Mobile Networks and Applications
https://doi.org/10.1007/s11036-022-02086-3

Deep learning for Depression Recognition from Speech

Han Tian[1,2] · Zhang Zhu[1,2] · Xu Jing[2]

Accepted: 14 December 2022

Abstract

In recent years, depression has been widely concerned, which makes people depressed, even suicidal, causing serious adverse consequences. In this paper, a multi information joint decision algorithm model is established by means of emotion recognition. The model is used to analyze the representative data of the subjects, and to assist in diagnosis of whether the subjects have depression. The main work is as follows: On the basis of exploring the speech characteristics of people with depressive disorder, this paper conducts an in-depth study of speech assisted depression diagnosis based on the speech data in the DAIC-WOZ dataset. First, the speech information is preprocessed, including speech signal pre emphasis, framing windowing, endpoint detection, noise reduction, etc. Secondly, OpenSmile is used to extract the features of speech signals, and the speech features that the features can reflect are studied and analyzed in depth. Then feature selection is carried out based on the influence of speech features and feature combination on depression diagnosis. Then, principal component analysis is used to reduce the dimension of data features. Finally, the convolutional neural network is used to modeling, testing and result analysis showed that the voice based diagnosis of depression was as high as 87%.

Keywords Depression recognition · CNN · speech · PCA

1 Introduction

Depression is a psychological disorder characterized by long duration and repeated attacks, with depression as the main clinical manifestation [1]. In recent years, the pressure from all aspects of life has led to a continuous rise in the incidence rate of depression. Middle aged and young people become the high risk population [2]. According to the statistics of the World Health Organization, there are 350 million patients with depression in the world. China is one of the countries with the heaviest burden of depression in the world. The number of depressive patients in China has reached 90 million, and about 280000 people commit suicide every year. In China, 50% to 70% of suicides and attempted suicides are suffering from depression, but the medical treatment rate of depression in China is lower than 10% [3].

At present, the diagnosis of depressive disorder is made through clinical consultation. The doctor can understand the psychological state of the visitors by asking them about the recent situation, or by making some authoritative questionnaires, and then the doctor can make judgments based on experience. People with a family history of depression are more likely to be diagnosed [4]. Different from other physiological diseases, depressive disorder cannot be clearly explained by some exact physical indicators. On the contrary, observation and questionnaire survey are based on the professional level of clinicians, and the diagnosis time is relatively long. At present, the number of clinicians is insufficient, and a certain proportion of registered doctors have not received vocational training [5]. In addition, the national mental health service institutions are at a stage of serious supply shortage. WHO estimates

Han Tian and Xu Jing are contributed equally to this work.

✉ Zhang Zhu
zhuzhang.zz@foxmail.com

Han Tian
hantian@hrbust.edu.cn

Xu Jing
Xujing@hrbust.edu.cn

[1] Department of Artificial Intelligence, Jinhua Advanced Research Institute, No.99 Huanchengnan Road, Jinhua, 321012, Zhejiang, China

[2] School of Measurement and Communication Engineering, Harbin University of Science and Technology, No.52 Xuefu Road, Harbin, 150080, Heilongjiang, China

Published online: 26 January 2023

that less than 50 per cent of patients worldwide have access to effective treatment. Due to the particularity of the diagnosis method of depression, the diagnosis process will be affected by the subjective experience of doctors on the one hand, and the patient's own conditions on the other hand, which may lead to unstable and accurate diagnosis results. At this stage, the diagnostic accuracy of depressive disorder is less than 50% [6]. In addition, the social prejudice against psychiatric patients also causes patients to be ashamed to seek treatment and intentionally conceal their illness. This phenomenon causes doctors to be unable to timely and accurately assess the psychological state of patients and miss the best time for diagnosis. Therefore, based on the existing clinical diagnosis methods, finding an objective and efficient diagnosis method of depression as an auxiliary method of clinical diagnosis has become a hot research direction of early detection of depression.

Depression is the most easily observed feature of patients with depressive disorder, so we can identify depressive disorder by identifying emotional state. According to the theory put forward by the psychologist Mehrabian, 55% of the emotional state can be revealed from facial expressions, 38% can be revealed from sounds, and the remaining 7% comes from people's language content [7]. In recent years, many researchers have found that the voice of patients with depression will be affected by their long-term unstable emotions. The clinical diagnosis is usually based on the changes of the patient's voice. For example, when speaking, the speaking speed will be slower than that of normal people, and there is no obvious emotional fluctuation [8]. In terms of language expression, they have obvious negative tendencies, such as "that painful day will not end before I die" and "I feel terrible!" Such pessimistic words and sentences frequently appear [9].

In this paper, a depression detection model is established by means of emotion recognition. The model is used to analyze the representative data of the subjects, and to assist in diagnosis of whether the subjects suffer from depression. By studying the judgment model of depression disorder based on deep learning and the judgment model of depression disorder based on speech information, the system can be used as an auxiliary means to diagnose depression disorder, which is efficient, convenient and practical.

2 Related works

In recent years, there has been an increasing amount of literature on assistant diagnosis of depression. According to clinical diagnosis, there are differences in speech between normal people and patients with depression. According to the statistics of clinical manifestations, the speaking habits of patients with depression will be more depressed, and the pauses between words will be more and longer. These clinical symptoms can effectively distinguish between normal people and depressed people.

In 2012, Kuan Ee Brian Ooi et al. proposed a judgment model of depressive disorder for young people, and specifically studied the role of four language features in the recognition task of depressive disorder [10]. The team's research was conducted on a data set of 97 boys and 94 girls, and the accuracy of the experimental results reached 73%. In 2015, Yasin Ozkanca et al. used neural networks to study the impact of prosodic features on the judgment of depressive disorder [11]. In 2017, Kl á raVicsi et al. focused on different paradigms and compared the speech obtained from interviews with that obtained from reading. The accuracy rates of the two paradigms were 83% and 86% respectively. The experiment found that prosodic characteristics did play a significant role in the judgment of depression [12]. In 2017, Wang Tianyang proposed that prosodic features and spectral features play an important role in the judgment of depressive disorder. And the accuracy of the model trained by the two types of features in the male and female databases reached 70% and 75% [13]. In 2018, Meysam Asgariden et al. proposed a new algorithm to extract speech features. The data set used by the researchers was 148 subjects including 77 normal people and 71 patients with depression. Compared with the previous research, the accuracy of this algorithm is improved by 9% [14]. In 2020, Yuan Jia and others discussed and analyzed the vocal music characteristics of patients with depressive disorder. Using the data set of 82 patients with depressive disorder and 57 healthy people, they obtained frequency interference, amplitude interference and noise harmonic ratio that can distinguish patients with depressive disorder from normal people [15]. Speech based recognition of depression is gradually combined with machine learning. Because of the complexity of the representation of depressive disorder, a single feature cannot completely describe it. Many researchers have established and combined a variety of speech features to build a more powerful model of depression recognition. In 2018, Lang He et al. proposed a speech based model combining manual extraction and deep learning methods, and proposed a method to improve data in order to solve the problem of small samples [16]. The above methods are applied to AVEC2013 and AVEC2014 datasets, and the results show that the method has good effectiveness and robustness in the judgment of depressive disorder. Alice othmani et al. proposed a convolutional neural network based on MFCC and spectrogram to identify depression. Finally, the features proposed by CNN were incorporated into the full connection layer [17]. This result is superior to other methods on DAIC-WOZ dataset.

