International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 4974-4984

https://doi.org/10.70517/ijhsa463426

Research on the development of mental health intervention technology based on multimodal image recognition and emotional reasoning

Zhige Lyu[1,*]
[1] Guilin Normal University, Guilin, Guangxi, 541199, China
Corresponding authors: (e-mail: xwz20032000@163.com).

Abstract Paying attention to the changing situation of college students' emotional state is the basis for orderly psychological intervention. This paper constructs a three-stage abnormal behavior detection model for college students, which includes target detection, multi-target tracking and abnormal behavior detection. The YOLOv5s detection module, which is small in size and fast in operation, is selected to detect students' behavioral and emotional information on the premise of ensuring the completeness of feature information extraction. Based on the deflationary dot product self-attention method, continuous emotion inference of students under multimodal fusion is realized. Combined with the emotion recognition reasoning results, psychological intervention for abnormal students is carried out. The results found that the area of ROC curve reaches 0.9, and the effect of behavioral-emotional recognition is good. The average accuracy of the model's emotional reasoning for five subjects was 99.54%, and it had a fast running speed and fine emotional classification effect. The scores of the 4 scales before and after the psychological intervention of abnormal students were P<0.01, and the mental health level was effectively improved after the intervention.

Index Terms YOLOv5s module, self-attention method, multimodal fusion, affective reasoning, psychological intervention

I. Introduction

Mental health is increasingly emphasized in modern society, and mental health interventions, as an important tool, are widely used to promote the mental health of individuals [1], [2]. Mental health intervention is an intervention to prevent, correct or rehabilitate people's mental health problems through professional psychological methods and techniques [3], [4]. It aims to promote the mental health and well-being of individuals by changing their cognition, emotion, and behavior [5].

The development of mental health interventions has continued for centuries, experiencing continuous exploration and refinement, and gradually forming an important branch of modern psychological interventions [6], [7]. Early mental health interventions were mainly focused on religion and philosophy [8]. In ancient religious beliefs, people believed that the mind and the body were interconnected, and psychological problems were treated through religious rituals and prayers [9], [10]. In ancient Greece, the philosopher Aristotle believed that reason was the key to solving psychological problems, and his views had a significant impact on later psychological interventions [11]. With the development of time, mental health intervention has gradually become an independent discipline and has been widely used in psychology, medicine, etc. With the development of science and technology, mental health intervention methods have become more diversified and advanced [12]-[15]. Modern mental health interventions can provide remote counseling and consultation through the Internet and mobile applications, making it convenient for individuals to receive psychological help at home [16]-[18]. In addition, mental health interventions have incorporated findings from neuroscience, applying biological and psychological knowledge to psychotherapy [19], [20]. The developmental history of mental health interventions not only reflects the progress of psychological research, but also reflects the deepening of human concern for mental health [21]-[23]. With the development of society and the improvement of people's awareness of mental health, the importance of psychological interventions has become more and more prominent, in which the multimodal image recognition and emotion reasoning methods based on multimodal image recognition and emotion reasoning methods have gradually been emphasized [24]-[26].

In this paper, we construct a multi-stage college student abnormal behavior detection model to identify the location information of students in the video. And the non-maximum suppression method of DIoU is utilized to solve the target tracking failure problem. YOLOv5s containing input, trunk network, neck network, and head network is

selected to handle the basic detection work in the target detection stage, fusing multi-scale features to improve the model computing speed. A model based on the deflationary dot product attention mechanism is designed to reason about the students' emotional changes by combining the video contextual information and the characters' emotional transmission. Classify the obtained multimodal fusion of emotional features using a three-layer perceptual machine to determine the emotional state that the student is in. Personalized psychological intervention strategies are adopted to intervene with students in abnormal emotional fluctuation states to improve their psychological health.

II. Model construction based on multimodal image recognition and emotion inference

This chapter systematically analyzes how to construct a model for detecting abnormal behaviors of college students based on multimodal image recognition and affective reasoning, which is described in detail below.

II. A. Modeling

II. A. 1) Overall framework of the model

This paper constructs a three-stage model for detecting abnormal behavior of college students. In which each stage model requires separate training and testing. In the target detection stage, the customized college student human target detection dataset (CS-Object) is used in this paper. For the multi-target tracking phase, the MOT16 public dataset is used for testing and is trained with the CUHK03 dataset for human weight recognition. The abnormal behavior detection phase is trained and tested using a customized college student abnormal behavior detection dataset (CS-Behavior). First is the person detection phase, where the video frames are input and the person feature information is extracted by feature extraction, which gives the location information of the detected person, including coordinates and width and height information. The detected personnel position information is used as an input to the multi-target tracking phase to get the personnel position information with ID information. The target frame information with ID information is used as an input, and since the behavior recognition is performed by continuous video frames, the processed video frames are fused as an input to the abnormal behavior recognition stage. After information extraction through the feature extraction network in this stage, the final fully connected layer is classified to get the category information. Target detection phase to solve the problem of target frame loss that occurs in target occlusion, this paper introduces the use of non-maximum suppression of DIoU (DIoU-NMS) to solve this problem, tracking phase due to the ID switching problem that occurs in the case where the target is occluded by street light poles, etc., this paper introduces the pedestrian re-recognition module to the multi-target tracking network, so that not only comparing the position information of the target frames for the tracking of the target between the front and back frames, the appearance information of the object in the target frame is introduced, and the ID switching problem is effectively reduced by the multi-target tracking network that incorporates the appearance information. In order to learn more information about small targets in the process of behavior recognition, the feature pyramid module is introduced, and by fusing the feature information between different feature layers, the small targets are mapped on a large sensory field and the target information is more obvious.

