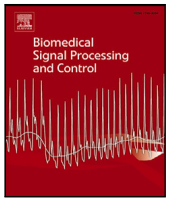
Biomedical Signal Processing and Control 88 (2024) 105675

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



A novel study for depression detecting using audio signals based on graph neural network

Chenjian Sun [a,b], Min Jiang [c,*], Linlin Gao [a,b], Yu Xin [a,b], Yihong Dong [a,b,**]

[a] Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, 315211, China
[b] Zhejiang Key Laboratory of Mobile Network Application Technology, Ningbo University, Ningbo, 315211, China
[c] Ningbo Smart Urban Management Center, Ningbo, 315041, Zhejiang, China

ARTICLE INFO

Keywords:
Automatic depression detection
Graph neural network
Audio recognition

ABSTRACT

Depression is a prevalent mental health disorder. The absence of specific biomarkers makes clinical diagnosis highly subjective. This makes it difficult to make a definitive diagnosis for the patient. Recently, deep learning methods have shown promise for depression detection. However, current methods tend to focus solely on the connections within or between audio signals, leading to limitations in the model's ability to recognize depression-related cues in audio signals and affecting its classification performance. To address these limitations, we propose a graph neural network approach for depression recognition that incorporates potential connections within and between audio signals. Specifically, we first use a gated recurrent unit (GRU) to extract time-series information between frame-level features of audio signals. We then construct two graph neural network modules sequentially to explore the potential connections within and between audio signals. The first graph network module constructs a graph using the frame-level features of each audio sample as nodes. The output is obtained as a graph-embedded feature vector representation after the graph convolution layers. Subsequently, the output graph embedding feature vector representation of the first graph network model is used as the nodes of the graph to construct the second graph network. The internal relationship between audio signals is encoded by the property of node neighborhood information propagation. In addition, we use a pre-trained emotion recognition network to extract emotional features that are highly correlated with depression. By further strengthening the connection weights among nodes in the second graph network through a self-attention mechanism, relevant cues are provided for the model to complete depression detection from audio signals. We conducted extensive experiments on three depression datasets, including DAIC-WOZ, MODMA, and D-Vlog. The proposed model achieves better results on several performance evaluation metrics such as accuracy, F1-score, precision, and recall compared to all the compared algorithms, validating its effectiveness.

1. Introduction

Major Depressive Disorder (MDD) is a prevalent psychiatric condition caused by various factors, including psychological, social, and physical factors. The fast-paced nature of today's society has exacerbated the number of patients suffering from depression, leading to increased rates of suicide [1]. This has resulted in significant burdens for individuals, families, and society at large. According to the World Health Organization, approximately 5% of adults worldwide suffer from this disorder, and it is projected to become the most prevalent disorder by 2030 [2].

Many researchers have utilized neuroimaging techniques for depression identification. Pan et al. [3] used functional magnetic resonance imaging (fMRI) and Seal et al. [4–6] used electroencephalography (EEG) for depression identification. However, considering that most depressed patients have emotional and cognitive problems, this largely reduces their willingness to seek diagnosis and treatment from specialized physicians actively. Therefore, the initial diagnosis of depression is made through the patient's audio signal. It can ensure the privacy of patients to a large extent and motivate them to actively participate in the diagnosis and treatment. At present, clinical interviews and scale tests such as the Hamilton Rating Scale for Depression (HAMD) [7] and the Beck Depression Inventory (BDI) [8] are the most common methods used for diagnosing depression. However, this diagnostic approach is dependent on several factors such as the patient's level of cooperation and expressive skills, the expertise of the physician, and the prevailing treatment environment. These factors may lead to

* Corresponding author.
** Corresponding author at: Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, 315211, China.
E-mail addresses: j3966@163.com (M. Jiang), dongyihong@nbu.edu.cn (Y. Dong).

https://doi.org/10.1016/j.bspc.2023.105675
Received 17 May 2023; Received in revised form 17 October 2023; Accepted 29 October 2023
Available online 4 November 2023

misdiagnosis, highlighting the need for more reliable and objective diagnostic tools.

In recent years, automatic recognition based on audio signals has garnered significant attention in the fields of computer vision and artificial intelligence. Researchers have achieved noteworthy results in emotion recognition, voice classification, and medical applications through the automatic recognition of various audio features. Emil Kraepelin, recognized as the father of modern psychiatry, characterized the voice of a depressed patient as being low, slow, hesitant, monotonous, sometimes stuttering or whispering, and often struggling to articulate words or becoming muted in the middle of a sentence [9]. A substantial body of research has shown [10–12] that there is a strong correlation between MDD and verbal behavior. Depression assessment based on patient-provided audio recordings protects the patient's privacy and encourages voluntary diagnosis and treatment. Consequently, depression detection based on audio signals is gaining attention among scholars. In the literature [13,14], comparing different acoustic features to detect depression has revealed that Mel Frequency Cepstral Coefficients (MFCCs) have high efficiency in detecting depression. Additionally, fundamental frequency (F0) is strongly correlated with social anxiety disorder and can be used as a potential disease-specific physiological marker for the condition [15]. Temporal frequency analysis based on the constant-Q transform (CQT) can provide variable spectral temporal resolution, which contains more information related to emotions [16].

In traditional machine learning methods, audio features are typically fed into classification or regression algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest, Decision Tree, and K-means, among others. For instance, Valstar et al. [17] used SVM linear regression to predict depression, achieving a prediction F1-score of 0.462, while Pampouchidou et al. [18] applied a Decision Tree approach that yielded a test F1-score of 0.52. However, traditional machine learning approaches are constrained by their dependence on prior knowledge of features and shallow model structures, resulting in suboptimal performance in the prediction of MDD. Convolutional neural network (CNN) models can capture the spatial characteristics of features, particularly for spectral, Mel Frequency Cepstral Coefficients (MFCC), or other acoustic features of speech signals. In most CNN-based approaches, high-level features are extracted from low-level audio features such as raw speech signals or spectrograms to perform depression prediction tasks [19,20]. Seneviratne et al. proposed a two-layer neural network architecture that combined CNN and Long Short-Term Memory (LSTM) to explore both spatial and temporal information of audio features [21]. CNNs and recurrent neural networks (RNNs) and their variants have achieved promising results in classification tasks based on audio signals. However, CNNs mainly focus on local regions when dealing with data, while RNNs mainly focus on local dependencies of sequences when dealing with sequence data. This localization leads to some limitations in the exploration of global correlations.

Graph neural networks have emerged as a powerful framework in recent years for using deep learning to directly learn from graph-structured data. Compared to traditional deep learning models, graph convolutional neural network (GCN) can take advantage of the connectivity between nodes to propagate and aggregate feature information from neighbor nodes through graph convolution operations. In this way each node can utilize the information of its neighbor nodes to better capture the node's contextual information in the graph, forming a more comprehensive feature representation and improving the quality of representation learning. As research in graph neural networks has progressed, researchers have started to explore their application in audio recognition tasks. For instance, Shirian et al. [22] used acoustic features of each time frame as nodes in a graph and connected adjacent nodes to enable the model to build graph convolutional networks and perform emotion recognition tasks. Similarly, Niu et al. [23] proposed a graph attention model based on the form of question–answer pairs in clinical interview data and used it for depression detection. As shown in Fig. 1(a), the graph classification models for building graph neural networks on individual audio samples can utilize the connectivity between nodes in the data structure to propagate and aggregate information between nodes through graph convolution operation to capture the global correlation between node features at the frame level. However, the method ignores the intra-class similarity and inter-class variability of audio features. In contrast, Chen et al. [24] constructed graph neural network models that used the features of the whole audio sample as node features, as shown in Fig. 1(b). While this node classification method explored the potential relationship between different samples, it ignored the connection of frame-level features within audio samples. To better mine audio signals for cues related to depression, as seen in Fig. 1(c), we propose a graph neural network method for depression detection based on audio signals. Our approach constructs graph networks between frame-level features within audio signals and between all audio samples, respectively. The intra-audio graph network can propagate and aggregate features from different frame nodes to learn the relationships between audio features at different time frame levels. The inter-audio graph network performs feature propagation and aggregation of feature-embedded representations of all audio to mine intra-class similarities and inter-class differences between different audio samples. The main contributions of our paper are as follows:

(1) In this study, we introduce a novel graph neural network model for detecting depression from audio signals. Our model simultaneously considers potential associations among frame-level features within audio signals, while also accounting for inter-class similarities and differences between audios. To the best of our knowledge, few studies have investigated MDD detection tasks by considering both intra- and inter-audio associations.

(2) In the process of detecting depression, we introduce an emotional pre-training model that leverages a self-attentive mechanism to extract emotional features that are highly correlated with depression. This approach enables the model to better learn depression-related information and ultimately improve its performance.

