Multimodal Deep Learning Approach for College Students' Mental Health Monitoring Using Online and Offline Data Integration

Jia Xu[1,*], Chunyan Huang[2]
[1]Information Technology Center, Zhejiang Business College, Hangzhou 310059, China
[2]School of Electronic Commerce, Zhejiang Business College, Hangzhou 310059, China
Email: Xujia1892@163.com; smallyellow98@sina.com
[*]Corresponding author

Keywords: Speech recognition, text extraction, facial expression recognition, psychological health monitoring, attention mechanism, emotion recognition

Received: June 12, 2025

In response to the difficulty of real-time monitoring and continuous tracking of college students' mental health in the era of new media, we collect the data from online student platforms and offline psychological interviews, and develop a college students' mental health monitoring system based on speech recognition, text extraction, and facial expression recognition, with the goal of achieving intelligent mental health management. The method is to first collect data from students' online network platforms and offline psychological interviews, mainly including multimodal information such as network text data, speech, video images, etc., to study automated speech recognition and text information extraction process methods. At the same time, for the micro expression recognition needs of video images, we propose a VGG19+SE+TA+LSTM network model, which extracts spatial features from four facial regions respectively. VGG19 is used as the convolutional neural network part on the traditional CNN+LSTM network structure, and channel and time attention mechanisms are introduced to enhance the network. The multi region features are fused as the features of a single frame image, and the multi frame image features are input in time series. The long short-term memory network (LSTM) based on time attention mechanism (TA) is used to extract temporal features. Experimental results have shown that integrating multiple modal data from online and offline sources can achieve the automation and intelligence of an intelligent monitoring system for college students' mental health. The fused feature algorithm improves the recognition rates of positive, negative, and neutral emotions by at least 8% and 4.8% respectively compared to the independent Fbank and MFCC feature algorithms, while the VGG19+SE+TA+LSTM network model improves the UF1 evaluation index by nearly 17.9% and 4.5% compared to the CNN+LSTM and VGG19+LSTM models, with providing emotional cognitive references for college students and effectively assisting college counselors in identifying the psychological emotions of college students.

Povzetek: Študija razvije inteligentni sistem za spremljanje duševnega zdravja študentov, ki združuje večmodalne podatke (besedilo, govor, video) z mikroizrazi prek modela VGG19+SE+TA+LSTM za avtomatsko prepoznavo čustev.

1 Introduction

According to the "2022 Survey Report on the Mental Health Status of Chinese College Students", psychological problems such as anxiety, depression, interpersonal communication disorders, and difficulties in self-awareness affect the learning outcomes and quality of life of college students. With the vigorous development of mobile Internet, new media has become an important platform for contemporary college students to obtain information, express themselves, and interact socially [1]. Digital mental health [2-3] has entered the fast lane of development, which provides new ideas for the prevention and research of college mental health. The combination of online and offline mental health intervention model emerged at the historic moment. Some colleges and universities use Internet technology, give full play to the advantages of new media technology, and expand the channels for college students to seek psychological help and counseling by integrating resources inside and outside the campus to build a new media mental health education platform or develop psychological counseling service APP. The campus mental health platform or APP provides convenient online psychological counseling. Through a comprehensive user feedback mechanism, universities can continuously track the usage and needs of college students. At the same time, the platform or APP can use new media virtual communities to build virtual psychological counseling communities and offline psychological counseling centers, achieving full coverage of online and offline psychological counseling [4]. The virtual community invites professional psychological consultants inside and outside the school to settle down in the mode of "Internet +psychological consultation", provides advice on college students' psychological problems through video, voice and other means, and

provides timely online professional psychological counseling services. In addition, the virtual community is also connected with offline psychological counseling centers, establishing an online appointment mechanism for medical treatment, so that college students with face-to-face in-depth counseling needs can quickly obtain professional services from offline psychological counseling centers.

Traditional monitoring and evaluation of mental health in universities mainly rely on subjective and static data collection methods such as psychological scales and daily observations, which cannot dynamically understand students' psychological states, and cannot efficiently and objectively evaluate and warn students of their psychological problems. There are certain limitations in dealing with mental health problems among college students, and it is urgent to introduce new concepts and methods to improve the effectiveness of mental health education. Therefore, many scholars have conducted research on monitoring and evaluating the mental health of college students based on artificial intelligence technology, such as using artificial intelligence to collect and comprehensively analyze students' psychological status through multiple channels [5], using multidimensional indicators [6-7] for mental health evaluation, realizing the transformation of mental health monitoring from static monitoring to dynamic management, and the transformation of mental health evaluation from subjective evaluation to big data algorithms [8-9]. At the same time, artificial intelligence technology can analyze and compare historical data of groups and individuals [10-11] on the basis of full data sampling big data analysis [12-13], and intelligently predict the psychological state of college students based on deep learning algorithm models [14-15]. The above studies have demonstrated that artificial intelligence, relying on real-time dynamic full sample data sampling and deep learning algorithms, can effectively integrate existing evaluation results to dynamically obtain a complete picture of the data, which helps to solve the inherent weakness of traditional evaluation methods that are biased and inefficient, and compensates for the subjective and inefficient limitations of traditional mental health monitoring and evaluation methods. The research content of this paper (including methods, datasets, models used, performance indicators, etc.) is presented in Table 1 below:

