ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Framework for Designing Digital English Classroom Activities Based on Multimodal Teaching Concepts

Dan Xie[1,*]

[1] School of Culture and Communication, Loudi Vocational and Technical College, Loudi, Hunan, 417000, China

SUMMARY: *The changes of students' learning condition in English class are transmitted through physiological data which are shown in many kinds of forms. Collecting, analyzing and using these data to identify the emotional changes in the students' learning process can provide a reference for teachers to reasonably design classroom activities. In this paper, we use Haar feature face detection method and OpenPose network structure feature recognition method to extract students' facial expression and behavioral posture feature data in English classroom. One model which recognizes emotion in learning has been built to complete the integration of multimodal data through the utilization of the multi-head attention mechanism. After that, it merges this with the time features that got from the long-short-term memory network to implement the recognition and classification of the emotions of students. This model, through continuous experiment, therefore, shows that the precision of its emotion recognition exceeds 85%. Through four times of examinations, the average score achievement of students in the experiment class, which is taught under the multi-modal teaching thought, rose to above 85 points. In all kinds of classroom activities, "acting English dramas and chanting English songs" has been proven to be the one with the biggest influence.*

KEYWORDS: *Haar features; OpenPose network; emotion recognition; multiple attention; long and short-term memory; English learning*

1 Introduction

Along with the progression of modern information technology, multi-modal teaching has gradually become a key point of study in the education field[1]. Multimodal teaching refers to the teaching mode that integrates the use of visual, auditory, tactile and other senses and symbolic resources for meaning construction in the teaching process [2]. In reality, the English classroom is still dominated by the traditional teaching mode, and there are some significant problems and limitations, which make the atmosphere of the English classroom dull and make it difficult to maintain students' learning interest [3, 4]. Firstly, the teaching process pays too much attention to the teaching of theoretical knowledge and neglects practical operation and skill training [5]. Secondly, there is a lack of teacher-student interaction and communication and discussion between students in the classroom, and students' participation is not high [6]. Meanwhile, the development and utilization of multimodal teaching resources are still relatively insufficient. Although some schools and teachers have begun to try to apply multimodal teaching in reading, writing and other aspects, such as using multimedia means to create a vivid teaching situation and stimulate students' enthusiasm for learning, this exploration

*xiedan_jc123@163.com
https://doi.org/10.65102/is2026584
Published: 30 April, 2026

is still relatively sporadic, and teachers' knowledge and mastery of the concept of multimodal teaching needs to be strengthened [7-9].

Under the condition that the attention to digital education is continuously increasing, the Ministry of Education (MOE) has put forward a number of policies whose goal is to promote the deep application of artificial intelligence and digital technology in the field of education[10]. In recent years, the “Artificial Intelligence Enabled Education Initiative” has been vigorously advocated, which explicitly calls for the extensive application of smart technologies in the education system to improve the quality of the supply of teaching resources and personalized learning experience [11, 12]. At the same time, the "intelligent learning assistant" system that is brought out by the intelligent education platform, together with the progress and application of artificial intelligence models, have the contribution to the more wide spreading of intelligent learning resources[13, 14]. This group of policies shows the urgent need that we must accelerate the modernization progress of education. The goal is to promote the effect of teaching and studying through technical and scientific ways, and also to promote individual study and promote the whole quality of students.

Under the wave of “digital technology + education”, modern digital technology has greatly promoted the development of multimodal teaching [15].A media network platform, it is designed to provide a very large number of language studying materials for teachers and students. These resources include teaching assist tools such as audio, video, cartoon movement, and pictures, together with multi-medium teaching course materials, so that the combination of linguistic modality and non-linguistic modality, greatly enriched the way and means of expression of meaning and communication, the traditional single-modal mode of teaching to the multimodal mode of teaching has become a major trend of teaching reform [16-18]. At the same time, emerging platforms such as webcasting, online cloud classroom, short videos, and self-media, which integrate static and dynamic resources into the teaching process create a good teaching atmosphere, fully mobilize the participation of learners' multiple senses, stimulate multi-level associations, and make the presentation of knowledge richer and more diversified, which greatly improves the motivation and participation of learners [19-21]. English teaching, as an important part of higher education, bears the important responsibility of cultivating internationalized talents and should comply with the national policy [22]. On the opposite side, the digital environment for English learning has the function of going beyond the traditional single method of teaching. It builds a deeply immersing teaching environment for learners, hence it promotes the level of teaching to reach the requirements of talent cultivation in the current age [23-24].

As digital teaching tools are widely used in English teaching, students' learning styles are gradually transformed, but the digital literacy differences between different students still exist, which is an important challenge for the transformation of the English classroom [25, 26]. Furthermore, in the course of using digital teaching tools to promote the transformation of junior middle school English classrooms, teachers' mastery of digital teaching resources is a key decisive factor that affects the quality of teaching [27]. In their teaching practice, some teachers still have problems such as improper selection of tools, lack of familiarity with their use, and inability to take multiple platforms into account, which leads to incoherent teaching sessions and decreased classroom teaching efficiency [28, 29]. Although the introduction of digital teaching tools into junior high school English classrooms can broaden and enrich the teaching content and the way of teaching presentation, the lack of integrated and systematic planning will easily lead to the fragmentation of teaching resources [30, 31].

In this paper, in the multimodal data extraction and preprocessing session, the Haar feature method is used to detect the four types of feature changes in the students' facial expressions, and the feature pixel sums in specific regions of the face are quickly calculated by the integral

map algorithm to obtain the facial change data during the students' learning process. Meanwhile, for the behavioral video images after denoising preprocessing, the OpenPose network structure is used to identify the behavioral pose images of the students, extract the behavioral feature data and predict whether the joints belong to the same person. In the multimodal data fusion and emotion recognition session, the learning emotion recognition model is constructed. The weights of various types of input modal data features are computed in parallel through the multi-head attention mechanism to complete the normalization process and feature fusion output. Combined with the long and short-term memory network to mine the temporal dependency relationship between each modal data feature, to capture the temporal dynamic characteristics of emotion change. The final output of emotion classification results.

2 Multimodal data extraction and emotion recognition modeling

2.1 Multimodal Data Extraction and Preprocessing

2.1.1 Face detection based on Haar features

The detection of faces acts as the beginning step in the procedure of recognizing facial expressions, and it possesses an important importance inside this whole process. Its core task is to accurately recognize the location of the face from the captured image and pass the recognized facial expression image to the image preprocessing stage. Therefore, in this study, Haar features are used to characterize faces in captured images.

