Hindawi
Journal of Sensors
Volume 2023, Article ID 9281230, 15 pages
https://doi.org/10.1155/2023/9281230

Research Article

Multiview Feature Fusion Attention Convolutional Recurrent Neural Networks for EEG-Based Emotion Recognition

Ruihao Xin,[1,2] Fengbo Miao,[1] Ping Cong,[1] Fan Zhang,[1] Yongxian Xin,[3] and Xin Feng[4,5]

[1]*College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, Jilin, China 130000*
[2]*College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China 130012*
[3]*College of Business and Economics, Australian National University, ACT, Canberra, Australia 2601*
[4]*School of Science, Jilin Institute of Chemical Technology, Jilin, Jilin, China 130000*
[5]*State Key Laboratory of Inorganic Synthesis and Preparative Chemistry, College of Chemistry, Jilin University, Changchun, Jilin, China 130012*

Correspondence should be addressed to Xin Feng; fengxin@jlict.edu.cn

Received 4 August 2022; Revised 9 November 2022; Accepted 7 April 2023; Published 29 April 2023

Academic Editor: Mohit Mittal

Emotion recognition is essential for computers to understand human emotions. Traditional EEG emotion recognition methods have significant limitations. To improve the accuracy of EEG emotion recognition, we propose a multiview feature fusion attention convolutional recurrent neural network (multi-aCRNN) model. Multi-aCRNN combines CNN, GRU, and attention mechanisms to fuse features from multiple perspectives deeply. Specifically, multiscale CNN can unite elements in the frequency and spatial domains through the convolution of different scales. The role of the attention mechanism is to weigh the frequency domain and spatial domain information of different periods to find more valuable temporal perspectives. Finally, the implicit feature representation is learned from the time domain through the bidirectional GRU to achieve the profound fusion of features from multiple perspectives in the time domain, frequency domain, and spatial domain. At the same time, for the noise problem, we use label smoothing to reduce the influence of label noise to achieve a better emotion recognition classification effect. Finally, the model is validated on the EEG data of 32 subjects on a public dataset (DEAP) by fivefold cross-validation. Multi-aCRNN achieves an average classification accuracy of 96.43% and 96.30% in arousal and valence classification tasks, respectively. In conclusion, multi-aCRNN can better integrate EEG features from different angles and provide better classification results for emotion recognition.

1. Introduction

Emotion plays a significant role in our life, affecting human cognition and decision-making [1]. At the same time, it is also a relatively complex psychological state [2]. How to recognize emotions has become one of the issues in the industry [3]. At present, the mainstream methods of emotion recognition include two-dimensional valence and arousal coordinate system [4] and discrete assessment method [5]. In the two-dimensional valence and arousal coordinate system method, valence represents the positive or negative direction of the emotion, and arousal represents the intensity of emotion [6]. In discrete assessment, emotions are divided into multiple discrete categories. For example, Zheng and Lu classify emotions into positive, neutral, and negative categories [7], and Shanmugam and Padmanaban classify emotions into eight types: joy, trust, fear, surprise, sadness, disgust, anger, and expectation [8]. Emotion recognition is of great significance. It can help humans understand their own emotions, and it can also help computers better understand human emotions [9] so that computers can better serve humans.

With the development of computer science and information technology, human-machine interaction technology has attracted more attention [10]. As the cornerstone of human-machine interaction technology, emotion recognition has

inevitably attracted the attention of the academic community [11]. Generally speaking, emotion recognition methods can be divided into two categories: one is based on external signals of the human body [12], such as expression, posture, and voice; the other is based on the internal motions of the body [13], such as EEG, ECG, and EMG. Compared with external signals such as facial expression, posture, and voice, emotion recognition results based on internal cues such as EEG are more reliable because humans cannot control them intentionally [14].

In the traditional EEG emotion recognition method, first, screen out the hand-made features that are more relevant to the emotion recognition task [15], and then, input these emotional features into the machine learning model for classification. However, because deep learning does not require manual feature making and has better learning effect [16], researchers of EEG emotion recognition mostly use deep learning methods for research in recent years [17, 18]. Based on the characteristics of EEG signals, it can be extracted from the time domain, frequency domain, time-frequency domain, and nonlinear dynamical system [19]. Differential entropy (DE) is a representative nonlinear dynamic feature commonly used in EEG emotion recognition tasks [20]. The research of Garcia-Martinez et al. confirmed the effectiveness and robustness of the DE feature in EEG emotion recognition tasks [21]. Zhu and Zhong [22] classified DE features by using the 2DCNN-BiGRU network and achieved 87.89% and 88.69%, respectively, in the arousal and valence classification results of the DEAP dataset. However, a single convolution scale makes this method limited in spatial feature extraction, resulting in feature loss. In Yin et al.'s study [23], by using the ERDL model and combining the characteristics of the frequency domain and time domain, the classification accuracy on the DEAP dataset reached 90.45% and 90.6%, respectively. Shen et al. [24] proposed a four-dimensional convolutional recurrent neural network (4D-CRNN) to integrate the frequency domain, spatial domain, and time domain information of multichannel EEG signals to improve the accuracy of emotion recognition based on EEG. The accuracy of arousal and valence classification in the DEAP dataset reached 94.22% and 94.58%, respectively. However, these two methods ignore the differences between features and the impact of different features on classification results. In the process of feature fusion, it is easy to cause feature redundancy by not distinguishing different features. Cui et al. [25] proposed a DE-CNN-bi-LSTM network to remove DE features on different time slices in different frequency bands. After that, CNN and bi-LSTM were used to learn spatial and temporal information, so the classification accuracy on the DEAP dataset reached 94.86% and 94.02%, respectively. However, this method could not effectively deal with tag noise, which affected the classification results.

Aiming at the limitations of these methods, we propose a recursive network model based on multiview feature fusion. According to different spatial features extracted from convolutions of different scales, different spatial features of different periods are weighted and fused through the multihead attention mechanism to magnify the actual features and reduce the impact of invalid features, and label smoothing is used to reduce the impact of label noise. In conclusion, the main contributions of this study are as follows:

(1) To solve the noise in prediction labels of a single subject in EEG sentiment analysis, the method of label smoothing has dramatically reduced the influence of label noise on the classification accuracy of models to achieve a better effect of sentiment classification

(2) Multiscale convolution makes the extracted spatial features more comprehensive, and convolution is closely combined with bidirectional GRU (bi-GRU) so that the model can learn more comprehensive time-frequency features

(3) This paper proposes a multiview feature fusion attention convolutional recurrent neural network, which integrates the weight of frequency, space, and time-domain features. It effectively improves the classification accuracy of emotion recognition

In this paper, Section 1 is the introduction part, which introduces some basic concepts in the field of EEG emotion recognition, and briefly summarizes the previous research work. This paper made some breakthroughs based on prior studies. Section 2 is the method part, which mainly introduces key concepts in the data set and model. Section 3 is the experimental results and analysis, and the experimental process and results are introduced and analyzed in detail. Section 4 discusses, compares, and analyzes the existing paper results, reflecting the research significance and value of this paper. Section 5 is the conclusion, reviewing and summarizing this paper.

Table 1: DEAP dataset.

