International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 5, October, 2024, pp. 2549-2557
Journal homepage: http://iieta.org/journals/ts

Automated Assessment of Student Mental Health Through Image Processing Technology

Fuwei Li

Intelligent Manufacturing College, Zibo Vocational Institute, Zibo 255300, China

Corresponding Author Email: 11136@zbvc.edu.cn

https://doi.org/10.18280/ts.410528

Received: 6 May 2024
Revised: 12 August 2024
Accepted: 23 September 2024
Available online: 31 October 2024

Keywords:
vocational education, mental health, image processing, temporal action detection, micro-expression recognition

ABSTRACT

Students in vocational education face high levels of academic and employment pressure, significantly impacting their mental health, academic performance, and career development. Traditional mental health assessment methods, relying on questionnaires or interviews, often lag in timeliness and are limited in their ability to reflect real-time changes in students' mental states. Recently, the application of image processing technology in mental health monitoring has gained attention, as it allows for faster, more accurate detection of emotional changes by capturing features like facial expressions and postural behaviors. However, existing approaches often focus on singular emotional features or are limited to static images, failing to leverage the combined potential of dynamic information and subtle facial expressions. This paper proposes a dual-analysis method based on temporal action detection and micro-expression recognition to comprehensively assess students' body language and emotional changes. This approach enables accurate monitoring of mental health status, providing technical support for a psychological support system tailored to vocational education students.

1. INTRODUCTION

With the rapid development of vocational education, student mental health issues have become a focus of attention for educators and researchers [1-3]. Vocational education students, influenced by the pressures of academics, skill improvement, and employment, are more susceptible to changes in their mental state, and psychological problems within this group often exhibit concealment and complexity [4, 5]. Traditional mental health assessment methods often rely on periodic surveys or counseling sessions, but these methods appear relatively passive and lagging in the face of the large-scale and dynamically changing psychological needs of students. Therefore, exploring emerging technologies, particularly image processing technology, to monitor and assess the mental health status of vocational education students in real time is of significant research value [6-8].

In relevant research on the automated assessment of mental health, the effectiveness of using image processing technology to recognize students' emotional changes and behavioral characteristics has been preliminarily verified [9-11]. These technological methods not only address the shortcomings of traditional methods but also allow for a more detailed analysis of subtle changes in students' mental states [12-15]. By capturing psychological changes from students' facial expressions, postures, and other behavioral features through automated systems, the efficiency of assessment is enhanced, providing educational institutions with more precise bases for psychological interventions. This image processing-based mental health monitoring approach has considerable application prospects and contributes to the establishment of a more targeted psychological support and guidance system.

However, current research methods still have certain limitations. On the one hand, most existing methods focus on recognizing a single expression or action, without fully utilizing the combined information of students' emotional states and body language [16-20]. On the other hand, current methods largely rely on static image analysis, unable to capture dynamic psychological changes in time series [21-23]. Additionally, the accuracy of emotion recognition is also affected by the complexity of micro-expressions and the subtle differences in body language, resulting in a high error rate. Therefore, a more comprehensive, accurate, and dynamically adaptive assessment method is urgently needed to meet the diverse needs of practical applications.

To address this, this paper proposes a novel automated assessment method for student mental health based on temporal action detection and micro-expression recognition technology. The research is primarily divided into two parts: the first part utilizes temporal action detection technology to analyze students' body language and capture the trend of emotional state changes; the second part uses micro-expression recognition to perform real-time analysis of students' subtle emotional changes. This method not only achieves dynamic monitoring of students' emotional changes but also enhances the accuracy and applicability of the assessment through multimodal information fusion, thus providing new technical support for the mental health service system in vocational education.

2. STUDENT BODY LANGUAGE ANALYSIS BASED ON TEMPORAL ACTION DETECTION

As a form of non-verbal communication, body language can reflect students' psychological and emotional states, such as anxiety, stress, or depression. In educational settings, students' mental health significantly impacts their learning abilities and overall growth. However, due to the hidden and complex nature of mental health issues, traditional survey or interview methods often fail to capture students' psychological fluctuations in a real-time and objective manner. The body language analysis based on temporal action detection proposed in this paper provides a novel, non-intrusive technical approach for the automated assessment of student mental health. The algorithm structure is shown in Figure 1.

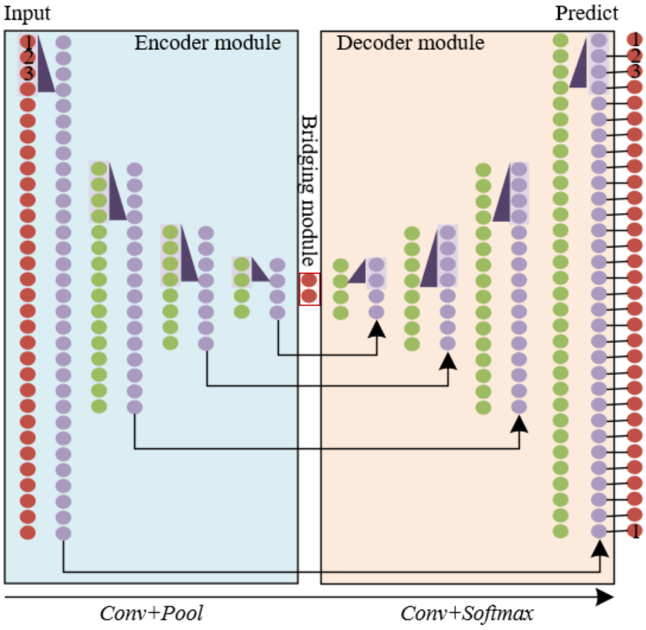

Figure 1. Structure of student body language analysis algorithm based on temporal action detection

