International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7184-7203

https://doi.org/10.70517/ijhsa463617

Multi-level Feature Fusion Transformer-CNN-LSTM Model for Dynamic Analysis and Personalized Prediction of Classroom Emotions of Non-native English Learners

Axin Huang[1,2,*] and Mary Jane C.Samonte[3]
[1] School of Graduate Studies, Mapua University, 1002 Metro Manila, Philippines
[2] GongQing Institute of Science and Technology, Gongqingcheng, Jiangxi, 332020, China
[3] School of Information Technology, Mapua University, 1002 Metro Manila, Philippines
Corresponding authors: (e-mail: ahuang@mymail.mapua.edu.ph).

Abstract The article proposes a novel cross-modal adversarial learning framework for analyzing the emotional dynamics of non-English learners during classroom engagement and predicting their individualized behaviors. The framework combines multilevel feature extraction and Transformer CNN-LSTM integrated model to handle multimodal data more efficiently and capture the complex relationship between emotions and behaviors. Low-level and high-level multilevel features are then extracted from the raw multimodal data. Meanwhile, Transformer is utilized to mine long-distance dependencies between multimodal data, CNN extracts local features, and LSTM is used to model dynamic changes in time series. In addition, the framework introduces adversarial training to learn shared features across modalities. Before 50 rounds of training, the CL-Transformer model loss function, emotion recognition accuracy, and behavior prediction accuracy converge, showing the fastest training speed and training results. The algorithm in this paper has more than 90% precision, recall, and F1 scores for emotion recognition and behavior prediction, and the recognition accuracy for different emotions is up to 0.96. In the fifth stage of the case study, the classroom emotion conversion rate and arousal is up to 0.66, and the model predicts that the probability of cell phone playing behavior is the highest for learners who are in angry moods, which is 64.7%. The learners' classroom emotional acceptance as well as behavioral integration have an impact on their classroom engagement.

Index Terms multilevel feature extraction, Transformer CNN-LSTM, cross-modal adversarial, behavioral prediction, emotion dynamic analysis

I. Introduction

The classroom is the main position of the implementation of the curriculum, student participation in the classroom for the observation of students' knowledge mastery, emotional changes, cognitive ability, growth and development provides a clear measure, can be a good reflection of the “students how to learn”, the implementation of the curriculum has an important significance of the value of educating people [1], [2]. At present, with the continuous investment of national education resources, students' enthusiasm for learning has not increased but decreased. Under the deepening of the new curriculum reform, the educational concept of “students as the main body” is gradually implemented and practiced [3]. Student participation is an important exploration of subjectivity, and at the same time also provides an important measure of teaching effectiveness [4]. However, although the subjective status of students has received widespread attention by educators, but the concrete implementation of classroom teaching practice, the long-established “knowledge-based” and “rationality first” value orientation is difficult to eliminate, teachers are still based on the structure of the subject to convey the subject curriculum, and teachers are still based on the structure of the subject, and the subject of the subject of the subject of the subject of the subject of the subject of the subject of the subject of the subject of the subject of the subject of the subject of the classroom. Teachers still convey the curriculum of their subjects according to the structure of the subjects, and students maximize the acceptance of institutionalized contents of the subjects, which affects classroom participation [5], [6]. Under this teaching method, students gradually lose their interest in learning, and what teachers and students interpret together is more of a passive and alien existence, which is the fundamental reason for the weak innovation ability of students in basic education and the frequent occurrence of high scores and low abilities [7], [8]. In addition to the singularity of the teaching form, the lack of students' initiative, academic pressure and other factors affecting the students' subjective position are also the main factors affecting students' participation in the classroom [9], [10]. The fundamental purpose of education is to cultivate people who can adapt to the modernized society, and the 21st century needs innovative talents, so how to break the oppression of test-based education on students' subjective

consciousness, develop students' thinking, and guide students to participate in the learning process independently is a problem that needs to be urgently explored in the field of education [11].

In the current educational context, students face pressure from various aspects such as coursework and interpersonal relationships, which leads to the prevalence of negative emotions, and emotions such as anxiety, depression, and frustration can affect students' learning outcomes and quality of life [12], [13]. As one of the main subjects of classroom participation, students' perception of teaching and emotional response is directly related to the observation of their participation in the process of emotional changes, which helps educators to better understand the students' learning status and psychological changes in order to students' classroom participation in the real emotional feedback, targeted adjustment of classroom interaction programs, teaching strategies, etc. [14], [15]. On the contrary, students' emotions directly affect classroom performance, classroom interaction effects, through understanding their emotional changes, timely regulation of emotions, mobilize subjective initiative, can improve classroom performance and interaction effects, and thus promote academic performance and social [16], [17]. In English teaching, classroom interaction is an essential part, and students' participation is an important way for educators to understand students' speaking and listening levels, and it is of great practical significance to analyze students' emotions in this process. However, emotions are categorized in various ways and individual emotions may change instantaneously, and accurate analysis of data that may produce different modalities, such as text, audio, image, and video, is the focus and difficulty of identifying students' emotions and dynamics of classroom participation [18]. Therefore, it is necessary to explore a method to comprehensively capture and analyze the change of emotional dynamics on students' classroom participation in order to provide a reference for personalized teaching and to solve the problem of students' lack of subjectivity.

In video as a modality for emotion analysis, Xinhan, N. [19] proposed an intelligent system for analyzing students' classroom status by combining neural network algorithm and emotion feature recognition, specifically, automatic recognition of students' status features by RFID, machine learning to construct a quantitative assessment model, logistic regression scoring algorithm, and training model, and neural network algorithm for calibrating the model performance. Ruiz, S. et al [20] proposed a twelve-emotion model and designed an emotion board in the PresenceClick system, where students can identify their emotional state and emotional evolution process based on monitored data.Deniz, S. et al [21] developed a software to save teachers' time for roll-calls and to analyze students' emotions in the classroom, which integrated the computer vision techniques and machine learning techniques.Llurba, C. et al [22] constructed a model to monitor and analyze the emotions of students in different school years, subjects and courses in the steps of designing code, recognizing facial emotions, tracking action units, and categorizing emotions.The model was input in the form of an image of the monitored facial emotions, and feature extraction in the image was carried out to achieve the emotion analysis.Huang , D. and Zhang, W. [23] used an improved hybrid face recognition model to detect students' facial expressions in surveillance videos, and then used the student expression recognition neural network algorithm to recognize the emotions presented on the face, which can accurately recognize students' emotional states in addition to detecting students' classroom attention. Among them, in the process of recognizing students' classroom emotions, there may be interfering factors such as unclear recognition of facial emotions, facial occlusion, etc. Yuan, Q. [24] captured students' emotion images and segmented them with multi-task convolutional neural network, extracted representative features in the images, composed a standard emotion database and classification criteria, and then extracted recognizable and representative features of facial emotion images by using a loss function. Combined with the results of image weight assignment from the attention mechanism network, the complete facial emotion is constructed, and finally the visual dynamic analysis of partial and global features is performed. While Ganesan, P. et al [25] addressed the occlusion region with a regenerative generative adversarial network and analyzed and predicted the full range of students' emotions through a visually presented interactive analysis dashboard that incorporates multiple disciplines' knowledge and teachers as inputs.

In addition, in emotion recognition with speech as modality, Liu, T. et al [26] introduced an ESERNet model, which mainly captures the long time dependency relationship in speech through the Transformer model as an emotional cue to recognize the emotion, and the Transformer model reduces the interference of emotional ambiguities due to noisy environments, sound quality of the recording equipment, etc. In the sentiment analysis with text as modality, Huddar, M. G. et al [27] extracted contextual features with bidirectional long and short-term memory network, extracted unimodal features in multimodality with evolutionary computational feature selection model, and fused unimodal two-by-two and then fused them three times, which formed a multilevel feature optimization and multimodal context fusion technique, which can be used not only in the sentiment analysis, but also in the sentiment classification.

In the above mentioned sentiment analysis, it is mostly based on the single modality analysis of video, speech and text, which makes it difficult to accurately identify the sentiment features in cases such as little change in mood,

facial paralysis, too large facial occlusion area in video, small amount of textual information, low clarity of images, and cluttered audio backgrounds, which leads to one-sided analysis results. Considering this, Li, M. et al [28] used a multimodal learning emotion analysis method implemented by deep neural network, which can automatically analyze students' classroom learning emotions in both video and voice modalities, and they also applied the PDA emotion scale to compare the learning status and emotions, so as to more accurately achieve the optimization of teaching methods. And Lu, Y. et al [29] pioneered the Long Short-Term Memory (LSTM) and Multi-scale Convolutional Neural Network (MSCNN) multimodal emotion recognition model by considering three modalities of text, image, and audio, and extracted the low-level and high-level features of the modalities using LSTM and MSCNN, respectively, and fused the two with an encoder of the Transformer model, and used the fused features as the input vector for final sentiment recognition. In multimodal sentiment analysis, multilevel feature selection and context extraction are the focus of the analysis. Among them, multilevel feature extraction is the process of transforming and processing different modal data several times, which can comprehensively extract the features in the modality. Xie, Y. et al [30] utilized a feature extraction module with different convolutional kernel scales to perform multilevel feature extraction for the recognition of facial expressions in the field environment, and fused the modules extracted several times to form a complete facial expression.

