Rapid Report

# Speech depression recognition based on attentional residual network

Xiaoyong Lu[1,2,*], Daimin Shi[3], Yang Liu[3], Jingyi Yuan[3]

[1] *Internet Education Data Learning Analysis Technology National and Local Joint Engineering Laboratory, 730070 Lanzhou, Gansu, China,* [2] *Key Laboratory of Behavior and Mental Health of Gansu Province, 730070 Lanzhou, Gansu, China,* [3] *School of Physics and Electronic Engineering, Northwest Normal University, 730070 Lanzhou, Gansu, China*

## TABLE OF CONTENTS

## 1. Abstract

**Background**: Depressive disorder is a common affective disorder, also known as depression, which is characterized by sadness, loss of interest, feelings of guilt or low self-worth and poor concentration. As speech is easy to obtain non-offensively with low-cost, many researchers explore the possibility of depression prediction through speech. Adopting speech signals to recognize depression has important practical significance. Aiming at the problem of the complex structure of the deep neural network method used in the recognition of speech depression and the traditional machine learning methods need to manually extract the features and the low recognition rate. **Methods**: This paper proposes a model that combines residual thinking and attention mechanism. First, depression corpus is designed based on the classic psychological experimental paradigm self-reference effect (SRE), and the speech dataset is labeled; then the attention module is introduced into the residual, and the channel attention is used to learn the features of the channel dimension, the spatial attention feedback the features of the spatial dimension, and the combination of the two to obtain the attention residual unit; finally the stacking unit constructs a speech depression recognition model based on the attention residual network. **Results**: Experimental results show that compared with traditional machine

Submitted: 21 April 2021 Revised: 19 August 2021 Accepted: 9 October 2021 Published: 30 December 2021

learning methods, this model obtains better results in the recognition of depression, which can meet the need for actual recognition application of depression. **Conclusions**: In this study, we not only predict whether person is depressed, but also estimate the severity of depression. In the designed corpus, the depression binary classification of an individual is given based on the severity of depression which is measured using BDI-II scores. Experimental results show that spontaneous speech can obtain better results than automatic speech, and the classification of speech features corresponding to negative questions is better than other tasks under negative emotions. Besides, the recognition accuracy rate of both male and female subjects is higher than that under other emotions.

## 2. Introduction

Depressive disorder is a common affective disorder, also known as depression, which is characterized by sadness, loss of interest, feelings of guilt or low self-worth, and poor concentration. Statistics from the World Health Organization show that there are approximately 350 million depressed patients in the world, of which more than 54 million are depressed in China [1]. At home, the lifetime prevalence of depression is as high as 6.8% [2]. It has brought great harm to society, families, and individuals.

Usually, the diagnosis of depression is based on scales, supplemented by the judgment of clinicians. The success of this treatment seems to rest a lot with the cooperation of the patient and the experience and professional level of the clinician [3]. In recent years, the number of depressed patients has been increasing. Accurate and rapid diagnosis of depression is a huge challenge for relevant medical staff. The use of computers to diagnose depression can provide clinicians with an initial entry point. The availability and effectiveness of speech have been proven to be a reliable objective indicator of depression [4]. For example, studies have found that compared with healthy people, there is a significant correlation between the clinical level of depression individuals and prosodic features [5, 6] and the subjective features of the speech of depressed individuals are low, slow, hesitant, and monotonous [7]. As speech is easy to obtain non-offensively at a low-cost, many researchers explore the possibility of depression prediction through speech. In the past decade, predicting emotion using speech has been a success. Corollary to that, the proof of speech brings useful in predicting clinical depression scores implies that speech is an important indicator for depression. There are also reports stating that the verbal behavior of depressed patients should have “decreased verbal activity productivity, a diminished prosody and monotonous, ’lifeless’ sounding speech”. Besides, the rise of deep learning technology has prompted researchers to explore different deep architecture and methods for evaluating depression [8, 9]. The traditional classification of depression mainly chooses better features [10] and classifiers [11] and so on. With the rapid development of deep learning, researchers have shifted their focus to the construction of deep neural networks to achieve automatic feature extraction and classification. Literature [12] proposed a depression recognition method based on dynamic convolution neural network (DCNN), which combines manually extracted features and high-level features to achieve the purpose of identifying depression, but manually extracting features require a lot of manpower and material resources. Literature [13] proposed an integrated learning method based on Convolutional Neural Network (CNN), which was evaluated on the AVEC2016 dataset and achieved good results. Literature [14] proposes a deep model that combines CNN and long and short-term memory network (LSTM) and introduces a random sampling strategy in the training phase to balance the samples so that the classification results do not appear to be biased. Literature [15] input the speech spectrogram and Mel-frequency cepstrum coefficient (MFCC) features into the CNN network, and finally joint fine-tuning to realize the recognition of depression. The complexity of the model, the large number of fine-tuning parameters and the time-consuming training have become the drawbacks of this method. Literature [16] uses a deep convolutional neural network DCNN to identify depression on the public dataset AVEC2017 to improve the accuracy of the experiment. The above methods have achieved good recognition rates on the open source English dataset, but none of them have been trained on the Chinese corpus. Given the high feature dimensionality of traditional depression recognition algorithms, insufficient feature extraction capabilities of classic deep learning models, and the existence of gradient disappearance when the depth of the deep network is too deep. This paper combines residual neural network through feature extraction and gradient prevention. With the advantage of descent and the advantage of the attention mechanism for obtaining refined features, the attention residual depression classification algorithm is proposed. The contributions of this work can be summarized by the following.

Establish depression corpus based on the classic paradigm of psychological self-reference effect (SRE).

Introduce the attention mechanism into the residual module and combine the two to obtain the attention residual unit, and finally stack the units to obtain a high-performance attention residual network.

Extract the MFCC features of the speech and establish a depressive recognition model to further recognize the depression.

The remaining paper is organized as follows: the proposed methodology is explained in Section 3, the database description is presented in Section 4, Section 5 demonstrates the experimental results and the conclusions of the paper are discussed in Section 6.

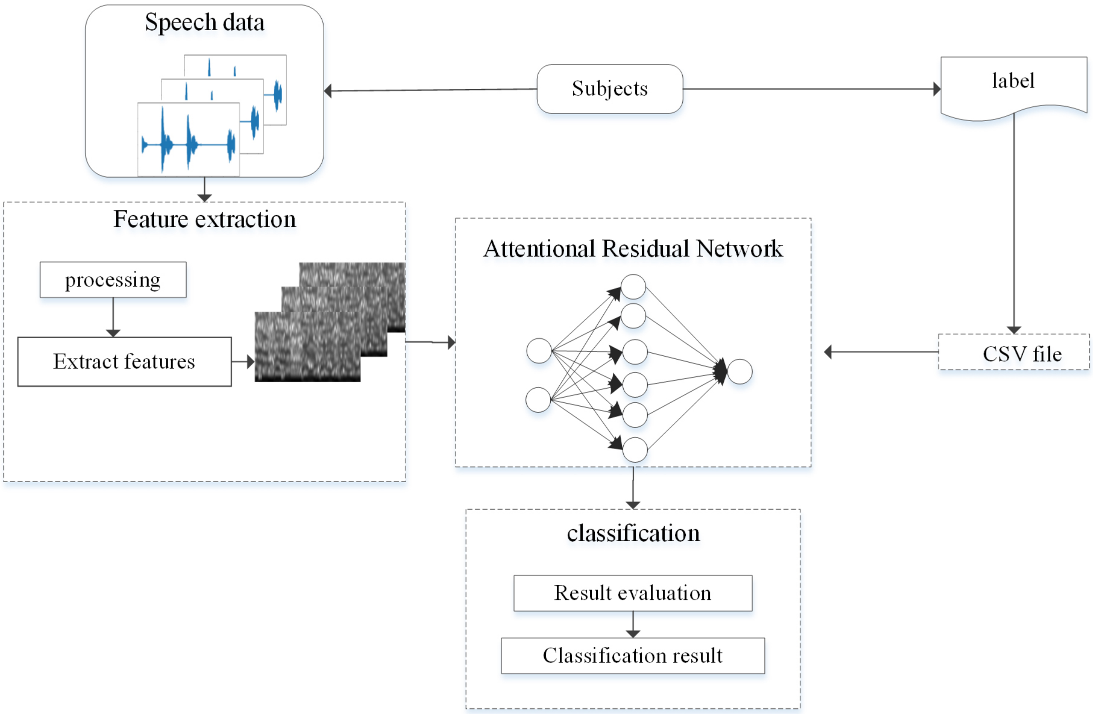

**Fig. 1. The architecture of our system of depression recognition.**

## 3. Recognition model of depression

In this section, three important stages: (i) design a corpus collection method, (ii) pre-processing and segmentation, and (iii) depression classifification are discussed. In the first stage, the corpus collection method is designed and the classic psychological experimental paradigm SRE is adopted. After designing corpus, the signal is segmented into audio segment. In the second stage, individual segmented are transformed into two-dimensional time-frequency images. Finally, attention residual is implemented for the classifification of speech using MFCC features as input. The overall stages of the proposed methodology is shown in Fig. 1.