In summary, the depression recognition model can be be divided into two categories. Traditional machine learning methods focus more on improving model accuracy through feature optimization, while deep learning methods focus more on model structure and super parameter optimization to improve accuracy, with less research compatible with both. In this paper, based on the in-depth study of speech features, convolutional neural network is used to train and optimize the model, so that the accuracy of speech based depression diagnosis is improved

3 Depression and data set

The assistant diagnosis system of depression presented in this paper is based on the information of voice. The data is all from the video data of the subjects. The video is divided into three modes to obtain the original voice information, facial expression information and text information. The original information cannot be directly processed by the computer system, so it is necessary to pre-process the original information. After pre-processing, the emotional features representing each mode are extracted. Then the useful emotional features are input into the classifier for judgment, and three classification results are obtained. Because the decision level fusion mode is selected in this paper, the last step of this system is to input the three classification results into the decision fusion level to get the final decision result of depression. The main process of depression judgment based on voice, facial expression and text fusion is shown in Figs. 1 and 2.

In The experimental data in this paper is from the DAIC-WOZ dataset, which contains audio and three-dimensional facial data of subjects with depressive disorder and non depressive disorder.

The DAIC-WOZ data set conducts interviews through the situational dialogue between the virtual interviewer Ellie and the interviewee, focusing on the interviewee's mood, life and status. This will stimulate the subjects to generate real emotions in specific situations, and show them in real time through facial expressions and voice. Each sample data in the data set includes the subject's voice audio, AU facial features of facial expressions, 3D facial features and eye features, as well as the text transformed from the interview content.

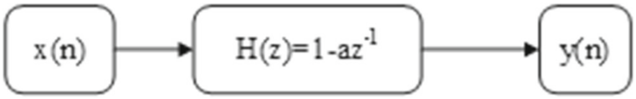


Fig. 2 Pre-emphasis diagram

The interview also conducted a PHQ-8 questionnaire survey on the subjects, which was marked according to the value of the PHQ-8 questionnaire. There are eight items in PHQ-8, which are designed to understand the duration of various sufferings suffered by respondents in the past two weeks. The days available are "0-1 days", "2-6 days", "7-11 days" and "12-14 days". The scores corresponding to the above four options are 0, 1, 2 and 3 [18]. The score obtained from the questionnaire ranges from 0 to 24, which is consistent with the clinical diagnosis. It is concluded that the total score of PHQ-8 is greater than or equal to 10, and it is considered that there are depressive symptoms; If the score is less than 10, it means that they have negative emotions, but have no depressive disorder. PHQ8 is suitable for the diagnosis of depression in the general population because of its strong sensitivity to depression. In the data set, the scores of the PHQ-8 questionnaire were given, and binary tags were also made to mark the depressed individuals as 1, and the non depressed individuals as 0.

In the DAIC-WOZ data set, the distribution of depressive samples and non depressive samples is shown in Table 1.

On one hand, there are few public data sets of depressive disorder including voice, facial expression and text at the same time; On the other hand, most of the current data sets are performance data sets, while the data of DAIC-WOZ data set are from real interviews, which is more realistic than the data obtained by performance. Based on a variety of factors, this paper finally chooses the DAIC-WOZ dataset for subsequent experimental research. The final test results of this paper are obtained on 47 test sets

Fig. 1 Flowchart for front-end processing of voice signal

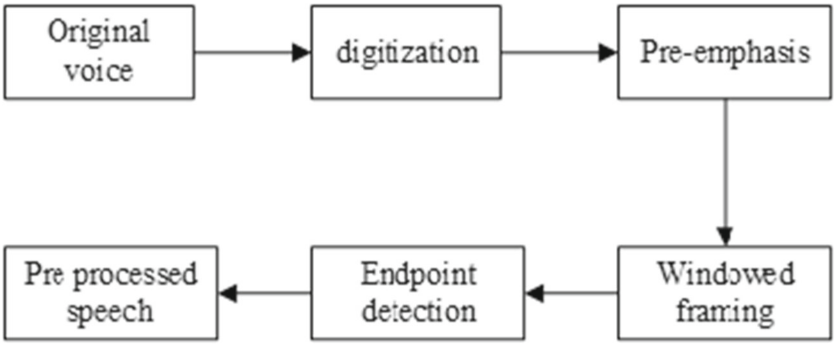


Table 1 Distribution of DAIC-WOZ Dataset

Data set partition	Number of depression	Number of non depression
Training set	30	77
Validation set	12	23
Testing set	14	33

4 Speech characteristics of depressed people

Alan Beck summarized the symptoms of depression as follows: emotional symptoms, cognitive performance, dynamic performance, somatic symptoms, delusions, hallucinations, etc. and the above symptoms are more or less reflected in the voice. Through clinical observation and statistics, the language performance of patients with depressive disorder can be summarized as: decreased autonomy and language ability; Speak slowly and pause frequently; The voice is small and weak; The tone of voice is monotonous; When you speak, you sigh and even cry. However, these characteristics generally do not exist in every individual at the same time, but appear in different stages of different people in a certain proportion. In order to find effective speech indicators, this paper makes a statistical comparison between the clinical language characteristics of depressed people and normal people, and makes Table 2.

The data is from Depression and tabulated after sorting

As the number of survey samples in Tables 1, 2 and 3 is large, it can be considered that the statistical results of the above language characteristics are highly reliable. It can be seen from Tables 1-3 that among the depressed people, the language features of slow speech, low mood, indecision and crying are very high, especially the first three features reach more than 60%. In the normal population, the proportion of these phonetic features is less than a quarter. The proportion of the above symptoms between the depressed population and the normal population is greatly different, which provides a direction for distinguishing between the depressed population and the normal population, and lays a realistic foundation for using voice signals to identify depressive disorders.

Table 2 Statistical table of clinical speech characteristics in patients with depression

Phonetic characteristics	Non depressive	Depressive
Speak slowly	25%	67%
Be down in spirits	16%	87%
Irresolute and hesitant	18%	64%
Crying	3%	23%

5 Speech pre-processing

Because sound is a continuous signal, and there is loss in the transmission process of sound and it is affected by noise, it cannot be directly processed by the computer. Therefore, it is necessary to conduct front-end processing on the voice signal, and then further study. The front-end processing flow of voice signal is shown in Fig. 1.

5.1 Digitization

The purpose of speech signal digitization is to discretize analog signal. Digitization includes five steps: amplification and gain control, pre filtering, sampling, A/D conversion and coding.

The main function of pre filtering is to filter the signal components whose intermediate frequency is higher than the sampling frequency of the input signal. Its essence is a band-pass filter. The reference human sound frequency range is 300Hz to 3400Hz, so the upper and lower cut-off frequencies are 3400Hz and 300Hz respectively.

Sampling quantization can realize signal discretization. The quantization accuracy is 16 bit.

Coding is the last step of speech signal digitization. In fact, the analog audio signal has been converted into digital signal after sampling and quantization, but it is encoded to reduce the amount of data.

5.2 Pre-emphasis

When the voice signal is at high frequency, the noise is relatively large; In low frequency, the noise is relatively small, which results in the signal to noise ratio of the high frequency part is smaller than that of the low frequency part. In the process of voice signal transmission, it is difficult to transmit the high-frequency part, which is easy to cause information loss. Therefore, weighting the signal before transmission can effectively solve this problem and facilitate further processing of voice signals.

The pre-emphasis is realized by a first-order filter. The pre-emphasis is to improve the high frequency, so it is a high pass filter. The pre-emphasis process is shown in Fig. 2.

