frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 15 August 2022
DOI 10.3389/fnins.2022.872311

OPEN ACCESS

EDITED BY
Varun Bajaj,
PDPM Indian Institute of Information Technology, Design and Manufacturing, India

REVIEWED BY
Chang Li,
Hefei University of Technology, China
Yu Zhang,
Lehigh University, United States

*CORRESPONDENCE
Yun Su
suyun@nwnu.edu.cn

SPECIALTY SECTION
This article was submitted to Neuroprosthetics, a section of the journal Frontiers in Neuroscience

RECEIVED 09 February 2022
ACCEPTED 20 July 2022
PUBLISHED 15 August 2022

CITATION
Su Y, Zhang Z, Li X, Zhang B and Ma H (2022) The multiscale 3D convolutional network for emotion recognition based on electroencephalogram. *Front. Neurosci.* 16:872311. doi: 10.3389/fnins.2022.872311

The multiscale 3D convolutional network for emotion recognition based on electroencephalogram

Yun Su[1]*, Zhixuan Zhang[1], Xuan Li[1], Bingtao Zhang[2] and Huifang Ma[1]

[1]School of Computer Science and Engineering, Northwest Normal University, Lanzhou, China, [2]School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China

Emotion recognition based on EEG (electroencephalogram) has become a research hotspot in the field of brain-computer interfaces (BCI). Compared with traditional machine learning, the convolutional neural network model has substantial advantages in automatic feature extraction in EEG-based emotion recognition. Motivated by the studies that multiple smaller scale kernels could increase non-linear expression than a larger scale, we propose a 3D convolutional neural network model with multiscale convolutional kernels to recognize emotional states based on EEG signals. We select more suitable time window data to carry out the emotion recognition of four classes (low valence vs. low arousal, low valence vs. high arousal, high valence vs. low arousal, and high valence vs. high arousal). The results using EEG signals in the DEAP and SEED-IV datasets show accuracies for our proposed emotion recognition network model (ERN) of 95.67 and 89.55%, respectively. The experimental results demonstrate that the proposed approach is potentially useful for enhancing emotional experience in BCI.

KEYWORDS

BCI, emotion recognition, EEG, 3D CNN, spatiotemporal features, deep learning

Introduction

Automatic EEG-based emotion recognition has been a research hotspot in the field of BCI and human–computer interaction over the past decade. Efficient emotion recognition methods based on EEG can prompt BCI to build a harmonious human-computer interaction environment, which can promote a natural, convenient, and friendly experience as communication between people (Zhang et al., 2015; Xie et al., 2019).

In general, there are two ways to recognize emotion, i.e., through non-physiological and physiological signals. Non-physiological signals (such as facial expressions, speech, gestures, etc.) can be artificially controlled (Yao, 2014). However, physiological signals (such as electroencephalogram (EEG), electrocardiograph (ECG),

electromyography (EMG), magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS) etc.) can show reliable and natural emotions without subjective control (Cai et al., 2018). EEG and MEG are the types of electrical signal produced by the brain that provides very useful information relating to emotional activity of the brain. Moreover, EEG and MEG have good temporal resolution and both are non-invasive. Some researchers (Xing et al., 2019; Sun et al., 2020) found that the MEG signal and fNIRS signal can realize emotional recognition, and could obtain the accuracy of high binary classification. However, MEG mainly reflects the inner structure of the brain. Nevertheless, EEG is used to check the function of the brain and mainly through brain waves reflecting the mood of the brain. As a result, EEG-based emotion recognition methods have become popular in current research.

Currently, traditional machine learning methods (Yoon and Chuang, 2013; Li et al., 2018) can effectively recognize emotions but require manual feature extraction and only consider the independence of a single feature in time or space. The two-dimensional convolutional neural network (2D-CNN) of deep learning can solve these problems. However, emotion recognition requires taking into account not only the time dependence between data points but also the spatial relevance between different electrodes of EEG signals (Yea-Hoon et al., 2018). In contrast to 2D-CNN (Mei and Xu, 2017), the emotion recognition method based on three-dimensional convolutional neural networks (3D-CNN) can meet these needs (Salama et al., 2018; Zhao et al., 2020). The 3D-CNN models can automatically extract spatiotemporal features. The existing emotion recognition model has achieved high accuracy, while most researchers believe that multiple smaller-scale kernels have the ability to increase non-linear expression more than a larger kernel. Therefore, how to define the convolutional kernel size in the convolutional network is still an interesting topic in emotion recognition research.

In this paper, we propose a four-class emotion recognition method based on a multiscale convolutional kernel 3D network, in which EEG-based emotional states can be efficiently recognized. First, we located the spatial position of the EEG signal electrode according to the 10–20 system diagram, the positional relationship between the positioning electrodes, and retained the spatial information of the EEG. The emotional recognition model based on three-dimensional EEG is generally used in the size of a consistent convolutional kernel. However, this paper attempts to use different smaller sizes of convolutional kernels, which are expected to increase the non-linear features of the data and the amount of data available. In addition, we join the double linear convolutional structure to the emotion recognition network model (ERN), and the EEG data are analyzed in parallel, thereby obtaining efficient recognition results.

According to existing studies, most researchers have applied two lengths of time windows, i.e., 1 s (1 s) and 2 s (2 s). To find the most suitable time window length for the ERN model, we compare the classification performance of 1 and 2 s time window lengths. The experimental results have shown that multiscale convolutional kernels with suitable time window lengths are more effective, which can improve the accuracy for emotion recognition. In summary, the main contributions of this paper are as follows. (1) We optimized the original dataset, designed a repositioned electrode topology, and constructed a 3D dataset for the model. (2) We enriched the emotion recognition method based on EEG and constructed a multiscale convolutional kernel 3D-CNN model to achieve more efficient emotion recognition performance.

The paper is structured as follows: related research is discussed, followed by consideration of the methodology adopted in our work. Experimentation and evaluation are addressed, and a discussion with results derived from the experimentation is presented. The paper closes with concluding observations and consideration of future work.

Related work

Currently, a variety of traditional machine learning-based emotion recognition methods have been documented in the literature, which also confirms the effectiveness and accuracy of traditional machine learning on emotion classification. However, machine learning-based emotion recognition methods require the specifically detailed design of classification models and manual extraction of temporal or spatial emotion features of EEG signals. For example, the traditional classifiers used in the literature include the support vector machine (SVM), and k-nearest neighbors (KNN) (Jenke et al., 2014). However, the use of traditional machine learning requires the manual extraction of relevant emotion features, limited to the temporal or frequency domains, with domain knowledge barriers and timeliness problems.

Emotion recognition methods based on deep learning can solve these problems. The deep network can extract different types of features at the same time, with the advantages of automatic detection features, and solve the dependence of artificial feature extraction. For example, Lin et al. (2017) proposed an emotion recognition method based on CNN, converting EEG data from the signal format into an image format containing time domain and frequency domain information, combined with the characteristics of other physiological signals, which were input into the pretrained AlexNet (Krizhevsky et al., 2012) network model for emotion recognition. Kwon et al. (2018) also proposed a sentiment classification method for extracting features based on the 2D CNN model, which were preprocessed before convoluting EEG signals by wavelet transformation considering both the time and frequency domains to improve the recognition performance.

In addition to CNN model, there are other methods effectively identify emotions. Liu et al. (2018) proposed that

the Residual Network-50 (ResNet-50) model can automatically learn deep semantic EEG information and classify the new features of the fusion of linear-frequency cepstral coefficients (LFCC). Xing et al. (2019) used a stack auto-encoder (SAE) to build and solve the linear EEG mixing model and the emotion timing model based on the long short-term memory recurrent neural network (LSTM-RNN). Liu et al. (2020) proposed a capsule network based on the multi-level feature boot, which can recognize multi-electrode EEG emotions. In the same year, Tao (Tao et al., 2020) proposed a convolutional recursive neural network (ACRNN) based on the attention mechanism, which can extract more discriminant features from the EEG signal, and improve the accuracy of emotional identification.

However, these models ignore the spatial structure of the EEG and the variations and distortion of the electrodes in each dimension. With the study of deep learning neural networks, methods for extracting spatiotemporal features have been proposed. Yang et al. (2018) implemented a hybrid neural network integrating a CNN and a recurrent neural network (RNN) so that network models could extract and integrate spatiotemporal features. Wang et al. (2018) proposed a simple and efficient preprocessing method that converts multiple electrodes of EEG signals into electrode topological maps containing topological location information. An et al. (2021) proposed an EEG emotion recognition algorithm based on 3D feature fusion and a convolutional auto-encoder (CAE). Zhao et al. (2020) proposed a 3D-CNN model to automatically extract the spatiotemporal features of EEG signals, introduced the preprocessing method for baseline signal and electrode topology relocation, compared the performance of the 2D convolutional kernel and 3D convolutional kernel in detail, and showed that the 3D-CNN model was more advantageous.