(3) Our proposed model achieves excellent results on all three datasets, namely, DAIC-WOZ, MODMA, and D-Vlog. These results provide further evidence of the validity and rationality of our approach.

The rest of this paper is as follows: Section 2 presents the related work. Section 3 introduces the graph neural network model proposed in this paper. Section 4 presents the dataset related to this paper and the analysis of experimental results. In Section 5, we provide an in-depth discussion of the proposed method. Finally, the conclusions and future research plans of this paper are discussed in Section 6.

2. Related work

2.1. Deep learning for depression detection

With the widespread success of deep learning in various fields, an increasing number of researchers are applying it to the medical field, leading to groundbreaking advancements. Deep learning methods, in comparison to traditional approaches, can automatically learn high-level abstract features by building multiple hidden layers, determining appropriate parameters without human intervention, and achieving better classification performance and score prediction ability. In the context of classification tasks based on audio signals, such as emotion recognition and MDD prediction, most current methods involve inputting the extracted spectral features and other low-level audio features into a deep learning model to learn high-level feature representations for classification prediction. Deep learning models can learn deeper feature information from audio signals, and their scalability is superior to traditional methods.

In the realm of MDD prediction tasks based on audio signals, common deep learning models include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Deep Belief Networks

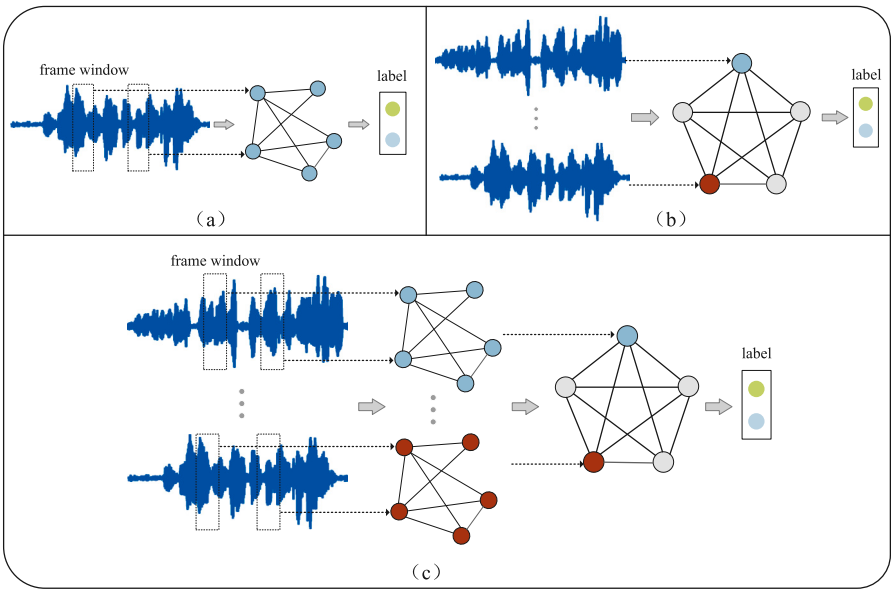


Fig. 1. Comparison of traditional and proposed graph neural network depression detection methods. (a) Graph classification depression classification method. (b) Node classification depression classification method. (c) Proposed depression classification method.

(DBN), among others. CNNs can effectively capture the spatial structure of audio features, making them ideal for advanced feature extraction methods such as spectrum or Mel Frequency Cepstral Coefficients (MFCC). Dong et al. [19] used a ResNet-based approach to learn deep features from the original signal and spectrum. To further learn the relationship between time and frequency from spectral features, Niu et al. [20] proposed a CNN model based on an attentional mechanism, the time–frequency channel attention block (TFCA). It emphasizes the timestamps, frequency bands, and channels associated with depression detection. Convolutional neural network-based methods [19] mainly focus on the spatial structure of spectral features, and [20] focus on temporal information to a certain extent through the attention mechanism. Nevertheless, there is a great limitation to learning long-time series. To circumvent this shortcoming, researchers have used Long Short-Term Memory (LSTM) [25], which provides a long-term memory function to extract the long-term temporal dependency of audio signals. Du et al. [26] proposed a new LSTM module, which combines the Inception module and the LSTM, to adapt to the irregular occurrence of bipolar disorder in different periods. Seneviratne et al. [21] designed a two-layer neural network architecture that extends the CNN-LSTM approach to address the problems of repeated sampling and discontinuous boundaries of neighboring sub-matrices in the traditional approach. However, CNNs and RNNs mainly focus on local dependencies, limiting their exploration of global correlations.

Inspired by the long-term dependency capability of graph neural networks in Natural Language Processing (NLP) tasks, we construct graph structure networks for each audio segment to explore the potential connections between different time-frame level features and reason about the mutual influence between different node features through graph attention networks.

2.2. Graph neural network for depression detection

Subsequently, Graph Neural Networks have achieved remarkable results in processing graph data, such as social networks, traffic networks, and citation networks. Researchers have also begun to explore the application of GNNs in the audio domain. Ji et al. [27] proposed a method based on graph convolutional networks for classifying baby cries. They constructed the graph with weighted edges based on the similarity between relevant nodes and fed it into a convolutional neural network to consider the short- and long-term effects of infant cry signals. Jung et al. [28] proposed the AASIST model, which is a graph neural network-based model for recognizing speech spoofing. The model is designed mainly as a heterogeneous stacked graph attention layer that uses heterogeneous attention mechanisms and stacked nodes to model time and frequency, considering both temporal and frequency dimensional features of audio data. Nie et al. [29] proposed a new correlation-based graph convolutional network (C-GCN) for automatic emotion recognition. They introduced a graph model to represent the correlation between videos while applying a multi-head attention mechanism to explore the hidden relationships between videos to enhance the correlation between classes. Shirian et al. [22] proposed a compact and efficient graph structure to represent audio data. The model models the speech signal as a cyclic or line graph, i.e., the acoustic features of each time frame of speech are treated as nodes in the graph, and neighboring nodes are connected, enabling the model to build a graph convolutional network-based architecture. However, this approach ignores the time series information in audio features. To address this, Liu et al. [30] proposed an automatic emotion recognition model, LSTM-GIN, which uses a Long Short-Term Memory (LSTM) to extract the time series information in the audio signal and uses it as an input to a graph isomorphic network for global emotion modeling.

Given the promising results achieved by Graph Neural Networks in audio data, researchers have extended their use to disease prediction tasks in the medical field. Rezaee et al. [31] proposed a hybrid approach by extracting depth features from heart sound signals and classifying them. Deep GCN attempts to determine the association between cardiovascular disease (CVD) and spectrograms to better identify CVD signals. Ghadiri et al. [32] proposed to enrich speech analysis for depression detection by fusing two methods, graph transformations of speech signals and representation learning of natural language processing. Niu et al. [23] proposed a hierarchical context-aware model based on a graph attention network (HCAG) for depression detection. HCAG formulates a clinical interview into several question–answer pairs that reflect the structure of depression assessment. Specifically, the hierarchical context-aware structure captures key information in the answers, while the GAT network can further aggregate sufficient relational and logical information between the interview questions. However, for each respondent, the questions to be answered are the same, so it is worth considering whether the same questions are effective in differentiating respondents' depressive conditions. Considering that most

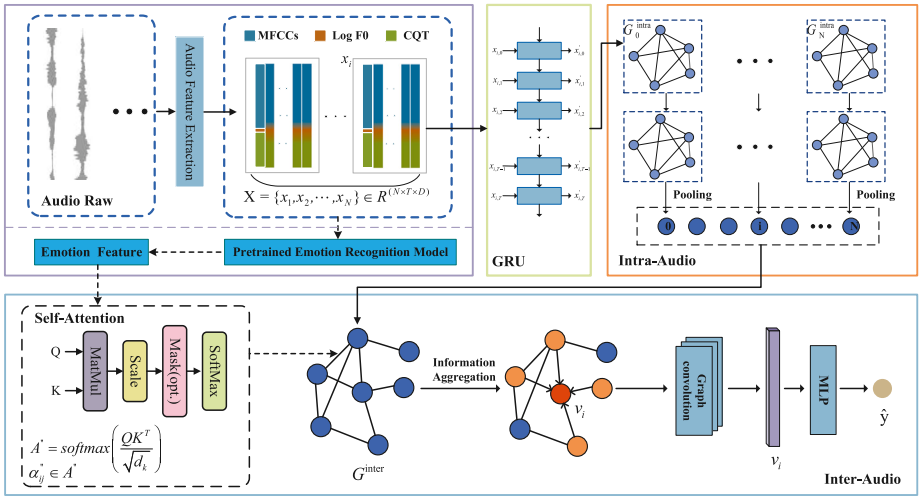


Fig. 2. General framework diagram of the proposed model in this paper. The model first extracts frame-level depression features and mood features from the audio's low-level features then constructs a graph convolutional network from the obtained embedded vector feature representations to extract feature information and perform depression assessment.

existing methods only consider intra-audio associations and ignore the heterogeneity between audio samples, Chen et al. [24] proposed a multimodal fusion model based on graph neural networks (MS2-GNN) to explore the heterogeneity/homogeneity among various psychophysiological modalities by fusing the commonalities and characteristics among different modalities and to explore the potential relationships among subjects using nodal classification methods. MS2-GNN performs a simple flattening operation on the features of different temporal frames within the audio, ignoring the connection of frame-level features within the audio samples.