5.3 Windowed framing

If the voice signal is divided into short frames, each frame can be regarded as a steady state signal, which can be processed using the steady state signal processing method. The framing is realized by adding windows. The object of each signal processing is the signal in the window. After processing, the next section is taken for processing and

Table 3 Example of a lengthy table which is set to full textwidth

Feature classification	Summary	Examples
Acoustic features	Subjective evaluation index used to determine audio clarity	Formant
Prosodic features	Describe the characteristics of speech at multiple levels and represent the non linguistic characteristics of speech	Energy, Duration
Spectral features	It is used to describe the relationship between the vibration of human vocal organs and the change of vocal tract shape	MFCC, LPCC

analysis. The function expression after windowing is:

$$s_w(n) = s(n) * w(n) \tag{1}$$

Where $s(n)$ is the nth sampling point in the selected window, and $w(n)$ is the corresponding weight. Different windowing methods are reflected in the value of $w(n)$.

Figure 3 shows the shapes of different window functions. Figure 4 shows the Fourier transform of different window functions. It can be seen from Fig. 4 that the main lobe width of Hamming window is larger than that of rectangular window. The rectangular window will cause spectrum leakage, so the rectangular window is not selected. The low pass effect of Hamming window is good, so Hamming window is selected for simulation in this paper.

The Hamming window function formula is as follows,

$$w(n) = \begin{cases} 0.54 - 0.46cos[2\pi n(N-1)] & 0 \le n \le n-1 \\ 0 & else \end{cases} \tag{2}$$

where N is the window length.

5.4 Endpoint detection

The collected voice signal does not contain useful information at all times. The purpose of endpoint detection is to locate the voice part in the segment, that is, to find the start time and end time of the voice, filter the noise part, filter the mute part, and find the content of a segment that really contains effective information.

The method of endpoint detection is short-time energy double threshold method, which takes short-time energy of each frame as the judgment criterion. This method needs to set two threshold values, and determine the sound segment by comparing the threshold value and short-term energy. When the short-time energy ratio is higher than the high threshold value, the frame is judged as having sound segment. The detection process is as follows: first, use the high threshold value to determine the approximate range of the sound segment; Then look forward for the first frame whose energy is higher than the high threshold value, take this as the starting point, and look backward until the first frame whose short-time energy is lower than the low threshold value appears, and then the position of the first segment with sound segment is determined. Next, repeat the above process to locate other sound segments.

6 Feature extraction

The extraction of emotional features is the data basis for the use of voice emotion analysis and research on

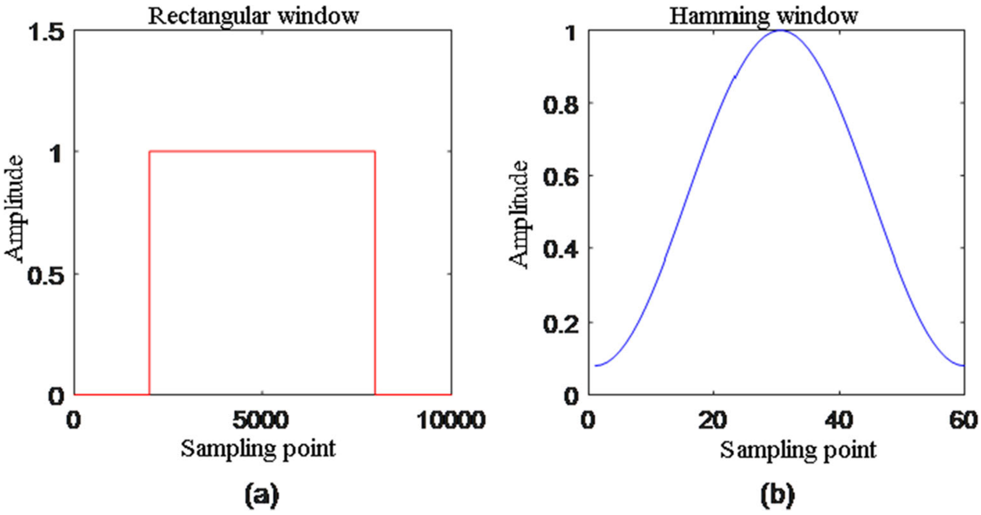


Fig. 3 Diagram of different window functions

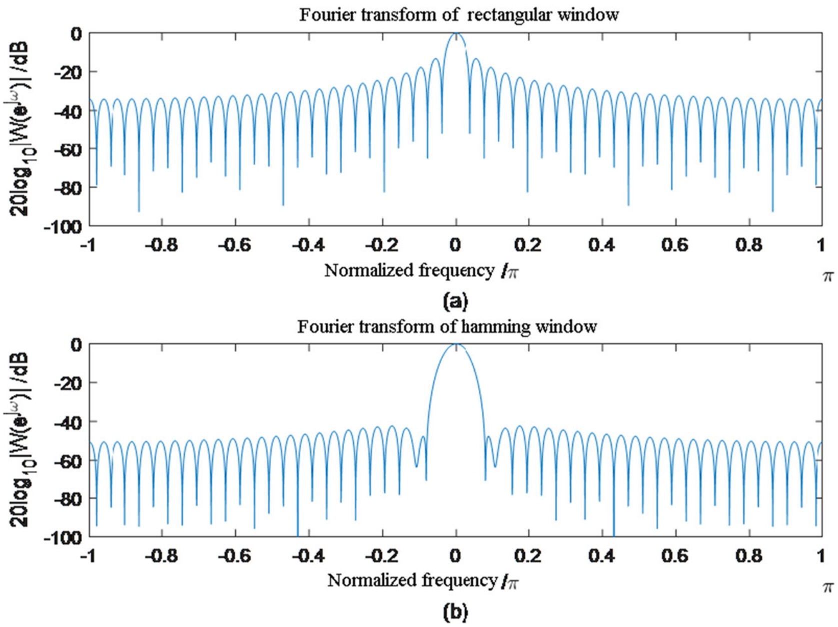


Fig. 4 Fourier transform diagram of different window functions

depression. The quality of the extracted voice emotional features is closely related to the success or failure of the depression judgment. The current research has not been able to determine which voice features can best reflect the specificity of depressed people. Therefore, it is necessary to extract different features containing emotional information as much as possible, and then use certain feature selection methods to select the features that are effective for depression judgment.

At present, speech features used in emotion recognition can be divided into three categories, namely, voice quality features, prosodic features and spectrum based correlation features. Table 3 summarizes the meanings and common features of the three types of voice features.

OpenSMILE is a highly encapsulated speech feature extraction tool: its bottom layer is written in C++language and supports running on a variety of systems. It can not only extract basic speech features such as frame energy, fundamental frequency, short-term jitter parameters, but also extract voice emotional feature parameters such as Mel frequency cepstrum coefficient. The above speech features are called low-level description factors. After a series of statistical operations such as mean, variance and regression coefficient, low-level description factors are transformed into feature vectors of certain dimensions, which can also represent the characteristics of speech. The process of using openSMILE toolbox to extract speech features is shown in Fig. 5.

Use the openSMILE toolbox to input a voice file in wav format, select the 'emobase2010. conf' command to process the data, and extract the voice emotional features. From this, 68 low-level description factors can be obtained. Finally, 1582 dimensional voice emotional features can be obtained through statistical operation, and the above features can be output as a file in csv format. Table 4 briefly introduces the extracted speech features.