Haar feature is a widely used feature description method in face detection and recognition. It captures local features such as image edges, texture, etc. by calculating the difference of pixel values in a specific region of the image.Haar features are classified into four categories: edge features, linear features, center features and diagonal features. Its feature values react to the process of image grayscale change, which can be a simple description of the face features through rectangular features, glasses are darker than the color of the cheeks, the color of the sides of the nose is darker than the bridge of the nose, and the mouth is darker than the surrounding color.

In addition, Haar employs an integral map for feature computation. Integral map (II) is a data structure used in image processing that quickly calculates the pixel sum of any rectangular region in an image. It is calculated as follows:

First, a matrix of the same size as the original image is created to store the integral map. For each pixel point (i, j) in the integral map, its value is equal to the sum of all pixel numbers inside the meta-image in the matrix region that extends from the top-left corner to (i, j), denoted:

$$I(i, j) = \sum_{x=0}^{i} \sum_{y=0}^{j} image(x, y) \tag{1}$$

where $image(x, y)$ is the pixel value of the original image at position (x, y).

The boundaries of the integral map are usually filled to 0.0, i.e:

$$I(0, j) = I(i, 0) = 0.0 \tag{2}$$

After that, each element of the integral map can be computed by adding the pixel values of

the corresponding positions of the original image plus the values of the three neighboring integral map pixels in the upper-left corner, above and to the left.

Once the integral map is constructed, the sum of pixels at any rectangular location can be quickly computed. For example, to calculate the sum of pixels within a rectangular region defined by four points (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) in an image. The following formula can be used:

$$Sum = I(x_4, y_4) + I(x_1, y_1) - I(x_2, y_2) - I(x_3, y_3) \tag{3}$$

In short, an integrogram is a fast algorithm for finding the sum of all pixels in an image by traversing the image only once, and this algorithm greatly improves the efficiency of the computation of image eigenvalues. Thanks to the use of integral map, Haar features can quickly calculate the pixel sum of any matrix region in the image. In addition, Haar features are robust to illumination changes, i.e., the computation performs well in complex backgrounds.

Therefore, in this study, Haar features are adopted to extract faces from the collected images. Research using OpenCV Haar features of XML files loaded face recognition model, the image into a gray image, using load classifier classifier. DetectMultiScale detect faces in images, scaleFactor parameter is used to specify the image in proportion to the size, minNeighbors is used to specify the number of neighboring elements that each candidate rectangle should retain, and minSize is used to specify the minimum size of the detected objects. A threshold judgment mechanism is added to each face, i.e., the width and height of the face region is checked to see if it is larger than 150 pixels, and if the face region is large enough, the detected region is cropped and resized.

2.1.2 Image pre-processing

In order to better satisfy the recognition of students' behavioral gestures in the classroom, the extracted video frame images need to be pre-processed: 1) to eliminate certain blurred images and retain some frame images with more distinctive features; and 2) since there is generally noise in the images, the images need to be de-noised. In practice, there are many types of noise in images, such as: Gaussian noise, white noise, etc., of which Gaussian noise is the most common type of noise. Therefore, in order to reduce the impact of noise in the image on the final recognition result, this chapter uses Gaussian filter to de-noise the image. The expressions of one-dimensional Gaussian distribution and two-dimensional Gaussian distribution are shown in the following equations (4) and (5):

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{4}$$

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

2.1.3 Behavioral signal extraction based on OpenPose network structure

Open Pose is suitable not only for single-person identification, but it can also be utilized in multi-person identification work. The external structure of the OpenPose model is presented in Figure 1. This model is divided into two stages: at the beginning, the first 12 layers of the VGG19 feature extraction network are used to extract features from the input image, which obtains the feature graph F. In the second step, the obtained feature map F is input into a two-

branch multi-stage convolutional neural network, in which the upper branch ($S(\cdot)$ part of Fig. 1) is used mainly to predict the body part positions in a set of 2D confidence maps, while the lower branch (the $L(\cdot)$ part of Fig. 1) is mainly used to predict a set of 2D vector fields of partial affinities, showing the affinities between joints (PAFs).