In the context of automated mental health assessment, students may exhibit different emotional responses at various points in a class, such as focus at the beginning, fatigue midway, or anxiety towards the end. Therefore, this paper constructs an encoder-decoder temporal convolutional network (TCN) for analyzing student body language. In the network model, the encoder module extracts deep features from the input video frames through convolutional layers and max-pooling layers. The bridging module plays an important role between the encoder and decoder modules, performing average pooling on the columns of the encoded features to further integrate the global information of the temporal data. In the decoder module, the up-sampling layers and convolutional layers map the features output from the encoder back to specific actions in the time sequence. By concatenating outputs from each layer of the encoder module, the decoder can progressively restore this temporal information to the initial action sequence, thereby capturing detailed changes in actions. A key feature of the encoder-decoder TCN in this paper is the simultaneous calculation of all frame features rather than processing them frame by frame, which offers distinct advantages in student mental health assessment.

In the task of assessing student mental health, the temporal information in videos of psychological interviews is crucial. The encoder module not only needs to extract action features from each frame but also to retain the temporal dependencies between frames, so that the subsequent decoding process can accurately evaluate mental health based on the trend in action features. In the encoder module, we first extract feature vectors A_s from each frame of the input video, where D_0 is the dimension of each frame's feature and s denotes the frame sequence number. For a complete video sequence, its feature vectors are arranged in the order of frames to form a video feature matrix. The row count of this matrix corresponds to the number of video frames, while the column count is the feature dimension D_0. To enable the network to process videos of varying lengths, we pad the feature matrix to unify the row count. Specifically, using the maximum frame count d_{MAX} in the dataset as a benchmark, we fill in the missing portions of video matrices with fewer frames than d_{MAX} with -1 to create input feature matrices of uniform dimension. This processing method ensures that feature matrices from videos of different lengths can be uniformly processed by the network.

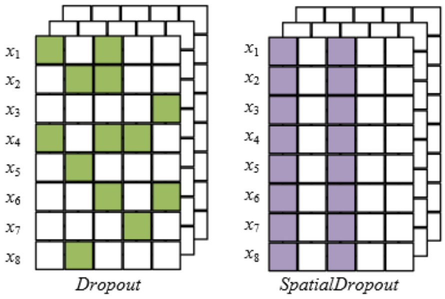

Figure 2. Diagram of *Spatial Dropout*

In the encoder module, each layer $R^{(u-1)}$ from the first to the fourth uses 64, 96, 128, and 160 one-dimensional convolutional kernels, respectively, each with a kernel size of 25 to capture local patterns within video frame features. The convolution operation in layer u is determined by specific weights Q and biases y, transforming video frame features into high-dimensional features suited for subsequent processing. Through multi-level feature abstraction, the convolution layers preserve latent mental health information in student action features, such as expressions of tension, unease, and attention state. To prevent model overfitting during training, a Spatial Dropout layer is added after each convolutional layer. Unlike standard Dropout, Spatial Dropout randomly zeroes out regions in the feature matrix, causing the model to discard a portion of features in each training iteration, thus enhancing generalization ability and preventing over-reliance on localized features, Figure 2 gives a diagram of Spatial Dropout. The dropout probability in this model is set to 0.3, meaning that 30% of the features are randomly dropped during each forward pass. This approach ensures that the model does not focus exclusively on specific actions or regions but learns and evaluates based on broader action patterns, effectively reducing the performance gap between the training and test sets.

In addition, each convolutional layer is followed by an activation function layer and a max-pooling layer. The activation function performs nonlinear transformation of features, enhancing the network's expressive power, while the max-pooling layer, with a stride of 2, reduces the dimensionality of feature vectors by half. This dimensionality reduction removes redundant information while retaining the core action features, increasing the network's robustness across different student behavior characteristics. Max-pooling also reduces computational costs, making the model more efficient for temporal analysis of video sequences. Specifically, assuming the convolution symbol is denoted by $*$, the activation function by d, and max-pooling by $MAXPL$, the signal $R^{(u)}$ in the u-th layer from the signal in the $R^{(u\text{-}1)}$-th layer is computed as:

$$R^{(u)} = MAXPL\left(d\left(Q * R^{(u-1)} + y\right)\right) \tag{1}$$

The bridging module's input is the output feature matrix of the encoder module, denoted as $X_{L\times V}$, where L represents the temporal length of features and V is the feature dimension at each time point. First, the bridging module performs global average pooling (GAP) on each column of the feature matrix, calculating an average value for each column, resulting in V global averages. These averages retain important information from the encoder module features while reducing noise, providing more representative feature expressions. In student body language analysis, these global averages capture the core information of each feature, providing a more robust feature input for the model's automated mental health assessment. Next, each global average is expanded vertically to form a column vector of dimension $L\times1$, resulting in V column vectors. These V column vectors are concatenated to the encoder module's output feature matrix, expanding the feature matrix dimensions from $X_{L\times V}$ to $X_{L\times 2V}$. This expansion enhances the feature matrix capacity, allowing the decoder module to delve into temporal information across richer feature dimensions. In mental health status analysis, this step allows the model to integrate both original and global features, capturing body language details more comprehensively and identifying latent psychological signals. Figure 3 provides an illustration of the feature concatenation.