### 3.1 Network architecture

In recent years, convolutional neural networks are widely used in emotion recognition and depression recognition field. The features of speech can better characterize depression. Since the information of the sound is input to the convolution in the form of images, many areas of the texture and contour features of the feature map have low pixel values and are sparse, which will help reduce the training time, and it can also reduce the difficulty of network learning and make it easier to converge. Given this idea, this paper proposed attention residual model, as shown in Fig. 2. The convolutional layer is used to extract deep-level features, and then the features are input to the stacked residual attention module for features in diverse dimensions the transformation and extraction. The output from the both architectures are flattened and fused in a fully-connected layer in order to generate the label prediction for binary (0 for non-depression and 1 for depression) and 24 depression severity levels. Due to different depths of residual convolution used, the number of stacked residual attention modules is also different, and the performance of the network also changes accordingly.

### 3.2 Residual module

The basic idea of ResNet is to fit a residual function, rather than fitting a direct mapping function of input and output. Take M(x) as the direct mapping function fitted by the stack of the feedforward neural network layer, where x represents the input of the layer, assuming that the stack of the feedforward neural network layer can be approximated as a complex variable function, then the feedforward neural network layer the residual function can be approximated, namely $M(x) – x$. Therefore, instead of the approximate function M(x), the residual function is approximated as $G(x) = M(x) – x$. By reconstructing the equation, the original function $M(x) = G(x) + x$ [17]. However, the identity mapping is not optimal, but the residual reconstruction helps to eliminate this problem [17]. The residual unit directly maps the low-level features to the high-level network through a short connection. This method directly detours the input information and outputs to ensure the integrity of the information. The structure diagram of the residual module is shown in Fig. 3.

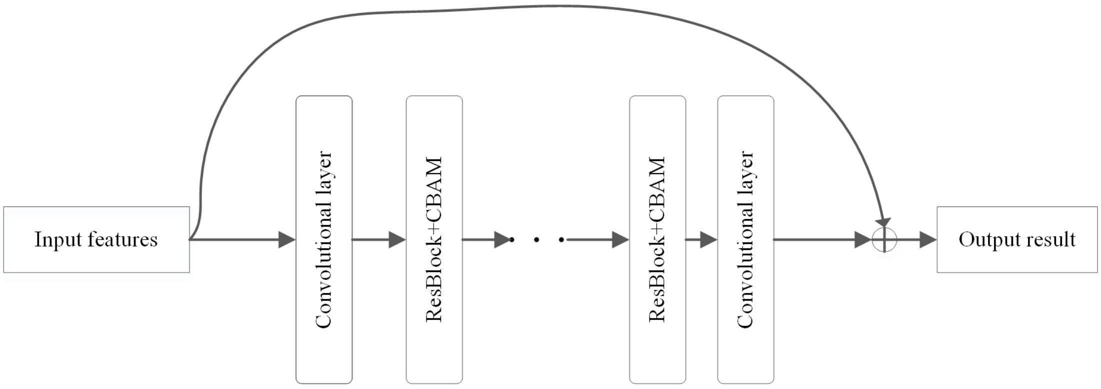

**Fig. 2. The pipeline of the proposed architecture for the recognition of depression.**

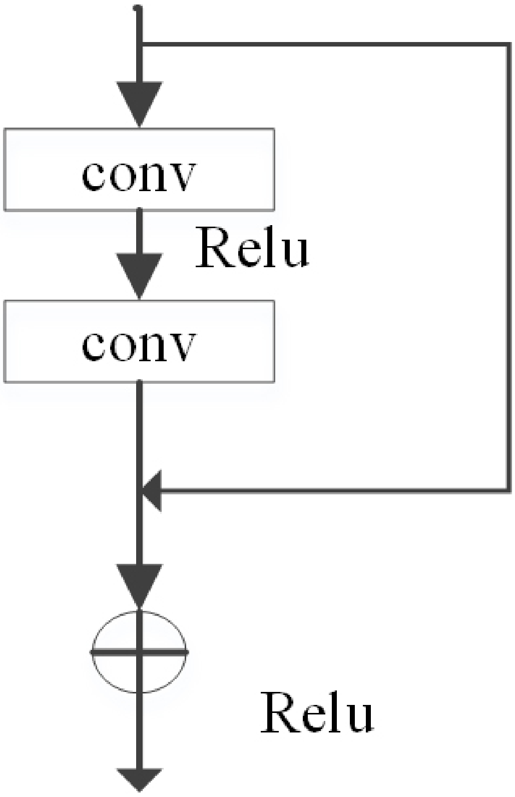

**Fig. 3. The detailed illustration of the residual module.**

### 3.3 Attention module

Attention modules include spatial attention modules and channel attention modules. According to the reference [18], the attention module of convolutional block attention module (CBAM) is shown in the figure below. The channel attention map (Fig. 4), which compresses the channel, and performs average pooling and maximum pooling in the channel dimension to generate the input features required by the spatial attention module.

The spatial attention maps (Fig. 5). The spatial attention module compresses the feature map in the spatial dimension to obtain a one-dimensional vector before performing operations. When compressing in the spatial dimension, not only the average pooling is considered but also the maximum pooling is considered to aggregate the spatial information of the feature map, send it to a shared network, compress the spatial dimension of the input feature map, and sum and merge it element by element.

### 3.4 Attention residual module

Through the attention module, the network can be more focused on feature extraction. In recent years, attention mechanism technology has gradually become a research hotspot. With the help of the attention mechanism, feature information can be paid more comprehensively. According to certain rules, the convolutional neural network can obtain local information according to window sliding, and the attention mechanism can improve the volume. The ability to express product. In this paper, the attention module is combined with the residual module to construct the attention residual module to improve the performance of the model (Fig. 6), after the feature map passes through the previous convolutional layer, it passes through the channel attention module and the spatial attention module in turn. Finally, the feature map and the residual module are added and output to the subsequent convolutional layer.

The convolutional layer has a certain number of channels. Adding channel attention can allow each channel of the convolution to learn features related to the recognition task while suppressing irrelevant features, to achieve the network focusing on the required features. The spatial attention module can better locate the fringe and contour positions of related features, allowing the network to focus on features that can recognize depression.