Different convolutional network recognition models set different convolutional kernel sizes, while most researchers (Szegedy et al., 2016) believe that multiple smaller scale kernels have the ability to increase non-linear expression more than a larger kernel. The different sizes of kernels in the network can increase the number of features that can be used and improve model performance. In this paper, we propose a four-class emotion recognition method based on a multiscale convolutional kernel 3D network, in which EEG-based emotional states can be efficiently recognized. The experimental results on the DEAP and SEED-IV datasets show that the proposed model has preferable performance than the other existing models in terms of recognition accuracy.

Materials and methods

The experimental process proposed in this paper is shown in **Figure 1**. In **Figure 1**, first, the original EEG signals are preprocessed. Then, the preprocessed data are converted from the 2D form to the 3D format and divided into two kinds of datasets: training data and testing data. Finally, we input preprocessed data and evaluated our ERN model by the recognition results of the four classes of labels.

Processing

To improve the recognition accuracy, it is necessary to preprocess the raw EEG data. First, the brain electrode data are selected and subsampled, and the noise artifacts are removed using a bandpass frequency filter of 4.0–45.0 Hz. A preprocessing method for the baseline EEG signal will affect the recognition result (Yang et al., 2018). Therefore, the specific practice of the data baseline signal processing is as follows. First, the first 3 s of baseline signals are extracted from all electrodes c of a single subject, and then cut into a fragment of the N-section fixed length L, thereby obtaining the $N \times C \times L$ matrix. Then, calculate the average of this $N \times C \times L$ matrix, obtain the Z matrix, and the structure of the Z matrix is $C \times L$. The last 60 s of the signals are divided into M fragments to obtain the matrix of $M \times C \times L$. Then the matrix of $M \times C \times L$ subtracts the average matrix Z. N and M are the number of data segments, C is the number of electrodes, and L is the data length. This calculation step can obtain all the data of a single subject and be repeated, and we can obtain all the pretreatment data.

The 32 electrodes of the EEG signals in the dataset are repositioned to a 2D electrode topology based on the International 10–20 System Diagram to acquire the spatial information of the EEG. During the recognition of emotional type based on EEG, Zhong et al. (2020) confirmed that both the position of signals acquisition and the interaction of EEG electrode position are conducive to improving the accuracy of emotion recognition based on EEG signals. Therefore, to retain the spatial information of the EEG, we located the spatial position of the EEG signal electrode according to the 10–20 system diagram (the positional relationship between the positioning electrodes), and retained the spatial information of the EEG. Based on the topological location information of the vulnerable electrodes during the original EEG emotion analysis experiment, Zhong and An (Zhong et al., 2020; An et al., 2021) proposed a solution by repositioning the 32 electrodes of the EEG signals in the dataset to the 2D electrode topology based on the international 10–20 system diagram, thereby preserving spatial information among electrodes. In this paper, 1-dimensional (1D) electrodes of the obtained dataset are repositioned into a 2D electrode topology.

As shown in **Figure 2A**, we choose the 32-electrode EEG data of the dataset, which is located in the International 10–20 system diagram (Sharbrough et al., 1991). According to the farthest distance between the two electrodes in **Figure 2A**, we set the size of the two-dimensional matrix and the size of the two-dimensional matrix is 9×9. Then, the selected 32 EEG signals are mapped to the 9×9 matrix. In **Figure 2B**, the

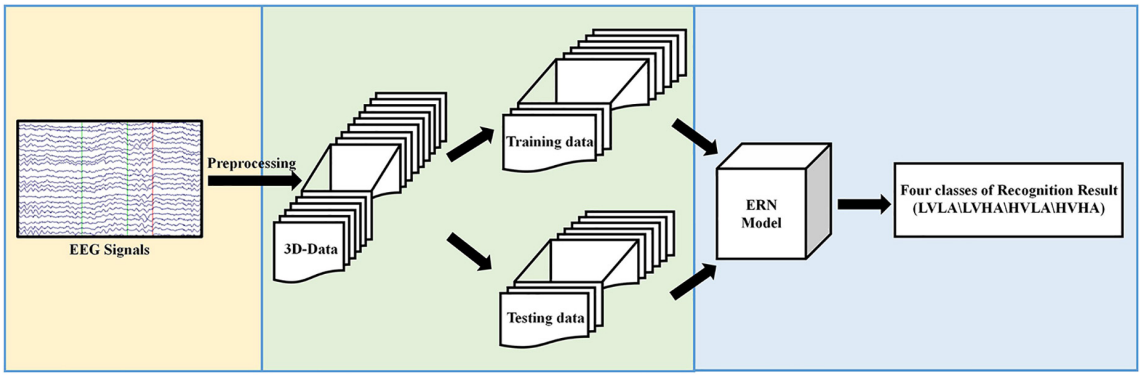

FIGURE 1
Overview of the proposed approach.

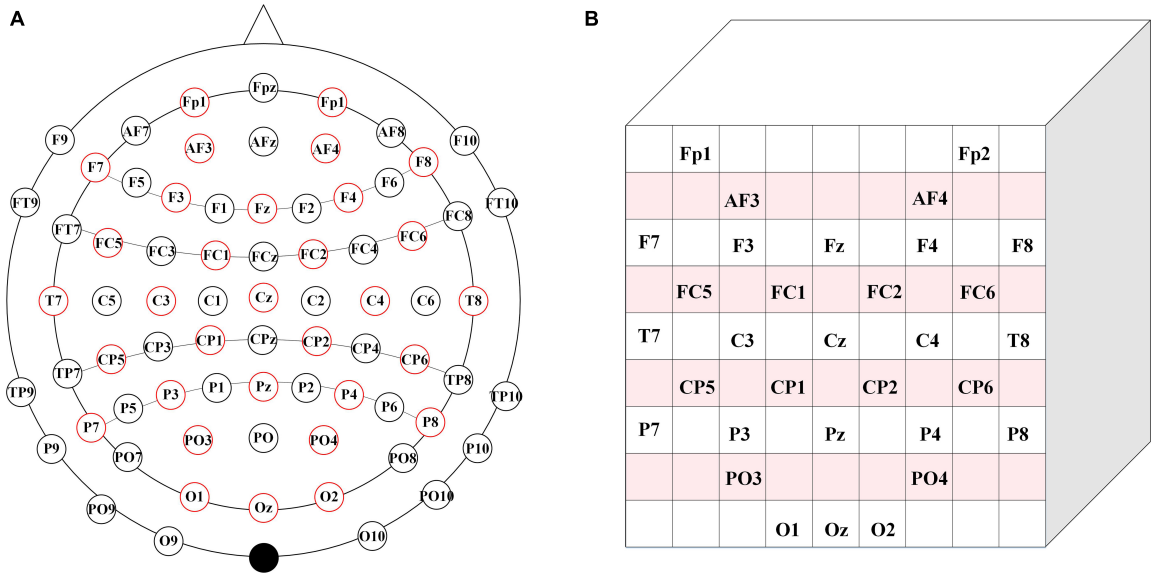

FIGURE 2
Brain electrode distribution diagram. **(A)** Brain electrode distribution in 10–20 system diagram. **(B)** The corresponding matrix.

positioning of each electrode is located based on the positional relationship between the various electrodes in **Figure 2A**. The blank position is represented as a topological position of the unselected physiological signal. Therefore, unused topologies are set to zero in the 9 × 9 matrix, and the matrix is normalized.

Model

We design the multiscale convolutional kernel 3D-CNN model based on the final obtained 3D EEG dataset. The reasons for the use of this network are as follows: First, the advantage of the convolutional network (Zeiler and Fergus, 2014b) is that it can calculate the eigenvalue rather than the original value with no need for the accurate mathematical expression between inputs and outputs. Second, the convolutional network can avoid the problem of gradient loss when reverse propagation occurs in the BP neural network. The CNN can be trained in parallel, which reduces the complexity of the network. In particular, the network can directly input multi-dimensional data directly, which avoids the complexity of data reconstruction during feature extraction and classification. The flexibility of the three-dimensional convolutional kernel is higher than that of the two-dimensional convolutional kernel, which helps to learn the advanced representation of learning information (Tran et al., 2015). The controllable range of the three-dimensional convolutional kernel is expanded to the spatial domain, which can utilize the interaction between the electrodes and increase the identification ability of the model.

