IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 March 2024, accepted 20 May 2024, date of publication 27 May 2024, date of current version 12 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405652

RESEARCH ARTICLE

Three-Branch Temporal-Spatial Convolutional Transformer for Motor Imagery EEG Classification

WEIMING CHEN, YIQING LUO, AND JIE WANG
College of Software, Jilin University, Changchun 130012, China

Corresponding author: Weiming Chen (chenwm02@163.com)

This work was supported by the College Students' Innovation and Entrepreneurship Training Program of Jilin Province under Grant S202310183433.

ABSTRACT In the classification of motor imagery Electroencephalogram (MI-EEG) signals through deep learning models, challenges such as the insufficiency of feature extraction due to the limited receptive field of single-scale convolutions, and overfitting due to small training sets, can hinder the perception of global dependencies in EEG signals. In this paper, we introduce a network called EEG TBTSCTnet, which represents Three-Branch Temporal-Spatial Convolutional Transformer. This approach expands the size of the training set through Data Augmentation, and then combines local and global features for classification. Specifically, Data Augmentation aims to mitigate the overfitting issue, whereas the Three-Branch Temporal-Spatial Convolution module captures a broader range of multi-scale, low-level local information in EEG signals more effectively than conventional CNNs. The Transformer Encoder module is directly connected to extract global correlations within local temporal-spatial features, utilizing the multi-head attention mechanism to effectively enhance the network's ability to represent relevant EEG signal features. Subsequently, a classifier module based on fully connected layers is used to predict the categories of EEG signals. Finally, extensive experiments were conducted on two public MI-EEG datasets to evaluate the proposed method. The study also allowed for an optimal selection of channels to balance accuracy and cost through weight visualization.

INDEX TERMS EEG classification, motor imagery, transformer, temporal-spatial convolutional network, data augmentation.

I. INTRODUCTION

The quest for efficient and accurate Motor Imagery Electroencephalogram (MI-EEG) classification has been at the heart of Brain-Computer Interface (BCI) research, driven by its potential to revolutionize assistive technologies and rehabilitative medicine [1]. BCIs based on EEG have broad prospects in many application fields in daily life because of their reliability and convenience, ranging from functional rehabilitation for patients with motor disorders [2], sleep stage classification [3], emotional regulation, to general intelligent applications such as brain-controlled systems [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

Despite remarkable strides in this domain, leveraging deep learning and advanced signal processing techniques, existing methodologies often face challenges, such as high inter-subject variability, limited robustness against noisy EEG data, and suboptimal generalization across diverse datasets.

The pioneering integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for MI-EEG classification, exhibits significant limitations [5], [6], [7], [8], [9]. For instance, although models such as EEG-inception [10] demonstrate superior performance in controlled datasets, they struggle to maintain consistency across varying experimental conditions or in real-world BCI applications owing to their complexity and

computational demands. Similarly, only attention mechanisms underscore the criticality of channel interdependencies but often overlook the dynamic temporal relationships intrinsic to EEG signals, impacting their adaptability and classification accuracy [11], [12], [13].

To address these challenges, we propose a Three-Branch Temporal-Spatial Convolutional Transformer framework called EEG TBTSCTnet to synergistically leverage the advantages of multi-branch CNNs and the Transformer. The entire framework is composed of a Data Augmentation module, Three-Branch Temporal-Spatial Convolution module, Transformer Encoder module, and classifier module, all serially connected. Initially, a Data Augmentation strategy is employed to expand the small sample data to mitigate the overfitting issue [14], [15], [16]. In the convolution module, three scales of temporal and spatial convolutions are utilized to capture local time and spatial features, respectively. The obtained features were then subjected to an average pooling layer, reducing the model complexity, and eliminating redundant information. Subsequently, the three feature matrices are concatenated and fed into the Transformer Encoder module, where a multi-head self-attention layer further learns global temporal dependencies. Finally, a simple fully connected layer and Softmax function were used to obtain decoding results. Detailed comparative experiments on two different modes of EEG signal datasets revealed the superior performance of the EEG TBTSCTnet.

The main contributions of this study are as follows:

- Development of EEG TBTSCTnet: A pioneering framework combining multi-branch CNNs with Transformer technology for advanced MI-EEG classification, addressing current challenges in the field.
- Holistic feature extraction and classification Approach: Data Augmentation, Three-Branch Convolution, and Transformer Encoder modules are incorporated to enhance model performance by effectively capturing both local and global EEG signal features.
- Balancing accuracy and cost: Weight visualization and recursive channel elimination reveal that fewer EEG channels do not significantly impact performance, allowing for an optimal selection of channels to balance accuracy and cost.
- Insensitivity to Hyperparameter Selection: Extensive experiments were conducted to investigate the influence of the Transformer module and attention parameters. The results indicate that the model is insensitive to the number of heads and the depth of the self-attention layers in the multi-head self-attention module when processing EEG data.