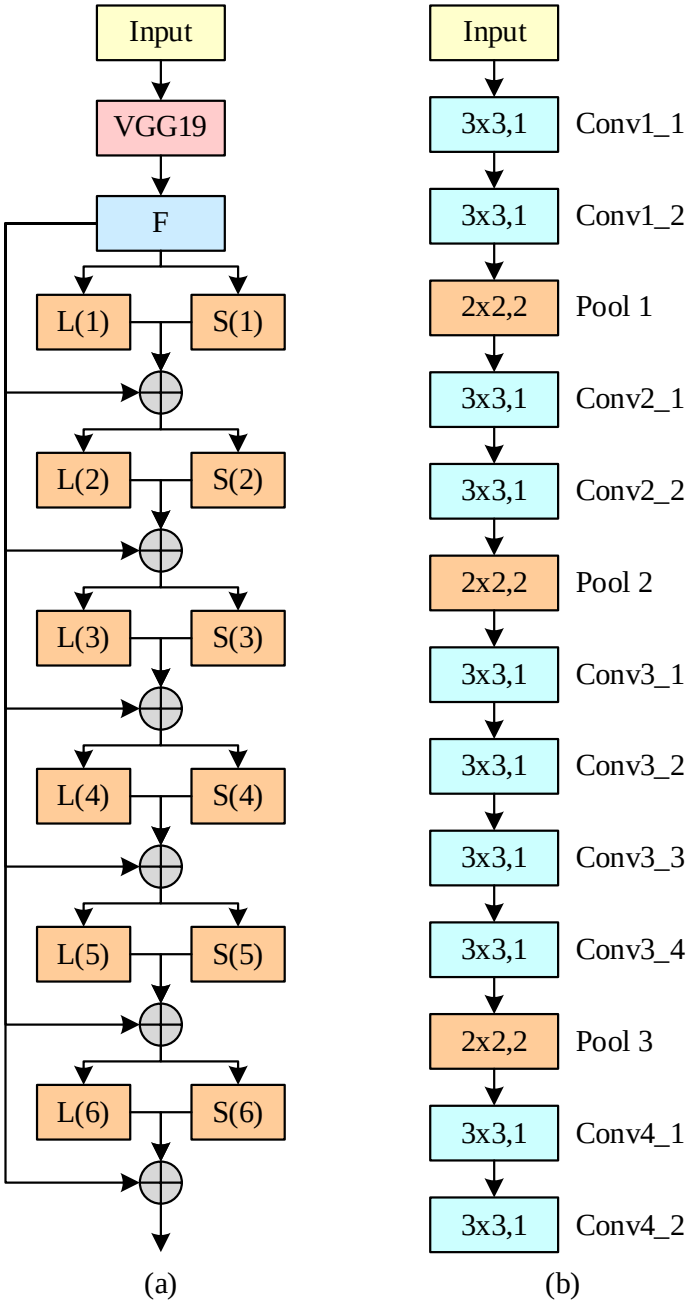

Figure 1: External structure diagram of OpenPose model

The application process of the OpenPose model is depicted in Figure 2. As can be inferred from Figure 2, the input for the initial phase of the OpenPose network is the feature map F, The data is processed by a series of Convolutional Neural Networks (CNNs) for generating the 2D confidence graph of the articulations. S^t and partial affinity L^t. Furthermore, beginning starting from the second stage, the network's input is constituted of a total of three constituent

parts., which are F, S^{t-1} and L^{t-1}, as shown in the following equation (6).

$$\begin{cases} S^t = \rho^t\left(F, S^{t-1}, L^{t-1}\right), & t \geq 2 \\ L^t = \phi^t\left(F, S^{t-1}, L^{t-1}\right), & t \geq 2 \end{cases} \tag{6}$$

This flow is completed through many repeated uses of the multi-step convolution nerve network, until the network gets to convergence. At last, in the prediction stage, the judgment of whether this joint belongs to the identical individual is completed through evaluating the affinity (PAF) between the joint pairs d_{j1} and d_{j1}, which is shown in the following equation (7).

$$E = \int_{u=0}^{u=1} L_c\left(p(u)\right) \cdot \frac{d_{j2} - d_{j1}}{\left|d_{j2} - d_{j1}\right|_2} du \tag{7}$$

At this place, $p(u)$ denotes the pixel point which is located between two continuous pixel points d_{j1} and d_{j2}, just as is shown in the equation (8) that is given below.

$$p(u) = (1-u)d_{j1} + ud_{j2} \tag{8}$$

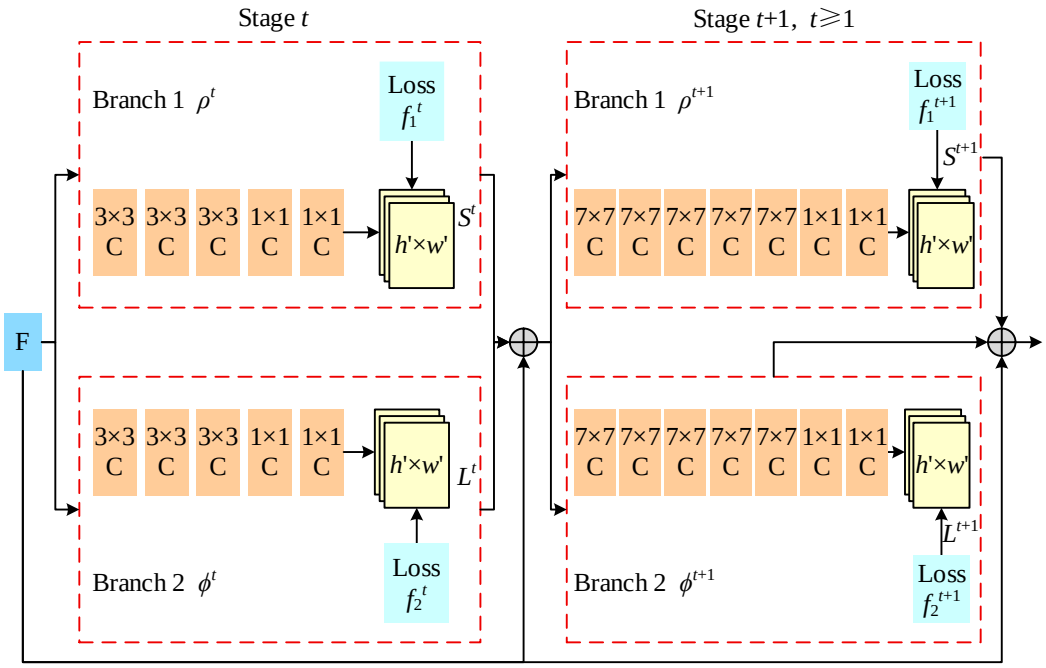

Figure 2: The application process of the OpenPose model

2.2 Multimodal Fusion and Emotion Recognition

Although students' emotional state can be identified through facial expressions and behavioral gestures. However, we know that human emotional expression is embodied in many ways, and using only one-sided features is not enough to fully and accurately express emotional information. Therefore, it is obviously not appropriate to consider only the information

contained in one side. Therefore, a specific explanation on multimodal fusion is developed by combining students' EEG signals and electrodermal signals.

2.2.1 Multi-attention mechanism design

At the core of the Multihead Attention Mechanism is the capture and fusion of the unique and complementary emotional information in each of the two signals, electroencephalographic (EEG) and galvanic skin response (GSR), whereas the time-domain features and frequency-domain features in the EEG signals are enriched with electrophysiological variations in brain activity, and the time-domain features in the GSR signals reflect the emotional fluctuations associated with the skin's electrical conductivity. These two signals have different importance in learning emotion recognition tasks, and the multihead attention mechanism is designed to accurately capture these differences and fuse them effectively.