Prosodic features include features that describe information such as voice intonation and energy. Prosodic features do not affect people's understanding and recognition of the content information, but people's most intuitive voice emotion perception. Therefore, it is believed that prosodic features play an important role in emotion discrimination, and are commonly used features in the field of voice emotion classification. The voice information conveyed by prosodic features, such as slow speech speed, long pauses, is corresponding to some clinical manifestations of patients

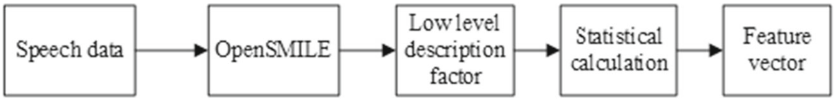


Fig. 5 Flowchart for extracting voice emotional features with openSMILE

Table 4 Summary of speech emotional features

Feature classification	Dimension	Characteristics of sound
F0	45	The pitch and the sound quality
F0 Envelope	42	Variation of fundamental frequency amplitude
Jitter	42	Irregular changes in sound quality in a short time
Sound intensity	55	Intensity and loudness of voice
Shimmer	43	Irregular variation of amplitude
Subband energy	20	Frequency distribution characteristics of energy
Formant	9	Timbre of sound
Zero crossing rate	5	Properties of roughly estimating sine wave spectrum
TEO	16	Nonlinear characteristics of signals
MFCC	681	Simulate human auditory characteristics
LPCC	14	Pulse sequence generated by sound source and transfer function parameters of sound track resonance system

with depression. Therefore, we can use prosodic features to identify depression. Here are some common prosodic features.

Speaking rate: describes the speed of speaking, which can be expressed by calculating the vocabulary in unit time. The definition of speech speed is shown in Formula (3-3):

$$t_{avg} = 1/m \sum_{i=1}^{m} t_i \tag{3}$$

Energy: The energy of voice signal changes with time. Energy is related to emotions. For example, people usually have low energy when they are depressed. Short time average energy E of voice signal is:

$$E = \sum_{m=-\infty}^{+\infty} [x(n)w(n-m)]^2 \tag{4}$$

Figure 6 shows the short-term energy diagram of a certain voice signal in the data set.

Zero Crossing Rate (ZCR): refers to the sum of the number of times the signal value changes from positive to negative or from negative to positive in each frame of voice signal. Zero crossing rate can be expressed as:

$$Z_n = \sum_{m=-\infty}^{+\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|w(n-m) \tag{5}$$

where

$$sgn = [x(n)] = \begin{cases} 1 & x(n) \geqslant 0 \\ -1 & x(n) < 0 \end{cases} \tag{6}$$

Spectral characteristics are used to describe the relationship between the vibration of human vocal organs and the change of vocal tract shape. Here are some common spectrum features.

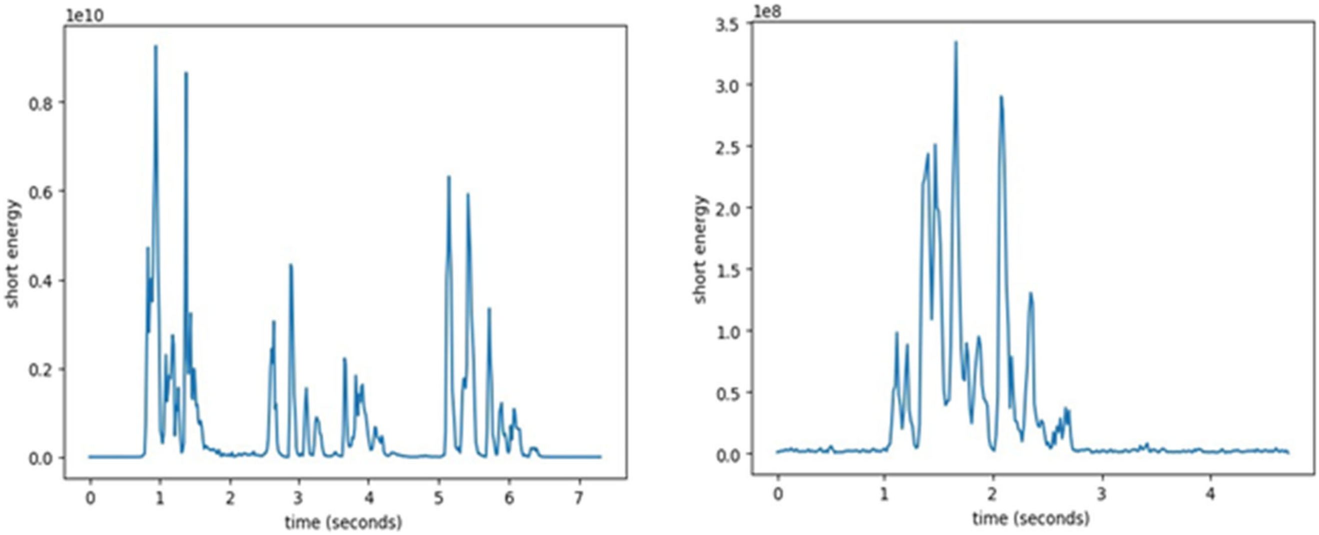


Fig. 6 Short-time signal energy map

Mel Cepstrum Coefficient: Mel Cepstrum Coefficient is from the perspective of simulating the human auditory system. Because the human auditory system is a nonlinear system for speech perception, it needs to be filtered, logarithmic and other operations to be as close as possible to the human ear's speech processing process. MFCC feature extraction flowchart is shown in Fig. 7, and Mel frequency is caculated by Eq. 7.

$$f_{mel} = 2595 \times \log\left(1 + \frac{f}{700}\right) \tag{7}$$

MFCC is calculated by Eq. 8.

$$C_n = \sum_{k=1}^{M} logx(k)cos\left[\frac{\pi(k-0.5)n}{M}\right] \tag{8}$$

Linear Predictive Coefficients (LPC): refers to the coefficients for linear prediction of speech signal values, that is, the weighted linear combination of several past values is used to fit the current value, and the weight value is the linear prediction coefficient.

7 Feature selection

There are many types of features extracted, and which features will have a positive impact on the experimental results of depression judgment needs to be determined through feature selection.

In actual research, there are two shortcomings in the discussion of identifying the effective features of depression:

(1) The types, dimensions and effectiveness of features obtained by different feature extraction methods are quite different. If these characteristics are directly input into the classifier training model, the training results may be far from the optimal solution. The reason for this result may be: on the one hand, different training tasks have different goals, so the corresponding voice features that can effectively reflect the essence are also very different. Some voice features may play an important role in a certain task, but they have no effect on the task of determining depression in this paper, or even have the opposite effect, so these features should be eliminated; On the other hand, in this training task, some features have a positive and positive effect on the training model when they appear alone, but when they participate in training with other features at the same time, the training result is not improved compared with that of a single feature, so such features are mutually redundant features, and only one of them can be retained when finally adopted.

(2) The effective features of depression judgment obtained from existing research indicate the feature categories, but each feature specifically contains many dimensions, some of which are valuable, but some of which not only have no positive impact on the judgment results, but even have adverse effects. And large dimension data will also increase the computational burden for model training.

In this section, we will find effective feature combinations based on the above two shortcomings, and use the accuracy rate to quantify the judgment ability of each feature combination. For the first disadvantage, through comparative analysis of different feature combinations, redundant features and useless features are filtered out to screen out the feature types that meet the requirements; For the second disadvantage, this paper uses principal component analysis and introduces the concept of information entropy to reduce the dimension of data.

7.1 Comparative analysis of different types of features

In real life, men tend to be rational when facing problems. The change range of voice and intonation is small, while women are more likely to be emotional, which is expressed in language. Therefore, the gender differences in language between men and women are reflected. Therefore, this chapter trains models applicable to men and women respectively for speech features.