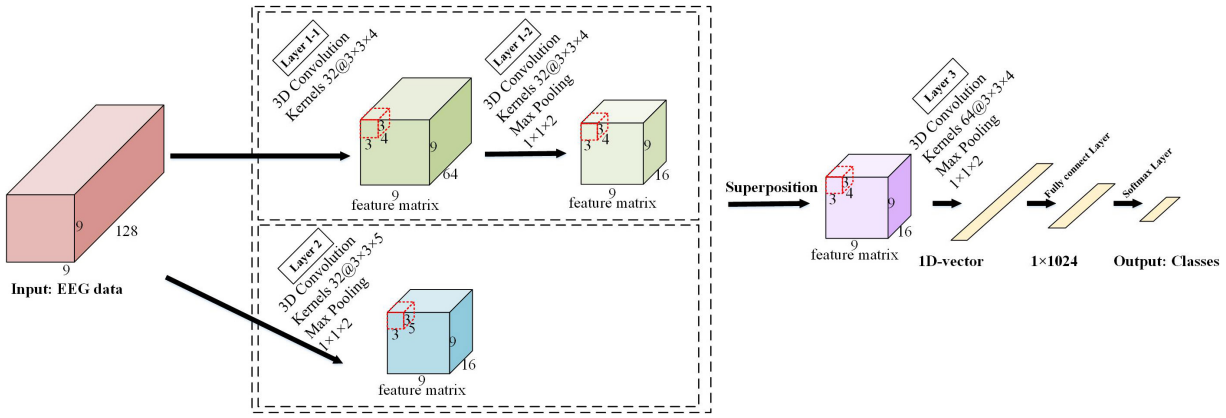

FIGURE 3
The framework of the ERN model.

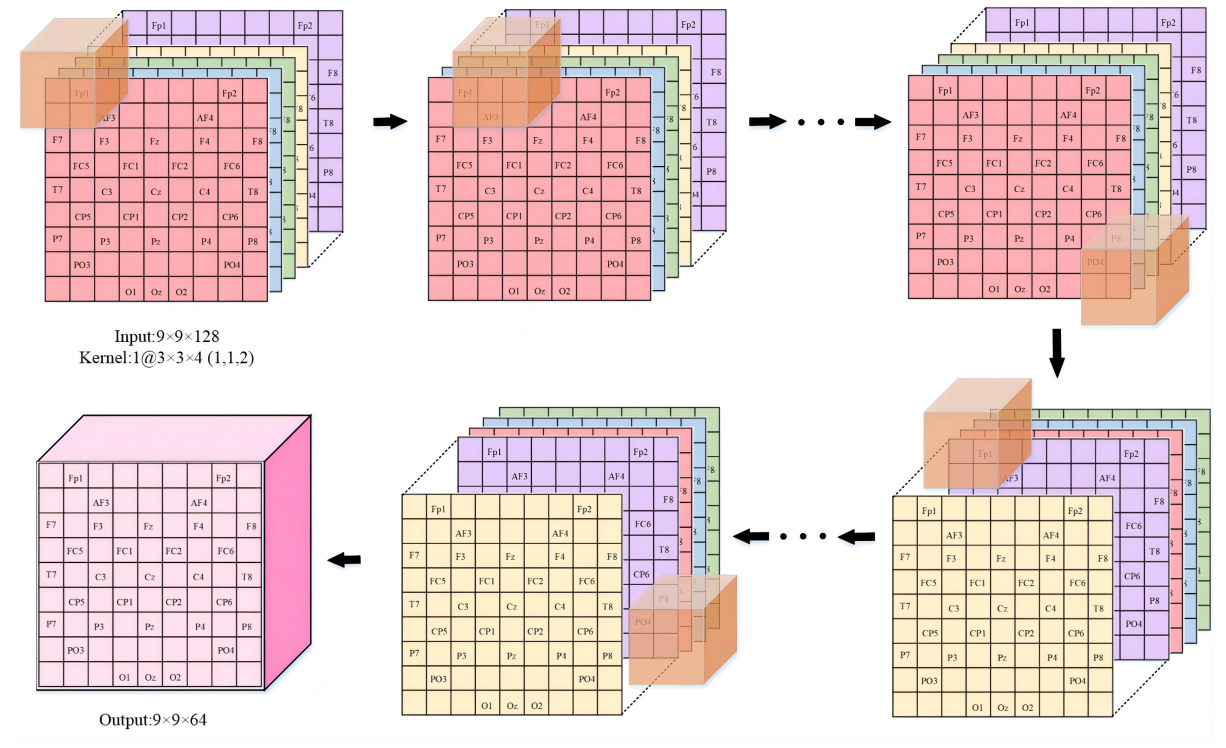

FIGURE 4
The feature representation learned in ERN. (Overview of the proposed approach detailed convolutional operations).

The detailed architecture of the ERN model is shown in Figure 3. The architecture of this 3D-CNN model consists of three convolutional layers, with the first convolutional layer implemented in parallel with the second convolutional layer in the model. The kernel size is 3 × 3 × 4 in the first convolutional and the third convolutional layers, where spatiotemporal features are generated by the local spatial topology of 3 × 3 and fragments of the temporally sampled point 4. The kernel size is 3 × 3 × 5 in the second convolutional layer and combines advanced spatial features via a local spatial topology of 3 × 3 and a temporal sampling point of 5. The first convolutional layer and the second convolutional layer are used to calculate the feature images, and the obtained feature images are superimposed to obtain a new feature image. Multiple 3 × 3 kernels have more non-linear functions than a larger convolutional kernel, which increases the non-linear expression and makes the judgment

function more efficient. Selecting more small convolutional kernels favors more accurate emotion recognition. Under the conditions of ensuring the same kernel, the depth of the network is improved, the parameters of the model are reduced, and the effect of the neural network is improved to some extent.

We provide a detailed description in **Figure 4** to visualize the feature representation learned by the hidden layers of the ERN model. Taking 3 × 3 × 5 as an example, the step size of the convolutional kernel is 1, the format of the input data is 9 × 9 × 128, and the resulting feature format is 9 × 9 × 64. The figure demonstrates the detailed process of the convolutional operation. Each convolutional kernel is used to compute features and moves with a fixed step on the dataset. We can calculate the size of the features according to the equation (1). $I_n(n = 1,2,3)$ is defined as the size of the input convolutional layer data, $O_n(n = 1,2,3)$ is the size of the output convolutional feature, $K_n(n = 1,2,3)$ is the convolutional kernel size, n is one of the three dimensions, N is the number of convolutional kernels, S is the moving step of the convolutional kernel, and P is padding value; this article set P is 1. The size of the three dimensions of the feature shown in **Figure 4** can be calculated separately by the equation (1). Equation (1) can calculate the size of the 3D features map.

$$O_n = \frac{I_n - K_n + 2P}{S} + 1 \quad (1)$$

A 3D maximum pooling layer is set behind each convolutional layer, and the kernel size is 1 × 1 × 2. The maximum pooling layer (Scherer et al., 2010) is used to extract the features more efficiently here; it can reduce the quantity of data on the time dimension and improve the robustness of the extracted features and provides a better generalization. The first maximum pooling layer and the second maximum pooling layer further squeeze the extracted spatiotemporal features to generate advanced high-level spatiotemporal features. The last pooling layer is followed by a fully connected layer, and the Softmax layer is deployed as the output. In the experiment, the input data size of the model is 9 × 9 × 128, where 9 × 9 is the 2D electrode topology and 128 is the number of continuous-time sampling points for one treatment. The number of feature maps for the last convolutional layer is 64, passing the 64 feature maps to the fully connected layer, which maps the input as vectors. The N in its output represented the number of labels in the task. The empty is set to zero in each convolutional layer to prevent the loss of information from the input data, and the ReLU activation function is used after each convolutional layer.

Experiment

We test the model in the public databases DEAP (Koelstra et al., 2012) and SEED-IV (BCMI, 1994). We use the PyTorch framework (Chaudhary et al., 2020) to implement this model and deploy it on the GeForce RTX 3060. The learning rate is set to 0.001 with the Adam AdaDelta Optimizer, and the probability of the dropout operation is set to 0.6. We use 10-fold cross-validation to evaluate the performance of the ERN model. The average accuracy of the 10-fold validation processes is taken as the final result.