The remainder of this paper is organized as follows. Related work is presented in Section II. A detailed description of the method is provided in Section III. The experiments and results are discussed in Section IV. The limitations and future work are presented in Section V, and conclusions are presented in Section VI.

II. RELATED WORKS

Recent studies in the field of MI-EEG classification have predominantly focused on developing more accurate and robust classification models to enhance the performance of BCI systems. This article reviews several key studies in this field, showcasing the latest technological advancements and methodologies.

Zhang et al. [10] introduced an EEG-inception CNN architecture for MI-EEG classification. This model, built on the inception-time network backbone, not only offers high accuracy but also processes raw EEG signals directly, eliminating the need for complex preprocessing steps. Li et al. [17] proposed a strategy based on FBCSP combined with a voting mechanism for three-class motor imagery classification, addressing the challenge of extending the CSP algorithm to multi-class MI scenarios. Their approach, which transforms a three-class problem into two binary-class problems, demonstrated an encouraging average classification accuracy of 68.6% with BCI competition IV Dataset 2a. Lawhern et al. [18] unveiled EEGNet, a versatile CNN for EEG-based BCIs, adept at classifying signals across multiple paradigms with minimal data, demonstrating robustness and high performance. Sakhavi et al. [19] utilized CNNs to enhance the extraction of temporal features from EEG signals by customizing parameters for individual subjects, significantly advancing BCI performance. Zheng et al. [20] introduced a Robust Support Matrix Machine (RSMM) for EEG classification, addressing EEG signal complexities with a novel classifier that utilizes matrix representation to improve BCI performance. By decomposing EEG data into a clean matrix with sparse noise, RSMM enhances classification accuracy and robustness against artifacts and noise, offering significant advancements over traditional classifiers. Yang et al. [21] devised a deep learning optimization framework for MI-EEG recognition, integrating CNN and RNN-LSTM to extract spatial, spectral, and temporal features, thereby enhancing system robustness and classification accuracy.

To address the limitations of single network models, an increasing number of hybrid network models have been proposed in recent years, all of which have demonstrated promising results. Altuwaijri et al. [22] developed a multi-branch CNN model incorporating squeeze-and-excitation (SE) attention blocks (MBEEGSE), adaptively modifying channel-wise feature responses by clearly specifying channel interdependencies and achieving commendable accuracy. Voinas et al. [23] focused on rehabilitation for stroke survivors and compared different feature extraction methods (WPD+HOS, CSP, FBCSP) for MI-EEG data classification of left and right wrist dorsiflexion, showing that the

WPD+HOS method achieved over 70% average accuracy, outperforming the CSP and FBCSP methods.

These studies highlight significant advances in MI-EEG signal classification using deep learning techniques, including efficient network architecture designs, incorporation of attention mechanisms, and optimizations for specific application scenarios. Together, they have propelled the development of BCI systems in terms of accuracy, robustness, and real-time capabilities, thereby laying a solid foundation for future research.

Therefore, inspired by the works above, we propose the TBTSCTnet framework as an efficient backbone.

III. METHOD

A. OVERVIEW

In this study, we introduce a novel neural network architecture called TBTSCTnet, which represents Three-Branch Temporal-Spatial Convolutional Transformer. This framework addresses the extraction of EEG signal features and classification of MI-EEG signals in an end-to-end fashion. TBTSCTnet harnesses the strength of Three-Branch Temporal-Spatial Convolutional Networks to capture information across various scales and contexts, coupled with the global correlation capturing abilities of the Transformer Encoder to learn global dependencies.

As shown in Figure 1, the model includes four modules: Data Augmentation, Three-Branch Temporal-Spatial Convolutional Network, Transformer Encoder, and Fully Connected Classification. Preprocessing precedes the input into TBTSCTnet, with a Segmentation and Reconstruction data augmentation technique applied for performance enhancement. Within the Three-Branch Temporal-Spatial Convolutional Network module, Temporal Convolutions utilize the temporal dimension for feature extraction, whereas Spatial Convolutions operate along the electrode channel dimension, utilizing average pooling to suppress noise interference [24]. The features were extracted using three sets of convolutional kernels and pooling layers of varying sizes. These extracted features are then concatenated to form a comprehensive matrix encapsulating the multiscale temporal-spatial information. This matrix is then fed into the Transformer Encoder with Multi-Head Attention to extract long-term temporal features. Finally, a simple two-layer fully connected network with a Softmax layer executes the classification.