### 3.5 Attention residual model is based on MFCC

One of the numerous advantages of the MFCC feature is its ability to withstand noise. Short-term spectrum analysis is the most common method of using MFCC to characterize speech signals. MFCCs can show the variation of the key bandwidth of the human ear with frequency, and they can capture the important speech features of speech by representing the signal on a Mel-frequency scale. In this paper, from the preprocessed speech fragments, librosa li-

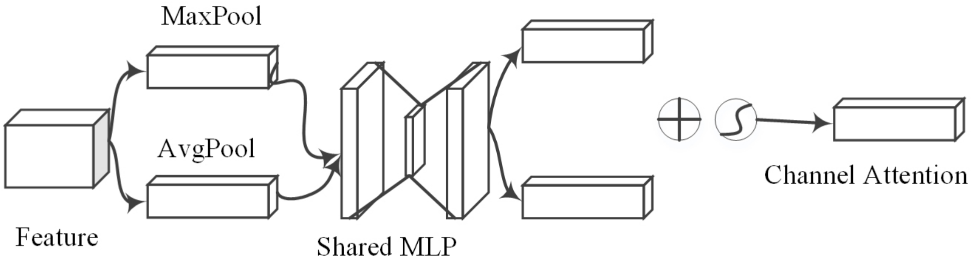

**Fig. 4. Diagram of channel attention.**

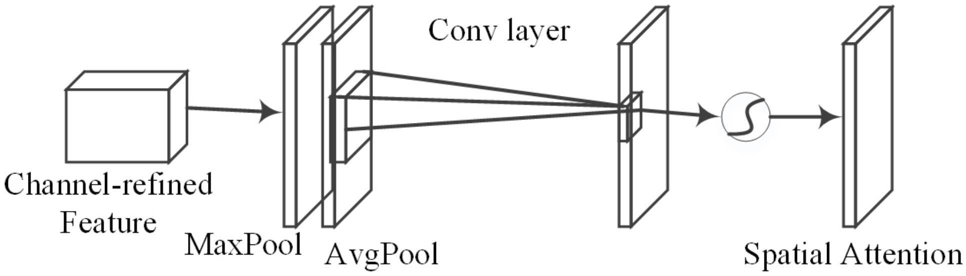

**Fig. 5. Diagram of spatial attention.**

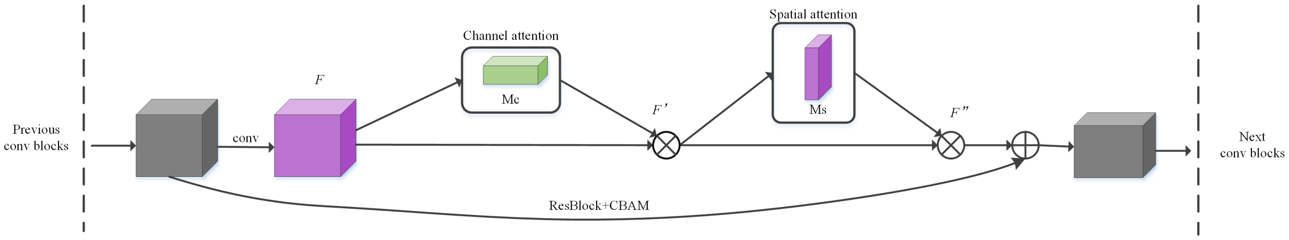

**Fig. 6. CBAM integrated with a ResBlock in ResNet.**

brary functions are used to generate MFCC features. The feature extraction process is shown (Fig. 7). An example of the MFCC of depressed and normal people generated from the speech fragments after the pre-processed speech samples are segmented (Fig. 8).

In the network structure (Fig. 9), the residual neural network has five convolution modules. Each convolution module consists of multiple convolution layers. Each convolution layer contains a convolution kernel, a batch normalization layer, and an activation function. In this paper, an attention mechanism is added after each convolution module. The features are input into two independent dimensions (channel and spatial). Then adaptive feature refinement is performed. After the average pooling layer and the flattening layer, the Softmax activation function is used to activate at last, and the classification result is obtained. The convolution kernel size of Conv1 is $7 \times 7$, and the step size is 2. Each of the remaining convolutional layers uses a $3 \times 3$ convolution kernel with a step size of 2. In the Pytorch framework, the construction of the network is realized.

In the residual network, the residual path is bypassed through three convolutional layers to deal with the gradient problem. The classification cross-entropy is used as the loss function, and the model does not use the discard rate because it will increase the training time. When the input and output sizes match, the residual path will be used directly. Otherwise, zero entries will be filled to increase the dimension with identity mapping. The convolutional layer of each stage is batch normalized and then activated in turn. The network weights are initialized as described in [17]. The algorithm steps are as follows:

**Input:** MFCC features.

**Output:** Recognition result.

**Step 1:** Segment the speech data and expand the sample size to train the attention residual network.

**Step 2:** Find the best learning rate value, at this time the loss function value is still falling.

**Step 3:** Freeze the previous attention residual layer and leave the last layer.

**Step 4:** Train the last layer.

**Step 5:** Unfreeze the previous attention residual layer.

**Step 6:** Looks for the best learning rate again, and the loss function value is still declining.

**Step 7:** Train the entire network until the network is overfitted.

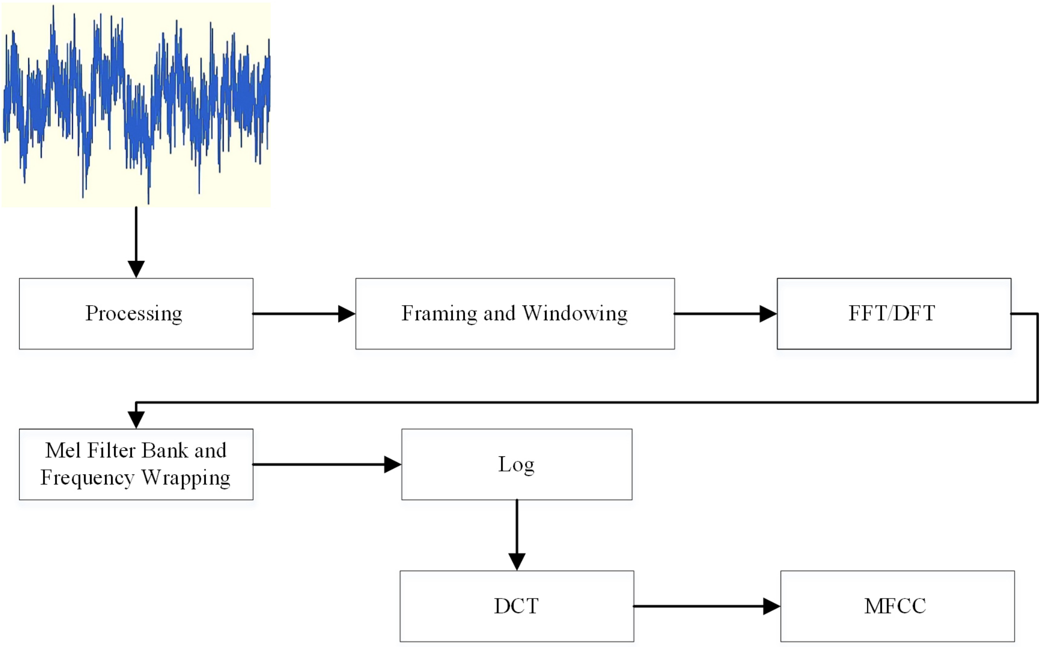

**Fig. 7. Steps for calculating MFCC coefficients.**

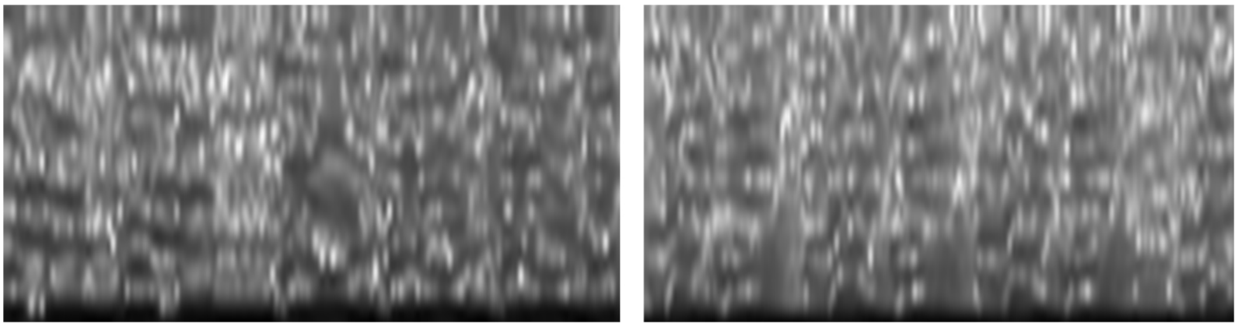

**Fig. 8. Normal MFCC and depressed MFCC of subjects.**

The Softmax [19] classifier is a generalization of the logarithmic model on multi-classification tasks. When the classification is 2, it degenerates into a logistic regression classification. In the two-classification task, the decision function of Softmax regression is:

$$\tilde{y} = \frac{\exp\left(\hat{y}_n\right)}{\sum_k \exp\left(\hat{y}_k\right)} \tag{1}$$

Among them, y represents the category label, k represents the sample, and n represents the number of neurons in the output layer.