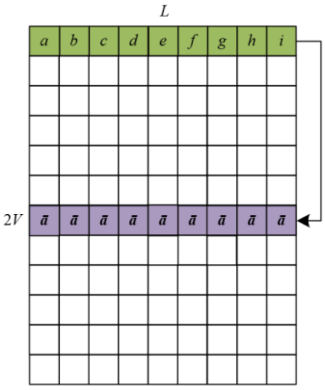

Figure 3. Diagram of feature concatenation

The decoder module consists of four layers, each comprising an up-sampling layer, a convolutional layer, and an activation function layer. This design effectively restores the spatial information of the input features while enhancing the model's ability to recognize various actions. Each layer in the decoder, $F^{(u)}$, is indexed from 4 to 1, reflecting symmetry with the encoder module. During decoding, the up-sampling layer plays a critical role by doubling the dimensions of the feature map, restoring spatial information lost in the encoding process due to pooling. This contrasts with the max-pooling layers in the encoder, which capture higher-level abstract features, while the up-sampling layers aim to retain the structural integrity of the original input image in the decoding phase, thus allowing for more accurate recognition of students' body language and their corresponding psychological states. Each convolutional layer uses a kernel size of 25, with the number of convolutional kernels gradually decreasing from 160 to 64, ensuring detailed feature capture and effectively modeling transitions between different action classes. The multi-level feature extraction capabilities of the convolutional layers enable the decoder to identify more complex body language patterns, which are often critical indicators of a student's psychological state.

Additionally, given that students' behaviors are influenced by various factors, the model requires adaptability to accurately assess mental health across different contexts. To achieve this, the decoder module applies a Spatial Dropout strategy, randomly dropping 30% of the feature regions. The decoder module utilizes a cross-entropy loss function to evaluate discrepancies between the predicted results and the actual labels. By calculating the cross-entropy between the output probability distribution and the actual labels, the network continuously adjusts parameters to improve the accuracy of student body language recognition. When the network detects the maximum probability for a particular micro-action, it identifies that the body language in the image corresponds to that action, which is crucial for the automated assessment of students' mental health.

Let v be the sample count and l the class count. The formula for the multi-class cross-entropy loss function is given by:

$$M = \text{-}\sum_{u=1}^{v} \widehat{b}_{u1} \log b_{u1} + \widehat{b}_{u2} \log b_{u2} + \ldots + \widehat{b}_{ul} \log b_{ul} \tag{2}$$

Finally, through a fully connected layer, the predicted output $\hat{B}_s = soft\ max(IF^{(1)}{}_s + z)$ is obtained. For image s, the vector $\hat{B}_s \in [0,1]^Z$ represents a vectorized Z-dimensional vector, where Z is the total number of micro-actions present in the student's psychological interview video, corresponding to the Z positions in the vector.

3. ANALYSIS OF STUDENTS' EMOTIONAL CHANGES BASED ON MICRO-EXPRESSION RECOGNITION

Analyzing emotional changes can help identify potential mental health issues. Studies indicate that prolonged negative emotional states may lead to more serious mental health concerns, such as depression and anxiety. Through continuous monitoring and analysis of students' micro-expressions, it is possible to identify students with significant emotional fluctuations and provide timely intervention and psychological support. For example, emotional change analysis might reveal that certain students exhibit persistent negative emotions towards specific subjects, which can not only assist teachers in adjusting their teaching strategies but also provide mental health counselors with tailored intervention insights. This study introduces a *VGG16-SE-TA-LSTM*-based micro-expression recognition algorithm for analyzing students' emotional changes. The structure of the micro-expression recognition algorithm is shown in Figure 4.

In the micro-expression recognition task, quick and effective feature extraction is crucial. When students face stress or anxiety, changes in their facial expressions are subtle, and capturing these changes requires a network that can quickly and accurately learn meaningful features. The algorithm leverages an improved VGG16 network structure to enhance the accuracy and robustness of emotional change analysis. The modified VGG16 network adds batch normalization layers after each convolutional layer, accelerating network convergence while preventing gradient vanishing. Given the limited number of micro-expression samples, which can lead to overfitting, the modified VGG16 network introduces Dropout layers after each pooling layer, reducing inter-feature dependency by randomly ignoring certain neurons during training and thus effectively lowering the risk of overfitting. Since spatial features of micro-expressions are critical for emotional expression, the original VGG16 network's three fully connected layers result in a large number of model parameters and high computational costs. To address this, the improved VGG16 network replaces these fully connected layers with a GAP layer. The GAP layer significantly reduces the model parameters by averaging the feature maps globally, while retaining essential spatial information.