B. PREPROCESSING AND DATA AUGMENTATION

To preserve the structural integrity of the EEG data, minimal preprocessing is applied before feeding the raw EEG trials into the model. A 6th-order Chebyshev filter is used to constrain the EEG signal frequency within the range of $[W1, W2]Hz$, with the aim of eliminating various high and low-frequency artifacts while retaining valuable rhythmic information. Specifically, $[W1, W2]$ is set to [8,30].

For the dataset BCI competition IV dataset 2a [25], TBTSCTnet model takes as input a motor imagery trial $X_i \in \mathbb{R}^{C\times T}$ consisting of C channels (EEG electrodes) and T time points. The objective of the TBTSCTnet model is to map the input MI trial X_i to its corresponding class y_i, given a set of m labeled MI trials $S = \{X_i, y_i\}_{i=1}^{m}$, where $y_i \in \{1, \ldots, n\}$ is the corresponding class label for trial X_i and n is the total number of defined classes for set S. For each Subject in this dataset, $T = 1000$ time points, $C = 22$ EEG channels, $n = 4$ MI classes, and $m = 48$ MI trials.

For BCI Competition IV Dataset 2b [25], the data dimensions for X_i, y_i, and S are consistent with those of Dataset 2a. However, EEG data originate from three electrodes C3, Cz, and C4, which are responsible for recording motor imagery. Consequently, the channel configuration differed, as did the number of trials per subject. For each Subject in this dataset, $T = 1000$ time points, $C = 3$ EEG channels, $n = 2$MI classes, and $m = 160$ MI trials were used.

Subsequently, a Segmentation and Reconstruction data augmentation technique is employed to enhance the performance of the model during the training phase. This augmentation strategy is specifically designed for EEG data, to address the challenges associated with limited labeled data for neural network training. During the Segmentation and Reconstruction process, the EEG data are organized into distinct segments, which are subsequently reconstructed to generate augmented training samples [26] [27]. Augmentation is conducted within each class, where the EEG signals were categorized based on the corresponding class labels.

C. NETWORK ARCHITECTURE

As shown in Figure 1, the input is a batch of preprocessed EEG trials $X_i \in \mathbb{R}^{C\times T}$,expanded by one dimension as the convolution channel, denoted by $X_i \in \mathbb{R}^{C\times 1\times T}$.

1) THREE-BRANCH TEMPORAL-SPATIAL CONVOLUTIONAL NETWORK

The Three-Branch Temporal-Spatial Convolutional Network (TBTSCN) is a pivotal component of TBTSCTnet. It is designed to effectively capture and process the temporal-spatial features of EEG signals at multiple scales. This method facilitates a deeper and thorough comprehension of EEG data, improving the accuracy of the models for MI-EEG classification. TBTSCN achieves this by employing convolutional layers that operate at different branches, each of which extracts distinct scale-specific features. These layers were then integrated to form a cohesive feature map that represented a wide range of temporal and spatial signal characteristics. This three-branch approach is instrumental for improving the accuracy and reliability of EEG signal classification, making it a valuable tool in the field of MI-EEG classification.

As shown in Figure 2, the structure of the TBTSCN consists of several convolutional layers that operate at different temporal branches. The input EEG signal is passed through