# 4. Experimental results and analysis

## 4.1 Corpus acquisition

The 25 depressive subjects were all graduates and undergraduates from Northwest Normal University. A well-trained trainer will perform a diagnostic evaluation of depression on the subjects according to the hamilton depression scale (HAMD) [20] and The beck depression inventory (BDI-II) [21]. During the test period, all subjects were free of drug dependence, abuse, and other mental illnesses, and no serious physical disorders, suicides, and other dangerous behaviors.

The 25 healthy subjects participated in the study as a control group. They are all students from Northwest Normal University recruited by advertisements. There was no significant differences in gender, age, and education level between normal subjects and those in the depression group. Before the start of the experiment, we re-evaluated both the depressed subjects and healthy subjects and the BDI-II evaluation to exclude those who did not meet the requirements and screen out those who met the requirements. Participant (Table 1) reports detailed demographic information.

The criteria for entry into the group for depression subjects: (1) age between 16–25; (2) BDI-II scale score between 14–28 score, indicating that the degree of depression of the patient is mild to moderate; (3) none other mental

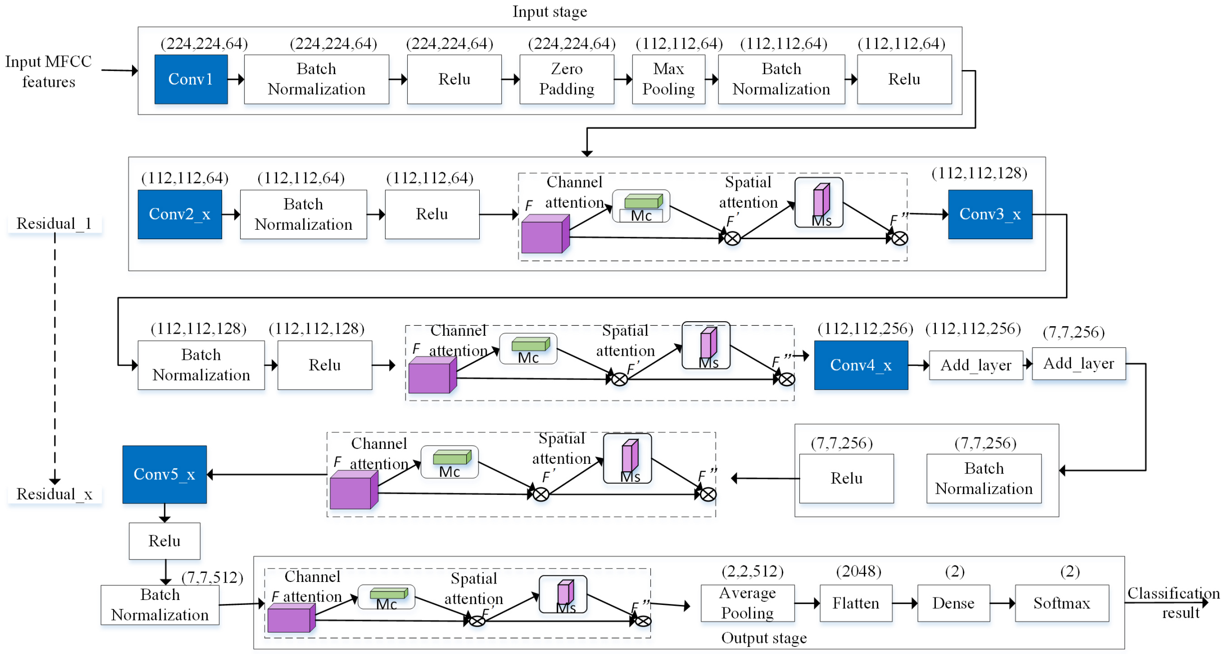

**Fig. 9. The neural network structure for the proposed technique.**

**Table 1. Demographic information of depressed patients and healthy controls*.**

| | Depression group (n = 25) | Healthy group (n = 25) |
|---|---|---|
| Age (years) | 20.4 ± 1.43 | 20.44 ± 1.15 |
| Education level (years) | 18 ± 0 | 18.04 ± 0 |
| Gender (Male/Female) | 10/15 | 12/13 |
| BDI-II | 16.52 ± 2.0 | 5.52 ± 1.97 |

*The presentation format of the data is the mean (standard deviation).

illnesses and suicidal behaviors occur; (4) no drug dependence and abuse and organic diseases; (5) no obvious signs of depression on the outside; (6) no aggressiveness, and the ability to understand and cooperate alone experiment.

Enrollment criteria for healthy control subjects: (1) age between 16–25; (2) BDI-II scale score below 14 points, indicating that the subject is not depressed; (3) HAMD and BDI-II amount the scores of the table evaluation are all within the normal range; (4) there is no history of physical impairment. All subjects have standard Mandarin, right-handed, and normal or corrected vision. Before the experiment, all subjects signed an informed consent form.

If the entire recording process continues, the participant's attention or reaction ability may be reduced. Therefore, the recording process is divided into multiple stages, with the participant taking a break and pause in the middle. The scene diagrams of the collected data (Fig. 10).

This study adopts an experimental design of 2 (subject type: healthy group, depression group) × 3 (emotional state: positive, neutral, negative) × 4 (task type: vocabulary reading, essay reading, interview, picture description) the way. The process of collecting data is shown in (Fig. 11).

**Fig. 10. Scene map of corpus collection subjects.**

The subjects were presented with vocabularies, short essays, pictures, and interview tasks to complete the corpus design. The participant type is between-group variables, and the task type is within-group variables. Besides, depressed subjects have damages of self-processing and self-recognition, negative self-reference, over-attention to

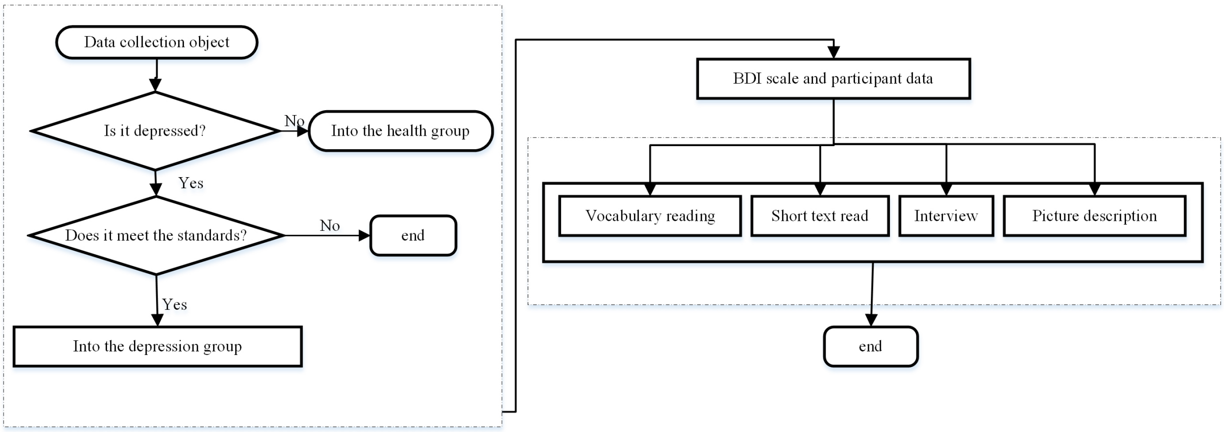

**Fig. 11. The flow chart of corpus design and collection.**

**Table 2. Recording duration of depressed patients and healthy controls*.**

| Tasks | Depression group (n = 25) | Healthy group (n = 25) |
|---|---|---|
| Vocabulary reading (minutes) | 1.88 ± 0.59 | 1.83 ± 0.43 |
| Essay reading (minutes) | 1.85 ± 0.22 | 1.86 ± 0.29 |
| Interview (minutes) | 7.27 ± 3.14 | 5.36 ± 1.85 |
| Picture description (minutes) | 1.84 ± 0.71 | 1.59 ± 0.67 |
| Four tasks (minutes) | 3.25 ± 1.64 | 2.67 ± 1.00 |

*The presentation format of the data is the mean (standard deviation).

self, and use the classic paradigm of SER to the corpus. Depressed patients have damages of self-processing and self-recognition, negative self-reference, over-attention to self, and use the classic paradigm of psychological self-reference effect to design depression corpus.