Whether it is LSTM, CNN, Transformer model, or multilevel feature extraction, all of them are fine identification and extraction of features in different modal data, but the advantages of each are not the same.LSTM has significant advantages in exploring the relationship of the memory over time, CNN can more accurately capture the features and relationships in the image, Transformer model can be the global contextual information is captured, and multilevel feature extraction is repeated many times for extraction and processing. These advantages are more friendly for the analysis of emotion dynamics in multimodal and cross-modal contexts, and the integrated model based on these advantages intends to perform multilevel accurate analysis of emotions by fusing different modalities and adversarial learning for accurate prediction of students' individualized behaviors in the classroom.

In this paper, we have developed a cross-modal adversarial learning framework by combining multi-feature extraction with various techniques such as Transformer, CNN, and LSTM. This framework carries a CNN-LSTM network structure that improves the generalization ability and robustness of the network through pooling operations, which contains memory units that can capture the time-dependence within the multimodal data, and in addition, power spectral densities are extracted for analyzing time-varying emotional signals of classroom participation. Transformer was used to extract the features of each modal data, and the similarities and differences between different modal data were analyzed with the assistance of the cross-modal matching module. To enhance the robust defense ability of the framework, adversarial training is completed with the help of the fast gradient notation method. Finally, under the constructed dataset, a study on the prediction of personalized behavior of non-English learners based on classroom participation emotions is launched.

II. Research objectives and methodology

II. A. Objectives of the study

There is a close relationship between classroom engagement emotions and learner behavior. Positive emotions promote focus, engagement, and creativity, while negative emotions may lead to distraction and reduced engagement. Currently, there exists more research on the recognition of learners' classroom emotions and the prediction of classroom behaviors by artificial intelligence techniques, but there is a lack of research on the prediction of learners' personalized classroom behaviors based on the dynamic changes of classroom engagement emotions. This paper takes this as a topic to develop related research.

Based on multi-feature extraction and Transformer CNN-LSTM model, this study mainly explores the following four issues:

First: to explore learners' mood changes in the classroom using the collected data.

Second: Explore whether there is a link between learners' emotions and their personalized behaviors.

Third: Whether the cross-modal adversarial learning framework constructed based on multi-feature extraction and Transformer CNN-LSTM model can be validated in emotion recognition and behavior prediction, and how accurate it is.

Fourth: Evaluate learners' engagement in the classroom based on the relationship between learners' emotions and behaviors.

In the emotion recognition part of the cross-modal adversarial learning framework based on multi-feature extraction and Transformer CNN-LSTM model designed in this paper, facial expression recognition based on non-physiological signals is selected as the method of emotion recognition, and the expression recognition technology is an important field for computers to understand the human emotions and human-computer interaction. That is, the expression states are selected from static pictures or video recordings as a way to obtain the emotional and

psychological states of people. In the research since then people have also added neutral expressions to this. The emotion recognition process in this paper can be divided into the following three steps: face detection and preprocessing, facial expression feature extraction and weight assignment, and expression classification and final emotion scoring.

II. B.Research methodology

(1) Literature research method

This study generally understands the relevant research directions and issues about this study through rough reading, and screens out the research methods and research ideas that need to be used in this study through intensive reading, determines the research questions of this study, and formulates the relevant research indexes.

(2) Empirical research method

Empirical research method is a research method based on data, through the collection and analysis of empirical data to verify or infer hypotheses. In this study, observation and data collection were conducted through the design of teaching experiments, and statistical analysis was used to explore the changes in learners' emotional levels in the classroom and the frequency of personalized behaviors exhibited by learners with different emotions.

(3) Data mining method

The data mining method utilizes techniques such as statistics, artificial intelligence and machine learning. Among them, machine learning method is a method that utilizes data and algorithms to enable computers to automatically learn and improve their performance by training models to learn from data and using these models to perform tasks such as prediction, classification and decision making. Based on data mining methods, the correlation between learners' classroom engagement emotions and personalized behaviors is explored, and the Transformer CNN-LSTM model is selected based on the analysis of experimental data for the prediction study of learners' personalized behaviors.

III. 3. CNN-LSTM network modeling

III. A. CBAM Attention Mechanism Module

CBAM [31] combines two attention mechanisms, channel attention and spatial attention, and is able to focus on both the channel dimension and the spatial dimension of the feature map. In this way, CBAM is able to capture the key information in the feature map more effectively and improve the model's ability to understand and express the image, thus achieving better performance in a variety of visual tasks.CBAM realizes the dynamic adjustment of the feature map through the attention patterns learned at different network layers, which enhances the model's attention to the important features.CBAM is able to capture more comprehensively the important information in the feature map, enabling the model to have a better understanding of and focus on all local regions of the input image.

CBAM consists of two sub-modules: the channel attention module and the spatial attention module.The network structure flow of CBAM is shown in Fig. 1. Assuming that the feature map $F \in R^{C\times H\times W}$ is given as an input, CBAM sequentially derives its one-dimensional channel attention map $M_c \in R^{C\times 1\times 1}$ and two-dimensional spatial attention map $M_s \in R^{1\times H\times W}$. The whole process of the attention mechanism can be summarized as follows:

$$F^{'} = M_c(F) \otimes F \tag{1}$$

$$F^{''} = M_s(F^{'}) \otimes F^{'} \tag{2}$$

where $\otimes$ denotes the Kronecker product between the two matrices, which is needed to guarantee the matching of the dimensions of the channel attention matrix and the spatial attention matrix during multiplication, and $F^{''}$ denotes the final output obtained by the CBAM module.

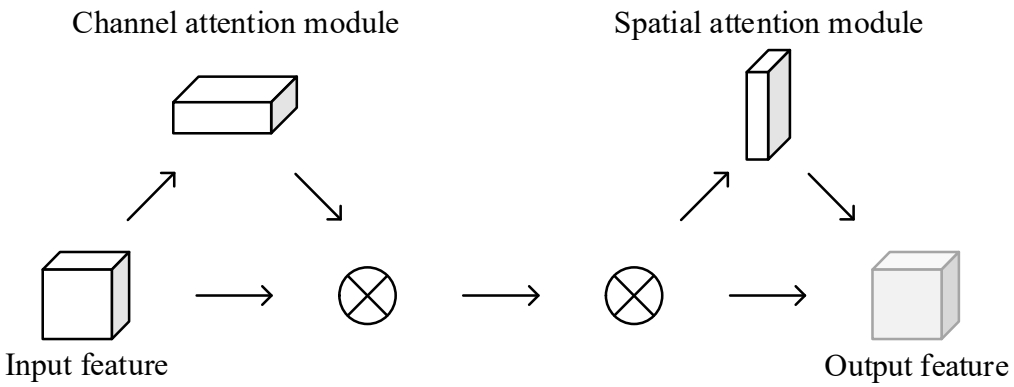

Figure 1: CBAM attention mechanism network structure

III. B. CNN-LSTM network construction

The CNN-LSTM network [32] used in this paper is divided into two parts, the CNN network and the LSTM network, and incorporates the CBAM attention mechanism to capture the feature information more comprehensively, and accesses the LSTM network after the CNN network to capture the temporal information in the features.

III. B. 1) CNN network model

The CNN network structure has a total of four convolutional layers, and the CNN structure in the CNN-LSTM network is shown in Figure 2. The first convolutional layer is a standard convolutional layer with a batch size of 64 and a convolutional kernel size of 4×4. The second convolutional layer is a standard convolutional layer with a batch size of 128 and a convolutional kernel size of 4 × 4. The third convolutional layer is a standard convolutional layer with a batch size of 256 and a convolutional kernel size of 4×4. The fourth convolutional layer is a standard convolutional layer with a batch size of 64 and a convolutional kernel size of 1×1. All the convolutional layers set the Relu function as the activation function to mitigate the overfitting of the model. After the convolutional layers the CBAM attention mechanism module is accessed, the CBAM attention mechanism module is placed after the convolutional layers in order to enhance the representation of the features and focus on the feature information from a more comprehensive perspective. After the CBAM attention mechanism module is a Maximum Pooling layer with a pooling kernel of 2×2 and a step size of 2, which improves the generalization ability and robustness of the network through pooling operations. The output of the pooling layer is then planarly unfolded through a Flatten layer and fed into a fully connected layer with 512 cells.

