A Multi-Branch Transformer Model Integrating Temporal Spectral and Spatial Features for Four-Class EEG Signal Classification

Xin-Tong Ye, Tian-Wei Shi*

Abstract—**This paper proposes a multi-branch Transformer model that integrates time-domain, frequency-domain, and spatial-domain feature extraction for the four-class EEG classification task. To fully exploit the multi-dimensional information of EEG signals, we design a parallel multi-branch architecture. In this architecture, time-domain features are extracted through a combination of convolutional neural networks (CNNs) and Transformer encoders; frequency-domain features are obtained via short-time Fourier transform (STFT) and further optimized by convolutional layers; spatial-domain features are processed using depthwise separable convolution and spatial attention mechanisms, dynamically adjusting the weights of spatial features to enhance the model's feature extraction capability. To further improve model performance, we introduce a dynamic weight adjustment mechanism, enabling the model to automatically learn and optimize the importance of time-domain, frequency-domain, and spatial-domain features and adjust the contribution of each feature branch according to different task requirements, significantly enhancing the model's accuracy and robustness. Additionally, we employ data augmentation and cross-validation strategies to enhance the model's generalization ability. Experimental results show that the proposed model achieves significant performance improvements on the BCI Competition 2a public dataset. During the cross-validation process, the model's highest accuracy and average accuracy on the validation set reached 93.36% and 92.79%, respectively, demonstrating its superior classification ability and strong generalization. The innovation of this paper lies in: by integrating time-domain, frequency-domain, and spatial-domain information, a novel multi-modal feature processing framework is proposed, and by combining spatial attention mechanisms and dynamic weight adjustment mechanisms, the accuracy and reliability of EEG classification are effectively improved.**

Index Terms—**Convolutional Neural Network, Short-Time Fourier Transform, Transformer, Brain Computer Interface，Motor Imagery**

Manuscript received March 1, 2025; revised May 31, 2025.

This work was supported by the Liaoning Province Science and Technology Major Project (2022JH1/10400005), Liaoning Province Department of Education University basic research funds, the Liaoning Province Intelligent Construction Iot Application Technology Key Laboratory Open Project (2024KFKT-08), and the National Natural Science Foundation of China (U1908218).

XinTong Ye is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (phone:+86-176-4124-2463, e-mail: 1925575202@qq.com).

TianWei Shi* is an Associate Professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (Corresponding author to provide phone: +86-139-9805-3962; e-mail: y17641242463@163.com.)

I. Introduction

BRAIN Computer Interface (BCI) technology, by directly decoding electroencephalogram (EEG) signals, has established an efficient communication channel between the human brain and external devices, demonstrating broad application prospects in intelligent healthcare, rehabilitation assistance, and virtual reality fields [1]. Among them, motor imagery, as an autonomous control paradigm without external stimulation, generates recognizable EEG signals through the brain's imagination of specific limb movements, providing a natural control method for BCI systems [2]. However, motor imagery EEG signals generally have low signal-to-noise ratios and significant inter-individual differences, which directly affect the accuracy of feature extraction and classification, becoming one of the core bottlenecks restricting the development of BCI technology [3].

Traditional EEG signal feature extraction methods, such as Wavelet Transform (WT) and Common Spatial Pattern (CSP), mainly rely on manual feature extraction and shallow classifiers. Wavelet Transform extracts local features through time-frequency analysis and is suitable for the analysis of time-frequency characteristics, but its limitations are significant. The manual design of feature extraction processes is vulnerable to noise and individual differences, and classification based on wavelets or CSP typically employs simple classifiers like Support Vector Machine (SVM), which struggle to effectively capture complex nonlinear relationships. For instance, Yang et al. used CSP for feature extraction of EEG signals and SVM as the classifier, achieving an accuracy rate of 78.7% [4]. Due to the frequency band selection issue of CSP, He et al. explored the limitations of traditional feature extraction methods (such as CSP), especially their robustness in the presence of individual differences and noise [5]. With the introduction of deep learning techniques, traditional methods have gradually been replaced by automated and adaptive feature extraction and classification methods. Feature extraction is a crucial step in EEG signal processing and directly affects classification performance. Traditional methods like CSP maximize inter-class differences through spatial projection, but they have poor robustness to individual differences and noise, particularly in complex multi-channel backgrounds [6]. In contrast, deep learning, especially Convolutional Neural Networks (CNN), can automatically learn multi-level and multi-scale features, avoiding the limitations of manual feature selection. Mahamune et al. proposed a

model combining CSP and CNN, achieving a classification accuracy of 75.03% on the BCI Competition IV 2a dataset, an improvement of over 10% compared to traditional methods [7]. Gao et al. proposed a model combining Gated Recurrent Unit (GRU) and CNN, achieving a classification accuracy of 80.7% on the BCI Competition IV 2a dataset and enhancing the robustness of deep learning in decoding small-scale EEG datasets [8].

Traditional EEG signal classification methods often rely on traditional classifiers such as support vector machines (SVM), which perform well when dealing with linear and low-dimensional data. Shi et al. proposed a pattern recognition method that used the improved squirrel search algorithm (ISSA) to optimize support vector machine (SVM), extracted the time domain features of EEG signals and directed them to SVM as feature vectors for classification and recognition, and achieved good results [9]. However, with the increase in data dimension and complexity, traditional classification methods are faced with overfitting, difficulty in feature selection, and low computational efficiency. In this context, deep learning methods have gradually become an important tool in the field of EEG signal classification.

In recent years, CNN has become a mainstream method for EEG signal processing due to its powerful spatial feature extraction capability and has been widely used in EEG signal classification [10]. By stacking convolution layers and pooling layers, CNN can effectively and automatically extract features and perform end-to-end classification [11]. For example, Craley et al. adopted multichannel CNN and short and long-term memory (LSTM) methods to further improve the robustness and accuracy of EEG signal classification [12]. However, it is difficult for CNN to effectively capture the global time dependence when processing long time series data [13]. Tan et al. found that CNNS are not good at modeling long-distance dependencies and obtaining global context information [14].

To overcome this shortcoming, the Transformer model has been developed. The Transformer model is a deep learning architecture based on a self-attention mechanism to capture long-term dependencies in EEG signals. With the self-attention mechanism, the Transformer can dynamically adjust the weights at different time points according to the global context of the signal, thus improving the accuracy of classification. The EEG classification method based on the Transformer proposed by He et al. shows its advantages in processing long-time series data [15].