Processing

DEAP dataset

The multimodal DEAP dataset is an open multimodal standardized dataset used to study the analysis of human emotional states. The dataset includes the 32 electrodes of EEG signals and the 8 electrodes of peripheral physiological signals when subjects watch music videos. After watching a video, subjects scored each video based on four psychological scales of arousal, valence, liking, and dominance. We select 32 electrodes of EEG signal data from the dataset for the analysis of the human emotional state.

The preprocessing step is as follows: First, the data are downsampled from 512 to 128 Hz, and then a bandpass frequency filter of 4.0–45.0 Hz to remove noise artifacts. At this time, the processed dataset data structure is 40 × 32 × 8,064 (video number × EEG electrode number × signal data), of which 8,064 signal data contained 384 baseline signals. The DEAP dataset is divided into two parts, as shown in **Table 1**. The data matrix refers to the EEG of 40 electrodes observed when each subject watched music videos. The label matrix refers to the four types of labels after each subject watches videos: arousal, valence, dominance, and liking.

In this dataset, each video of each stimulus is 60 s so that the first 3 s is the baseline signals of the unstimulated, and the last 60 s is the signals of the stimulus in the 63 s signals of each stimulus trial. Therefore, we need to carry out baseline signal processing for each trial signal after preprocessing. The processing step is as follows: For each trial signal (32 × 8,064), cut the baseline signal (32 × 384) of the first 3 s to 3 segments (32 × 128) and calculate the mean value of the baseline signals (32 × 128). Then, the signal data of the last 60 s are cut into 60 segments (32 × 128), and the mean of the baseline signal is substracted

TABLE 1 The DEAP dataset and SEED-IV dataset.

Matrix name	Matrix structure representation
DEAP Dataset	
$Data_{40 \times 40 \times 8,064}$	$Data_{video \times channel \times fixed\ point\ in\ time}$
$Label_{40 \times 4}$	$Label_{video \times value}$
SEED-IV Dataset	
$Data_{15 \times 62 \times *}$	$Data_{video \times channel \times fixed\ point\ in\ time}$
$Label_{15 \times 1}$	$Label_{video \times value}$

and merged to obtain the processed signal (32 × 7,680). Next, each electrode needs to be repositioned to the two-dimensional topological location to learn the spatial properties of the data. 128, 384, 7,680, and 8,064 are the time points and 32 stands for the number of electrodes. To extract the spatiotemporal features, the EEG data are mapped into a 9 × 9 matrix based on the International 10–20 system diagram. Finally, the matrix is cut into fragments with a length of 1 s (9 × 9 × *), and the 3D electrode topology was obtained (7,680 9 × 9 × *), of which the symbol "*" represents the size of the time window and 7,680 represents the number of matrixes.

Specifically, for the selected labels, the valence describes the degree of pleasure associated with the stimulus, represented by continuous values ranging from 1 (negative) to 5 (neutral) to 9 (positive). Arousal represents the degree of waking to the stimulus with the same range, with 1 and 9 indicating negative and positive, respectively. As shown in Table 2, we set the distribution of 4 label values based on EEG arousal and valence markers: low valence vs. low arousal (LVLA), low valence vs. high arousal (LVHA), high valence vs. low arousal (HVLA), and high valence vs. high arousal (HVHA). As shown in Table 2, we set the values of the four types of labels based on arousal and valence and set 5 as the threshold. After processing, the label structure is 40 × 1 (number of videos × label value).

SEED-IV dataset

We also use the SEED-IV dataset as a standardized dataset to study the model recognition performance of this paper, which is a well-formed multimodal dataset for emotion recognition. In the SEED-IV dataset, a total of 15 subjects participated in the experiment. For each subject, the test, respectively, was performed on three different days and each test contained 24 trials. In each trial, his or her EEG signals are saved when the subject watches each film clip.

The preprocessing step is as follows: First, the same 32-electrode EEG data as DEAP in the SEED-IV dataset were selected to analyze the human emotional state. Then, the data are downsampled from 1,000 Hz to 128 Hz using a bandpass frequency filter of 4.0–45.0 Hz to remove noise artifacts. The SEED-IV dataset is divided into two parts, as shown in Table 1. The data matrix refers to the physiological data of 62 electrodes observed that include the EEG signal and peripheral physiological signal. In the data matrix, the length of the movie clips resulted in the different lengths of the EEG data in each trial. The label data matrix refers to the four types of labels played when a subject watches the film clips: happy, sad, neutral, and fear.

In the SEED-IV dataset, the length of the data varies in each trial. Therefore, we need to select the data length suitable for the model after preprocessing in each stage of each subject and obtain 15 matrixes, each with 32 rows and 128 columns (32 × 128), and 15 is the number of movie clips. In each 32 × 128, the 32 is the number of EEG electrodes and 128 is the data of 1 s. Then, 32 EEG electrodes are mapped into the matrix with 9 rows and 9 columns, and 15 3D-matrixes (9 × 9 × 128) are obtained. Finally, we combine the data from three stages of 15 subjects and obtain 675 (15 × 3 × 15, subject number × stage number × video number) segments of the total dataset (9 × 9 × 128). After processing, the label structure is 15 × 1 (number of videos × label value).

TABLE 2 Label values in the DEAP.

Label	LVLA	LVHA	HVLA	HVHA
Valence	≤ 5	≤5	> 5	>5
Arousal	≤ 5	> 5	≤ 5	> 5
Value	0	1	2	3

Optimal time window in model

We test and select the time window length more suitable for the experiment to obtain the best recognition result of the model and apply the 1 s time window. To improve the accuracy of the data in the experimental input model, one of the solutions of this paper is the application of the time window. The available window size is not necessarily fixed, it can constantly expand until certain conditions are met, it can be constantly reduced until a minimum window to meet the conditions is found, and it can be a fixed size.

In the literature, people (Zhao et al., 2020) confirmed that the average classification accuracy of a 1 s period based on EEG was superior to other periods, and selected the 1 s length as the most appropriate time window length. However, someone (Candra et al., 2015) believed that a length of 2 s was the most appropriate time window length. To address this problem, we compare all EEG classification performances at two different time window lengths: 1 and 2 s. In Figure 5, the left side of the figure shows the recognition result trend of 1 s time window data, and the right side of the figure shows the recognition result trend of 2 s time window data. The classification accuracy of the 1 s time window is better than that of the 2 s time window with the same number of iterations, and no overfitting occurs in the first 500 epochs. Therefore, we compare the results of the different time windows, select the time window that is more suitable for the model, and improve the accuracy of the experiment.

Since we have chosen the more suitable size of the time window for the ERN model, and the next step is the analysis of whether the ERN model needs to set the overlapping window. Taking the DEAP dataset as an example, we set 0.5 s overlapping windows to process the dataset. Depending on the identification result after using the overlapping window, there may be multiple problems. First, the baseline processing method requires each second of data to subtract the mean of the baseline data. This