Array name	Array shape	Array contents
Data	40*40*8064(63*128)	Video/trial*channel*data
Labels	40*4	Video/trial*label

2. Methods

2.1. Dataset and Preprocessing. DEAP dataset [26] is a multichannel dataset collected by Koelstra et al., as shown in Table 1, who invited 32 subjects (including 16 males and 16 females) to watch 40 music videos used to study human emotion. The subjects invited for the experiment are in good physical condition and mental health and can generally respond to the stimulation of the video material. Each music video lasts for 1 minute, regarded as an experiment. For each experiment, the first 3 seconds (3 s) is the video conversion time, and the last 60 s is the music video play time, so the duration of each sample is 63 s, and the video conversion time of the first 3 s is the baseline time of the experiment. After playing the video, each subject had to score the video in valence, arousal, and other dimensions, ranging from 1 to 9. We selected five as the threshold and regarded the

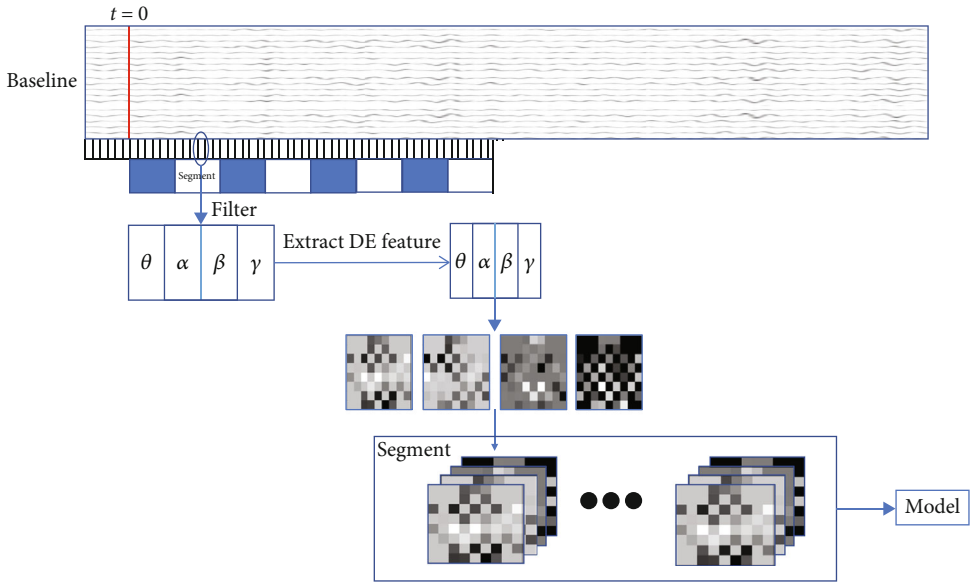


Figure 1: Data preprocessing flow chart.

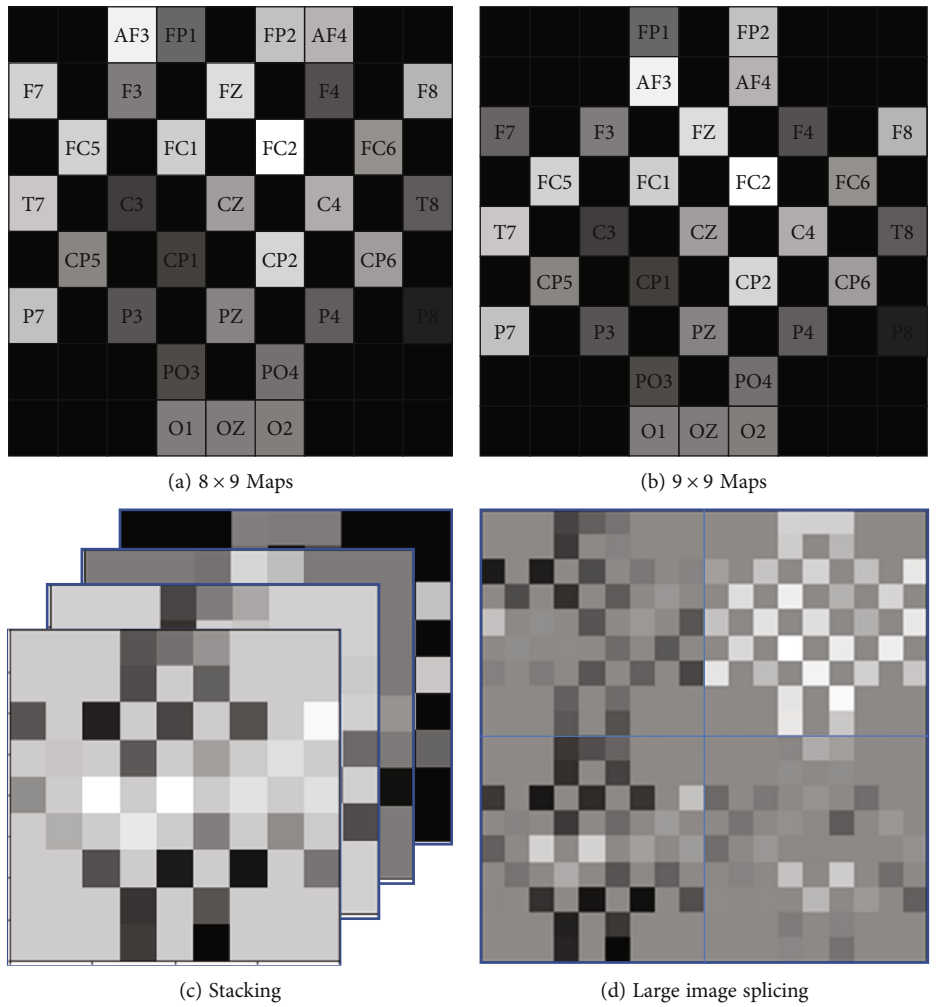
emotion recognition task of the DEAP dataset as two binary classification problems. EEG signals were sampled using a 32-channel electrode. The sampling frequency is 512 Hz, and electrodes were placed by the international standard lead 10-20. In addition, the experiment collected not only 32-channel EEG signals but also 8-channel ECG and EOG, a total of 40 physiological signal channels. Only the first 32 EEG signals were used in this paper. The website provides the preprocessed data by downsampling the frequency from 512 Hz to 128 Hz and removing noise such as ocular artifacts.

The 63 s EEG signal was collected for each trial. First, we cut the 63 s signal to 0.5 s. Then, each segmented EEG signal block was filtered to obtain four frequency bands of θ, α, β, and γ, and then, DE features were extracted from each frequency band. Since Yang et al. [27, 28] have proved that considering baseline signals can improve the classification effect of the model, we will carry out baseline correction according to the method in the paper. The DE features of each frequency band of the baseline signal were averaged. Then, the DE features of the stimulus signal were differentiated from the average of the corresponding frequency band to obtain the baseline-corrected DE features. The processed DE features were mapped into a two-dimensional map for each frequency band according to the electrode distribution. For each block, the 2D maps of the four frequency bands are concatenated to form a new feature matrix. Finally, the feature matrix of the block is combined according to the segment (1segment = 6 blocks) and sent into the model as a sample. The data preprocessing process is shown in Figure 1.

Electrodes can be converted into two kinds of 2D maps: one type is 8×9, as shown in Figure 2(a), and the other is a 9×9, as shown in Figure 2(b). Likewise, 2D maps of the four frequency bands can be jointed in two ways. One is stacked splicing, that is, to form a three-dimensional matrix. In this paper, 8×9 maps are spliced stacked, as shown in Figure 2(c). Or it can be assembled into a large picture. The picture jointed in this way is still a two-dimensional matrix. In this paper, 9×9 graphs are jointed in a large picture, as shown in Figure 2(d).

For a single subject, 60 seconds of stimulus signal data collected from 40 music videos watched is 40*60*2, divided into 800 samples for 4800 stimulus signals; each sample contains 6 time period information, and each period information contains 4 frequency bands. Taking the superposition result as an example, the characteristic of each frequency band is an 8*9 mapping matrix. There are a total of 32 subjects, and in the arousal classification, there are a total of 25600 samples, including 10860 low arousal samples and 14740 high arousal samples. Each sample contains 6 time period information. Each period information is a 4*8*9 feature matrix. In the valence classification, there are a total of 25600 samples, including 11,40 low-valence samples and 14160 high-valence samples, and each sample contains 6 time period information. Each period information is a 4*8*9 feature matrix. Because subjects respond differently to emotional stimuli, the number of positive and negative samples for (low/high) arousal and (low/high) valence will not be the same.

2.2. Spatial Feature Extraction Based on 2D-CNN. Convolutional neural networks are often used to process 2D data and usually consist of three parts: convolutional layers, pooling layers, and activation function layers [29]. The convolution layer performs the inner product operation on the input data through the convolution kernel. By setting the size and number of the convolution kernel, the model can extract different types of data features. At the same time, in the convolutional layer, the number of parameters that the model needs to be trained is reduced through "sparse connection" [30] and "weight sharing," thereby reducing the difficulty of training.



(a) 8 × 9 Maps

(b) 9 × 9 Maps

(c) Stacking

(d) Large image splicing

Figure 2: Electrodes conversion.