Table 1: Comparison table of research on college student mental health monitoring systems

Dimension	Specific content	Comparison
method	1.Multimodal data fusion (online platform data + offline psychological interviews) 2.Speech recognition and text extraction technology 3.Improved VGG19+SE+TA+LSTM network model (for video images)	single-mode speech data single-mode text data single-mode facial expression image
dataset	1. Online: web text data 2. Offline: speech data, video and image data	single-mode dataset
algorithm	1. Multimodal feature fusion (Fbank/MFCC + visual features) 2. Spatial-Temporal Attention Mechanism (SE+TA)	Independent Fbank features Independent MFCC features Independent video frame image features
performance indicators	1. Emotion recognition rate 2. Model UF1 and UAR indicators	CNN+LSTM model VGG19+LSTM model

2 A psychological health service system that integrates online and offline services

Digital empowerment provides new possibilities for mental health education. Through the Internet, big data, artificial intelligence and other technical means, we can achieve personalized, accurate and efficient mental health education, and greatly improve the coverage and service quality of mental health education. This article aims to organically combine traditional offline psychological counseling services with online psychological service platforms to achieve an integrated online and offline psychological health intervention model, forming a new type of psychological health service system with complementary advantages. Specifically, it includes the following aspects:

(1) Offline services: Set up a dedicated psychological counseling room, equipped with a professional team of psychological counselors, regularly carry out individual counseling, group counseling and other activities, and enhance students' mental health awareness and self-adjustment ability through holding psychological health lectures, themed activities and other forms.

(2) Online services: Utilizing the school's official website, APP, and other platforms to build an online psychological service platform, providing students with 24-hour uninterrupted psychological assistance hotline, online consultation, psychological testing, psychological classes, appointments for offline psychological clinics, psychological quality development training, and online "cloud psychological counseling" and other functions. In addition, the popularization of psychological knowledge and mutual support can be achieved through the establishment of mental health education courses, online psychological mutual aid groups, and other means.

(3) Integration of both: organically integrating online and offline services, such as referring offline cases to online for follow-up; During the online consultation process, students who require in-depth intervention should be promptly guided to receive professional treatment offline; Regularly publish offline activity notifications through online platforms to expand coverage and increase participation.

This article gathers online student behavior data and offline psychological interview data, collects digital multimodal data (including online text, voice chat, and video images) from online psychological counseling

platforms, mental health apps, and offline psychological interviews, constructs a psychological emotion recognition model based on deep learning technology, optimizes speech recognition, text extraction, and expression recognition algorithms, improves data analysis effectiveness, and aims to achieve real-time monitoring and intelligent evaluation of college students' mental health. This project was approved by Ethics Committee of Zhejiang Business College. Before collecting data, the research team first obtained informed consent from all students and informed them of the research purpose. At the same time, in order to protect the privacy of participants, the team anonymized the collected data, set limited access to the data, and only allowed team members participating in the research to process and analyze it. The system framework diagram is shown in Figure 1 below.

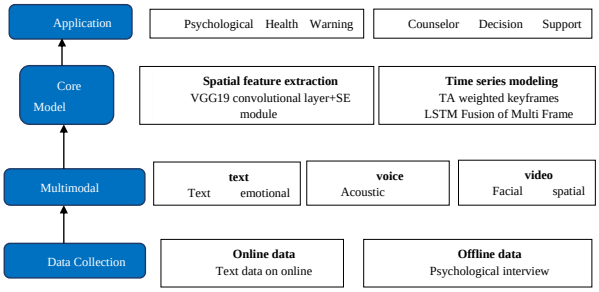

Figure 1: Framework diagram of college student mental health monitoring system

3 Speech recognition for monitoring the mental health of college students

Language data can reflect the health status of the brain, and sound features such as pitch, speed, intonation, volume, and prosody can reveal a person's psychological state [16]. We need to combine linguistic and psychological knowledge to analyze and model the biological and social cognitive information contained in language data and language related data, in order to achieve accurate prediction and evaluation of psychological disorders. By utilizing mobile devices and social platforms to collect voice and language data, digitizing it as a potential resource for early artificial intelligence screening of individuals with psychological disorders. This article studies an intelligent audio detection method for psychological disorders based on the mixed features of Fbank and MFCC [17], which mainly includes the following modules:

(1) Data collection module, used to collect audio datasets of online voice messages or offline psychological interviews with students; A clinically validated speech dataset consisting of 210 participants was constructed, with each participant receiving 15 seconds of audio as the dataset. The dataset categorizes emotions into three types: positive, negative, and neutral, with a ratio of 1:1:1.

(2) The audio preprocessing module is used for preprocessing audio, including audio clipping stage, feature extraction stage, and feature stitching stage. Extract Fbank and MFCC features [18] from the 15S audio, then concatenate them to obtain a 649 * 39 audio feature matrix.