II. A. 2) YOLOv5s detection module

The model magnitude of YOLOv5 in the limit state is nearly 90% smaller than YOLOv4. The detection accuracy of YOLOv5 is comparable to that of YOLOv4, which uses DarkNet as its backbone feature extraction network.YOLOv5 can be categorized into YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x based on the structure of the network and the parameters of the model.YOLOv5s has the smallest model size and the fastest operation speed. . Since target detection is only one part of the three-phase network architecture proposed in this paper, and to ensure the real-time performance of detection, YOLOv5s is chosen as the basic detection network in the target detection phase.

The network architecture of YOLOv5s is mainly composed of four parts, which are input, backbone network, neck network, and head network.

1) Trunk network

The main task is to extract the information of the video frame target. It is mainly composed of Focus module, Conv module, C3 module and SPPF module.Focus module is proposed in YOLOv5, which is a special kind of downsampling. The purpose is to reduce the dimensional information of the features in the feature layer in order to retain the information in the low-latitude network, which reduces the amount of data by sacrificing some of the information in a way that is useful for preventing overfitting in the training process. By slicing the high-resolution image and splitting it into multiple low-resolution images, using the form of spaced column sampling and splicing, the original input is 660*660*4 video frame data, and 330*330*16 feature layer data is generated by slicing, and after splicing, a convolution is performed to get the final 330*330*84 feature layer.

Each Conv module is composed of Conv2d, BatchNorm2d and activation function, and the role of the convolution layer is to perform the convolution operation on the input image, and then obtain the feature information. Not only that, but also to organize the feature map information that has been acquired. Normalization is done to make every

batch of data in a uniform format. Activation Function Through nonlinear features, the neural network can learn and understand more complex information, so that the generated feature function can better cope with the complex transformation of the function. Nowadays, during the research, Sigmoid and ReLU are used more, with formulas such as (1) and (2), and in this paper, in order to improve the performance of the model, SiLU activation function is used, with formulas such as (3).

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$\mathrm{Re}\,LU(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{2}$$

$$SiLU(x) = x * sigmoid(x) = \frac{x}{1+e^{-x}} \tag{3}$$

C3 is the more important feature extraction network module. The new CSPDarket53 is used in the backbone network, which integrates Darknet53 and CSPDensNet (C3) and extracts features through different deep convolutional layers, aiming to enhance the learning capability of the CNN, speed up the inference speed of the network, eliminate the computational upper limit, and reduce the memory usage. Figure 1 shows the network structure of C3. From the figure, it can be seen that when entering C3, all the convolutional modules in it only play the roles of boosting the dimension and lowering the dimension; boosting the dimension is to obtain more feature information, while lowering the dimension is to allow the network to better understand the feature information. The feature map then has two branches, through a convolutional layer and a Bottleneck module, which uses a residual module designed to connect the input of the latter to the output of the convolution to prevent the gradient from vanishing. Eventually the two branches are fused.

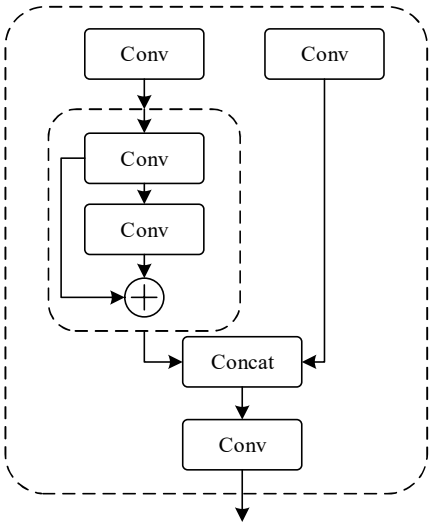

Figure 1: C3 structure

The last module in the backbone network, SPPF, is an improvement on the original SPP called spatial pyramid pooling. Its main purpose is to transform the original parallel SPP structure into a more efficient serial structure, which significantly reduces the amount of computation and increases the speed of modeling. Despite the change in structure, SPPF is functionally consistent with traditional SPP. It is capable of multi-scale feature extraction and encoding of images, rescaling them to a fixed size regardless of the size of the input image, and generating fixed-length feature vectors. This module also greatly fuses multi-scale features, fusing features from feature maps of different scales under the same representation. This powerful feature makes SPPF play an indispensable role in YOLOv5, providing strong support for model performance and computational speed.