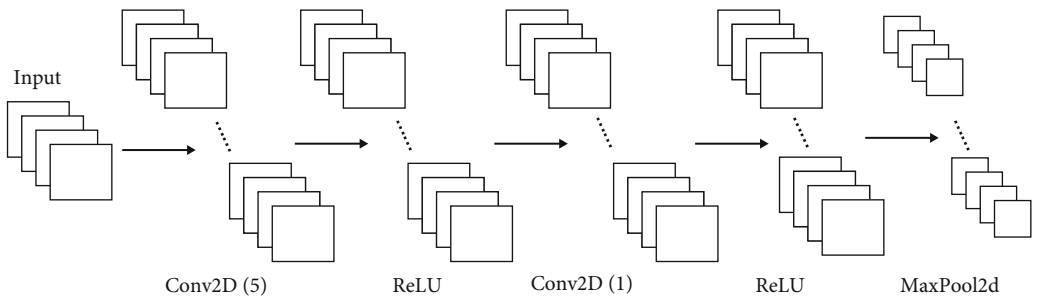


Figure 3: Part of the convolutional neural network model.

In addition, the pooling layer can further reduce the data provided by the model to the next layer of the network, reducing the difficulty of model training [31]. The activation function layer transforms the data to reduce the training difficulty and enhance the correlation between data. Part of the convolutional neural network model used in this paper is shown in Figure 3.

2.3. Time Series Feature Extraction Based on GRU. RNN (recurrent neural network) has certain advantages when dealing with time series data. When RNN processes the information at each moment, it can effectively preserve the original timing of the data, and the training parameters will not increase due to the increase in the sequence length. This paper uses an improved cyclic structure GRU (gated recurrent unit) model, as shown in Figure 4.

The GRU has a reset gate and an update gate. The reset gate determines the degree to which the input information will be combined with the previously memorized information. The update gate determines how much of

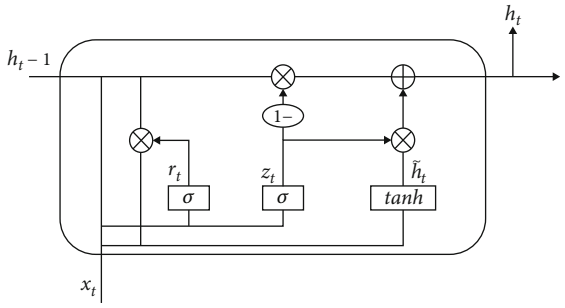


Figure 4: GRU (gated recurrent unit) model.

the previously memorized information can be retained to the current time step. The specific formula is as follows:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_t - 1, x_t]), \\ r_t &= \sigma(W_r \cdot [h_t - 1, x_t]), \\ \tilde{h}_t &= \tan h(W \cdot [r_t * h_t - 1, x_t]), \\ h_t &= (1 - z_t) * h_t - 1 + z_t * \tilde{h}_t, \end{aligned} \tag{1}$$

where h_t is the hidden state at time t, x_t is the input at time t, r_t and z_t are the reset gate and update gate, respectively, $\tilde{h}_t$ is the candidate hidden state, σ is the sigmoid function, and $*$ is the Hadamard product.

In this paper, GRU is used to obtain the time series characteristics of data, and a comparative test is carried out for GRU and bi-GRU in Section 3.4.

2.4. Feature Fusion Based on Multihead Attention. Usually, scaled dot-product attention consists of three parts: Q (query), K (key), and V (value). The structure is shown in Figure 5(a). Assume that the dimensions of the input Q and K are d_k, and the dimension of V is d_v. Then, calculate the transposed multiplication of Q and K, divide by $\sqrt{d_k}$, pass the result through the Softmax function to get the weight, and multiply the weight by V to get the output matrix. The specific formula is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)V. \tag{2}$$

This paper uses multihead attention [32], and its structure is shown in Figure 5(b). Multihead attention can combine the information learned by different heads, which can be regarded as parallel processing of multiple scaled dot-product attention. Q, K, and V are first subjected to a linear transformation and then input to the scaled dot-product attention. Here is the scaled dot-product attention for head times and stacking the obtained results. The spliced result is then subjected to a linear transformation to obtain the value as the output of multihead attention.

2.5. AdamW and Label Smoothing

2.5.1. AdamW Optimization Algorithm. Adam optimization algorithm has been widely used in various deep learning models since its appearance, but experiments found that Adam has specific problems. Such as slow model convergence, nonconvergence, and other problems, various improved versions of Adam appeared. Different parameters in the Adam optimization algorithm adaptively learn at different learning rates. The formula is as follows:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) {g_t}^2, \end{aligned} \tag{3}$$

where g_t represents the gradient, the subscript t represents time, m_t is the first-order moment variable of the gradient, v_t is the second-order moment variable of the gradient, and β_1 and β_2 are the exponential decay rates (decay factors) of the moment estimation. When the values of m_t and v_t approach the 0 vectors, the result will be biased. This problem is solved by performing bias correction on m_t and v_t. The formulas for the bias correction value m_t and v_t of m_t' and v_t' are as follows:

$$\begin{aligned} m_t' &= \frac{m_t}{1 - \beta_1^t}, \\ v_t' &= \frac{v_t}{1 - \beta_2^t}. \end{aligned} \tag{4}$$

AdamW [33] adds a regular term to Adam's loss function and adds the result of the gradient of the regular term when calculating the gradient so that the gradient of the overall loss function is calculated when updating the model parameters, thereby updating the parameters. AdamW's loss function is

$$L = \text{loss} + \frac{1}{2}\|\theta\|^2. \tag{5}$$

Then, the formula for AdamW parameter update is

$$\theta_t = \theta_{t-1} - \eta\left(\frac{\alpha m_t'}{\sqrt{v_t'} + \xi} + \omega\theta_{t-1}\right), \tag{6}$$

where θ is a parameter in the model, η is the learning rate, α is 0.001, ξ is 10^{-8}, and ω is an actual number.

2.5.2. Label Smoothing. There are usually some noisy labels in machine learning samples, and these labels will have a certain impact on the prediction results. Label smoothing prevents the model from believing too much in the labels of the training samples by assuming that the labels may be wrong during training [34]. The formula looks like this:

$$y_i = \begin{cases} 1 - \varepsilon, & \text{if } i = \text{true}, \\ \dfrac{\varepsilon}{K - 1}, & \text{otherwise}. \end{cases} \tag{7}$$

Among them, ε is a defined hyperparameter, which generally takes a value of 0.1, K is the number of categories of

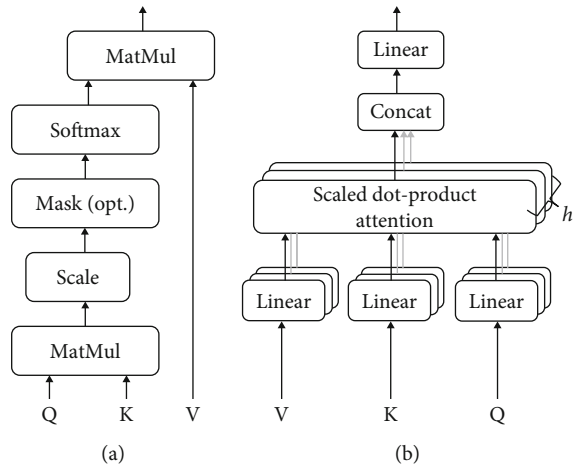


Figure 5: (a) Scaled dot-product attention. (b) Multihead attention.

