DOI: 10.1111/nyas.15366

ORIGINAL ARTICLE

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Audio multi-feature fusion detection for depression based on graph convolutional networks

Guangsheng Luo[1,2] | Xianda Ma[1,3,4] | Jun Yea[3,4] | Yang Liu[1] | Yiwei Xia[1] | Chengrun Li[1] | Yifang Kuang[1,3] | Ruyuan Zhang[4] | Siyu Lou[5] | Kai Yu[6] | Mengyue Wu[7] | Weidong Li[1,4,8]

[1]College of Electrical and Electronic Engineering, Shanghai University Of Engineering Science, Shanghai, China

[2]College of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai, China

[3]Bio-X Institutes, Key Laboratory for the Genetics of Development and Neuropsychiatric Disorders (Ministry of Education), Shanghai Key Laboratory of Psychotic Disorders, and Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai, China

[4]Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China

[5]Eastern Institute for Advanced Study, Shanghai, China

[6]X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[7]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China

[8]WLA Laboratories, World Laureates Association, Shanghai, China

Correspondence
Yang Liu and Weidong Li, Center for Brain Health and Brain Technology, Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, China. Email: yang.liu1@sjtu.edu.cn and liwd@sjtu.edu.cn

Mengyue Wu, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China. Email: mengyuewu@sjtu.edu.cn

Guangsheng Luo, Xianda Ma, and Jun Ye share first authorship.

Funding information
STI2030-Major Projects, Grant/Award Number: 2022ZD0204900

Abstract

Depression is a prevalent mental disorder, and early detection and diagnosis are crucial for its prevention and treatment. Speech-based depression detection represents an efficient and convenient approach within the current landscape of computer-aided detection methods. However, challenges remain in effectively and reliably extracting features and classifying speech patterns to distinguish individuals with depression from those without. This paper introduces an audio feature set for depression analysis, referred to as SJTU-LWDLab DACD. Based on this feature set, we propose a novel method for identifying patients with depression using summed graph convolutional networks to mitigate inaccuracies that arise from the loss of spatial features, such as height and depth, during the structured fusion of multiple depression audio features. Experimental results demonstrate that the accuracy of depression recognition in speech can reach 92.4%. The method proposed in this paper provides objective indicators and a foundation for the auxiliary identification of depression.

KEYWORDS
audio feature, DACD data set, depression detection, SGCNs, structured fusion

INTRODUCTION

According to data from the World Health Organization in 2024, approximately 3.8% of the global population suffers from depression, including 5% of all adults (4% of men and 6% of women) and 5.7% of adults over 60 years old, with this number increasing annually. Concurrently, the 2024 China Mental Health Survey reports that the lifetime prevalence of depressive disorders among Chinese adults is 6.8%, with depression accounting for 3.4%. In China, approximately 95 million people are affected by depression, and around 280,000 individuals commit suicide each year, of which 40% suffer from depression. Depression not only inflicts significant trauma on individuals and families but also results in substantial societal loss. Despite its severe impact, depression is a condition that can be effectively treated and improved. In clinical practice, pharmacotherapy can facilitate the recovery of patients with depression. Additionally, various forms of psychotherapy (such as cognitive behavioral therapy, interpersonal psychotherapy, behavioral therapy, and psychodynamic therapy) and physical therapies (such as repetitive transcranial magnetic stimulation) are available.[1] Less than half of people affected by depression in China and less than 10% in many other countries receive treatment. The main reasons for these low rates are as follows:

1. For an extended period, depression has been misunderstood by the public, often labeled as "small-minded," "unable to think clearly," and "pretentious," resulting in insufficient attention to the condition. In reality, depression is a serious mood disorder. Mental illness, like other psychological disorders, is frequently not well understood by society. While it is considered normal for individuals with physical illnesses to seek medical treatment, psychological disorders, particularly mental illnesses, are often stigmatized as "mentally abnormal." The rapid development of society, high work-related stress, changes in social identity and interpersonal relationships, and even the impact of physical illnesses can increase the likelihood of depression. The peak incidence of depression occurs between the ages of 20 and 60, with adolescents, professionals, menopausal individuals, and the elderly being particularly vulnerable. Given that depression affects a broad spectrum of individuals and can lead to severe consequences, it warrants significant attention.
2. Individuals with depression often hesitate to seek help. They experience impaired functioning, negative emotional states, decreased interest in activities, and a lack of motivation to improve their current situation. Furthermore, many individuals with depression lack sufficient knowledge about the condition and are more likely to conceal their illness and avoid consulting a doctor.
3. There is a lack of relevant resources and limited medical facilities for mental health care. Due to insufficient attention to mental illness, there is a relative scarcity of financial investment, as well as professional medical personnel and psychological service providers. Additionally, effective screening for depression poses challenges. Currently, the diagnosis of depression primarily relies on self-report measures and various clinical assessment tools, which can be subjective and time-consuming, resulting in inconsistent diagnoses and delayed treatment.

As a medium for individuals to express their thoughts and emotions, speech has been shown to be associated with neurocognitive dysfunction in patients with depression.[2] Increasingly, global researchers have begun to use machine learning and deep learning methods to study the detection of depression through speech, achieving excellent performance in depression detection. Several groups have shown that speech features are significantly related to the severity of depression and can be used to distinguish between depressed and nondepressed individuals.[3–5] Vázquez-Romero and Gallardo-Antolín[6] discussed the performance of an automatic depression detection system based on the ensemble learning of convolutional neural networks (CNNs). Scherer et al.[7] employed a combination of the normalized amplitude quotient, quasi-open quotient, peak slope, open quotient neural network, and support vector machine (SVM) with a radial basis kernel, achieving a classification accuracy of 75%. Kwon and Kim[8] proposed a divide-and-conquer approach to detect 76 clinically depressed patients (38 with severe depression and 38 with moderate depression), as assessed by the Montgomery–Asberg Depression Rating Scale using speech, achieving an accuracy of 78%. Existing studies indicate that the speech of depressed patients exhibits the following characteristics: a slower speech rate, more frequent pauses, and longer pauses.[5] Compared with normal individuals, the changes in vocal features are diminished, the voice lacks intonation, and the voice sounds dull, with breath sounds being more prominent in the depressed group. From the perspective of prosodic features, the fundamental frequency (F0) of patients with depression exhibits minimal changes, including in bandwidth, amplitude, and energy, indicating that the frequency variations in the voices of these patients are relatively limited. Spectral features are also associated with the severity of depression. Studies have demonstrated that the degree of change in the energy of the sound spectrum below 500 Hz and in the range of 500–1000 Hz correlates with the increasing severity of depression. Therefore, the feature extraction of speech, along with the capture of relatively objective acoustic characteristics that individuals are less likely to conceal deliberately, will contribute to a deeper understanding of depression.