In addition, CNN and Transformer are emerging forces in EEG signal classification with their powerful automatic learning and characterization capabilities. By combining the model of CNN and Transformer, Li et al. achieved better performance in processing EEG signals, overcoming the deficiency of medium and long-distance dependence of CNN and the deficiency of local features in the Transformer [16]. Lu et al. proposed integrating CNN and Transformer into a single framework model in parallel, which can better interact and integrate local and global features [17].

Traditional methods such as CSP are prone to the problem of decreasing classification accuracy under the condition of large noise and individual differences. In contrast, deep learning-based methods are more robust and can automatically adjust feature extraction and classification strategies, thereby improving classification accuracy. Traditional methods generally only focus on spatial features or temporal features, while deep learning-based methods can handle both spatial features and temporal features at the same time, further improving the accuracy and generalization ability of classification.

In this paper, a four-classification model based on the fusion of CNN and Transformer is proposed. The spatial features of EEG signals are extracted by CNN；the temporal features are captured by the Transformer's axial attention mechanism. CNN can automatically extract spatial features from EEG signals, avoiding the process of relying on manual feature selection in traditional methods. Meanwhile, Transformer enhances the modeling capability of timing information by modeling timing features with axial attention. The model can not only model the spatial features of the input signal but also capture the timing dependence of the signal through the self-attention mechanism of the Transformer, thus fully integrating the advantages of spatial feature extraction and timing modeling, and improving the accuracy and robustness of feature extraction.

The four-classification model based on the fusion of CNN and Transformer proposed in this paper shows high classification accuracy and robustness in motion imagination tasks. By combining CNN and Transformer, we not only overcome the limitations of traditional methods but also improve the adaptive and classification performance of EEG signal processing. The experimental results show that the model can process EEG signals in complex backgrounds and achieve excellent performance in motion imagination BCI tasks, which have great practical application potential.

II. Method

The aim of this study is to comprehensively analyze EEG signals through a multi-branch deep learning network and to improve the classification accuracy by using the fusion of time-domain, frequency-domain, and spatial-domain features. In the process of EEG signal processing, the preprocessing steps, including bandpass filtering, artifact removal, and standardization, are first carried out to ensure signal quality and remove noise. Then, we use a short-time Fourier transform (STFT) to extract the local frequency features of the signal from the perspective of the frequency domain. In order to further improve the performance of the model, we use a deep neural network architecture with three branches: time domain, frequency domain, and spatial domain, and capture the information of EEG signals in these three dimensions by convolutional neural network (CNN) and depth-separable convolution. The features extracted from each branch are fused by concatenation, and the final classification task is completed by the fully connected layer.

In addition, the study introduces Transformer encoders to enhance the modeling capability of timing signals, using their self-attention mechanism to capture long-term dependencies. In the training process of the model, adaptive learning rate adjustment and cross-validation techniques are combined to avoid overfitting and ensure the generalization ability of the

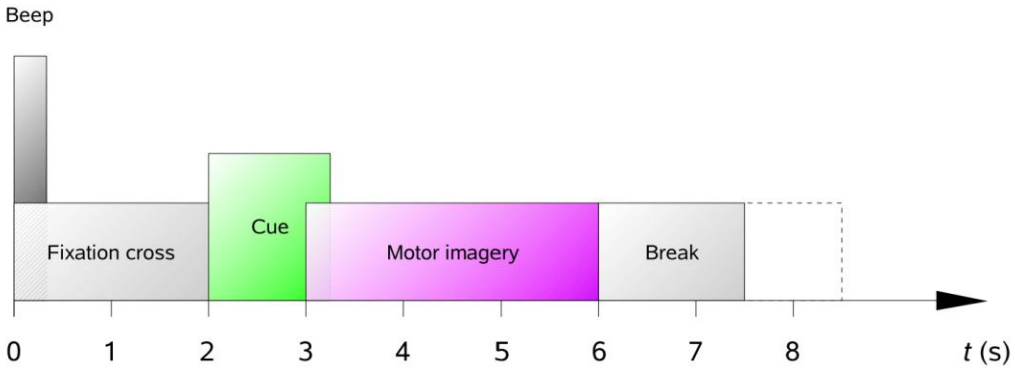

Fig. 1. Timing diagram of a complete test of BCI Competition IV 2a data set

model. Through these innovative methods, this study expects to achieve higher accuracy and stronger robustness in EEG signal classification tasks.

A. Data Acquisition

The BCI Competition IV 2a dataset contains EEG recordings from nine individuals performing four distinct motor imagery tasks: left hand (type 1), right hand (type 2), foot (type 3), and tongue (type 4). The experiment was structured into six sessions, each comprising 48 trials. Participants were seated comfortably in an armchair facing a 21-inch LCD screen. At the start of each trial (t = 0 s), a fixed "+" symbol appeared on the black screen, accompanied by a brief auditory cue. Two seconds later (t = 2 s), an arrow pointing left, right, downward, or upward was displayed for approximately 1.25 s, corresponding to the four motor imagery tasks: left hand, right hand, feet, and tongue. During this phase, subjects imagined executing the indicated movement until the "+" symbol disappeared at t = 6 s. Following this, they rested briefly before the screen turned black again. Each trial lasted about 8 seconds on average. The timing of a full trial is illustrated in Figure 1.

B. Data Preprocessing

In this study, EEG signals first need to undergo data preprocessing to eliminate noise and artifacts. We use the MNE library to load raw EEG data and perform a series of processing on it. First, bandpass filtering is applied to the EEG signal with a frequency range of 8 Hz to 30 Hz. This pass-through filter can effectively remove low-frequency noise (such as myoelectric artifacts) and high-frequency noise (such as electrical interference), thereby preserving the main signal components associated with the electrical activity of the brain. Then, after filtering, we normalize the EEG signal to ensure that the signal amplitude is in the same range for each channel, usually using a normalization method of zero mean and unit variance. In addition, we also used artifact removal techniques, especially the interference of EEG signals by electrical ocular artifacts (EOG). By labeling and excluding the electrodes containing the ophthalmic electrical signal (e.g., EOG-left, EOG-central, EOG-right), we can significantly improve the quality of the signal and reduce noise caused by non-brain-derived activity.

After data preprocessing is completed, EEG signals are divided according to the event markers in the experimental design to generate signal segments with multiple time windows. In order to enhance the robustness of the model, we also use data enhancement techniques to add Gaussian noise to the signal to simulate different electrical disturbances and limit the enhanced signal to the amplitude range of [-1, 1] through signal clipping. The enhanced data is used together with the original data for model training, thus increasing the diversity of training data and helping to improve the generalization ability of the model. The EEG preprocessing process is shown in Figure 2.