In this section, support vector machine, K nearest neighbor, and random forest are used to recognize 14 single

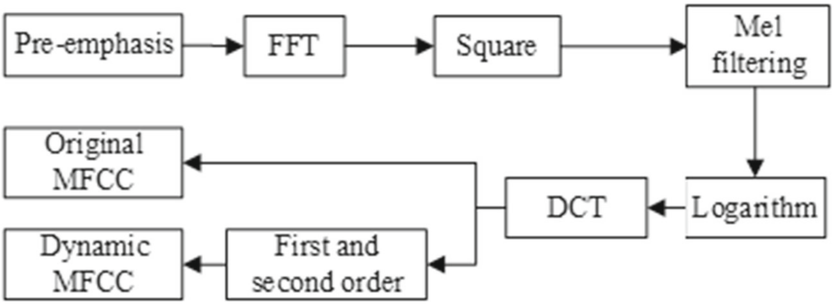


Fig. 7 MFCC Feature Extraction Flowchart

class features in the dataset, and the accuracy is the average of the three classifiers. The results are shown in Table 5.

It can be seen from the data in the table that the classification accuracy of most of the 14 types of features is 50% higher than that of random classification. This shows that voice features have a positive effect on the judgment of depression.

In the judgment of male depressive disorder, the characteristics of higher accuracy of judgment are sound intensity and jitter, which may be related to the more deep voice of men with depressive disorder. In the judgment of female depressive disorder, Mel cepstrum coefficient and LPC coefficient have a high recognition accuracy, which is significantly different from the speech features with a high accuracy rate of men. This shows that this paper is targeted at male and female training models, which can improve the accuracy rate.

Because researchers use different paradigms and emotional stimulation methods in the research process, the effective feature classes obtained are not necessarily consistent. In this paper, we need to find a combination of characteristics that can better distinguish whether or not we have depression in a relatively general sense.

7.2 Comparative analysis of feature combination

In this section, for men and women, two types of characteristics are combined to make a decision on depression. There are 91 combinations of characteristics. Here, seven groups of characteristics with the highest accuracy of decision are selected. The results after sorting out are shown in Tables 6 and 7.

Table 5 Classification Result of Full Feature Set for 14 Classes of Features

	Accuracy	
Features	Male	Female
Intensity	57.5%	56.7%
MFCC	54.8%	57.3%
Mel logarithmic power	54.4%	59.5%
LPC spectral power	54.0%	57.2%
LPC	53.8%	51.4%
F0 envelope	53.3%	52.4%
F0	53.1%	51.%
Jitter	57.2%	53.1%
JitterDDP	47.5%	51.6%
Shimmer	50.3%	52.1%
Subband engergy	45.9%	53.7%
TEO	47.6%	55.2%
Formant	47.1%	47.3%
ZCR	46.3%	47.2%

The decision accuracy rate in Tables 3, 4, 5 and Table 3-6 is higher than that of single feature decision in Tables 3-5, which indicates that after the combination of different features, information complementation will be formed between features to provide more information when identifying depression and improve the decision accuracy. However, whether the combination of the two features contains the most saturated information needs further verification.

In male data, among the feature combinations with high accuracy, LPC spectral frequency and jitter occurrence frequency are higher than LPC spectral frequency and jitter single category feature accuracy. Although LPC spectral frequency and jitter classification accuracy are high, and may carry more judgment information, in order to further improve the accuracy, it is still necessary to supplement information of other features.

In the female data, among the feature combinations with high decision accuracy, Mel's logarithmic power appears more frequently. However, compared with the classification accuracy of Mel's logarithmic power of 59.5% in a single category of features, the classification accuracy of Mel's logarithmic power is only slightly improved compared with the combined classification accuracy of other features, which indicates that other features have little complementary effect on it.

In addition, for the resonance peak with low decision accuracy rate in single feature, the accuracy rate is improved by about 10% after feature combination, which is not much different from the accuracy rate of other high accuracy feature combination. This result shows that when a certain feature appears alone, it does not carry much effective information for the judgment of depression, but when it is combined with other features, the combination effect of the two features is excellent. Therefore, in the process of searching for the best feature combination, we should not

Table 6 Top Seven Results of Classification Rate for Male Combination of Two Types of Features

Feature set		
Feature1	Feature2	Accuracy
MFCC	LPC spectral power	59.3%
MFCC	Jitter	59.0%
LPC spectral power	Jitter	58.9%
LPC spectral power	TEO	58.7%
Jitter	Intensity	58.4%
Formant	Intensity	58.2%
Formant	LPC spectral power	57.9%

Table 7 Top Seven Results of Classification Rate for Combination of Two Types of Features for Women

Feature set		
Feature1	Feature2	Accuracy
Mel logarithmic power	Formant	61.5%
Mel logarithmic power	ZCR	61.1%
Mel logarithmic power	Subband energy	60.6%
Mel logarithmic power	TEO	60.2%
Mel logarithmic power	Jitter	60.5%
TEO	LPCC	60.1%
TEO	Formant	58.4%

only focus on the features with high decision accuracy, but also try multiple combinations to find the feature set combination with the highest degree of information complementarity.

In order to further find the best feature combination, this paper selects seven feature combinations with the highest decision accuracy rate from 364 feature combinations based on three types of feature combinations. The results are shown in Tables 8 and 9.

In male data, the combination of the two types of features is the same, and the frequency of voice intensity and jitter is also high, which indicates that the information contained in these two types of features is not easy to be replaced by other features. After the combination of resonance peak, zero crossing rate and other features, the classification accuracy is significantly improved. Although these two types of features carry insufficient information, they play an important role in the judgment of depressive disorder. When selecting features, the important role of these features cannot be ignored.

The judgment accuracy rate of female data has improved slightly. After the combination of formant, zero crossing rate and other features with other features, the judgment accuracy rate has improved significantly, which is similar to the results of male data.

To sum up, the decision accuracy of the combination containing three types of features is only slightly improved compared with the combination of two types of features. It is speculated that the complementarity between the combination information containing three types of features is close to saturation, and adding feature classes cannot significantly improve the decision accuracy. Therefore, considering the model efficiency comprehensively, this paper finally selects the combination of three types of features for model training.

Table 8 Top Seven Results of Classification Rate for Male Combination of Three Classes of Features

Feature set			
Feature1	Feature2	Feature3	Accuracy
Intensity	LPCC	Formant	62.4%
Intensity	Jitter	ZCR	62.1%
LPCC	Jitter	Formant	62.1%
Intensity	Jitter	LPCC	61.8%
Intensity	Jitter	ZCR	61.6%
Intensity	LPCC	ZCR	61.4%
Intensity	Jitter	F0 envelope	61.1%

8 Feature dimension reduction based on PCA

The conclusion in Section 3.4.2 shows that proper combination of feature categories can effectively improve the accuracy of identifying depression, but due to the possible existence of useless and redundant features, the accuracy of classification is not very high. This section introduces the concept of information entropy according to the classification results of the full feature collection, screens feature combinations that carry more information, and then reduces the dimension of features in combination with Principal Component Analysis (PCA) as the input of the subsequent speech signal based depression recognition model.

Principal component analysis takes a long time and takes up a large amount of memory when processing high-dimensional data, so it is obviously unrealistic to analyze all feature combinations with principal component analysis. Therefore, the concept of information entropy is introduced here, and a set of feature combinations are left by information entropy screening, and then the dimension of the feature combination is reduced by principal component analysis. Hereinafter referred to as PCA algorithm. The specific steps of the PCA algorithm are shown in Fig. 8 and explained as follows:

(1) Convert the data into a feature matrix $A_{m\times n}$, where m represents the number of samples, n represents the number of features, and a_{ij} is the value of feature i of sample j.

(2) Use formula (3-10) to calculate the information entropy value $H(a_i)$ of each feature combination, compare the information entropy value of the feature combination selected in the previous section and leave the feature combination with the largest information entropy.

(3) PCA is used to reduce the dimension of the feature combination with the maximum information entropy to obtain the final dimension reduced data.