Despite these groundbreaking research results, the application of speech-based depression diagnosis in clinical settings remains challenging due to insufficient accuracy and robustness, as well as a lack of large clinical data sets for training. In recent years, graph convolutional networks (GCNs) have garnered significant attention for their ability to model graph-structured data by leveraging the underlying graph topology.[9,10] GCNs have been successfully applied across various fields, including computer vision, natural language processing, social network analysis, and bioinformatics.[11–14] They provide a powerful framework for modeling graph-structured data and have demonstrated promising results in a wide range of applications. As research in this field continues to advance, it is anticipated that GCNs will play an increasingly important role in diverse areas. Recent studies have shown the effectiveness of GCNs in depression diagnosis[15,16] by analyzing brain imaging data and identifying biomarkers associated with depression. These biomarkers can be utilized to develop more accurate and objective diagnostic tools, thereby facilitating earlier intervention and improving patient outcomes. However, there is currently no research on the application of GCNs in speech-based depression detection; thus, we aim to employ GCNs to achieve high-precision and reliable discrimination of depression-related speech data. The innovations of this paper are as follows:

1. We collected a substantial amount of real speech data related to depression and created the SJTU-LWDLab DACD data set. From this data set, we extracted numerous valuable features of

depression and labeled the features corresponding to patients with depression and those without depression separately.

2. To address the issue where judgments become inaccurate due to the loss of spatial features, such as height and depth, following the structured fusion of multiple depression audio features, we propose a depression classification method based on the summation graph convolutional networks (SGCNs) for this data set. This method resolves inaccuracies related to the classification of depression speech caused by pronunciation features,[17] vocabulary features, acoustic features, and background noise spatial features.[18]

METHODOLOGY

Audio feature extraction

We used the speech toolkit openSMILE to extract frame-level features from the original waveform with a frame window size of 25 ms and a frame interval of 10 ms. We used the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) introduced by Eyben et al.[19] to extract frame-level features with 88 dimensions. Since depressed audio data do not have a satisfactory diversity of linguistic information, only nonlinguistic acoustic features were extracted. Therefore, the focus here was on achieving accurate classification by using only audio sentiment attributes. Several sound components can be employed by utilizing audio–emotional characteristics. It is crucial to test diverse feature combinations to identify which ones have a greater relation with depression, indicated by improved precision and F1 scores.[20,21] The selection of speech signals is the basis for the evaluation of reliable automatic systems for detecting depressive states. In this study, in the absence of medical and scientific protocols, the most relevant acoustic features were considered to determine specific changes in speech quality and to assess the presence of depression. The estimated acoustic features include:

1. Jitter: We used it to evaluate basal-periodic changes in psychological frequency and to indicate the instability of vocal fold oscillatory patterns. Jitter was estimated by using Praat (whose name derives from the command form of the Dutch word “praaten,” i.e., “to speak”).[22,23]
2. MFCC: Mel-frequency cepstral coefficients (MFCC) is a widely utilized technique in speech signal processing and audio feature extraction. In this paper, we conducted experiments employing MFCC, zero-crossing rate, speech probability, fundamental frequency (F0), and frame energy,[24] analyzing the differences in the vocal characteristics of patients with depression.
3. Derivatives of cepstral coefficients, which are useful to investigate the properties of the dynamic behavior of the speech signal; and spectral centroid, which evaluates the modifications of the signal frequency over time.[25]

The features we extracted from the real audio files are shown in Table 1. The collected audio is a doctor’s interview with a patient. In this paper, by using the original audio and different concepts in machine learning, such as data reassignment, we mitigated the harmful effects of bias.

Audio feature selection and processing

Our data processing was divided into two steps: the first step was speaker separation, to extract the audio clips and mark the speakers (doctors or respondents); the second step was audio separation, to extract the audio clips of the respondents. If the respondents had multiple audio clips, they were merged into one clip. We used a total of 1135 actual patient samples, including 588 depressed and 547 nondepressed patients, with a total duration of 106 h. All audio data come from the relevant affiliated hospitals of Shanghai Jiao Tong University. They were completed in stages by doctors during working hours in Chinese hospitals. All patients provided written informed consent and were compensated by a certain amount.

We extracted features in two stages: frequency domain feature extraction and time domain feature extraction. The frequency domain feature extraction consisted of a peak frequency, which identifies the most obvious frequency peaks in the speech spectrogram, corresponding to the main frequencies in the audio signal; an energy spectral density, which calculates the power spectral density in each frequency interval and extracts the energy distribution of the frequency bands; a spectral average, which calculates the average of the whole spectrogram that represents the overall spectral characteristics; and a spectral equalization, which divides the spectrogram into multiple frequency bands and calculates the energy of each frequency band. Spectral equalization was used for sound equalization processing. The time domain feature extraction contained: a short-time over-zero rate that calculates the number of over-zero points between neighboring audio samples, which is used to represent the frequency characteristics of the signal; a short-time energy that calculates the energy of the audio signal within each time window, which is used to represent the amplitude characteristics of the signal; and a short-time autocorrelation that calculates the autocorrelation coefficient of the audio signal under different time delays, which is used to represent the periodicity characteristics of the signal. Finally, the above frequency domain and time domain features were combined into features and represented by an unstructured graph network to further learn the characteristics of depression using the GCNs model.

We use both labeled and unlabeled training data when constructing the graph. Since audio data have different sound and speech features, speech data using low-level descriptors (LLDs), spectrograms, and so on were extracted by spectrogram convolution, and then node embeddings of the graph were constructed to demonstrate the effectiveness of our graph approach and that our model is not bound to any particular embedding. During the audio-to-feature embeddings, each audio segment was divided into nonoverlapping 960 ms segments. For each segment, a log-Meier spectrogram was computed using frames with a length of 25 ms and an overlap of 10 ms, 64 spaced frequency bins, and log-transforming the magnitude of each bin. This yields a log-Meier

TABLE 1 Features for audio feature extraction.

Type	Name
Energy features	Energy contour Band-energy-ratio RMS energy
Time domain features Frequency domain features	Equalization of root energy onset time Autocorrelation Zero crossing rate Spectral center of mass Mel-frequency cepstral coefficients (MFCCs) linear prediction coefficients (LPCs) CEPSTRAL coefficients Spectral flatness Spectral flux Spectral centroid Peak-to-sidelobe ratio (PSR) of Hilbert envelope of LP residual Spectral roll-off Log mel spectrum energy pitch Harmonic ratio bandwidth Band-energy-ratio Delta-spectrum magnitude Normalized autocorrelation peak strength (NAPS) of zero-frequency filtered signal
Music theory features	Detuning Formants
Sensory features	Loudness Sharpness Chroma

spectrogram of size 96 × 64 as input to the pretrained GCN model. We used the 128-dimensional features extracted from the audio to process each log-Meier spectrogram and averaged over all segments to form a final vector representation of each audio segment.

Construction of the initial graph

For the graph neural network (GNN) construction of depressed speech audio files, we generally needed to convert the audio files into graph structures first, and then use the GCNs to extract features and train models.[25] We extracted valuable features from audio signals, including MFCCs, spectrograms, short-time Fourier transforms, and pitch features. The audio signal is segmented into multiple time windows, each with a duration of 25 ms, while the frame shift is typically set between 10 and 15 ms. Each time window represents a frame, and each frame corresponds to a node. The features of each node consist of the extracted characteristics from the corresponding frame. Adjacent frames are connected to capture the temporal relationships between nodes, and sliding windows are employed to link additional nodes to the current node. The MFCC features, along with other long-term features (e.g., segment duration and voice activity detection results), are concatenated into a comprehensive feature vector, thereby integrating information at different levels before inputting it into the GCN model. Assuming that the audio is divided into N frames, there will be N nodes in the graph, and each node $\boldsymbol{V}$ contains the audio feature vector $\boldsymbol{X_i}$ of the frame. Given the feature vectors $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ of two nodes $\boldsymbol{V_i}$ and $\boldsymbol{V_j}$, the similarity calculation is as $S\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|}$. The element $A(i, j)$ of the adjacent matrix $\boldsymbol{A}$ of the graph can be expressed as:

$$A\left(i, j\right) = \begin{cases} 1, & \text{if } S\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right) \geq \theta \\ 0, & \text{if } S\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right) < \theta \end{cases}, \tag{1}$$

where θ is the hyperparameter with a value of 0.5 threshold. By calculating the feature similarity of depressed speech nodes, a cosine similarity graph structure is constructed to represent depressed and nondepressed patients. Considering the real data set scenario, the speech in this article is divided into only two categories: depressed and nondepressed, and only for female patients, because we found that more women suffer from depression than men.