Figure 3 shows the design flow of the multi-head attention mechanism. The multi-head attention mechanism designed in this study consists of the following main steps:

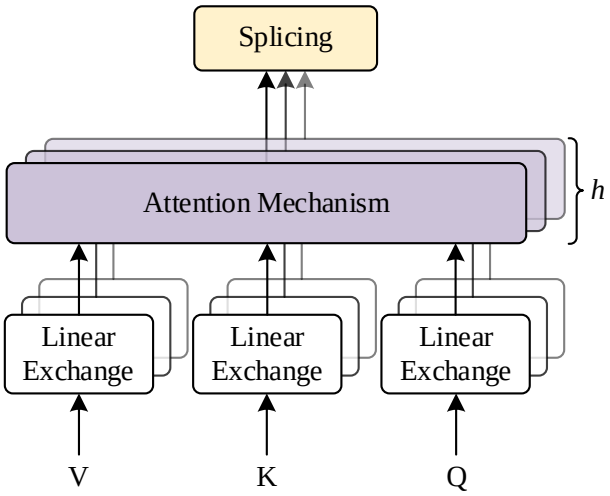

Figure 3: Design process of the multi-head attention mechanism

1) Multi-head structure construction. Construct multiple parallel attention heads, each responsible for a different subset of features, to capture more diverse information.

2) Calculation of Attention Weights. Within each attention head, the attention weights are computed by introducing a trainable parameter matrix. The parallel is computed using dot product attention, where the dot product between the query matrix Q and the key matrix K is first computed. Specifically, for each element in the sequence, its query vector is dot-producted with all key vectors to obtain an attention score matrix.

3) Weight normalization and feature fusion. To ensure the reasonableness of the weight assignment, the calculated attention weights were normalized using the SoftMax function. Subsequently, the extracted features of the EEG and dermatoglyphic signals were weighted and fused according to these weights, and the output of each head was calculated.In the end, the results that come from all heads go through a splicing process to produce the final feature representation.

Specifically, three matrices, Q, K, and V, are first defined, where Q is the matrix composed of query vectors, K is the matrix composed of key value vectors, and V is the matrix composed of output value vectors. Then the three matrices Q, K, and V for EEG signals and dermatoglyphic signals are:

$$Q_i = W_{qi} X_i \tag{9}$$

$$K_i = W_{ki} X_i \tag{10}$$

$$V_i = W_{vi} X_i \tag{11}$$

where W_{qi}, W_{ki} and W_{vi} are the weight matrices and $i = 1, 2, 3$ denote the three features.

After that the attention scores between each pair of signals need to be calculated:

$$Attention\left(Q_i, K_i, V_i\right) = soft\max\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j \tag{12}$$

where $i, j \in \{1, 2, 3\}$ and $i \neq j$; d_k is the dimensionality of the key vectors for the purpose of stability; and T denotes the transpose matrix.

Finally, the attentional outputs between all signals need to be fused, and these outputs are cascaded to obtain the final fused output features.

$$O = Concat\left(Attention\left(Q_i, K_i, V_i\right)\right) \tag{13}$$

where O is the fused output feature matrix.

2.2.2 Learning emotion recognition model design

This study focuses on constructing an efficient and accurate model for learning emotion recognition, which involves the design of the model structure, the selection of the loss function, the application of optimization algorithms based on the determination of the model evaluation method.

Long Short-Term Memory (LSTM) network is one special kind of recurrent neural network which is made for solving problems that have relation with long-sequence dependencies.The basic principle of LSTM lies in having a unique cellular structure, where each cell contains three control gates, namely, the input gate, the forgetting gate, and the output gate, as well as a memory cell for storing historical information. Therefore, when we construct the LSTM emotion recognition model, it is necessary that we design the input layer, the hidden layer, and the output layer. The input layer undertakes the responsibility of receiving the pre-processed and feature-extracted electroencephalogram (EEG) and electrodermal signals. For enabling the model to process these features in an effective way, the present research adopts appropriate feature encoding and normalization methods in the input layer, so that it can map feature values into the range which the model is able to handle.

The hidden layer is the core part of the LSTM model, and it undertakes the work of distinguishing the sequential connections between the input features.In this study, multiple LSTM structures are used to construct the hidden layer to increase the depth and complexity of the model. Each layer of LSTM contains multiple LSTM units, each of which controls the flow and update of information through input gates, forgetting gates, and output gates, and this structure enables the model to capture temporal dynamics in long sequential data and efficiently handle complex patterns in emotion recognition tasks.

In the output layer, this study employs a fully connected layer and a Softmax activation function to output the sentiment classification results. The whole connected layer possesses the function that maps the hidden layer's output to the dimension numbers of the sentiment sort

categories. At the same time, the Softmax function has the function of transforming the output into a probability distribution for the realizing of sentiment category evaluation. The design of the concrete learned mood identification model is displayed in Figure 4.

