A Neural Network for EEG Emotion Recognition that Combines CNN and Transformer for Multi-scale Spatial-temporal Feature Extraction

Zhangfang Hu, Haoze Wu, Lingxiao He

Abstract—**In recent years, emotion recognition based on EEG signals has received significant attention and research interest. EEG signals have the advantages of universality, spontaneity, and difficulty in deception, making them capable of accurately reflecting genuine emotional states. In this field, researchers have conducted binary (high/low) and ternary (low/medium/high) classification studies on the valence and arousal levels in the DEAP dataset.However, in order to better identify deep and intrinsic emotions, a clear definition of emotions becomes particularly important. Therefore, this study refers to Russell's Circumplex Model, which arranges emotions in a circular manner based on their valence and arousal levels. The study proposes placing emotion labels from the DEAP dataset within the two-dimensional emotional space of the circumplex model. Emotions are defined as four labels - Excited, Afraid, Sad, and Relaxed - based on a linear distribution of valence and arousal levels.Furthermore, a hybrid deep learning model combining CNN and Transformer is proposed for multi-scale spatial-temporal feature extraction. This model is employed for the classification of the four emotions. Finally, the model achieves an average accuracy of 91.26% on the four-class emotion classification task.**

Index Terms—**EEG signals, Circumplex Model, CNN, Transformer,Emotion Classification**

I. INTRODUCTION

EMOTION is a psychological state and response of individuals to stimuli based on subjective experiences in a particular context [1]. As a higher brain function, emotions profoundly influence our learning, work, and daily lives, making emotion analysis crucial. In 1997, the concept of Affective Computing (AC) was introduced by Professor Picard from MIT, which expanded the scope of emotion research beyond traditional fields by empowering computers with the ability to classify human emotions, thereby promoting more natural human-computer interaction [2][3].

E. Kroupi et al. conducted a classification study on three emotions: sadness, joy, and neutrality, using Linear Discriminant Analysis (LDA). The experimental results showed that neutral emotions were more prone to misclassification, while joy and sadness were relatively easier to recognize accurately [4]. Muhammad Zubair et al. employed Gaussian kernel support vector machines (SVM) for binary classification of EEG-based emotions. They compared SVM algorithms with different kernel functions and found that the Gaussian kernel SVM performed the best in terms of classification accuracy [5]. Additionally, Xie Qiao et al. proposed a classification method based on the XGBoost model and random forest model for EEG-based emotion recognition. The experimental results showed that the average recognition rates of this classification model reached 77.19% and 79.06% for the arousal and valence dimensions, respectively. This indicates that combining classifiers has advantages over using a single classifier in emotion classification tasks [6].

With the development of deep learning techniques, researchers have shifted their focus from machine learning to deep learning and applied it to the field of EEG-based emotion recognition, achieving significant progress.

In 2015, Zheng Weilong et al. divided the EEG signals into five rhythms: δ、θ、α、β and γ, and then extracted the DE features from different frequency bands. They conducted classification research on three emotional states: positive, neutral, and negative, using SVM and Deep Belief Networks (DBN). The average classification accuracy of DBN was found to be 86.08%, while SVM achieved an average accuracy of 83.99%. This indicates that DBN outperforms SVM in emotion classification tasks [7].In 2016, N. Thammasan et al. used DBN for classifying music-induced EEG emotions. They found that DBN effectively improved the classification accuracy of FD, PSD, and discrete wavelet features, with the highest accuracy reaching 86%. This further confirmed the effectiveness of DBN in emotion classification [8].In 2017, Li Jinpeng et al. extracted DE features from a 62-channel EEG signal and mapped them into an 8x9 feature map based on electrode placement. They then expanded the feature map to 20x20 using sparsity. Finally,

Manuscript received on March 15, 2024; revised on June 18, 2024. This work was supported in part by the Youth Fund Program of the National Natural Science Foundation of China (Grant No. 61703067), the Chongqing Basic Science and Frontier Technology Research Program (Grant No. Cstc2017jcyjAX0212), and the Science and Technology Research Program of Chongqing Municipal Education Commission (KJ1704072).

Zhangfang Hu is a Professor at the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (e-mail: 3565207151@qq.com).

Haoze Wu is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (corresponding author phone: 157-5610-5026; e-mail: s220431102@stu.cqupt.edu.cn).

Lingxiao He is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (e-mail: s220431028@stu.cqupt.edu.cn).

they built a 5-layer CNN network to classify the three emotional states, achieving an average recognition rate of 88.2%. In the same year, Wen Zhiyuan et al. directly used the raw EEG signals from 32 channels as inputs for CNN-based emotion recognition. They rearranged the EEG channels based on Pearson correlation coefficients and found that the highest average recognition rate was achieved when using the maximally adjacent arrangement. The average recognition rates for the valence and arousal dimensions in binary classification reached 77.98% and 72.98%, respectively [10].

In 2018, K. Yea-Hoon et al. utilized wavelet transform to convert EEG signals into color maps with a resolution of 42×200, and then used CNN to accurately extract and classify the features of EEG signals on these color maps. This innovative method resulted in an average recognition rate of 73.4% for four emotional states, providing a new perspective for emotion recognition research [11]. In 2020, Du X et al. proposed a hybrid model (ATDD-LSTM) based on attention mechanism and LSTM, which effectively characterized the spatial features of functional relationships between EEG signals at different electrodes and automatically selected suitable EEG channels for emotion recognition [12]. In 2021, An Y et al. introduced an emotion recognition model that combined spatiotemporal convolutional networks, leveraging CNN to extract spatial features and using LSTM to capture temporal features, effectively improving the accuracy of emotion recognition [13]. Also in the same year, Gao Z et al. designed an emotion recognition model based on Multi-layer Convolutional Neural Network (MNCNN) and differential entropy, achieving a classification accuracy of 91.45% on the SEED dataset [14]. In 2022, Li Yang et al. proposed a model called Bidirectional Domain Adversarial Neural Network (BiDANN), which enhanced the accuracy of emotion classification by extracting asymmetrical features from the left and right hemispheres of the brain [15].

In 2023, Yonghao Song et al. presented a compact Convolutional Transformer (EEGC) aimed at encapsulating local and global features within a unified EEG classification framework. The EEGC is an efficient decoding method for EEG data, combining the strengths of CNN and Transformer to achieve significant performance improvements across different EEG datasets, and visually demonstrating the model's ability to represent global features [16].

Currently, most researchers have focused on binary or ternary classification of EEG emotion signals, with limited exploration of four-class emotion classification. Because the training of quadruple classification requires a large number of data samples [17], and the commonly used DEAP dataset only provides labels such as valence and arousal, which are insufficient for four-class emotion classification. Therefore, this study redefined the four emotional labels (Excited, Afraid, Sad, Relaxed) based on the valence and arousal labels in the DEAP dataset, and increased the number of training samples using a sliding window method, resulting in a training sample size four times larger than the original dataset, facilitating better deep learning tasks.