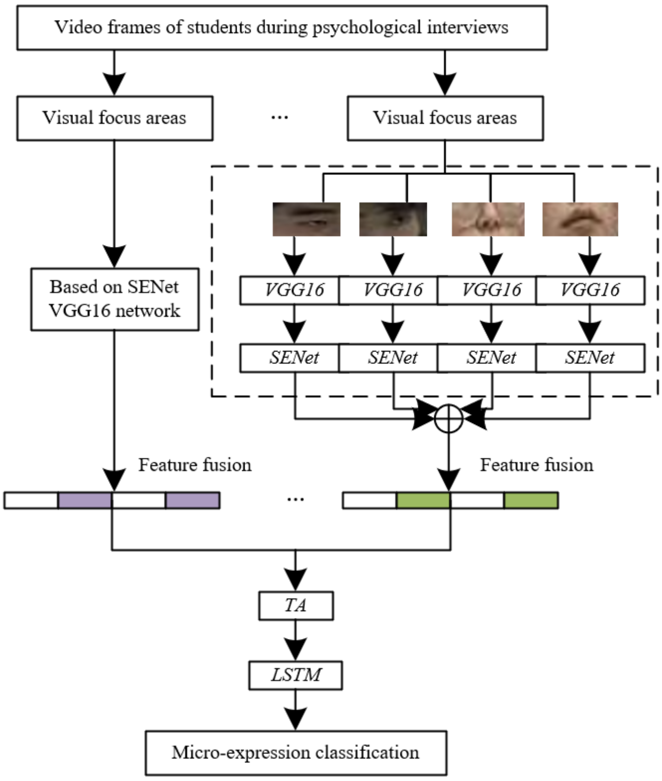

Figure 4. Micro-expression recognition algorithm structure

The Long Short-Term Memory (LSTM) in the micro-expression recognition algorithm effectively processes temporal data, allowing the extraction of time-dependent features in micro-expression video sequences. The LSTM network consists of several gates, including the input gate, forget gate, and output gate, which work together to control information flow and state updates. In analyzing students' emotional changes, the forget gate determines which prior states are no longer relevant and can be disregarded, maintaining the model's simplicity and effectiveness. The input gate selectively updates the current state variable at each moment. The memory cell z_{s-1} at time s in the LSTM holds essential temporal information. During micro-expression analysis, the input gate weights regulate the introduction of current frame features, ensuring that new inputs significantly impact the model. Meanwhile, the forget gate allows the model to discard irrelevant historical information, increasing sensitivity to current emotional states. This mechanism enables LSTM to handle long-sequence data while capturing continuity and trends in emotions over time. Suppose the current input at time s is a_s and previous hidden state is g_{s-1}. The forget gate's weight and bias are denoted by q_d and y_d, and the sigmoid function by δ. When the weight d_s=1, the forget gate allows the LSTM to retain all information from the previous state variable z_{s-1}; when d_s=0, it ignores z_{s-1}. The weight definition for the forget gate is as follows:

$$d_s = \delta\left(q_d * \left[g_{s-1}, a_s\right] + y_d\right) \tag{3}$$

Suppose the input gate parameters are denoted by q_z and y_z and the activation function by tanh; the input at time s is a_s, with the specific input gate calculation as:

$$\tilde{z}_s = \tanh\left(q_z * \left[g_{s-1}, a_s\right] + y_z\right) \tag{4}$$

Let the input gate parameters be q_u and y_u, and the control weight definition as:

$$u_s = \delta\left(q_u * \left[g_{s-1}, a_s\right] + y_u\right) \tag{5}$$

The update formula for the state variable z_s at the current time is:

$$z_s = d_s z_{s-1} + u_s \tilde{z}_s \tag{6}$$

Assuming the output gate parameters are q_p and y_p, the output gate control variable p_s can be calculated as:

$$p_s = \delta\left(q_p \left[g_{s-1}, a_s\right] + y_p\right) \tag{7}$$

The LSTM output can be obtained using the following formula:

$$g_s = p_s \cdot \tanh\left(z_s\right) \tag{8}$$

In students' learning processes, emotional states and mental health often manifest through micro-expressions, and these subtle changes are essential for teachers and mental health experts to understand students' emotional states promptly. To ensure efficiency in feature selection and information processing while maintaining sensitivity to emotional shifts within complex temporal data, the channel and temporal attention mechanisms are introduced in this study.

The channel attention mechanism (SE module) assigns appropriate weights to crucial feature channels by analyzing correlations among them, allowing the model to highlight features essential for emotion recognition while suppressing irrelevant or distracting features. In the context of student emotional changes, specific micro-expression features, such as eyebrow movement and mouth corner lifts, may more accurately reveal the student's genuine emotional state. By introducing SENet modules after each convolutional layer of the VGG16, the model effectively filters and weights the 512 feature channels, ensuring that significant features are emphasized in subsequent emotion analysis, thereby improving emotion recognition accuracy. For example, when students appear confused during class, specific micro-expression changes might occur, and the channel attention mechanism can help the model focus on these variations, thereby identifying negative emotions promptly.

The SENet module utilizes the Squeeze-and-Excitation (SE) mechanism to recalibrate convolutional features. In the Squeeze part, GAP is first applied to the feature map output from the VGG16 network, compressing each feature channel's spatial information into a global descriptor, forming a vector of length *Z*, where *Z* is the number of channels in the feature map. Let *I* represent the output tensor from the previous convolutional structure D_{se}, which undergoes the Squeeze operation as detailed by the formula:

$$c_z = D_{tw}\left(i_z\right) = \frac{1}{Q \times G}\sum_{q=1}^{Q}\sum_{g=1}^{G} i_z\left(q,g\right) \tag{9}$$

Next, the Excitation part processes this global feature through two fully connected layers. The first fully connected layer compresses the *Z*-dimensional vector to *Z*/*e* dimensions, where *e* is a scaling parameter. This step reduces the number of parameters and introduces a non-linear transformation, enhancing the model's representation capacity. The second fully connected layer restores the compressed feature vector to its original *Z*-dimension and generates a weight vector *t* in the range of 0 to 1 using a Sigmoid activation function. This weight vector *t* represents the importance of each channel and is used to weight the initial feature map, highlighting important features and suppressing less relevant ones. Assuming the weights of the two fully connected layers are represented by Q_1 and Q_2, and the ReLU and sigmoid functions are represented by σ and δ, the calculation can be expressed as follows:

$$t = D_{ra}\left(c,Q\right) = \delta\left(h\left(c,Q\right)\right) = \delta\left(Q_2\sigma\left(Q_1 c\right)\right) \tag{10}$$