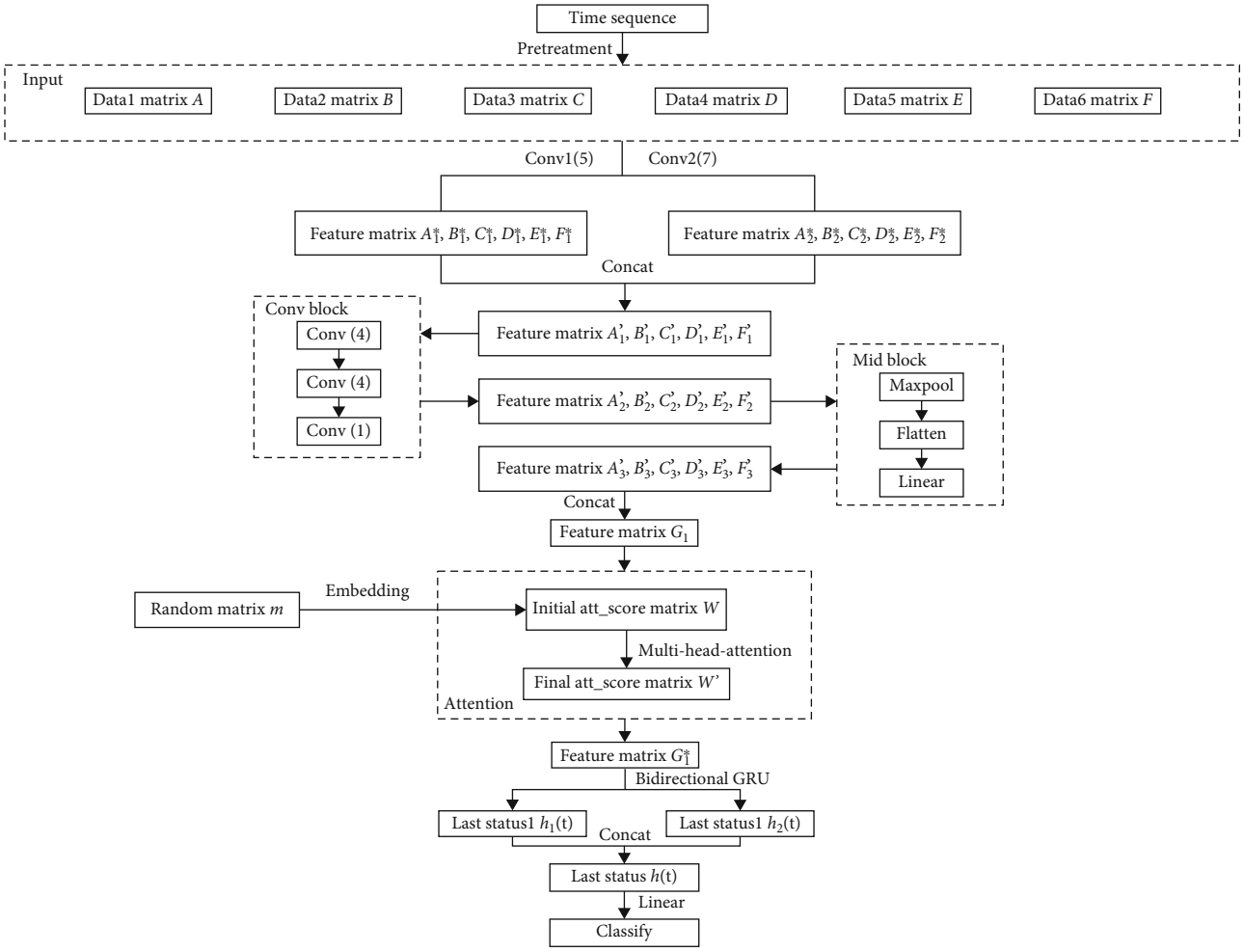


Figure 6: Model flow chart.

the classification problem, and y_i is the sample label. When training samples, it is usually not guaranteed that all sample labels do not contain noise interference. For example, if the label of a sample is wrong, then the sample may harm the training results during training. By letting the model know that the label of the sample is not necessarily correct, the trained model can better identify a small number of wrong samples.

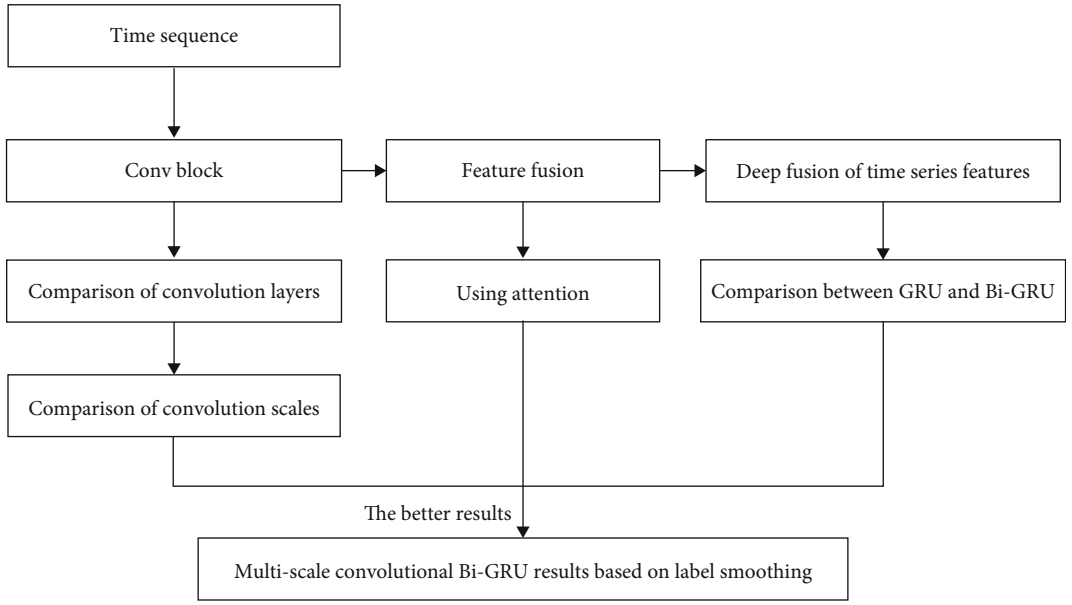


Figure 7: Overall experimental process.

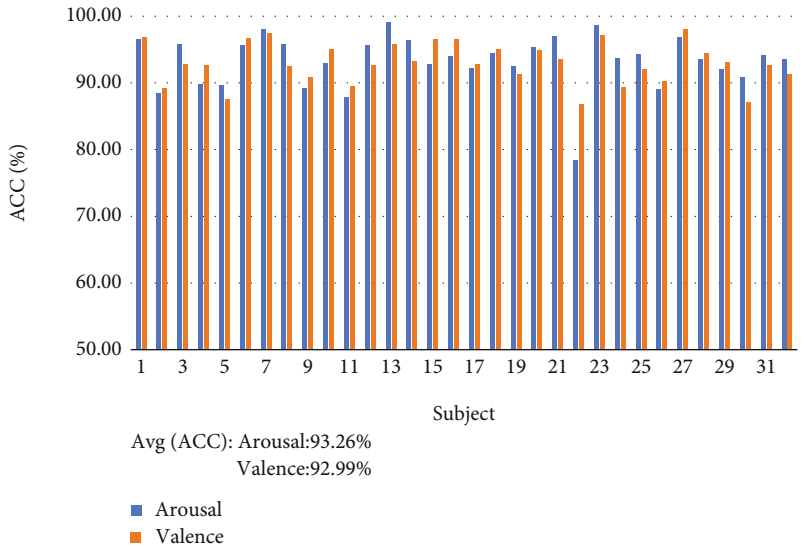


Figure 8: Two-layer convolution model results (the model is 5+1, there are two layers of the convolutional network, the size of one convolution kernel is 5, and the size of one convolution kernel is 1).

2.6. Multiview Feature Fusion Attention Convolutional Recurrent Neural Network Model. In this paper, after preprocessing the EEG time series data, the original sequence is divided into six data segments according to the period, thus effectively preserving the time series of the data. At the same time, to make the extracted features more comprehensive, multiscale convolution is used to extract spatial-domain features, and the extracted features are highly abstracted through convolution blocks. For the abstracted data, the extracted frequency-domain and spatial-domain features are weighted from the time series perspective through the attention mechanism, and the weighted data is classified through the bidirectional GRU model. The specific process is shown in Figure 6.