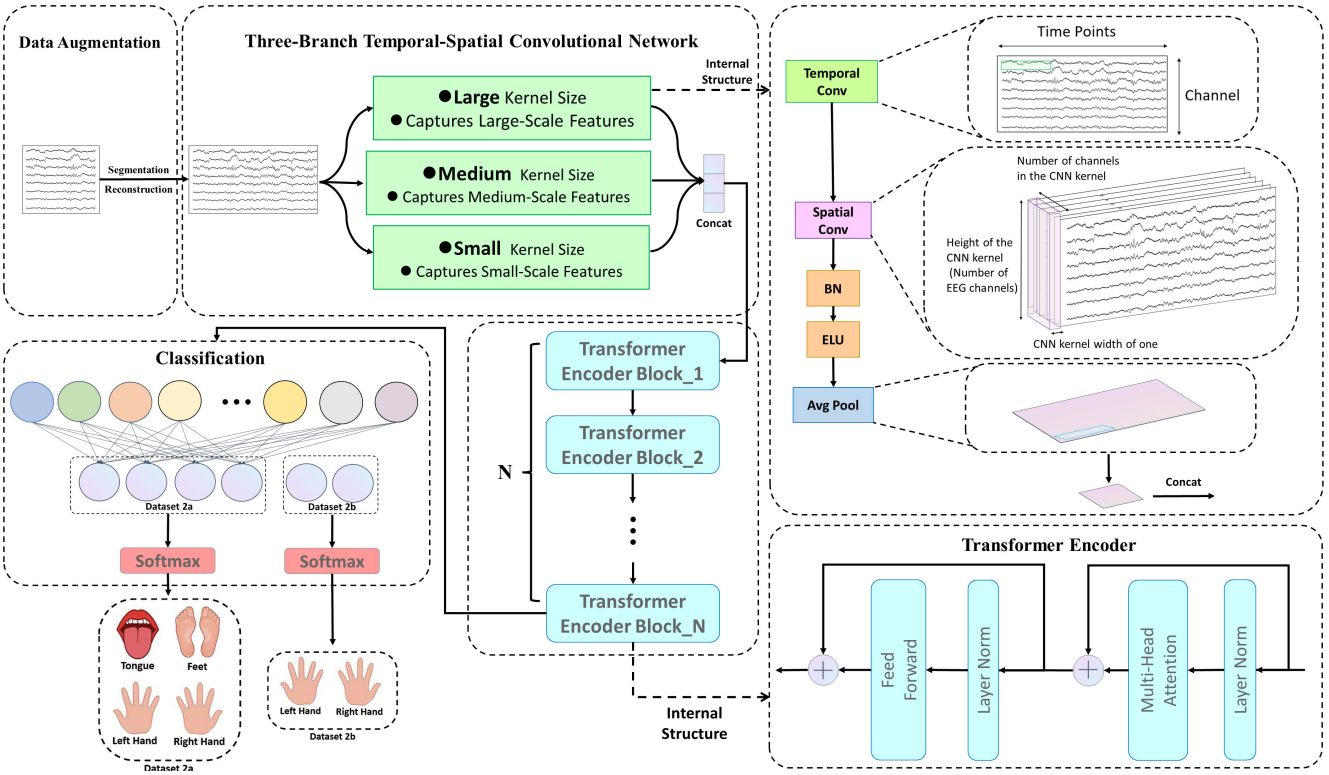


FIGURE 1. The framework of Three-Branch Temporal-Spatial Convolutional Transformer (TBTSCTnet), including a Data Augmentation module, a Three-Branch Temporal-Spatial Convolutional Network module, a Transformer Encoder module, and a classifier module, where BN stands for Batch Normalization and ELU stands for Exponential Linear Unit.

three parallel temporal convolutional layers, each with a distinct filter size: (1,70), (1,40), and (1,20), all with a stride of (1,1), allowing the network to capture various time-dependent features from the input. The outputs of these layers are then passed through the respective spatial convolution layers with a filter size of (ch,1) and a stride of (1,1), further refining the feature extraction process. Following this, batch normalization [28] and exponential linear unit (ELU) [29] activation functions were applied. After the initial temporal and spatial convolutions, the network integrated three parallel branches of the average pooling layers. Each branch corresponds to a different filter size, specifically tailored to the features extracted at each branch. The first pooling layer utilize a pool size of (1,30) with a stride of (1,20), the second employs a pool size of (1,60) with the same stride, and the third uses a pool size of (1,80) with a stride of (1,20). This design allows the network to down-sample the feature maps in a manner that preserves the critical spatial information across multiple branches, which is essential for accurately capturing the dynamics of EEG signals. Finally, they were concatenated to a $(3\times m, k)$ tensor, which is used as the input for the next module.

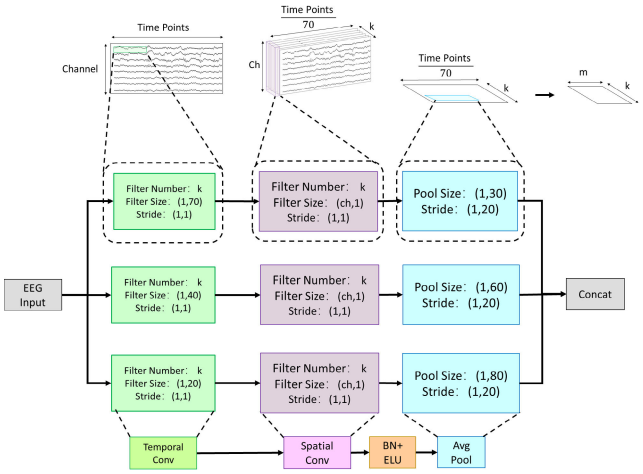


FIGURE 2. Architecture of the Three-Branch Temporal-Spatial Convolutional Network Module in TBTSCTnet, where BN stands for Batch Normalization and ELU stands for Exponential Linear Unit.