Table 9 Top Seven Results of Classification Rate for Three Classes of Feature Combination for Women

Feature set			
Feature1	Feature2	Feature3	Accuracy
TEO	Mel logarithmic power	LPCC	67.3%
Mel logarithmic power	TEO	ZCR	67.0%
Mel logarithmic power	TEO	Formant	66.9%
jitter	Intensity	Mel logarithmic power	66.8%
Mel logarithmic power	LPCC	ZCR	66.5%
LPCC	Mel logarithmic power	Formant	66.3%
Mel logarithmic power	F0 envelope	jitterDDP	66.1%

The amount of information can measure the uncertainty of events. The self information of an event has two meanings: it can represent the uncertainty before the event occurs, and it can also represent the amount of information brought by the event.

The self information can be obtained from Eq. 9:

$$I(a_i) = log\left(\frac{1}{p_i}\right) = -log(p_i) \tag{9}$$

If the signal sent by the source has n values, that is $X = a_1, a_2, a_3...a_n$, the signal set, the corresponding probability is $p_1,\ p_2,\ p_3,\ ...p_n$. And satisfy $\sum_{i=1}^{n} p_i\ =\ 1$. At this time, the average uncertainty of the source should be the statistical average of the uncertainty of a single symbol $-\log p_i$, called information entropy, recorded as:

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{10}$$

Information entropy can measure the amount of information. When information entropy is applied to features, it can be explained that if the information entropy of a feature is larger, it means that the information content of the data it contains is larger, which plays an important role in data classification. On the contrary, it means that the data has less role in classification.

According to the analysis of the meaning of information entropy, the greater the information entropy of the feature combination, the greater the amount of information contained in the original data, which belongs to the feature combination that should be retained; The lower the information entropy of the feature combination, the less data information the feature combination contains. When making feature selection, the feature combination is the combination that should be discarded.

In the feature combination containing three features obtained from the comparison experiment in the previous section, the feature combination with the maximum information entropy is left according to the value of information entropy. The specific feature combination is: male feature combination is Mel cepstrum coefficient, sound intensity, zero crossing rate; The combination of female characteristics is Mel logarithmic power, sound intensity and LPC coefficient.

Combined with its physical significance, the selected features are analyzed as follows:

(1) The sound intensity and zero crossing rate selected from the male data belong to prosodic features, while the Mel cepstrum frequency belongs to spectral features; The sound intensity selected from the female data belongs to prosodic characteristics, while the Mel logarithmic power and LPC belong to spectral characteristics. It can be seen that the feature combination selected from male and female data contains both prosodic features and spectral features, which indicates that when identifying depression, both prosodic features and spectral features are indispensable, and information needs to be obtained from different angles to achieve better results.

(2) The feature combination selected by male and female data includes voice intensity, which shows that

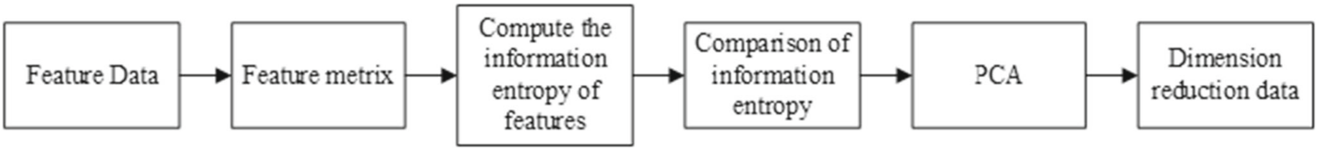

Feature Data
Feature metrix
Compute the information entropy of features
Comparison of information entropy
PCA
Dimension reduction data

Fig. 8 Flowchart of PCA algorithm

the voice intensity of people with depression is significantly different from that of normal people, indicating that voice intensity is an irreplaceable indicator to distinguish between depressed people and normal people.

(3) Mel cepstrum coefficient and Mel logarithmic power appear in the feature combination with good judgment effect for men and women respectively. After hearing the voice, people can more accurately identify the speaker and his mental state through the processing of the auditory organs. The high performance of Mel cepstrum coefficient and Mel logarithmic power shows that it can simulate the human ear better, accurately reflect the speaker's emotional characteristics, and identify depression.

The next step is to reduce the dimension of the selected feature combination. In this paper, principal component analysis is used.

Principal component analysis focuses on the main properties of things. It combines the original variables linearly through linear transformation and maps them from n-dimensional features to k dimensional features $(k < n)$. The k-dimensional data are orthogonal features reconstructed and are called principal components.

Set the training parameters of PCA to 0.9, and the training parameters represent 90% information of the reduced dimension data. The dimensional comparison and accuracy comparison of features before and after dimension reduction are shown in Table 10.

According to the experimental results in the above table, the feature dimension can be effectively reduced and the time complexity can be reduced by using the IPCA algorithm; Compared with the accuracy before dimensionality reduction, the decision accuracy obtained from the dimensionality reduced data is slightly improved, which shows that the dimensionality reduced by IPCA algorithm can improve the model performance in both computational complexity and decision accuracy, proving the effectiveness of the algorithm.

9 Model and experiments

In this paper, the voice based depression recognition model uses convolutional neural network training model. The training model is to obtain the optimal network parameters, so that the final model parameters can be used in various data sets as much as possible. The training process of convolutional neural network model is mainly divided into two parts: forward propagation of convolutional neural network and backward propagation of convolutional neural network. First, each convolution kernel is randomly initialized with weights, input data, and forward propagation of convolutional neural network is completed through convolution layer, pooling layer and full connection layer; After the output is obtained, compare the error between the output value of the network and the real value. When the error is within the expected range, the training is ended. When the error is outside the expected range, the error is returned to the network, and the errors of the full connection layer, pooling layer and convolution layer are calculated in turn. According to the updated weight of the error, the backward propagation of the convolution neural network is completed; Finally, on the basis of the new weight, repeat the above process until the error is within the allowable range. The training process of convolutional neural network is shown in Fig. 9.

The convolution neural network structure is designed as follows:

There are three convolution layers, and the first convolution layer has 20 convolution cores which is 5 × 5; The second convolution layer has 40 convolution cores which is 5 × 3; The third convolution layer has 60 convolution cores which is 4 × 3. There are two pool layers which is 2 × 2. The third convolution layer is followed by the full connection layer, and the full connection layer gets the final category as required. The number of training samples in this experiment is 107, and the testing samples is 47. This experiment is divided into two categories, so the output of the full connection layer corresponds to whether is depression.