Most of the above methods unilaterally explore the potential relationship between patients and diseases from audio signals, do not consider the correlation between audio and audio at the same time, and do not fully explore the MDD-related clues in the subject's audio signals.

3. Methods

Our proposed approach for depression detection of audio signals is based on a graph neural network, and the general framework of our model is presented in Fig. 2. Initially, we preprocess the audio data and extract the low-level audio features, which serve as the input of the Intra-Audio and Emotion-Feature modules. The Intra-Audio module comprises gated recurrent units (GRUs) and GCNs, which extract time series information within audio features and potential associations of frame-level features and output an embedded feature vector representation of each sample. The Emotion-Feature module is a pre-trained emotion feature extraction network that extracts depression-related emotions from low-level audio features. Subsequently, we construct graph neural networks from the outputs of the Intra-Audio and Emotion-Feature modules to explore the inter-class similarity and differences of the audios. Specifically, we utilize embedding feature vector representations as nodes, construct edges using node feature similarity, and strengthen the connections between nodes by extracting depression-related cues from emotional features through a self-attentive mechanism. Finally, we input the extracted high-level features into the MLP layers for the final depression prediction.

3.1. Data preprocessing

For audio recordings of clinical interviews, our focus is on detecting depression from the patient's voice signal alone, and thus we do not consider the semantic information of the interviewer's voice. Consequently, we remove the interviewer's audio fragments and retain only the participant's audio fragments for depression detection. To maintain the continuity of the audio, we do not stitch the voice fragments after removing the interviewer, but instead perform audio feature extraction on the participant's audio alone. For non-interview datasets, audio feature extraction is performed directly. The low-level audio feature extraction process involved sampling the audio signal using Librosa, an audio signal processing library in Python, with a fixed frame window and a frame interval at the original audio sampling rate. To ensure the duration of all discourses was consistent, we extracted time-fixed audio features for each audio. We extracted frame-level audio features for each audio, including MFCCs, logarithmic fundamental frequency (log F0), and constant-Q transform (CQT) features.

3.2. Emotional feature

Psychological studies have suggested that depressed emotions can directly affect an individual's emotional expression and perception and that cognitive biases and deficits caused by depression can affect emotion regulation abilities, such as the habitual upregulation of negative emotions during emotional expression. Considering the effect of depression on mood, we use CompactSER [22], a compact graph-structured emotion recognition network that has been pre-trained with the dataset IEMOCAP, to extract emotional features from audio signals.

In CompactSER, for a given set of audio discourse data, feature extraction, and graph structure transformation are performed to obtain a set of graphs $\{G_1, G_2, \ldots, G_N\}$ with their true labels $\{y_1, y_2, \ldots, y_N\}$. The specific process is as follows:

$$\begin{aligned} \mathrm{H}^{(k+1)} &= \mathrm{U}\left(\mathrm{MLP}\left(\mathrm{U}^{T}\mathrm{H}^{(k)}\right)\right), \\ \mathcal{L} &= \mathrm{D}^{-\frac{1}{2}}\mathrm{L}\mathrm{D}^{-\frac{1}{2}}. \end{aligned} \tag{1}$$

where MLP refers to a multilayer perceptron whose parameters are learnable, $\mathcal{L}$ is the Laplacian matrix of the normalized graph, D is the degree matrix, $L = D - A$, where A is the adjacency matrix of the graph. λ_ς is the ς-th eigenvalue of the Laplacian matrix L corresponding to the eigenvector u_ς, $\Lambda = \mathrm{diag}\left(\lambda_\varsigma\right)$, and $U = [u_1, u_2, \ldots, u_N]$.

The goal is to classify the entire graph. Therefore, a function is needed to obtain the graph-level embedding representation $h_{G_i} \in R^m$ from the node-level embedding, which can be obtained by pooling the node-level embedding $H_i^{(K)}$ at the last layer before passing it to the classification layer.

$$h_{G_i} = \mathrm{sumpool}\left(H_i^{(K)}\right) = \sum_{j=1}^{M} \mathrm{h}_j^{(K)}. \tag{2}$$

The graph embedding layer is followed by a fully connected layer that is used as the classification layer of the model and the model is trained using a cross-entropy loss function. The trained model is removed from the classification layer and used as a sentiment feature extractor for our model.

3.3. Intra-audio correlation

3.3.1. Time series information

In this paper, the initial audio features are defined as $X = \{x_1, x_2, \ldots, x_N\} \in R^{N\times T\times D}$, where N denotes the number of samples, T denotes the timestamp of each audio sequence, and D denotes the audio feature dimension of each frame. To avoid the gradient disappearance and explosion problems of traditional RNNs in the backpropagation process, this paper uses the GRU to extract the time series information of audio signals. The input of the GRU-based time series extraction network is as follows:

$$x'_i = GRU\left(\left[x_{i,1}, x_{i,2}, \ldots, x_{i,T}\right]\right), \quad i \in \{1, 2, \ldots, N\}. \tag{3}$$

where i denotes the ith sample. $x'_i \in R^{T\times D'}$ is the feature matrix extracted by the GRU network. T is the number of frames in the feature sequence. D' is the dimension of the characteristics of the audio.

3.3.2. Intra-audio graph convolutional network

Building graph: building the graph for the ith audio sample $G_{\text{intra}}(i) = \left\{\left(V_{\text{intra}}(i), E_{\text{intra}}(i)\right) \mid i \in \{1, 2, \ldots, N\}\right\}$, where $V_{\text{intra}}(i)$ is the set of all nodes, each audio signal frame constitutes a node in the graph, and the frame-level acoustic features are used as node feature vectors. $E_{\text{intra}}(i)$ is the set of all edges between nodes. In this paper, two time windows, p and s, are used to construct edges. For each discourse vertex v, it is connected to the first p time frame nodes and the last s time frame nodes.

To explore the association between different time-frame level features within audio, this paper establishes the connection between different time frames by graph convolutional neural network, to explore their potential connections. The information propagation process of the graph neural network is to propagate information from neighboring nodes to the target node through predefined graphs. We establishes the connections between different time frames based on the time sequence of frames and determines the connection strength through the attention mechanism.

We use the graph attention network (GAT) to propagate the neighboring node information. Specifically, GAT takes all nodes as input and updates the target node features through the neighboring nodes $\mathcal{N}_i$ in the graph. The process of updating node h_i^{l+1} is as follows:

$$h_i^{l+1} = \sigma\left(\sum_{j\in\mathcal{N}_i} w_{ij} W_1^l h_j^l\right), \quad i \in \{1, 2, \ldots, T\}. \tag{4}$$

where W_1^l is the trainable weight matrix, l represents the lth graph convolution layer, h_i^{l+1} is the node representation of the ith time-frame node of the current sample after the update of the lth graph convolution layer, σ is the nonlinear activation function Relu, and w_{ij} is the attention coefficient between node i and node j. The calculation process is as follows:

$$w_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a\left[W_2 h_i \parallel W_2 h_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(a\left[W_2 h_i \parallel W_2 h_k\right]\right)\right)}. \tag{5}$$

where $a \in R^{2D'\times 1}$ and $W_2 \in R^{D'\times D'}$ are the learnable training parameters, which are calculated and normalized to obtain the weight coefficients between node i and node j.

After the graph convolution layer, each frame node obtains enough information from its neighboring time-frame nodes, and to obtain the feature representation of the whole graph, we use graph averaging pooling to obtain the graph embedding representation. The embedding representation z_i for each audio sample is as follows:

$$z_i = \text{mean}\left(\left\{h^L \mid h^L \subseteq G_{\text{intra}}(i)\right\}\right), \quad i \in \{1, 2, \ldots, N\}. \tag{6}$$

3.4. Inter-audio correlation

3.4.1. Inter-audio graph convolutional network

Building graph: after going through the Intra-Audio module, the embedding representation of each sample was obtained. We reconstruct a graph $G_{inter} = (V_{inter}, E_{inter})$ with the embedding features of N samples, where V_{inter} is the set of all nodes. E_{inter} denotes the set of edges between all nodes. The adjacency matrix of G_{inter} is represented as $A' \in R^{N\times N}$, where α'_{ij} is the element corresponding to the adjacency matrix A', which denotes the weights of the edges between node i and node j.