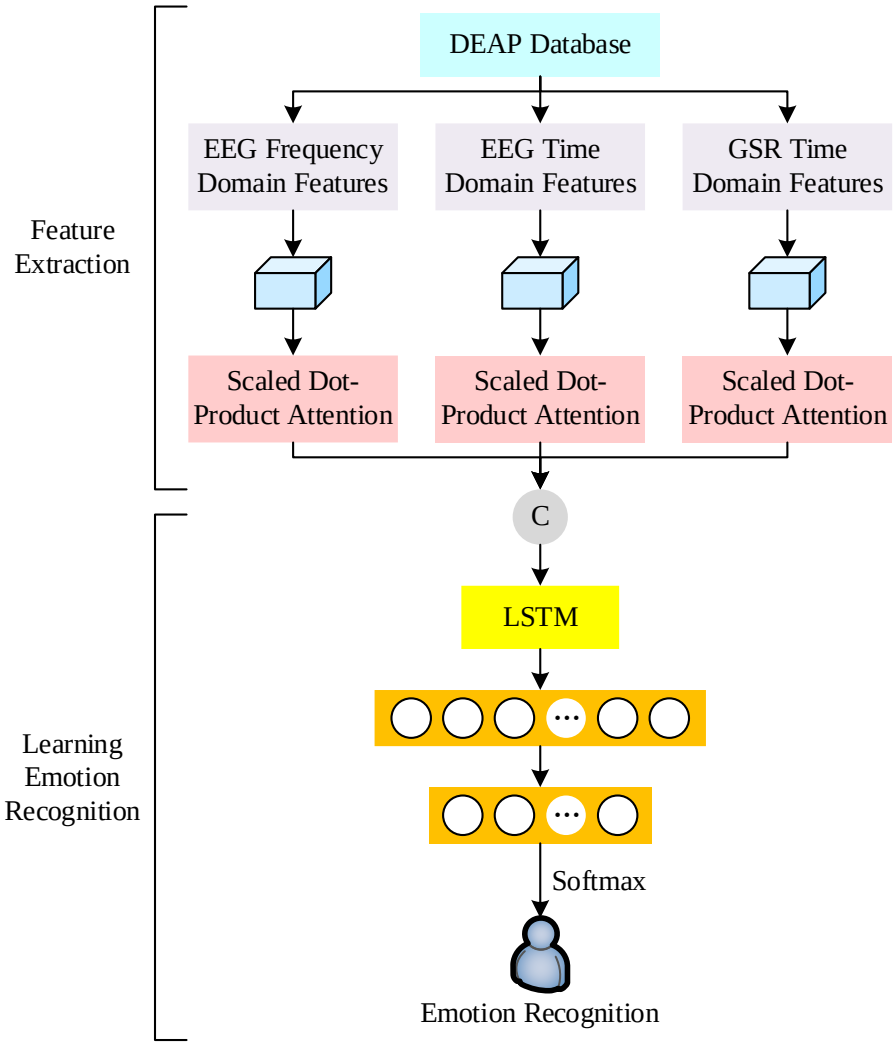

Figure 4: Physiological signal learning model for emotion recognition

3 The role of multimodal affective teaching in the digital English classroom

By extracting and analyzing students' multimodal physiological data while learning, students' emotional changes are identified in a timely manner. Under the guidance of this multimodal teaching concept, teachers design and continuously adjust the content and type of digital English classroom activities according to students' emotional state, making the diverse digital English classroom activities more in line with students' learning and enhancement needs. Multi-modal emotion-based teaching holds an important position in digital English classroom activities. Its influence mainly gets embodiment in a number of aspects. First of all, it promotes students' enthusiasm and motivation in the aspect of learning. Secondly, it can assist to enhance students' self-confidence and self-esteem. Thirdly, it can make students have more participation inside the classroom. At last, it promotes the growth of students' feeling cognition and emotional intelligence quotient.

3.1 Increasing students' interest and motivation in learning

Traditional teaching methods often focus on the inculcation of knowledge and ignore the emotional needs of students.Nevertheless, under the frame of the multimodal emotion teaching method, teachers pay attention not only to students' grasping of knowledge but also put emphasis on students' emotion participation in the study process. Through making active and attractive teaching activities and using visual teaching methods, teachers are able to hold students' attention and stimulate their passion for study. This good mood experience lets students have more willingness to take an active part in classroom activities, therefore it promotes students' study motivation.

The carrying out of multi-mode emotion teaching methods is also clearly seen in the building of the classroom environment. A happy and friendly classroom environment can effectively reduce students' study pressure and let them be more confident when putting forward their opinions and questions. Under this kind of environment, students already are not any longer the passive acceptors of knowledge. On the contrary, they become active participating persons, and their enthusiasm and motivation for study are without exception strengthened.

3.2 Enhancement of students' self-confidence and self-esteem

Multimodal affective teaching methods can promote students' self-confidence and self-esteem through actively paying attention to their emotional needs. Under the background of digital English classroom activities, some students have big probability to produce inferiority feeling, and encounter the problem of lacking self-confidence and self-respect, this is because their English basic knowledge is not good. The multimodal affective teaching strategy emphasizes paying attention to students' emotional needs, including understanding students' interests, hobbies, difficulties and psychological needs, actively communicating with students and helping them solve problems, so as to enhance their self-confidence and self-esteem.

3.3 Increase student participation in the classroom

Multimodal affective teaching emphasizes the enhancement of students' learning effect by increasing their classroom participation. In digital English classroom activities, multimodal affective teaching strategies can make students actively participate in the classroom according to their emotional needs by organizing rich and colorful teaching activities, creating a cooperative learning classroom atmosphere, and so on, so as to improve their participation in classroom activities. By increasing students' participation in classroom activities, it can effectively improve students' learning effect and language use ability.

3.4 Developing students' emotional literacy and emotional intelligence

Multimodal affective teaching aims to cultivate students' emotional literacy and emotional intelligence through classroom teaching activities. In digital English classroom activities, multimodal affective teaching strategy can introduce vivid and interesting teaching contents in time according to the students' emotional changes, organize role-playing, speech contests, group discussions and other teaching activities, and cultivate the students' sense of morality, aesthetics, rationality and other advanced emotions. At the same time, the multimodal emotion teaching strategy can also target the cultivation of students' emotion regulation ability and emotion communication ability, and improve students' self-knowledge and self-management ability. By cultivating students' emotional literacy and emotional intelligence, multimodal affective teaching strategies can promote students' comprehensive development and lifelong development.

4 English learning practices based on multimodal emotion recognition

4.1 Effectiveness test of emotion recognition model

4.1.1 Emotion recognition based on multimodal fusion

For evaluating the recognition effect of the constructed learning emotion recognition model, the dataset collected by ourselves is utilized to carry out emotion recognition. At the same time, one similar Arousal model which belongs to this type is selected to act as the reference model. After that, same emotion identification experiments are carried out on the dataset which is collected by ourselves. The emotion identification results of the two models which are based on multimodal fusion are shown in Figure 5. The self-acquisition dataset includes multimodal data such as facial expression features, behavioral posture features, and EEG signal features of 10 students during English classroom activities. The emotion recognition accuracy of the learning emotion recognition model in this paper after multimodal data fusion always maintains above 85%, with a maximum of 91.91% and a minimum of 87.47%, with little fluctuation in the emotion recognition accuracy for different students. Comparatively, the emotion recognition accuracy of the Arousal model is lower, mostly below 70%, and fluctuates greatly, with the highest being 80.74% for Student 2 and the lowest being 44.51% for Student 9. The stability and accuracy of the recognition performance of the Arousal model are not as good as that of the learned emotion recognition model in this paper.