In the field of EEG emotion recognition, models are transitioning from traditional machine learning to deep learning, with CNN models making significant contributions [18]. While CNNs have the advantage of capturing local receptive field information, they often ignore global information and have limited ability to extract temporal information, thus constraining their performance. Therefore, this study designed a hybrid module that combines CNN and Transformer to leverage the advantages of perceiving global information, and created a parallel network model capable of extracting multi-scale resolution features to better perceive deeper features of EEG signals. This integration successfully improved the training efficiency and accuracy of the four-class classification.

The main contributions of this paper are as follows:

1.The reconstruction of emotion labels in the DEAP dataset, enabling it to be used for a four-class classification task.

2.The proposal of a multi-feature extraction module based on Transformer, named C-T Block. This module effectively combines CNN and Transformer, utilizing the advantages of local perception from CNN and global perception from Transformer, thereby extracting more feature information and improving the accuracy and efficiency of emotion recognition.

3.The introduction of a network model for extracting deep features to capture different depths of brainwave signals. The incorporation of this model effectively enhances the spatial, spectral, and temporal resolution of brainwave signals, achieving satisfactory results.

The structure of the paper is as follows: The first part introduces relevant research on EEG-based emotion recognition. The second part describes the preparation for experiments, including preprocessing the DEAP dataset and calibrating emotion labels. The third part presents the proposed methods for experiments, including the fusion module of CNN and Transformer and a novel parallel network model. The fourth part presents the experimental results and provides a detailed analysis and comparison of the proposed deep learning models on the DEAP dataset. The fifth part is the conclusion.

II. Experiment preparation

A. DEAP data set

The DEAP database was developed in collaboration by Queen Mary University of London, Trent University, University of Geneva, and others, focusing on the study of physiological signals for emotion analysis [19]. The database collected EEG data from 32 participants, who were asked to watch 40 music videos to elicit different emotions. The research team meticulously recorded participants' physiological changes, including peripheral physiological signals, EEG signals, and facial expression data from 22 participants. The data collection included a 3-second baseline recording during video transitions and a 60-second experimental recording while watching the videos. Participants provided subjective ratings for valence, arousal, dominance, and liking based on the videos they watched. The EEG signal acquisition followed the 10-20 system method recommended by the International Federation of Clinical Neurophysiology [20], ensuring precise electrode placement, which provides strong support for related studies, as shown in Figure 1.

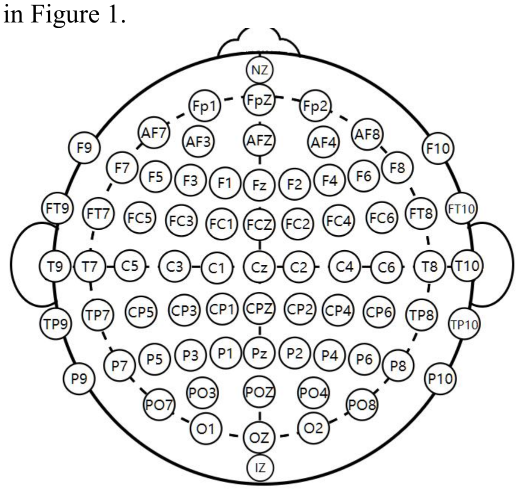

Fig. 1. 10-20 System channel

In Figure 1, the combination of letters and numbers represents different electrode positions for EEG channels. After the experiments, the research team generated data files for each of the 32 participants based on their respective ratings, enabling further analysis and study. These data files not only provide detailed records of various data during the experiment but also serve as rich resources for subsequent research. To improve data processing efficiency, the research team downsampled the original data by reducing the sampling frequency from the original high frequency to 128 Hz. This downsampling retains the main characteristics of the data while reducing the complexity of analysis. Table 1 provides a detailed description of the DEAP dataset.

In Table 1, based on the duration of EEG signal collection and the specific positions of electrodes, the data file for each participant is presented as a specific data structure: a 40×40×8064 data matrix. This matrix details 8064 data points generated on 40 channels by 40 different music videos, providing a comprehensive reflection of the participants' physiological responses while watching each video. Additionally, each data file also includes a 40×4 label matrix. This matrix records four rating metrics corresponding to each video, namely valence, arousal, dominance, and liking scores. These labels serve as the basis for researchers to evaluate participants' emotional responses, enabling researchers to understand the impact of different videos on participants' emotions. Among the 40 channels, there are 32 EEG channels and 8 other channels, including common signals like EOG and ECG. In this experiment, only the data from the 32 EEG channels were used in the study.Table 1 provides a detailed description of the DEAP dataset.

TABLE Ⅰ

DETAILED DATA FORMAT OF DEAP DATASET

Type	Discription
Subject	32(16 men and 16 women)
Number of people	32
Sample rate	128Hz
Triggering condition	40 different movie clips
Data shape	(40,32,8064)

B. *Data preprocessing*

In the data preprocessing stage, the analysis and processing of EEG signals posed significant challenges due to the inclusion of unrelated signals such as EOG (electrooculogram). To address this challenge, this study employed the ICA method to remove EEG artifacts from the filtered signals.

As EEG data collection is difficult, time-consuming, and carries high ethical and safety risks [21], the available EEG signal data is limited, resulting in a small database that does not meet the requirements for large-scale deep network analysis. This presents a high risk of overfitting, making data augmentation necessary. This study utilized a sliding window approach for data augmentation, designing a 10-second window as shown in Figure 2, and then sequentially sliding it at intervals of 2.5 seconds. The processed DEAP dataset resulted in 63 seconds of EEG signals. Starting from 0 seconds, a set of EEG data was extracted for each sliding window, representing the data from 0 to 10 seconds as the first set, and from 52.5 seconds to 62.5 seconds as the last set. With 21 sliding steps, one set of EEG data was transformed into 21 sets, thereby providing a larger data source for the training process, as illustrated in Figure 2.

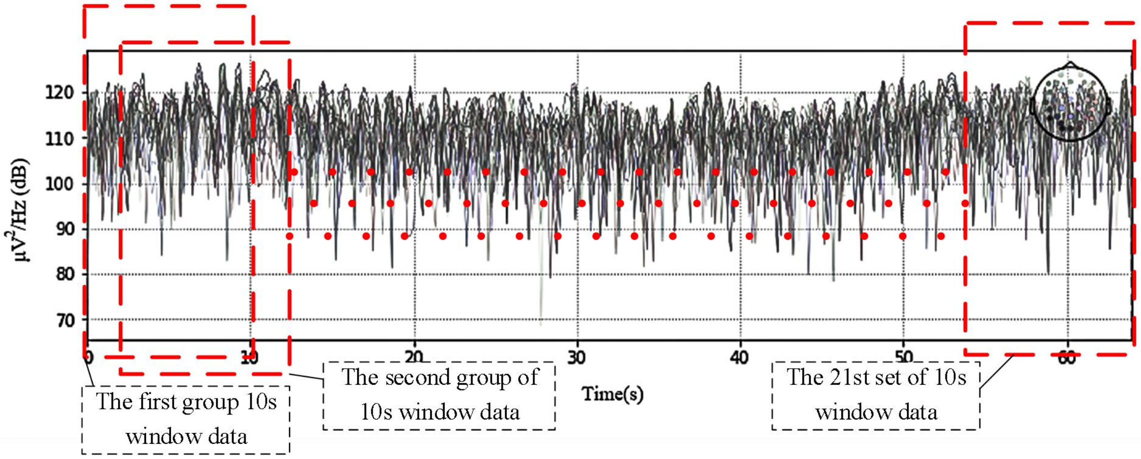

Fig. 2. Processing EEG data by sliding window

After applying a sliding window to the EEG signals, this study integrated the EEG data from 32 subjects, resulting in 40×32×21 training data points. Here, 40 represents the signals collected from each subject while watching 40 movies, 32 represents the 32 subjects, and 21 is the expansion factor for the data. In total, 26880 training data points were obtained.