To obtain the topology of the feature space and to be able to better capture the dependencies between different nodes in the graph structure and avoid unrelated neighborhood interference. In this paper, we use cosine similarity to construct the graph structure, and the process of generating a graph based on cosine similarity is as follows:

$$\begin{aligned} \alpha'_{ij} &= \frac{z_i^T z_j}{\|z_i\| \times \|z_j\|}, \\ \alpha'_{ij} &= \begin{cases} \alpha'_{ij}, & \text{if } \alpha'_{ij} \geq \varepsilon \\ 0, & \text{otherwise} \end{cases}. \end{aligned} \tag{7}$$

where z_i denotes the feature vector representation of the ith node, ε is the threshold hyperparameter, and α'_{ij} is the cosine similarity of node i to node j, both the weights of node i and node j edges.

3.4.2. Feature weight fusion

To fully and flexibly integrate emotional features and depressive features, we use the self-attentive mechanism to calculate the potential connections between emotional features among different samples and fuse them into depressive features. Compared with simple feature splicing, our weight fusion method is more flexible, not only can fully explore depression cues and emotion cues among different samples but also can flexibly adjust the weight proportion of both, thus can more fully fuse depression features and emotion features. The process of the self-attentive mechanism is as follows:

$$\begin{aligned} A'' &= \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right), \\ Q &= \left[h_{G_1}, h_{G_2}, \ldots, h_{G_N}\right]^T W_Q, \\ K &= \left[h_{G_1}, h_{G_2}, \ldots, h_{G_N}\right]^T W_K. \end{aligned} \tag{8}$$

where W_Q and W_K are learnable parameters. Q, K and d_k are the Query, Key vectors and their feature dimensions we defined. A'' is the weight matrix of the feature vectors, α''_{ij} denotes is the value of A'' corresponding to row i and column j, both the weight between node i and node j.

The next key step is features aggregation, where a node aggregates information from its neighboring nodes to the target node. In this paper, the graph convolution layer is a spectral-based graph convolution that combines the entire structure and individual components, using the Chebyshev spectral graph convolution operator proposed by Defferrard et al. [33]. Spectral-based graph convolution can be defined as the product of a signal $x \in R^m$(scalar at each node) with a filter $g_\theta = diag(\theta)$, parameterized by $\theta \in R^m$:

$$g_\theta \star x = U g_\theta(\Lambda) U^T x, \tag{9}$$

where $\star$ is the convolution operator on the graph. U is the matrix constructed by the eigenvectors of the Laplace matrix $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U\Lambda U^T$ where I is the identity matrix. And using the K-order Chebyshev polynomials to approximate $g_\theta(\Lambda)$, expressed as:

$$g_\theta \star x \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L}) x, \tag{10}$$

where $\tilde{L}$ is the rescaled graph Laplacian, θ_k are the trainable parameters. $T_k(\tilde{L}) = 2\tilde{L}T_{k-1}(\tilde{L}) - T_{k-2}(\tilde{L})$ with $T_0(\tilde{L}) = 1$ and $T_1(\tilde{L}) = \tilde{L}$. Suppose we restrict the hierarchical convolution operation to $K = 1$, i.e., a linear function, and hence a linear function on the Laplace spectrogram. Under these approximations, the expression reduces to:

$$g_\theta \star x \approx \theta_0 x + \theta_1 (L - I)x = \theta \left(I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right) x, \tag{11}$$

with a single parameter $\theta = \theta_0 = -\theta_1$. Thus the graph convolution layer receives the node feature matrix H^l and the association matrix A. The final expression for updating the node features is as follows:

$$\begin{aligned} H^{l+1} &= \sigma \left(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} H^l W^l \right), \\ A &= \beta A' + (1 - \beta) A''. \end{aligned} \tag{12}$$

where $\widetilde{A} = A + I$ is the adjacency matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$, W^l is the learnable parameter, A' and A'' are the weight matrices of depression features and emotion features, respectively, and β is an importance coefficient ranging from 0 to 1, a hyperparameter.

3.4.3. Classifier

After the multilayer graph convolution layers, the learned features are fed into our classifier for depression detection. In this paper, the MLP layer is finally used as our final classifier:

$$\hat{y} = \mathrm{softmax} \left(MLP \left(H^L \right) \right). \tag{13}$$

For this task depression detection is a binary classification task, so we use cross entropy as a loss function to train the model parameters:

$$\mathrm{loss} = \sum_{c \in [0,1]} \mathrm{P}(c \mid \mathrm{y}) \log \mathrm{P}(\mathrm{c} \mid \hat{y}). \tag{14}$$

where 0 and 1 denote the sample labels of health and depression, respectively, $\mathrm{P}(c \mid \mathrm{y})$ is the true label distribution, and $\mathrm{P}(\mathrm{c} \mid \hat{y})$ is the estimated probability distribution of label c.

3.5. Algorithm and complexity analysis

In this part, we give the overall framework algorithm of the proposed method in this paper, as shown in Algorithm 1. The time series information of audio features is first extracted by GRU. Then the graph attention network is utilized to extract the potential relationships between the features of different time frames within the audio. The obtained embedded feature vector representations are used as nodes to construct the graph neural network again, and the edges are constructed using the depression features and emotion features. Finally, the extracted high-level features are fed into the MLP layer for final depression prediction. The main time complexity lies in the optimization of step 6, which has a complexity of $O(N L_1 |E_{intra}| D)$. The second is the optimization for the step 12 part, which has a complexity of $O(L_2 |E_{inter}| D)$. L_1 and L_2 denote the number of layers of the two graph convolutional networks, and D is the feature dimension.

Algorithm 1 The learning process of the method in this paper.

Input: low-level audio features $X = \{x_1, x_2, \cdots, x_N\} \in R^{N \times T \times D}$, graph convolution layers L_1 and L_2.
Output: predicted labels $\hat{y}$.

1: initialize the parameters θ;
2: X extracts the sentiment features h_G by the pre-trained emotion feature extractor;
3: extract time series information by Eq. 3.
4: for l =1 to L_1 do
5: for i=1 to N do
6: perform operation $h_i^l = \mathrm{GAT}(h_i^{l-1})$ by Eq. 4,5 where $h_i^0 = x_i$;
7: end for
8: end for
9: use average pooling to get all sample embedding representations z of frame-level features after the graph attention layer;
10: use the self-attention mechanism to extract the weight matrix A'' for the emotional features. The extracted depression feature weights A' and emotion features are weight fused to get the final weights A by Eq. 12;
11: for l=1 to L_2 do
12: perform operation $H^{l+1} = GCN\left(AH^l\right)$ by Eq.12 where $H^0 = z$;
13: end for
14: Calculate the predicted label $\hat{y}$;
15: Calculate the cross-entropy loss function loss, backpropagate and update the parameters θ;
16: **return** $\hat{y}$.

4. Experiments

4.1. Datasets

DAIC-WOZ [34]: the DAIC-WOZ is a clinical interview dataset for depression, which contains 189 subjects. The interview transcripts were presented as a question and answer session between Ellie, the virtual interviewer, and the subject, and each subject's transcript had a PHQ-8 score representing depression status and contained a binary label indicating depression and health: 133 depressed (PHQ-8 $\geq$ 10) and 56 healthy (PHQ-8 $<$ 10). For this experiment, we extracted individual audio segments for each interview transcript, resulting in a total of 4903 audio samples. To address the issue of imbalanced depression and health ratios in the dataset, we randomly down-sampled all audio discourse. This resulted in a dataset of audio samples from 2914 subjects, which was used for further analysis in our experiment.

MODMA [35]: the MODMA is an open dataset for psychiatric disorder analysis published by Lanzhou University. It contains 23 subjects with MDD and 29 healthy subjects. MDD patients were recruited from inpatients and outpatients who met the diagnostic criteria for depression, the Diagnostic and Statistical Manual of Mental Disorders (DMS), and healthy controls were recruited through posters and excluded from other disorders. Each subject was required to complete 29 audio recording tasks through interviews, readings, and picture descriptions in different mood states, with the duration of each recording varying from a few seconds to several tens of seconds. From these, 6 abnormal sample files were excluded, resulting in a final sample of 1502 (661 depressed and 841 healthy).

D-Vlog [36]: the D-Vlog was composed by Jeewoo et al. through 961 Vlogs (approximately 160 h) posted on YouTube between 2020 and 2021. Depressed videos were identified by the keywords "depression daily vlog", "depression journey", "depression vlog", etc., and non-depressed videos were identified by the keywords 'daily vlog', 'day of vlog', 'talking vlog', etc. We ensured that the sample videos were in Vlog format (a person speaking directly to the camera), and examined the transcripts automatically generated from the video voice content to determine whether the speaker was depressed, with 555 depressed samples and 406 healthy samples. Due to privacy issues by design, we did not obtain the original audio for this dataset. Therefore, the relevant experimental content was completed using the low-level audio features provided by the authors.