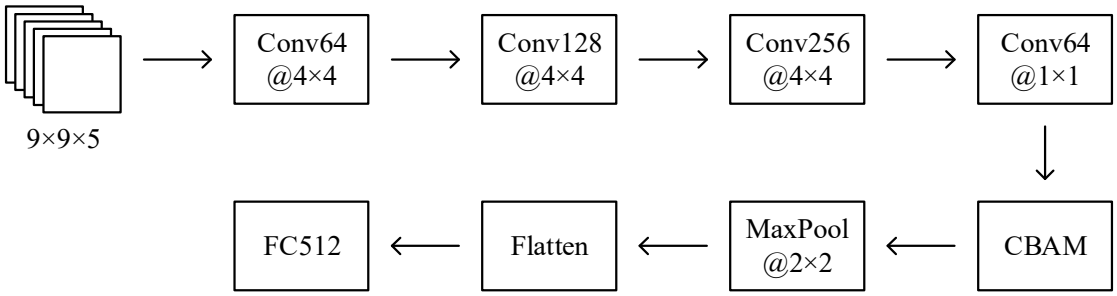

Figure 2: CNN network structure in CNN-LSTM

III. B. 2) LSTM network modeling

The LSTM network consists of a series of LSTM units, each of which includes three gating units (i.e., input gate, output gate, and forgetting gate) and a memory unit. They are interconnected and control the input, output and forgetting of information through the gating mechanism.

In this paper, an LSTM network layer with 128 memory cells is used to capture the temporal dependence within the data.The computational flow of the LSTM network can be represented by the following equation:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = \sigma(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * tanh(C_t) \tag{8}$$

$$y_t = W_{ho} * h_t + b_o \tag{9}$$

where σ denotes the activation operation with sigmoid activation function, f_t, i_t and o_t denote the forgetting gate, input gate and output gate, respectively, C_t denotes the state of the memory cell, $\tilde{C}_t$ denotes the content updated by the memory cell, w denotes the weight matrix, and b denotes the bias vector.

The LSTM structure in the CNN-LSTM network designed in this paper is shown in Fig. 3. Taking y_n as the final output feature representation of the CNN-LSTM network in this paper, the labels of the EEG data segment X_n are computed by means of linear transformation:

$$Out = Ay_n + b = [Out_1, Out_2, \cdots\cdots, Out_N] \tag{10}$$

where A denotes the transformation matrix and N denotes the number of categories for sentiment classification. Then the sentiment classification operation is performed on the output by softmax classifier and the final result represents the probability that the output classification result is the current category.

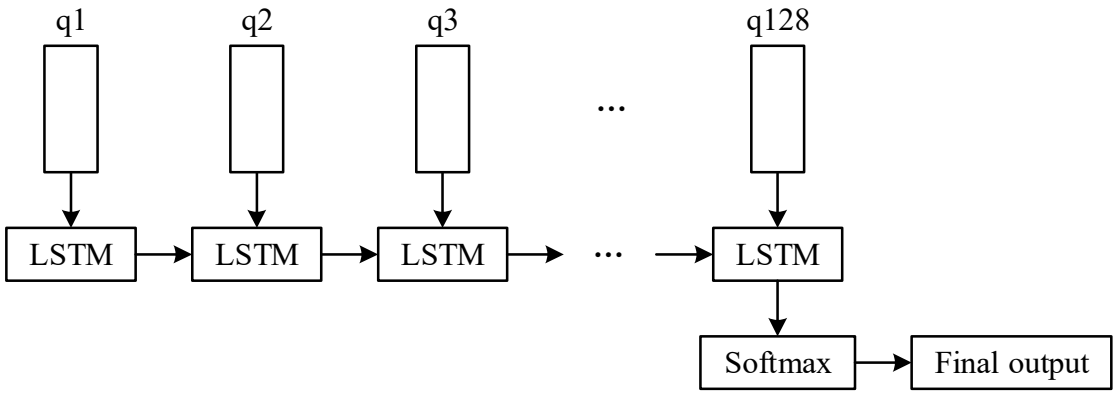

Figure 3: LSTM network structure in CNN-LSTM

IV. Cross-modal adversarial learning framework construction

Based on the above paper, a new CL-Transformer model is proposed by combining the better-performing CNN-LSTM model and Transformer structure, and incorporating cross-modal comparative learning and adversarial training to construct a cross-modal adversarial learning framework. To ensure the comparability of the study, this paper also extracted the power spectral density (PSD) features, and compared the power spectral density (PSD) features with the differential entropy (DE) features to determine their effectiveness in the dynamic analysis of classroom participation emotions.

IV. A. Power spectral density feature extraction

PSD [33] is the energy distribution of a signal in the frequency domain, which describes the intensity or power of the signal at different frequencies. In EEG signal analysis, changes in power spectral density can provide important information about brain activity. Different frequency ranges are often associated with specific brain waves (e.g., alpha, beta, theta, etc.), and variations in these waves may be related to cognitive tasks, emotions, or other neural activities.

In this paper, we use an estimation method of the Fourier transform (Welch's method) to estimate the PSD by segmenting the time-domain signal and calculating the power spectrum of each segment, and then averaging it. The calculation formula is shown in:

$$PSD(f)=\frac{1}{N\cdot\Delta f}\sum_{i=1}^{N}|X_i(f)|^2 \tag{11}$$

where N is the number of segments into which the signal is divided, Δf is the frequency resolution, and $X_i(f)$ represents the Fourier transform of the i th segment.

IV. B. Transformer structure

The core components of Transformer [34] contain the self-attention mechanism, multi-head attention, positional encoding, encoder-decoder structure, residual concatenation and layer normalization. The attention mechanism allows the model to dynamically allocate attention by assigning different weights to different positions of the input data.The self-attention mechanism in the Transformer model is a special kind of attention mechanism, which is used to process sequential data. Self-attention computation can be divided into the following for the data in the sequence, the original input data is first mapped to the Query, Key, Value vector space by linear transformation. The specific steps are as follows:

(1) Calculate the dot product of the Query vector of each element and the Key vectors of all elements, i.e., the similarity between the elements:

$$\text{similarity}\ (Q_i,K_j)=Q_i\cdot K_j \tag{12}$$

where, i denotes the position of the query element and j denotes the position of the target element.

(2) Use softmax function to convert similarity into weights:

$$\text{Attention weight}(Q_i,K_j)=softmax\left(\frac{\text{similarity}(Q_i,K_j)}{\sqrt{d_k}}\right) \tag{13}$$

where $\sqrt{d_k}$ is the scaling of the Key vector dimension.

(3) The obtained attention weights are weighted and summed with the Value vector to generate the final representation of the query elements:

$$\text{Output}(i)=\sum_{j=1}^{n}\text{Attention weight}(Q_i,K_j)\cdot V_j \tag{14}$$

where n is the length of the sequence and V_j is the Value vector corresponding to element j.

Multi-head attention means that the model sets multiple attention heads, each head produces an output, and finally these outputs are spliced and processed. Positional coding will consider the positional information of different elements in the model, the encoder is responsible for the input information and the decoder is responsible for generating the information, similar to the input and output in neural networks. In addition to this, residual concatenation and layer normalization are usually included in Transformer with the aim of speeding up the training process and mitigating the problem of vanishing gradients.

The specific operation process of the Transformer model is as follows: first, the input data is flattened by “nn.Flatten”, and then projected by “nn.Linear” to map the input data to the d_model dimension. Linear” to map the input data to the d_model dimension. Next, a Transformer encoder layer is defined using “nn.TransformerEncoderLayer”. This encoder layer contains the Multihead Self-Attention Mechanism (MHA) and Feedforward Neural Network (FNN) inside. In this experiment, d_model was set to 128 and nhead was set to 4, indicating the dimensionality of the model and the number of heads in the MHA. Subsequently, the whole Transformer model was constructed by stacking multiple Transformer encoder layers using “nn.TransformerEncoder”.

IV. C. Cross-modal comparative learning

Cross-modal comparative learning [35] is a neural network learning approach in which the model will improve the learning performance by comparing and recognizing the similarities and differences between different types or modalities of data, such as images, text, or sound, in order to obtain rich feature representations. To perform cross-modal retrieval, it is first necessary to prepare two or more different modalities of data, such as images and their corresponding textual descriptions. Then feature extraction is performed on these different modal data separately. After extracting the features of each modal data, the cross-modal matching module is used to evaluate the similarity

between these features. These different modal data are then mapped to the shared feature space by a mapping function.

IV. D. Classroom Emotion Recognition Based on CL-Transformer Modeling

The CL-Transformer model integrates the CNN-LSTM structure and the Transformer structure to capture the spatiotemporal features of the EEG signal more comprehensively. In the model, power spectral density and differential entropy features are initially extracted by the CNN model, respectively, and subsequently passed to the LSTM structure and the Transformer structure, where the features are processed independently and fused at a later stage. With this design, the CL-Transformer model can understand the multilevel and multidimensional features of EEG signals more comprehensively and provide richer information for emotion recognition tasks.The structure of the CL-Transformer model is shown in Fig. 4.

