaus YFA, Zhang F, Turabzadeh S. (2014). Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. pp. 73–80.

6. Al Jazaery M, Guo G. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Trans Affect Computing*. (2018) 12:262–8. doi: 10.1109/T-AFFC.5165369

7. Zhuhong SHAO, Yuanyuan S. H. A. N. G. Channel-wise attention mechanism-relevant multi-branches convolutional network-based depressive disorder recognition. *J Image Graphics*. (2022) 27:3292–302. doi: 10.11834/jig.210397

8. Valstar M, Schuller Björn, Smith K, Eyben F, Jiang B, Bilakhia S, et al. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. pp. 3–10.

9. Valstar M, Schuller Björn, Smith K, Almaev T, Eyben F, Krajewski J, et al. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. pp. 3–10.

10. 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

11. Kim JH, Poulose和 A, Ha DS. CVGG-19: customized visual geometry group deep learning architecture for facial emotion recognition. *IEEE Access*. (2024) 12:41557–78. doi: 10.1109/ACCESS.2024.3377235

12. Casado CÁlvarez, Cañellas ML, López MB. Depression recognition using remote photoplethysmography from facial videos. *IEEE Trans Affect Computing*. (2023) 14:3305–16. doi: 10.1109/TAFFC.2023.3238641

13. Ishimaru M, Okada Y, Uchiyama R, Horiguchi R, Toyoshima I. A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network. *Diagnostics*. (2023) 13:727. doi: 10.3390/diagnostics13040727

14. Li Y, Wang Y, Yang Xu, Im S-K. Speech emotion recognition based on Graph-LSTM neural network. *EURASIP J Audio Speech Music Process*. (2023) 2023:40. doi: 10.1186/s13636-023-00303-9

15. Kakuba S, Poulose和 A, Han DS. Deep learning approaches for bimodal speech emotion recognition: advancements, challenges, and a multi-learning model. *IEEE Access*. (2023) 11:113769–89. doi: 10.1109/ACCESS.2023.3325037

16. Uddin MdA, Joolee JB, Sohn K-A. Deep multi-modal network based automated depression severity estimation. *IEEE Trans Affect computing*. (2022) 14:2153–67. doi: 10.1109/TAFFC.2022.3179478

17. Fang M, Peng S, Liang Y, Hung C-C, Liu S. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomed Signal Process Control*. (2023) 82:104561. doi: 10.1016/j.bspc.2022.104561

18. Gimeno-Gómez D, Bucur A-M, Cosma A, Martínez-Hinarejos C-D, Rosso P. Reading Between the Frames: Multi-modal Depression Detection in Videos from Non-verbal Cues. In: *European Conference on Information Retrieval*. Springer Nature Switzerland, Cham (2024). p. 191–209.

19. Sundararajan M, Taly A, Yan Q. (2017). Axiomatic attribution for deep networks. In: *International conference on machine learning*. (pp. 3319–28). PMLR.

20. Li J, Chen Q, Ding Z, Liu Z. Recognition of major depressive disorder based on facial behavior and speech fusion features. *J Beijing Univ Posts Telecommunications*. (2023) 46(1):32.

21. Silva WJ, Lopes L, Galdino MKC, Almeida AA. Voice acoustic parameters as predictors of depression. *J Voice*. (2024) 38(1):77–85. doi: 10.1016/j.jvoice.2021.06.018

22. Ittichaichareon C, Suksri S, Yingthawornsuk T. Speech recognition using MFCC. In: *International conference on computer graphics, simulation and modeling*. Vol. 9. p. 2012.

23. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. *LREC*. (2014), 3123–8.

24. Ringeval F, Schuller Björn, Valstar M, Cummins N, Cowie R, Tavabi L, et al. (2019). AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. pp. 3–12.

25. Eyben F, Weninger F, Gross F, Schuller B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In: *Proceedings of the 21st ACM international conference on Multimedia*. pp. 835–838.

26. Tiwari V. MFCC and its applications in speaker recognition. *Int J Emerging Technol*. (2010) 1(1):19–22.

27. Baltrušaitis T, Robinson P, Morency L-P. (2016). Openface: an open source facial behavior analysis toolkit. In: *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 1–1.

28. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. (2009) 114:163–73. doi: 10.1016/j.jad.2008.06.026

29. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*. (2020).

30. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Dev*. (2014) 7:1247–50. doi: 10.5194/gmd-7-1247-2014

31. Ji J, Wei K, Zhang Z, Deng C. ACFun: abstract-concrete fusion facial stylization. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, .

32. Paissan F, Libera LD, Ravanelli M, Subakan C. Listenable maps for zero-shot audio classifiers. *arXiv preprint arXiv:10.48550/arXiv.2405.17615*. (2024).