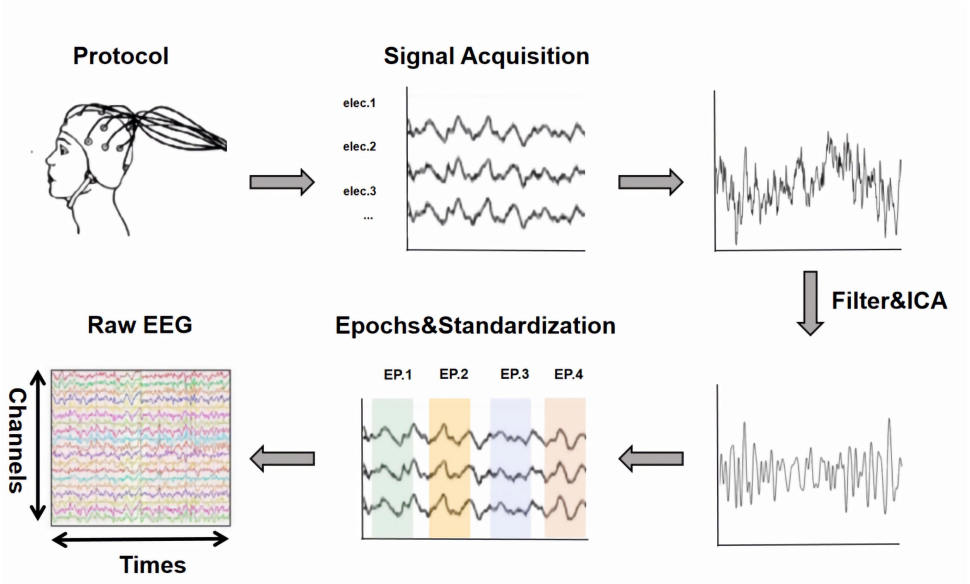

Fig. 2. Flow chart of EEG preprocessing

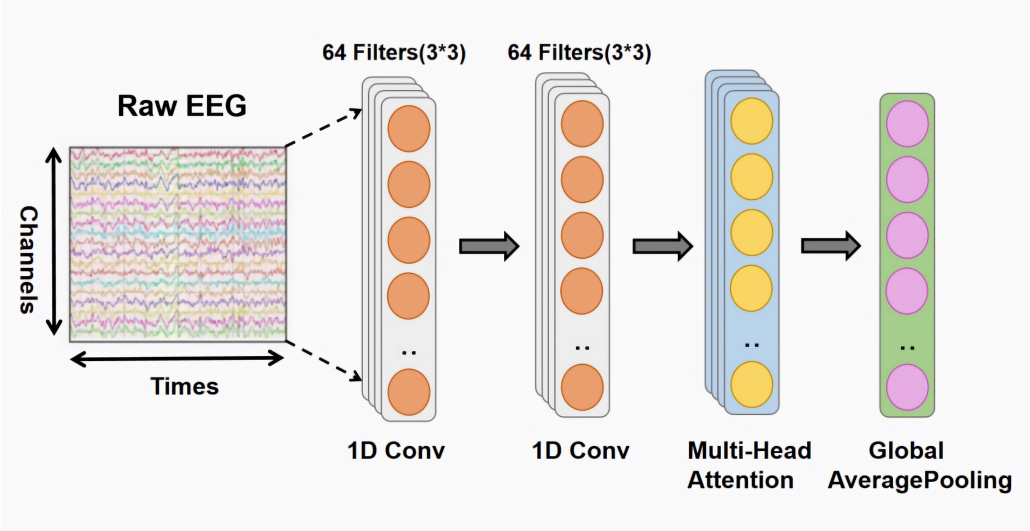

Fig. 3. Time domain feature extraction diagram

C. Feature Extraction

In order to further improve the accuracy of EEG signal classification, a multi-branch deep learning network structure combining time-domain, frequency-domain, spatial-domain features is proposed. In this network, the time-domain branch, the frequency-domain branch and the spatial-domain branch, respectively, process different features from the EEG signal and fuse them to form the final classification decision. This method makes full use of the information of EEG signals in different feature Spaces to enhance the classification ability.

1)Time domain features: First, the time-domain branch uses a one-dimensional convolutional neural network (Conv1D) to process the EEG signals. This branch can capture the local dependence of the signal in the time dimension through the convolution operation so as to identify the short-term change pattern in the EEG signal. After the convolution operation, we also introduce the ReLU activation function to enhance the nonlinear representation capability. In the process of capturing time domain features, we have added a Transformer encoder to the time domain branch in order to handle long time series dependencies. The ability to model EEG signal sequences can be improved by using the self-attention mechanism to effectively capture the long-range dependent information of the signal. Time domain feature extraction is shown in Figure 3.

First, the local time features of EEG signals are extracted by a one-dimensional convolution layer. Through sliding window operation, Conv1D is able to learn short-term dependencies in time dimension to extract local patterns in EEG signals [18]. Given input EEG segment$X \in RC \times T$ (where C represents the number of channels and T represents the time step), one-dimensional convolution is calculated as follows:

$$h_i^{(l)} = f(\sum_{j=0}^{k-1} \varpi_j^{(l)} x_{i+j} + b^{(l)}) \qquad (1)$$

where: $h_i^{(l)}$ is the feature after the layer convolution, K is the size of the convolution kernel, $\varpi_j^{(l)}$ is the weight of the convolution kernel and $b^{(l)}$ is the bias term, $f(\bullet)$ is the activation function (ReLU). Through the convolutional layer, the model extracts the features of the short-term dependence relationship.

To capture long-term time dependencies, introduce a Transformer Encoder. This module models long-term dependence through multi-head self-attention mechanisms:

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

Among them, Q, K, and V are the query, key, and value matrix, respectively, and d_k is the dimension of the key. Different time relationships are learned in parallel by multiple attention heads, and the results are finally concatenated.

2)Frequency domain features: To extract useful frequency-domain features from EEG signals, we use the short-time Fourier transform (STFT). STFT is an analytical method that can provide time and frequency localization information and is especially suitable for analyzing non-stationary signals such as EEG[19]. With STFT, we are able to convert the EEG signal into a spectrogram to reveal the frequency composition of the signal and its properties over time.