The channel weight *t* is then multiplied with the original tensor *I* on a channel basis, as follows:

$$\tilde{A} = D_{SC}\left(I,t\right) \tag{11}$$

The introduced Temporal Attention (TA) mechanism assigns weights to different frames, concentrating attention on key frames that contain crucial feature information, thereby enhancing overall recognition accuracy. Specifically, in the VGG16-SE-TA-LSTM-based micro-expression recognition algorithm, the VGG16 network initially extracts feature sequences $D(a)=[d(a_1), d(a_2),...d(a_S)]$ from the input image frame sequence, where *S* is the sequence length. These features are then input into the temporal attention module, which computes the hidden state at each time step $G=[g_1, g_2,...,g_s]$, with g_s representing the hidden vector at time step *t*. Suppose the network weights of the fully connected layer are represented by Q_x, and the hidden vector of the sequence is represented by g_s and g_u. The following formula calculates the relevance among frames in the micro-expression sequence:

$$DF\left(g_s,g_u\right) = g_s^S Q_x g_u \tag{12}$$

To evaluate each frame's importance within the entire sequence, the TA mechanism uses a fully connected layer to calculate frame-to-frame correlations in the micro-expression sequence. At time step *s*, β_s represents the influence of the time sequence on the time step vector g_s. Specifically, β_s is calculated by applying the softmax function to the hidden vector g_s, generating a set of weights that reflect each time step's importance to the current time step *s*. The influence of the *u*-th time step on predicting the current time step *s*, $x_{s,u}$, can be calculated as follows:

$$x_{s,u} = \text{softmax}\left(DF\left(g_s,g_u\right)\right) = \frac{\exp\left(DF\left(g_s,g_u\right)\right)}{\sum_{u=1}^{S}\exp\left(DF\left(g_s,g_u\right)\right)} \tag{13}$$

Finally, a weighted sum yields the attention weight β_s for each frame image:

$$x_s = \sum_{u=1}^{S} x_{s,u} g_u \tag{14}$$

Detailed execution steps of the VGG16-SE-TA-LSTM-based micro-expression recognition algorithm:

(1) Preprocessing video frames is one of the fundamental steps. Each frame image is segmented into four key regions: left eye, right eye, nose base, and lips. These areas are where micro-expressions are most evident and frequent.

(2) The modified VGG16 network extracts spatial features from each key region. Compared to the original VGG16, this enhanced version is better suited for micro-expression recognition tasks. Modifications include introducing a channel attention mechanism. Through adaptive recalibration of the feature channels' weights, this module strengthens the representation of essential features. Specifically, the SE module first captures each channel's global information via GAP, then produces inter-channel weights through two fully connected layers. These weights act on the initial feature map, highlighting critical channel information.

(3) After obtaining the spatial features of each frame image, the features from the four key regions are combined to form the comprehensive feature of each frame. These composite features are then input into the temporal and channel attention modules. The TA module assigns corresponding weights to various video frames, focusing on key frames with significant micro-expression features. The SE module further assigns weights to different feature channels, ensuring the model emphasizes channels relevant to micro-expression recognition. This dynamic adjustment of attention to each frame and channel effectively suppresses irrelevant information and improves recognition accuracy.

(4) The network, which integrates spatial and temporal features, is then processed by the LSTM network. LSTM captures long- and short-term dependencies within the feature sequence, enabling a better understanding of the dynamic changes in micro-expressions. The trained VGG16-SE-TA-LSTM network efficiently identifies micro-expressions, supporting the analysis of student emotional changes.

4. EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, incorporating multiple modules significantly improves the performance of the micro-motion detection system. The baseline ResNet50 model achieves an mAP@0.5 of 73.2% without any additional modules. Adding the Spatial Dropout module increases detection performance to 74.5%. The performance then slightly improves to 75.1%

after including the connected feature fusion module. Introducing the transition module and max interpolation upsampling further raises the mAP to 75.6% and 75.9%, respectively. Notably, using mean interpolation upsampling results in the most significant improvement, reaching 76.5%. Finally, combining multiple modules for comprehensive micro-expression detection and action recognition yields a mAP of 78.1%, which further rises to 79.6% with mean interpolation upsampling. This analysis indicates that module combinations significantly enhance the micro-motion detection system's performance, particularly with the inclusion of Spatial Dropout and mean interpolation upsampling, which improve the model's generalization and fine-detail capture.