Perform spatial feature extraction to obtain feature matrix $A_1^*, B_1^*, C_1^*, D_1^*, E_1^*, F_1^*$ and feature matrix $A_2^*, B_2^*, C_2^*, D_2^*, E_2^*, F_2^*$. Concat the two spatial features to obtain matrix $A_1', B_1', C_1', D_1', E_1', F_1'$, and transfer the matrix to the Conv block. Abstract the spatial features through three-layer convolution in the Conv block to obtain feature matrix $A_2', B_2', C_2', D_2', E_2', F_2'$. The abstract matrix is subjected to maximum pooling, flatten and linear network layers to obtain feature matrix $A_3', B_3', C_3', D_3', E_3', F_3'$, and six matrices are Concat to obtain matrix G_1. At the same time, the random initialization matrix m is used as the initial weight matrix W of attention after passing through the embedding layer. After G_1 is input, the final weight matrix W' is obtained through multihead attention

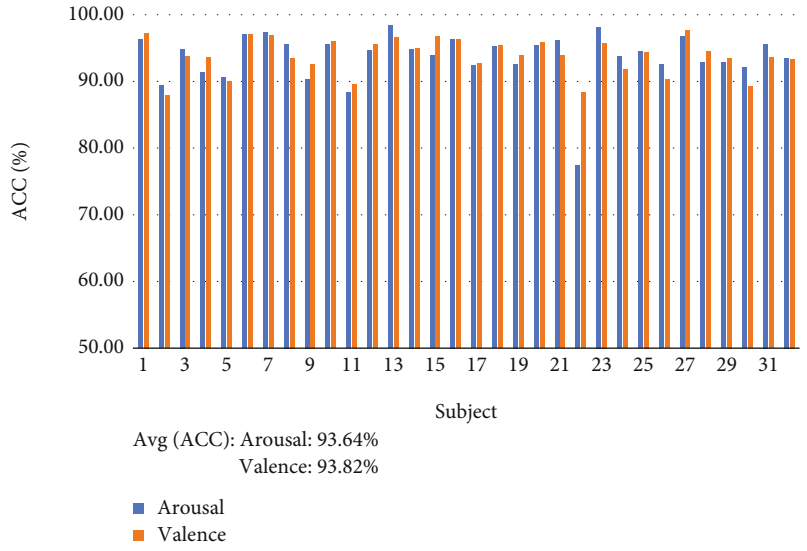


Figure 9: Deep convolution model results (the model is 5+4+4+1, and there are four layers of the convolutional network).

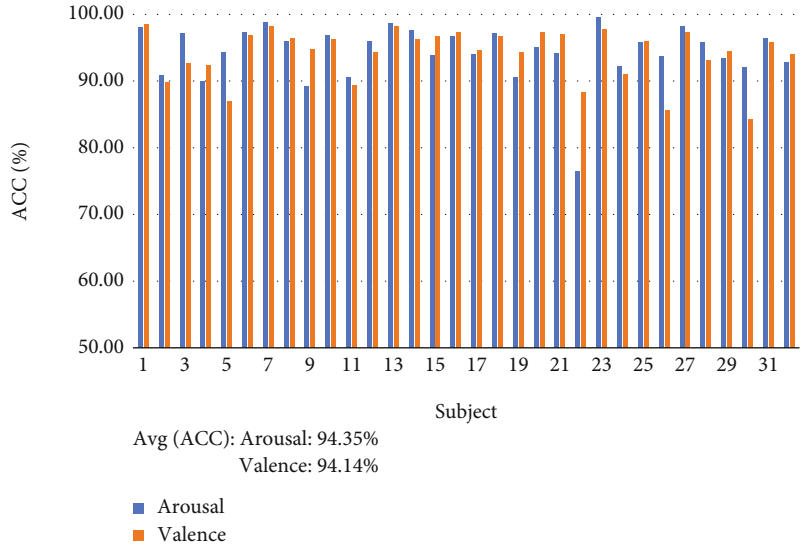


Figure 10: Results of deep convolution classification based on attention (the model is 5+4+4+1+Att, and attention is used in the model).

training, and the feature matrix G_1^* is obtained after weighting. The weighted matrix is passed into the bidirectional GRU model to extract the time series features of the data. Concat the states $h_1(t)$ and $h_2(t)$ at the last moment of forward and reverse to obtain the final output state matrix $h(t)$, and pass $h(t)$ through the linear network layer to achieve classification.

3. Experiment and Analysis

The batch size for training multi-aCRNN is 128, the dropout is 0.5, the maximum number of epochs is 500, the learning rate is $5 * 10^{-5}$, and the number of heads is 8. PyTorch implements the model, NVIDIA-SMI 460.67, CUDA Version: 11.2, python version 3.7.0, PyTorch version 1.11.0.

Table 2: Comparison of model classification results before and after adding attention.

Model	ACC (a)	Num (a)	ACC (v)	Num (v)
5+4+4+1	93.64%	3	93.82%	4
5+4+4+1+Att	94.35%	2	94.14%	6

Five cross-validations were used for each experiment in this paper. They were performed for the average classification accuracy (ACC) and the number of subjects (Num) with average classification accuracy below 90% for arousal (a) and valence (v) analysis. The overall experimental process is shown in Figure 7. The idea is to compare the number and scale of convolution layers, the use of attention, the

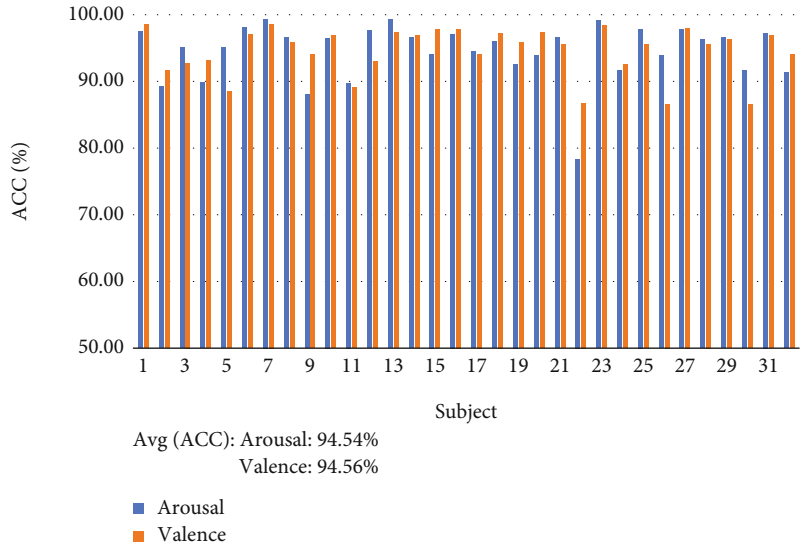


Figure 11: Attention-based deep convolution (the model is 7+4+4+1+Att).

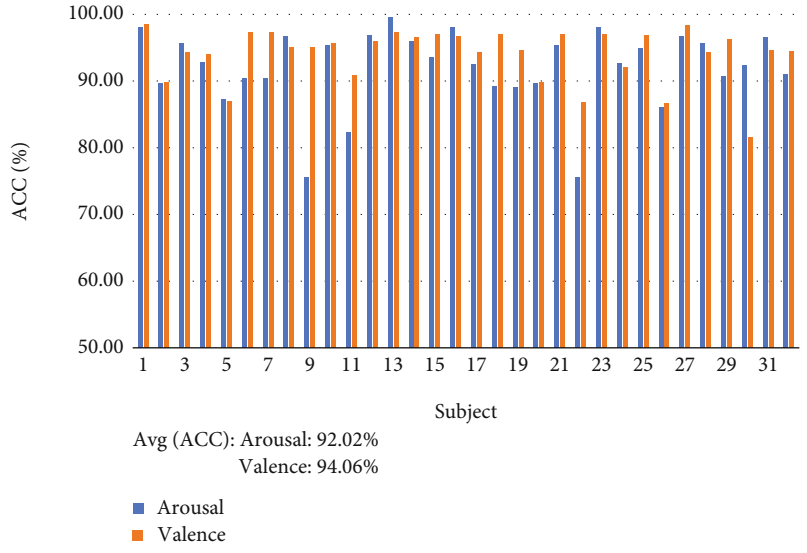


Figure 12: Attention-based deep convolution (the model is 7+5+5+1+Att).

comparison between GRU and bi-GRU, and the use of label smoothing.