Table 10 Dimension reduction before and after contrast

		Before PCA		After PCA	
		Dimension	Accuracy	Dimension3	Accuracy
	intensity	55		42	
Male	MFCC	681	61.3%	578	64.1%
	ZCR	5		5	
	intensity	55		41	
Female	MFCC	336	66.1%	245	69.7%
	LPCC	14		12	

Fig. 9 Training flow chart of CNN

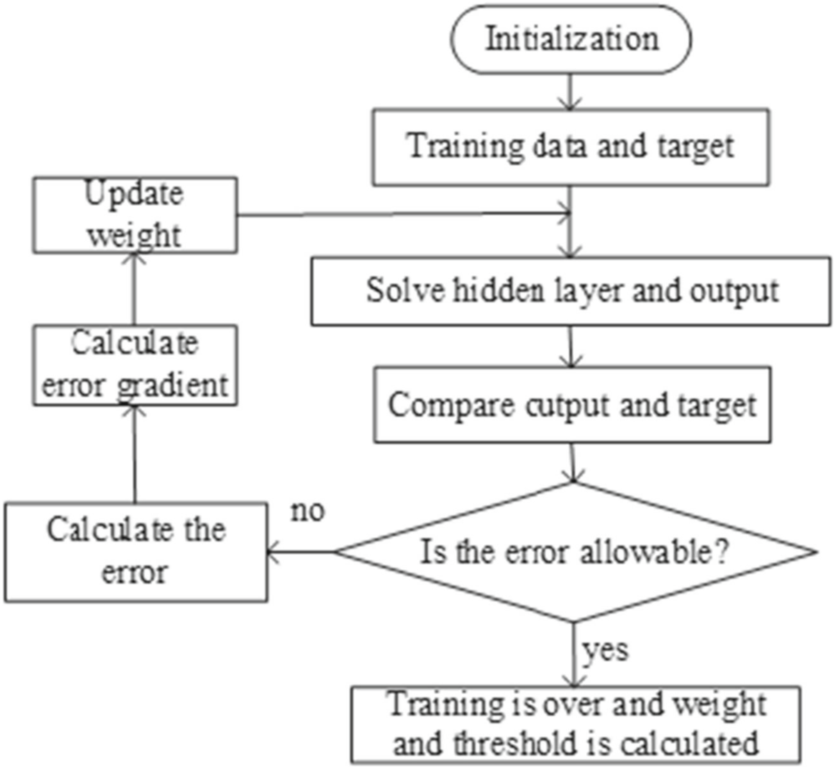
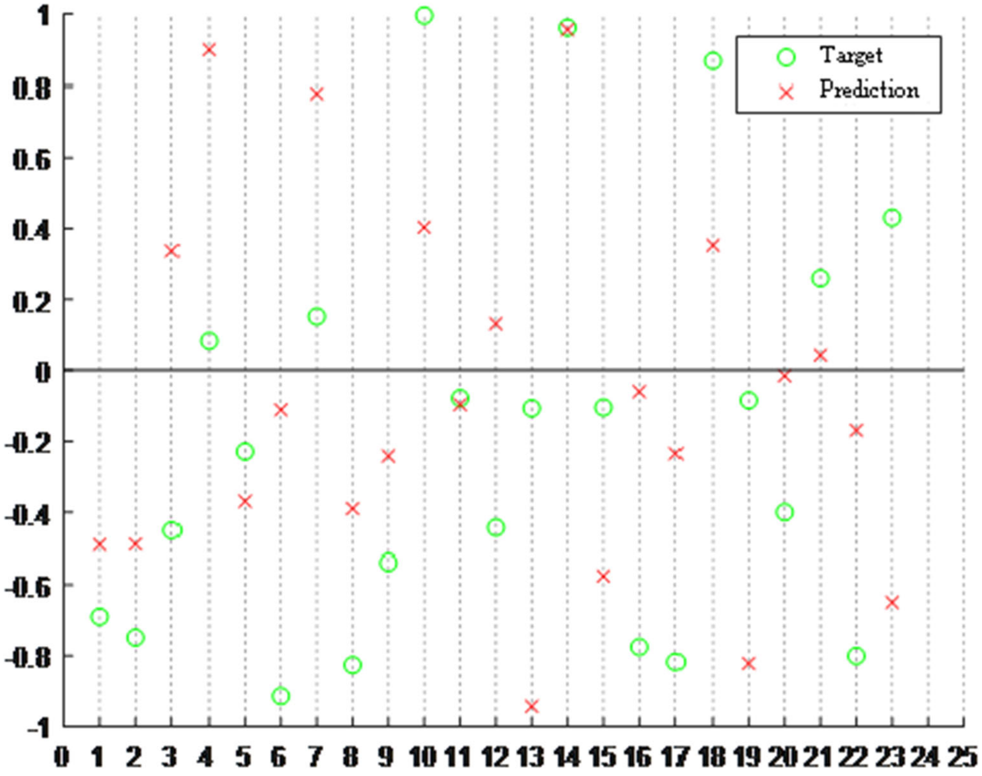


Fig. 10 Comparison of male voice judgment results



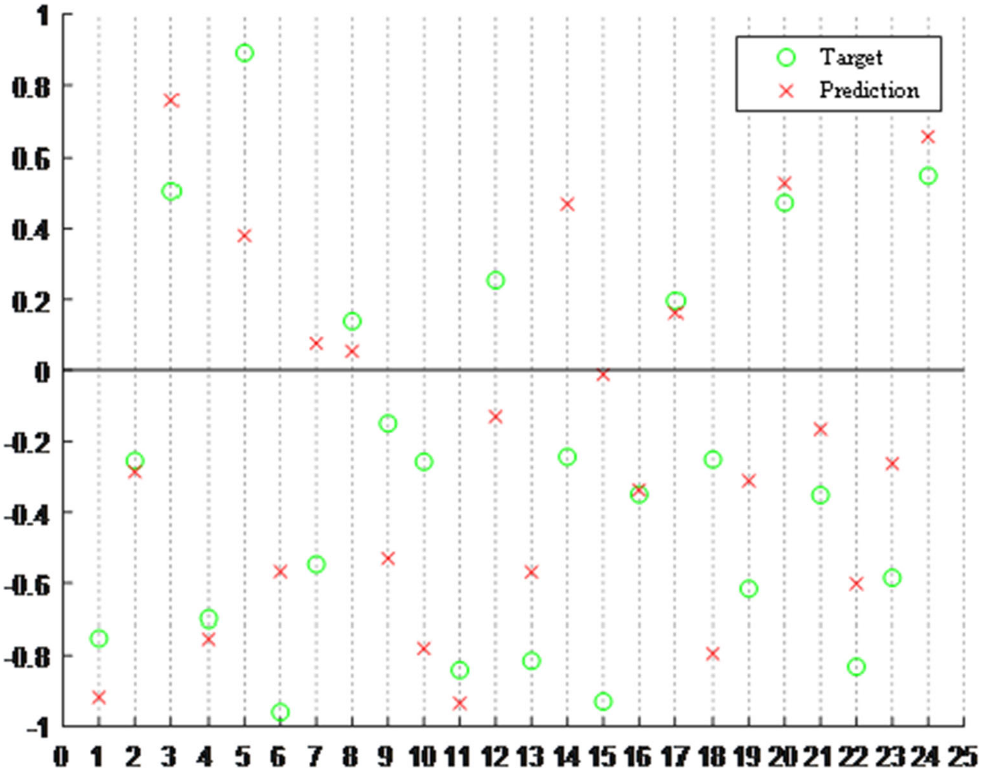


Fig. 11 Comparison of Female Voice Judgment Results

Before the training data is input into the model, the weight of the model is 0. If the initial values are all 0, the model cannot start training. Therefore, the model weight is required to initialize and assign values. In this step, the random function can generate weights and offsets.

The speech feature data is input into the network. After training, the number of iterations of the convolution neural network is set to 200. The results are shown in Figs. 10 and 11.