2) Neck network

There are two important concepts in this phase, graphical features and semantic features. In the process of neural network extracting features, the network level goes from shallow to deep, in the shallow neural network, the extracted features such as texture, contour, color, etc., these features are called graphical features, with the

deepening of the network layer, it will be fusion of these features, common is the fusion of the color and the contour, which can distinguish some of the objects in the graph, these features are called semantic features. Overall, the lower feature layers have fewer convolutions, lower semantic information and more noise, but higher resolution and broader feature information. After many convolutions, the upper feature layers have stronger semantic information but are less perceptive of details.

The role of the neck network is to fuse the shallow feature layer with the deeper semantic features. The human dataset we use has many different scales. In the process of extracting the target features from the backbone network, another corresponding branch is extracted, which is then path-aggregated with the branch extracted from another feature layer after up-sampling. In the neck network, PAN is used to obtain multi-level target feature mapping and CSP structure is used to splice and fuse features between neighboring layers.

Then, combined with the FPN structure of the feature pyramid that integrates feature information from top to bottom, the detection capability of the human body can be further improved.

3) Head network

Also known as Detect module, it consists of just three 1*1 convolutions corresponding to three detection feature layers.

In order to address the difference in size of the detected targets, the input frame is divided into grid regions of size 40*40, 60*60 and 120*120. The smaller grids are responsible for detecting smaller targets, while the larger grids are responsible for detecting larger targets. The grid is responsible for predicting a certain target object when its center falls into the grid. Suppose (c_x, c_y) is the reference point of the grid cell in the upper-left corner of the frame image, (p_w, p_h) is the width and height of the bounding frame, and (t_x, t_y) is the x-coordinate offset and y-coordinate offset of the center of the target at the predicted width and height of the grid in the upper-left corner of the grid . In earlier versions of YOLO, the prediction equation for the prediction box was defined as:

$$\begin{aligned} b_x &= \delta(t_x) + c_x \\ b_y &= \delta(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{4}$$

where (b_x, b_y) are the coordinates of the centroid, (b_w, b_h) are the width and height of the prediction, and δ is the sigmoid that keeps the predicted centroid from exceeding the area of the corresponding mesh, with offsets limited to 0 to 1.

In YOLOv4, it was found that the prediction box in the original algorithm has a design flaw. For example, when the target center of the prediction box is infinitely close to the upper left or lower right corner of the grid ($\delta(t_x, t_y)$ is infinitely close to 0 and 1) and $\delta(t_x)$ and $\delta(t_y)$ need to be close to infinity. In YOLOv5, the offsets are scaled from (0, 1) in the original algorithm to (-0.5, 1.5) to make it easier to access the upper left and lower right corners of the grid. The predicted centroids as well as the width and height are then adjusted as follows:

$$\begin{aligned} b_x &= \delta(2 * \delta(t_x) - 0.5) + c_x \\ b_y &= \delta(2 * \delta(t_y) - 0.5) + c_y \\ b_w &= p_w (2 * \delta(t_w))^2 \\ b_h &= p_h (2 * \delta(t_h))^2 \end{aligned} \tag{5}$$

II. B.Analysis of Emotional Reasoning Methods

II. B. 1) A multimodal approach to affective reasoning

In this section, the sample is formally defined as $\left(V, P_m, S_n, E_{m,n}\right)$, where $V = \left\{\{P_i\}_{i=1}^{M}, \{S_j\}_{j=1}^{N}\right\}$ is a video containing M characters and N semantic segments , $P_m \in \{P_i\}_{i=1}^{M}$ is the target character, $S_n \in \{S_j\}_{j=1}^{N}$ is the target segment where the emotion moment is located, and $E_{m,n}$ is the emotion category of the target character P_m labeled on the emotion segment S_n , which can be one of the main emotions or fine-grained emotions.

Considering that the target person P_m may be missing visual signals, or missing audio-textual signals (non-speaker), one or more modalities may be missing from S_n for direct recognition of emotions. Therefore, the goal of this section is: in addition to the multimodal signals of the target segment S_n , to perform multimodal emotion inference on the target person P_m by utilizing contextual information in the video V , as well as external knowledge such as the character's personality.

II. B. 2) Self-Attention Based Modeling Framework

In this paper, the following inference strategies are considered: on a psychological level, first, characters' emotions are continuously changing, so the emotional context of a character plays an important role in inferring the emotion of his or her current state. Second, characters' emotions are transmitted and influenced by each other, e.g., both happy and sad moods may be rapidly contagious, while conversing with an angry person may cause fearful emotions. Finally, a character's a priori knowledge, such as personality and past history, can portray the character better, as different characters have different emotions for the same event. Therefore, in order to perform character-level multimodal affective reasoning, this chapter proposes a model based on an attention mechanism to adequately model a character's affective context, affective propagation between characters, and a priori knowledge about a character's personality.

The attention method used in this chapter is the deflated dot product attention used in Transformer:

$$Att(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{6}$$

where Q , K , V are query vectors, key vectors, and value vectors, respectively, and d is the dimension of the query vectors and key vectors. Suppose a video $V=\left\{\{P\}_{i=1}^{M},\{S_j\}_{j=1}^{N}\right\}$ has N clips and M characters, and the task of this chapter is to identify the emotion category $E_{m,n}$ of the target character P_m in the target emotion clip S_n . The sections are described in detail below.