In this study, we applied STFT to the time series signals of each EEG channel, using a sliding window of 32 sampling points for the Fourier transform. The selection of this window size is based on the consideration of the spectral characteristics of the signal and the experimental requirements. After STFT calculation, the signal is converted into a two-dimensional matrix, where each row represents a frequency component within a time window, and the columns correspond to a different frequency range. This spectrum map provides a rich set of frequency-domain features for subsequent deep-learning models, helping them capture potential frequency patterns in the signal. Frequency domain feature extraction is shown in Figure 4.

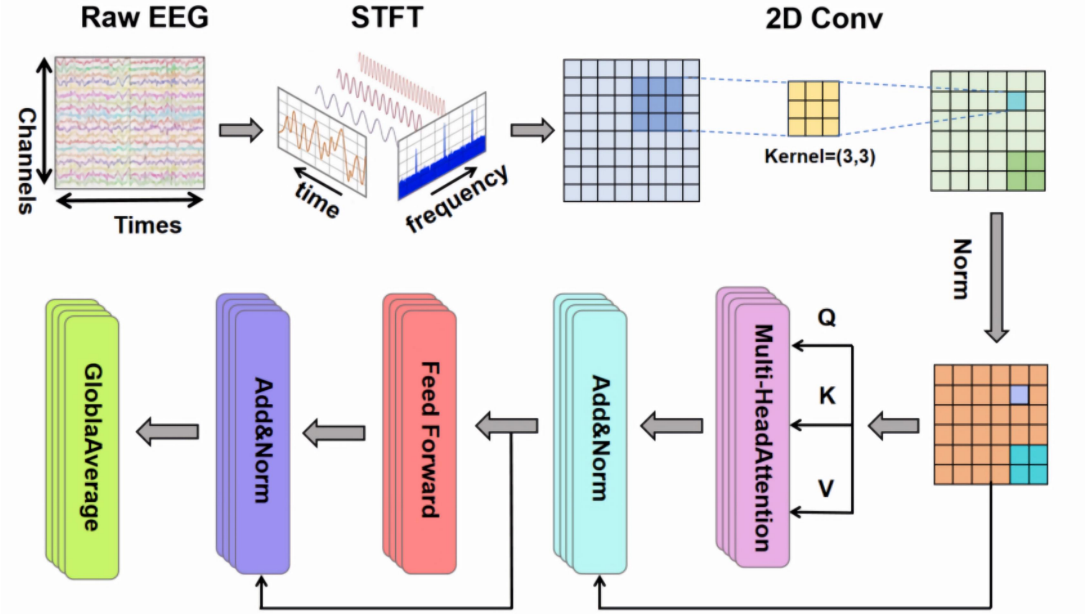

Fig. 4. Frequency domain feature extraction diagram

First, the original time domain signal is converted to the frequency domain by a short-time Fourier transform (STFT). The goal of STFT is to analyze the frequency component of a signal over time, as follows:

$$h_{i,j}^{(l)} = f(\sum_{m=0}^{k-1}\sum_{n=0}^{k-1} \varpi_{m,n}^{(l)} x_{i+m,j+n} + b^{(l)}) \quad (4)$$

where: $h_{i,j}^{(l)}$ is the feature after the l layer convolution, K is the size of the convolution kernel, $\varpi_{m,n}^{(l)}$ is the weight of the convolution kernel, $b^{(l)}$ is the bias term, $f(\cdot)$ is the activation function (ReLU). Through the convolutional layer, the model extracts the features of the short-term dependence relationship.

In order to capture long-term time dependencies, a Transformer Encoder must be introduced.

3)Frequency domain features: The branch is processed based on the spectral diagram obtained by the short-time Fourier transform. We apply two-dimensional convolutional neural networks (Conv2D) to spectral graphs to extract features from both spatial and frequency dimensions. The convolution operation can not only extract the frequency components of the signal but also capture the changes in the spectrum between different times Windows. Through layer-by-layer convolution and pooling operation, the model gradually extracts more complex features in the frequency domain, which helps to identify frequency patterns in EEG signals.

Spatial branches are processed by Depthwise Separable Convolution. Deep separable convolution can reduce computational complexity while maintaining strong feature extraction capability. This branch processes the spatial information of EEG signals to capture the contribution of different spatial regions to the signals. In addition, we introduce a spatial attention mechanism to further enhance the model's focus on key spatial regions, thereby improving the model's sensitivity to spatial features in the signal. Spatial feature extraction is shown in Figure 5.

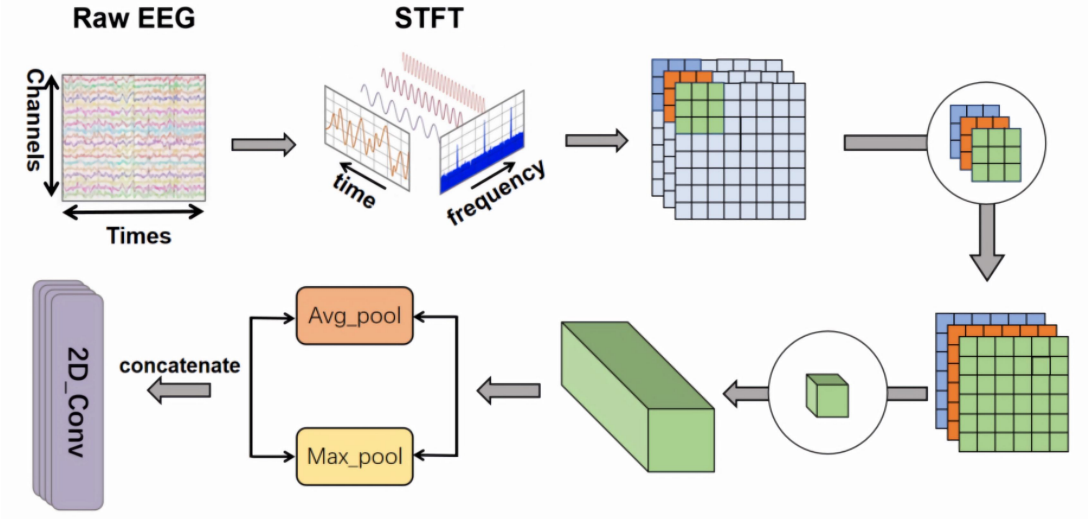

Fig. 5. Spatial feature extraction map

By dividing standard convolution into two steps, the depth separable convolution reduces the computational complexity of the model and can extract spatial features effectively. The first is deep convolution, where each input channel (electrode) has an independent convolution kernel. Then, a point-by-point convolution is performed, combining the outputs of each convolutional channel. The formula for depth-separable convolution is as follows:

$$Y_{i,j} = \sum_{m,n} x_{i+m,j+n} * \varpi_{m,n}^{(d)} \tag{6}$$

$$z_{i,j} = \sum_{m,n} x_{i+m,j+n} * \varpi_{m,n}^{(p)} \tag{7}$$

where: x is the input signal (such as EEG signal), y is the feature after deep convolution, $\varpi^{(\mathrm{d})}$ and $\varpi^{(p)}$ is the convolution kernel of deep convolution and point-by-point convolution, z is the final convolution feature. This method not only ensures feature extraction but also reduces the amount of computation, which is especially suitable for multi-electrode data processing.