(3) The audio feature vector determination module is used to determine the audio feature vector based on the audio dataset and the audio channel model; Cut the audio into a certain length, extract the Fbank and MFCC features of the audio separately, and then concatenate the Fbank and MFCC features. Compared to simple concatenation and weighted concatenation algorithms for different features, deep learning models have better feature fusion performance. We use convolutional neural networks (CNN) and deep learning models to fuse these two features, constructing an end-to-end neural network model that receives both Fbank and MFCC as inputs and outputs the final fused features or directly uses them for classification/regression tasks. The core pseudocode is as follows:

```
class FeatureFusionModel(nn.Module):
    def __init__(self, fbank_dim, mfcc_dim, output_dim):
        super(FeatureFusionModel, self).__init__()
        self.fc1 = nn.Linear(fbank_dim + mfcc_dim, 128)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(128, output_dim)

    def forward(self, fbank, mfcc):
        x = torch.cat((fbank, mfcc), dim=1)
        x = self.relu(self.fc1(x))
        x = self.fc2(x)
        return x
```

Figure 2: Fbank and MFCC feature fusion

In the aforementioned deep learning model, the EigenFusionModel class takes Fbank and MFCC as inputs, concatenates them, and processes them through a fully connected layer to ultimately output fused features.

(4) Two-dimensional convolutional neural network model construction module, used to construct a two-dimensional convolutional neural network model; The network fully connected stage includes a first deep convolution stage, a second deep convolution stage, and a network fully connected stage. The first deep convolution stage includes a first audio feature convolutional layer, a second audio feature convolutional layer, and a first audio feature pooling layer. The second deep convolution stage includes a third audio feature convolutional layer, a fourth audio feature convolutional layer, and a second audio feature pooling layer. The network fully connected stage includes an audio feature input layer, an audio feature hidden layer, and an audio feature input layer, as shown in Figure 3 below.

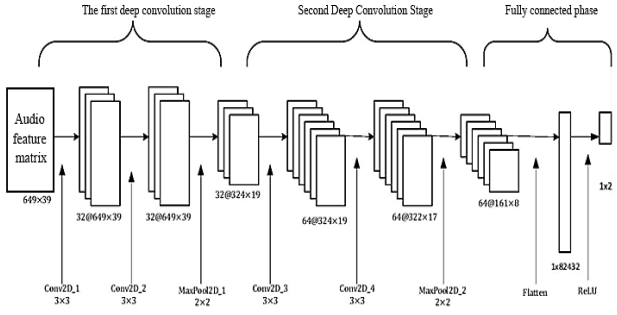

Figure 3: Two-dimensional convolutional neural network model

The two-dimensional convolutional neural network in the above figure consists of two deep convolution stages and one fully connected stage. First, the audio feature matrix Vec_{audio1} with a size of 649 × 39 is normalized by subtracting the average value and then dividing it by the maximum value. Then, in the first deep convolution stage: the first convolution layer (Conv2D_1)+the second convolution layer (Conv2D_2)+the first pooling layer (MaxPool2D_1), the convolution kernel size of the first and second convolution layers is set to 3 × 3, the number of convolution kernels is set to 32, the stride is set to 1, and the zero padding at the boundary is set to 1. The pooling layer adopts the maximum pooling method, and the pooling region kernel size is 2 × 2, with a stride of 2. The output feature vector Vec_{audio2} with 32 channels and a size of 324 × 19 is obtained. Next, through the second deep convolution stage: the third convolutional layer (Conv2D_2)+the fourth convolutional layer (Conv2D_2)+the second pooling layer (MaxPool2D_2), the convolution kernel size of the third and fourth convolutional layers is set to 3 × 3, the number is set to 32, the step size is set to 1, and the boundary zero padding is set to 1. The second pooling layer adopts the maximum pooling method, and the pooling region kernel size is 2 × 2, the step size is 2, the output channel number is 64, and the feature vector Vec_{audio3} with a size of 161 × 8 is output. The output of each convolutional layer is normalized by subtracting the average value and dividing it by the maximum value. Restore the distribution to its original input state. Flatten Vec_{audio3} into a feature vector Vec_{audio4} with a size of 1 × 82432, which serves as the input vector for the fully connected stage. The structure of the fully connected stage includes an input layer, a hidden layer, and an output layer. ReLU is used as the activation function, and the Dropout method is used to randomly inactivate a certain number of neurons to reduce overfitting. The inactivation probability is p=0.3, and the final output is a label feature vector Vec_{output} with a size of 1 × 2. Then use the sigmoid function to process Vec_{output} to obtain Vec_{target}, and determine whether it is a patient based on the two values of Vec_{target}.

(5) A label vector determination module is used to obtain label vectors based on the audio feature vector, video feature vector, and two-dimensional convolutional neural network model; The specific steps are as follows:

a. Set the number of convolution kernels in the first audio convolutional layer, the second audio convolutional layer, the third audio convolutional layer, and the fourth audio convolutional layer to 32, 32, 64, and 64, respectively. The size of the convolution kernels is set to 3 × 3, the step size is set to 1, and the boundary zero padding is set to 1. Normalize the outputs of each convolutional layer.

b. Both the first audio pooling layer and the second audio pooling layer adopt the maximum pooling method, with the pooling region kernel size set to 2 × 2 and the stride set to 2.

c. Flatten the output feature matrices of the first audio deep convolution stage and the second audio deep convolution stage into 1D feature vectors.

d. Take the 2D feature vector output by the audio preprocessing model as the input vector of the 2D convolutional neural network to obtain the 1D label vector.