Multimodal embedding vectors. In this chapter, three encoders E_a , E_t , E_v are used for the multimodal input features $\{(a_{i,j},t_{i,j},v_{i,j})\}_{i=1,j=1}^{M,N}$ are encoded to obtain a multimodal embedding vector $\left\{\left(f_{i,j}^{(a)},f_{i,j}^{(t)},f_{i,j}^{(v)}\right)\right\}_{i=1,j=1}^{M,N}$ in uniform dimensions, and $f_{i,j}^{(k)}\in\square^{256}$, where $k\in\{a,t,v\}$ represents sound, text and visual modalities.

Character embedding vectors. In this chapter, a two-layer perceptual machine E_p is used to encode the character traits p_i of the character P_i to obtain the character embedding vector $f_i^{(p)}$. Subsequently, it is used as a global knowledge to augment the multimodal embedding vectors, yielding

$$\square_{i,j}^{(k)} = [f_{i,j}^{(k)};f_i^{(p)}] \tag{7}$$

where $k\in\{a,t,v\}$ represents sound, text and visual modalities.

Attentional mechanisms at the modal level: character-guided temporal and inter-character attentional enhancement was performed at the modal level prior to fusion of the multimodal embedding vectors. First, at each semantic segment S_j , a self-attention operation is performed on the intercharacter

$$h_{i,j}^{(k),1} = \square_{i,j}^{(k)} + Att\left(\square_{i,j}^{(k)},\square_{:,j}^{(k)},\square_{:,j}^{(k)}\right) \tag{8}$$

In this way, relationships and emotion propagation between pairs of characters are implicitly modeled, guided by the character traits $f^{(p)}$. Next, capturing the contextual affective associations of each character P_i in its time, i.e., performing self-attention between contextual semantic fragments, yields

$$h_{i,j}^{(k),2} = h_{i,j}^{(k),1} + Att\left(h_{i,j}^{(k),1},h_{i,:}^{(k),1},h_{i,:}^{(k),1}\right) \tag{9}$$

If the modality k is missing for the character P_i on the semantic segment S_j , then $\square_{i,j}^{(k)}$ is set to 0 in the attention mechanism at the modality level.

Multimodal feature fusion. This section takes an early fusion approach to multimodal feature fusion and connects the character embedding vectors to the multimodal features again after the fusion so as to directly utilize the character's personality traits at a higher level, resulting in the multimodal fusion features on each semantic segment:

$$h_{i,j} = [h_{i,j}^{(a),2};h_{i,j}^{(t),2};h_{i,j}^{(v),2};f_i^{(p)}] \tag{10}$$

Character-level attention mechanism. For target emotion segment S_j , the character-level self-attention is used to further capture its high-level inter-character emotion propagation and enhance the character-level multimodal representation. Thus, the final enhanced multimodal representation is:

$$\hat{h}_{i,j} = h_{i,j} + Att\left(h_{i,j},h_{:,j},h_{:,j}\right) \tag{11}$$

Ultimately, the augmented multimodal fusion features $\hat{h}_{m,n}$ about the target person P_m under the target emotion moment S_n are fed into a three-layer perceptual machine for emotion classification.

II. C.Algorithm Validation

II. C. 1) Experimental environment

In order to achieve better experimental results, the experiment was chosen to be conducted in the hardware environment of STRIX-RTX2080Ti model CPU with 8 cores and a memory size of 32G. The phython programming language was chosen for programming, and then tensorflow was used to build the model framework.

II. C. 2) Data sources

The data were obtained from the College Student Human Object Detection dataset (CS-Object), the MOT16 public dataset, and the College Student Abnormal Behavior Detection dataset (CS-Behavior). In total, the dataset contains over 5,000 emotion videos containing a variety of common emotions such as anger, fear, sadness, surprise, and emotionlessness.

II. C. 3) Data pre-processing

The raw data thus collected is susceptible to external interference noise, making a large amount of redundant information in the data, which may increase the amount of emotional feature extraction calculations, leading to a reduction in the accuracy and efficiency of feature extraction. Based on this, a preprocessing operation is proposed for the video signal. The specific preprocessing methods are pre-emphasis, windowing and frame splitting.

1) Pre-emphasis

The purpose of pre-emphasis is to make the spectrum of the video signal uniformly distributed in low to high frequencies, thus making the video signal spectrum smoother and facilitating the accurate analysis of the subsequent spectrum and channel parameters. According to the characteristics of the captured original speech emotion, it is proposed to use a first-order digital filter to pre-emphasize the video signal, and the specific processing formula is:

$$H(z) = 1 - \mu z^{-1} \tag{12}$$

In Eq. (12), μ denotes the pre-emphasis coefficient, and the value range of μ is $0.95 \le \mu \le 1$.