The experiment was conducted in a professional recording studio. Participants held professional recording equipment R-26. All tasks and instructions were in the same computer program. Subjects sat about 1 m away from a 21-inch computer screen directly in front of the screen, and the task requirements were displayed on the screen. After each task material is presented, the participant needs to answer according to the prompts. The principal examiner operates the program from the side and uses the recording device to record the subject's speech. The sampling rate of the speech file is 44.1 Khz and saved as a WAV file in the form of mono format. The corresponding duration and statistical information about each recording task (Table 2).

## 4.2 Data preprocessing

To obtain optimized features, the noise of the device, the participant's coughing, misreading, and re-reading parts, or the subject's interference with the participant, such as correcting the wrong pronunciation and prompting the specific details of the task is removed. To avoid the personal characteristics of the subject's speech from affecting the accuracy of the experiment, Python was used to programmatically cut the subject's long silent period of speech. On the one hand, to ensure the accuracy of the experiment, on the other hand, to increase the corpus. Each participant's interview was cut into a 4-second segment. According to the literature [13, 15], the results obtained from the 4 s speech segment are relatively objective. A total of 196 samples in the original data set are defined as data set A, and the samples after 4 s segmentation are defined as data set B (Table 3). Based on the score of the BDI-II scale, the training data set is labeled. The sample rate of the original recorded speech samples is 44.1 kHz. This research has carried out re-sampling processing to reduce the sampling rate to 16 kHz.

**Table 3. Distribution table of original data and segmented data.**

| Dataset A | | Dataset B | |
|---|---|---|---|
| Classification | Subjects | Classification | Subjects |
| Depression | 96 | Depression | 4755 |
| Healthy | 100 | Healthy | 3910 |

## 4.3 Experimental environment

To reasonably evaluate the effectiveness of the method in this paper, the hardware, and software environments of the experiment are as follows. Hardware environment: processor: 56 Intel(R) Xeon(R) CPU@2.00GHz; GPU: two NVIDIA GTX TITAN XP, 25 GB; memory: 16 GB. Software environment: operating system: Ubuntu18.04.3; programming language: Python3.6; deep learning framework: Pytorch1.3 and Sklearn, CUDA10.0.

### 4.4 Model parameters

In the process of adapting the model to our dataset, we adjusted two hyperparameters. Hyperparameters express the high-level properties of the model, and there is no way to learn in the general learning process. Algorithm 1 presents the method of parameter fine-tuning. The model parameter values (Table 4).

**Table 4. Model hyperparameter configuration.**

| Hyperparameter | Value |
|---|---|
| Learning rate | 0.001 |
| Batch sizes | 16 |
| epochs | 20 |
| Momentum optimization | 0.9 |
| Weight decay | 0.003 |

**Algorithm** The method of parameter fine-tuning

```
given Learning rate $\eta$ = 0.0001, $\gamma$ = 0.001, $\lambda \in \Re$
initialize time step t$\leftarrow$ 0, parameter vector $\theta_t$ = 0
repeat
t$\leftarrow$ t + 1
$\bigtriangledown$f$_t$($\theta$t-1)$\leftarrow$ SelectBatch($\theta_{t-1}$)
g$_t$ $\leftarrow$ $\bigtriangledown$f$_t$($\theta_{t-1}$) + $\theta_{t-1}$
$\theta_t$ $\leftarrow$ $\eta$*$\bigtriangledown$f$_t$($\theta_{t-1}$) + $\lambda\theta_{t-1}$
$\theta_t$'$\leftarrow$ $\gamma$*$\bigtriangledown$f$_t$($\theta_{t-1}$) + $\lambda\theta_{t-1}$
until stopping criterion is met
return optimized parameters $\theta_t$'
```

### 4.5 Evaluation metrics

To evaluate the classification performance of selected features, the evaluation principle is that the accuracy of depression classification can be determined by true positives, true negatives, false positives, and false negatives. Among them, Accuracy is called the accuracy rate, which refers to the number of all samples that are identified as depression, that is, the number of depressed ones that are identified as depression, and the proportion of healthy ones that are identified as depression. Precision is called accuracy or accuracy rate refers to the proportion of samples whose real category is depression among the samples that are predicted to be depressed. The recall is called the recall rate, which represents the proportion of samples that are correctly predicted to be depressed among all samples whose real category is depression, F1 score is a comprehensive indicator, which is the harmonic average of precision and recall. The AUC metric is used to evaluate the classifier's ability to discriminate between positive and negative examples.

## 5. Comparison experiment settings

Choosing a classifier with good practical results in the experimental classification process. This paper uses Gaussian Process (GP), Support Vector Machines (SVM), and k-Nearest Neighbor (KNN). Compare and analyze the classification results with the AdaBoost algorithm.

### 5.1 Result analysis

In this study, we not only predict whether the person is depressed but also estimate the severity of depression. In the designed corpus, the depression binary classification of an individual is given based on the severity of depression which is measured using BDI-II scores. These predictions are performed under the following section.

### 5.2 Performances on depression binary classification

After the collected data is preprocessed, extracted MFCC features are input into the classification algorithm for comparison. Keeping the output results of each verification. To verify the influence of the attention residual mechanism on the classification results, this paper gives a set of comparative experiments. The residual module is used as a reference model to compare with the stacked attention residual module (Fig. 12).

It can be seen from the above results that compared to the baseline model, the classification performance is significantly improved, which shows that after the attention mechanism is added, the four evaluation indicators have reached relatively ideal results, and the recognition of depression is greats help. This is because after the attention mechanism is added, more accurate features can be obtained without changing the size of the input features, which is more conducive to classification. On the other hand, it highlights the effectiveness of this method. In this paper, after adding the attention mechanism, the trained model is used to classify the samples, and the resulting confusion matrix (Fig. 13).

Among them, the abscissa represents the predicted label, and the ordinate represents the actual label. The sample design in this paper was carried out under different emotional stimuli, and different stimuli under different tasks were investigated with the help of the attention residual algorithm (Table 5).

Under the stimulation of positive emotions, the vocabulary reading task has the best prediction effect, with the highest prediction accuracy rate of 76% and the recall rate of 88%; under the activation of neutral emotion, the interview task has the best effect, and its highest prediction accuracy rate is 84.5%, and the recall rate is 85%; among the negative materials, the picture description has the best effect, its prediction accuracy rate is 89%, and the recall rate is 92%. Overall, under different tasks and starting conditions, the predictive effect of speech features on whether an individual is depressed is basically above 60%. From an emotional point of view, speech features can better distinguish whether individuals are prone to depression in the positive, middle, and negative emotions. This shows that compared with normal people, depressive individuals do have emotional abnormalities, and this abnormality is re-