(6) A module for determining patients with psychological abnormalities (negative emotions), used to identify patients with negative emotions based on the label vector.

Finally, experimental comparisons were conducted for different feature representations, and the above algorithm was applied to speech recognition operations. The speech data of 210 subjects in the data acquisition module were selected as the recognition content. The emotion classification results are "positive", "negative", and "neutral", with 70 samples for each classification. The sampling frequency is set to 16kHz and quantized to 16 bits. The training sample is set to 160, and the test sample is set to 50. First, train the model with training samples, conduct 5 experiments per group, calculate the mean as the result, and finally use the mean of the recognition and classification accuracy as the final recognition rate for each speech. The results are shown in Table 2, which compares the recognition rates of Fbank features, MFCC features, and their fusion features.

Table 2: Comparison of recognition rates of three feature extraction algorithms

	Positive	Negative	Neutral
Fbank features	78.6	77.8	82.3
MFCC features	82.4	81.6	85.5
Fusion features	88.1	86.7	90.3

From the results in the table above, it can be seen that using MFCC features can achieve better recognition performance compared to Fbank features. After feature fusion, the fused features can better present speech characteristics and achieve higher recognition rates.

4 Psychological prediction analysis based on text analysis

(1) Extraction of Online Consultation Text Data

Select the online consultation module built on the school's official website and online psychological service APP platform, and use the web crawler software "Octopus Collector" to crawl user consultation texts and related information, including user ID, question title, question

content, question time, question status, questioner's gender, age, psychologist's answer, doctor's title, number of likes, and answer time. Then, the extracted text data is preprocessed, including: ① manual data cleaning to remove invalid text: removing questions that users only submit question titles but do not provide detailed descriptions of the question content, or have incomplete content expression and no doctor answers. ② Text segmentation: Use Jieba segmentation tool in Python to segment user questions. Due to the particularity of Chinese text, in order to achieve better segmentation results, it is necessary to use a complete segmentation vocabulary list for semantic segmentation. This article adds professional vocabulary such as psychiatric and psychological related terms and drug names to a custom vocabulary list to obtain more accurate segmentation results. ③ Removing stop words: There are still a large number of words in the question content that have no practical significance for topic analysis, such as "I", "Hello", "You" and other words that frequently appear in user questions but cannot provide reference for topic recognition. Therefore, they were removed in the study.

(2) Convert voice data into text

For speech data, in addition to distinguishing psychological abnormalities based on speech features in the third section, in order to improve the effectiveness of psychological monitoring, it is necessary to further convert it into text for analysis. The collection of voice data is mainly carried out by various personnel such as teachers from the psychological counseling center, selecting voice messages or interview dialogue data generated by college students for online "cloud psychological counseling" or offline due to their own psychological problems, mainly covering students' basic information, voice recordings, etc. In order to prevent the leakage of students' personal privacy, anonymization was carried out in the data preprocessing stage, blocking the 4-7 digits of the phone number and identifying and deleting the real name in the voice data, ensuring the privacy and security of the data. The data style is shown in Table 3 below:

Table 3: Speech data styles

Number	file name	duration
01	2024-03-05-198****0245.wav	00:35:36
02	2024-03-05-158****3598.wav	00:45:03
03	2024-03-05-153****9845.wav	00:15:42
……	……	……

Due to the fact that the speech information in the dataset we collected is mostly in Chinese, we need to utilize existing mainstream Chinese language speech recognition open-source tools both domestically and internationally for recognition. This article selects the ASRT model and three open-source tools, namely iFLYTEK and Baidu AI's speech recognition, to test the recognition performance of Chinese. The experiment found that iFLYTEK's tool had the best performance, and the comparison results are shown in Table 4:

Table 4: Comparison results of word error rates in speech recognition

Tool/Time Dimension	<10min	10-30min	10-30min
ASRT	10.3%	15.3%	19.8%
Baidu AI	5.2%	7.8%	12.1%
iFLYTEK	4.8%	7.6%	11.5%

In order to reduce the error rate after converting speech into text, we propose a speech recognition model based on multi tool fusion. This model integrates open-source tools to complement each other's strengths and weaknesses, fully leveraging their respective advantages to further improve the accuracy of speech recognition. The method is to use ASRT, iFLYTEK, and Baidu AI for speech recognition of the same speech data, conduct detailed comparative analysis through experiments, and compare the final text obtained through string comparison. For the differences in recognition, based on the principle of minority obeying majority, the content recognized by most speech recognition tools is selected as the final result. If all three are different, the result recognized by iFLYTEK with the lowest overall word error rate in the experiment is selected. After model processing, we can obtain the speech to text conversion result with the lowest error rate.

(3) Text sentiment analysis

The text data obtained after the above (1) and (2) processing is stored in Chinese, and then the ROST CM software [19] is used to study the emotional tendencies of patients with psychological disorders on online health platforms through text mining and sentiment analysis methods. The software will output three emotional tendencies: neutral emotion, positive emotion, and negative emotion, and score each emotion. For example, 0 represents neutral emotion, -1~-100 represents negative emotion,+1~+100 represents positive emotion, and the numerical value represents the degree level.