2) TRANSFORMER ENCODER

As shown in Figure 3, it begins with an input feature map processed through linear layers to generate queries, keys, and values [30]. These are essential components of the attention mechanism that allow the model to focus on different parts of the input sequence. The queries, keys, and values are then passed through multiple heads in the multi-head attention mechanism, where each head captures different aspects of the input data. The attention outputs from all the heads were concatenated, scaled, and normalized using Softmax. The final step in Transformer Encoder is a feed-forward neural network that processes the concatenated output to produce the final output of the Transformer Encoder block. This architecture is particularly effective for capturing both local and global

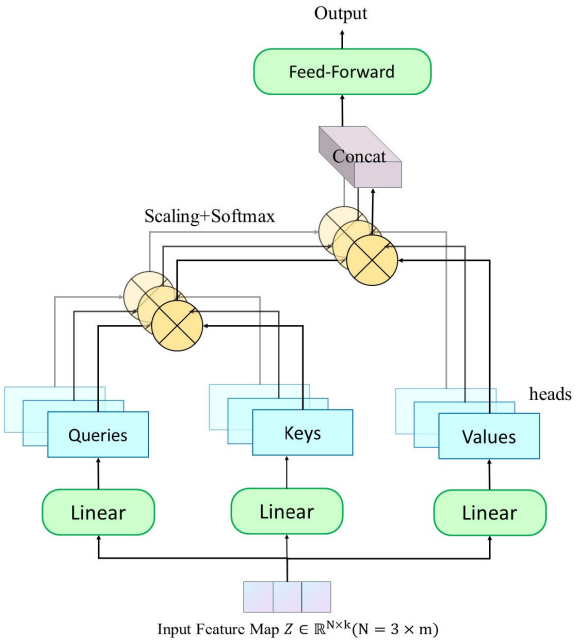


FIGURE 3. Architecture of the Transformer Encoder Module in TBTSCTnet.

dependencies within the input data, making it suitable for MI-EEG classification tasks. The dimensions are transformed as follows:

a) Linear projections for queries, keys, and values for each head h

$$\begin{cases} Q_h = ZW_h^Q \\ K_h = ZW_h^K \\ V_h = ZW_h^V \end{cases} \tag{1}$$

where $W_h^Q \in \mathbb{R}^{k\times d_q}$, $W_h^K \in \mathbb{R}^{k\times d_k}$, $W_h^V \in \mathbb{R}^{k\times d_v}$ are weight matrices for queries, keys, and values, respectively.

b) Scaled dot-product attention for each head:

$$\text{Attention}\,(Q_h, K_h, V_h) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h \tag{2}$$

where d_k is the dimensionality of the keys. The output dimension for the attention matrix of each head is $\mathbb{R}^{N\times d_v}$.

c) Concatenation of all heads' outputs and final projection:

$$\begin{cases} MultiHead\,(Q, K, V) = Concat\,(head_1, \ldots, head_h)\, W^O \\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{cases} \tag{3}$$

where $W^O \in \mathbb{R}^{hd_v\times d_{model}}$ is the final projection matrix.

3) CLASSIFICATION

Final classification layer is a classifier composed of two fully connected layers that precede a Softmax function [31]. The initial dense layer serves to interpret the rich, feature-laden representations delivered by the preceding network stages, distilling this information into a format suitable for classification. Subsequently, a second dense layer refines the distilled information, further tuning the discriminative capabilities of the network. The Softmax function operates as the final component of this architecture, converting the output of the final dense layer into a probability distribution across the anticipated classes. This setup ensures that the network output can be interpreted as the likelihood of each class, thus enabling a decision-making process based on the probabilistic assessment of the input EEG signals.

IV. EXPERIMENTS AND RESULTS

A. EVALUATION INDICATORS

The performance of classification models accuracy (Acc) and kappa coefficient are two widely recognized metrics for evaluating the performance of classification models. Accuracy is defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined. It is calculated using the formula:

$$Acc = \frac{TP + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

The kappa coefficient, also known as Cohen's kappa, measures the agreement between two raters who classify N items into C mutually exclusive categories. The kappa score accounts for the possibility of agreement occurring by chance, providing a more robust understanding of the classifier's performance, especially in imbalanced datasets. It is calculated as

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement. A kappa value of 1 implies perfect agreement, whereas a value of 0 indicates that the agreement is no better than chance.

B. EXPERIMENTAL SETUP

The experimental setup for TBTSCTnet is conducted in a controlled environment furnished by local platforms. The hardware infrastructure is centered around an AMD Ryzen 7940HS CPU coupled with 32GB of RAM and an Nvidia RTX 4060 GPU, featuring 8GB of memory. The software framework is grounded in a Windows 11 operating system, with Python 3.11 as the programming language milieu. All the experimental procedures were executed using PyTorch as the backend.