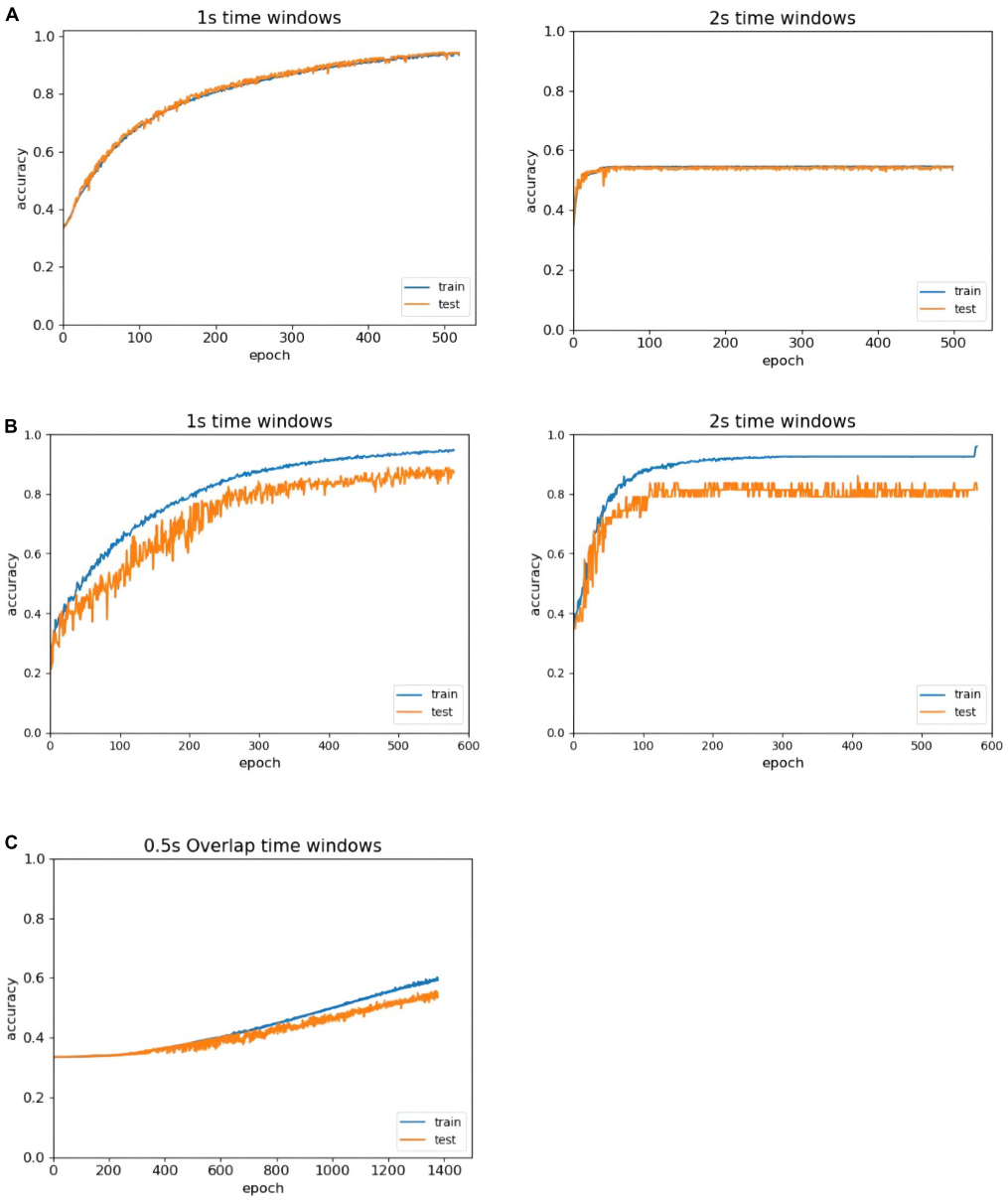

FIGURE 5
Accuracy comparison of different time window lengths. **(A)** Comparison of different time windows in the DEAP without overlapped time window. **(B)** Comparison of different time windows in the SEED-IV without overlapped time window. **(C)** The recognition of overlapped window in the DEAP.

method can disconnect the time continuity of data and eliminate the advantage of high time resolution. Second, the amount of data doubled after using the 0.5 s overlapping time window. About another data set SEED-IV, the data length is 128 after the process of downsampling. The pre-processed data in the SEED-IV have high time continuity, so there are no setup 0.5 s overlapping windows to process this dataset. The overlapping window processing generates a large amount of available data. But in the same number of iteration, the recognition efficiency by overlapping window processing is much lower as shown in Figure 5C. To achieve out high-efficiency recognition accuracy, we have not used overlapping windows.

Results and discussions

Accuracy

To estimate the accuracy of the emotion recognition for the ERN model, we conducted a comparative analysis that

compared the results obtained from our proposed approach applied over the DEAP and SEED-IV datasets with other methods. All of the classification procedures were conducted under 10-fold cross-validation (Yu et al., 2015), and the average value of 10 accuracies, F_1 values and AUC (the area under the ROC curve) values were calculated as the evaluation standard of model accuracy. In the process of training, the optimal weight parameters are obtained and set by random training data to avoid the overfitting problem caused by optimization.

Table 3 shows the 10-fold cross-validation accuracy in the DEAP and SEED-IV datasets. The 2D-CNN only utilizes high features of EEG time resolution in EEG and neglects space information. In contrast to 2D-CNN (Mei and Xu, 2017), the emotion recognition method based on three-dimensional convolutional neural networks (3D-CNN) can meet the need (Salama et al., 2018; Zhao et al., 2020) for spatial information. Not only does the 3D-CNN extract the time characteristics based on the 1 s time window, but it can also obtain the spatial feature between the electrodes. The previous literature report and the ERN model experiments in this article show that our model has the ability to improve the accuracy of EEG-based emotion recognition.

In **Table 4**, the average accuracy of our model for four emotion classes is up to 95.67% in DEAP and 89.55% in SEED-IV datasets, which is higher than previously reported models where 93.72 and 87.71% accuracies were achieved. It is known that the model based on deep learning proposed by Qiu et al. (2018) and Zhao et al. (2020) currently has the best performance in the DEAP dataset and SEED-IV dataset. However, the results of the ERN model are approximately 1.95 and 1.84% higher than those models. Compared with the CNN model (Mei and Xu, 2017; Salama et al., 2018; Tao et al., 2020; Zhao et al., 2020; Yin et al., 2021) and compared with other methods (Zangeneh et al., 2019; Song et al., 2020; An et al., 2021; Li S. et al., 2021) in the DEAP dataset, our model adopts a simpler and more efficient structure and has the best performance. Our model has higher speed efficiency and better identification performance than other models (Qiu et al., 2018; Zheng et al., 2019; Acharya et al., 2020) in the SEED-IV dataset.

For classification, cross-validation is not effective protection against overfitting or overhyping. It would be better to use techniques such as lockboxes, blind analyses, pre-registrations, or nested cross-validation to limit overhyping. We use the lockbox (Hosseini et al., 2020) approach to determine whether overhyping has occurred in the CNN model. The lockbox approach is a new technique that can be used to determine whether overhyping has occurred. The lockbox is accessed just one time to generate an unbiased estimate of the model's performance. In the DEAP and SEED-IV datasets, 90% of the data are set aside in the hyperparameter optimization set and the remaining 10% of the data are set aside in a lockbox. With the 10-fold cross-validation approach, the hyperparameters in the ERN model can be iteratively modified on the hyperparameter optimization set. When the average accuracy in the model is good enough, the model is tested on the lockbox data.

As shown in **Table 5**, the training result on the hyperparameter optimization set and the testing result on the lockbox set are 98.59 and 95.67% in the DEAP, 93.05 and 89.55% in the SEED-IV. According to the identification result, we can obtain the following conclusions. The theta band is in the state of sleep and a less responsive emotional state, so the recognition rate of emotion is lower than that of the other three waveforms. The excitement state of alpha, beta, gamma waveforms increased successively, so the recognition accuracy of emotion is higher.

TABLE 3 Ten-fold cross-validation accuracy in the DEAP and SEED-IV datasets.

Fold ID	DEAP	SEED-IV
Fold 1	94.78%	88.56%
Fold 2	95.83%	85.07%
Fold 3	95.42%	92.54%
Fold 4	97.08%	94.78%
Fold 5	93.87%	94.04%
Fold 6	96.67%	82.29%
Fold 7	96.78%	85.46%
Fold 8	93.75%	89.48%
Fold 9	95.85%	90.78%
Fold 10	96.67%	92.47%
Mean	95.67%	89.55%

TABLE 4 Comparison of ERN model with previous studies.

Research	Year	Method	Accuracy
DEAP dataset			
Mei and Xu	2017	2D-CNN	73.10%
Salama et al.	2018	3D-CNN	88.49%
Zangeneh et al.	2019	HcF+KNN+MSVM	86.01%
Song et al.	2020	DGCNN	90.4%
Zhao et al.	2020	3D-CNN	93.53%
Tao et al.	2020	ACRNN	93.72%
Li S. et al.	2021	The binary gray wolf optimization algorithm+SVM	90.48%
Yin et al.	2021	GCNN+LSTM	90.53%
An et al.	2021	3D Feature Fusion+CAE	90.76%
Our model	2021	3D-CNN	95.67%
SEED-IV Dataset			
Zheng et al.	2015	DBN	86.08%
Qiu et al.	2018	CAN	87.71%
Zheng et al.	2019	EmotionMeter	85.11%
Acharya et al.	2020	LSTM	87.22%
Our model	2021	3D-CNN	89.55%

F1 and *ROC*

We also calculated the F_1 value and *ROC* value to analyze the performance of the ERN. In Formulas (2) and (3), F_1 is the unweighted average of multiple categories of F_{1_m} (m = 0,..., C, C = 3), where m means four classes of emotion: LVLA, LVHA, HVLA and HVHA. F_{1_m} is calculated from $Precision_m$ and $Recall_m$, and $Precision_m$ and $Recall_m$ are calculated from FN_m, TP_m, TN_m and FP_m. In Formulas (4) and (5), FN_m and TP_m, represent the number of (incorrectly) recognized samples of a certain category, FP_m and TN_m, represent the number of (incorrectly) recognized samples of other categories except the m-th emotion category. According to Formula (6) and Formula (7), the true positive rate (*TPR*) and false positive rate (*FPR*) are calculated by FN_m, TP_m, TN_m and FP_m, and the *ROC* curves of the model are calculated to obtain AUC_m (the area under the *ROC* curve) of each category.