SGCNs model framework

GCN message-passing method can learn expressive representations in many tasks,[26] but it does not work well in depressive speech message-passing because the sum of multiple data sets includes the aggregation of noise, which interferes with the learning of representations. In this paper, we chose the most suitable message-passing method for our task with similar GCNs (SGCNs) and aggregated rich distributional information in depressive speech classification, where statistical information is valued.[27] GCN with good performance was designed, and comprehensive guidance was provided for the design space batch normalization, culling, activation, aggregation, layer connectivity, preprocessing layer message passing layer, and postprocessing layer.[28,29] The first layer of the SGCNs model is $\boldsymbol{H}^{(1)} = \boldsymbol{\sigma}(\boldsymbol{\hat{A}XW}^{(0)})$. Among them,

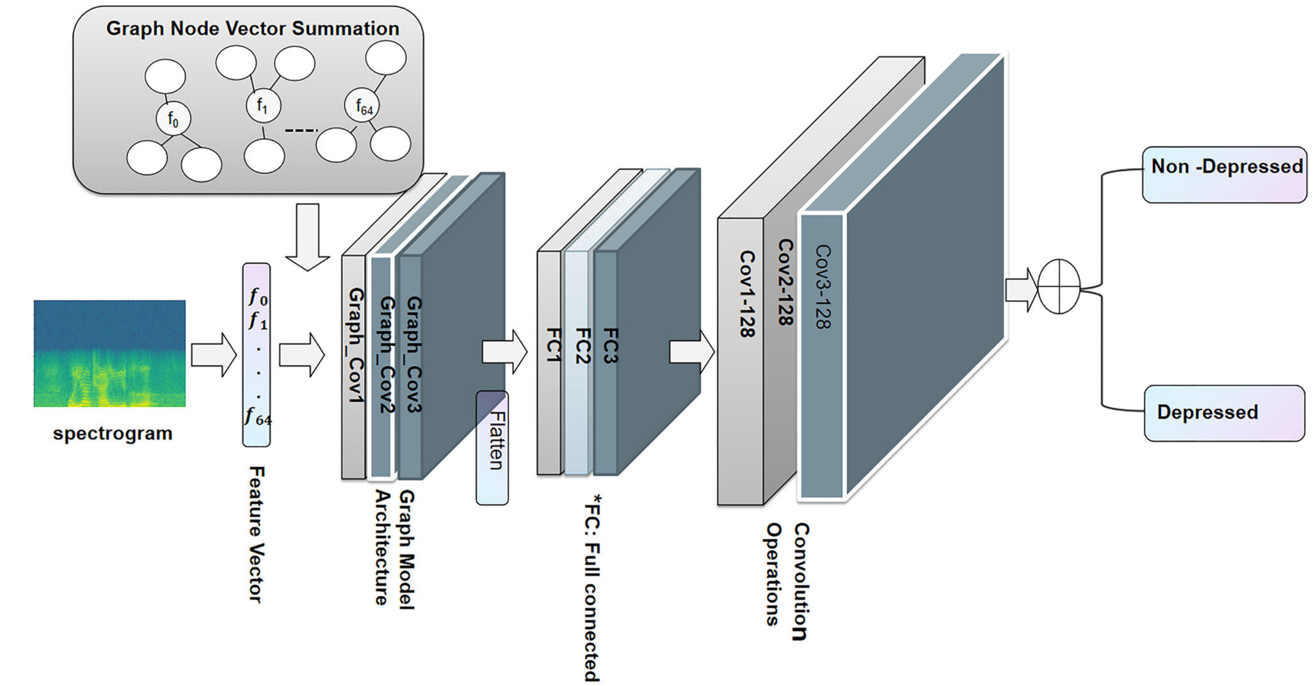

FIGURE 1 Audio classification framework using summed graph convolutional networks (SGCNs).

X is the node feature matrix of the 0th layer, that is, $H^{(0)} = X$. $\widehat{A}$ is the normalized adjacency matrix, which acts on the node feature matrix and aggregates the features of the node and its neighbors. $\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \in \mathbb{R}^{N\times N}$ is the self-loop normalized adjacency matrix, and $\tilde{A} = A + I$ is the original adjacency matrix with self-loop connections added. $\tilde{D} = D + I, X^{(k-1)} \in \mathbb{R}^{N\times D}, W^{(k)} \in \mathbb{R}^{D\times D}$, where D and I are the degree matrix and the identity matrix, respectively. $W^{(0)}$ is a learnable weight matrix, and σ is a nonlinear activation function.

In addition, $L2$ regularization technology was introduced to control the similarity between node features, prevent features from gathering toward the center of the space, and alleviate the smoothing problem. The final SGCN model is shown in Equation (2):

$$H^{(k+1)} = \sigma\left(\hat{A}H^{(k)}W^{(k)}\right) + b, \tag{2}$$

where b is a hyperparameter to prevent overfitting. In this paper, we only did two classification tasks, extracting a complete representation from multiple vectors or multiple features to complete the task, so we used the summation as the OUTPUT function represented as $y_G = OUTPUT(H_v^{(K+1)} \mid v \in G)$. The loss function is shown in Equation (3):

$$\mathcal{L}_{OSS} = -\sum_{i\in L}\sum_{j=1}^{c} Y_{ij}\ln y_G. \tag{3}$$

Here, Y_{ij} is the true labeled value and y_G is the model-predicted value. The SGCN-based depression audio classification framework is shown in Figure 1. The steps of spectrogram generation were divided into preprocessing the audio signals, splitting the frames, applying the window function, performing the Fourier transform, calculating the power spectral density, and combining the spectrograms. First, preprocessing of depression and nondepression raw audio signals was conducted, including loading audio files, sample rate conversion, noise removal, and audio normalization. The audio signals were then divided into short frames containing tens to hundreds of milliseconds each, using a sliding window technique, and a window function was applied to each frame using a Hamming window to reduce spectral leakage. Next, a fast Fourier transform (FFT) is applied to the audio signal in each window to convert the signal from the time domain to the frequency domain, and the power spectral density of each window is obtained by taking the square of the mode of the complex result obtained from the FFT. Finally, the power spectral densities of all windows are arranged in chronological order to form a spectrogram.