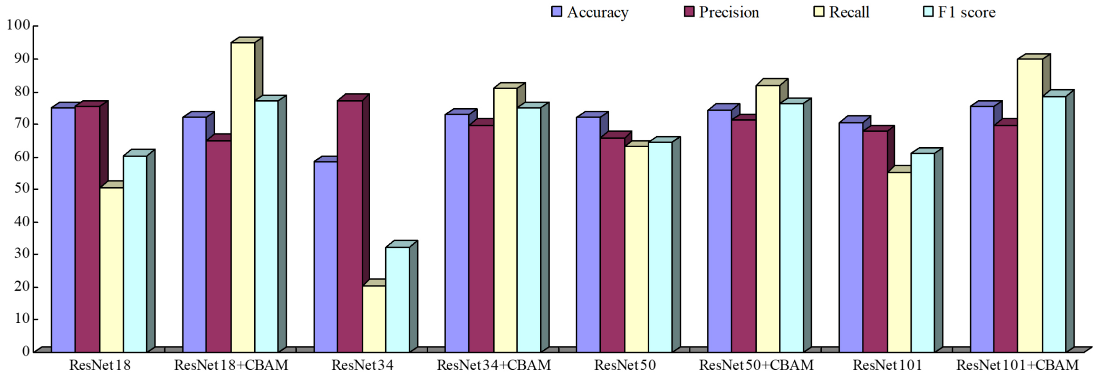

**Fig. 12. MFCC prediction results (%).**

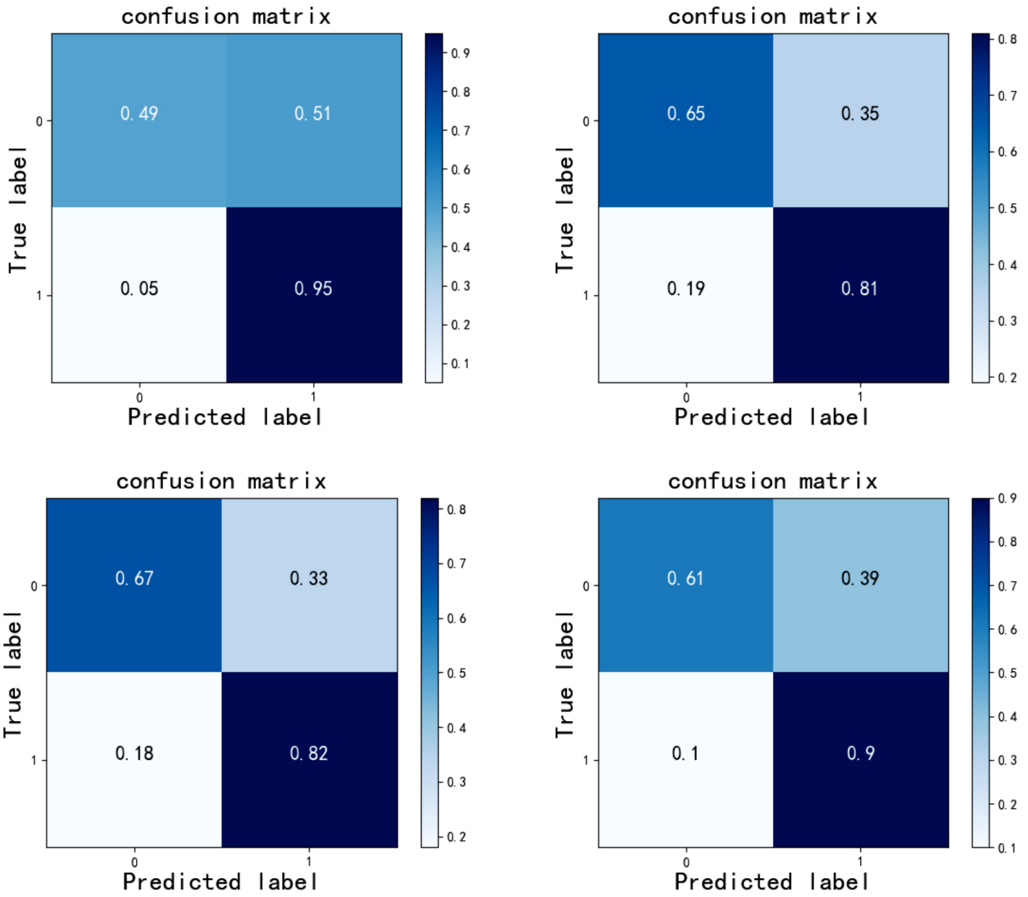

**Fig. 13. Confusion matrix of different algorithms on Database: ResNet50+CBAM demonstrates good performance in identifying depression.**

flected in the individual's speech characteristics. From the perspective of the task, the model under the picture description task is more ideal under the positive and negative emotional valence. The possible reason is that the picture description task requires the subjects to describe their own emotional experience based on the pictures displayed on the screen, which can better reflect the individual emotions. The interview task also created an emotional environment. The lower accuracy rate than the picture description may be because there is no stronger emotional stimulation. The overall situation of the vocabulary reading task is not as good as the above two tasks. It may be because the task requires the participant to read aloud according to the text displayed on the screen, which does not stimulate the individual's inner experiences and self-feelings. The effect is not as good as the above tasks. It is consistent with the results obtained in the literature [4, 22, 23]. Spontaneous speech (such as interviews, picture descriptions) can obtain better results than automatic speech (such as reading aloud), and the classification of speech features correspond-

**Table 5. Model classification results of different tasks under different emotional stimuli (%).**

| Models | Metrics | Vocabulary reading | | | Interview | | | Picture description | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Neutral | Negative | Positive | Neutral | Negative | Positive | Neutral | Negative |
| ResNet18 + CBAM | Accuracy | 76 | 63.5 | 65.5 | 69.5 | 82 | 71 | 77 | 70 | 89 |
| | Precision | 70.1 | 62.6 | 60.7 | 65.6 | 74.2 | 63.7 | 80.7 | 68.9 | 86.8 |
| | Recall | 88 | 67 | 88 | 82 | 98 | 74 | 71 | 73 | 92 |
| | F1 score | 78.6 | 64.7 | 71.2 | 72.89 | 84.4 | 82.7 | 75.5 | 70.9 | 89.3 |
| | AUC | 79 | 68 | 69 | 78 | 93 | 84 | 71 | 84 | 85 |
| ResNet34 + CBAM | Accuracy | 59 | 63 | 65.5 | 66.5 | 84.5 | 70.5 | 64.5 | 81 | 82 |
| | Precision | 56.8 | 60.8 | 61.8 | 62.2 | 84.2 | 69.2 | 62.8 | 77.7 | 80.2 |
| | Recall | 75 | 73 | 81 | 84 | 85 | 74 | 71 | 87 | 85 |
| | F1 score | 64.7 | 66.4 | 70.1 | 71.5 | 84.6 | 71.5 | 66.7 | 82.1 | 82.5 |
| | AUC | 69 | 74 | 62 | 80 | 88 | 82 | 68 | 85 | 89 |
| ResNet50 + CBAM | Accuracy | 72 | 66.5 | 63 | 70 | 84.5 | 76.5 | 65.5 | 65.5 | 74 |
| | Precision | 65.3 | 68.9 | 61.6 | 64.5 | 93.7 | 70.9 | 65.9 | 63.48 | 72.6 |
| | Recall | 94 | 60 | 69 | 89 | 74 | 90 | 64 | 73 | 77 |
| | F1 score | 77.1 | 64.2 | 65.1 | 74.8 | 82.7 | 79.3 | 64.9 | 67.9 | 74.8 |
| | AUC | 73 | 68 | 63 | 86 | 92 | 85 | 65 | 71 | 85 |
| ResNet101 + CBAM | Accuracy | 62 | 50 | 63 | 76.5 | 85 | 77.5 | 73 | 77 | 78 |
| | Precision | 58.7 | 50 | 63.3 | 73 | 93.8 | 72 | 73.9 | 72.5 | 71.9 |
| | Recall | 81 | 67 | 62 | 84 | 75 | 90 | 71 | 87 | 92 |
| | F1 score | 68.1 | 57.3 | 62.6 | 78.1 | 83.3 | 80 | 72.4 | 79.1 | 80.7 |
| | AUC | 76 | 76 | 67 | 84 | 89 | 89 | 66 | 88 | 88 |

ing to negative questions is better than other tasks. This may imply that depression-prone individuals and healthy people have the different cognitive processing methods and negative self-referencing, which make different emotional experiences when facing spontaneous negative tasks, which are reflected in the speech. When designing the sample, both male and female groups were covered, due to the potential difference in depression performance between male and female speech. This paper examines the two genders, and the results of the experiment are shown in (Fig. 14). It is obvious from the figure that the accuracy of women is higher than that of men, which is consistent with the results of [24], indicating that gender can affect the classification performance of the model.

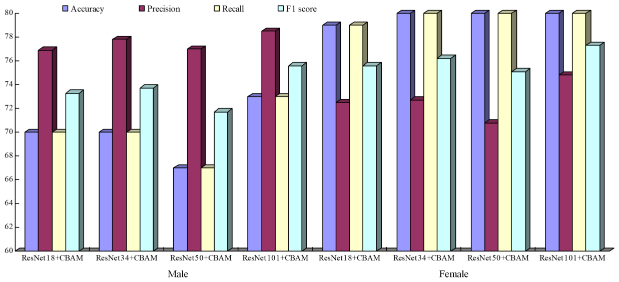

**Fig. 14. Model classification results of different tasks in different genders (%).**

The results obtained are compared with the results obtained using different machine learning algorithms. Also, this article compares the experimental results obtained with those obtained in other documents. The results are shown in (Fig. 15). According to Fig. 15, in the depression corpus designed in this paper, the MFCC features of speech are extracted and input into the proposed attention residual network and several different machine learning algorithms. The overall performance of the attention residual network is better than that of the machine. The learning algorithm can further illustrate that the algorithm in this paper has better generalization ability and robustness.