$$F_1 = \frac{1}{C}\sum_{m=0}^{C} F_{1_m} \tag{2}$$

$$F_{1_m} = \frac{2 \times Precision_m \times Recall_m}{Precision_m + Recall_m} \tag{3}$$

$$Precision_m = \frac{TP_m}{TP_m + FP_m} \tag{4}$$

$$Recall_m = \frac{TP_m}{TP_m + FN_m} \tag{5}$$

$$TPR = \frac{TP_m}{TP_m + TN_m} \tag{6}$$

$$FPR = \frac{FP_m}{TN_m + FP_m} \tag{7}$$

After several iterations, the values of the four classes are shown in Table 6. In the DEAP dataset, the LVLA, LVHA, HVLA, and HVHA values are 97.51, 98.62, 98.03, and 98.76, respectively. In the SEED-IV dataset, the LVLA, LVHA, HVLA, and HVHA values are 99.92, 98.97, 99.94, and 99.85, respectively. As shown in Figure 6, the solid red line is the average ROC curve of the four categories, and the average AUC is 98.23 for DEAP and 99.33 for SEED-IV.

In the operation process, mass test data and the large threshold distance between the two samples result in the ROC curve not being smooth in Figure 6. The higher values of the four categories indicate that the model constructed in this paper has better performance, among which the data of the fourth category (HVHA) have higher identifiability. Each column of the confusion matrix represents the predictive category, and the total number of each column indicates the number of data predicted for this category. Each line represents the real category of the data, and the total number of each line of data represents the number of data instances of the category in Figure 7. The matrix verifies that our model is stronger than others in predicting complex labels.

From our experimental results shown in Tables 4–6, Figures 6, 7, it can be seen that our results are superior to those of previous studies reported in the literature over the same EEG datasets from DEAP and SEED-IV. There are 3 possible reasons. (1) We conduct a simple and efficient preprocessing method, including data baseline signal processing, EEG electrode topological mapping, and 1 s time window length selection. (2) A 3D convolutional structure is a necessary technique to study emotional recognition based on EEG signals, as this structure can identify space information of the electrodes to quickly extract spatiotemporal features. (3) The multiscale convolution kernel not only reduces the computation of the model but also improves the identification ability of the model because multiple smaller scale kernels have the ability to increase non-linear expression more than a larger kernel. These all enhance the operating efficiency and improve the recognition performance of the ERN model.

TABLE 5 Comparison of ERN model with different bands.

Modality	Train result (%)	Test result (%)
EEG signals (DEAP)		
Theta	87.81	84.40
Alpha	92.47	88.74
Beta	97.01	90.91
Gamma	94.39	88.31
EEG	98.59	95.67
EEG signals (SEED-IV)		
Theta	83.08	82.09
Alpha	90.48	86.57
Beta	85.65	85.07
Gamma	92.45	88.06
EEG	93.05	89.55

Fisher

The parietal lobe of the human brain and non-human primate brain have been associated with attention based on the evidence of clinical and physiological (Joseph, 1990). Studies in literature show that the lateral Intraparietal area (LIP) plays

TABLE 6 F_1 values and AUC values of the four-class.

Dataset	m	0	1	2	3	Mean
DEAP	F_{1_m}	94.82	95.06	94.52	96.742	95.29
	AUC_m	97.51	98.62	98.03	98.76	98.23
SEED-IV	F_{1_m}	95.652	96.97	92.683	85.714	92.75
	AUC_m	99.92	98.97	99.94	99.85	99.67

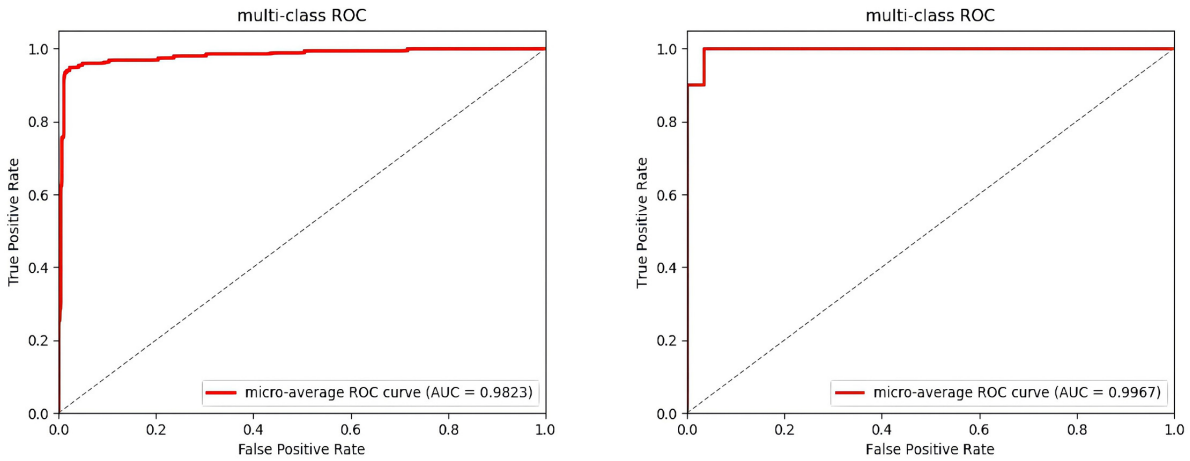

FIGURE 6
Average ROC of the four-classes. (The ROC of the DEAP in the left, the ROC of the SEED-IV in the right).

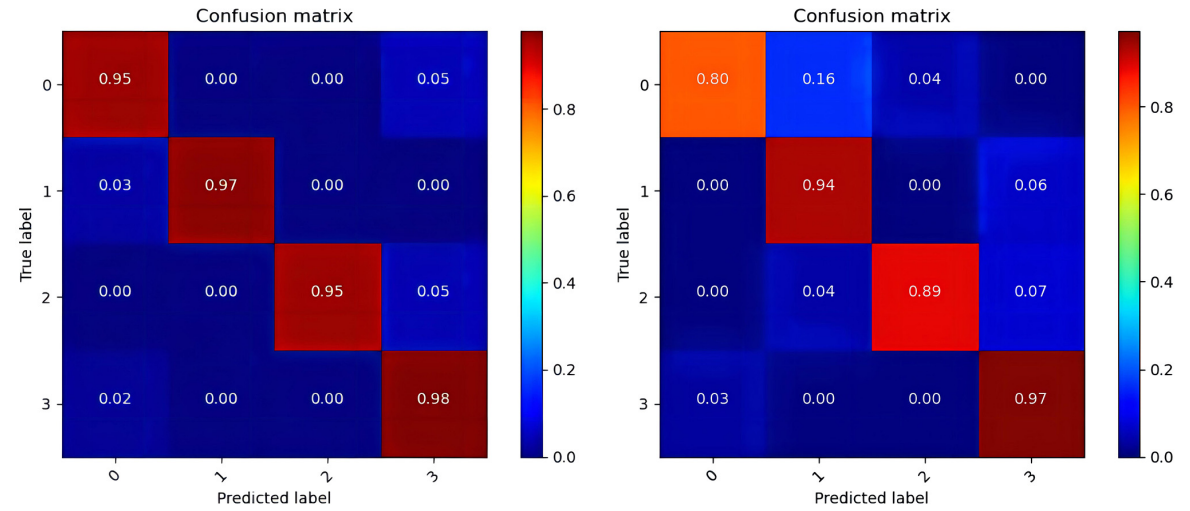

FIGURE 7
The confusion matrixes of datasets. (The confusion matrix of the DEAP in the left, the confusion matrix of the SEED-IV in the right).

an independent role in target selection and visual attention generation (Simon et al., 2002). These findings can be validated by the distribution of Fisher's selected EEG electrodes. We score the 32 electrodes of the applied dataset using the *Fisher* (Li R. et al., 2021) scorer and rank the 32 electrodes from highest to lowest to obtain the top 8, top 16 and top 24. In Formula (8), $\mu_{m,v}$ is the average of each electrode of the m-th class, μ_v is the average of each electrode, $\sigma_{m,v}$ is the variance of each electrode of the m-th class and n_m is the number of samples of the m-th class (m = 0,...,M, M = 3).

$$Fisher = \frac{\sum_{m=0}^{M} n_m(\mu_{m,v} - \mu_v)^2}{\sum_{m=0}^{M} n_m \sigma_{m,v}^2} \quad (8)$$

We can sort the impact sequence of different electrodes based on emotions by the recognition results. According to the output results of the ERN model, we can study the recognition performance of the model for many EEG electrodes. **Table 7** shows that the emotion recognition results of different quantities of electrodes in the sort. Experimental results showed that there is a 4.7% difference between the data of the first 8 electrodes and the whole dataset in DEAP and a 5.97% difference between the data of the first 8 electrodes and the whole dataset in SEED-IV. This indicates that the model can quickly obtain high recognition performance with less EEG electrodes data. We can use the ERN model for portable emotional identification (Cai et al., 2018) based on EEG and the model meets simple

TABLE 7 Evaluation of recognition results based on *Fisher*.