The video signal after pre-emphasis can be expressed as:

$$Y(n) = X(n) - \mu * X(n-1) \tag{13}$$

In Eq. (13), $Y(n)$ and $X(n)$ denote the pre-emphasized video signal and the sampled value of the video signal at n moments, respectively.

2) Windowing and Framing

The role of subframing is to multiply the original video signal $s(n)$ by a movable window function of finite window length $w(n)$ to obtain the windowed video signal $s_w(n)$, the specific expression is:

$$s_w(n) = s(n) * w(n) \tag{14}$$

Among them, after the video signal is processed by frame splitting, the length of each frame is 12-32ms, and the frame shift is 0-0.6 times of the frame length.

The operation of adding windows to the video signal can effectively avoid the spectral leakage phenomenon during frame splitting. At present, the common window function is divided into Hamming window and rectangular window. They are denoted as:

$$w(n) = \begin{cases} 1, & 0 \le n \le (N-1) \\ 0, & n = else \end{cases} \tag{15}$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \le n \le (N-1) \\ 0, & n = else \end{cases} \tag{16}$$

Among the two window functions, the Hamming window is more adapted to analyze the frequency characteristics of short-time video signals. Therefore, the Hamming window is chosen as the window function. If the window length of Hamming window is denoted as N and the sampling period is denoted as T_s, the frequency resolution of the video signal can be expressed as:

$$\Delta f = \frac{1}{NT_s} \tag{17}$$

Based on Eq. (17), it can be seen that the window length increases with the decrease of frequency resolution with the same sampling period, and the two are inversely proportional to each other. For better video signal emotion recognition, it is proposed to set the window function length to 258 points and the frame shift to 129 points.

The fast Fourier transform FFT is performed on each frame of the short-time frequency domain signal obtained above, from which the frequency domain data are obtained; finally, the obtained FFT frequency domain data are spliced with the frequency domain features to perform the splicing operation, from which the preprocessed frequency domain features CFFD are obtained.

III. Performance validation and application analysis of emotion detection inference models

This chapter combines behavioral detection tests and performance comparison experiments to investigate the model's ability to reason about students' emotion detection. Based on the results of students' emotion detection reasoning, targeted psychological interventions are conducted to compare the changes in mental health before and after the interventions.

III. A. Model Testing and Evaluation

III. A. 1) Overall Behavior Detection and Identification Evaluation

The model is tested for behavioral detection using the dataset to verify the performance usability of the model. Figure 2 shows the ROC curve obtained by the model during the detection process.The ROC curve area of 0.9 implies that the model performs well on the behavioral detection task with high precision and accuracy, is able to identify most of the abnormal behaviors, and has a good ability to distinguish between normal and abnormal behaviors.

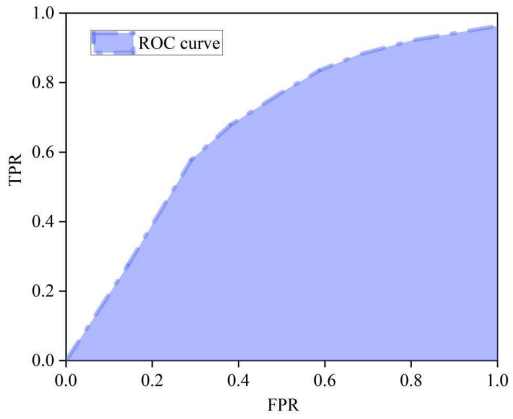

Figure 2: ROC curve

III. A. 2) Analysis of abnormal behavior detection results

Combined with the detection results, the students' behaviors in the dataset were classified according to their emotional states, which were classified into normal state (no emotion and small amplitude of emotional fluctuations) and abnormal state (the appearance of emotions such as anger, rage, fear, etc.). The video times of normal and abnormal states were further counted to analyze students' emotional performance at different times and on different objects during the stressful time frame of the end of the semester. Figure 3 shows the comparison of the number of times the two groups experienced mood swings on different time periods. Figure 4 is a comparison of the number of times that the two groups focused on entertainment information during mood swings in each time period. Comparing the time of the two groups' mood swings in Figure 3, it can be found that the abnormal group has far more abnormal mood swings than the normal group in the time range of 0:00-6:00, accounting for up to 4%. The rest of the time, the abnormal group and the normal group in the daily study and life of the number of abnormal mood swings is relatively the same, there is no excessive difference. Observing the number of times the two groups pay attention to entertainment information behavior during mood fluctuations, it can be seen that the number of times the abnormal group relieves their emotions by paying attention to entertainment information is much higher than that of the normal group when they experience large mood fluctuations such as anger and rage. Especially

from 0:00 a.m. to 3:00 a.m., the proportion of the abnormal group with emotional fluctuations and paying attention to entertainment information reaches up to 1.63%, which is much higher than that of the normal group, which is 0.25%. Combined with the results of the model detection, it is found that the abnormal student group may have a greater psychological crisis late at night and need to be intervened. It also verifies that the model can effectively recognize students' expressions and make emotional reasoning.