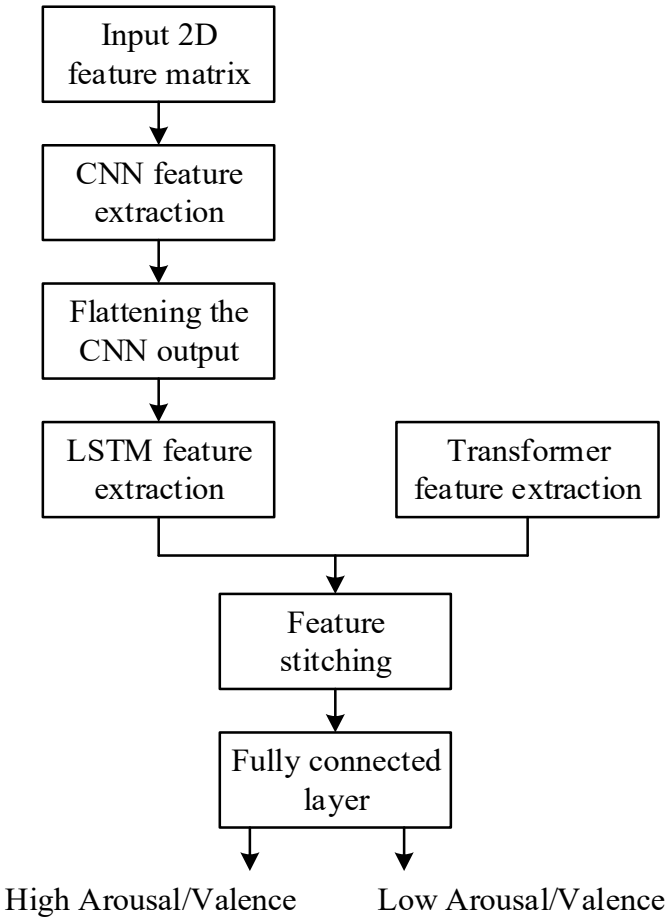

Figure 4: CL-Transformer model structure diagram

IV. D. 1) Cross-modal feature fusion comparison learning

In order to improve the sample information fusion ability of cross-modal features, contrast learning is added to cross-modal feature fusion, and a collection of adversarial negative samples of text and audio is introduced to construct the negative sample space. Adversarial contrast learning between and within modalities allows the adversarial negative samples in the negative sample space of text and audio to constantly track indistinguishable samples, effectively enhancing inter-modal interactions and amplifying the differences in the fusion representations between samples in order to effectively improve the mining and extraction of sample information.

In order to ensure that the input sequence length is consistent, each modal representation is first processed as the same sequence length. Define H_k^t as the text representation vector, H_k^v as the speech representation vector, and M_k^j as the fusion representation vector. The expressions are as follows:

$$\hat{H}_k^a = Conv(H_k^a, x^a), a \in \{t, v\} \tag{15}$$

$$\hat{M}_k^j = Conv(M_k^j, x^f) \tag{16}$$

$\hat{M}_k^j$ - is the fused representation vector obtained after j Transformer layers.

x^a - the size of the convolution kernel for modality a.

x' - the size of the convolution kernel of the fused modality.

First K samples are randomly selected to construct a group. Then a positive sample pair is randomly selected from the K samples and the remaining K-1 become negative sample pairs. Positive sample pairs are modal sample pairs consisting of text and speech in the same sample, and negative samples are modal sample pairs consisting of text and other samples in the same sample. For each sample pair self-supervised comparison loss:

$$L^{tv,j} = -\log \frac{\exp(M_k^{'} H_k^{'})}{\exp(M_k^{'} H_k^{'}) + \sum_{x=1}^{X} M_k^{'} H_x^{'}} \tag{17}$$

$L^{n,j}$ - Contrast loss for text-to-speech modal fusion performed on the encoder's Layer j Transformer.

IV. D. 2) Confrontational training

Adversarial training adds a modest amount of perturbation I_{inter} to the input image I to generate an adversarial image I_{adv} to force the model to learn deeper levels of feature attention. Adversarial image I_{adv} consists of adversarial examples I_{FGSM} generated by Fast Gradient Symbology Method (FGSM) [36] and adversarial examples I_{PGD} generated by Projected Gradient Descent (PGD).

FGSM increases image interference by attacking the model gradient. It is computed as:

$$I_{FGSM} = I + \grave{o} \cdot sign(\nabla_I \mathrm{L}(\theta, I, y_{gt})) \tag{18}$$

I denotes the source image. $\grave{o}$ takes a random number in the range (0, 0.01] at each iteration to increase randomness. ∇_I is the model gradient. θ denotes the model parameters, and y_{gt} is the ground-truth mask map. L denotes the loss function.

FGSM attacks the model only once based on the gradient. However, when the model is complex, this method may not achieve optimal results. Projected gradient descent is a multi-step variant of FGSM. One small step at a time, each iteration projects the perturbation to a defined range with the following formula:

$$\begin{cases} I_{PGD}^0 = I \\ I_{PGD}^{n+1} = Clip(I_{PGD}^n + \alpha \cdot sign(\nabla_l \mathrm{L}(\theta, I_{PGD}^n, y_{gt}))) \end{cases} \tag{19}$$

Clip denotes cropped image. The small step attack by PGD makes the model more urgent to learn robust defense capability to avoid misidentifying the tampered region.

V. Constructing a dataset of non-English language learners' classroom behaviors

V. A. Sources of data sets

(1) Classroom Participation Mood

The data for the classroom participation emotion dynamic analysis experiments all come from the public service platform of educational resources, and the study has collected a total of about 265 lessons from the same region, the same subject and the same grade from the public service platform, and after screening, we retained 157 lessons with obvious emotions and clear sound quality for generating the classroom participation emotion dynamic analysis dataset.

This paper draws on the acquisition and preprocessing process of the IEMOCAP dataset as a way to construct the classroom emotion recognition dataset used in this study. Specifically, after obtaining the classroom video data, the audio data in the video is first extracted, and then the audio is subjected to batch preprocessing operations, which include denoising, endpoint detection, and image-audio cutting, and then the cut image-audio segments are called Baidu Speech Recognition to obtain the text translation in batch, and finally error correction and emotion labeling are performed manually.

(2) Personalized Behavior Prediction

In this paper, we used Python to write a crawler program to crawl the image libraries of the three major search engines of Baidu, Bing and Google, and collected about 10,000 high-definition photos in total. In order to further

expand the dataset, the recording of surveillance cameras was also started after the beginning of the new semester. More than 100 classrooms were selected for monitoring and the videos were recorded every day for almost 3 months in high definition (1080p) quality. At the end of the recording, a video frame-cutting technique was used to capture an image from it every 5 minutes, resulting in over 70,000 images.

V. B. Labeling of data sets

(1) Classroom Engagement Emotions

Discrete Emotion Description Model It describes emotions into discrete categories, such as happy, angry, and sad. This dataset refers to a wide range of human emotions as well as combining the actual emotional performance in the traditional classroom in secondary schools, and finally identifies 7 categories of emotions: angry, disgusted, scared, happy, sad, surprised, and calm. The emotions of each sample in the dataset were jointly labeled by 5 people, and the principle of minority to majority was used to determine the final emotion labels of the samples.

(2) Personalized Behavior Prediction

The web capture and video recording datasets labeled in this project are divided into five categories of personalized behaviors: listening, raising hands, taking notes, playing cell phones and sleeping. At the same time, a set of labeling rules is specified to satisfy the following conditions on the premise that the labeling frame closely follows the target profile:

Hand-raising category: labeling the target with the hand raised and labeling the upper body.

Listening category: labeling the target with the head up looking at the board, labeling the upper body.

Taking notes category: mark the target of writing with a pen in hand, mark the pen and hand.

Playing with cell phone category: target with cell phone in hand, mark cell phone and hand. Targets that look down, sideways, or backward, label the upper body.

Sleeping: Targets sleeping on their stomachs, label the upper body.

With the exception of targets with particularly severe occlusion (showing only one arm or only half of the head) that were not labeled, the rest of the targets were labeled within the rules. The labeling rule is based on the assumption that in most normal classrooms, the teacher is usually in the front of the class and the learners at the bottom are looking up at the podium. Targets with their heads down are naturally classified as playing with their cell phones because it is impossible to tell exactly what they are doing. Similarly sideways and backward looking and indiscernible movements are classified as playing with a cell phone.

V. C. Data enhancement

The dataset of classroom scenarios involved in this project is very difficult to label due to the fact that the average number of learners are more than 30 and the project staff has limited energy. And in order to train a high-precision model, a large amount of data is needed to support it. In order to expand the dataset when the number of annotations reaches about 5000, this project adopts the data enhancement technique, which can generate more data on the basis of the existing annotated data, so as to increase the size and diversity of the dataset.

The data enhancement technique mainly consists of operations such as horizontal, left, right, and flip of the video frame image, horizontal and vertical panning, and randomly changing the brightness and darkness of the image. Through these operations, more data samples can be generated, making the size of the dataset expand from the original 5,000 to more than 12,000. The expanded dataset can better provide data support for subsequent model training, and can also improve the generalization ability of the model in terms of quantitative advantages.

In addition, the data enhancement technique can also effectively solve the problem of overfitting, by performing some transformations and perturbations on the original data, the diversity of the dataset can be increased, so as to reduce the sensitivity of the model to specific samples. At the same time, data enhancement can also improve the generalization ability of the model, so that the model can be better adapted to different scenes and lighting conditions, thus improving the reliability and stability of the model in practical applications.