Fourier transform of graph structure audio features

We utilized graph frequencies to represent node-related data, employing the eigenvectors and eigenvalues of the graph's Laplacian matrix as basis functions and frequencies. Spectral graph convolution leverages the spectral information of the graph to capture the relationships among nodes, thereby enhancing the model's capacity to learn graph structures. By applying the Fourier transform to the time series signal, we extracted time-domain features, which were subsequently integrated with the graph structure. The Fourier transform in the continuous domain is defined as:

$$F(\omega) = \mathcal{F}[f(t)] = \int f(t)\, e^{-i\omega t}dt. \tag{4}$$

In the equation above, $\boldsymbol{f}(\boldsymbol{t})$ is the integral of the signal and $e^{-i\omega t}$ is the basis function, which satisfies:

$$\boldsymbol{\Delta}\, e^{-i\omega t} = \frac{\partial^2}{\partial t^2}\, e^{-i\omega t} = -\omega^2 e^{-i\omega t}. \tag{5}$$

The $\boldsymbol{\Delta}$ is the Laplacian operator, that is, a second-order differential operator in n-dimensional Euclidean space $\boldsymbol{\Delta f} = \boldsymbol{\nabla}^2\, \boldsymbol{f}$, which is a transformation of the function $\boldsymbol{f}$. We transfer the Fourier transform to the graph structure. With the Laplacian matrix, we only need to calculate its eigenvector. Therefore, as long as the eigenfunction $e^{-i\omega t}$ of the Laplacian operator is mapped to the eigenvector of the graph Laplacian matrix to solve the eigenvector, we can get the Fourier transform on the graph as follows:

$$\boldsymbol{F}\,(\lambda_l) = \hat{\boldsymbol{f}}\ \,(\lambda_l) = \sum_{i=1}^{N} \boldsymbol{f}\ \,(\boldsymbol{i})\,\boldsymbol{u}_l\,(\boldsymbol{i})\,, \tag{6}$$

where $\boldsymbol{f}$ is the transformation of the nodes on the graph, such as returning the node embedding, $\boldsymbol{f}(\boldsymbol{i})$ corresponds one-to-one to the nodes on the graph, and $\boldsymbol{u}_l(\boldsymbol{i})$ represents the $\boldsymbol{i}$th component of the $\boldsymbol{l}$th eigenvector. In other words, the graph Fourier transform of node embedding $\boldsymbol{f}(\boldsymbol{i})$ under eigenvalue $\boldsymbol{\lambda}_l$ is the inner product operation with the eigenvector $\boldsymbol{u}_l$ corresponding to $\boldsymbol{\lambda}_l$. Using matrix multiplication, the Fourier transform on the graph is generalized to a matrix form:

$$\begin{pmatrix} \hat{\boldsymbol{f}}\,(\lambda_1) \\ \hat{\boldsymbol{f}}\,(\lambda_2) \\ \vdots \\ \hat{\boldsymbol{f}}(\lambda_N) \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1\,(\boldsymbol{1}) & \boldsymbol{u}_1\,(\boldsymbol{2}) & \dots & \boldsymbol{u}_1\,(\boldsymbol{N}) \\ \boldsymbol{u}_2\,(\boldsymbol{1}) & \boldsymbol{u}_2\,(\boldsymbol{2}) & \dots & \boldsymbol{u}_2\,(\boldsymbol{N}) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{u}_N\,(\boldsymbol{1}) & \boldsymbol{u}_N\,(\boldsymbol{2}) & \dots & \boldsymbol{u}_N\,(\boldsymbol{N}) \end{pmatrix} \begin{pmatrix} \boldsymbol{f}\,(\boldsymbol{1}) \\ \boldsymbol{f}\,(\boldsymbol{2}) \\ \vdots \\ \boldsymbol{f}\,(\boldsymbol{N}) \end{pmatrix}. \tag{7}$$

That is, the matrix form of the Fourier transform of $\boldsymbol{f}$ on the graph is: $\hat{\boldsymbol{f}} = \boldsymbol{U}^{\top}\ \boldsymbol{f}$. In short, given the input audio feature node embedding $\boldsymbol{f}$, multiplying it with $\boldsymbol{U}^{\top}$ on the left can get the output embedding $\hat{\boldsymbol{f}}$ after the Fourier transform on the graph.

SGCNs classification

Sun and other related researchers used GNNs to capture the relationship between the various segments of the speaker's speech in the conversation for speech emotion classification.[30,31] Figueredo and Wolf regarded each audio channel as a node and built a speech graph for speech enhancement tasks.[32] GNN is also used to fuse information from multiple heterogeneous modalities.[33] From these studies, we can see that graph structure learning using speech signal representation has matured. Unlike previous studies, our implementation regards each audio sample as a graph node and directly models the time step as a node or edge in the graph structure to capture the timing characteristics.

Our goal is to use the SGCNs method to classify patients into depression patients and healthy people. In this paper, we use random forest, SVM, Gaussian mixture model, and naive Bayes to conduct comparative studies.[34–37] The comparison verifies the effectiveness and reliability of our method. Our data for each patient were collected in interviews conducted by trained doctors and manually labeled into three categories: the first category of responses to all questions, with long question-and-answer interview durations; the second category of the initial part of each question, which was open-ended and asked each volunteer to respond subjectively according to his experience, feelings, and methods; and the third category of the end of each question, by extracting the last 5 s of the question and then cascading them to prepare a sample file. We extracted multiple features from the obtained speech samples to learn classification using GNNs. For each node in the graph, first, we obtained some data: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_n \in \boldsymbol{R}^{\boldsymbol{p}}$, and the labels of the first $\boldsymbol{l}$ points: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l \in \boldsymbol{1}..\boldsymbol{C}$. Then, we defined the initial label matrix $\boldsymbol{Y} \in \boldsymbol{R}^{\boldsymbol{n}\times\boldsymbol{C}}$, with $\boldsymbol{Y}_{\boldsymbol{lm}} = 1$ if $\boldsymbol{x}_l$ has label $\boldsymbol{l} = \boldsymbol{m}$ and 0 otherwise. This process can be iterated several times, with each iteration updating the features of that node by aggregating the features of its neighbors.

DATA AND RESULTS

Data sets

All of our experiments use actual data from patients and healthy individuals with a history of depression diagnosed by a medical professional in a hospital. The interview time for each individual varies from 10 min to several hours, and each patient is manually labeled. According to the survey, our data set is currently the only and largest voice data set on depression in women in the world. We named it the Depression Audio Consultation Dataset from Li Weidong Laboratory, Shanghai Jiao Tong University (SJTU-LWDLab DACD).

Depression, Anxiety, and Stress in the Context of Conversational Agents-Wizard of Oz (DAIC-WoZ) is a data set for the study of emotion and mental health. It was developed in collaboration with the National Institute of Mental Health (NIMH) Data Archive and the Technical University of Denmark. The DAIC-WoZ data set is a publicly available English-language depression data set containing recordings and transcripts of 142 participants, such as the patient's age and gender, each of whom was labeled with a Patient Health Questionnaire (PHQ)-8 score. The PHQ-8 questionnaire is a popular depression screening questionnaire, and participants with a PHQ-8 score greater than or equal to 10 are considered depressed. The DAIC-WoZ data set consists of a training set (30 depressed and 77 nondepressed patients), a developmental set (12 depressed and 23 nondepressed patients), and an unpublished test set. Before the interview, each participant completed a psychosis questionnaire (PHQ-8) from which a binary truth classification (depression, nondepression) was derived. Each patient was asked to answer a series of questions that addressed their emotional state, depression, anxiety, and other mental health issues. Related studies on using the PHQ questionnaire to screen depression can be found in Refs. 38 and 39.