IEMOCAP [37]: the IEMOCAP is a sentiment database with a wide range of applications in automatic emotion recognition, collected by the SAIL lab at the University of Southern California. The corpus involves 10 different participants and contains approximately 12 h of data, with 5 binary conversations on 10 topics. Each of these binary conversations was further divided into discourses, and each of these discourses was tagged with an emotion label: happy, sad, neutral, angry, excited, and frustrated. This paper uses this dataset to train our emotion feature extractor.

4.2. Evaluation metrics

To evaluate the performance of the model, we use the classification performance metrics of accuracy (acc), precision (pre), recall (rec), and F1-score to evaluate the model performance. The evaluation metrics are defined as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+FP+TN+FN} \\ \text{Precision} &= \frac{TP}{TP+FP}, \\ \text{Recall} &= \frac{TP}{TP+FN}, \\ F1 &= \frac{2\times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}}. \end{aligned} \tag{15}$$

Where TP (True Positive) and TN (True Negative) denote correctly predicted positive and negative samples, respectively, and FP (False Positive) and FN (False Negative) denote incorrectly predicted samples into positive and negative classes, respectively. F1-score combines precision and recall, and when the F1-score value is higher, the classification ability is better.

4.3. Results

4.3.1. Comparison with existing methods

SVM [17]: this method uses a traditional machine learning support vector machine (SVM) to perform a linear.

Depaudionnet [38]: this method combines a one-dimensional convolutional neural network CNN and LSTM to encode depression-related features in the vocal tract, and extracts depression-related cues from the spectrogram for depression detection.

Yoon et al. [36] used acoustic and visual features extracted from data to detect depression and to effectively capture the relationship between different modalities, the model used a cross-attention mechanism to learn feature representations.

MSCDR [39]: MSCDR extracts linear predictive coding (LPC) and Mel cepstral coefficients (MFCC) features to describe the process of speech generation and speech perception, respectively, and then captures depressive features sequentially through the one-dimensional convolutional neural network and the long short-term memory network for classification.

DEPA [40]: the DEPA uses a self-supervised and pre-trained approach to learn the embedding representation of depressed audio and uses the trained encoder to do the downstream depression detection task.

Ghadiri et al. [32] enriched depressive cues in speech signals by fusing two methods, graph transformation and natural language processing representation learning, which were fused to generate the final class labels.

HCAG [23]: HCAG is a hierarchical context-aware graph attention model for MDD that constructs graph models of clinical interview transcripts as sequences of question–answer pairs to capture contextual information more effectively.

MS^2-GNN [24]: MS^2-GNN explores the heterogeneity/homogeneity among various psychophysiological modalities by fusing the commonalities and characteristics among different modalities and uses node classification to explore the potential relationships among subjects.

Chen et al. [41]: this method uses speech data to construct a decision tree model for depression screening. The Chi-squared automatic interaction detector (CRT) algorithm and the classification and regression tree (CHAID) algorithm with better results were selected to construct the decision tree model according to the type of data.

TAMFN [42]: to fully mine and fuse multimodal features, TAMFN captures more temporal behavioral information by combining local and global temporal information and mines the temporal importance between different modalities to guide multimodal feature fusion.

CAIINET [43]: CAIINET proposes a neural network based on contextual attention and information interaction mechanisms, correlations, and interactions between acoustic and visual features extracted at local and global scales, to carry out depression detection tasks.

We evaluated our model on three depression datasets, namely DAIC-WOZ, MODMA, and D-Vlog depression data. In this experiment, to obtain a stable and reliable model, we adopted a ten-fold cross-validation method to randomly divide the data into 10 copies, with the training, validation, and testing data assigned in the ratio of 8:1:1. Finally, we obtained the average and standard deviation of the 10 results. Table 1 presents the performance of the proposed methods in this paper and the comparison methods on different datasets. All of the results of the comparison methods in which the relevant code is not available are taken from the results presented in their original paper. Compared with the existing methods, our methods demonstrate better performance on all evaluation metrics across the three datasets.

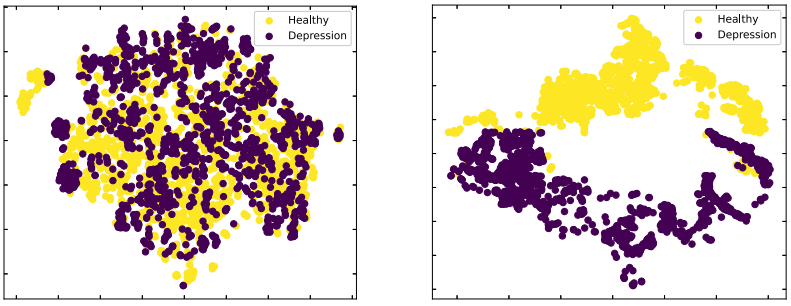
(i) Comparison of non-graph neural network methods: for the dataset DAIC-WOZ, one of the best classification performances of non-graph convolutional methods is DEPA [40], with an F1-score of 90%. Compared to other traditional machine learning [17] and neural networks [36,38,39] comparison methods performance is improved by up to 33.8%, the effect is the best besides us. It illustrates that DEPA [40] predicts the reconstructed spectrogram center segment by training the encoder–decoder model to be able to learn the contextual information in the given spectrogram very well, and thus obtain a more efficient representation of the audio embedding for downstream classification tasks. In contrast, our approach not only learns effective audio embedding representations but also captures dissimilarities among audio through information propagation aggregation among audio embedding nodes, which further improves the overall evaluation metric F1-score by 2.23%. The best of our proposed methods in MODMA compared to non-graph neural network methods similarly improves by 5.15% on the comprehensive evaluation metric F1-score. The 31.92% improvement in performance on the dataset D-Vlog can be attributed to the fact that this data is one of the latest depression datasets on which there is very little relevant work for the time being, and secondly, we are the first relevant work to use a graph neural network approach on this dataset, thus illustrating the effectiveness of graph neural networks.

(ii) Comparison of graph neural network methods: among all the compared methods Ghadiri et al. [32], HCAG [23] and MS2-GNN [24] are graph neural network methods. Whether comparing the graph classification methods [23,32] or the node classification methods [24], the methods proposed in this paper have improved the performance of the depression classification task. HCAG [23] uses clinical Q&A transcripts to construct a hierarchical graphical attention model with a sequence of Q&A pairs, which enhances the model's ability to retrieve depression cues from the contextual information of the speech signal. However, the fact that all the questions in the Q&A were from the interviewer rather than the patient and were all the same is questionable as to whether the model is effective in recognizing the depressive condition of the interviewee. In contrast our model avoids possible negative effects and can take into account potential links between the audios, improving by 12.23% in the F1-score. Our model outperforms the best comparison method MS2-GNN [24] in dataset DAIC-WOZ and dataset MODMA in accuracy by 3.08% and 3.86% respectively. Compared to MS2-GNN [24], which simply extracts the time series in the features and then performs a spreading operation on the features to get the audio embedding representation, our method fully exploits the correlation of different time frames within the audio and can learn a better audio embedding representation.

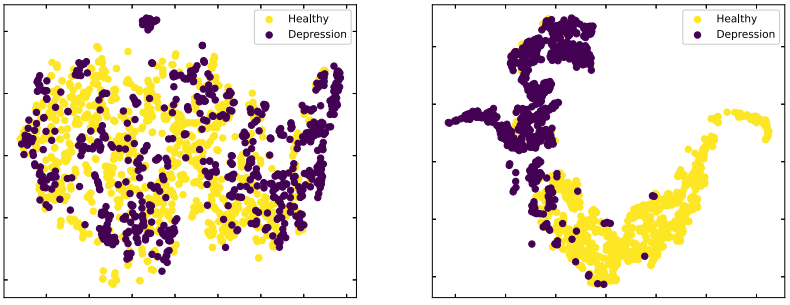
Our model outperforms all compared graph neural network methods and non-graph neural network methods. The main reason is that our model can fully mine the depression-related cues in the patient's speech signal from both intra-audio and inter-audio and at the same time, our emotion feature module can further help the model to mine the correlation of inter-audio depression cues and improve the depression recognition ability of the model.

Table 1
Depression detection results of the method proposed in this paper and other comparative methods on the data set DAIC-WOZ, MODMA, and D-Vlog, where the best results are shown in bold.