To optimize the TBTSCTnet model, the Adam optimization algorithm is employed for the TBTSCTnet model, with the batch size adjusted to 72 for Dataset 2a and 80 for Dataset 2b, maintaining a learning rate of 0.0001. A cross-entropy loss function is employed to refine the training process.

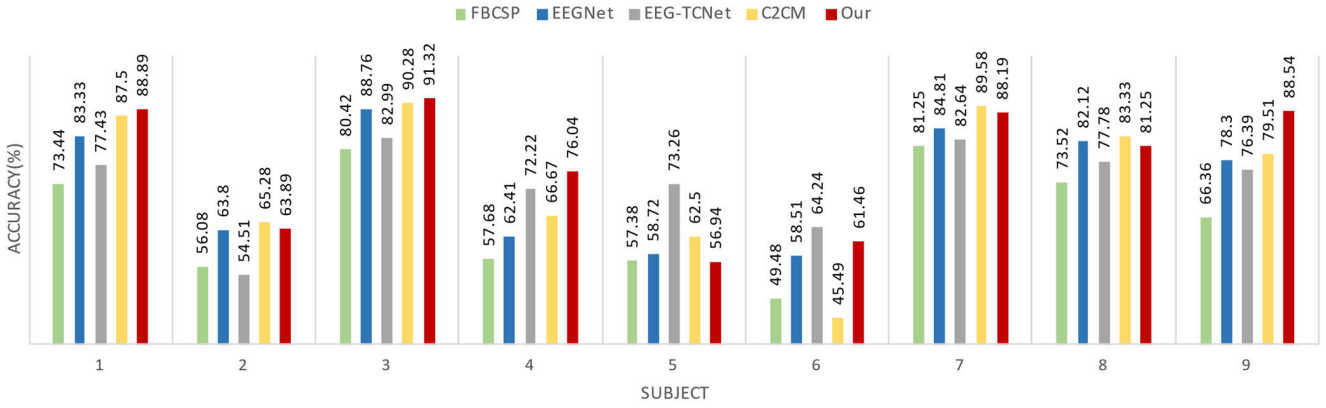


FIGURE 4. Comparison of classification accuracy of the proposed model with other methods on the test set in Dataset 2a.

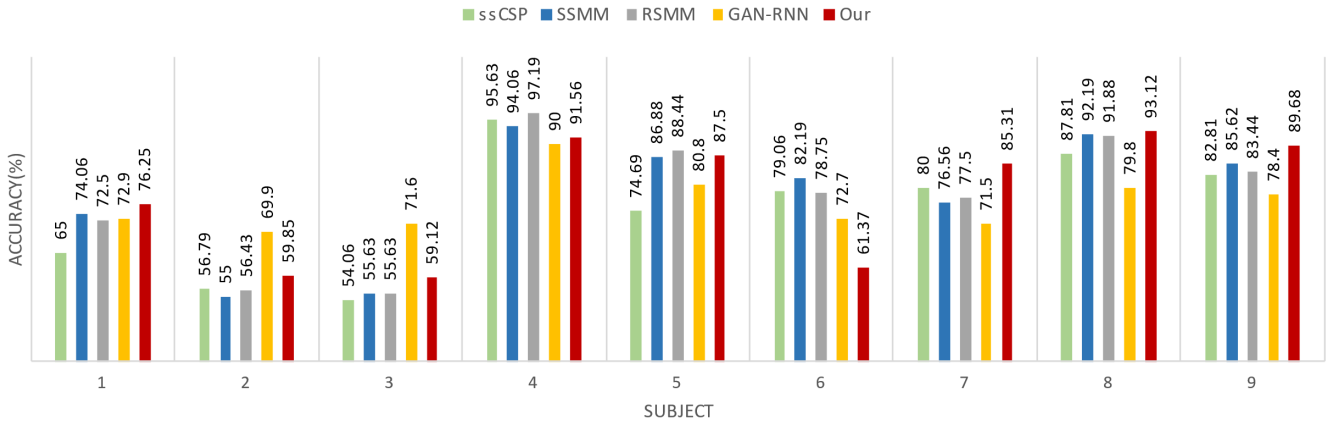


FIGURE 5. Comparison of classification accuracy of the proposed model with other methods on the test set in Dataset 2b.

TABLE 1. Comparison of classification accuracy of our model with other methods on the test set in Dataset 2a.