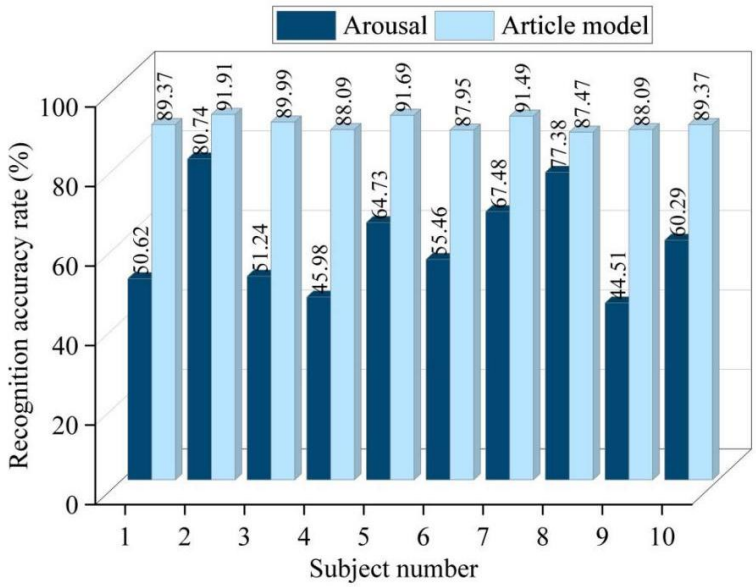

Figure 5: Emotion recognition results of 2 models based on multimodal fusion

4.1.2 Unimodal-based emotion recognition

In order to analyze the reason for such a large difference in recognition accuracy between the 2 models, the 2 models were used to perform unimodal emotion recognition on which face expression feature data and EEG signal feature data respectively. Figure 6 shows the comparison of emotion recognition results based on face expression.Figure 7 gives a showing of the results of emotion identification which depends on EEG signals. Under the situation of single-modality emotion recognition, the emotion-recognition model which is learned by this paper reaches a recognition accuracy that is more than 90% for the two sorts of single-modality

feature data. While the accuracy of emotion recognition of facial expression feature modal data of Arousal model is between 40% and 57%, the accuracy of recognition of EEG signal feature modal data is between 52% and 60%.The difference in the recognition correctness between the two models may come from the circumstance that the learning-based emotion recognition model put forward in this paper uses a multi-head attention mechanism. This mechanism is utilized for the extraction and standardization of the combination of various kinds of feature modal data, hence therefore making the relation among different data types more reasonable. By comparison, the Arousal model has difficulty in reaching the same accuracy level as the learning-based emotion recognition model, when we talk about the fusion processing of many kinds of modal data.

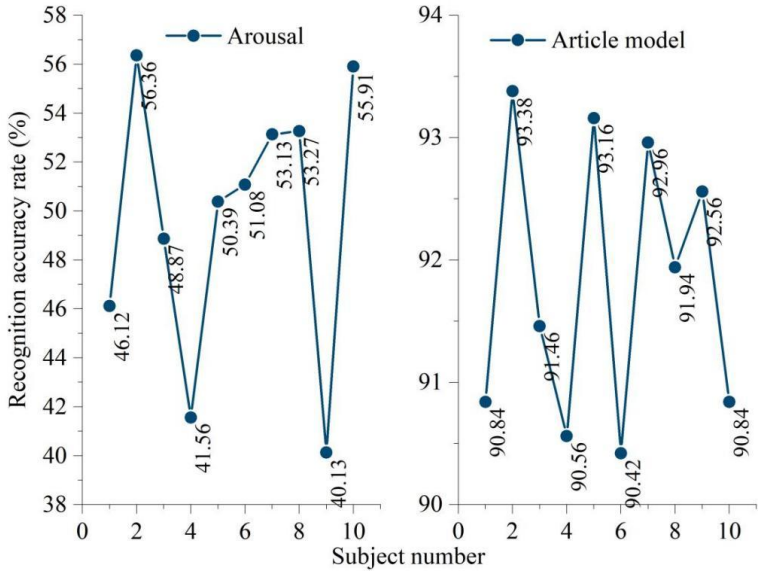

Figure 6: Result of emotion recognition based on facial expressions

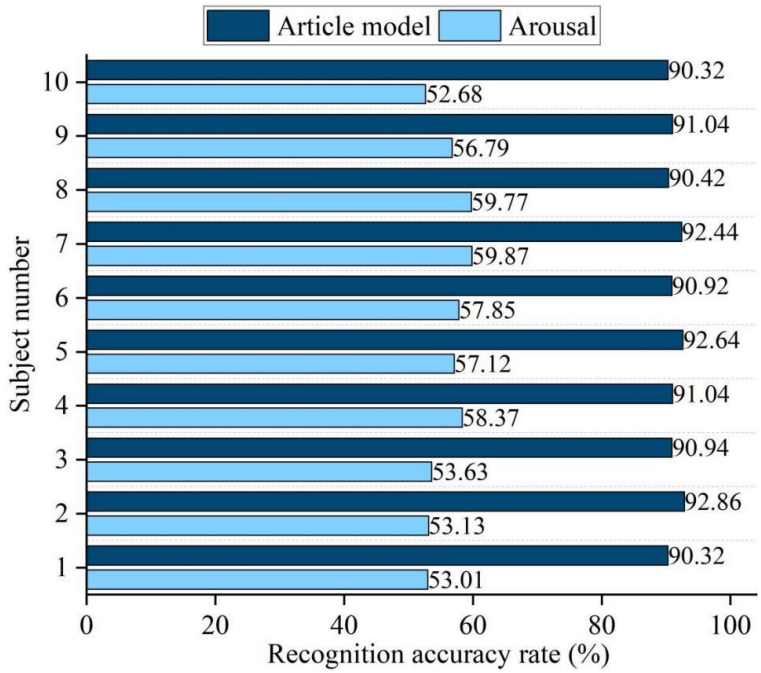

Figure 7: Result of emotion recognition based on brain electrical signals

4.2 Empirical Study of English Learning Effectiveness Supported by Multimodal Teaching Concepts

4.2.1 Comparison of English learning effects between the two classes before the experiment

After we finished the evaluation of the effect of the Learning Emotion Recognition Model, one comparative experiment was carried out. The goal of this experiment was to study the influence of putting into practice the multimodal teaching idea, which is built upon the Learning Emotion Recognition Model, in digital English classroom teaching activities.Two first-year classes of advanced translators in the School of Foreign Languages at the University of R were chosen as the experimental subjects. class A was used as the experimental class to assist in the design of English classroom activities using the multimodal teaching concept based on the Learning Emotion Recognition Model, and class B was used as the control class to complete the design of English classroom activities using the traditional method. There were 35 students in each of the two classes, and tests were set up before and after the experiment to examine students' English learning interest and motivation (L1), English learning self-confidence and self-esteem (L2), English classroom engagement (C1), emotional literacy and emotional intelligence levels (E1), and to determine the effects of the classroom activity design under different teaching concepts on the students.