3.1. Comparison of the Number of Convolutional Layers. Usually, in a convolutional neural network, the number of convolutional layers determines the degree of abstraction of features. Thereby, a more accurate prediction result can be obtained. Therefore, this section compares the two-layer and four-layer convolution results. The results are shown in Figures 8 and 9:

It can be seen from Figures 8 and 9 that the deep convolution is better than the two-layer convolution to a certain extent for the experimental results, especially in the valence result—the average accuracy of using deep convolution increases by 0.83%. At the same time, judging from the classification results of a single subject, when using two-layer convolution for classification, the arousal classification results of 7 subjects were lower than 90%. The valence classification results of 6 subjects were lower than 90%. In contrast, in the four-layer convolution, there are three subjects whose arousal classification results are lower than 90% and

Table 3: Comparison of classification results of different scale convolution kernel models.

Model	ACC (a)	Num (a)	ACC (v)	Num (v)
5+4+4+1+Att	94.35%	2	94.14%	6
7+4+4+1+Att	94.54%	5	94.56%	5
7+5+5+1+Att	92.02%	9	94.06%	6

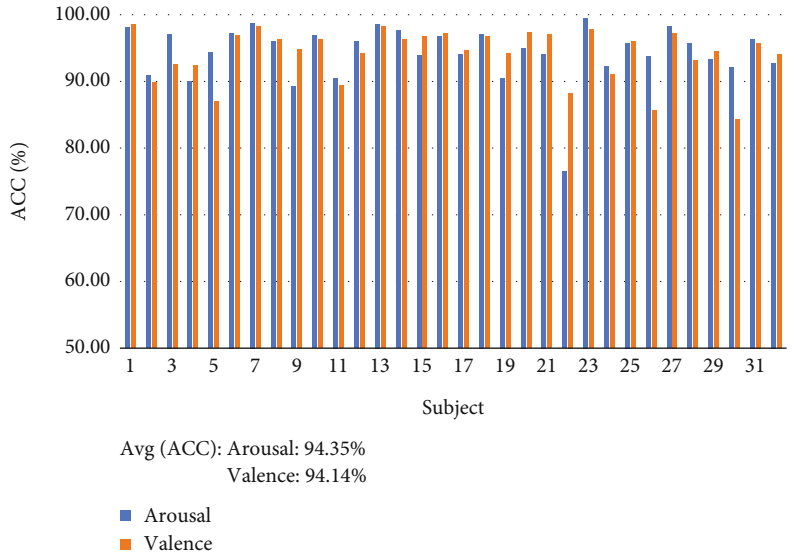


Figure 13: GRU model results (the model is 7+4+4+1+Att+GRU).

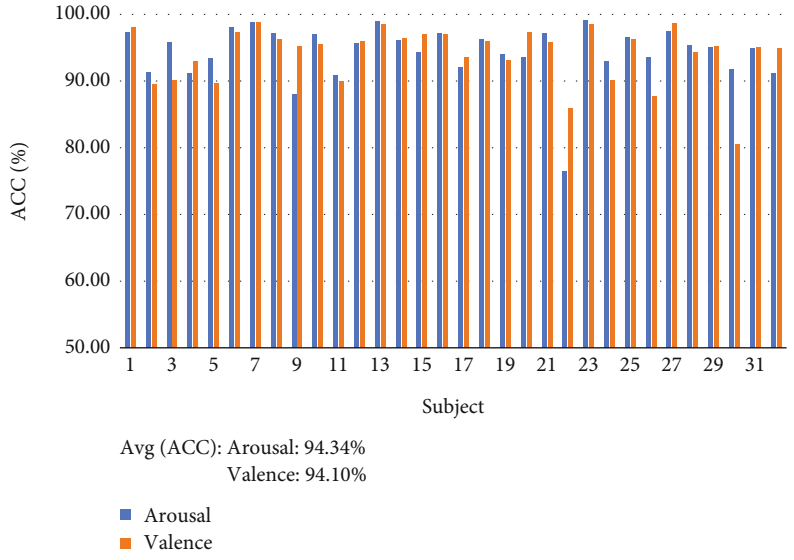


Figure 14: Bi-GRU model results (the model is 7+4+4+1+Att+bi-GRU).

four subjects whose valence classification results are lower than 90%, greatly reducing the degree of classification error.

3.2. Attention-Based Deep Convolution Classification Results. For the features extracted by the convolution layer for different periods, the traditional Concat method cannot sufficiently distinguish the effectiveness of these features. Aiming at this problem, we use the attention mechanism for feature fusion so that the frequency-domain and spatial-domain features of different periods can be distinguished by increasing the weight to achieve a better classification prediction effect. The results are shown in Figure 10.

Comparing Figures 10 and 9, after adding attention, the overall classification accuracy has been improved to a certain extent. It can be seen from Table 2 that the average accuracy of arousal classification increased by 0.71%, and the average accuracy of valence classification increased by 0.32%. For a single subject, using weighted feature fusion resulted in greater progress in the more difficult second subject to classify. At the same time, only two subjects had an accuracy rate below 90% for the arousal classification accuracy rate. Still, six subjects had an accuracy rate below 90% in the valence classification results. Since the weighted fusion is

Table 4: Comparison of classification results between GRU and bi-GRU models.

Model	ACC (a)	Num (a)	ACC (v)	Num (v)
GRU	94.35%	2	94.14%	6
Bi-GRU	94.34%	2	94.10%	5

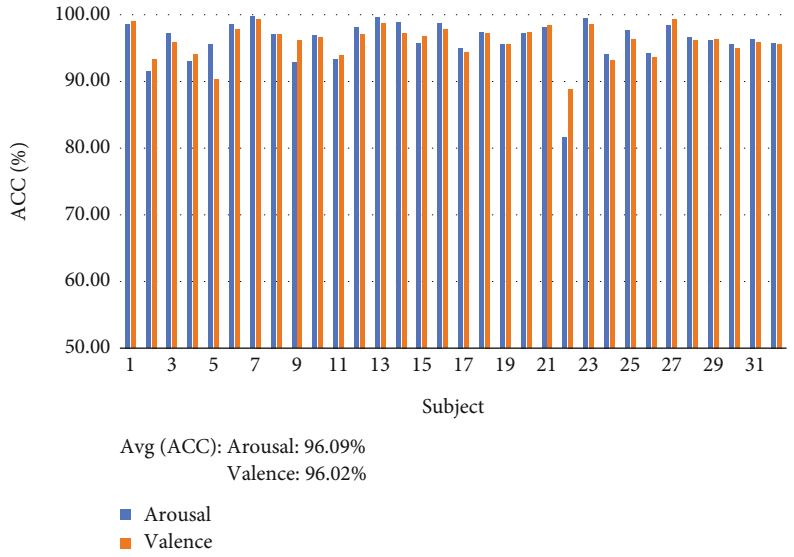


Figure 15: Single-scale convolutional bi-GRU results based on label smoothing (the model is 7+4+4+1+bi-GRU+lab).

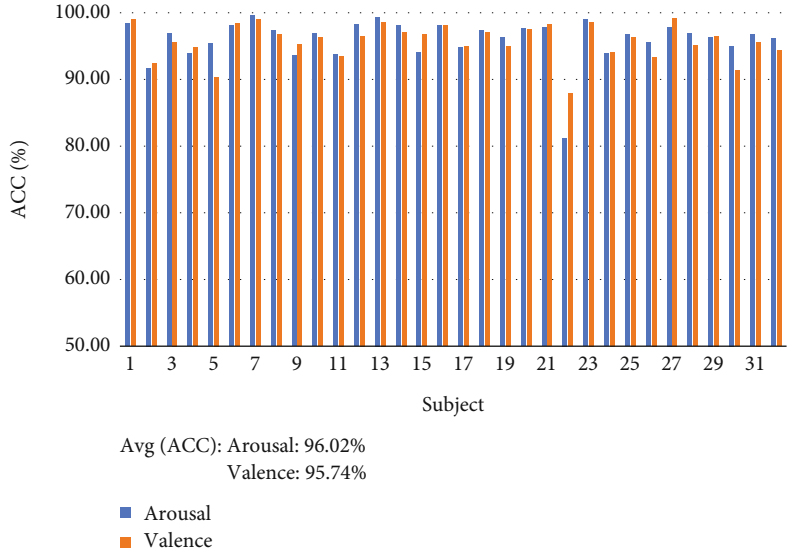


Figure 16: Multiscale convolutional GRU results based on label smoothing (the model is 5&7+4+4+1+GRU+lab).

carried out on the results of the convolution of six time periods, it has much to do with the convolution process. It is guessed that the locality of the initial convolution causes the attention of the model to be limited during learning, so the overall accuracy of valence classification has improved, and some single-subject results have declined.