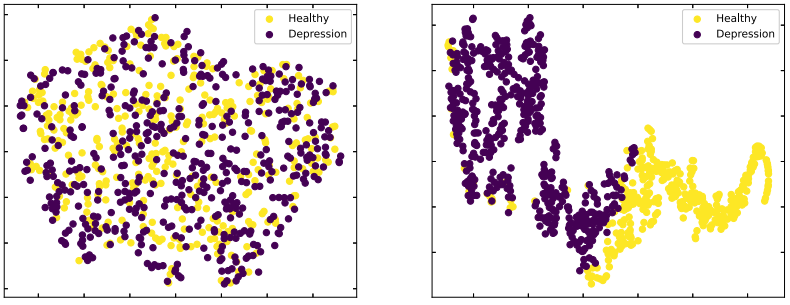
Dataset	Method	Acc (%)	Pre (%)	Rec (%)	F1 (%)
DAIC-WOZ	SVM [17]	58.93 ± 2.69	59.86 ± 2.73	57.42 ± 4.18	57.42 ± 4.18
	Depaudionnet [38]	72.13 ± 2.35	70.97 ± 3.27	75.72 ± 2.64	73.26 ± 2.95
	Yoon et al. [36]	–	62.57	52.63	55.45
	MSCDR [39]	77.1	–	–	66.0
	DEPA [40]	–	91.0	89.0	90.0
	Ghadiri et al. [32]	61.0	61.1	66.7	63.4
	HCAG [23]	–	77.0	83.0	80.0
	MS2-GNN [24]	89.13	80.0	85.71	82.76
	Ours	**92.21 ± 1.86**	**92.36 ± 2.53**	**92.18 ± 1.55**	**92.23 ± 2.01**
MODMA	Chen et al. [41]	83.4	83.5	76.8	80.0
	MSCDR [39]	85.7	–	–	84.0
	MS2-GNN [24]	86.49	82.35	87.5	84.85
	Ours	**90.35 ± 2.46**	**88.25 ± 3.26**	**90.33 ± 3.67**	**89.15 ± 2.89**
D-Vlog	Yoon et al. [36]	–	65.4	65.57	63.5
	TAMFN [42]	–	66.02	66.5	65.82
	CAIINET [43]	–	66.57	66.98	66.56
	Ours	**93.91 ± 1.43**	**91.9 ± 1.92**	**98.48 ± 1.34**	**95.05 ± 1.19**



(a) DAIC-WOZ(Initial representation) (b) DAIC-WOZ(After learning representation)



(c) MODMA(Initial representation) (d) MODMA(After learning representation)



(e) D-Vlog(Initial representation) (f) D-Vlog(After learning representation)

Fig. 3. The original features of three datasets DIAC-WOZ (a), MODMA (c), and D-Vlog (e) visualized by t-Distributed Stochastic Neighbor Embedding(t-SNE) projection onto a two-dimensional space. (b), (d) and (f) are the embedding representations of the datasets DIAC-WOZ, MODMA, and D-Vlog, respectively, after learning from the model proposed in this paper. Where the yellow nodes indicate healthy samples and the purple nodes indicate depressed samples.

4.3.2. Visualization

We performed feature visualization analysis on three datasets and visualized the original features of the three datasets and the embedding representations after model learning in this paper by the t-SNE algorithm, respectively. These visualizations are presented in Fig. 3. The discrete distribution of the raw features of the three datasets is evident from (a), (c), and (e), with different types of node features overlapping and interspersed without clear delineation. This reflects the difficulty in distinguishing between the features of the depressed group and the features of the healthy control group. However, the visualization of the embedding representation of the original features after learning by our model is shown in (b), (d), and (f), and it is clear that the distribution of features in the depression group and the healthy control group has significantly improved. This improvement increases the intra-class similarity and inter-class separability to a large extent, enabling clearer division of nodes from different classes. These results demonstrate the effectiveness of our proposed model.

4.4. Ablation experiments

In this section, we examine in detail the Intra-Audio module, Emotion-Feature module, and Inter-Audio module of the proposed model in this paper to better understand the contribution of each module of the model. The comparative changes in the model are shown below.

w/o audio_intra: the model in this mode will remove the Intra-Audio module, and we directly take the last hidden layer state of the GRU output instead of the graph embedding representation module.

w/o audio_inter: In this case, the model will remove the Inter-Audio module and just replace it with an FC layer as the classifier, the rest of the modules remain the same.

w/o emotion: In this case, the model removes the pre-trained emotion feature extractor, the model will not fuse the emotion features, and the rest of the modules remain consistent.

w/o self-attention: The model does not use the self-attention weight fusion method proposed in this paper in that case. It is replaced with a simple concatenated link for fusing depression features and emotion features, and the rest of the modules are kept the same.

We conducted ablation experiments between different modules on the three datasets to verify the usefulness of different modules for our proposed model. Table 2 demonstrates the comparison results between the full model and different versions of the model in different performance metrics, and it can be seen that the full model outperforms the other versions in all metrics. The performance of the full model is compared with other versions on the datasets DAIC-WOZ, MODMA, and D-Vlog. Comparing w/o audio_intra version, the full model improves in accuracy by 4.23%, 1.94%, and 3.27% respectively. Compared to the w/o audio_inter version, the full model improves the accuracy by 10.65%, 5.59%, and 18.27% respectively. The model in this paper improves the accuracy by 0.92%, 0.74%, and 1.91% compared to the w/o emotion version. The performance of the self-attention fusion module of our model improves the accuracy of the three datasets by 0.58%, 1.55%, and 1.27%, respectively, compared to the simple splicing feature fusion approach. Our model not only considers the potential connection between frame-level features within the audio signal and between the audio signals but also the emotional features that are highly correlated with depression can enhance the model's ability to capture depression-related cues in the audio signal.

4.5. Parameter study

4.5.1. β-value

In this section, we first analyze the hyperparameter β in detail. In our proposed model, β represents the weight magnitude of depression features and emotional features in the connection between different audio signals. Selecting an appropriate fusion ratio is critical in exploring depression cues from emotional features. To this end, we conducted several experiments on the three datasets, and the impact of different values of the hyperparameter β on the model's performance is presented in Fig. 4.

Table 2
Ablation experiments on each module, the best results are shown in bold.

Dataset	Method	Acc (%)	Pre (%)	Rec (%)	F1 (%)
DAIC-WOZ	w/o audio_intra	87.98	87.86	88.33	88.10
	w/o audio_inter	81.56	82.79	80.75	81.73
	w/o emotion	91.29	92.16	89.98	91.06
	w/o self-attention	91.62	92.20	90.92	91.52
	ours	**92.21**	**92.36**	**92.18**	**92.23**
MODMA	w/o audio_intra	88.4	86.56	88.52	87.53
	w/o audio_inter	84.75	86.28	78.8	81.81
	w/o emotion	89.6	87.57	89.43	88.35
	w/o self-attention	88.79	88.01	89.04	88.41
	ours	**90.35**	**88.25**	**90.33**	**89.15**
D-Vlog	w/o audio_intra	90.64	89.48	93.81	91.57
	w/o audio_inter	75.64	76.62	82.34	79.16
	w/o emotion	92.00	91.25	96.00	93.45
	w/o self-attention	92.64	91.65	97.20	94.35
	ours	**93.91**	**91.90**	**98.48**	**95.05**

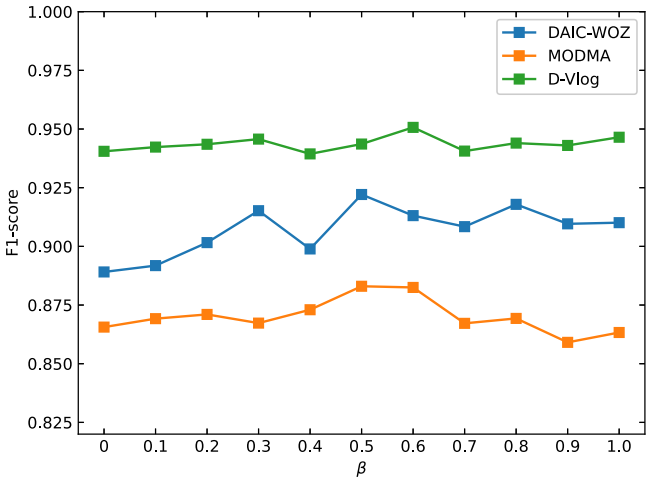


Fig. 4. Sensitivity analysis of hyperparameter β.

We conducted experiments to evaluate the impact of varying values of the hyperparameter β, within the range of [0,1], on the model performance. The performance of the classification exhibits a general trend of increasing and then decreasing with an increase in β. For the DAIC-WOZ dataset, the best classification performance is achieved when $\beta = 0.5$. However, for the MODMA and D-Vlog datasets, the best classification performance is achieved when $\beta = 0.6$. The difference between the best and worst results is about 3%, prompting the conclusion that excessive or insufficient incorporation of emotional features can affect the final classification performance. Large values of β render emotional features are largely ineffective, while small values weaken the strength of similarity connections between the nodes' features.