After obtaining the integrated EEG data, this study further applied the Short-Time Fourier Transform (STFT) to these data. STFT has numerous advantages in processing EEG signals, including high time-frequency resolution, adaptability to non-stationary signals, intuitive visualization, and convenience for feature extraction and signal processing. The STFT formula is as follows:

$$STFT_{\chi}(\mathrm{t},\omega)=\int_{-\infty}^{\infty}\chi(\tau)\omega^{*}(\tau-t)e^{-j\omega\tau}d\tau \qquad (1)$$

In the equation:ω represents the angular frequency, *x(τ)* represents the value of the original signal in the time domain, *ω(τ-t)* represents the window function shifted at time *t*, * denotes conjugation.

Through STFT, this paper can effectively capture the temporal and frequency features of the signal, transform them into spectrogram data, and input these data into a deep learning network for further analysis and processing. This series of operations help to more comprehensively explore the information in EEG signals, providing strong support for subsequent research and applications.As shown in Figure 3, it is the time-frequency graph obtained by STFT transformation of EEG signals.

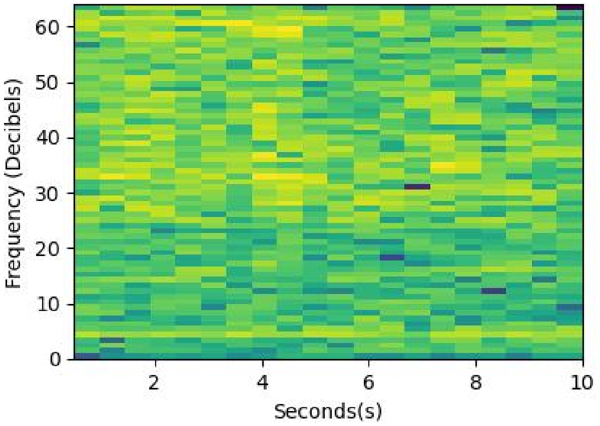

Fig. 3. Time-frequency diagram

C. Affective labeling

Russell's circular model of emotions, established on the dimensions of valence and arousal, provides a unique perspective for the study of emotions [22]. The valence dimension measures the positive or negative tendency of emotions, while the arousal dimension reveals the intensity or activation state of emotions [23]. When these two dimensions are combined, they constitute a circular emotional space that covers states ranging from calmness to highly positive or negative emotions. During the observation of subjects watching emotionally-induced videos, this paper found that their emotional states exhibited a continuous distribution pattern on Russell's circular model of emotions, indicating a significant correlation between the valence and arousal dimensions in the model and the subjects' physiological data. For the DEAP dataset, this discovery provides strong support for the re-calibration of labels in this paper. By applying Russell's circular model of emotions, it is possible to more accurately understand and classify the emotional labels in the dataset, as shown in Figure 4.

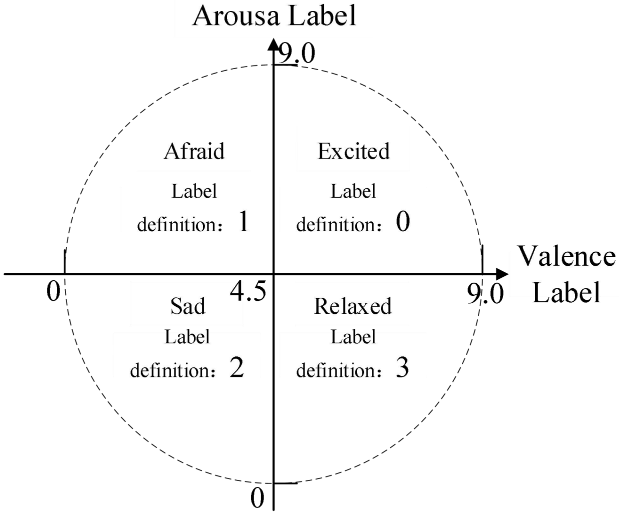

Fig. 4. DEAP tags define Valence and arousal

In the rating matrix (40×4) of the DEAP dataset, each video is scored by subjects on a continuous 9-point scale for valence, arousal, dominance, and likability [24]. This paper primarily focuses on the two dimensions of valence and arousal, and integrates these labels into Russell's circular model of emotions. Specifically, based on the original categorical labels and utilizing a label encoding method, we define labels with a valence score greater than 4.5 and an arousal score greater than 4.5 as "Excited" with a label of 0. Labels with a valence score less than or equal to 4.5 and an arousal score greater than 4.5 are classified as "Afraid" with a label of 1. Labels with a valence score less than or equal to 4.5 and an arousal score less than or equal to 4.5 are labeled as "Sad" with a label of 2. Finally, labels with a valence score greater than 4.5 and an arousal score less than or equal to 4.5 are categorized as "Relaxed" with a label of 3.

III. Proposed method

The text describes a neural network architecture for processing preprocessed EEG data of video stimuli. The data is fed into a deep convolutional module consisting of 5 layers, where each layer passes the convolved data to a feature extraction module. Upon receiving the data, the feature extraction module performs deep feature extraction tasks through 5 parallel channels composed of C-T modules. Due to variations in input data sizes for each channel, average pooling layers are incorporated with different kernels to ensure uniform feature sizes across all channels. Finally, the extracted features are inputted into a classification module, which flattens and combines the features before passing them through fully connected and softmax layers for classification output. The proposed network structure is illustrated in Figure 5.