VI. Results and analysis

VI. A. CL-Transformer model training performance evaluation

This experiment is constructed in Ubuntu 20.04.2 LTS operating system environment based on PyTorch deep learning framework. The experimental environment is as follows:

CPU: Intel Xeon Silver 4210 CPU @ 2.20GHz × 2

Memory: 128G

Programming language: Python 3.8

CPU acceleration library: CUDA 11.4

In the training process, the input images are all of size 640 × 640, the learning rate is set to 0.001, and the bit-size is set to 64. The Adam optimizer is used for optimization, the number of training iterations is set to 200, and the weight decay is set to 0.001. The source code with the set parameters is compiled, i.e., the weight data of the model that can be obtained through training, and the per-image The inference running time is about 1 ms. Long Short-Term Memory Network (LSTM), Convolutional Neural Network (CNN), Residual Network (ResNet), and Random Forest (RF) models are chosen as comparisons, and are trained under the same experimental environment to verify the superiority of the training effect of the models in this paper.

Figure 5 shows the loss function training of each model for 200 iterations. From the data in the figure, it is understood that all models complete convergence before 100 rounds of iterations. The model in this paper converged to a loss value of around 0.08 at 40 rounds. Compared to the comparison models, the CNN model has the lowest loss value, but it needs 80 iterations to fully converge.The ResNet model converges in the 55th iteration, showing the fastest convergence speed, but its loss value is around 0.3. In summary, the model in this paper shows great advantages in both the convergence speed and convergence effect on the loss function.

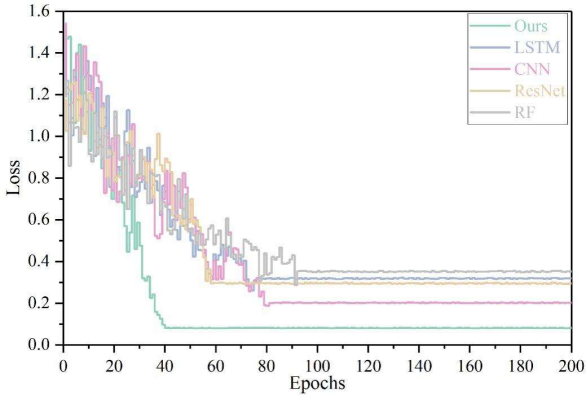

Figure 5: The loss value training of the various models of 200 times

In addition, this section validates the accuracy of emotion recognition as well as behavior prediction for non-English learners in the above dataset, and the training results of emotion recognition accuracy and personalized behavior prediction accuracy of different network models are shown in Figures 6 and 7, respectively. According to the experimental data, it is found that the introduction of cross-modal feature fusion comparative learning and adversarial training in the CL-Transformer network can obviously speed up the convergence speed of the model and achieve better convergence results. This is due to the fact that the combination of the two improves the discriminative nature of classroom participation emotion features and enhances the extraction of image emotion features.The emotion recognition accuracy of the CL-Transformer network model in the classroom emotion dataset eventually stabilized at around 0.97 and above, with only 45 iterations. Since it is more difficult to extract useful emotion features from the shallow network, it took 80 to 100 rounds of training for the comparison model before the emotion recognition accuracy converged to between 0.7 and 0.9. In the personalized behavior prediction dataset, the model in this paper also shows the best training results, with the personalized behavior prediction accuracy oscillating around 0.94 at 40 rounds of iterations. During the training process, the CBAM module is placed after the CSP (Cross Satge Prtiala) module, which means that every time the model passes through the CSP, a computation of the attention mechanism is performed. This process performs a multidimensional attentional weighting of the targets on the feature map, highlighting key features of the learner's classroom behavior and reducing the attention to irrelevant features. In this way, the algorithm in this paper is able to extract the key features related to learners' classroom behaviors from the feature map more efficiently, thus improving the accuracy of learners' classroom emotion recognition and behavior prediction.

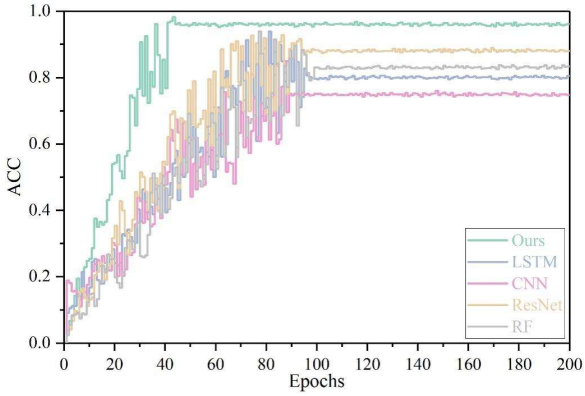

Figure 6: Emotional recognition accuracy of different network models

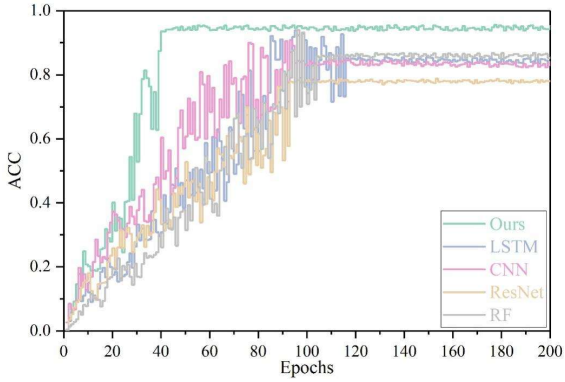

Figure 7: Behavior prediction accuracy of different network models

VI. B. Feasibility Analysis of Emotion Recognition and Behavior Prediction

This section continues to test the classroom participation emotion recognition effect and personalized behavior prediction effect under the above different algorithms using precision rate, recall rate and F1 score as indicators. The test results of classroom participation emotion recognition set behavior prediction of different algorithms are shown in Fig. 8 and Fig. 9, respectively.From them, it can be seen that the performance of classroom participation emotion recognition of this paper's algorithm outperforms that of the four contrasting algorithms, with the measured values of 94.31%, 97.39%, and 95.83% in precision rate, recall rate, and F1 scores, respectively. The measured values of the comparison algorithms did not exceed 90%, which is 7.56% to 17.13% lower than this paper's algorithm. For personalized behavior prediction, this paper's model is far ahead of the comparison model with an F1 score of 95.96%. The algorithmic model designed in the study demonstrated superior performance in the tasks of identifying classroom participation emotions and predicting personalized behaviors, with significant model validity and feasibility.

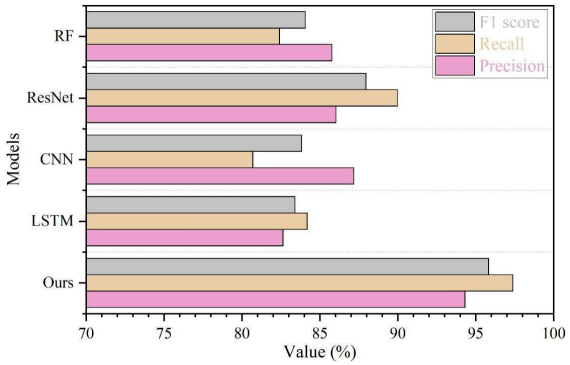

Figure 8: Classroom participation in emotional recognition test results

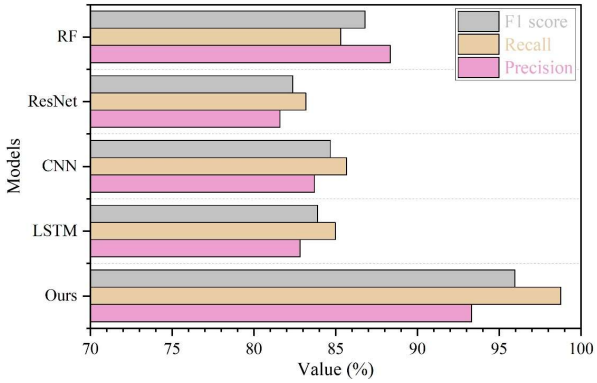

Figure 9: Personalized behavior prediction test results

To further validate this paper's model in recognizing the emotions of non-English learners in classroom participation, the above emotion analysis dataset is classified in this section into anger, disgust, fear, happiness, sadness, surprise, and calmness, a total of seven types. The emotions of each category are identified using the model of this paper and the emotion recognition confusion matrix is drawn.

Figure 10 shows the emotion recognition confusion matrix, and the diagonal line in the figure shows the accuracy of this paper's model on each emotion recognition. It can be seen that the recognition accuracy of this paper's model in seven kinds of classroom participation emotions has been more than 0.9, in which the recognition accuracy of happiness is the highest, up to 0.96. The probability of each emotion to be recognized as other emotions is not more than 0.04, which indicates that this paper's model can better achieve the recognition of emotions such as angry, disgusted, afraid, happy, and so on.