Before the experiment, the descriptive statistics outcomes of the sub-groups inside the experimental class and the control class are shown in Table 1. Before we start to do the experiment, the average score values of the four test projects among students in the two classes were roughly in the range from 35 to 37. In addition, the standard deviation of all these score values was not larger than 0.3. This points out that there existed not big difference between the starting English learning study results and the related language ability levels of students in the experiment class and those in the comparison class. This kind of circumstance is very consistent with the requirements that are established for this experiment.

Table 1: Descriptive statistics of group divisions before the experiment

Testing items	Class	N	Mean	Std.Deviation	Std.Error Mean
L1	Experimental Class	35	35.48	0.39	0.26
	Control class	35	35.29	0.40	0.28
L2	Experimental Class	35	36.37	0.37	0.16
	Control class	35	36.40	0.37	0.17
C1	Experimental Class	35	35.48	0.35	0.19
	Control class	35	35.47	0.32	0.19
E1	Experimental Class	35	36.59	0.31	0.25
	Control class	35	36.56	0.32	0.24

4.2.2 Comparison of English learning effects between the two classes after the experiment

When the experiment has been completed, the results which are got from descriptive statistics for the grouping of the experimental class and the control class are shown in Table 2. Table 3 has shown the results of the independent samples test which was done between the experimental class and the control class after the experiment. In the experiment class, the average marks of all four test projects after the experiment were 85 or higher. By comparison, the average score points of the control class have only increased to be lower than 46. The difference in the mean scores of the four items between the two classes after the experiment is significantly higher than

that before the experiment, while the probability of significance of the two-tailed t-test Sig.(2-tailed) for both classes on these four scores is 0.01, which indicates that the two classes after the experiment have a high level of interest and motivation in learning English (L1), self-confidence and self-esteem in learning English (L2), participation in the English classroom (C1), and level of affective literacy and affective intelligence (E1) are significantly different on four variables. The class which took part in the experiment has obtained more excellent outcomes when put in comparison with the control group.

Table 2: Descriptive statistics of group divisions after the experiment

Testing items	Class	N	Mean	Std.Deviation	Std.Error Mean
L1	Experimental Class	35	85.38	0.37	0.10
	Control class	35	45.27	0.62	0.52
L2	Experimental Class	35	86.44	0.31	0.16
	Control class	35	45.01	0.66	0.79
C1	Experimental Class	35	87.19	0.30	0.15
	Control class	35	45.26	0.65	0.53
E1	Experimental Class	35	88.35	0.37	0.17
	Control class	35	45.16	0.61	0.62

Table 3: Independent sample test between 2 classes after the experiment

		Levences test for equality of variances		T-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std.Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
L1	Equal variances assumed	85.48	0.18	-3.24	35.00	0.01	-3.28	0.12	-1.03	-0.21
	Equal variances not assumed			-3.24	30.29	0.01	-3.28	0.12	-1.09	-0.20
L2	Equal variances assumed	80.21	0.36	-3.18	34.27	0.01	-3.01	1.36	-2.46	-1.25
	Equal variances not assumed			-3.18	30.05	0.01	-3.01	1.36	-2.46	-1.25
C1	Equal variances assumed	81.27	0.54	-3.27	31.74	0.01	-3.16	0.74	-3.28	-0.73
	Equal variances not assumed			-3.27	30.01	0.01	-3.16	0.74	-3.28	-0.73
E1	Equal variances assumed	82.19	0.63	-3.64	31.62	0.01	-3.27	0.51	-3.47	-0.18
	Equal variances not assumed			-3.64	31.62	0.01	-3.27	0.51	-3.47	-0.18

4.3 Analysis of the effectiveness of digital English classroom activities

For the purpose of evaluating the effect degree of the design of digital English classroom activities which are led by the multimodal teaching idea, one investigation has been done by us among the students that are in the experimental class. The aim of this investigation was to collect the first 10 classroom activity items which the students of the experimental class considered to be the most effective.Table 4 shows the data of analyzing the effectiveness of the 10 digital English classroom activities under the guidance of multimodal teaching concept.Among the 10 digital English classroom activities, students in the experimental class think the most effective one is A4 (staging English plays and songs), with a mean score of 4.75, followed by A1 (watching English TV programs and videos), with a mean score of 4.18.These two types of activities can quickly bring out students' emotional experiences, and therefore

enable them to maintain a relatively positive emotional state in the English classroom that is held by school. Therefore, this promotes students' enthusiasm and motivation in English study, and nurtures their emotional ability and emotional intelligence, in other aspects.

Table 4: Data on the effectiveness of 10 classroom activities(N=35)

Ranking	Activity	Minimum value	Maximum value	Average value	Standard deviation
1	A4	3.00	5.00	4.75	0.14
2	A1	3.00	5.00	4.18	0.21
3	A6	2.00	5.00	3.65	0.24
4	A3	2.00	4.00	3.46	0.27
5	A7	2.00	4.00	3.39	0.29
6	A9	2.00	4.00	3.27	0.30
7	A2	2.00	3.00	2.55	0.30
8	A10	1.00	3.00	2.30	0.32
9	A5	1.00	2.00	1.45	0.33
10	A8	1.00	2.00	1.24	0.35

5 Conclusion

In this research article, we have established a model which is used to identify the emotions that relate to learning. This model carries out analysis on the emotion undulations of students through the integration and processing of their physiological multi-modal data in the course of learning activities. According to the outcome of the analysis, we next design appropriate digital English classroom activities.The recognition accuracy of the model is [87.47,91.91] % in emotion recognition under multimodal fusion. After the idea of model-supported multiform teaching had been put into practice in the making of digital English class activities and teaching doing, the average mark of students in the experiment class went up from about 35-37 points to more than 85 points.In this controlled experiment, the activities of “staging English plays and songs” (4.75 points) and “watching English TV programs and videos” (4.18 points) were designed to be the most effective in improving the four qualities of the students. Through the inspection of the features of students' multi-modal physiology data, we can thus quickly make the judgment of students' emotional conditions. According to these results, classroom activities can be designed to adapt to the emotional conditions of students. This method can stimulate students to participate in the digital English class room, thus promoting their overall quality and study ability.