Table 1. Impact of different modules on micro-motion detection results

***ResNet*50**	**Spatial Dropout**	**Connected Feature Fusion**	**Bridging Module**	**Max Interpolation Upsampling**	**Mean Interpolation Upsampling**	***mAP*@0.5**
	√					73.2%
√	√					74.5%
√	√	√				75.1%
√	√		√			75.6%
√	√			√		75.9%
√	√				√	76.5%
√	√	√	√	√		78.1%
√	√	√	√		√	79.6%

Table 2. Comparison of different micro-behavior detection algorithms on micro-motion detection

Method	**Acc**	**mAP@0.5**
Kernel *SVM*	63.8	82.3%
3*D-CNN*	75.4	68.9%
EfficientNet	78.5	75.2%
The Proposed Algorithm	82.3	77.9%

From the results in Table 2, it is clear that different micro-behavior detection algorithms show varying levels of performance. While the Kernel Support Vector Machine (SVM) algorithm achieves a high mAP@0.5 of 82.3%, its accuracy (Acc) is low at 63.8%. The 3D-Convolutional Neural Network (CNN) model shows an improvement in accuracy to 75.4%, but its mAP@0.5 drops to 68.9%. EfficientNet performs well on both metrics, achieving an accuracy of 78.5% and an mAP@0.5 of 75.2%. However, the proposed algorithm achieves the best results on both metrics, with an accuracy of 82.3% and an mAP@0.5 of 77.9%. This indicates that the proposed algorithm, based on temporal action detection and micro-expression recognition, excels in capturing students' body language and subtle emotional changes, as shown by its high accuracy and mAP@0.5. The Kernel SVM's strong mAP@0.5 is offset by low accuracy, suggesting a lack of comprehensive detection ability. The 3D-CNN struggles to capture temporal information, leading to lower mAP@0.5. Although EfficientNet achieves a balanced performance across metrics, it still falls short of the proposed algorithm's overall efficacy.

Table 3. Recognition performance of different micro-expression recognition algorithms across four datasets

Metric **Model**	**ATNET**	**GAN**	**EfficientNet**	**H-SVM**	**CNN-LSTM**	**The Proposed Algorithm**
Unweighted F1 (1)	0.6215	-	0.5896	0.6124	0.5369	0.7156
Unweighted Avg Recall (1)	0.6241	-	0.5869	0.6231	0.5547	0.7125
Unweighted F1 (2)	0.5481	0.6235	0.5412	0.5389	0.5478	0.6895
Unweighted Avg Recall (2)	0.5326	0.6215	0.5348	0.6215	0.5589	0.7256
Unweighted F1 (3)	0.7895	0.6652	0.7584	0.7262	0.4789	0.6626
Unweighted Avg Recall (3)	0.7451	0.6452	0.7415	0.7412	0.5123	0.6689
Unweighted F1 (4)	0.4879	0.6125	0.4326	0.4256	0.5148	0.7156
Unweighted Avg Recall (4)	0.4758	0.5896	0.4751	0.4223	0.5125	0.7156

As shown in Table 3, the performance of different micro-expression recognition algorithms varies significantly across four datasets. The proposed algorithm outperforms others in unweighted F1 and unweighted average recall, especially on the CASME II dataset (unweighted F1 of 0.7156 and unweighted average recall of 0.7125) and the AffectNet dataset (unweighted F1 and unweighted average recall both at 0.7156). By comparison, ATNET performs well on the CASME II and EMO-DB datasets (unweighted F1 of 0.6215 and 0.7895, respectively) but falls short on other datasets. Generative Adversarial Network (GAN) stands out on the SMIC dataset (unweighted F1 of 0.6235 and unweighted average recall of 0.6215) but underperforms on other datasets. EfficientNet and H-SVM deliver relatively balanced performance across multiple datasets, though they do not achieve top results in unweighted F1 or unweighted average recall. CNN-LSTM performs well on some datasets but is overall less effective than other advanced algorithms. In conclusion, the proposed algorithm, based on temporal action detection and micro-expression recognition, consistently performs better across different datasets than other common algorithms, particularly in capturing subtle emotional changes and analyzing trends in emotional states. Although ATNET and GAN perform well on specific datasets, they lack stability and consistency across datasets. While EfficientNet and H-SVM are balanced in performance, they fail to reach optimal scores in unweighted F1 and unweighted average recall.

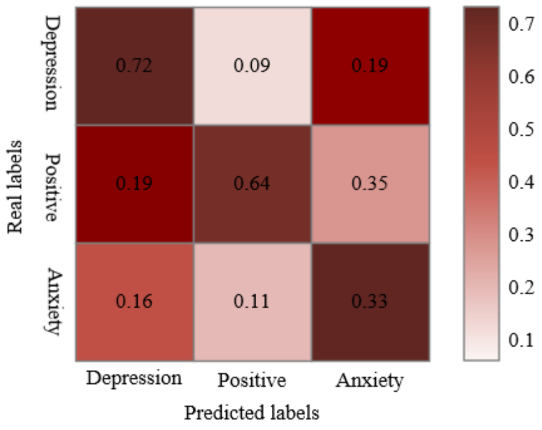

Figure 5. Confusion matrix of the proposed micro-expression recognition algorithm

The experimental results presented above demonstrate the significant effectiveness of the micro-expression recognition achieved using the VGG16-SE-TA-LSTM model on the fused dataset. The analysis of the confusion matrix reveals notable differences in the recognition accuracy among various micro-expression categories. In the confusion matrix depicted in Figure 5, the "anxiety" category achieves the highest recognition accuracy at 0.73, indicating the model's strong reliability in detecting "anxiety" emotions. The "depression" category also shows a relatively high recognition accuracy, although it is lower than that of "anxiety." In contrast, the "positive" category exhibits a lower recognition accuracy, potentially due to the subtler facial muscle changes associated with positive emotions, which are harder to capture. The experimental results indicate that the combined use of temporal action detection technology and micro-expression recognition techniques has high effectiveness in assessing student mental health. Notably, the VGG16-SE-TA-LSTM model excels in capturing and analyzing "anxiety" micro-expressions, which is likely due to the more pronounced facial muscle changes associated with this emotion, allowing the model to extract more valuable information. Consequently, this model is highly effective in detecting more apparent negative emotions. However, the lower recognition rate for "positive" emotions suggests the need for further model improvements to enhance sensitivity to subtle facial changes. Overall, the proposed method shows great potential for real-time analysis of students' emotional changes and mental health status, particularly for the timely recognition and intervention of negative emotions.