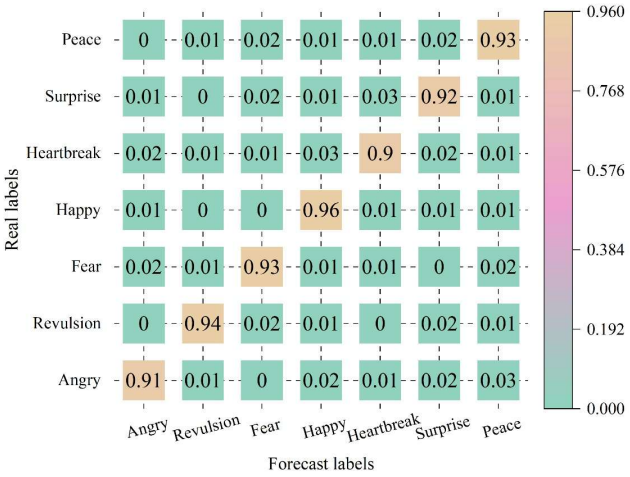

Figure 10: Emotional identification confusion matrix

VI. C. Interaction Analysis of Learners' Classroom Emotions and Individual Behaviors

In order to better analyze the emotional changes of non-English learners in classroom participation, this paper utilizes a cross-modal adversarial learning framework to identify learners' classroom emotions and calculates the classroom emotion conversion rate and classroom excitement to determine the level of classroom emotions. The overall classroom mood is divided into: excitement, calmness, and depression.

The formula for classroom mood conversion rate is as follows:

$$S=\frac{f_s}{T}O_c \tag{20}$$

f_s is the number of emotional transformations, T represents the classroom duration, and O_c is the sampling frequency of instructional images, which was taken as O_c = 10 times/minute in this study.

The formula for classroom excitement is as follows:

$$E = \frac{\sum f_{\varepsilon}}{T} \tag{21}$$

where, f_{ε} is the number of images judged to be excited emotions. Combining the above dimensions, the following formula for classroom mood level determination is obtained:

$$M = \begin{cases} Excitation, E>0.5\&S>0.5 \\ Smoothness, \text{other} \\ Down, E<0.3\&S<0.3 \end{cases} \tag{22}$$

This section of the experiment counted the mood changes of the non-English learners throughout the classroom session, dividing the 45-minute classroom into 10 phases. The initial stage was before the beginning of the class, and from the first stage (0-5 minutes), the division was done every 5 minutes until the final stage (40-45 minutes).

The rate of classroom mood conversion and classroom excitement were calculated in this lesson to determine the level of classroom mood in this case. Figures 11 and 12 show the results of calculating the classroom mood conversion rate and classroom excitement as well as the changes in the classroom mood level, respectively. Combining the two figures, it can be seen that the classroom mood conversion rate and classroom excitement level are continuously changing throughout the whole classroom session. In the initial stage before the beginning of the class, the classroom mood conversion rate is 0.33, the excitement level is 0.48, and the mood level is in a calm state. Combined with the images, this is the classroom rest stage at the end of the last class. Thereafter, in the first stage, the classroom mood is more excited, and at the second stage, it transforms into calm, and the analysis shows that there is a clear interactive behavior in the classroom at this time. In the third and fourth stages, the classroom mood is relatively low, combined with the classroom video analysis can be seen at this time the classroom in the teacher to explain the state of knowledge. In the fifth stage, the classroom mood conversion rate and excitement were calculated to be 0.66, the classroom resumed the state of excitement, and students began to discuss the problem. The reasons for the change of classroom mood in the following stages are basically consistent with the previous analysis.

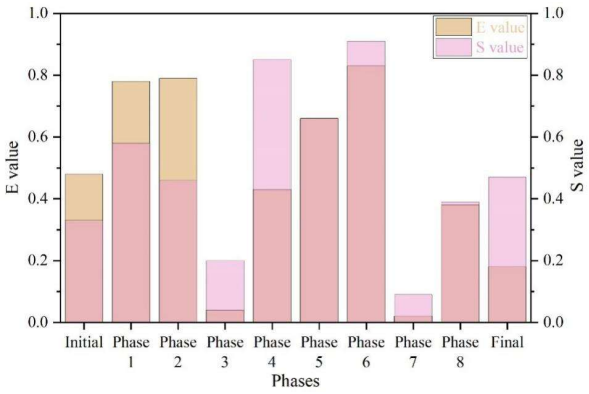

Figure 11: The rate of mood conversion and the calculation of classroom excitation

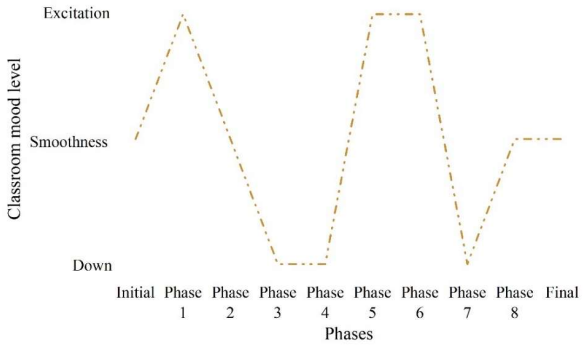

Figure 12: Mood level changes in the classroom

In order to deeply analyze the existence of personalized behaviors of non-English learners with different emotions, the experiments recorded the learning videos of each non-English learner in a classroom, and statistically analyzed the frequency of different personalized behaviors of the population of learners with the tendency of angry, disgusted, scared, happy, sad, surprised, and calm emotions. The cross-modal adversarial learning framework based on the CL-Transformer model in this paper is further adopted to predict the personalized behaviors that learners are about to display by identifying their classroom emotions.

The results of the analysis of learners' emotionally inclined classroom behaviors are shown in Figure 13. As can be seen from the figure, learners with different emotions have significant differences in the frequency of personalized behaviors they display in the classroom. According to statistics, learners with negative emotions (anger, disgust, fear, and sadness) in the classroom tend to play with cell phones, sleep classroom behaviors, and raise their hands, take notes, and listen to the lecture less frequently. For example, learners with angry emotions have a frequency of 43 and 27 for playing cell phones and sleeping, respectively, while the frequency of listening to lectures is 1, and there are no other types of behaviors. This indicates that learners with negative emotions are often not interested in class content and easily feel sleepy. The learners with positive emotions (happy, surprise), on the other hand, showed more class participation behaviors (raising hands, taking notes, listening). For example, learners with happy emotions have a frequency of 54 listening behaviors, and there is no cell phone playing or sleeping type of behaviors. Lastly, learners with calm emotions, this category of learners showed all classroom behaviors, but the frequency of behaviors in the category of raising hands, taking notes, and listening to lectures was much higher than the behaviors in the categories of playing cell phones and sleeping.

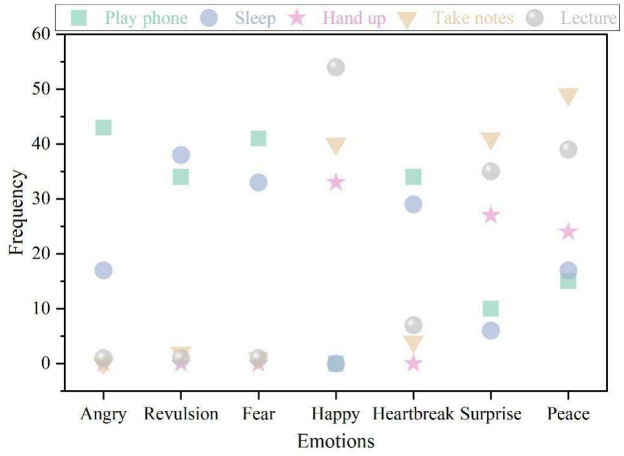

Figure 13: The results of the analysis of the behavior of the learner

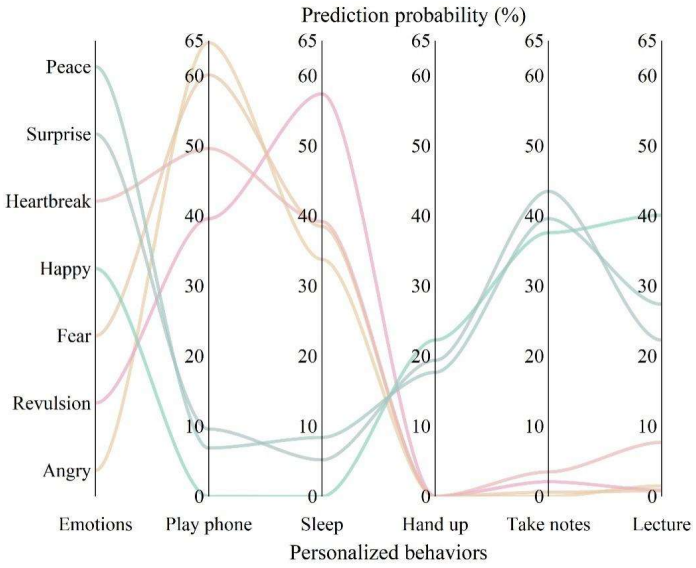

Figure 14: The behavior of different mood learners in this paper model results

In addition, based on the analysis of the emotional tendencies of non-English learners, the personalized behavior prediction of learners with different emotions is carried out. Figure 14 shows the results of personalized behavior prediction for learners with different emotions under the model of this paper. From the data in the figure, it can be concluded that for learners with negative emotions, the model predicts a higher probability of cell phone playing and sleeping behaviors. Taking learners with angry emotions as an example, the model predicts that the probability of their appearing to play with cell phones is 64.7%, 33.8% of the probability of sleeping, and only 1.5% of the probability of their appearing to listen to lectures. As for positive and calm emotions, the model shows that the probability of this type of learners to show hand raising, note taking and listening type of behaviors is higher than that of cell phone playing and sleeping type of behaviors.