5 Real time psychological crisis warning based on facial recognition emotion analysis

At present, research has been conducted based on video using computer image processing technology to extract feature information from raw input facial emotion images, and classify facial emotion features according to human emotional expression, in order to achieve psychological state recognition, such as elderly depression recognition [20]. Due to the temporal characteristics of offline psychological interviews with college students, the addition of LSTM network can achieve better recognition performance compared to traditional CNN network image feature extraction. For example, in reference [21], VGG was used as the convolutional neural network part in the traditional CNN+LSTM network, which proved that this method can effectively model spatiotemporal interactions

and identify salient features. Micro expressions are unconscious facial information that is difficult to artificially pretend or control, so they can better reflect people's true psychological state than macro expressions. This article proposes a VGG19+SE+TA+LSTM network model based on the requirements of micro expression recognition. VGG19 is added to the traditional CNN+LSTM network structure, and channel and temporal attention mechanisms are introduced to enhance the network. The VGG19+SE+TA+LSTM network model uses a visual geometry group network (VGG19) based on channel attention mechanism (SENet) to extract spatial features of the main facial regions, and the multi region features are fused as the features of a single frame image. The multi frame image features are input in time sequence to a long short-term memory network (LSTM) based on temporal attention mechanism (TA) to extract temporal features. The VGG19+SE+TA+LSTM network designed in this article is shown in Figure 4. The micro expression recognition algorithm mainly consists of four steps: 1) Preprocess the micro expression video frames, and capture the main areas of micro expression changes in each frame of the image: left eyebrow eye, right eyebrow eye, nose bottom, and lips; 2) Extract spatial feature information of each key region through VGG19 network, and fuse the features of these four regions as the spatial features of each frame image; 3) Embedding time and channel attention modules into the network to assign corresponding weights to different video frames and feature channels; 4) Utilize trained networks to achieve micro expression recognition.

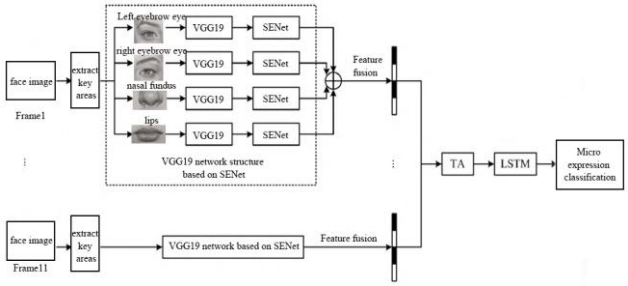

Figure 4: VGG19-SE-TA-LSTM Network

In the above figure, we use VGG19 to extract features of four important facial regions that reflect micro expression changes, and fuse the four features as spatial features of a single frame image. After obtaining spatial features through the VGG19 network, each frame of the image is input into the LSTM network in chronological order to obtain temporal features, and finally subjected to micro expression classification. This article adopts a unidirectional LSTM network, in which the hidden layer contains 256 nodes, followed by a fully connected layer for feature transformation, and then uses BN and Dropout layers to accelerate training convergence and prevent overfitting. Finally, the fully connected layer converts the feature vectors into the dimension of label vectors.

Channel Attention Module (SENet): A single facial region outputs 512 feature channels through the VGG19 convolutional network. After each VGG19 network, a Channel Attention (SENet) module [22] is introduced to improve the network's attention to important feature channels, suppress the influence of useless feature channels, and enhance the network's adaptability and performance by assigning weights to feature channels.

Time Attention Module (TA): Due to the short occurrence time of micro expressions, in order to highlight the role of keyframes in micro expression recognition, this paper introduces a time attention mechanism [23], which assigns corresponding weights to different frames and focuses attention on the keyframes in the sequence.

The output of the VGG19 network module is $F(x) = (f(x_1), f(x_2), \cdots, f(x_T))$ that the length of the sample sequence is T. $F(x)$ will be used as the input of the time attention module to obtain the hidden state $H = (h_1, h_2, \cdots, h_t)$, where h_t represents the hidden vector of the t frame of the sample sequence. Use a fully connected layer to calculate the frame-to-frame correlation in a micro expression sequence, as shown in formula (1) below.

$$s(h_t, h_i) = h_t^T W_a h_i \tag{1}$$

In the formula W_a represents the network weight matrix of the fully connected layer, h_t and h_i represents the hidden vector of the sequence.

At time step t, a_t represent the degree of influence of the entire time series on the time step vector h_t, where each element $a_{t,i}$ represents the magnitude of the effect of the i-th time step in the sequence on predicting the current time step t. $a_{t,i}$ use the normalized exponential function (softmax) to calculate, as shown in formula (2) below:

$$a_{t,i} = soft\max(s(h_t, h_i)) = \frac{\exp(s(h_t, h_i))}{\sum_{i=1}^{T} \exp(s(h_t, h_i))} \tag{2}$$

Finally, the weighted sum can obtain the attention weight a_t of each frame, and assign corresponding attention weights to different frames in the micro expression sequence, as shown in formula (3):

$$a_t = \sum_{i=1}^{T} a_{t,i} h_i \tag{3}$$

In order to verify the experimental comparison results of the algorithm in this paper, the experimental dataset uses the Institute of Psychology of the Chinese Academy of Sciences to build a CAS(ME)^3 face image database. This database provides approximately 80 hours of video, comprising over 8000000 frames, including 1030 manually annotated micro expressions and 3364 macro expressions. Such a large sample size can effectively validate the intelligent analysis method of micro expressions, while avoiding database bias. To ensure the accuracy of the samples in each experiment, considering that the research object of this article is college students, subjects with age too old (>35 years old) or too young (<18 years old) in the fused dataset, as well as samples with blurred images or mixed expressions, were excluded.