The spatial attention mechanism enhances the model's focus on key spatial regions by calculating the importance of each location and giving greater weight to important locations. Common spatial attention mechanisms include extracting spatial features using Avg Pooling and Max Pooling, then combining the two to generate a spatial attention map, and finally weighing the original features. Spatial attention is calculated as follows:

$$A = \sigma(Conv(Concatenate(AvgPool(X), MaxPool(X)))) \tag{8}$$

where, X is the input feature (the output from the convolution layer), $AvgPool(X)$ and $MaxPool(X)$ is the average and maximum pooling operation, respectively, A is the spatial attention diagram generated by the convolution operation, used to weight the input feature.

Through the above process, spatial convolution and spatial attention mechanisms extract the spatial features of EEG signals. Finally, we carry out Global Average Pooling on the spatial features to aggregate the features of each spatial location into a single value, which is used as the output feature of the spatial branch. The global average pooling formula is as follows:

$$F_{spatial} = \frac{1}{H \times M} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j} \tag{9}$$

where H and M are the height and width of the spatial feature map, respectively, $X_{i,j}$ is the eigenvalue of the spatial position (i, j).

D.Data fusion and classification

After the feature extraction stage is completed, we fuse the output features of the time domain, frequency domain , spatial domain branches. Specifically, the outputs of the time, frequency, and spatial branches are spliced into a high-dimensional feature vector and fed into a fully connected layer for final classification. Before fusing features, we introduced a dynamic weight learning module, which is able to automatically learn the importance of different branches and dynamically adjust the weights based on each branch's contribution to the final classification result. The dynamic weights are calculated by an additional fully connected layer, and the output features of the time domain, frequency domain, and airspace branches are weighted to form a weighted feature vector.

The fully connected layer maps these weighted features to the classification space through the multi-layer neural network and finally outputs the category label. In order to avoid overfitting and ensure the stability of the model on different data sets, we use a 5-fold cross-validation. Each break is trained using a different training set and a different verification set. In addition, an adaptive learning rate adjustment mechanism (such as ReduceLROnPlateau) is adopted in the training process, which automatically adjusts the learning rate according to the change in the performance of the verification set to accelerate convergence and avoid falling into the local optimal solution. To enhance the generalization ability of the model, we used Dropout layers during training to randomly discard the output of some neurons, reducing the risk of overfitting.

In the time domain branch, we introduce a Transformer encoder to improve the modelling ability of the model for time series data. Through its self-attention mechanism, Transformer can capture long-term dependencies in sequence data, which is particularly important for long-term changes in EEG signals [20]. The Transformer encoder in each branch uses a multi-head self-attention mechanism to learn important features in the signal in multiple subspaces, enhancing the model's timing modeling capabilities. Feature fusion is shown in Figure 6.

First, the features of the time domain, frequency domain , and airspace branch are Concatenated to form a higher-dimensional feature vector. Assuming that the output features of the time domain, frequency domain, airspace branches are f_{time} 、 f_{freq} and $\mathrm{f}_{spatial}$ and respectively, then the fused features can be expressed as:

$$\mathrm{f}_{combined} = [\mathrm{f}_{time}, \mathrm{f}_{freq}, \mathrm{f}_{spatial}] \tag{10}$$

where, f_{time} is the output of the time domain branch, f_{freq} is the output of the frequency domain branch, $\mathrm{f}_{spatial}$ is the output of the spatial branch. The linked feature $\mathrm{f}_{combined}$ will contain the feature information for all three branches.

In order for the model to automatically learn weights based on the importance of each branch, we introduce a dynamic weight learning module that adjusts the contribution of each branch by calculating the weight of each branch. The weights output by this module are normalized by the softmax activation function, ensuring that each weight is in the range [0,1] and that the ownership weight adds up to 1. The formula for weight learning is as follows:

$$\mathrm{W}_{combined} = \mathrm{soft\,max}(\mathrm{W}_{fusion} \cdot \mathrm{f}_{combined}) \tag{11}$$

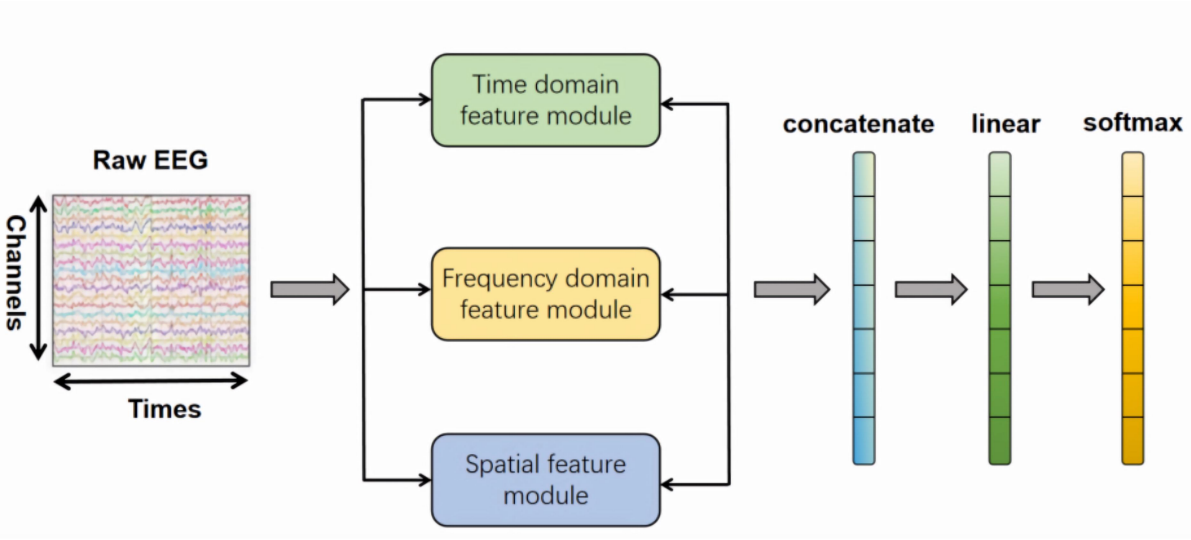

Fig. 6. Spatial feature extraction map

where, W_{fusion} is the learnable weight matrix used to calculate the weights, the shape is $[d_{combined},3]$, where $d_{combined}$ is the dimension of the fused features. $W_{combined}$ is the weight vector of the calculated three branches, representing the importance of the time domain, frequency domain and spatial features. With the softmax function, the resulting weights can be expressed as:

$$W_{combined} = [\alpha_{time}, \alpha_{freq}, \alpha_{spatial}] \tag{12}$$

where, α_{time}, α_{freq} and $\alpha_{spatial}$ are the weights of the time domain, frequency domain and spatial domain branches, respectively.

Using the resulting weights α_{time}, α_{freq} and $\alpha_{spatial}$ weighted summation of the features of each branch, the final fusion feature representation is obtained:

$$f_{final} = \alpha_{time} \cdot f_{time} + \alpha_{freq} \cdot f_{freq} + \alpha_{spatial} \cdot f_{spatial} \tag{13}$$

where: f_{final} is the final weighted feature representation that will serve as the input to the classifier.

E. Training

The Adam algorithm is used as the optimizer and the cross-entropy loss function is used to measure loss . The optimizer formula is as follows:

$$L = -\sum y \times \log(y\_pred) \tag{14}$$

where L is the value of the loss function, y is the One-Hot encoding of the real label and *y_pred* is the predicted output value.

III. Experiment and Results

A. Experimental Settings

The EEG signal preprocessing and classification models of the experimental environment and MI task were constructed using Python, MNE library, and TensorFlow library, respectively, and ran on a Lenovo Yoga 4s notebook equipped with an AMD 5800H processor and 16GB memory. On the experimental data of a single subject, 200 epochs were carried out in the training process, the batch size was set to 32, and the learning rate was 0.001. The data set is divided into the training set and the test set by the ratio of 8:2, and the training set is divided by the 50-fold cross-validation. The training set is used for model training and parameter tuning, while the test set is only used to evaluate the final performance of the model.

The algorithm proposed in this paper is combined with bidirectional convolutional LSTM(ConvLSTM)[21], compact multi-branch one-dimensional convolutional Neural network (CMO-CNN)[22], fused compact convolutional neural network (CCNN), gated recurrent unit (GRU)[23], convolutional neural network (CNN) and long short-term memory (LSTM) combined with CLRNet[24] for comparison. Accuracy is selected as the evaluation parameter. Here, TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives.

TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

B. Comparative test

The accuracy comparison results between the proposed algorithm and other comparison methods on the data set of BCI Competition IV 2a are shown in Table 1. The highest classification accuracy and the average classification accuracy reached 91.78% and 90.57%, respectively.

TABLE I
ACCURACY (%) COMPARISON RESULTS ON THE BCI COMPETITION IV 2A DATA SET

Subject	ConvLSTM	CMO-CNN	CCNN-GRU	CLRNet	Ours
1	74.56	81.44	87.55	90.45	**92.33**
2	72.59	83.57	85.44	91.56	**93.76**
3	75.21	84.25	88.31	89.54	**93.23**
4	73.66	82.95	84.82	89.33	**93.59**
5	74.25	84.28	87.21	88.74	**92.90**
6	71.89	83.95	85.65	90.58	**91.69**
7	74.81	80.84	87.95	89.12	**92.00**
8	72.56	82.94	86.75	88.66	**93.25**
9	72.47	83.78	85.33	88.25	**93.06**
10	74.06	83.25	86.78	90.77	**92.79**
11	71.28	81.58	85.89	89.65	**91.97**
12	75.19	83.55	86.57	90.55	**92.89**
AVG	73.54	83.00	86.52	89.77	**92.79**
SD	1.34	1.14	1.10	1.01	**0.66**

As can be seen from Table 1, in the data set of BCI Competition IV 2a, the maximum classification accuracy of the CMO-CNN algorithm is 84.28%, the average classification accuracy is 83.00%, and the sample standard deviation is 1.14. The highest classification accuracy of the CCNN-GRU algorithm is 88.31%, the average classification accuracy is 86.52%, and the sample standard deviation is 1.10. The maximum classification accuracy of the CLRNet algorithm is 91.56%, the average accuracy is 89.77%, and the sample standard deviation is 1.01. The maximum classification accuracy of the proposed algorithm is 93.76%, the average accuracy is 92.79%, and the sample standard deviation is 0.66. Compared with the other three comparison methods, the maximum accuracy of the proposed algorithm on the dataset of BCI Competition IV 2a is increased by 9.48%, 5.45%, and 2.2%, respectively, and the average classification accuracy is increased by 9.79%, 6.27%, and 3.02%, respectively. The sample standard deviation is also significantly better than the other three groups of comparison experiments. Since the algorithm combines the time domain, frequency domain, and spatial domain to form a multi-modal feature processing framework, as well as the CNN-Transformer algorithm with spatial domain attention mechanism and dynamic weight adjustment mechanism, the proposed algorithm has significant advantages in data set classification tasks, and can better capture and learn relevant features in EEG signals of MI tasks. It has better four classification performance and accuracy stability.