Experimental settings

Our experiment was divided into three parts. The first part was to separate the conversation and then convert the segments into text data using a third-party automatic speech recognition tool. The text is analyzed for the respondent's sociological attributes and questionnaire results, which include speaker number, start time stamp, text, and end time stamp. The sampling rate of the audio was 16,000 Hz for residual blocks for frame-level feature extraction and 5 ms per frame for long short-term memory (LSTM) for contextual information extraction, and all segments were concatenated to the complete audio. The second part extracts useful feature information from the audio signal to achieve the classification task, and these features contain MFCC, delta features, pitch, and so on. In the field of audio processing, multiple features are fused and compared separately using multiple classifiers and our proposed SGCNs classifier method. In the third part, we compared the audio features of depressed patients from the DAIC-WoZ data set and our own data set, and illustrated the differences of patient features between the two data sets.

Evaluation metrics

This study employed precision, recall, F1-score, and the area under the ROC curve (AUC) as metrics to evaluate model performance. Precision denotes the proportion of accurately identified depressed individuals among all samples, thus reflecting the model's accuracy. Recall refers to the ratio of true depression samples detected by the model compared to the actual depression samples, reflecting the model's sensitivity. The F1-score represents the harmonic mean of precision and recall, demonstrating the balance between these two metrics and reflecting the robustness of the model. AUC serves as an additional indicator of model performance and robustness, with larger AUC values indicating superior model performance and robustness. Additionally, we performed a statistical analysis of data samples from patients with depression and nondepressed individuals, calculating metrics such as mean, median, standard deviation, and *p*-value. To assess whether significant differences exist between the samples, we conducted a *t*-test comparing the data from patients with depression and nondepressed individuals.

Statistics of sociological attribute features

Because of the lack of expressive power of LLD features or sociological attribute features alone, this paper adopted a combination of two approaches for depression and nondepression classification. In the experiment, we analyzed the data of sociological attributes of the interview of depression, including age, educational background, job, marital status, place of birth, alcohol and drug use, kinship relationship, life stress, and effect size of sexual abuse. Our research has found that four social attributes, drug, alcohol, mania, and social life, have a great impact on depression. The statistical results of the distribution of these attributes are shown in Figure 2 (the data can be found in Supplementary Material).

TABLE 2 Statistics of social attributes of depressed (D) patients and nondepressed (ND) patients.

Social attributes	Label	Average value	Median values	Standard deviation	*p*-value
Age	ND	50.3	51	5.7	0.018
	D[a]	51.1	52	5.7	
Education	ND	5.4	5	2	0.065
	D	4.9	4	2.3	
Marriage status	ND	1.41	1	0.95	0.75
	D	1.37	1	0.91	
Work status	ND	2.6	1	2	0
	D	3.6	4	2.2	

[a] Only including 40–60 years old.

Our experimental subjects were females 30–60 years old. There are 11 categories of educational background: uneducated, education kindergarten, primary school, junior high school, high school, specialist/technical school, adult/radio and television education/technical school and colleges, undergraduate, master's degree, and doctoral degree. There are five categories of marital status, namely, married, separated, divorced, widowed, and unmarried. There are eight categories of work status, which are gainfully employed, temporarily laid off or on sick leave, unemployed, looking for work, retired from gainful employment, permanently disabled, homemaker/home-based, in school, and others. To determine whether there is a significant difference between the depressed and nondepressed subjects, we tested the statistical significance of the difference between the two groups separately through statistical analysis. The statistical distribution is listed in Table 2, and the specific statistical histograms are shown in Figure 3.

The total number of depressed and nondepressed subjects was 1135. To compare whether there is a significant difference between the means of the two independent samples of depressed and nondepressed, we did the unpaired *t*-test and calculated the *p*-value. From Table 2, the average age of individuals experiencing depression within the 30–60 years age range is 51.1 years. There are 11 categories of educational attainment, with an average depression rate of 4.9%. Additionally, there are five categories of marital status, which show an average depression rate of 1.37%. Furthermore, the analysis includes eight categories of employment, with an average depression rate of 3.6%. Our further statistical analysis indicates that divorced individuals and those with lower levels of education are at a higher risk of experiencing depression.

Comparison of experimental results for classifier methods

In this experiment, we used different methods to perform classification verification on multiple feature sets, using accuracy, sensitivity, and specificity as evaluation indicators.

First, we screened the data for individuals by selecting those individuals with all the labeling, physiological, and behavioral information

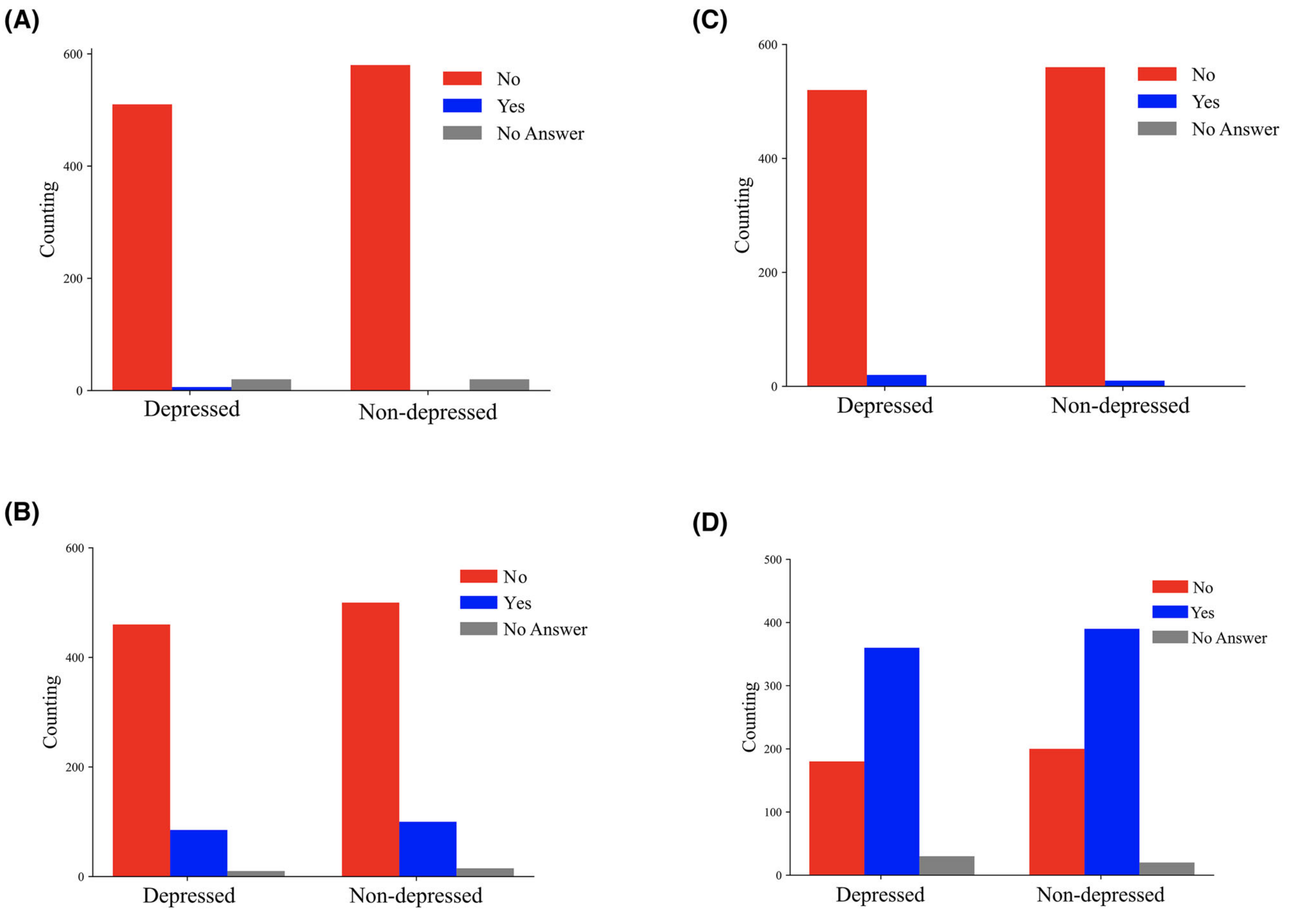

FIGURE 2 Results of statistical analysis on the answers to the PHQ questionnaire for patients with depression and those without depression. Panels (A)–(D) show the distribution of depression and those without depression in terms of drug, alcohol, mania, and social life attributes, respectively.