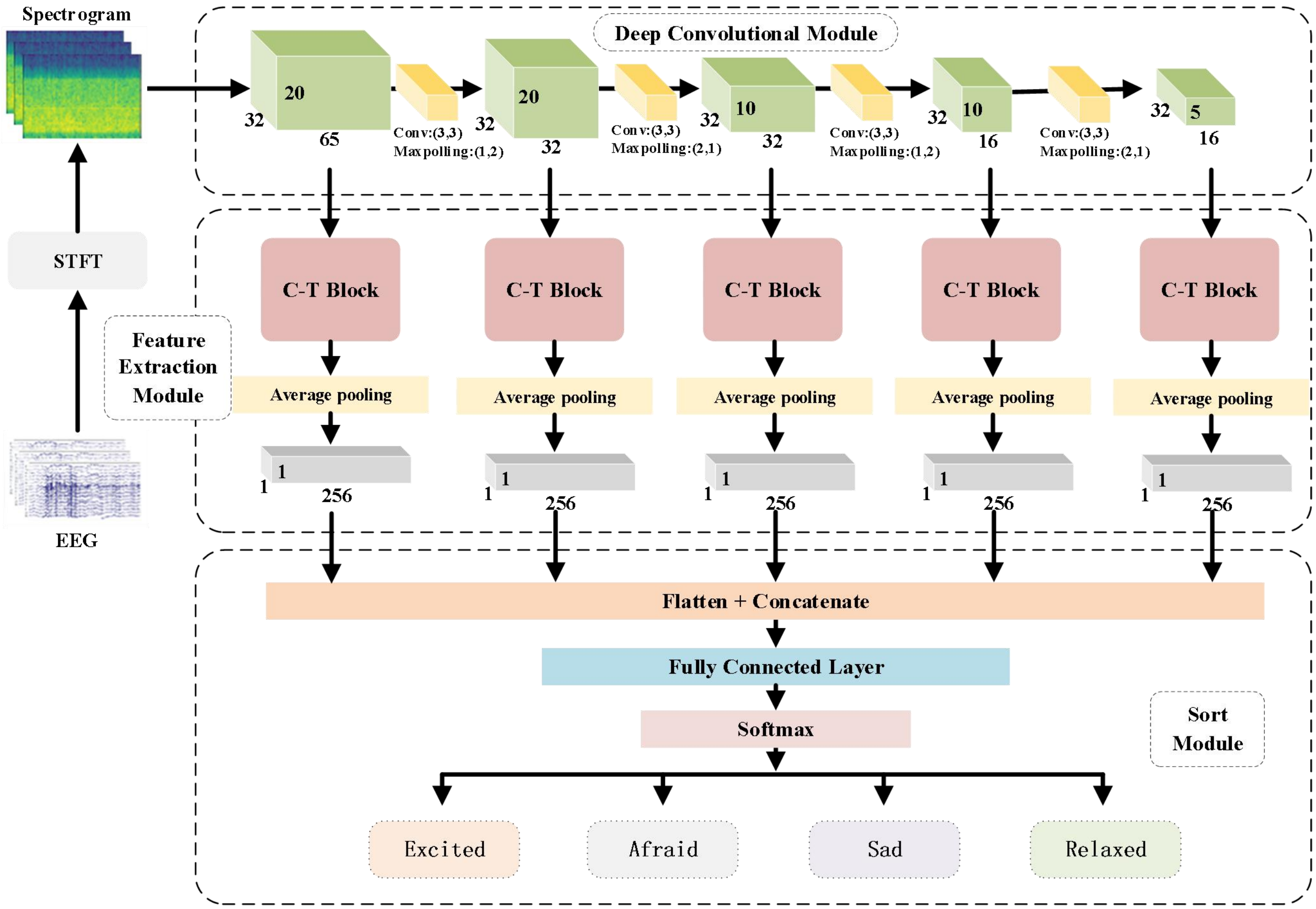

Fig. 5. Overall system framework

Below are detailed descriptions of the three modules: deep convolutional module, feature extraction module, and classification module.

A. *Deep convolutional module*

The effectiveness of multiscale EEG detection has been confirmed by multiple researchers. The deep convolutional module in this study is inspired by the feature pyramid technique in image processing. It utilizes deep convolutional layers and pooling layers to divide the input signal into multiple scales with different resolutions. The designed deep convolutional module consists of five consecutive layers, where each layer is designed based on the previous layer's resolution being halved in order to capture features effectively. This module automatically learns weights and extracts valuable features within each channel while downsampling the signal by half.

In this module, the shape of the input data is (32, 20, 65), indicating a depth of 32, height of 20, and width of 65. All experiments in this study correspond to Figure 1. In each layer, convolution is performed first with a kernel size of 3×3, stride of 1, and padding of 1. The output calculation of the convolutional layer is as follows:

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * \omega_{ij}^l + b_j^l\right) \qquad (2)$$

In the equation: χ_i^{l-1} is the region corresponding to the *i*-th convolution kernel of layer 1, χ_i^l is the *j*-th feature map of layer 1, *M* is the feature input map, ω is the weight matrix of the convolution kernel, *b* is the bias, *f* is the activation function, and * is the convolution operation.

After the convolution operation, the output passes through a max pooling layer with a kernel size of 2×1. When entering the next pooling layer, the kernel size is switched to 1×2. The calculation for the pooling layer is as follows:

$$x_j^l = down_{\max}(x_j^{l-1}) \qquad (3)$$

In the equation: $down_{max}$ represents the max pooling function, and χ_j^l is the output feature map of the pooling layer.

Through this operation, the data volume of layer x becomes half of layer x-1. Each convolutional layer independently extracts features without information exchange between different channels, aiming to capture information of different depths in EEG signals. The depth information extracted from each channel is then transmitted to the next module.

B. *Feature extraction module*

Most current researches use simple concatenation for feature fusion, without considering the different impacts of various features on classification results. Additionally, there is a lack of interaction and fusion of information among different features, which leads to many networks not significantly improving the robustness and accuracy of classification results after extracting multiple features. This paper proposes the C-T module to address this issue, which consists of CNN and Transformer. Firstly, the outputs of the deep convolutional module are separately connected to the CNN layer and Transformer layer. Then, the extracted feature information from these two layers is combined to facilitate the interaction of local and global features and weight adjustment. The structure of the C-T module is shown in Figure 6.

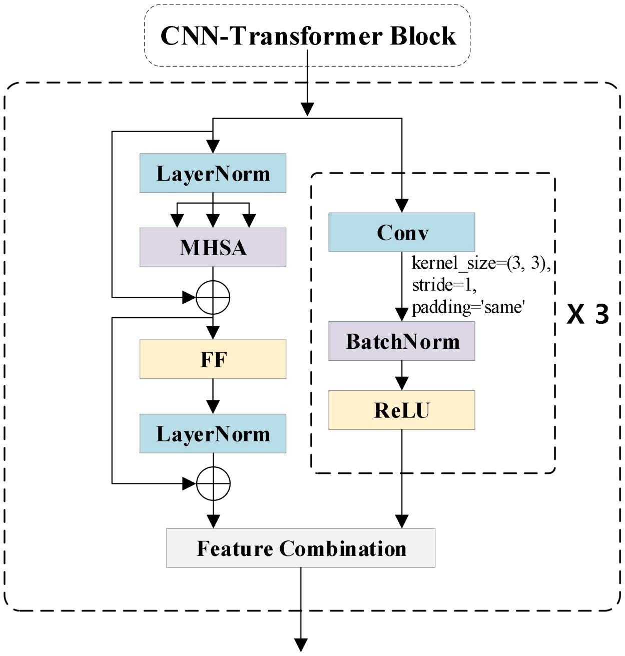

Fig. 6. Design CNN-Transformer hybrid module

The CNN part consists of three convolutional layers with a kernel size of 3x3, a stride of 1, and padding of 1. Following each convolutional layer is a Batch Normalization (BN) layer and a ReLU activation layer. Each convolutional layer conducts deeper feature extraction on the input features, until the three layers of convolution complete the extraction of local features. The role of the BN layer is primarily to optimize the training process of the neural network, improve the model's performance and generalization ability. The normalization formula is as follows:

$$\mu_\beta = \frac{1}{\mathrm{m}}\sum_{i=1}^{m} x_i \tag{4}$$

$$\mathrm{y} = \frac{x_i - \mu_\beta}{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_i - \mu_\beta\right)^2 + \varepsilon}} * \gamma + \beta \tag{5}$$