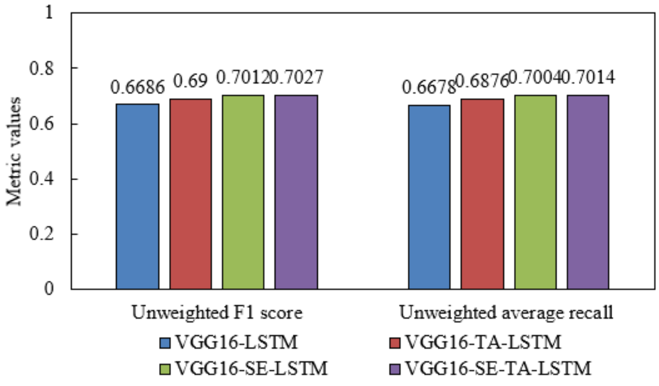

Figure 6. Comparison of recognition results of different micro-expression recognition algorithms

The results shown in Figure 6 illustrate significant differences in the performance of various micro-expression recognition algorithms based on the unweighted F1 score and unweighted average recall. The VGG16-SE-TA-LSTM algorithm outperforms all others on both metrics, achieving scores of 0.7027 for unweighted F1 and 0.7014 for unweighted average recall, slightly higher than the VGG16-SE-LSTM, which has an unweighted F1 score of 0.7012 and an unweighted average recall of 0.7004. This demonstrates that the LSTM model, which integrates the SE module and the TA mechanism, exhibits strong performance in capturing micro-expression features. In comparison, while the VGG16-TA-LSTM (unweighted F1 of 0.69 and unweighted average recall of 0.6876) and VGG16-LSTM (unweighted F1 of 0.6686 and unweighted average recall of 0.6678) algorithms also show good performance, they fall short of the top scores. Thus, it can be concluded that the algorithm proposed in this paper, based on temporal action detection and micro-expression recognition technology, excels in micro-expression recognition tasks, particularly due to the introduction of the SE module and TA mechanism, which significantly enhance the model's recognition capabilities. The VGG16-SE-TA-LSTM model effectively integrates temporal information and attention mechanisms, allowing for more effective capture and processing of subtle changes in micro-expressions. By combining temporal action detection and micro-expression recognition techniques, the proposed algorithm not only improves the precision and stability of micro-expression recognition but also demonstrates its superiority and potential in practical applications, providing more accurate and real-time emotional analysis for student mental health assessments.

5. CONCLUSION

This paper proposed a novel automated assessment method for student mental health based on temporal action detection and micro-expression recognition technology. The research is divided into two main parts: 1) Utilizing temporal action detection technology to analyze students' body language and capture trends in their emotional states. By analyzing students' physical movements in classrooms, during breaks, and in other activity settings, the study assesses emotional fluctuations and psychological states. Experimental results indicate that different modules significantly impact micro-motion detection outcomes, particularly in terms of detection accuracy and response speed. 2) Applying advanced micro-expression recognition technology for real-time analysis of subtle emotional changes in students. This study compared the impact of various micro-behavior detection algorithms on micro-motion detection results, revealing that certain algorithms excel in specific contexts. Furthermore, this paper compared different micro-expression recognition algorithms across four datasets, demonstrating that the proposed micro-expression recognition algorithm exhibited high recognition accuracy and robustness. Confusion matrix analysis further validated the algorithm's ability to recognize different emotional categories.

The proposed method effectively captured students' emotional changes and provided real-time, dynamic mental health assessments. This holds significant importance for educational institutions in understanding and intervening in students' mental health issues promptly. Additionally, the method's multidimensional data analysis can provide scientific support for psychological counseling and personalized education. Despite demonstrating high accuracy and stability in experiments, the method has limitations in practical applications. For instance, variations in body language and micro-expressions among students from different subjects and age groups may impact the accuracy of assessments. Furthermore, issues related to privacy protection and ethics during the data collection process remain important challenges that need to be addressed further.

REFERENCES

[1] Yan, W., Zhang, X., Wang, Y., Peng, K., Ma, Y. (2024). Unraveling the relationship between teachers' and students' mental health: A one-to-one matched analysis. The Journal of Experimental Education. https://doi.org/10.1080/00220973.2024.2306412

[2] Saito, A.S., Creedy, D.K. (2021). Determining mental health literacy of undergraduate nursing students to inform learning and teaching strategies. International Journal of Mental Health Nursing, 30(5): 1117-1126. https://doi.org/10.1111/inm.12862

[3] Kotera, Y., Ting, S.H., Neary, S. (2021). Mental health of Malaysian university students: UK comparison, and relationship between negative mental health attitudes, self-compassion, and resilience. Higher Education, 81(2): 403-419. https://doi.org/10.1007/s10734-020-00547-w

[4] Ning, X., Wong, J.P.H., Huang, S., Fu, Y., et al. (2022). Chinese university students' perspectives on help-seeking and mental health counseling. International Journal of Environmental Research and Public Health, 19(14): 8259. https://doi.org/10.3390/ijerph19148259