and features for further processing. Then, in the feature set processing phase, we took three steps:

1. Processing vacant data: If more than half of the data in a channel within the feature set was missing, the channel was deleted. If less than half of the data in a row was missing, the vacancies were filled with the average of the other channels in that row.
2. Feature channel pooling: The weights of different channels accumulated from high to low until the sum of weights reached a threshold of 0.5, and then took the average of the accumulated channel values.
3. Filtering outliers: Data with absolute values greater than 100 were filtered out and replaced with the median of the feature.

When constructing the adjacency matrix, we first screened the feature information and removed information irrelevant to the task, such as the number of patients and speech duration, to reduce interference. The SGCNs model parameters were as follows: GraphConvolution1—input feature length; GraphConvolution2—32 hidden nodes; GraphConvolution3—16 hidden nodes; GraphConvolution4—4 hidden nodes; Dropout—0.5; Output layer—softmax. The learning rate was set to 0.001, and the Adam optimizer was used. We used random forest (RF), SVM, CNN, BiLSTM, and our method to conduct comparative experiments on different feature sets, among which RF was performed 200 times. The experimental results are shown in Table 3. Based on the experimental analysis, since the eGeMAPS feature is a set of audio features for emotion and gender recognition, it has certain advantages in audio feature classification tasks. The eGeMAPS features cover many aspects of audio information, including pitch, speaking rate, sound intensity, sound quality, and so on, thus providing a comprehensive description of the audio, and they are relatively insensitive to changes in factors such as noise and room acoustics. This makes them suitable for complex real-world voice-recording environments. eGeMAPS features can also be used for gender recognition. They contain gender-related audio features, such as fundamental frequency, resonance peaks, and so on, which can effectively distinguish between male and female voices. Next, we only used the eGeMaps audio features of depressed speech, and verified the classification effect of depression and nondepression features using SVM and CNN classifiers, respectively.

From the experimental results in Table 3, it can be seen that the classification effect using only audio features cannot achieve the ideal effect. Since low-level audio features have a significant impact on

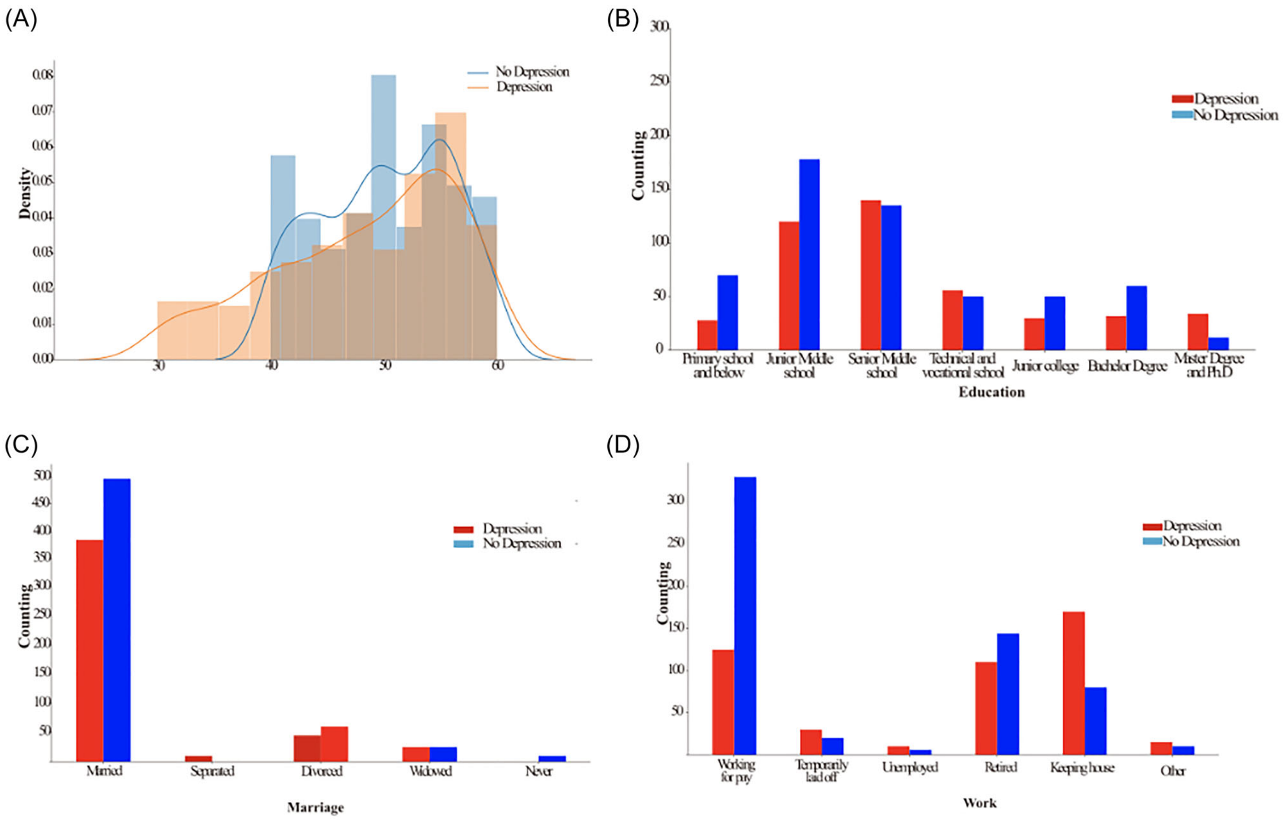

FIGURE 3 Statistical analysis of (A) age, (B) education, (C) marriage status, and (D) work status of depressed and nondepressed patients.

TABLE 3 Comparison of different classification methods: RF, SVM, CNN, BiLSTM, and SGCNs.