In the equation: x represents the batch input data, m represents the current batch size, ε represents a small value added to the variance to prevent division by zero, γ represents the trainable scale parameter, and β represents the trainable bias parameter.

The Transformer part consists of a Multi-Head Self-Attention (MHSA) module, a feed-forward module, and two layers of LayerNorm normalization modules[25]. The two normalization layers are placed before the MHSA module and after the feed-forward module. The calculation method for the attention mechanism is as follows:

$$Attention(\mathrm{Q}_{itf}, \mathrm{K}_{itf}, V_{its}) = Soft\max\left(\frac{Q_{itf}K_{itf}^{T}}{\sqrt{d_k}}\right)V_{its} \tag{6}$$

In the equation: Q_{itf}, K_{itf}, and V_{itf} represent the query vector, key vector, and value vector inputs in the attention module, *Softmax(·)* represents the Softmax function, and d_k represents the dimension of K_{itf}.

In the feature extraction module, this paper utilizes 5 C-T modules. Since the input data size of each module is different, each module will be adjusted internally based on the input data size. The output data size will also change according to the input data. Therefore, an average pooling module is added after the C-T modules to unify the output data size of the 5 layers of C-T modules, resulting in feature data with an output size of (256, 1, 1).

C. Classification Module

The module first flattens and combines the 5 sets of data output by the previous module to form a one-dimensional data of length 1280, which is then input to the fully connected layer. It consists of four FC (fully connected) layers with sizes of 1280, 320, 80, and 4 respectively. Finally, it directly inputs to the LogSoftmax layer. LogSoftmax calculates the logarithm of the predicted probability for each class. While stable gradient descent calculation, LogSoftmax heavily penalizes highly incorrect classes, further optimizing training time. The formula for LogSoftmax calculation is as follows:

$$LogSoft\max(x_i) = \log\left(\frac{\exp(x_i)}{\sum_j \exp(x_i)}\right) \tag{7}$$

In the equation: x_i represents the i-th element of the input vector x.

IV. EXPERIMENT

A. Experimental setup

The article was deployed on an NVIDIA RTX 2080 GPU server and the model was trained and tested using the PyTorch framework. During training, Adam optimizer was chosen with a learning rate of 0.0001, and cross-entropy loss function was utilized. For the DEAP dataset, a batch size of 16 was selected for training with 50 iterations. To prevent overfitting, a dropout rate of 0.2 was implemented.

In this study, classification accuracy and cross-entropy loss function value were used as evaluation metrics for the model. The cross-entropy loss function is a metric that measures the degree of error of the model. A higher value indicates a poorer model performance. The expression for the cross-entropy loss function is:

$$L = -\frac{1}{N}\sum_{\mathrm{i}}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{8}$$

In the equation:M represents the number of categories of the classification, i represents the number of categories of the classification, y_{ic} represents the probability of the class i of the real label, and p_{ic} represents the probability of the class i predicted by the model.

B. Ablation experiment

1) Different depth convolution modules

This study primarily utilizes a deep convolutional module as the overall framework, which derives five parallel channels from the deep convolutional module. The study investigates the impact of varying depths of convolutional layers on EEG emotion classification tasks. Specifically, experiments were conducted with three, four, five, and six layers of deep convolution.

Table 2 presents the average accuracy and average loss values under different deep convolutional modules. Figures 7 and 8 illustrate the training curves of accuracy and loss for each deep convolutional module. The dashed lines represent the training set curves, while the solid lines represent the validation set curves.

When the depth of convolutional layers is five (The lines marked with circular dots in Figures 7 and 8), the performance significantly outperforms other layer configurations. The average accuracy is 3.27 percentage points higher than that of three-layer convolution, and the average loss is reduced by 0.067 compared to four-layer and six-layer convolutional models. The effect of three-layer (The lines marked with downward triangle points in Figures 7 and 8) and four-layer convolution (The lines marked with star-shaped points in Figures 7 and 8) is relatively similar, possibly due to a reduced amount of extracted deep features caused by a smaller number of layers. When the number of convolutional layers reaches six (The lines marked with square points in Figures 7 and 8), the performance is much worse than that of the five-layer convolutional model, likely due to the excessive number of layers leading to learning redundancy and overfitting.

In summary, in the model proposed in this paper, after a series of experiments and comparative analysis, the model performs best when the number of deep convolutional layers is set to 5.

2) Different feature extraction modules

In the feature extraction module, the C-T module plays a major role. To determine the impact of the scale of the convolutional kernel and the inclusion of the Transformer on the classification task, this study conducted analysis and verification by modifying the C-T module. The experimental settings are as follows: without the inclusion of Transformer and with a 3x3 convolutional kernel, with the inclusion of Transformer and a 3x3 convolutional kernel, with the inclusion of Transformer and a 1x1 convolutional kernel, and with the inclusion of Transformer and a 5x5 convolutional kernel.

Table 3 presents the average accuracy and average loss values for different feature extraction modules. Figures 9 and 10 display the training accuracy and loss curves for different feature extraction modules. The dashed line represents the training set curve, while the solid line represents the validation set curve.

When Transformer is not used (The lines marked with downward triangle points in Figures 9 and 10), it is evident that the training accuracy and loss curves are significantly affected. The average accuracy is 14.08 percentage points lower than that of a 3x3 convolutional kernel, and the average loss value is higher by 0.127. However, the inclusion of Transformer leads to a noticeable improvement, possibly because it captures more features from the global context of the time slices. Despite the improvement achieved with the inclusion of Transformer, the results are poorest when using a 1x1 convolutional kernel (The lines marked with star-shaped points in Figures 9 and 10). This may be due to the small size of the convolutional kernel, which reduces the receptive field. A smaller receptive field requires a deeper network to compensate, but the convolutional network in the C-T module consists of only three layers, resulting in poor performance. The use of a 3x3 convolutional kernel (The lines marked with circular dots in Figures 9 and 10) outperforms a 5x5 kernel (The lines marked with square points in Figures 9 and 10) by 3.23 percentage points in average accuracy and a decrease of 0.121 in average loss value. A 3x3 convolutional kernel performs significantly better, and using a 5x5 kernel requires a larger computational cost, increasing the overall computation burden.