The basic emotion types in the database are divided into 8 categories: happiness, sadness, disgust, surprise, contempt, fear, repression, and tension. We classify emotions into positive, negative, surprise, or neutral, and the specific classification operation is: re label happiness as positive; Re label disgust, repression, anger, contempt, sadness, and fear as negative; The category of surprise remains unchanged; Other types of microexpressions are considered as neutral types. The final emotion classification dataset consists of 864 micro expressions, including 153 samples of positive emotions, 303 samples of negative emotions, 179 samples of surprise emotions, and 229 samples of neutral emotions. At the same time, in order to avoid the problem of small samples that restrict the application of deep learning in micro expression analysis and prevent overfitting of deep learning models, we enlarged the data to 10 times through image flipping, translation, scaling, mirror transformation, etc. for model training. The deep learning development platform used in the experiment is TensorFlow 2.0 framework. The experimental parameter settings are as follows: the optimizer introduces an adaptive Adam method; The initial learning rate is set to 0.0001, the attenuation factor is set to 0.8, and the minimum value is set to 0.000001; Set the batch processing quantity to 2 and the epoch to 200. This article sets up three sets of experiments to compare the training and recognition performance of traditional CNN+LSTM model, VGG19+LSTM model, and VGG19+SE+TA+LSTM model with attention mechanism under the same experimental data and environment, and observe the performance changes of the three models. Due to uneven distribution of labels, using traditional evaluation metrics such as Accuracy, Precision, Recall, and F1 will lead to excessive optimism towards those with large sample sizes. Use unweighted F1 score (UF1) and unweighted average recall (UAR) as performance metrics to avoid overfitting of the proposed method to a certain category. The calculation formulas for UF1 and UAR are as follows:

$$UF1 = \frac{1}{C}\sum_{i}^{C}\frac{2TP_c}{2TP_c + FP_c + FN_c} \tag{4}$$

Among them, C represents the number of categories, which are divided into four categories: positive, negative, surprise, and neutral. Therefore, C=4; TP_c, FP_c, FN_c refers to true positives, false positives, and false negatives in the classification results.

$$UAR = \frac{\sum_{i=1}^{N_c}\mathrm{Re}\,call_i}{N_c} \tag{5}$$

In the above formula, N_c is the number of samples c, Recall refers to the recall rate, and the calculation formula is as follows:

$$\mathrm{Re}\,call_c = TP_c / (TP_c + FN_c) \tag{6}$$

The experimental results are shown in Table 6 and Figure 5. The VGG19+SE+TA+LSTM model with attention mechanism has better training and recognition performance than the other two models. By summarizing literature and consulting with psychological experts, it was found that students with abnormal mental health usually exhibit negative micro expressions such as tension, anger, disgust, fear, and suppression, and occasionally show micro expressions of "surprise". Based on the micro expression classification results of the model, the mental health risk level is classified. Defining "positive" and "neutral" emotions as low-risk states; Define occasional 'surprise' as a medium risk state; And negative emotions are defined as high-risk states.

Table 6: Comparison of experimental results of different algorithms

Evaluation indicator / model	UF1 (fusion)	UAR (fusion)	UF1 (CASME)	UAR (CASME)	UF1 (CASME II)	UAR (CASME II)	UF1 (CAS(ME)^2)	UAR (CAS(ME)^2)
CNN+LSTM	0.52231	0.54822	0.48 12	0.50 14	0.5048	0.5036	0.5574	0.5652
VGG19+LSTM	0.65722	0.67833	0.5342	0.5624	0.5863	0.5846	0.6254	0.6315
VGG19+SE+TA+LSTM	0.70188	0.70366	0.6589	0.6654	0.6945	0.6956	0.7089	0.7046