3.3. A Comparative Study of Convolution at Different Scales. Since the convolutional layer is limited by the size of the convolution kernel when performing feature extraction, the size of the convolution kernel of the model is experimentally explored in this section. It can be seen from Figures 11 and 12 that different convolution kernel sizes have a significant impact on classification accuracy and classification stability. It can be seen from Table 3 that the first layer of convolution is the feature extraction of the original data, so by comparing the convolution kernels with the convolution kernel size of 5 and 7, it can be seen that the convolution kernel of 7 scales is used in the classification results better than five-scale convolution, and the accuracy of valence classification is improved by 0.42%. This is because large-scale convolution has a more extensive perception range when extracting features from the original data, which can significantly reduce the limitations of spatial feature extraction. However, this is only limited to the extraction of initial features. The effect of using a larger-scale convolution in the middle layer convolution is significantly reduced. This is because the middle layer convolution is a reabstraction of features. A larger-scale

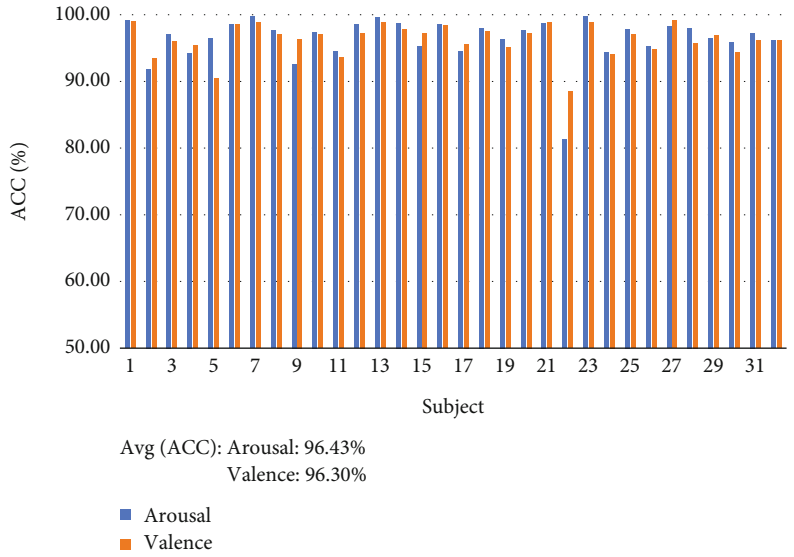


Figure 17: Multiscale convolutional bi-GRU results based on label smoothing (the model is 5&7+4+4+1+bi-GRU+lab).

Table 5: Comparison of multiscale and GRU and bi-GRU experimental results based on label smoothing.

Model	ACC (a)	Num (a)	ACC (v)	Num (v)
7+4+4+1+bi-GRU+lab	96.09%	1	96.02%	1
5&7+4+4+1+GRU+lab	96.02%	1	95.74%	1
5&7+4+4+1+bi-GRU+lab	96.43%	1	96.30%	1

convolution will affect the initial feature information and cover the part of the effective information in the initial features, resulting in a decrease in the classification results.

3.4. GRU and Bi-GRU Comparative Experiment. For time series, the time series features are extracted through a recurrent neural network to optimize the model. In this section, a comparative experiment is carried out on the GRU and bi-GRU networks, and the experimental results are shown in Figures 13 and 14. It can be seen from Table 4 that the classification accuracy results of GRU and bi-GRU are almost the same, and the classification results for a single subject are also relatively close. For this phenomenon, the use of GRU and bi-GRU will be further explored in Section 3.5.

3.5. Multiscale Fusion Model Based on Label Smoothing. In Section 3.4, GRU and bi-GRU are explored, but the experimental results are relatively close and cannot clearly show the pros and cons of the model. At the same time, we explore convolutional networks of different scales in Section 3.3 and find that larger initial convolution kernels are more effective for extracting spatial features but whether retaining the features of smaller-scale convolutions at the same time will promote emotion recognition to a certain extent. The multiscale fusion model based on label smoothing will be explored in this section. The results are shown in Figures 15–17 and Table 5.

After label smoothing, the model's accuracy has been significantly improved, and the results for a single subject are also more stable. In the single-scale Bi-GRU fusion model (7+4+4+1+Bi-GRU+lab), the accuracy of arousal and valence reached 96.09% and 96.02%, respectively, and there is only one subject with a classification accuracy below 90%. Since some subjects have certain errors in the experiment, it is difficult to improve the classification accuracy of some subjects. These noises will also affect the classification results of other subjects and even lead to the model's accuracy. The training results are getting worse and worse. It can be seen that AdamW and label smoothing can significantly promote the fitting of the model and the calibration of the network, which can dramatically reduce the impact of label noise.

Table 6: Comparison of experimental results between splicing and stacking preprocessing methods.

Methods	ACC (a)	Num (a)	ACC (v)	Num (v)
Splicing	94.96%	2	94.60%	4
Stacking	96.43%	1	96.30%	1

It can be seen from Table 5 that the training results of bi-GRU are significantly better than GRU. In the multiscale network, the results of bi-GRU (5&7+4+4+1+bi-GRU+lab) reach 96.43% and 96.30%, respectively, compared with 96.02% and 95.74% of the GRU network (5&7+4+4+1 +GRU+lab), the results of using bi-GRU are improved by 0.42% and 0.56%, respectively, and the overall optimization was achieved. It shows that in the EEG time series data,

Table 7: Comparison of results with other papers.

Nos.	Model	Information	ACC (a)	ACC (v)
1	2DCNN-BiGRU (2021) [22]	Spatial+temporal	87.89%	88.69%
2	ERDL (2020) [23]	Frequency+temporal	90.45%	90.6%
3	4D-CRNN (2020) [24]	Frequency+spatial+temporal	94.22%	94.58%
4	DE-CNN-BiLSTM (2022) [25]	Frequency+spatial+temporal	94.86%	94.02%
5	Multi-aCRNN (ours)	Frequency+spatial+temporal	96.43%	96.30%

the reverse time series information also has a certain effect, promoting the overall experimental results. At the same time, by comparing the experimental results of single-scale and multiscale, it can be seen that the results of using 5-scale and 7-scale convolution kernels at the same time to extract features from the original data are significantly better than the results of using 7-scale convolution kernels alone. Indicating the fusion of different scale features is more helpful for the model to learn more comprehensive and effective information to achieve better classification results.

3.6. Contrast of Splicing and Stacking Preprocessing. This section experimentally explores two different preprocessing methods, splicing and stacking. It can be seen from Table 6 that the stacking preprocessing method improves the arousal and valence classification results by 1.47% and 1.7%, respectively, compared with the splicing preprocessing method. And, for a single subject, the stacking results are more stable. At the same time, the splicing method has two subjects with an accuracy of less than 90% in arousal classification and four subjects with an accuracy of less than 90% in valence classification. So in this experiment, the stacking preprocessing method is used for the experiment.

4. Discussion

As shown in Table 7, when the model learns spatiotemporal information based on extracting frequency information, it tends to get a better experimental result. However, when the model does not distinguish the obtained information and trains all the information, it will affect the training effect of the model. We perform weight training on the extracted features through the attention mechanism, amplify effective information, reduce invalid information, and use label smoothing to reduce the impact of noise in the label on the final classification result. It can be seen by comparing the papers using the same feature information that multi-aCRNN (ours) outperforms the 4D-CRNN model by 2.21% and 1.72% on the arousal task and the valence task, respectively, and surpasses the DE-CNN-BiLSTM model by 1.57% and 2.28%. We can conclude that selectively training on frequency, spatial, and temporal information is more conducive to emotion recognition, and reducing label noise positively affects emotion recognition.