In summary, in this study's model, the inclusion of Transformer performs better than without it, and a 3x3 convolutional kernel performs better than a 1x1 or 5x5 kernel.

TABLE II

VALUES OF AVERAGE ACCURACY AND AVERAGE LOSS FUNCTION UNDER DIFFERENT DEPTH CONVOLUTIONAL MODULES

Depth umber of convolution layers	The deepest data size	Average Acc	Average Loss
3 floors	(32,10,32)	87.99%	0.408
4 floors	(32,10,16)	86.36%	0.588
5 floors	(32,5,16)	91.26%	0.332
6 floors	(32,5,8)	86.86%	0.592

TABLE Ⅲ
AVERAGE ACCURACY AND LOSS FUNCTION VALUES UNDER DIFFERENT FEATURE EXTRACTION MODULES

Convolution kernel size	Join Transformer	Average Acc	Average Loss
3×3	×	77.18%	0.459
1×1	√	52.56%	2.444
3×3	√	91.26%	0.332
5×5	√	88.03%	0.453

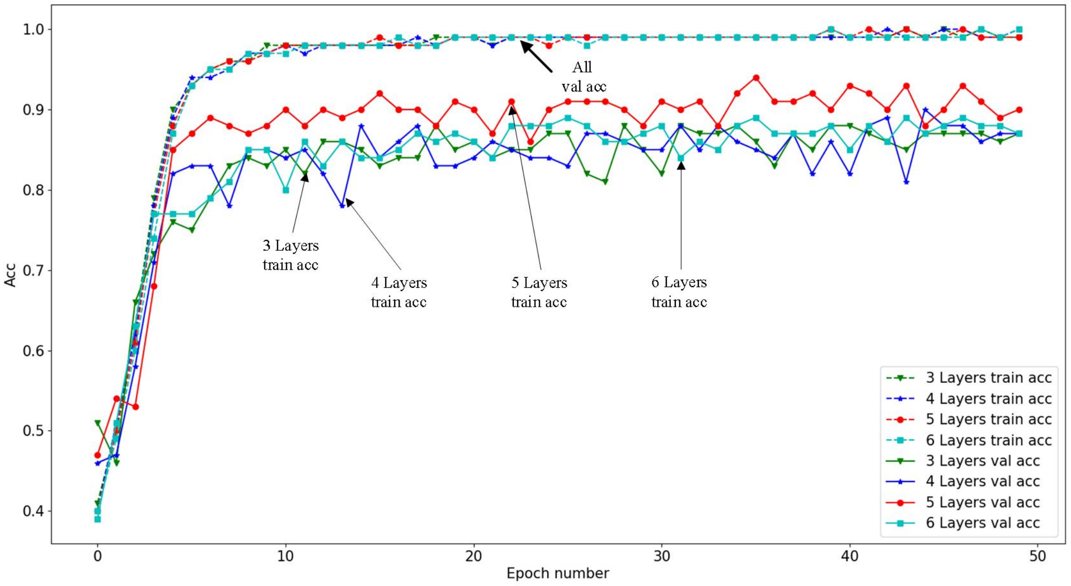

Fig. 7. Acc curves of convolution modules with different depths

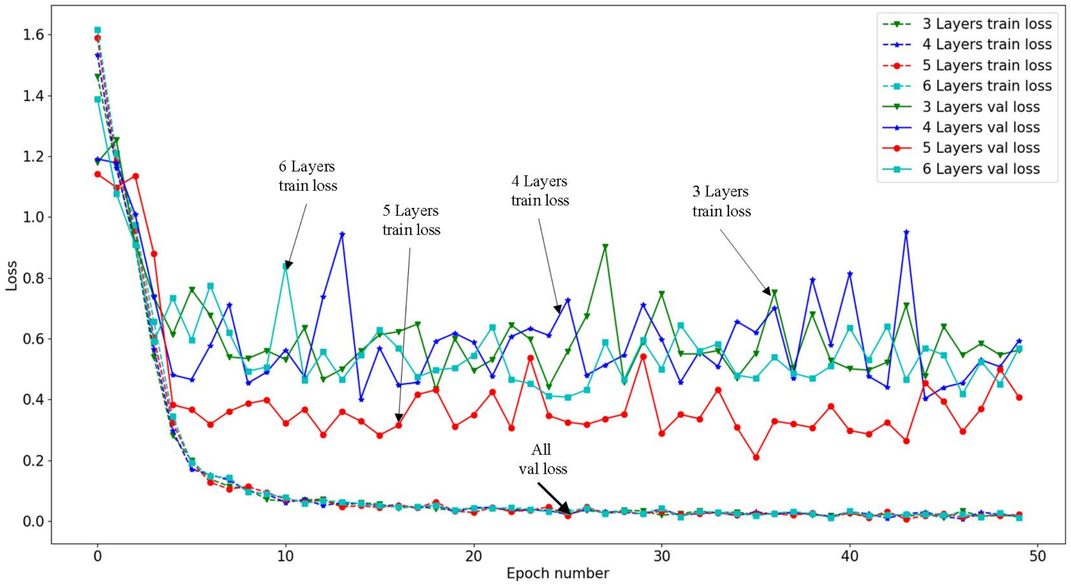

Fig. 8. Loss curves of convolutional modules with different depths

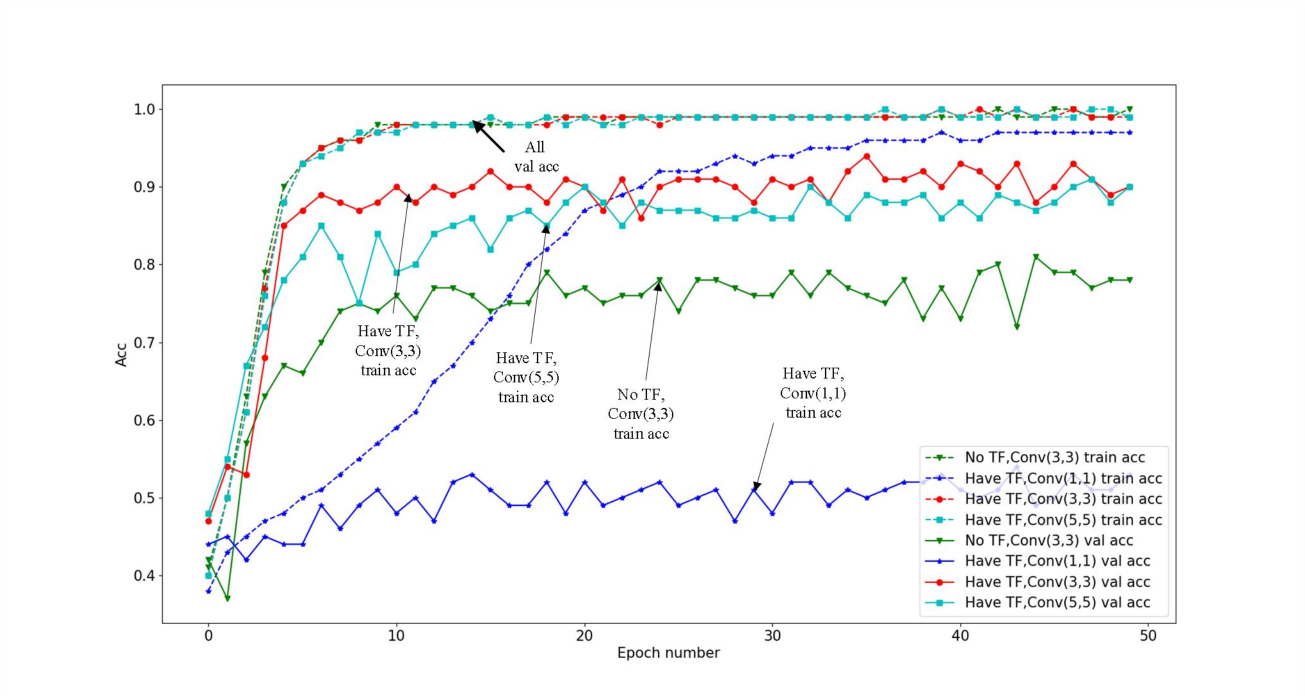

Fig. 9. Acc curves under different feature extraction modules

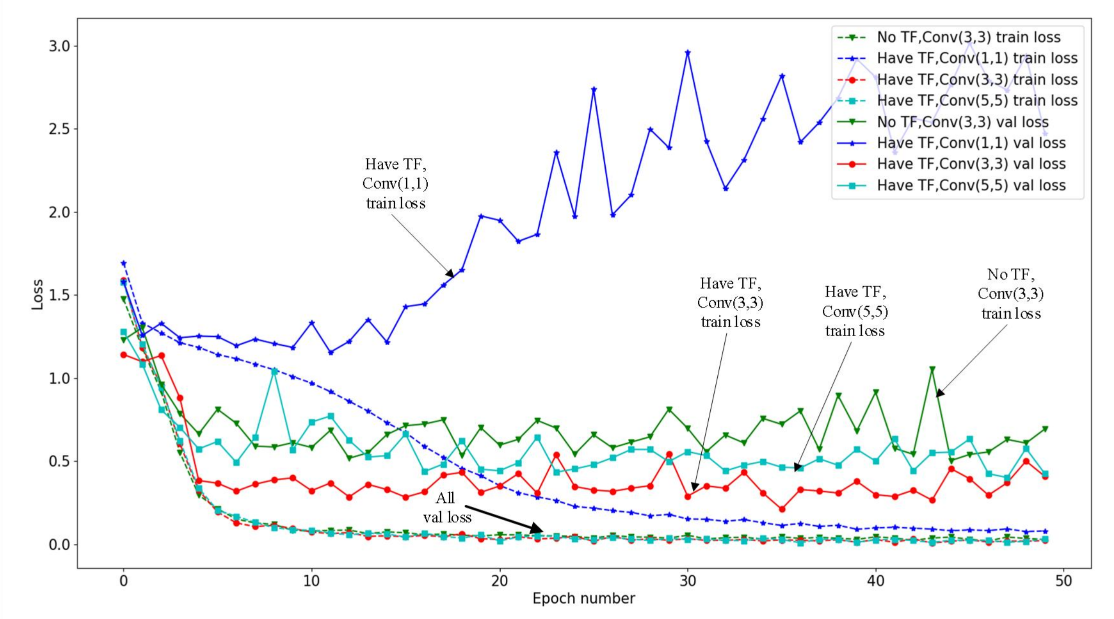

Fig. 10. Loss curve under different feature extraction modules

C. Experimental results and analysis

1) Experimental result

As shown in Figure 11 and Figure 12, are the accuracy and loss function values of the training set and validation set during the training of the four-class task. After 50 times of iterative training, the accuracy and loss function values of the final training and test tend to be stable. The average accuracy of this result on the DEAP dataset is 91.26%.