Features	Feature dimension	Classifier	Accuracy	Sensitivity	Specificity
MFCC	13	RF	55.6	61.7	54.9
Prosody	103	RF	71.1	72.6	69.5
Phonation	29	RF	69.2	69.1	69.4
Phonological	108	RF	70.8	69	72.7
EGeMaps	88	RF	71	71.1	70.9
		SVM	67	60.0	75.0
		CNN	59	53.0	65.0
ComParE	6373		73.9	73.3	74.5
SJTU-LWDLab DACD	-	BiLSTM	89.2	89.0	90.6
	-	**SGCNs**	**92.4**	**91.1**	**93.7**

Note: The values in bold highlight the results for the model of this paper.
Abbreviations: BiLSTM, bi-directional long short-term memory; CNN, convolutional neural network; MFCC, mel-frequency cepstral coefficient; RF, random forest; SGCN, summed graph convolutional network; SJTU-LWDLab DACD, Depression Audio Consultation Dataset from Li Weidong Laboratory, Shanghai Jiao Tong University; SVM, support vector machine.

distinguishing depression from nondepression, we extracted some low-level features from the audio signals of the real data set. These features contain basic information such as pitch, volume, and spectrum, which are used to compare the classification effects of nongraph-structured data and graph-structured data. From the experimental comparison, it can be seen that our model has the best effect. To verify which specific features impact depression, we conducted a statistical analysis of the input high-dimensional audio features and found that the best feature combination is alcohol, mania, and social life. The other differences were not obvious, and the experiment was relatively random. At the same time, we selected two feature sets, eGeMAPS and ComParE, for confusion experiments. The experimental results are shown in Figure 4. As can be seen from Figure 4A, the predicted true positive (TP) and true negative (TN) are both very high. As can be seen from the ROC

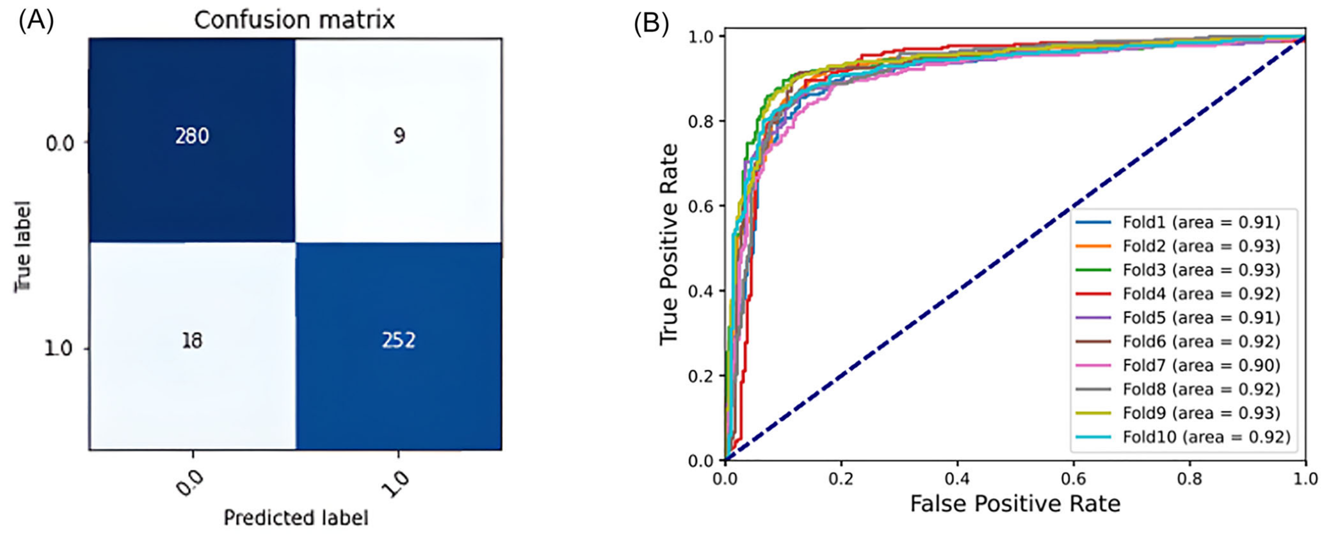

FIGURE 4 Experiments on summed graph convolutional networks (SGCNs) on the ComParE feature set: (A) the confusion matrix and (B) the ROC curve of ComParE features.

TABLE 4 Accuracy values for depressed and nondepressed patients in the public data set, respectively.

Data set	Method	Feature	Model	Accuracy (D/ND)
DAIC–WOZ	LSTM-CNNs-CRF[40]	Mel-spectrogram	CNN-LSTM	52/70
	DRNN[41]	MFCC	RNN	86/85
	Deep learning–based[42]	MFCC, spectrogram	CNN	87/82
	Ours (SGCNs)	**MFCC**	**GCN**	**91/92**

Note: The values in bold highlight the results for the model of this paper.
Abbreviations: CNN, convolutional neural network; CRF, conditional random field; D, depressed; DAIC–WOZ, Depression, Anxiety, and Stress in the Context of Conversational Agents-Wizard of Oz; DRNN, deep recurrent neural network; GCN, graph convolutional network; LSTM, long short-term memory; MFCC, mel-frequency cepstral coefficient; ND, nondepressed; RNN, recurrent neural network; SGCN, summed graph convolutional network.

curve in Figure 4B, the TP rate can reach more than 90%. To ensure the model's reliability and make full use of the audio data, we performed cross-validation 10 times. The division of the data set was random and uniform. As can be seen from Figure 4B, our ROC curve area is between 91% and 93%, indicating that our method has very good reliability and robustness in different audio feature sets.

Experimental comparison with public data sets

In addition, we conducted a comparison experiment between English-speaking depressed patients and Chinese-speaking depressed patients using the SJTU-LWDLab DACD data set and the public data set DAIC-WOZ. The comparison of our modeling approach with other method approaches on the DAIC-WOZ data set is shown in Table 4. Experimental results indicate that the audio features of the real patients and the public data set patients are different, and the classification accuracy of our method on the public data set for depressed and nondepressed patients exceeds that of the other model methods, which fully proves the validity and accuracy of our method.

DISCUSSION

This study had several notable shortcomings. It did not fully consider and control for confounding variables, identify covariates that are strongly correlated with depression, or achieve precise matching between the experimental and control groups. This oversight is significant, as depression is a complex mood disorder with a multifaceted etiology, influenced by both biological and social factors. Although the sample included both male and female participants, the limited size restricted the possibility of separate analyses for each gender, despite the known substantial differences in vocal characteristics between men and women. Future studies should account for these vocal characteristics separately for each gender. Additionally, this study primarily focused on the differences between patients with depression and healthy individuals. In clinical practice, depression is often misdiagnosed as other mental disorders, such as bipolar disorder, complicating the diagnosis process. Therefore, it is essential to stratify the research into more detailed categories and comprehensively examine the role of voice in predicting depression across different populations, including individuals with physical illnesses and those with other mental disorders. This approach will facilitate the development of more rigorous conclusions. Based on the findings herein, this study has established that voice possesses a cross-feature predictive effect on depression. Future research should aim to utilize more effective methods for processing high-dimensional data to identify the most salient features relevant to voice. Enhancing the model's generalization capabilities can be achieved by reducing feature dimensionality.

CONCLUSION

This study gathered a substantial data set of speech recordings from patients diagnosed with clinical depression, known as the SJTU-LWDLab DACD data set. Utilizing this data set, we introduced an innovative speech data classification approach that employs the SGCN for individual depression identification. This method addressed the issue of low accuracy in depression detection caused by inadequate spatial features during the fusion of multiple audio attributes. This research underscores the importance of speech as a rapid identification and diagnostic tool for clinical depression. Considering the findings and limitations of this study, future research could focus on speech features strongly correlated with depression, enabling more precise exploration and positioning. Additionally, it is suggested to implement advanced computational classification algorithms and conduct robust repeatability verification to enhance the model's accuracy, generalization capability, and the overall applicability of the results.

AUTHOR CONTRIBUTIONS

G.L. for manuscript writing; X.M. and J.Y. for software code; Y.L. for overall control; Y.X. and C.L. for data analysis; Y.K. and R.Z. for data set samples; S.L. and K.Y. for experiments; and M.W. and W.L. for review.