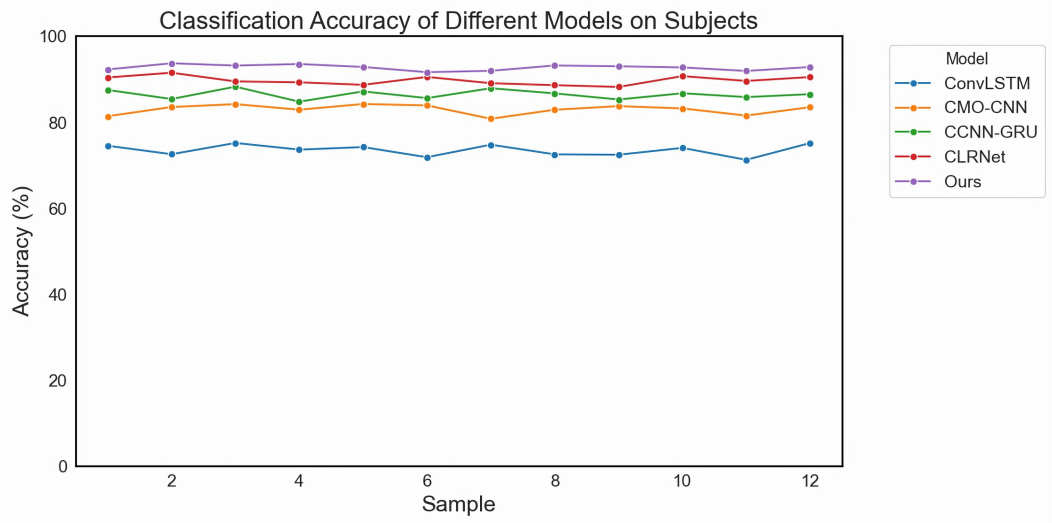

Fig. 7. Accuracy trends of five subjects across 12 rounds on the BCI Competition IV 2a dataset

Figure 7 shows the accuracy trends of ConvLSTM, CMO-CNN, CCNN-GRU, CLRNet, and the algorithms proposed in this chapter running for 12 rounds on the BCI CompetitionIV 2a dataset. It can be observed that the algorithm presented in this chapter exhibits the highest accuracy and the best stability.

C. Ablation experiment

1) *Ablation experiment settings*

In order to explore the influence of each part of the deep learning-based multi-feature fusion Transformer algorithm proposed in this chapter, an ablation experiment is conducted on the dataset of BCI Competition IV 2a. This experiment evaluates the effect of the time domain, frequency domain, space domain, and dynamic weight adjustment mechanism on the algorithm performance by gradually removing different modules in the algorithm. The specific algorithm is as follows:

a. *SSFM(Spectral Spatial Fusion Model)*: Only frequency domain (STFT) and DSC depth can be used to separate the convolution spatial features, and the time-domain feature branches are removed to evaluate the impact of time-domain features on the algorithm.*TSFM(Temporal-Spatial Fusion Model)*: Only time and spatial features are used, and frequency domain (frequency domain features extracted by STFT) is removed so as to evaluate the impact of frequency domain features on algorithm performance.

b. *T-SFM (Temporal-Spectral Fusion Model)*: Only the time domain and spatial domain features are used, and the frequency domain (frequency domain features extracted by STFT) is removed to evaluate the importance of frequency domain features.

c. *SFFM(Full Multi-Domain Model)*: Remove dynamic adjustment weights: Three-domain features (time domain + frequency domain + airspace) are still used, but dynamic weight adjustment is not carried out. Instead, the three-domain features are simply spliced and input into the full connection layer to evaluate the impact of dynamic weight adjustment on algorithm performance.

d. *FMDM(Static Feature Fusion Model)*: contains three domain features (time domain + frequency domain + airspace) and adopts a dynamic weight adjustment mechanism for final classification. This is a complete algorithm proposed in this chapter, which is expected to achieve the best results on classification tasks.

2) *Ablation experiment results*

Table 2 presents the comparison results of the ablation experiments conducted on the BCI Competition IV 2a dataset. The algorithm proposed in this chapter (FMDM) achieves the highest accuracy and the smallest standard deviation.

One-way ANOVA results indicated that there were significant differences among the algorithms ($F = 22.00$, $P < 0.05$). Firstly, under the condition of $\alpha = 0.05$, the P values obtained by each algorithm were all less than 0.05. This proved that there were significant differences among the algorithms, preliminarily demonstrating the important role of each module in the FMDM algorithm. Secondly, T-tests were conducted on each algorithm, and the results showed that there were significant differences between the SSFM and FMDM algorithms ($P = 0.00006$), indicating the significant applicability of the time-domain feature branch in the FMDM algorithm; there were significant differences between the TSFM and FMDM algorithms ($P = 0.0002$),

TABLE II
ACCURACY (%) COMPARSION RESULTS OF THE ABLATIONG EXPERIMENT BASED ON THE BCI COMPETITION IV 2A DATE SET

Subject	SSFM	TSFM	T-SFM	SFFM	FMDM
1	**92.57**	90.66	90.04	91.56	92.33
2	91.98	91.78	89.58	92.39	**93.76**
3	91.54	92.01	90.79	92.44	**93.23**
4	89.33	92.54	91.28	91.08	**93.59**
5	90.28	91.85	90.77	91.84	**92.90**
6	91.64	90.65	88.86	**92.59**	91.69
7	90.85	91.54	89.94	91.33	**92.00**
8	91.28	90.55	88.86	91.78	**93.25**
9	92.37	90.98	89.25	90.28	**93.06**
10	90.55	**92.88**	91.14	92.36	92.79
11	89.64	91.37	89.59	90.05	**91.97**
12	90.67	90.57	90.56	91.56	**92.89**
AVG	91.05	91.45	90.06	91.61	**92.79**
SD	1.02	0.79	0.85	0.82	**0.66**

indicating the strong applicability of the frequency-domain feature branch in the FMDM algorithm; although there was no significant difference between the T-SFM and FMDM algorithms (P = 1.18), the performance of the FMDM algorithm (AVG = 90.79, S.D. = 0.66) was better than that of the T-SFM (AVG = 90.06, S.D. = 0.85). This proved the spatial-domain feature branch's applicability in the FMDM algorithm. There were significant differences between the SFFM and FMDM algorithms (P = 0.0008), indicating the important role of the dynamic weight adjustment mechanism in the FMDM algorithm.

IV. Conclusion

In this paper, we propose a multi-branch Transformer model that integrates feature extraction across the time, frequency, and spatial domains for four-class EEG classification. By designing a parallel multi-branch architecture, the model effectively exploits the multi-dimensional information inherent in EEG signals, thereby enhancing classification performance. Specifically, time-domain features are extracted through a combination of convolutional neural networks (CNN) and Transformer encoders; frequency-domain features are derived via short-time Fourier transform (STFT) and further refined using convolutional layers; spatial features are dynamically adjusted by integrating depth-separable convolutions with a spatial attention mechanism, enhancing the model's feature extraction capability. A dynamic weight adjustment mechanism is introduced to enable the model to automatically learn and optimize the relative importance of time, frequency, and spatial features, allowing adaptive weighting of each feature branch according to task demands. This significantly improves the accuracy and robustness of the model. Furthermore, the generalization ability is strengthened through data augmentation and cross-validation strategies.Experimental results demonstrate that the proposed model achieves substantial performance improvements on the BCI Competition IV 2a dataset, attaining peak and average accuracies of 93.36% and 92.79%, respectively, on the validation set. Compared to several state-of-the-art algorithms, the maximum classification accuracy improves by 9.48%, 5.45%, and 2.2%, while the average accuracy increases by 9.79%, 6.27%, and 3.02%, respectively. Additionally, the model exhibits significantly lower standard deviation in classification performance, confirming its enhanced stability and superiority.