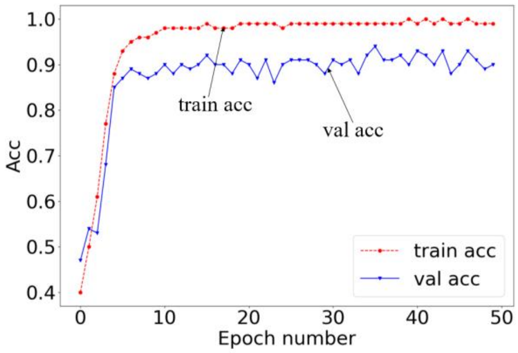

Fig. 11. Training accuracy curve of neural network

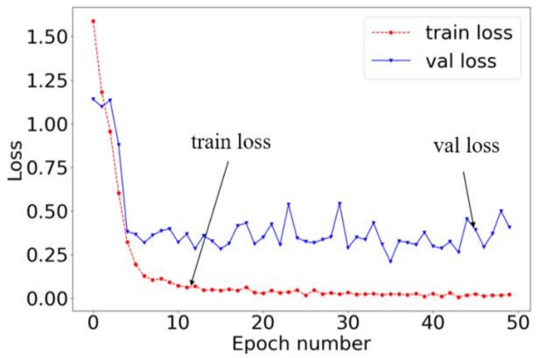

Fig. 12. Training loss function curve of neural network

2) *Experimental analysis*

To evaluate the effectiveness of the proposed method, a comparison was made with existing state-of-the-art methods in this study. The compared methods include Support Vector Machine (SVM) [5], a CNN-based approach that integrates multimodal data [26], a 3DCNN design using single-variable convolutional layers and multi-variable convolutional layers [27], an RNN model that incorporates Gated Recurrent Units (GRU) with skip connections [28], and a hybrid model that combines CNN and Bidirectional Long Short-Term Memory (Bi-LSTM) [29]. The aim was to demonstrate the differences among different methods within the same topic.

The proposed model was evaluated on the DEAP dataset in this study, providing objective evidence of its performance. Table 4 presents the four-class classification accuracies of the aforementioned models as well as the proposed model. By comparing these results, a clearer understanding of the performance differences among different methods can be obtained, thus validating the effectiveness of the proposed method.