About the Author

Dan Xie was born in Loudi, Hunan Province, China in 1983. She is an associate professor in Loudi Vocational and Technical College. She received double bachelor's degrees in Literature and Economics from Jiangxi University of Finance and Economics, her master's degree from Central South University. Her research interests include English language teaching and translation.

References

[1] Lee, X. (2023). A Review of Middle School English Writing Instruction from the Perspective of Multi-modal Theory. International Journal of Social Science and Education Research, 6(6), 393-398.

[2] Zeng, R., & Wen, L. (2020, October). The Construction of an Effective Multi-modal English Oral Output Teaching Mode in the Cloud Environment. In Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education (pp. 420-424).

[3] Min, H. (2016). A study on silence phenomenon in college English classroom. International Journal of Education and Research, 4(6), 451-458.

[4] Niu, J., & Liu, Y. (2022). The Construction of English Smart Classroom Teaching Mode Based on Deep Learning. Computational Intelligence and Neuroscience, 2022(1), 9037010.

[5] Huang, R. (2022, March). Motivating EFL Students in Learner-centered Classroom. In 8th International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC 2021) (pp. 132-137). Atlantis Press.

[6] Li, Y., & Qu, C. (2019). College English Education Platform Based on Browser/Server Structure and Flipped Classroom. International Journal of Emerging Technologies in Learning, 14(15).

[7] Tseng, Y. H. (2025). Motivating rural EFL students in multimodal teaching. Linguistics and Education, 87, 101425.

[8] Freyn, A. L., & Gross, S. (2017). An empirical study of Ecuadorian university EFL learners' comprehension of English idioms using a multimodal teaching approach. Theory and Practice in Language Studies, 7(11), 984-990.

[9] Pan, J., & Zhang, L. Research on the Application of Multimodal Teaching Model in English Reading Teaching in Primary Schools——A Case Study of H Primary School in T City. International Journal of Social Sciences in Universities, 1.

[10] Xie, L., He, X., & Zhang, H. (2025). How Can the Diffusion of Digital Education Policies Be Advanced in Regions With Diverse Educational Conditions?—Qualitative Comparative Analysis of Chinese Policy. European Journal of Education, 60(3), e70188.

[11] Lameras, P., & Arnab, S. (2021). Power to the teachers: an exploratory review on artificial intelligence in education. Information, 13(1), 14.

[12] Yang, Y., Yuan, Y., Zhang, G., Wang, H., Chen, Y. C., Liu, Y., ... & Katabi, D. (2022). Artificial intelligence-enabled detection and assessment of Parkinson’s disease using nocturnal breathing signals. Nature medicine, 28(10), 2207-2215.

[13] Ma, X. (2024). High-quality Curriculum Resources Support the High-quality Development of Rural Education. Region-Educational Research and Reviews (10).

[14] Zhou, L., Meng, W., Wu, S., & Cheng, X. (2023). Development of Digital Education in the Age of Digital Transformation: Citing China's Practice in Smart Education as a Case Study. Science Insights Education Frontiers, 14(2), 2077-2092.

[15] Qiao, Y., Lv, N., & Zhang, J. (2023, December). Research on the Multi-modal Digital Teaching Model Centered on Higher Vocational Learners. In 2023 International Conference on Applied Psychology and Modern Education (ICAPME 2023) (pp. 54-61). Atlantis Press.

[16] Sarkar, S. (2012). The role of information and communication technology (ICT) in higher education for the 21st century. Science, 1(1), 30-41.

[17] Nie, H. (2023). College English teaching reform and innovative methods under the new media platform based on the IoT. Entertainment Computing, 47, 100578.

[18] Damayanti, I. L., Febrianti, Y., Suryatama, K., Dewi, F., & Lubis, A. H. (2025). " IS KANCIL KIND?": EXPLORING THE INTERPLAY OF VERBAL AND VISUAL TEXTS IN A PICTURE BOOK FOR TEACHING ENGLISH TO YOUNG LEARNERS. TEFLIN Journal: A Publication on the Teaching & Learning of English, 36(1).

[19] Wei, Y. (2024). Chinese and English text classification techniques incorporating CHI feature selection for ELT cloud classroom. Open Computer Science, 14(1), 20240007.

[20] Yangtao, C. H. E. N. (2025). Leveraging We-Media for Enhancing English Conversational Skills. Sino-US English Teaching, 22(7), 217-221.

[21] ChanLin, L. J. (2020). Engaging university students in an ESL live broadcast. The Electronic Library, 38(1), 28-43.

[22] Linsen, L. (2021). Reform and Innovation of Practical English Talents Training Mode in Colleges and Universities. Frontiers in Educational Research, 4(13).

[23] Jiang, L. (2017). The affordances of digital multimodal composing for EFL learning. Elt Journal, 71(4), 413-422.

[24] Jiang, L., & Ren, W. (2021). Digital multimodal composing in L2 learning: Ideologies and impact. Journal of Language, Identity & Education, 20(3), 167-182.

[25] Shopova, T. (2014). Digital literacy of students and its improvement at the university. Journal on Efficiency and Responsibility in Education and Science, 7(2), 26-32.

[26] Budiman, R., & Syafrony, A. I. (2023). The digital literacy of first-year students and its function in an online method of delivery. Asian Association of Open Universities Journal, 18(2), 176-186.

[27] Rahmawati, S., Abdullah, A. G., & Widiaty, I. (2024). Teachers' digital literacy overview in secondary school. International Journal of Evaluation and Research in Education, 13(1), 597-606.

[28] Zulkifli, N. A., & Hayati, M. (2021, September). Teacher Challenge and Tech Issues in Online Schools. In Eighth International Conference on English Language and Teaching

(ICOELT-8 2020) (pp. 14-18). Atlantis Press.

[29] Ourn, V., & Chhorn, T. (2025). Teachers' Challenges in Transforming the Digital Teaching in Cambodian Context. Journal of Education Innovation and Language Teaching (JEILT), 1(1), 76-87.

[30] Zhang, S. (2025). Analysis of the Impact of the Digital Tools on TESOL Teaching. In SHS Web of Conferences (Vol. 220, p. 04009). EDP Sciences.

[31] Zhang, C. The Impact and Evaluation of Digitalization in Physical Education on Students' Physical Fitness Development in Higher Education Institutions. International Journal of Education and Economics, 8.