Dataset	The number of electrodes	The name of electrodes	Accuracy
DEAP	8	O2\PO4\AF3\F3\F7\FC5\FC1\C3	90.31%
	16	O2\PO4\AF3\F3\F7\FC5\FC1\C3\T7\CP5\CP1\P3\P7\PO3\O1\Oz	92.07%
	24	O2\PO4\AF3\F3\F7\FC5\FC1\C3\T7\CP5\CP1\P3\P7\PO3\O1\Oz\ Pz\Fp2\AF4\Fz\F4\F8\FC6\FC2	93.03%
	32	O2\PO4\AF3\F3\F7\FC5\FC1\C3\T7\CP5\CP1\P3\P7\PO3\O1\Oz\ Pz\Fp2\AF4\ Fz\F4\F8\FC6\FC2\Cz\C4\T8\CP6\CP2\P4\P8\Fp1	95.67%
SEED-IV	8	Oz\O2\FP2\AF3\AF4\F7\F3\Fz	83.58%
	16	Oz\O2\FP2\AF3\AF4\F7\F3\Fz\F4\F8\FC5\FC1\FC2\FC6\T7\C3	85.07%
	24	Oz\O2\FP2\AF3\AF4\F7\F3\Fz\F4\F8\FC5\FC1\FC2\FC6\T7\C3\Cz\C4\ T8\CP5\CP1\CP2\CP6\P7	86.57%
	32	Fp1\FP2\AF3\AF4\F7\F3\Fz\F4\F8\FC5\FC1\FC2\FC6\T7\C3\Cz\C4\ T8\CP5\CP1\CP2\CP6\P7\P3\PZ\P4\P8\PO3\PO4\O1\Oz\O2	89.55%

and fast needs. According to the identification results of the first 8, 16 and 24 electrodes in the table, most of the effective EEG electrodes are distributed in the frontal and parietal lobes (such as, O2, PO4, AF3, F3, F7, FC5, FC1, and so on). Therefore, the frontal and parietal lobes have a large effect on emotional identification.

Ablation experiments

The generic dimension and volume type of the kernel have 3 × 3, 5 × 5, and 7 × 7, where a plurality of 3 × 3 stacked approximately a 5 × 5 or 7 × 7. Because the activation function is set after the convolutional layer, Krizhevsky et al. (2012) believes that the recognition capability of the model can be controlled by the volume of kernel. Multiscale small kernel subscriptions have diversely increased the network capacity so that the decision function is more distinguished for different categories.

We assume that the size of the 3D convolutional kernel is M × N × K. When using a 3D kernel, it can be divided into three steps: First, complete the convolution of the M × 1 × 1 content; second, complete the convolution of the 1 × N × 1 content; finally, complete the convolution of the 1 × 1 × K content. The total convolution process can increase the non-linear expression of the model because the local convolution of each small step is completed and pass through the non-linear function.

TABLE 8 Comparison of different kernels.

Kernel Size	DEAP (%)	SEED-IV (%)
Conv3 × 3*	95.67	89.88
Conv3 × 3	92.88	85.13

*represents the multiscale kernel.

We can clearly see that the entire process has three non-linear transformations. Therefore, the non-linear characteristics of the results eventually increase, making the decision function more decisive and helping the model increase the accuracy of emotion recognition. The size of conv3 × 3 (the size of convolutional kernel is 3 × 3) compared to conv5 × 5 and conv7 × 7 significantly reduces the number of parameters. Simonyan and Zisserman (2014) replaces a conv7 × 7 with three conv3 × 3, which is considered to further decompose the characteristics mentioned by the 7 × 7 larger volume kernels. The Regularization of the multiscale small kernel can improve the model performance.

In addition to the small conv3 × 3, there is the small conv2 × 2. However, Zeiler and Fergus (2014a) studies conv2 × 2 and is unable to find the central point of the convolution, which causes the characteristics of the padding process to constantly offset. As the number of layers deepens, conv2 × 2 makes the distance of feature offset increasingly obvious. Thus, this paper expects to apply the multiscale convolutional kernel to increase the feature amount of the model calculation and improve the model recognition.

According to the above research results, this paper takes the types of convolutional kernels of conv3 × 3. The same dataset is carried out by different types of kernels and each classification result is compared. Conv3 × 3* represents the multiscale kernel, and conv3 × 3 represents the same-scale kernel. **Table 8** shows that the multiscale small kernel subscriptions diversely increase the network capacity so that the decision function is more distinguished for different categories.

Conclusion

In this paper, we have presented the ERN model which uses the multiscale 3D-CNN to recognize emotions based

on EEG. We obtain the optimal parameters through random training data and design experiments to compare the promotion classification performance at different time windows. Then, based on the 1 s time window dataset with better classification performance, an effective multiscale convolutional kernel 3D-CNN model based on EEG signals is implemented to simultaneously extract spatial and temporal features, and achieves higher accuracy of emotion recognition.

In the comparative analysis using the EEG signals in the DEAP and SEED-IV datasets, we have demonstrated the superior accuracy, F_1, and AUC values of emotion recognition for the ERN model based on multiscale 3D-CNN. From the experimental results, we show that this model can achieve higher performance, which helps to efficiently recognize the emotional state of the subjects so that BCI technology can quickly and accurately convert the neural electrical signal into commands that can be identified by the computer, greatly improving human-machine interaction.

The limitations of the model include the exploratory interpretability of the convolutional model. The calculation process in the convolution network is similar to a black box, and it is especially difficult to understand how the method works on feature learning. If the feature representation learned by the hidden layers can be visualized, it will be more conducive to the optimization of the convolutional network.

While we have achieved superior accuracy when compared to alternative methods as discussed in this paper using EEG data, we consider that there are further potential improvements. Our projected future directions for research include addressing subject-independent emotion recognition (the model can be trained using data acquired from a limited number of participants and can be applied to a subject who has never experienced the system prior to the experiment.) used our model to conduct the fusion of multimodal signals such as EEG and EMG studies and the investigation of other methods to improve our model with respect to the recognition accuracy. In addition, we can also use the model to challenge other tasks, such as emotion recognition based on multimodal physiological signal fusion, which could facilitate the performance of real-time emotion recognition and enhance emotional experience in the field of BCI.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://www.eecs.qmul.ac.uk/mmv/datasets/deap/ and https://bcmi.sjtu.edu.cn/~seed/index.html.

Author contributions

YS and ZZ designed this project, carried out most of the experiments and data analysis, and revised the manuscript. All authors analyzed the results and presented the discussion and conclusion, contributed to the article and approved the submitted version.

Funding

This research was supported in part by the National Natural Science Foundation of China (NSFC) (Grant nos. 61862058 and 61962034), the Gansu Provincial Science and Technology Department (Grant no. 20JR10RA076), and the Cultivation plan of major Scientific Research Projects of Northwest Normal University (NWNU-LKZD2021-06).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Acharya, D., Goel, S., Bhardwaj, H., Sakalle, A., and Bhardwaj, A. (2020). "A long short term memory deep learning network for the classification of negative emotions using EEG Signals," in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK. doi: 10.1109/IJCNN48605.2020.9207280

An, Y., Hu, S., Duan, X., Zhao, L., Xie, C., and Zhao, Y. (2021). Electroencephalogram emotion recognition based on 3D feature fusion and convolutional autoencoder. *Front. Comput. Neurosci.* 15: 743426.

BCMI (1994). *Center for Brain-like Computing and Machine Intelligence(BCMI) laboratory, China*. Shanghai: Shanghai Jiao Tong University.