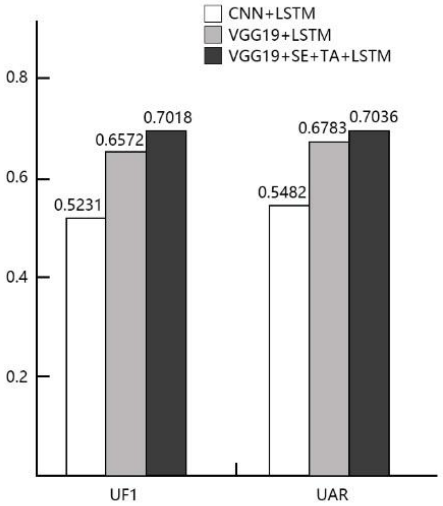

Figure 5: Comparison of recognition effects of different models

6 Discussion and summary

In today's society, mental health issues have become a global concern, especially in universities where students face multiple challenges such as academic pressure, interpersonal relationships, and future planning. The importance of mental health is self-evident. Therefore, how to effectively monitor and evaluate mental health has become an urgent issue for educators and professionals. The innovative improvements in network architecture proposed in this article contain: (1) dual enhancement of attention mechanism: the SE (Channel Attention) module dynamically calibrates the channel feature response to solve the problem of low channel information utilization in facial micro expression recognition of VGG19. TA (time attention) optimized the capture ability of LSTM for long sequence keyframes, and the synergy of the two significantly improved the UF1 index (17.9% higher than CNN+LSTM). (2) Multi region feature fusion strategy: Compared with traditional whole face input, four facial features were extracted in different regions and fused in time sequence, effectively alleviating the problem of local facial features being diluted by global information. At the same time, a trade-off between performance and cost was made through experimental comparison of different network models. In terms of computational efficiency, although VGG19+SE+TA+LSTM increased the number of parameters by about 15% compared to the basic VGG19+LSTM, it reduced redundant calculations by 30% through SE channel compression. At the same time, this paper controlled the overfitting of the model. In the experiment, it was found that the TA mechanism had a regularization effect on small-scale psychological datasets with sample sizes<1000, and the accuracy fluctuation of the test set was reduced by 2.3%. Of course, there are still limitations to this study, such as (1) the impact of sample diversity on data bias risk. Currently, UF1 improvement (4.5%) is based on campus scene data, and the generalization of micro expression recognition for cross-cultural/age groups needs to be verified. (2) the problem of label sparsity: the proportion of "neutral" emotion samples in psychological interview data reaches 62%, which may lead to insufficient sensitivity of UAR indicators to minority (negative emotions). Future research and improvement directions: (1) Optimization of computational costs: lightweight attention modules (such as ECA Net) can be explored to replace SE+TA combinations; (2) Multi modal collaboration bottleneck: Currently, text/speech modalities are only used for auxiliary decision-making, and future research on cross modal attention fusion mechanisms is needed.

This article explores a new practical model - multimodal emotional computing mental health services that integrate online and offline, with universities as the background, revealing its potential and value in improving students' mental health levels.

(1) Online mental health screening: overcoming spatial constraints and achieving comprehensive coverage

With the help of computer and Internet technology, colleges and universities have designed a psychological health assessment APP that is easy to operate and protects privacy, students can conduct psychological tests and self-assessment at any time and any place. This method is not only convenient and fast, but also can avoid the psychological pressure that face-to-face communication may bring, encouraging more students to actively participate. Meanwhile, big data analysis can help us more accurately identify students who may be experiencing psychological distress, providing a basis for subsequent interventions.

(2) Offline intervention: professional guidance, personalized care

Offline intervention is a deep response to the results of online screening. Each university has established specialized psychological counseling centers, equipped with professional psychological counselors, to provide one-on-one counseling services for students in need. Offline intervention emphasizes personalization and depth, which can provide more targeted assistance and meet the individual needs of students.

(3) Integration strategy: Complementary Advantages, Improving Service Quality

The key to the integration of online and offline lies in how to effectively combine the advantages of both. Online screening can serve as a preliminary "warning system" to identify students who may have problems, and then be deeply intervened by offline services. Meanwhile, offline services can also guide students to use online resources for self-learning and adjustment, forming a virtuous cycle. In addition, online platforms can also serve as feedback mechanisms to collect students' evaluations and suggestions on offline services, continuously optimizing service content and methods.

This article takes the author's university as an example. After implementing an integrated online and offline mental health service for one year, the number of students participating in screening significantly increased, and through offline intervention, many students were successfully helped to improve their psychological state. They regularly promote mental health knowledge through online platforms, and hold mental health weeks offline, providing free consultation and workshops. This model not only improves the accessibility and effectiveness of services, but also enhances students' awareness and importance of mental health. Looking ahead to the future, we still need to continue to innovate, and there is still huge room for development in the integration of online and offline mental health services. Future research directions include: (1) Text sentiment analysis using ROST CM is outdated and lacks rigor. Consider adopting transformer-based models for Chinese sentiment analysis, e.g., BERT variants trained on Chinese datasets. Also, quantify the sentiment prediction performance on a labeled subset. (2) Compare the emotional results (positive/negative/neutral) recognized by the system with the evaluation of professional psychologists, establish evaluation criteria such as accuracy, sensitivity, and specificity, and design a double-blind experiment to verify the consistency between the system evaluation and the professional evaluation. (3) Regarding the risk of false negatives, the system failed to identify the real psychological issues. In

the future, research plans to adopt multimodal data cross validation and set sensitivity thresholds to prevent them.

Acknowledgment

This work is supported by the research result of scientific research project of Zhejiang Research Institute of Education Science “Research on the Construction of Multi modal Curriculum Resources and Personalized Teaching in Vocational Colleges Empowered by Embodied Intelligence” in 2025 (Grant No.:2025SCG340).

References

[1] Sun L, Yang Z. The problems and causes of college students' mental health education based on new media environment[J]. Applied & Educational Psychology,2024,5(3). DOI:10.23977/APPEP.2024.050322

[2] Kolenik T, Gams M. Intelligent Cognitive Assistants for Attitude and Behavior Change Support in Mental Health: State-of-the-Art Technical Review. Electronics. 2021; 10(11):1250. https://doi.org/10.3390/electronics10111250

[3] Kolenik, Tine & Gams, Matjaz. (2021). Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality? IEEE Technology and Society Magazine. 40. 80-86. DOI:10.1109/MTS.2021.3056288.