In summary, the CNN-Transformer model presented here, which combines multi-modal feature extraction from time, frequency, and spatial domains with spatial attention and dynamic weighting mechanisms, effectively captures key EEG signal characteristics and substantially improves classification accuracy and robustness in four-class tasks.

References

[1] Y. Chen, F. Wang, T. Li, L. Zhao, A. Gong, W. Nan, et al., "Considerations and discussions on the clear definition and definite scope of brain-computer interfaces," *Frontiers in Neuroscience*, vol. 18, Art. no. 1449208, 2024.

[2] A. Singh, A. A. Hussain, S. Lal, and H. W. Guesgen, "A comprehensive review on critical issues and possible solutions of motor imagery based electroencephalography brain-computer interface," *Sensors*, vol. 21, no. 6, p. 2173, 2021.

[3] Z. Xue, Y. Zhang, H. Li, H. Chen, S. Shen, and H. Du, "Instrumentation, measurement, and signal processing in electroencephalography-based brain–computer interfaces: situations and prospects," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[4] J. Yang, Z. Ma, and T. Shen, "Multi-time and multi-band CSP motor imagery EEG feature classification algorithm," *Applied Sciences*, vol. 11, no. 21, p. 10294, 2021.

[5] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 399–410, 2019.

[6] J. Sun, M. Wei, N. Luo, Z. Li, and H. Wang, "Euler common spatial patterns for EEG classification," *Medical & Biological Engineering & Computing*, vol. 60, no. 3, pp. 753–767, 2022.

[7] R. Mahamune and S. H. Laskar, "An automatic channel selection method based on the standard deviation of wavelet coefficients for motor imagery based brain-computer interfacing," *International Journal of Imaging Systems and Technology*, vol. 33, no. 2, pp. 714–728, 2023.

[8] S. Gao, J. Yang, T. Shen, Z. Xue, Y. Zhang, and H. Li, "A parallel feature fusion network combining GRU and CNN for motor imagery EEG decoding," *Brain Sciences*, vol. 12, no. 9, p. 1233, 2022.

[9] M. Shi, C. Wang, X. Z. Li, M. Q. Li, L. Wang, and N. G. Xie, "EEG signal classification based on SVM with improved squirrel search algorithm," *Biomedical Engineering/Biomedizinische Technik*, vol. 66, no. 2, pp. 137–152, 2021.

[10] S. Rajwal and S. Aggarwal, "Convolutional neural network-based EEG signal analysis: A systematic review," *Archives of Computational Methods in Engineering*, vol. 30, no. 6, pp. 3585–3615, 2023.

[11] I. Rakhmatulin, M. S. Dao, A. Nassibi, and D. Mandic, "Exploring convolutional neural network architectures for EEG feature extraction," *Sensors*, vol. 24, no. 3, p. 877, 2024.

[12] J. Craley, E. Johnson, C. Jouny, and A. Venkataraman, "Automated inter-patient seizure detection using multichannel convolutional and recurrent neural networks," Biomedical Signal Processing and Control, vol. 64, p. 102360, 2021.

[13] Ö. Türk, "Classification of electroencephalogram records related to cursor movements with a hybrid method based on deep learning," International Journal of Imaging Systems and Technology, vol. 31, no. 4, pp. 2322–2333, 2021.

[14] X. Tan, K. Gao, B. Liu, Y. Fu, and L. Kang, "Deep global-local transformer network combined with extended morphological profiles for hyperspectral image classification," *Journal of Applied Remote Sensing*, vol. 15, no. 3, p. 038509, 2021.

[15] Y. He, X. Wang, Z. Yang, L. Xue, Y. Chen, J. Ji, et al., "Classification of attention deficit/hyperactivity disorder based on EEG signals using a EEG-Transformer model," *Journal of Neural Engineering*, vol. 20, no. 5, p. 056013, 2023.

[16] C. Li, X. Huang, R. Song, R. Qian, X. Liu, and X. Chen, "EEG-based seizure prediction via Transformer guided CNN," *Measurement*, vol. 203, p. 111948, 2022.

[17] W. Lu, L. Xia, T. P. Tan, and H. Ma, "CIT-EmotionNet: convolution interactive transformer network for EEG emotion recognition," *PeerJ Computer Science*, vol. 10, p. e2610, 2024.

[18] C. Huang, Y. Xiao, and G. Xu, "Predicting human intention-behavior through EEG signal analysis using multi-scale CNN," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 5, pp. 1722–1729, 2020.

[19] S. M. Beeraka, A. Kumar, M. Sameer, S. Ghosh, and B. Gupta, "Accuracy enhancement of epileptic seizure detection: a deep learning approach with hardware realization of STFT," Circuits, Systems, and Signal Processing, vol. 41, pp. 461–484, 2022.

[20] Y. Zheng, X. Zhao, and L. Yao, "Copula-based transformer in EEG to assess visual discomfort induced by stereoscopic 3D," *Biomedical Signal Processing and Control*, vol. 77, p. 103803, 2022.

[21] L. Li and N. Sun, "Attention-Based DSC–ConvLSTM for Multiclass Motor Imagery Classification," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 8187009, 2022.

[22] X. Liu, S. Xiong, X. Wang, T. Liang, H. Wang, and X. Liu, "A compact multi-branch 1D convolutional neural network for EEG-based motor imagery classification," *Biomedical Signal Processing and Control*, vol. 81, p. 104456, 2023.

[23] S. Lian and Z. Li, "An end-to-end multi-task motor imagery EEG classification neural network based on dynamic fusion of spectral-temporal features," Computers in Biology and Medicine, vol. 178, p. 108727, 2024.

[24] C. Zhang, H. Chu, and M. Ma, "Decoding algorithm of motor imagery electroencephalogram signal based on CLRNet network model," Sensors, vol. 23, no. 18, p. 7694, 2023.