4.5.2. Size of hidden layers

We also analyzed the impact of the number of convolutional layers of the graph neural network on the model, and compared the results obtained by using different numbers of hidden layers in the two graph neural network modules across the three datasets. The performance impact of using different sizes of hidden layers in the two graph convolution modules is presented in Fig. 5. The performance curve initially increases as the number of convolutional layers increases, but then starts to decrease as the number of layers continues to increase. This is due to the occurrence of the over-smoothing problem, where the hidden

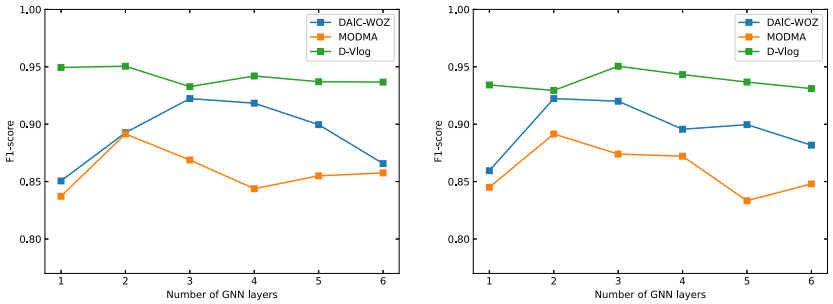


Fig. 5. Results for different numbers of layers of the graph neural network, where (a) and (b) indicate the number of layers of the intra-audio and inter-audio modules, respectively.

layer features of each node converge to the same position in the spatial representation as the number of convolutional layers increases. For the three datasets, the best performance is achieved when the number of graph convolution layers in the two graph neural network modules is set to 2 or 3. The detailed analysis of the hyperparameters above highlights the importance of adjusting hyperparameters to enhance the overall performance of the model.

5. Discussions

This paper proposes a novel approach that utilizes a graph convolutional network to better explore depression-related cues in audio signals, to aid in the early diagnosis and treatment of depression patients. The effectiveness of our proposed model is validated using several datasets. To extract low-level features from the original audio signal, we employ a transformation technique that results in time frames and feature dimensions in the form of a two-dimensional spectrogram. We then extract time series information from these low-level features and learn the embedding representation of each audio sample using the first stage graph convolutional network. By exploiting the properties of information aggregation from the neighborhood with graph neural networks, we can explore the potential relationship between different time-frame level features to learn depression-related cues for subsequent classification tasks. To explore intra- and inter-class similarities and differences between unused audio samples, we construct a second-stage graph convolutional network based on the embedding representation of each audio sample. Additionally, we introduce emotional features through a self-attentive mechanism to further explore depression-related cues in the audio signal. We believe that these emotional features, which are highly correlated with depression, can help the model better identify depression-related cues in audio signals. The effectiveness of our proposed approach is demonstrated in Table 2.

To investigate the interplay between frame-level features in audio, we propose a graph neural network model wherein distinct time frames serve as nodes, and a sliding window of size p + s is utilized. Our investigation focuses on assessing the impact of varying values of p and s on the proposed method's performance. The results of connecting past p time frame nodes and future s time frame nodes to each frame node are depicted in Fig. 6. It is evident from the findings that the optimal performance is achieved when the past time window is set to 10, and the future time window is set to 12, exhibiting a disparity of approximately 3% from the worst results. Despite implementing the graph attention mechanism to amplify the connections between relevant nodes and weaken those between irrelevant nodes, the different time windows still affect the final classification accuracy. This can be attributed to the fact that small time windows may overlook crucial contextual information, while large time windows may introduce extraneous noise information during information aggregation. Thus, constructing graph models to adjust the sliding window size of various time-frame node connections for audio signal-based classification and

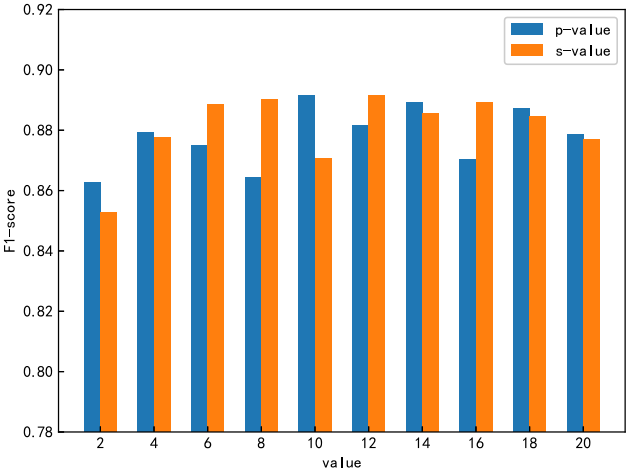


Fig. 6. Prediction results of depression on the dataset MODMA by setting different sizes p and s.

prediction tasks offers a promising means of enhancing the model's performance.

To explore the optimal number of time frames under different datasets, we explored the effect of setting different numbers of frame-level features on the model performance for the datasets DAIC-WOZ, MODMA and D-Vlog, shown in Fig. 7. From the results, it can be seen that for the dataset DAIC-WOZ and the data MODMA, with the increase of the number of nodes of the frame-level features, the accuracy firstly has a slight improvement and gradually tends to stabilize. However, as the number of time frames continues to increase, the performance instead decreases slightly. For the dataset D-Vlog, with the increase in the number of frame-level feature nodes, the accuracy first shows an increasing trend and then tends to level off. The main reason may be that the duration of the audio signal is not consistent across datasets, where most samples in dataset D-Vlog have more time frames than datasets DAIC-WOZ and MODMA. To keep having as many samples as possible to train the model, we fill in the samples with an insufficient number of time frames. Inevitably, noise is introduced, causing the model performance to degrade when the number of feature nodes at the frame level is too high. Considering that the demand of the model for temporal and spatial resources increases rapidly as the number of nodes increases in the graph neural network, the number of frame-level node features was eventually set to 300, 300, and 400 for the datasets DAIC-WOZ, MODMA, and D-Vlog, respectively.

Given that the DAIC-WOZ and MODMA datasets utilized in our study are in different languages and recorded using different methods, we adopt a training and testing strategy where one dataset is used for training and the other for testing. Specifically, we use all the data in the training dataset and randomly select 10% of the original dataset

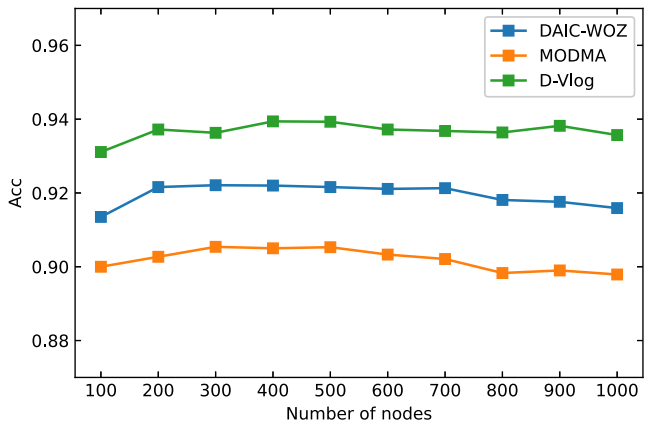


Fig. 7. Depression detection results on three datasets with different numbers of nodes of frame-level features.

Table 3
Cross-dataset test results for the DAIC-WOZ dataset and MODMA dataset.

Train	Test	Acc (%)	F1 (%)
DAIC-WOZ	MODMA	48.72	60.86
MODMA	DAIC-WOZ	61.64	56.25

for testing. The obtained results are presented in Table 3. The test results reveal that regardless of whether the DAIC-WOZ dataset is used to train the MODMA dataset for testing or the MODMA dataset is used to train the DAIC-WOZ dataset for testing, the final classification performance is approximately 30% lower than that of the training and testing data from the same dataset. This may be attributed to the significant differences between the two datasets in terms of language and recording methods, leading to large discrepancies in the learned feature embeddings between the training and testing sets and ultimately impacting the model's classification performance. At present, our audio signal-based depression detection model is trained and tested using data from the same dataset, thus limiting its practical application. Therefore, it is of great research significance to design a cross-domain model for depression detection with cross-database, cross-cultural, and cross-linguistic capabilities to facilitate its practical applications.

Despite achieving promising results on multiple depression datasets, the proposed method in this paper has certain limitations. Firstly, the audio signal is not a natural graph data structure, and consequently, connections between frame-level features within the audio are constructed based on temporal order, similar to most current research. While we employ an attention mechanism to minimize noise interference from irrelevant nodes, the introduction of certain noise information is unavoidable. Secondly, our model is a transduction learning model and hence requires retraining if new data points are added to the training or testing sets. In real-world applications, inductive learning is a more feasible approach, enabling the model to handle new data points without requiring retraining. Finally, while our model performs well on tests within the same database, its performance on cross-database and cross-language tests is comparatively lower. In practice, the data characteristics of the training and testing sets may differ significantly, posing a significant challenge for our prediction task. This is an issue we aim to address in our future work.