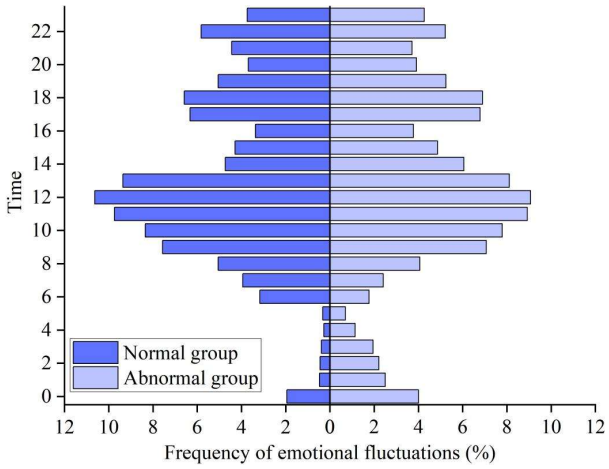

Figure 3: Comparison of the frequency of emotional fluctuations

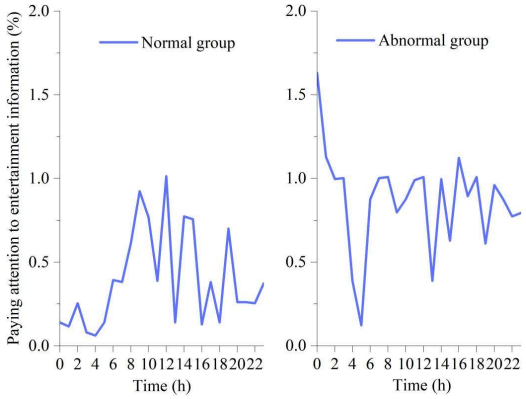

Figure 4: Comparison of number of times entertainment information is followed

III. B. Algorithm performance comparison

III. B. 1) Comparison of Accuracy of Emotional Reasoning

In order to assess the advantage of this paper's approach based on multimodal affective reasoning and self-attention framework on the accuracy of students' affective state reasoning, different classification algorithms (including support vector machine SVM, C4.5 and neural network BP) are selected to classify the processed video data of five students for affective classification, and to compare the classification accuracies of different algorithms. Table 1 shows the 10-fold cross-validated classification accuracy comparison. Where C4.5 algorithm uses J48 classifier in WEKA toolbox with default parameters.SVM uses LibSVM package with radial basis kernel function with parameters C and g, which are derived from lattice optimization.BP neural network uses MultilayerPerceptron classifier in WEKA toolbox with default parameters.

From the data in Table 1, it can be seen that the method in this paper can obtain relatively good inference prediction results, and the average classification accuracy of the five subjects is as high as 99.54%, which is higher than the accuracy of 97.17% for C4.5. While BP and SVM only obtain 94.54% and 83.03% classification accuracy. The method in this paper has a shorter modeling time and high accuracy, good tolerance to outliers and noise, and has a greater advantage in sentiment inference. For the C4.5 method, since the process of constructing the tree is based on the localized division criterion under each node, it is significantly affected by the randomness of the samples, which can easily lead to overlearning. Similarly, the BP neural network also has overlearning problem. The SVM is less accurate on this sample set, probably because the SVM is a 2-class classifier and the dataset is a multi-class sentiment dataset.

Table 1: Accuracy of several classifiers with 10-fold cross-validation

Subject	Article method	C4.5	BP	SVM
Subject 1	100.00%	99.10%	94.51%	79.46%
Subject 2	100.00%	98.54%	92.05%	86.17%
Subject 3	99.53%	96.21%	96.63%	84.41%
Subject 4	100.00%	96.03%	96.56%	79.67%
Subject 5	98.16%	95.95%	92.95%	85.42%
Average accuracy rate	99.54%	97.17%	94.54%	83.03%

III. B. 2) Speed Analysis of Emotional Reasoning

Fig. 5 shows the features of this paper's method that appear most frequently in the sentiment inference rules. The analysis of the inference rules obtained by this paper's method concludes that many attention features have no relation to sentiment classification, and the most used attention features are Beta wave to Theta wave absolute power ratio, Beta wave absolute power, Beta wave maximum power, skewness, Kolmgolov entropy, Alpha wave absolute power, variance, Shannon entropy, peak mean and Skewness, and the Beta wave to Theta wave absolute power ratio feature appeared much more often than other features in the rule set for the five subjects, reaching 2533, 2815, 3934, 3271, and 3694 times, respectively. In emotion inference, the method in this paper does not extract all the features of the video frames to reduce the amount of storage after preprocessing, thus obtaining a faster emotion inference speed and quickly recognizing the emotional state of students.