Cai, H., Zhang, X., Zhang, Y., Wang, Z., and Hu, B. (2018). A case-based reasoning model for depression based on three-electrode EEG data. *IEEE Trans. Affect. Comput.* 11, 383–392. doi: 10.1109/TAFFC.2018.2801289

Candra, H., Yuwono, M., Chai, R., Handojoseno, A., and Su, S. (2015). Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2015, 7250–7253. doi: 10.1109/EMBC.2015.7320065

Hosseini, M., Powell, M., Collins, J., Mahan, H., Michael, P., John, C., et al. (2020). I tried a bunch of things: The dangers of unexpectedoverfitting in classification of brain data. *Neurosci. Biobehav. Rev.* 119, 456–467. doi: 10.1016/j.neubiorev.2020.09.036

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 327–339. doi: 10.1109/TAFFC.2019.2901673

Joseph, R. (1990). *The parietal lobes*. Boston, MA: Springer.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). *ImageNet classification with deep convolutional neural networks*. Red Hook; NY: NIPS. Curran Associates Inc.

Kwon, Y. H., Shin, S. B., and Kim, S. D. (2018). Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors* 18:1383. doi: 10.3390/s18051383

Li, R., Jared, J. S., Ahmed, H., IIyevsky, T. V., Wilbur, R. B., and Bharadwaj, H. M. (2021). The perils and pitfalls of block design for EEG classification experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* 43:316–333. doi: 10.1109/TPAMI.2020.2973153

Li, S., Lyu, X., Zhao, L., Chen, Z., Gong, A., and Fu, Y. (2021). Identification of emotion using electroencephalogram by tunable Q-factor wavelet transform and binary gray wolf optimization. *Front. Comput. Neurosci.* 15:732763. doi: 10.3389/fncom.2021.732763

Li, Y., Lei, M., Zhang, X., Cui, W., Guo, Y., Huang, T. W., et al. (2018). Boosted convolutional neural networks for motor imagery EEG decoding with multiwavelet-based time-frequency conditional granger causality analysis. *arXiv* [Preprint]. doi: 10.48550/arXiv.1810.10353

Lin, W. Q., Li, C., and Sun, S. Q. (2017). Deep convolutional neural network for emotion recognition using EEG and peripheral physiological signal. *Lecture Notes Comp. Sci.* 10667, 385–394. doi: 10.1007/978-3-319-71589-6_33

Liu, N. J., Fang, Y. C., Li, L., Hou, L. M., Yang, F. L., and Guo, Y. K. (2018). "Multiple feature fusion for automatic emotion recognition using EEG signals," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Piscataway, NJ. doi: 10.1109/ICASSP.2018.8462518

Liu, Y., Ding, Y., Li, C., Cheng, J., Song, R., Wan, F., et al. (2020). Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network. *Comput. Biol. Med.* 123:103927. doi: 10.1016/J.COMPBIOMED.2020.103927

Mei, M., and Xu, X. (2017). "EEG-Based Emotion Classification Using Convolutional Neural Networks," in *Proceedings of the International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Chengdu.

Qiu, J. L., Li, X. Y., and Hu, K. (2018). "Correlated Attention Networks for Multimodal Emotion Recognition," in *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Piscataway, NJ.

Salama, E., El-Khoribi, R., Shoman, M., and Wahby, S. M. (2018). EEG-based emotion recognition using 3D convolutional neural networks. *Int. J. Adv. Comput. Sci. Appl.* 9, 329–339. doi: 10.14569/IJACSA.2018.090843

Scherer, D., Andreas, M., and Behnke, S. (2010). "Evaluation of pooling operations in convolutional architectures for object recognition," in *Artificial Neural Networks – ICANN 2010. ICANN 2010. Lecture Notes in Computer Science*, eds K. Diamantaras, W. Duch, and L. S. Iliadis, (Berlin: Springer), 92–101. doi: 10.1007/978-3-642-15825-4_10

Sharbrough, F., Chatrian, G. E., Lüders, H., Nuwer, M., Picton, T. W., et al. (1991). American electroencephalographic society guidelines for standard electrode position nomenclature. *J. Clin. Neurophysiol.* 8, 200–202.

Simon, O., Mangin, J. F., Cohen, L., Le Bihan, D., and Dehaene, S. (2002). Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. *Neuron* 33, 475–487. doi: 10.1016/S0896-6273(02)00575-5

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint]. doi: 10.48550/arXiv.1409.1556

Song, T., Zheng, W., Song, P., and Cui, Z. (2020). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Chaudhary, A., Chouhan, K. S., Gajrani, J., and Sharma, B. (2020). "Deep learning with PyTorch," in *Machine learning and deep learning in real-time applications* (Hershey, PA: IGI Global), 61–95.

Sun, Y., Ayaz, H., and Akansu, A. N. (2020). Multimodal affective state assessment using fnirs + eeg and spontaneous facial expression. *Brain Sci.* 10:85. doi: 10.3390/brainsci10020085

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Chengdu. doi: 10.1109/CVPR.2016.308

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., et al. (2020). "EEG-based Emotion Recognition via Channelwise Attention and Self Attention," in *Proceedings of the IEEE Transactions on Affective Computing*, Chengdu. doi: 10.1109/TAFFC.2020.3025777

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Chengdu. doi: 10.1109/ICCV.2015.510

Wang, Y., Huang, Z., Mccane, B., and Neo, P. (2018). "EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition," in *Proceedings of the International Joint Conference on Neural Networks*, Chengdu. doi: 10.1109/IJCNN.2018.8489715

Xie, Y., Liang, R., Liang, Z., Huang, C., and Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE ACM Trans. Audio Speech Lang. Proc.* 27, 1675–1685. doi: 10.1109/TASLP.2019.2925934

Xing, X., Li, Z., Xu, T., Shu, L., and Xu, X. (2019). SAE+LSTM: a new framework for emotion recognition from multi-channel EEG. *Front. Neurorobot.* 13: 37. doi: 10.3389/fnbot.2019.00037

Yang, Y., Wu, Q., Ming, Q., Wang, Y., and Chen, X. (2018). "Emotion Recognition from Multi-ChannelEEG through Parallel Convolutional Recurrent Neural Network," in *Proceedings of the International Joint Conference on Neural Networks*, Piscataway, NJ. doi: 10.1109/IJCNN.2018.8489331

Yao, Q. (2014). *Multi-Sensory Emotion Recognition with Speech and Facial Expression*. Ph.D. thesis. Mississippi: University of Southern Mississippi.

Yea-Hoon, K., Sae-Byuk, S., and Shin-Dug, K. (2018). Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors* 18:1383. doi: 10.3390/s18051383

Yin, Y. Q., Zheng, X. W., Hu, B., Zhang, Y., and Cui, X. C. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Appl. Soft Comput.* 100:106954. doi: 10.1016/j.asoc.2020.106954

Yoon, H. J., and Chuang, S. Y. (2013). EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm. *Comp. Biol. Med.* 43, 2230–2237. doi: 10.1016/j.compbiomed.2013.10.017

Yu, S., Yoshimoto, J., Shigeru, T., Masahiro, T., Shinpei, Y., Yasumasa, O., et al. (2015). Nested 10-fold cross-validation. *PLoS One* 10:e0123524. doi: 10.1371/journal.pone.0123524.g002

Zangeneh, S. M., Maghooli, K., Setarehdan, S. K., and Motie, N. A. (2019). A novel eeg-based approach to classify emotions through phase space dynamics. *Signal Image Video Process.* 13, 1149–1156. doi: 10.1007/s11760-019-01455-y

Zeiler, M. D., and Fergus, R. (2014b). "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014. Lecture Notes in Computer Science*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Zurich: ECCV). doi: 10.1007/978-3-319-10590-1_53

Zeiler, M. D., and Fergus, R. (2014a). "Architecture of Convolutional Neural Networks (CNNs) demystified," in *Proceedings of the 12th European conference on computer vision*, Chapel Hill, NC.

Zhang, M., Bont, C. D., and Li, W. (2015). "Emotional Engagement for Human-Computer Interaction in Exhibition Design," in *Human-Computer Interaction: Design and Evaluation(HCI). Lecture Notes in Computer Science*, eds M. Kurosu (Berlin: Springer), doi: 10.1007/978-3-319-20901-2_51

Zhao, Y., Yang, J., Lin, J., Yu, D., and Cao, X. (2020). "A 3D Convolutional Neural Network for Emotion Recognition based on EEG Signals," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Piscataway, NJ. doi: 10.1109/IJCNN48605.2020.9207420

Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2019). EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Trans. Cybernet.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-Based emotion recognition with deep neural networks. *IEEE Trans. Autonom. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhong, Q., Zhu, Y., Cai, D., Xiao, L., and Zhang, H. (2020). Electroencephalogram access for emotion recognition based on a deep hybrid network. *Front. Hum. Neurosci.* 14:589001. doi: 10.3389/fnhum.2020.589001