Method	Subject01	Subject02	Subject03	Subject04	Subject05	Subject06	Subject07	Subject08	Subject09	Average	Kappa
FBCSP[32]	73.44	56.08	80.42	57.68	57.38	49.48	81.25	73.52	66.36	66.18	0.59
EEGNet[18]	83.33	63.80	88.76	62.41	58.72	58.51	84.81	82.12	78.30	73.42	0.67
EEG-Inception[10]	77.43	54.51	82.99	72.22	73.26	64.24	82.64	77.78	76.39	73.50	0.68
C2CM[19]	87.5	65.28	90.28	66.67	62.5	45.49	89.58	83.33	79.51	74.46	0.66
Our	**88.89**	**63.89**	**91.32**	**76.04**	**56.94**	**61.46**	**88.19**	**81.25**	**88.54**	**77.39**	**0.71**

TABLE 2. Comparison of classification accuracy of our model with other methods on the test set in Dataset 2b.

Method	Subject01	Subject02	Subject03	Subject04	Subject05	Subject06	Subject07	Subject08	Subject09	Average	Kappa
ssCSP[33]	65.00	56.79	54.06	95.63	74.69	79.06	80.00	87.81	82.81	75.09	0.50
SSMM[34]	74.06	55.00	55.63	94.06	86.88	82.19	76.56	92.19	85.62	78.00	0.56
RSMM[20]	72.50	56.43	55.63	97.19	88.44	78.75	77.50	91.88	83.44	77.97	0.60
GAN-RNN[21]	72.90	69.90	71.60	90.00	80.80	72.70	71.50	79.80	78.40	76.4	0.61
Our	**76.25**	**59.85**	**59.12**	**91.56**	**87.5**	**61.37**	**85.31**	**93.12**	**89.68**	**78.20**	**0.61**

Dropout rates of 0.5 within the Three-Branch Convolution module and 0.3 in the Transformer encoder were implemented to prevent overfitting. The self-attention mechanism is set to execute six times with 10 heads, balancing efficiency and robustness in model training. Finally, the number of epochs is set to 500.

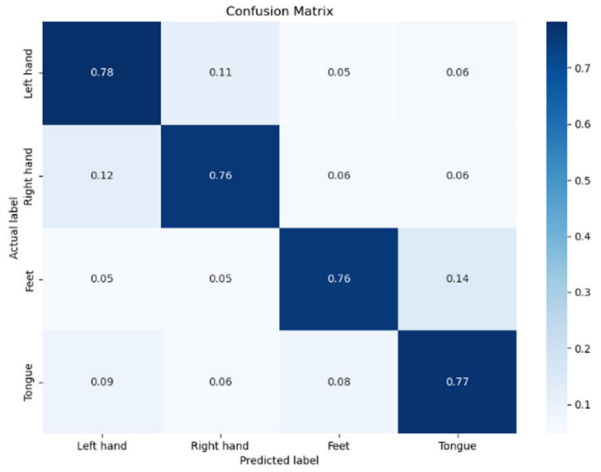


FIGURE 6. Confusion matrices of dataset 2a.

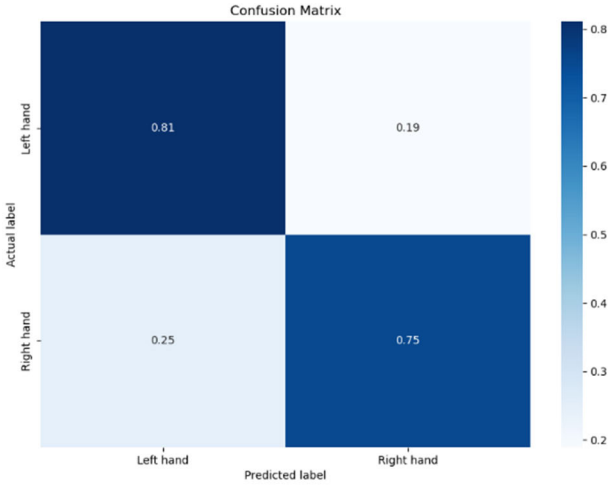


FIGURE 7. Confusion matrices of dataset 2b.

C. DATASETS

To assess the robustness of our EEG-based classification method, we used two distinct datasets, each with unique characteristics.

The first dataset from BCI Competition IV (Dataset 2a), courtesy of the Graz University of Technology, encompasses recordings from nine individuals performing four types of motor imagery tasks. It features two separate sessions for each subject with a 250 Hz sampling rate using twenty-two channels. For our analysis, we extracted a specific timeframe from each trial and applied a band-pass filter to isolate the relevant frequency band.

The second dataset, also from BCI Competition IV (Dataset 2b), included EEG data from nine participants focusing on two motor imagery tasks with a similar sampling rate but utilizing three bipolar electrodes. Multiple sessions provided a substantial number of trials for both training and testing, with a specific segment for each trial after band-pass filtering.

These datasets, with varied configurations, were instrumental in demonstrating the versatility of our classification approach.