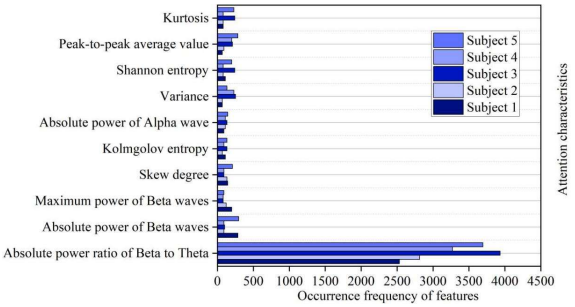

Figure 5: Feature that occurs most frequently in emotional reasoning rules

III. B. 3) Analysis of the effect of different levels of reasoning for the same emotion

The classification of fear emotion reasoning of subject 1 is used as an example to determine whether the method in this paper can successfully recognize different degrees of reasoning about the same emotion. Figure 6 shows the fear grading confusion matrix under the four classifications of subject 1. Among them, Class1-4 represent no fear state, mild fear state, moderate fear state, and severe fear state. It can be seen that the model of this paper, which integrates the multimodal emotion inference method and attention feature extraction, has a better inference effect for subtle emotion recognition. Among them, the recognition inference ability for the no-fear state is the strongest, reaching 0.85, and the most confusing ones are mild fear and moderate fear. Reasoning for different levels of the same emotion can quickly determine which stage of the emotional state the student is currently in, so as to deduce his or her possible mental health problems and provide targeted interventions.

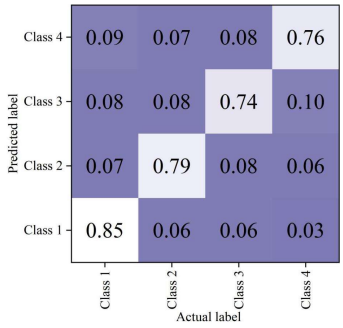

Figure 6: Confusion matrix of fear classification under the four categories

III. C. Comparison experiment before and after mental health intervention

After validating the emotion recognition reasoning effect of the model constructed in this paper, it was applied as a psychological intervention aid to students identified as having abnormal emotional states. After eight weeks of psychological intervention treatment, the results of the AQ and ATHQ scales and BAT tests of the subjects before and after four weeks of the start of the formal treatment were counted, and the results were compared using the nonparametric test of two paired samples. Table 2 shows the results of the comparison of scores on different scales before and after the psychological intervention treatment. Figure 7 shows the results of the comparison of the change in the means of the different scales before and after the treatment. Before and after treatment AQ-Fear decreased from about 59.70 to about 28.56, AQ-Anger decreased from about 16.46 to about 6.54, ATHQ decreased from about 38.42 to about 27.41, and BAT decreased from about 20.45 to about 7.26, and scores on the four clinical scales decreased significantly and reached statistical significance (p=0.00 <0.01, a significant difference). It indicates that using the model of this paper for students' emotion recognition reasoning and combining it with targeted psychological interventions can effectively reduce the frequency and degree of negative emotions such as anger and fear in students with abnormal emotional states and improve students' mental health after multiple interventions and treatments.

Table 2: Scale scores before and after psychological intervention treatment

Different scales	Before the intervention	After the intervention	P value
AQ-Fear	59.70±10.58	28.56±7.11	0.00
AQ-Anger	16.46±4.75	6.54±3.72	0.00
ATHQ	38.42±7.36	27.41±6.24	0.00
BAT	20.45±5.89	7.26±3.00	0.00

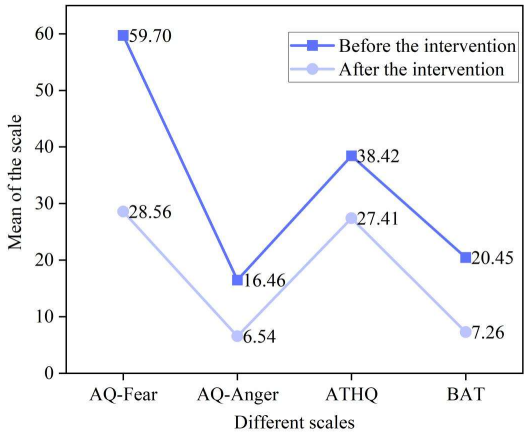

Figure 7: Changes in mean values of different scales before and after treatment

IV. Results

This paper constructs an abnormal behavior detection model for college students based on multimodal image recognition and affective reasoning to analyze the emotional state that students are in and provide targeted psychological interventions. The model has a ROC area of 0.9 in the process of behavior detection, which can effectively distinguish normal and abnormal behaviors of students. The model's correct rate of emotional reasoning was 99.54%, higher than the 97.17%, 94.54%, and 83.03% of the comparison methods, and it has excellent reasoning ability and reasoning speed and reasoning accuracy. The scores of AQ-Fear, AQ-Anger, ATHQ, and BAT4 scales after psychological intervention decreased to 28.56, 6.54, 27.41, and 7.26, respectively, and the difference between the scores of the scales before and after the intervention was significant (P<0.01). Using the model in this paper, students in abnormal emotional states can be detected and their mental health can be improved through timely intervention. In the future, the real-time processing efficiency of the model can be further optimized to ensure the timeliness of recognition reasoning and improve the effect of intervention.

Funding

This work was supported by Key Research Project of Guilin Normal College in 2023 (Project No. KYA202307): "Construction of a Mental Health Monitoring Framework System for Primary Schools".