The influence of gender on results is shown in (Fig. 16). In the gender case, compared with normal individuals, patients too much associate negative emotions with self for processing, leading to self-negative prejudice. They negatively encode individual experiences, tending to summarize the negative experience and avoid the positive factors in the situation, which will distort the perception and bias the information. It can be seen from the figure that the accuracy rates are all above 60%, and the highest accuracy rate can reach 80%. Besides, the recognition accuracy rate of females is higher than that of males regardless of self or others, which is in line with previous research results. Besides, the area under curve (AUC) value of the classification results (Fig. 17). Evaluate the model and fine-tune the parameters with the aid of AUC. Draw a receiver operating characteristic (ROC) curve to characterize the specificity and sensitivity of the proposed feature.

When AUC is larger, it indicates that the prediction accuracy of the classification model is higher, and the classification effect is better. In general, the AUC value is between 0.80 and 0.83, indicating that the proposed model has a high accuracy rate of classification prediction.

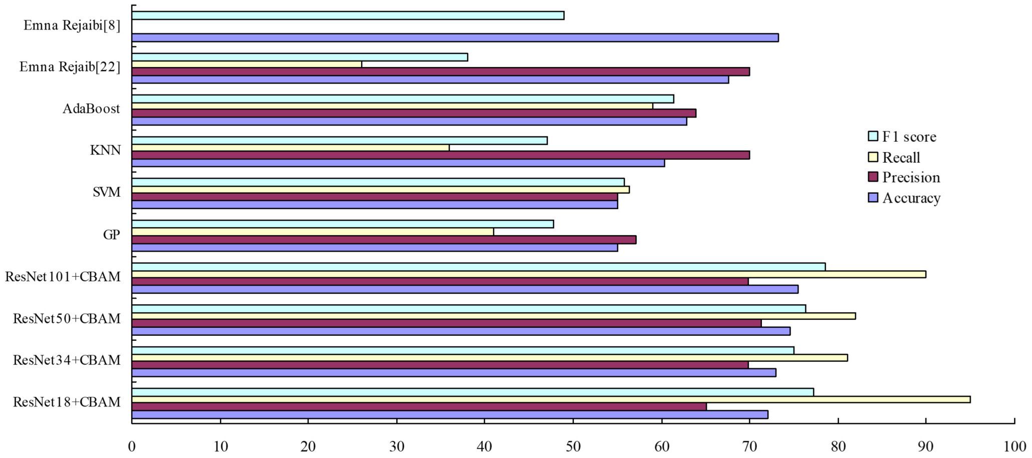

**Fig. 15. Comparison and analysis of the results of input MFCC under different algorithms (%).**

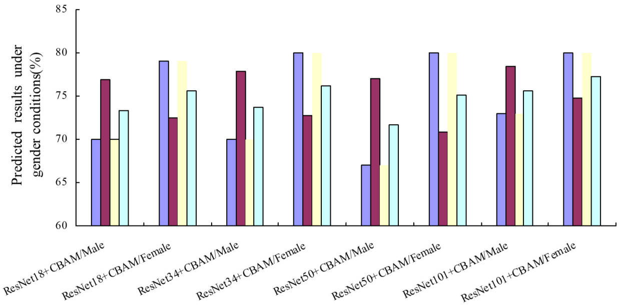

**Fig. 16. The predicted results under the gender of different models.**

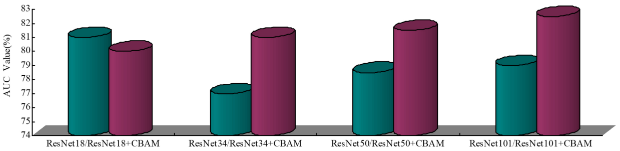

**Fig. 17. The AUC value of the classification results.**

### 5.3 Performances on depression severity levels prediction task

The Fig. 18 shows the confusion matrix of multi-class performance of gender in the depression recognition task under the premise of different emotional valences. From the results of the confusion matrix, it can be seen from the confusion matrix that under negative emotions, the recognition accuracy rate of both male and female subjects is higher than that under other emotions. This also reveals that depressed individuals have a negative bias in the processing of self-reference information, which has its cognitive roots, that is, individuals have formed a negative self-schema in their minds. Besides, as can be seen from Fig. 18 in the recognition rate of moderate, the model performs well. The proposed method has the highest recognition rate of mild, while has a low recognition rate of severe. Similar to the result of neutrality recognition, the mild and severe recognition rate of each algorithm is similar. In conclusion, the proposed method has satisfactory results in terms of minimal and mild recognition rate. Furthermore, for minimal and mild, because of the large number of samples and obvious characteristics of them, they are in a relatively high recognition rate in all kinds of conditions. In addition, in the experimental results, we can see the most emotions of moderate are misclassified as mild training samples is the largest, which results in a majority of samples of moderate shifting to the mild samples. Looking at the confusion matrix as a whole, the method proposed in this paper can classify subjects with depression propensity and healthy subjects.

## 6. Conclusions

This paper introduces a novel method, based on ResNet and CBAM model for depression classification with limited amount of training dataset. In the present work, speech based ResNet model is proposed to make use of frequency invariant amplitude response and progressive resolution of the signals. To utilize gradient information efficiently, and fifine-tune the weights with limited training. What's more, the idea of attention mechanism to design an attention residual model to classify speech depression. Input the MFCC features into the attention residual network of different depths to obtain more speech features, capture the correlation of the spatial position, and use the residual block to overcome the problem of gradient drop when the model depth becomes deeper, and to improve the accuracy of feature extraction. What's more, in multi-classification tasks, good recognition results were also obtained. The proposed method can be applied clinically as it has several advantages and shows promising performance even with limited training dataset. The advantages of the proposed framework are concluded as follows: (i) higher performance in detection of depression of speech, (ii) provides better performance for