ACKNOWLEDGMENTS

This work was supported by STI2030-Major Projects (2022ZD020 4900).

CONFLICT OF INTEREST STATEMENT

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 STATEMENT

The desensitized data supporting the results of this study can be found in the Supporting Information. Audio recordings and MFCC characterization data can be obtained upon written request to the corresponding authors.

CODE AVAILABILITY

https://github.com/IvanBai-Nebula/Speech-depression-model-GCN.

ORCID

Guangsheng Luo https://orcid.org/0000-0003-4278-2971

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/nyas.15366.

REFERENCES

1. Lépine, J.-P., & Briley, M. (2011). The increasing burden of depression. *Neuropsychiatric Disease and Treatment*, *7*(sup1), 3–7.
2. Trifu, R. N., NEMES, B., Bodea-Hațegan, C., & Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *Journal of Evidence-Based Psychotherapies*, *17*(1).
3. Cohen, A. S., Kim, Y., & Najolia, G. M. (2013). Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders. *Schizophrenia Research*, *146*(1–3), 249–253.
4. Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and Cognition*, *56*(1), 30–35.
5. Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry*, *72*(7), 580–587.
6. Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, *22*(6), 688.
7. Scherer, S., Stratou, G., Gratch, J., & Morency, L.-P. (2013). Investigating voice quality as a speaker-independent indicator of depression and PTSD. In *INTERSPEECH* (pp. 847–851).
8. Kwon, N., & Kim, S. (2021). Depression severity detection using read speech with a divide-and-conquer approach. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 633–637). IEEE.
9. Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, *7*, 100943–100953.
10. Li, M., Sun, X., & Wang, M. (2023). Detecting depression with heterogeneous graph neural network in clinical interview transcript. *IEEE Transactions on Computational Social Systems*, *11*(1), 1315–1324.
11. Shou, Y., Meng, T., Ai, W., Yang, S., & Li, K. (2022). Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. *Neurocomputing*, *501*, 629–639.
12. Xia, Y., Liu, L., Dong, T., Chen, J., Cheng, Y., & Tang, L. (2024). A depression detection model based on multimodal graph neural network. In *Multimedia tools and applications.*
13. Salekin, A., Eberle, J. W., Glenn, J. J., Teach-man, B. A., & Stankovic, J. A. (2018). A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 1–26.
14. Yang, W., Liu, J., Cao, P., Zhu, R., Wang, Y., Liu, J. K., Wang, F., & Zhang, X. (2023). Attention guided learnable time-domain filterbanks for speech depression detection. *Neural Networks*, *165*, 135–149.
15. Qin, K., Lei, D., Pinaya, W., Pan, N., Li, W., Zhu, Z., Sweeney, J. A., Mechelli, A., & Gong, Q. (2022). Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites. *EBioMedicine*, *78*, 103977.
16. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., & Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (pp. 3–9).
17. Alghifari, M. F., Gunawan, T. S., Nordin, M. A. W., Kartiwi, M., & Borhan, L. (2019). On the optimum speech segment length for depression detection. In *2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)* (pp. 1–5). IEEE.
18. Patil, M., & Wadhai, V. (2021). Selection of classifiers for depression detection using acoustic features. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (pp. 1–4). IEEE.
19. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, J., Busso, C., … Truong, K. P. (2015). The Geneva minimalistic acoustic

parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190–202.

20. Liu, Z., Hu, B., Yan, L., Wang, T., Liu, F., Li, X., & Kang, H. (2015). Detection of depression in speech. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 743–747). IEEE.
21. Huang, Z., Epps, J., Joachim, D., Stasak, B., Williamson, J. R., & Quatieri, T. F. (2020). Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated CNNs. In *INTERSPEECH* (pp. 4561–4565).
22. Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*, *211*, 106433.
23. Huang, Z., Epps, J., & Joachim, D. (2019). Investigation of speech landmark patterns for depression detection. *IEEE Transactions on Affective Computing*, *13*(2), 666–679.
24. Alsarrani, R., Esposito, A., & Vinciarelli, A. (2022). Thin slices of depression: Improving depression detection performance through data segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6257–6261). IEEE.
25. Lai, Y., Zhang, L., Han, D., Zhou, R., & Wang, G. (2020). Fine-grained emotion classification of Chinese microblogs based on graph convolution networks. *World Wide Web*, *23*, 2771–2787.
26. He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z. m., Pan, X., & Dang, W. (2022). Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, *80*, 56–86.
27. Chen, T., Hong, R., Guo, Y., Hao, S., & Hu, B. (2022). Ms2-GNN: Exploring GNN-based multimodal fusion network for depression detection. *IEEE Transactions on Cybernetics*, *53*(12), 7749–7759.
28. Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., & Hadid, A. (2021). Towards robust deep neural networks for affect and depression recognition from speech. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event* (pp. 5–19). Springer.
29. Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., & Schuller, B. (2019). A hierarchical attention network-based approach for depression detection from transcribed clinical interviews.
30. Sun, G., Zhao, S., Zou, B., & An, Y. (2022). Speech-based depression detection using unsupervised autoencoder. In *2022 7th International Conference on Signal and Image Processing (ICSIP)* (pp. 35–38). IEEE.
31. Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. *Measurement: Sensors*, *25*, 100587.
32. Figueredo, A. J., & Wolf, P. S. A. (2009). Assortative pairing and life history strategy: A cross-cultural study. *Human Nature*, *20*, 317–330.
33. Lian, Z., Liu, B., & Tao, J. (2021). CTNET: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 985–1000.
34. Morales, M. R., & Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 136–143). IEEE.
35. Di, Y., Wang, J., Li, W., & Zhu, T. (2021). Using i-vectors from voice features to identify major depressive disorder. *Journal of Affective Disorders*, *288*, 161–166.
36. Pestian, J. P., Sorter, M., Connolly, B., Cohen, K. B., McCullumsmith, C., Gee, J. T., Morency, L.-P., Scherer, S., & Rohlfs, L., STM Research Group. (2017). A machine learning approach to identifying the thought markers of suicidal subjects: A prospective multicenter trial. *Suicide and Life-Threatening Behavior*, *47*(1), 112–121.
37. Gong, Y., & Poellabauer, C. (2017). Topic modeling based multi-modal depression detection. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (pp. 69–76).
38. Beaman, J., Lawson, L., Keener, A., & Mathews, M. L. (2022). Within clinic reliability and usability of a voice-based Amazon Alexa administration of the Patient Health Questionnaire 9 (PHQ 9). *Journal of Medical Systems*, *46*(6), 38.
39. Lawson, L., Beaman, J., & Mathews, M. (2024). Within clinic reliability and usability of a voice-based Amazon Alexa administration of the General Anxiety Disorder 7 (GAD 7). *Journal of Medical Systems*, *48*(1), 70.
40. Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354.
41. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, *71*, 103107.
42. Othmani, A., Zeghina, A.-O., & Muzammel, M. (2022). A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. *Computer Methods and Programs in Biomedicine*, *226*, 107132.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Luo, G., Ma, X., Yea, J., Liu, Y., Xia, Y., Li, C., Kuang, Y., Zhang, R., Lou, S., Yu, K., Wu, M., & Li, W. (2025). Audio multi-feature fusion detection for depression based on graph convolutional networks. *Ann NY Acad Sci.*, 1–12. https://doi.org/10.1111/nyas.15366