TABLE IV
COMPARISON WITH OTHER METHODS

Ref.	Classifier	Average Acc
Zubair and Yoon[5]	SVM	49.70%
Kwon *et al.*[26]	CNN	73.43%
Chao and Dong[27]	3D CNN	76.77%
Asghar *et al.*[28]	GRU in RNN	80.10%
Singh *et al.*[29]	CNN and BI-LSTM	88.19%
Our Proposed	CNN and Transformer	91.26%

In Table 4, for the four-class emotion recognition task, the proposed method in this study achieved an average accuracy that is 3.07% higher than the second-ranked method in terms of classification accuracy, showing promising results. Due to the non-stationary nature of EEG signals and their strong background noise, Zubair *et al.* found that the performance of the traditional SVM method was slightly inferior compared to deep learning methods. Additionally, while Kwon and Chao improved upon traditional CNN, the singular feature extraction capability of CNN did not fully leverage its advantages. Singh *et al.*'s hybrid model of CNN and Bi-LSTM performed better than traditional CNN; however, Bi-LSTM processes information in a step-by-step iterative manner, which may lead to information decay or loss over multiple time steps and may not effectively capture long-range dependencies, unlike Transformers. Therefore, the proposed method in this study is considered superior to Singh *et al.*'s approach. The comparative results indicate that the proposed method in this study shows better performance in EEG-based emotion recognition tasks.

V. CONCLUSION

In this paper, a method for restructuring the emotional labels of the DEAP dataset is proposed, along with a C-T module for extracting deep features and a deep learning model for the four-classification of emotion recognition in EEG signals. Firstly, the preprocessed EEG signals are subjected to windowing, resulting in a fourfold increase in the training data. Subsequently, the EEG signals are transformed into time-frequency maps using short-time Fourier transform and input into the deep convolutional module. The deep convolutional module extracts feature information at different depths from the EEG signals, which is then input into the feature extraction module. In the feature extraction module, the use of the C-T module effectively enhances the capability to extract global and local information from the EEG signals. Finally, the data from the feature extraction module is integrated for the four-classification task.

Comparing with current state-of-the-art methods, the proposed method in this paper demonstrates a higher classification accuracy, providing full validation of the effectiveness of the algorithm. In future work, the plan is to apply this deep learning network model to more datasets to comprehensively evaluate its effectiveness and robustness, as well as to further explore its performance potential.

REFERENCES

[1] Kim, Min-Ki, et al. "A review on the computational methods for emotional state estimation from the human EEG." Computational and Mathematical Methods in Medicine 2013 (2013).

[2] Picard, Rosalind W. "Affective computing: challenges." International Journal of Human-Computer Studies 59.1-2 (2003): 55-64.

[3] Khosrowabadi, Reza, et al. "EEG-based emotion recognition using self-organizing map for boundary detection." 2010 20th International Conference on Pattern Recognition. IEEE, 2010.

[4] Kroupi, Eleni, Jean-Marc Vesin, and Touradj Ebrahimi. "Subject-independent odor pleasantness classification using brain and peripheral signals." IEEE Transactions on Affective Computing 7.4 (2015): 422-434.

[5] Zubair, Muhammad, and Changwoo Yoon. "EEG based classification of human emotions using discrete wavelet transform." IT Convergence and Security 2017: Volume 2. Springer Singapore, 2018.

[6] Xie Oiao, Zhen-Tao Liu, and Xue-Wen Ding. "Electroencephalogram emotion recognition based on a stacking classification model." 2018 37th Chinese Control Conference (CCC). IEEE, 2018.

[7] Zheng, Wei-Long, and Bao-Liang Lu. "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks." IEEE Transactions on Autonomous Mental Development 7.3 (2015): 162-175.

[8] Thammasan, Nattapong, Ken-ichi Fukui, and Masayuki Numao. "Application of deep belief networks in eeg-based dynamic music-emotion recognition." 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016.

[9] Li, Jinpeng, Zhaoxiang Zhang, and Huiguang He. "Hierarchical convolutional neural networks for EEG-based emotion recognition." Cognitive Computation 10 (2018): 368-380.

[10] Wen, Zhiyuan, Ruifeng Xu, and Jiachen Du. "A novel convolutional neural networks for emotion recognition based on EEG signal." 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). IEEE, 2017.

[11] Kwon, Yea-Hoon, Sae-Byuk Shin, and Shin-Dug Kim. "Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system." Sensors 18.5 (2018): 1383.

[12] Du, Xiaobing, et al. "An efficient LSTM network for emotion recognition from multichannel EEG signals." IEEE Transactions on Affective Computing 13.3 (2020): 1528-1540.

[13] An, Yi, Ning Xu, and Zhen Qu. "Leveraging spatial-temporal convolutional features for EEG-based emotion recognition." Biomedical Signal Processing and Control 69 (2021): 102743.

[14] Gao, Zhongke, et al. "Core-brain-network-based multilayer convolutional neural network for emotion recognition." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-9.

[15] Li, Yang, et al. "From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition." IEEE Transactions on Affective Computing 13.2 (2019): 568-578.

[16] Song, Yonghao, et al. "EEG conformer: Convolutional transformer for EEG decoding and visualization." IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2022): 710-719.

[17] Huo, Yonghua, et al. "Traffic anomaly detection method based on improved GRU and EFMS-Kmeans clustering." Computer Modeling in Engineering & Sciences 126.3 (2021): 1053-1091.

[18] Bai, Zhongli, et al. "Domain-adaptive emotion recognition based on horizontal vertical flow representation of EEG signals." IEEE Access (2023).

[19] Koelstra, Sander, et al. "Deap: A database for emotion analysis; using physiological signals." IEEE Transactions on Affective Computing 3.1 (2011): 18-31.

[20] Torres, Edgar P., et al. "EEG-based BCI emotion recognition: A survey." Sensors 20.18 (2020): 5083.

[21] Wang, Chenxi, et al. "Dynamic model-assisted transferable network for liquid rocket engine fault diagnosis using limited fault samples." Reliability Engineering & System Safety 243 (2024): 109837.

[22] Russell, James A. "A circumplex model of affect." Journal of Personality and Social Psychology 39.6 (1980): 1161.

[23] Xin, Ruihao, et al. "Multiview Feature Fusion Attention Convolutional Recurrent Neural Networks for EEG-Based Emotion Recognition." Journal of Sensors 2023 (2023).

[24] Singh, Khushboo, Mitul Kumar Ahirwal, and Manish Pandey. "Subject wise data augmentation based on balancing factor for quaternary emotion recognition through hybrid deep learning model." Biomedical Signal Processing and Control 86 (2023): 105075.

[25] Yin, Jin, et al. "A GAN guided parallel CNN and transformer network for EEG denoising." IEEE Journal of Biomedical and Health Informatics (2023).

[26] Kwon, Yea-Hoon, Sae-Byuk Shin, and Shin-Dug Kim. "Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system."Sensors 18.5 (2018): 1383.

[27] Chao, Hao, and Liang Dong. "Emotion recognition using three-dimensional feature and convolutional neural network from multichannel EEG signals." IEEE Sensors Journal 21.2 (2020): 2024-2034.

[28] Asghar, Muhammad Adeel, et al. "Semi-skipping layered gated unit and efficient network: hybrid deep feature selection method for edge computing in EEG-based emotion classification." IEEE Access 9 (2021): 13378-13389.

[29] Singh, Khushboo, Mitul Kumar Ahirwal, and Manish Pandey. "Subject wise data augmentation based on balancing factor for quaternary emotion recognition through hybrid deep learning model." Biomedical Signal Processing and Control 86 (2023): 105075.