[4] Kolenik, Tine. (2022). Methods in Digital Mental Health: Smartphone-Based Assessment and Intervention for Stress, Anxiety, and Depression. DOI: 10.1007/978-3-030-91181-2_7.

[5] Rui L. Early Warning Model of College Students' Psychological Crises Based on Big Data Mining and SEM. International Journal of Information Technologies and Systems Approach (IJITSA), 2023, 16(2):1-17. DOI:10.4018/IJITSA.316164

[6] Jingjing L, Guangyuan S, Jing Z, et al. Prediction of College Students' Psychological Crisis Based on Data Mining. Mobile Information Systems, 2021. DOI:10.1155/2021/9979770

[7] Li X. Research on the Application of Data Mining Technology in College Students’ Mental Health Education in the Network Age. Security and Communication Networks, 2022. DOI:10.1155/2022/4449066

[8] Panpan L, Feng L. An Assessment and Analysis Model of Psychological Health of College Students Based on Convolutional Neural Networks. Computational Intelligence and Neuroscience, 2022, 20227586918-7586918. DOI:10.1155/2022/7586918

[9] Cai B, Wang D. Prediction of psychological intervention for college students in digital entertainment media environment based on artificial intelligence and parallel computing algorithms. Entertainment Computing, 2025, 52100858-100858. DOI: 10.1016/J.ENTCOM.2024.100858

[10] Tine K, Gü S, Nter, et al. Computational Psychotherapy System for Mental Health Prediction and Behavior Change with a Conversational Agent [J]. Neuropsychiatric disease and treatment,2024,202465-2498. DOI: https://doi.org/10.2147/NDT.S417695

[11] Tine Kolenik. Intelligent cognitive system for computational psychotherapy with a conversational agent for attitude and behavior change in stress, anxiety, and depression. Informatica, 2025,49(2):451-454. DOI: https://doi.org/10.31449/inf.v49i2.8738

[12] International T O. Artificial Intelligence-Based Prediction of Individual Differences in Psychological Occupational Therapy Intervention Guided by the Realization of Occupational Values. Occupational therapy international, 2024, 9853562-9853562. DOI:10.1155/2024/9853562

[13] Li Y, Shuo S, Yu D. Graph Neural Network on Psychological Prediction of College Students Special Education. Journal of autism and developmental disorders, 2023, 54(4):1622-1622. DOI:10.1007/S10803-023-06068-6

[14] Vikas K, Praveen K, Masoud M. An intelligent disease prediction system for psychological diseases by implementing hybrid hopfield recurrent neural network approach. Intelligent Systems with Applications ,2023,18. DOI:10.1016/J.ISWA.2023.200208

[15] Computational N A I. Construction of a Prediction Model for College Students' Psychological Disorders Based on Decision Systems and Improved Neural Networks. Computational intelligence and neuroscience, 2023, 9813150-9813150. DOI:10.1155/2023/9813150

[16] Fadilah A N, Habibie H, Kristina A S, et al. Analysis of the mental health of pharmacy students at A number of public and private universities in Indonesia. Exploratory Research in Clinical and Social Pharmacy, 2024, 16100500-100500. DOI:10.1016/J.RCSOP.2024.100500

[17] Khan A W, Qudous U H, Farhan A A. Speech emotion recognition using feature fusion: a hybrid approach to deep learning. Multimedia Tools and Applications, 2024,83(31):75557-75584. DOI:10.1007/S11042-024-18316-7

[18] Singh K M. Identification of Speaker from Disguised Voice Using MFCC Feature Extraction, Chi-Square and Classification Technique. Wireless Personal Communications,2024,(prepublish):1-15. DOI:10.1007/S11277-024-11542-0

[19] Jia H, Wang X. The Performance Characteristics and Generation Logic of Citizen National Identity in News Communication of Major Scientific and Technological Achievements: Analysis of Short Video Comment Text Based on ROST CM 6.0 Software. Journal of Yangtze River Normal University, 2023, 39 (05): 65-74. DOI:10.19933/j.cnki.ISSN1674-3652.2023.05.008

[20] Tsai H H, Li R J, Shieh Y W. The Application of Emotion Valence Ratios in Facial Emotion

Recognition for Detecting Depression Among Older Adults in Institutional Settings. Studies in health technology and informatics, 2024, 318194-195. DOI:10.3233/SHTI240924

[21] Chouhayebi H, Mahraz A M, Riffi J, et al. Human Emotion Recognition Based on Spatio-Temporal Facial Features Using HOG-HOF and VGG-LSTM. Computers,2024,13(4):101. DOI:10.3390/COMPUTERS13040101

[22] Fu R, Tian M. Classroom Facial Expression Recognition Method Based on Conv3D-ConvLSTM-SEnet in Online Education Environment. Journal of Circuits, Systems and Computers,2023,33(07). DOI:10.1142/S0218126624501317

[23] Saheed K Y, Omole I A, Sabit O M. GA-mADAM-IIoT: A new lightweight threats detection in the industrial IoT via genetic algorithm with attention mechanism and LSTM on multivariate time series sensor data. Sensors International,2025,6100297-100297. DOI:10.1016/J.SINTL.2024.100297