[5] Bonsaksen, T., Chiu, V., Leung, J., Schoultz, M., et al. (2022). Students' mental health, well-being, and loneliness during the COVID-19 pandemic: A cross-national study. Healthcare, 10(6): 996. https://doi.org/10.3390/healthcare10060996

[6] Kotera, Y., Tsuda-McCaie, F., Maughan, G., Green, P. (2022). Cross-cultural comparison of mental health in social work students between UK and Ireland: Mental health shame and self-compassion. The British Journal of Social Work, 52(6): 3247-3267. https://doi.org/10.1093/bjsw/bcab240

[7] Kotera, Y., Andrzejewski, D., Dosedlova, J., Taylor, E., Edwards, A.M., Blackmore, C. (2022). Mental health of Czech university psychology students: Negative mental health attitudes, mental health shame and self-compassion. Healthcare, 10(4): 676. https://doi.org/10.3390/healthcare10040676

[8] Lee, J., Jeong, H.J., Kim, S. (2021). Stress, anxiety, and depression among undergraduate students during the COVID-19 pandemic and their use of mental health services. Innovative Higher Education, 46: 519-538. https://doi.org/10.1007/s10755-021-09552-y

[9] Yang, F., Song, Y., Yang, Y., Wang, R., Xia, Z. (2024). The influence of study abroad experience on the destination loyalty of international students: Mediating effects of emotional solidarity and destination image. Journal of Vacation Marketing, 30(2): 245-260. https://doi.org/10.1177/13567667221127391

[10] Aledo-Ruiz, M.D., Martínez-Caro, E., Santos-Jaén, J.M. (2022). The influence of corporate social responsibility on students' emotional appeal in the HEIs: The mediating effect of reputation and corporate image. Corporate Social Responsibility and Environmental Management, 29(3): 578-592. https://doi.org/10.1002/csr.2221

[11] Cui, L., Kong, W., Sun, Y., Shao, L. (2022). Expression identification and emotional classification of students in job interviews based on image processing. Traitement du Signal, 39(2): 651-658. https://doi.org/10.18280/ts.390228

[12] AlZu'bi, S., Abu Zitar, R., Hawashin, B., Abu Shanab, S., et al. (2022). A novel deep learning technique for detecting emotional impact in online education. Electronics, 11(18): 2964. https://doi.org/10.3390/electronics11182964

[13] Alkhalaf, S., Areed, M.F., Amasha, M.A., Abougalala, R.A. (2021). Emotional intelligence robotics to motivate interaction in E-learning: An algorithm. International Journal of Advanced Computer Science and Applications, 12(6): 173-183. https://doi.org/10.14569/IJACSA.2021.0120619

[14] Wu, S. (2021). Simulation of classroom student behavior recognition based on PSO-kNN algorithm and emotional image processing. Journal of Intelligent & Fuzzy Systems, 40(4): 7273-7283. https://doi.org/10.3233/JIFS-189553

[15] Arabi-Mianrood, H., Shahhosseini, Z., Tabaghdehi, M.H. (2022). The association between body image, emotional health, relationships, and unhealthy dietary behaviors among medical sciences students: A structural equation modeling analysis. Neuropsychopharmacology Reports, 42(4): 485-491. https://doi.org/10.1002/npr2.12291

[16] Sakellariou, C. (2023). The effect of body image perceptions on life satisfaction and emotional wellbeing of adolescent students. Child Indicators Research, 16(4): 1679-1708. https://doi.org/10.1007/s12187-023-10029-x

[17] Stephens, L.E., Bowers, E.P., Schmalz, D.L., Duffy, L.N., Lenhoff, J. (2023). A mixed method approach to evaluating eating-related psychopathologies in collegiate student-athletes. Journal of American College Health, 71(6): 1761-1774. https://doi.org/10.1080/07448481.2021.1947304

[18] Jia, N., Jing, H. (2021). Analysis of Cultural Education and Behavior in Colleges and Universities Based on Image Recognition Technology. Wireless Communications and Mobile Computing, 2021(1): 6195212. https://doi.org/10.1155/2021/6195212

[19] Zhou, J., Ran, F., Li, G., Peng, J., Li, K., Wang, Z. (2022). Classroom learning status assessment based on deep learning. Mathematical Problems in Engineering, 2022(1): 7049458. https://doi.org/10.1155/2022/7049458

[20] Shou, Z., Yan, M., Wen, H., Liu, J., Mo, J., Zhang, H. (2023). Research on students' action behavior recognition method based on classroom time-series images. Applied Sciences, 13(18): 10426. https://doi.org/10.3390/app131810426

[21] Lin, J., Li, J., Chen, J. (2022). An analysis of English classroom behavior by intelligent image recognition in IoT. International Journal of System Assurance Engineering and Management, 13(Suppl 3): 1063-1071. https://doi.org/10.1007/s13198-021-01327-0

[22] Pinna, R., Cicotto, G., Jafarkarimi, H. (2023). Student's co-creation behavior in a business and economic

bachelor's degree in Italy: Influence of perceived service quality, institutional image, and loyalty. Sustainability, 15(11): 8920. https://doi.org/10.3390/su15118920

[23] Chen, W., Fan, X., Dai, F., Chen, T. (2023). Student behavior identification during practice and training based on video image. Traitement du Signal, 40(1): 249-256. https://doi.org/10.18280/ts.400124