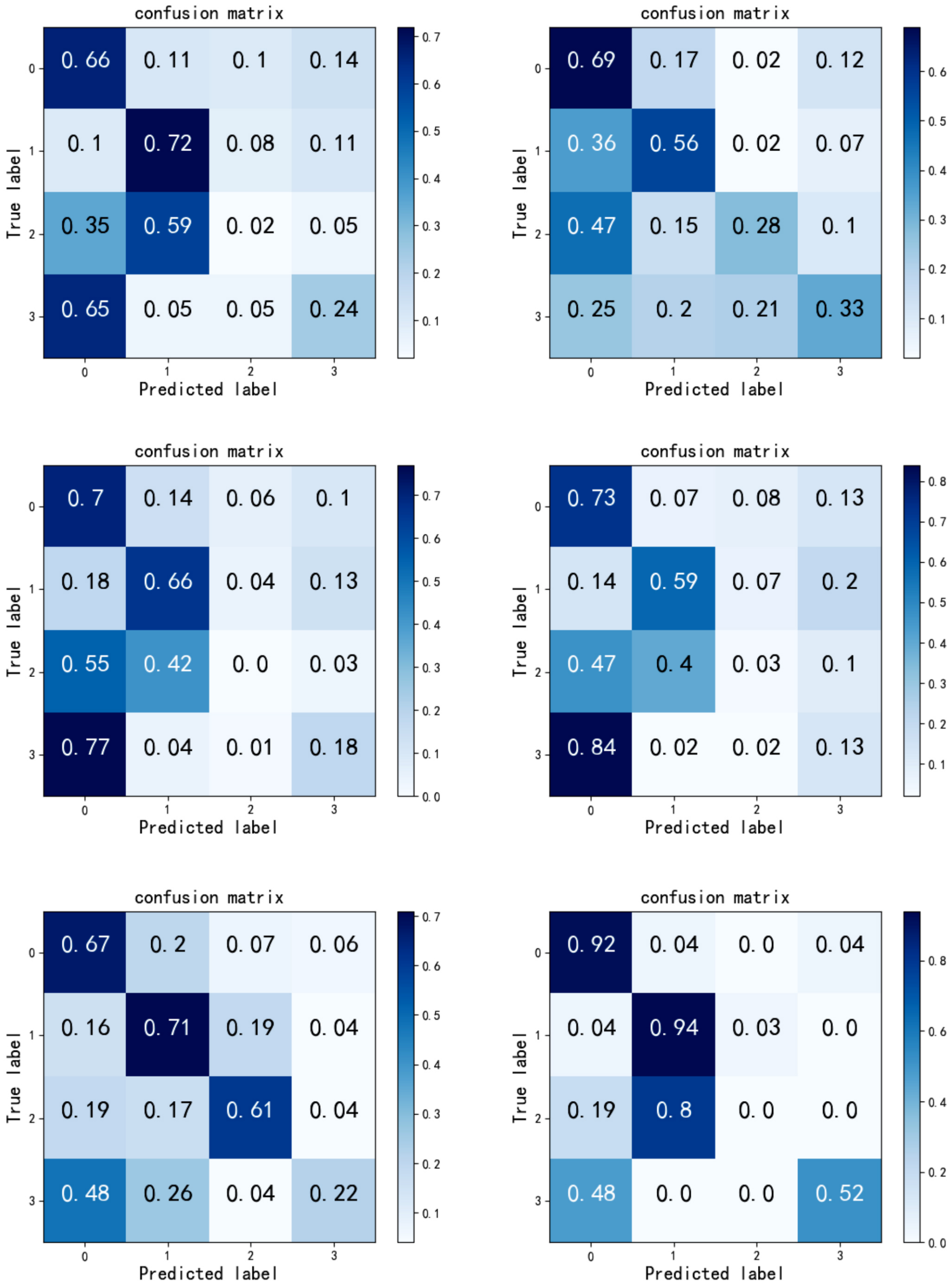

**Fig. 18. The left column is the multi-classification results of men under positive, neutral, and negative emotions; the right column is the multi-class results of women under positive, neutral, and negative emotions.**

real-time acquired dataset, and (iii) the diversified data input forms highlight the effectiveness of this method in the task of identifying depression. Due to the encouraging results of the work, this research is deserving of continuing. Given the research deficiencies, in the future, the existing corpus will be augmented with data to expand the size of the corpus and examine the influence of gender on the recognition of depression under different tasks.

## 7. Author contributions

XL and DS conceived and designed the experiments; DS performed the experiments; YL and JY contributed the data collection.

## 8. Ethics approval and consent to participate

We were obtained with the informed consent of all participants.

## 9. Acknowledgment

Thanks to all the peer reviewers for their opinions and suggestions.

## 10. Funding

This work was supported by the National Science Foundation of China (NSFC) under grant No. 31860285 and No. 31660281.

## 11. Conflict of interest

The authors declare no conflict of interest.

## 12. References

[1] Pan, Wei,Wang, Jingying,Liu, Tianli, *et al*. Depression recognition based on speech analysi. Kexue Tongbao/Chinese Science Bulletin. 2018; 63: 2081–2092.

[2] Huang Y, Wang Y, Wang H, Liu Z, Yu X, Yan J, *et al*. Prevalence of mental disorders in China: a cross-sectional epidemiological study. the Lancet. Psychiatry. 2019; 6: 211–224.

[3] Blais M, Baer L. Understanding Rating Scales and Assessment Instruments. Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health. 2010; 1–6.

[4] Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Communication. 2015; 71: 10–49.

[5] Cummins N, Epps J, Ambikairajah E. Spectro-temporal analysis of speech affected by depression and psychomotor retardation. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013; 7542–7546.

[6] Nicholas C, Jyoti J, Abhinav D, Vidhyasaharan S, Roland G, Julien E. Diagnosis of depression by behavioral signals: a multimodal approach. Computer Science. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. 2013; 11–20.

[7] Xie Z, Zinszer BD, Riggs M, Beevers CG, Chandrasekaran B. Impact of depression on speech perception in noise. PLoS ONE. 2019; 14: e0220928.

[8] Alice O, Daoud K, Kamil B, Emna R, Romain A, Abdenour H. Towards robust deep neural networks for affect and depression recognition. ICPR CAIHA 2020 workshop. 2020; arXiv. (in press)

[9] Salekin A, Eberle JW, Glenn JJ, Teachman BA, Stankovic JA. A Weakly Supervised Learning Framework for Detecting Social Anxiety and Depression. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018; 2: 81.

[10] Cai H, Qu Z, Li Z, Zhang Y, Hu X, Hu B. Feature-level fusion approaches based on multimodal EEG data for depression recognition. Information Fusion. 2020; 59: 127–138.

[11] Zhengyu L, Dongyu W, Lan Z, Bin H. A Novel Decision Tree for Depression Recognition in Speech. 2020; arXiv: 2002.12759v1.

[12] Eyben F, Weninger F, Gross F, Schuller B. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (MM '13). Association for Computing Machinery: New York, NY, USA. 2013; 835–838.

[13] He L, Cao C. Automated depression analysis using convolutional neural networks from speech. Journal of Biomedical Informatics. 2018; 83: 103–111.

[14] Ma X, Yang H, Chen Q, Huang D, Wang Y. DepAudioNet: An Efficient Deep Model for Audio-based Depression Classification. The 6th International Workshop. IEEE Press: Washington D.C., USA. 2016; 35–42.

[15] Chlasta K, Wołk K, Krejtz I. Automated speech-based screening of depression using deep convolutional neural networks. Procedia Computer Science. 2019; 164: 618–628.

[16] Chao L, Tao J, Yang M, Li Y. Multi task sequence learning for depression scale prediction from video. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). 2015; 526–531.

[17] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[18] Woo S, Park J, Lee J, Kweon IS. CBAM: Convolutional Block Attention Module. Computer Vision – ECCV 2018. 2018; 86: 3–19.

[19] Rao Q, Yu B, He K, Feng B. Regularization and Iterative Initialization of Softmax for Fast Training of Convolutional Neural Networks. 2019 International Joint Conference on Neural Networks (IJCNN). 2019; 1–8.

[20] Nixon N, Guo B, Garland A, Kaylor-Hughes C, Nixon E, Morriss R. The bi-factor structure of the 17-item Hamilton Depression Rating Scale in persistent major depression; dimensional measurement of outcome. PLoS ONE. 2020; 15: e0241370.

[21] Westhoff-Bleck M, Winter L, Aguirre Davila L, Herrmann-Lingen C, Treptau J, Bauersachs J, *et al*. Diagnostic evaluation of the hospital depression scale (HADS) and the Beck depression inventory II (BDI-II) in adults with congenital heart disease using a structured clinical interview: Impact of depression severity. European Journal of Preventive Cardiology. 2020; 27: 381–390.

[22] Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A. MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control. 2019; 71: 103107.

[23] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, *et al*. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine. 2001; 18: 32–80.

[24] Vlasenko B, Sagha H, Cummins N, Schuller B. Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. Interspeech 2017: Stockholm, Sweden. 2017; 3266–3270.

**Abbreviations:** SER, Self-reference effect; MFCC, Mel-frequency cepstrum coefficient; DCNN, Dynamic Convolution Neural Network; CNN, Convolutional Neural Network; AVEC, Audio/Visual Emotion Challenge and Workshop; LSTM, Long and short-term memory network; CBAM, Convolutional Block Attention Module; HAMD, Hamilton Depression Scale; BDI-II, The Beck Depression Inventory; SGD, Stochastic Gradient Descent; AUC, Area Under Curve; ROC, Receiver Operating Characteristic.

**Keywords:** Automatic recognition of depression; Residual neural network; Attention mechanism; MFCC

**Send correspondence to:** Xiaoyong Lu, Internet Education Data Learning Analysis Technology National and Local Joint Engineering Laboratory, 730070 Lanzhou, Gansu, China, Key Laboratory of Behavior and Mental Health of Gansu Province, 730070 Lanzhou, Gansu, China, E-mail: luxy@nwnu.edu.cn