References

[1] Barhalescu, M. (2025). THE IMPORTANCE OF MENTAL HEALTH IN MODERN SOCIETY. Journal of Marine Technology & Environment, 1.

[2] Drake, R. E., & Wallach, M. A. (2020). Employment is a critical mental health intervention. Epidemiology and Psychiatric Sciences, 29, e178.

[3] Fazel, M., Hoagwood, K., Stephan, S., & Ford, T. (2014). Mental health interventions in schools in high-income countries. The Lancet Psychiatry, 1(5), 377-387.

[4] Phillips, E. A., Gordeev, V. S., & Schreyögg, J. (2019). Effectiveness of occupational e-mental health interventions. Scandinavian journal of work, environment & health, 45(6), 560-576.

[5] Lindow, J. C., Hughes, J. L., South, C., Minhajuddin, A., Gutierrez, L., Bannister, E., ... & Byerly, M. J. (2020). The youth aware of mental health intervention: impact on help seeking, mental health knowledge, and stigma in US adolescents. Journal of Adolescent Health, 67(1), 101-107.

[6] Purtle, J., Nelson, K. L., Counts, N. Z., & Yudell, M. (2020). Population-based approaches to mental health: history, strategies, and evidence. Annual Review of Public Health, 41(1), 201-221.

[7] DeRubeis, R. J. (2019). The history, current status, and possible future of precision mental health. Behaviour Research and Therapy, 123, 103506.

[8] Kurhade, C. S., Jagannathan, A., Varambally, S., & Shivanna, S. (2022). Religion-based interventions for mental health disorders: A systematic review. Journal of Applied Consciousness Studies, 10(1), 20-33.

[9] Gonçalves, J. P., Lucchetti, G., Menezes, P. R., & Vallada, H. (2015). Religious and spiritual interventions in mental health care: a systematic review and meta-analysis of randomized controlled clinical trials. Psychological medicine, 45(14), 2937-2949.

[10] Cook, C. C. (2020). Spirituality, religion & mental health: exploring the boundaries. Mental Health, Religion & Culture, 23(5), 363-374.

[11] Moss, J. (2017). Aristotle's Ethical Psychology. The Cambridge Companion to Ancient Ethics, 124-142.

[12] Soklaridis, S., Lin, E., Lalani, Y., Rodak, T., & Sockalingam, S. (2020). Mental health interventions and supports during COVID-19 and other medical pandemics: A rapid systematic review of the evidence. General hospital psychiatry, 66, 133-146.

[13] Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. Journal of medical Internet research, 21(7), e12869.

[14] Jung, S. J., & Jun, J. Y. (2020). Mental health and psychological intervention amid COVID-19 outbreak: perspectives from South Korea. Yonsei medical journal, 61(4), 271-272.

[15] Van Agteren, J., Iasiello, M., Lo, L., Bartholomaeus, J., Kopsaftis, Z., Carey, M., & Kyrios, M. (2021). A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. Nature human behaviour, 5(5), 631-652.

[16] Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., ... & Baumeister, H. (2018). Internet-and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. European Psychologist.

[17] Bidargaddi, N., Schrader, G., Klasnja, P., Licinio, J., & Murphy, S. (2020). Designing m-Health interventions for precision mental health support. Translational psychiatry, 10(1), 222.

[18] Graham, A. K., Lattie, E. G., Powell, B. J., Lyon, A. R., Smith, J. D., Schueller, S. M., ... & Mohr, D. C. (2020). Implementation strategies for digital mental health interventions in health care settings. American Psychologist, 75(8), 1080.

[19] Javanbakht, A., & Alberini, C. M. (2019). Neurobiological models of psychotherapy. Frontiers in behavioral neuroscience, 13, 144.

[20] Wittchen, H. U., Härtling, S., & Hoyer, J. (2015). Psychotherapy and mental health as a psychological science discipline. Verhaltenstherapie, 25(2), 98-109.

[21] Russell, E., & Patrick, K. (2018). Mental health needs our attention. CMAJ, 190(2), E34-E34.

[22] Kohrt, B. A., Ottman, K., Panter-Brick, C., Konner, M., & Patel, V. (2020). Why we heal: The evolution of psychological healing and implications for global mental health. Clinical Psychology Review, 82, 101920.

[23] Zhang, X., Li, J., Xie, F., Chen, X., Xu, W., & Hudson, N. W. (2022). The relationship between adult attachment and mental health: A meta-analysis. Journal of Personality and Social Psychology, 123(5), 1089.

[24] Aguilera, A. (2015). Digital technology and mental health interventions: Opportunities and challenges. Arbor, 191(771), a210.

[25] De Witte, N. A., Joris, S., Van Assche, E., & Van Daele, T. (2021). Technological and digital interventions for mental health and wellbeing: an overview of systematic reviews. Frontiers in digital health, 3, 754337.

[26] Berle, D., Moulds, M. L., Starcevic, V., Milicevic, D., Hannan, A., Dale, E., ... & Brakoulias, V. (2016). Does emotional reasoning change during cognitive behavioural therapy for anxiety?. Cognitive behaviour therapy, 45(2), 123-135.