5. Conclusions

This paper proposes a multiview feature fusion attention convolutional recurrent neural network model for EEG sentiment analysis. This method extracts more comprehensive spatial feature information through multiscale convolution and combines the frequency-domain features and spatial-domain features of EEG data. The weight fusion is carried out from the time series perspective so that the model learns more accurate information for classification prediction. Through the comparison experiment between GRU and bi-GRU networks, the bi-GRU network is determined as the network layer for temporal feature extraction. At the same time, the noise in the label is smoothed, which effectively reduces the impact of label noise on the classification results, realizes emotion recognition in complex practical situations, and is verified on the DEAP dataset. The multi-aCRNN model achieved 96.43% and 96.30% on the arousal task and valence task, respectively. At the same time, this paper conducts an experimental comparison of stacking and splicing of motor conversion methods to understand the impact of different feature combination methods on sentiment analysis. These experiments can be the basis for further research on EEG characteristics and better experimental research on emotion analysis.

Although the experiments in this paper effectively fuse multiview features and obtain high classification accuracy, the model still has shortcomings in classification tasks. In future work, we will extend multi-aCRNN to a multiclass classifier to obtain more accurate emotional state localization and achieve more accurate and good classification results.

Data Availability

The experimental data in this paper comes from the public data set DEAP, and the data set source link is http://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

XF and RH conceived the project, designed the experiments, and drafted the manuscript. RH, FB, PC, and ZF collected the data and conducted the experiments. RH, FB, and XF proofed and polished the manuscript and organized this project.

Acknowledgments

This work is supported by the Natural Science Foundation of Jilin Province (YDZJ202301ZYTS401 and YDZJ202301-ZYTS288), the Science and Technology Project of the Education Department of Jilin Province (JJKH20220245KJ and JJKH20220226SK), and the National Natural Science Foundation of China Joint Fund Project (U19A200496).

References

[1] K. Ajith, R. Menaka, and S. S. Kumar, "EEG based mental state analysis," *Journal of Physics Conference Series*, vol. 1911, no. 1, article 012014, 2021.

[2] A. H. Brooke and N. A. Harrison, "Neuroimaging and emotion," in *Stress: Concepts, Cognition, Emotion, and Behavior*, vol. 1, pp. 251–259, Elsevier, 2016.

[3] M. Iqbal, S. M. Ali, M. Abid, F. Majeed, and A. Ali, "Artificial neural network based emotion classification and recognition from speech," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 434–444, 2020.

[4] C. Wu, F. Wu, Y. Huang, S. Wu, and Z. Yuan, "THU_NGN at IJCNLP-2017 task 2: dimensional sentiment analysis for Chinese phrases with deep LSTM," in *Proceedings of the IJCNLP 2017, Shared Tasks*, pp. 47–52, Taipei, Taiwan, 2017.

[5] L. A. Moctezuma, T. Abe, and M. Molinas, "Two-dimensional CNN-based distinction of human emotions from EEG channels selected by multi-objective evolutionary algorithm," *Scientific Reports*, vol. 12, p. 1, 2022.

[6] M. Yen, Y. Huang, L. Yu, and Y. L. Chen, "A two-dimensional sentiment analysis of online public opinion and future financial performance of publicly listed companies," *Computational Economics*, vol. 59, no. 4, pp. 1677–1698, 2022.

[7] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[8] S. Shanmugam and I. Padmanaban, "Twitter emotion analysis for brand comparison using naive Bayes classifier," in *International Conference on Soft Computing and Its Engineering Applications*, pp. 199–211, Springer, 2021.

[9] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Computers in Human Behavior*, vol. 93, no. 4, pp. 309–317, 2019.

[10] D. Gorecky, M. Schmitt, M. Loskyll, and D. Zühlke, "Human-machine-interaction in the industry 4.0 era," in *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 289–294, Porto Alegre, Brazil, 2014.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[12] S. Wang, X. Wang, Z. Wang, and R. Xiao, "Emotion recognition based on static human posture features," in *International Conference on Computing, Control and Industrial Engineering*, pp. 529–539, Springer, Singapore, 2022.

[13] M. M. Rahman, A. K. Sarkar, M. A. Hossain et al., "Recognition of human emotions using EEG signals: a review," *Computers in Biology and Medicine*, vol. 136, no. 2, article 104696, 2021.

[14] R. N. Duan, J. Y. Zhu, and B. L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84, San Diego, CA, USA, 2013.

[15] L. Jin, S. Gao, Z. Li, and J. Tang, "Hand-crafted features or machine learnt features? Together they improve RGB-D object recognition," in *2014 IEEE International Symposium on Multimedia*, pp. 311–319, Taichung, Taiwan, 2014.

[16] N. H. Nguyen, D. T. A. Nguyen, and J. Hu, "The application of machine learning and deep learning in sport: predicting NBA players performance and popularity," *Journal of Information and Telecommunication*, vol. 6, no. 2, pp. 217–235, 2022.

[17] N. G. Paterakis, E. Mocanu, M. Gibescu, B. Stappers, and W. V. Alst, "Deep learning versus traditional machine learning methods for aggregated energy demand prediction," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pp. 1–6, Turin, Italy, 2017.

[18] Q. Chen, Q. Xie, Q. Yuan, H. Huang, and Y. Li, "Research on a real-time monitoring method for the wear state of a tool based on a convolutional bidirectional LSTM model," *Symmetry*, vol. 11, no. 10, p. 1233, 2019.

[19] Y. P. Lin, C. H. Wang, T. P. Jung et al., "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[20] W. Zheng, J. Zhu, Y. Peng, and B. Lu, "EEG-based emotion classification using deep belief networks," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Chengdu, China, 2014.

[21] B. I. García-Martínez, A. Martínez-Rodrigo, R. Alcaraz, and A. Fernández-Caballero, "A review on nonlinear methods using electroencephalographic recordings for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 801–820, 2021.

[22] Y. Zhu and Q. Zhong, "Differential entropy feature signal extraction based on activation mode and its recognition in convolutional gated recurrent unit network," *Frontiers in Physics*, vol. 8, no. 1, article 629620, 2021.

[23] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Applied Soft Computing*, vol. 100, no. 1, article 106954, 2021.

[24] F. Shen, G. Dai, G. Lin, J. Zhang, W. Kong, and H. Zeng, "EEG-based emotion recognition using 4D convolutional recurrent neural network," *Cognitive Neurodynamics*, vol. 14, no. 6, pp. 815–828, 2020.

[25] F. Cui, R. Wang, W. Ding, Y. Chen, and L. Huang, "A novel DE-CNN-BiLSTM multi-fusion model for EEG emotion recognition," *Mathematics*, vol. 10, no. 4, pp. 582–582, 2022.

[26] S. Koelstra, C. Mühl, M. Soleymani et al., "DEAP: a database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[27] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," in *The 25th International Conference on Neural Information Processing*, p. 433, Springer, 2018.

[28] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Rio de Janeiro, Brazil, 2018.

[29] H. Gao, S. Lin, Y. Yang, C. Li, and M. Yang, "Convolution neural network based on two-dimensional spectrum for hyperspectral image classification," *Journal of Sensors*, vol. 2018, Article ID 8602103, 13 pages, 2018.

[30] Y. Yan, Y. Mao, and B. Li, "SECOND: sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[31] M. Sun, Z. Song, X. Jiang, J. Pan, and Y. Pang, "Learning pooling for convolutional neural network," *Neurocomputing*, vol. 224, no. 8, pp. 96–104, 2017.

[32] A. Vaswani, N. M. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 5, no. 6-pp. 6000–6010, Curran Associates Inc, Long Beach, California, USA, 2017.

[33] J. P. Amorim, P. H. Abreu, M. Reyes, and J. Santos, "Interpretability vs. complexity: the friction in deep neural networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Glasgow, UK, 2020.

[34] J. Wang, P. Zhang, Q. He, Y. Li, and Y. Hu, "Revisiting label smoothing regularization with knowledge distillation," *Applied Sciences*, vol. 11, no. 10, pp. 4699–4699, 2021.