VI. D. Classroom Engagement Measures Based on Emotions and Behaviors

In this paper, emotional and behavioral factors were numerically detected and extracted through the quantitative model of emotion and hierarchical analysis, respectively. Relevant questionnaires were designed and scored by experts to determine the influence weights of each factor on learners' classroom participation. The multilevel comprehensive evaluation system of classroom participation is shown in Table 1.

Table 1: Multi-stage comprehensive evaluation system of classroom participation

Second factors	Weighting	First factors	Weighting
Emotional acceptance	0.75	Angry	0.0513
		Revulsion	0.0169
		Fear	0.0416
		Happy	0.4021
		Heartbreak	0.0613
		Surprise	0.2651
		Peace	0.1617
Behavior integration	0.25	Play phone	0.0457
		Sleep	0.0441
		Hand up	0.5171
		Take notes	0.1758
		Lecture	0.2173

This section analyzes the results of quantifying the emotional acceptance of classroom participation by non-English language learners through actual tests and data statistics. The study conducted multiple tests, and the results of four representative sets of tests, all 30 minutes in length, are presented. The formula for calculating emotional acceptance is as follows:

$$Emo = \frac{1}{n}\left[w_1, w_2, ..., w_7\right]\begin{bmatrix} t_1 m_1 \\ t_2 m_2 \\ ... \\ t_7 m_7 \end{bmatrix} \tag{23}$$

n indicates the number of included emotion frames, $\left[w_1, w_2, ..., w_7\right]$ is the weight of the seven emotions, $\left[t_1, t_2, ..., t_7\right]^T$ is the number of times the corresponding emotion appeared during the detection period, and $\left[m_1, m_2, ..., m_7\right]^T$ is the corresponding emotion score.

Figure 15 shows the results of emotion detection and calculation in classroom participation of non-English learners. The four groups of tests represent four different kinds of classroom performance. The first group is classroom performance in an ordinary state, where the learners' acceptance of the knowledge and the classroom is average, and they do not show active listening, but they are not resistant to the classroom itself. The second group is obviously integrated into the classroom, and its happy mood accounted for 89%, which to some extent indicates that the learner has a high degree of welcome to the content taught by the teacher, and the learner is in a kind of active and positive learning state, and his emotional state is in a high degree of activation, so from the quantitative results, it can also be seen that the learner scores higher in the emotional perspective.

The third and fourth groups are two experiments of negative listening experience. Since negative emotion is a comprehensive manifestation, the learners in the third group of experiments were in a negative state as a whole,

without some kind of negative emotion dominating, which was mapped to the classroom performance as a relative deviation from the classroom content or resistance to the knowledge imparted by the teacher, and the scores were negative. In the fourth group, compared to the third group, the learners' anger accounted for most of the test frames, indicating that the learners' listening experience was poor, in a highly tense emotional state of resistance, questioning, deviation and even criticism of knowledge or class, and the effect of listening to the class under this kind of emotion was also significantly worse.

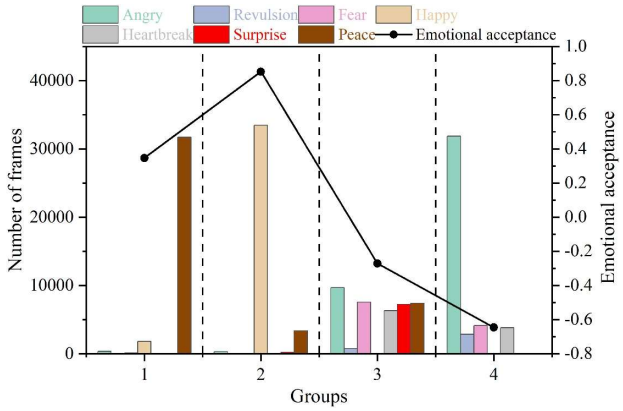

Figure 15: Emotional acceptance test and calculation results

This paper gives the results of the personalized behavioral integration measure quantification for non-English language learners, which is calculated based on the following formula for judging behavioral integration:

$$Act = \frac{1}{n}\left[w_1, w_2, ..., w_5\right]\begin{bmatrix} t_1 \\ t_2 \\ ... \\ t_5 \end{bmatrix} \tag{24}$$

n indicates the number of frames containing a particular behavior in the image, $\left[w_1, w_2, ..., w_5\right]$ is the weight of the five behaviors, and $\left[t_1, t_2, ..., t_5\right]^T$ is the number of occurrences of the five behaviors.

Also listed are representative results of four groups of tests all of 30 minutes in duration. Figure 16 shows the statistics of the number of occurrences of various behaviors and their scoring results. From the comparison of the scores of the second and fourth groups, it can be clearly seen that when the learner is out of the classroom (e.g., sleeping, playing with cell phone actions accounted for the main part) the learner's classroom behavioral integration scores are significantly lower. The learners are not engaged in the lesson itself at the moment, but rather in what is called “desertion”.

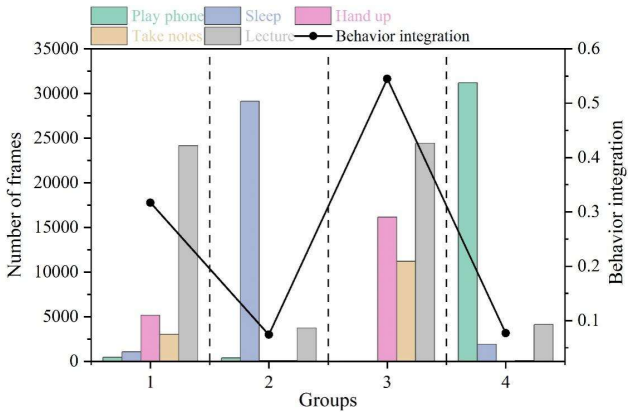

Figure 16: Behavior integration test and calculation results

As a result of the quantitative scores above, this section identifies emotional acceptance and behavioral integration as important detectors of the impact of classroom engagement for non-English learners. This gives a quantitative formula for classroom engagement:

$$C_p = 100 \times \left(0.75Emo + 0.25Act\right) \tag{25}$$

In this section, 30 learners were tested and scored while learning the same course content using the model of this paper to identify the classroom emotions and behaviors of each learner. Figure 17 shows the results of quantitative classroom engagement scores for each learner, with (a)-(c) showing the learner emotional acceptance score, learner behavioral integration score, and learner classroom engagement score, respectively. By identifying classroom emotions and behaviors through the model in this paper, the learners' emotional acceptance as well as behavioral integration scores ranged from 3.31 to 9.26 and 2.72 to 9.34, respectively. Using the above formula, the learners' classroom engagement scores were obtained between 3.54 and 8.85, which corroborated with the results based on teachers' classroom feelings. Through the post-class interview survey, the classroom emotions as well as behavioral performance identified in this paper are basically consistent with the learners' situation at that time. For learners with low emotional acceptance scores, they mostly showed confusion, incomprehension or even aversion to the knowledge content, which led to behaviors such as playing cell phones and sleeping, thus leading to lower classroom engagement.

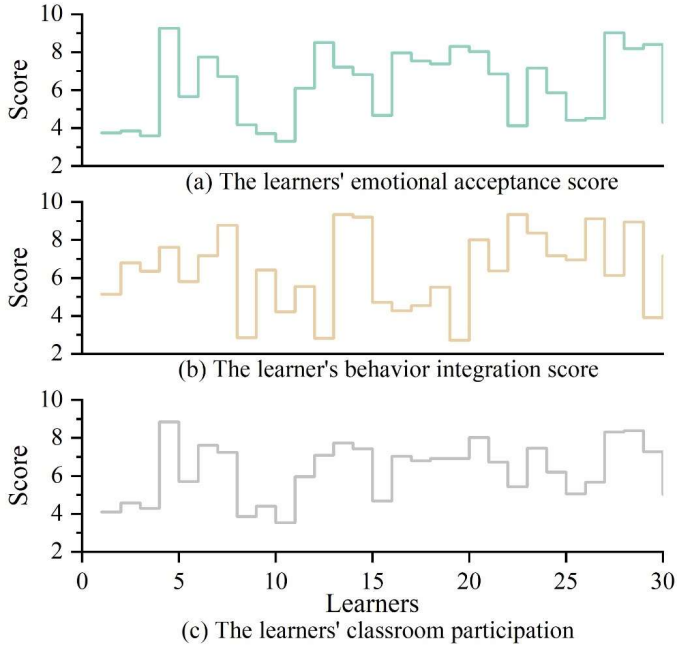

Figure 17: The results of the participants' participation in classroom participation

VII. Conclusion

A cross-modal adversarial learning framework is designed to predict learners' personalized behaviors through their classroom engagement mood changes. The Transformer CNN-LSTM integration model is used as the central architecture of the framework to achieve multi-level feature extraction of classroom emotion and behavior data. At the same time, cross-modal feature fusion and comparison is utilized to perform adversarial comparative analysis between and within modalities to improve the ability of sample information mining. This is then combined with adversarial training to learn deeper levels of feature attention and enhance defense capabilities. The research results are as follows:

(1) The model in this paper converges successively in loss function, behavior prediction accuracy and emotion recognition accuracy before 50 rounds of iteration, in which the loss value stabilizes around 0.08 and the accuracy are all higher than 0.94.