This section compares the predicted results with the real results. To make the comparison effect intuitive, generate a random number between 0-1 to represent the real result or the sample value with a decision result of 1, and generate a random number between - 1 and 0 to represent the real result or the sample value with a decision result of 0. The green dots in the figure represent the real results of the data in the test set, and the red cross points represent the prediction results of the data in the test set determined by the above model. The predicted value and the result value of the same sample are connected by a dotted line perpendicular to the x-axis. It can be seen that if two points of different colors on the same dotted line are on the same side of the x-axis, the

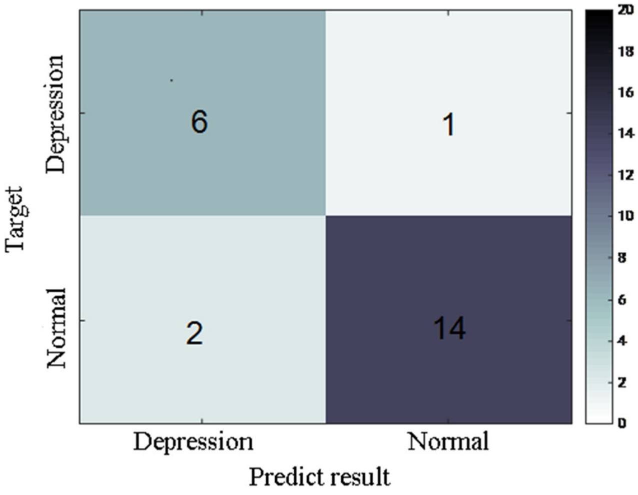


Fig. 12 Confusion matrix of male

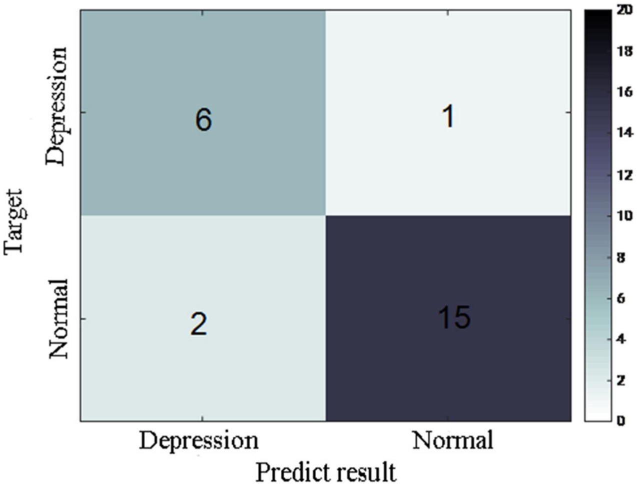


Fig. 13 Confusion matrix of female

prediction result is correct; If it is located on the opposite side of the x-axis, it indicates that the prediction result is incorrect.

As can be seen from Figs. 12 and 13, in male data, there are 20 groups of decision results on the same side of the x-axis and 3 groups of decision results on the opposite side of the x-axis; In female data, 22 groups of decision results are located on the same side of the x-axis, and 2 groups of decision results are located on the opposite side of the x-axis. Through calculation, it can be concluded that the accuracy of male and female voice based decision models for depression is 87.0% and 87.5% respectively, the model sensitivity is 85.7%, and the model specificity is 87.5% and 88.2% respectively. According to the practical significance of accuracy, sensitivity and specificity, the following results are obtained from the analysis: due to the differences between male and female physiology and psychology, the data of male and female are input into the model for separate training, and the accuracy of the model is improved, which makes the model more targeted and effectively improves the judgment results of both sexes; The model performs well in sensitivity and specificity, which indicates that the model has a high recognition ability for normal people and depressed people. According to the experimental data, the model confusion matrix is drawn as shown in Figs. 3, 4, 5, 6, 7, 8, 10, 11, 12 and 13.

10 Conclusion

This paper studies the decision model of depression based on speech. First of all, the speech signal of the interview is preprocessed so that it can be transformed into a form that can be processed by the computer; Secondly, OpenSMILE is used to extract speech emotional features, and then use the principal component analysis method to reduce the dimension of large dimension speech features. This paper introduces the concept of information entropy to select feature combinations first, and then reduce the dimension specifically to reduce the computational complexity; Finally, input the filtered voice features into the CNN network created, depression recognition result is given by CNN network. Convolutional neural network proposes that each neuron does not need to perceive all the input information, only perceives the local input information, and then combines these local information at a higher level to obtain all the representation information. The nerve units in different layers are connected locally, that is, the nerve units in each layer are only connected with some nerve units in the previous layer. Each nerve unit only responds to the area inside the receptive field, and does not care about the area outside the receptive field at all. This local connection mode ensures that the learned convolution check input spatial local mode has the strongest response. The weight sharing network structure makes it more similar to the biological neural network, reducing the complexity of the network model and the number of weights. According to the testing result, the accuracy of male is 87% and the accuracy of female is 87.5%. This result shows that the prediction result of speech model has important practical significance in assisting depression diagnosis.

Acknowledgements This work was partially supported by research funding of Jinhua Advanced Research Institute (G202209) and research funding of Jinhua Advanced Research Institute (G202207).

Author Contributions Tian Han provided the idea of this paper. Zhu Zhang Established the model, analyzed the results and finished the manuscript. Jing Xu extracted the features and selected the Features.

Funding This work was partially supported by research funding from Jinhua Advanced Research Institute (G202209) and research funding of Jinhua Advanced Research Institute (G202207).

Availability of data and materials Data in this paper will be made available on reasonable request

Declarations

Ethics approval Not applicable (The data in this paper is from public data sets and there is not other ethical content.)

Consent to participate Consent

Consent for Publication Consent

Conflict of Interests No potential conflict or competing of interest was reported by the authors.

References

1. Naujokat E, Perkuhn M, Harris M, Norra C (2009) Depression detection system
2. Rafiqul IM, Ashad KM, Ashir A, Kamal ARM, Hua W, Anwaar U (2018) Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 6(1):8–18
3. Ryder AG, Chentsova-Dutton YE (2012) Depression in china: integrating developmental psychopathology and cultural-clinical psychology. J Clin Child Adolesc Psychol 41(5):682–694
4. He YC, Zhang B, Qu W, Ning J, Quan HY, Xia Y, Yao Y, Han M. (2014) Value of serum monoamine neurotransmitters and their metabolites in diagnosis of comorbid anxiety and depression and major depressive disorder. Journal of third military medical university 36(08):806–810
5. Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synthetic Emotions 1(1):68–99
6. Mph R (2012) Epidemiologic evidence concerning the bereavement exclusion in major depression—reply:bereavement and the diagnosis of major depressive episode in the national epidemiologic survey on alcohol and related conditions. JAMA Psychiat 69(11):1179–1181
7. Subhashree R, Rathna GN (2016) Speech emotion recognition: Performance analysis based on fused algorithms and gmm modelling. Indian Journal of Science and Technology 9(11)

8. France DJ, Shiavi RG (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 47(7):829–837
9. Jiang H, Hu B, Liu Z, Yan L, Wang T, Liu F, Kang H, Li X (2017) Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Communication
10. Ooi KEB, Lech M, Allen NB (2014) Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system. Biomed Signal Process Control 14:228–239
11. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2010) Front-end factor analysis for speaker verification. IEEE Transactions on Audio Speech, and Language Processing 19(4):788–798
12. Huang Z, Epps J, Joachim D (2022) Investigation of speech landmark patterns for depression detection. IEEE Trans Affect Comput 13(2):666–679
13. Lorenzo-Trueba J, Henter GE, Takaki S, Yamagishi J, Morino Y, Ochiai Y (2018) Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis. Speech Comm 99:135–143
14. Asgari M, Shafran I (2018) Improvements to harmonic model for extracting better speech features in clinical applications. Computer Speech and Language 47:298–313
15. Rajisha TM, Sunija AP, Riyas KS (2016) Performance analysis of malayalam language speech emotion recognition system using ann/svm. Procedia Technol 24:1097–1104
16. Lang H, Cui C (2018) Automated depression analysis using convolutional neural networks from speech. J Biomed Inform 83:103–111
17. Nie W, Ren M, Nie J, Zhao S (2021) C-gcn: Correlation based graph convolutional network for audio-video emotion recognition. IEEE Trans Multimedia 23:3793–3804
18. James CM, Adam PV, Douglas EF, William RL (2012) Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 72(7):580–587

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.