6. Conclusion

In this study, we propose a novel approach for recognizing MDD that utilizes a graph neural network. Our approach is designed to consider both intra-audio and inter-audio associations to comprehensively explore potential connections between audio signals and improve the identification of depression-related cues in audio data. To further enhance the performance of our model, we incorporate an emotion feature extractor that is capable of extracting highly depression-related emotion information from patients' audio signals. This feature extractor strengthens the connection between depressed patients and improves the overall performance of the model. To validate the effectiveness of our proposed model, we conducted extensive experiments on multiple datasets. However, depression databases often suffer from low data volume and unbalanced data distribution, which poses a challenge for practical applications. Additionally, patients may come from different regions and speak different languages, which presents additional challenges for cross-database, cross-cultural, or cross-linguistic depression identification. Therefore, our future research will focus on addressing these challenges and developing methods for more robust and reliable MDD recognition across diverse populations and datasets.

CRediT authorship contribution statement

Chenjian Sun: Conception and design of the study, Data collection, Analysis and/or interpretation, Methodology, Experimental design, Writing–original draft. **Min Jiang:** Conceptualization, Analysis and/or interpretation of data, Writing – review & editing. **Linlin Gao:** Experimental supplements, Language polishing. **Yu Xin:** Methodology Experimental analysis, Discussion, Writing – original draft. **Yihong Dong:** Conception and design of the study, Data collection, Analysis and/or interpretation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Zhejiang Province, China (No. LY20F020009) and Natural Science Foundation of Ningbo, China (No. 2023J114).

References

[1] R.C. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K.R. Merikangas, A.J. Rush, E.E. Walters, P.S. Wang, The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-r), JAMA 289 (23) (2003) 3095–3105.
[2] C.D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030, PLoS Med. 3 (11) (2006) e442.
[3] J. Pan, H. Lin, Y. Dong, Y. Wang, Y. Ji, MAMF-GCN: Multi-scale adaptive multi-channel fusion deep graph convolutional network for predicting mental disorder, Comput. Biol. Med. 148 (2022) 105823.
[4] S. Soni, A. Seal, S.K. Mohanty, K. Sakurai, Electroencephalography signals-based sparse networks integration using a fuzzy ensemble technique for depression detection, Biomed. Signal Process. Control 85 (2023) 104873.
[5] A. Seal, R. Bajpai, M. Karnati, J. Agnihotri, A. Yazidi, E. Herrera-Viedma, O. Krejcar, Benchmarks for machine learning in depression discrimination using electroencephalography signals, Appl. Intell. 53 (10) (2023) 12666–12683.
[6] S. Soni, A. Seal, A. Yazidi, O. Krejcar, Graphical representation learning-based approach for automatic classification of electroencephalogram signals in depression, Comput. Biol. Med. 145 (2022) 105420.
[7] M. Hamilton, The hamilton rating scale for depression, in: Assessment of Depression, Springer, 1986, pp. 143–152.
[8] A.T. Beck, R.A. Steer, R. Ball, W.F. Ranieri, Comparison of beck depression inventories-IA and-II in psychiatric outpatients, J. Pers. Assess. 67 (3) (1996) 588–597.
[9] H. Jiang, B. Hu, Z. Liu, L. Yan, T. Wang, F. Liu, H. Kang, X. Li, Investigation of different speech types and emotions for detecting depression using different classifiers, Speech Commun. 90 (2017) 39–46.

[10] H. Chen, D. Jiang, H. Sahli, Transformer encoder with multi-modal multi-head attention for continuous affect recognition, IEEE Trans. Multimed. 23 (2020) 4171–4183.
[11] S.A. Qureshi, S. Saha, M. Hasanuzzaman, G. Dias, Multitask representation learning for multimodal estimation of depression level, IEEE Intell. Syst. 34 (5) (2019) 45–52.
[12] A. Seal, R. Bajpai, J. Agnihotri, A. Yazidi, E. Herrera-Viedma, O. Krejcar, DeprNet: A deep convolution neural network framework for detecting depression using EEG, IEEE Trans. Instrum. Meas. 70 (2021) 1–13.
[13] P. Lopez-Otero, L. Dacia-Fernandez, C. Garcia-Mateo, A study of acoustic features for depression detection, in: 2nd International Workshop on Biometrics and Forensics, IEEE, 2014, pp. 1–6.
[14] N. Cummins, J. Epps, M. Breakspear, R. Goecke, An investigation of depressed speech detection: Features and normalization, in: Twelfth Annual Conference of the International Speech Communication Association, 2011.
[15] J.W. Weeks, A. Srivastav, A.N. Howell, A.R. Menatti, "Speaking more than words": Classifying men with social anxiety disorder via vocal acoustic analyses of diagnostic interviews, J. Psychopathol. Behav. Assess. 38 (2016) 30–41.
[16] P. Singh, G. Saha, M. Sahidullah, Non-linear frequency warping using constant-Q transformation for speech emotion recognition, in: 2021 International Conference on Computer Communication and Informatics, ICCCI, IEEE, 2021, pp. 1–6.
[17] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 3–10.
[18] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, 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, 2016, pp. 27–34.
[19] Y. Dong, X. Yang, A hierarchical depression detection model based on vocal and emotional cues, Neurocomputing 441 (2021) 279–290.
[20] M. Niu, B. Liu, J. Tao, Q. Li, A time-frequency channel attention and vectorization network for automatic depression level prediction, Neurocomputing 450 (2021) 208–218.
[21] N. Seneviratne, C. Espy-Wilson, Speech based depression severity level classification using a multi-stage dilated CNN-LSTM model, 2021, arXiv preprint arXiv:2104.04195.
[22] A. Shirian, T. Guha, Compact graph architecture for speech emotion recognition, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6284–6288.
[23] M. Niu, K. Chen, Q. Chen, L. Yang, Hcag: A hierarchical context-aware graph attention model for depression detection, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 4235–4239.
[24] T. Chen, R. Hong, Y. Guo, S. Hao, B. Hu, MS ^{2}-GNN: Exploring GNN-based multimodal fusion network for depression detection, IEEE Trans. Cybern. (2022).
[25] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[26] Y. Gong, C. Poellabauer, Topic modeling based multi-modal depression detection, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017, pp. 69–76.
[27] J. Chunyan, M. Chen, L. Bin, Y. Pan, Infant cry classification with graph convolutional networks, in: 2021 IEEE 6th International Conference on Computer and Communication Systems, ICCCS, IEEE, 2021, pp. 322–327.
[28] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J.S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 6367–6371.
[29] W. Nie, M. Ren, J. Nie, S. Zhao, C-GCN: Correlation based graph convolutional network for audio-video emotion recognition, IEEE Trans. Multimed. 23 (2020) 3793–3804.
[30] J. Liu, H. Wang, Graph isomorphism network for speech emotion recognition, in: Interspeech, 2021, pp. 3405–3409.
[31] K. Rezaee, M.R. Khosravi, M. Jabari, S. Hesari, M.S. Anari, F. Aghaei, Graph convolutional network-based deep feature learning for cardiovascular disease recognition from heart sound signals, Int. J. Intell. Syst. (2022).
[32] N. Ghadiri, R. Samani, F. Shahrokh, Integration of text and graph-based features for detecting mental health disorders from voice, 2022, arXiv preprint arXiv:2205.07006.
[33] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.
[34] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The Distress Analysis Interview Corpus of Human and Computer Interviews, Tech. rep., University of Southern California Los Angeles, 2014.
[35] H. Cai, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, et al., Modma dataset: A multi-modal open dataset for mental-disorder analysis, 2020, arXiv preprint arXiv:2002.09283.
[36] J. Yoon, C. Kang, S. Kim, J. Han, D-vlog: Multimodal vlog dataset for depression detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, no. 11, 2022, pp. 12226–12234.
[37] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.
[38] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: An efficient deep model for audio based depression classification, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 35–42.
[39] M. Du, S. Liu, T. Wang, W. Zhang, Y. Ke, L. Chen, D. Ming, Depression recognition using a proposed speech chain model fusing speech production and perception features, J. Affect. Disord. 323 (2023) 299–308.
[40] P. Zhang, M. Wu, H. Dinkel, K. Yu, Depa: Self-supervised audio embedding for depression detection, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 135–143.
[41] X. Chen, Z. Pan, A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health, Int. J. Environ. Res. Public Health 18 (12) (2021) 6441.
[42] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, B. Hu, TAMFN: Time-aware attention multimodal fusion network for depression detection, IEEE Trans. Neural Syst. Rehabil. Eng. (2022).
[43] L. Zhou, Z. Liu, X. Yuan, Z. Shangguan, Y. Li, B. Hu, CAIINET: Neural network based on contextual attention and information interaction mechanism for depression detection, Digit. Signal Process. 137 (2023) 103986.