(2) The model's precision, recall and F1 score on emotion recognition are 94.31%, 97.39% and 95.83%, respectively, and the recognition accuracy for anger, disgust, fear, happiness, sadness, surprise and calm among them is more than 0.9.

(3) Classroom emotions affect the personalized behaviors of non-English learners, the CL-Transformer model predicts that the likelihood of learners (angry emotions), listening behavior is 1.5%.

(4) By using the model in this paper to identify the learners' classroom emotions and predict their personalized actions, the learners' classroom engagement scores were calculated to be between 3.54 and 8.85, which is consistent with the teachers' feelings about the learners' engagement.

The results of the study show the superior performance of the CL-Transformer model based on cross-modal feature fusion and adversarial training in emotion recognition and behavior prediction, which provides a data basis for the analysis of emotional dynamics and individualized behavior in the field of education.

References

[1] McCleary, D. F., McCreary, B., & Coles, J. (2019). Cognitive Variables, Classroom Behaviors, and a Participation Intervention on Students' Classroom Participation and Exam Performance. International Journal of Teaching and Learning in Higher Education, 31(2), 184-194.

[2] Mazer, J. P. (2017). Associations among classroom emotional processes, student interest, and engagement: A convergent validity test. Communication Education, 66(3), 350-360.

[3] Shurbanovska, O. (2018). Scale for Higher Education Teaching–Teacher-oriented or student-oriented. Empirical studies in psychology.

[4] Hard, B. M., & RaoShah, T. (2022). Developing collaborative thinkers: Rethinking how we define, teach, and assess class participation. Teaching of Psychology, 49(2), 176-184.

[5] Liu, D. (2021, April). Student-Oriented Classroom Teaching Model from the Perspective of Constructivism. In 2021 2nd Asia-Pacific Conference on Image Processing, Electronics and Computers (pp. 1025-1028).

[6] Severe, E., Stalnaker, J., Hubbard, A., Hafen, C. H., & Bailey, E. G. (2024). To participate or not to participate? A qualitative investigation of students' complex motivations for verbal classroom participation. Plos one, 19(2), e0297771.

[7] Balliu, V. (2017). Modern teaching versus traditional teaching-Albanian teachers between challenges and choices. European Journal of Multidisciplinary Studies, 2(4), 20-26.

[8] Gumartifa, A., Syahri, I., Siroj, R. A., Nurrahmi, M., & Yusof, N. (2023). Perception of Teachers Regarding Problem-Based Learning and Traditional Method in the Classroom Learning Innovation Process. Indonesian Journal on Learning and Advanced Education (IJOLAE), 5(2), 151-166.

[9] van Balen, J., Gosen, M. N., de Vries, S., & Koole, T. (2024). Taking learner initiatives within classroom discussions with room for subjectification. Classroom Discourse, 15(2), 123-142.

[10] Ahmad, C. V. (2021). Causes of students' reluctance to participate in classroom discussions. ASEAN Journal of Science and Engineering Education, 1(1), 47-62.

[11] Chu, G., & Chai, X. (2018, February). Thinking and Practice of" Student Oriented" Educational Philosophy. In 6th International Conference on Social Science, Education and Humanities Research (SSEHR 2017) (pp. 654-657). Atlantis Press.

[12] Bhujade, V. M. (2017). Depression, anxiety and academic stress among college students: A brief review. Indian Journal of Health & Wellbeing, 8(7).

[13] Grilo, A. P. D. S., Pina-Oliveira, A. A., & Puggina, A. C. G. (2021). COMPETENCE IN INTERPERSONAL COMMUNICATION: RELATIONSHIPS WITH SOCIAL CHARACTERISTICS AND ANXIETY TRAIT. Revista Mineira de Enfermagem, 25.

[14] Goetz, T., Keller, M. M., Lüdtke, O., Nett, U. E., & Lipnevich, A. A. (2020). The dynamics of real-time classroom emotions: Appraisals mediate the relation between students' perceptions of teaching and their emotions. Journal of Educational Psychology, 112(6), 1243.

[15] Huang, Y., Deng, W., & Xu, T. (2024). A Study of Potential Applications of Student Emotion Recognition in Primary and Secondary Classrooms. Applied Sciences (2076-3417), 14(23).

[16] Canovi, A. G., Rajala, A., Kumpulainen, K., & Molinari, L. (2019). The dynamics of class mood and student agency in classroom interactions. The Journal of Classroom Interaction, 4-25.

[17] De Neve, D., Bronstein, M. V., Leroy, A., Truyts, A., & Everaert, J. (2023). Emotion regulation in the classroom: A network approach to model relations among emotion regulation difficulties, engagement to learn, and relationships with peers and teachers. Journal of youth and adolescence, 52(2), 273-286.

[18] Noroozi, O., Pijeira-Díaz, H. J., Sobocinski, M., Dindar, M., Järvelä, S., & Kirschner, P. A. (2020). Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review. Education and Information Technologies, 25, 5499-5547.

[19] Xinhan, N. (2021). Intelligent analysis of classroom student state based on neural network algorithm and emotional feature recognition. Journal of Intelligent & Fuzzy Systems, 40(4), 7171-7182.

[20] Ruiz, S., Urretavizcaya, M., Rodríguez, C., & Fernández-Castro, I. (2020). Predicting students' outcomes from emotional response in the classroom and attendance. Interactive Learning Environments, 28(1), 107-129.

[21] Deniz, S., Lee, D., Kurian, G., Altamirano, L., Yee, D., Ferra, M., ... & Oh, P. (2019, January). Computer vision for attendance and emotion analysis in school settings. In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0134-0139). IEEE.

[22] Llurba, C., Fretes, G., & Palau, R. (2024). Classroom Emotion Monitoring Based on Image Processing. Sustainability, 16(2), 916.

[23] Huang, D., & Zhang, W. (2021). Research on learning state based on students' attitude and emotion in class learning. Scientific Programming, 2021(1), 9944176.

[24] Yuan, Q. (2022). Research on classroom emotion recognition algorithm based on visual emotion classification. Computational intelligence and neuroscience, 2022(1), 6453499.

[25] Ganesan, P., Jagatheesaperumal, S. K., Gobhinath, I., Venkatraman, V., Gaftandzhieva, S., & Doneva, R. (2024). Deep Learning-based Interactive Dashboard for Enhancing Online Classroom Experience through Student Emotion Analysis. IEEE Access.

[26] Liu, T., Wang, M., Yang, B., Liu, H., & Yi, S. (2025). ESERNet: Learning spectrogram structure relationship for effective speech emotion recognition with swin transformer in classroom discourse analysis. Neurocomputing, 612, 128711.

[27] Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi‐level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. Computational Intelligence, 36(2), 861-881.

[28] Li, M., Liu, M., Jiang, Z., Zhao, Z., Zhang, J., Ge, M., ... & Wang, Y. (2022). Multimodal emotion recognition and state analysis of classroom video and audio based on deep neural network. Journal of Interconnection Networks, 22(Supp04), 2146011.
[29] Lu, Y., Chen, Z., Zheng, Q., Zhu, Y., & Wang, M. (2023). Exploring multimodal data analysis for emotion recognition in teachers' teaching behavior based on LSTM and MSCNN. Soft Computing, 1-8.
[30] Xie, Y., Tian, W., Zhang, H., & Ma, T. (2023). Facial expression recognition through multi-level features extraction and fusion. Soft Computing, 27(16), 11243-11258.
[31] YUHONG ZHANG,QIAOHUI ZHANG,JUNJIE GU,HONGXIAO YU & DEMIN XU. (2024). U-NET MODEL BASED ON CBAM ATTENTION MECHANISM FOR CORONARY ANGIOGRAPHY SEGMENTATION. Journal of Mechanics in Medicine and Biology(09).
[32] Xinyao Hu,Shiling Yu,Jihan Zheng,Zhimeng Fang,Zhong Zhao & Xingda Qu. (2025). A hybrid CNN-LSTM model for involuntary fall detection using wrist-worn sensors. Advanced Engineering Informatics(PA),103178-103178.
[33] Sameh Abdelazim,David Santoro & Fred Moshary. (2025). FPGA Programming Challenges When Estimating Power Spectral Density and Autocorrelation in Coherent Doppler Lidar Systems for Wind Sensing †. Sensors(3),973-973.
[34] Peter Uhe,Chris Lucas,Laurence Hawker,Malcolm Brine,Hamish Wilkinson,Anthony Cooper... & Christopher Sampson. (2025). FathomDEM: an improved global terrain map using a hybrid vision transformer model. Environmental Research Letters(3),034002-034002.
[35] Mao WeiYang & Xia Jshardrom. (2022). Cross-modal representation learning based on contrast learning. (eds.)Shanghai University (China).
[36] Ning Ding & Knut Möller. (2024). Minimally Distorted Adversarial Images with a Step-Adaptive Iterative Fast Gradient Sign Method. AI(2),922-937.