D. COMPARISON OF CLASSIFICATION RESULTS

1) IN OUR EVALUATION USING THE BCI COMPETITION IV, DATASET 2A

The classification performance of each subject and the average results are shown in Figure 4 and Table 1. The FBCSP method, which is based on machine learning, achieved a mean accuracy of 66.18% across subjects, indicating a lack of robustness and lower performance compared to deep learning approaches. EEGNet, despite its compact framework, showed an impressive capability for temporal feature extraction and maintained good generalization. The EEG-Inception model adeptly captured features across different time scales through its dual-scale inception structure. C2CM, while incorporating

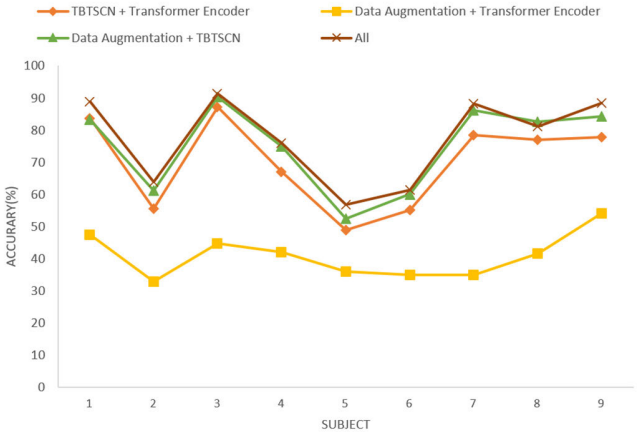


FIGURE 8. Ablation Study Results on Accuracy Across Subjects for TBTSCTnet.

hand-crafted features with deep models, did not outperform our method, except for one instance. This underlines the potential limitations of models that do not adapt across diverse subjects, despite parameter fine-tuning. In comparison with recent studies, the method proposed in this study significantly enhances MI-EEG classification tasks. It employs TBTSCN for feature extraction from EEG signals and utilizes Transformer Encoder to analyze the combined features, effectively reducing the generalization issues caused by inter subject variability. This approach has achieved high classification performance for most subjects, with an average accuracy of 77.39%, surpassing the average accuracy of existing deep learning models.

To further analyze the impact of TBTSCTnet on the recognition of each class of MI-EEG, the confusion matrices for dataset 2a under the model are calculated, as shown in Figure 6. In this figure, the horizontal axis of the confusion matrix represents the categories of motor imagery predicted by the model. The diagonal elements indicate the proportion

TABLE 3. Average accuracy comparison of TBTSCTnet components in ablation study.

Conditions	Average Accuracy (%)
TBTSCN + Transformer Encoder	70.10
Data Augmentation + Transformer Encoder	41.04
Data Augmentation + TBTSCN	75.03
All	**77.39**

of correct classifications for the four types of motor imagery tasks by the model, while the remaining results represent the proportion of misclassifications. For most subjects, the proposed model achieves a prediction accuracy of over 76% when predicting tasks involving the left hand, right hand, both feet, and the tongue, demonstrating strong stability.

2) IN OUR EVALUATION USING THE BCI COMPETITION IV, DATASET 2B

As shown in Figure 5 and Table 2, the performance data highlight that our TBTSCTnet model has achieved superior classification accuracy in comparison to the RSMM, GAN-RNN, ssCSP, and SSMM models for most subjects. Particularly noteworthy is the model's performance for Subject 4 and Subject 8 where our model demonstrates its robustness with accuracy rates of 91.56% and 93. 12%, respectively. Across all subjects, our model presents an average accuracy of 78.20%, which is in line with a kappa coefficient of 0.61, reflecting a meaningful level of agreement beyond chance. These results reinforce the capability of our model to accurately classify MI-EEG signals.

To further analyze the impact of TBTSCTnet on the recognition of each class of MI-EEG, the confusion matrices for dataset 2b under the model are calculated, as shown in Figure 7.

E. ABLATION STUDY

After verifying the overall framework, ablation studies were conducted. This process involved systematically removing each component from the model to assess the efficacy of several key feature transfer components. The experimental results for Dataset 2a are presented in Figure 8 and Table 3.

1) DATA AUGMENTATION

Based on the comparison shown in Figure 8, it is evident that Data Augmentation plays a beneficial role in enhancing accuracy for all the subjects. As indicated in Table 3, the average accuracy of TBTSCTnet with Data Augmentation is 7.29% higher than that of TBTSCTnet without Data Augmentation. This clearly demonstrates the value of incorporating Data Augmentation in the model to improve its performance across various subjects. In conclusion, the strategic integration of Data Augmentation into the TBTSCTnet framework is pivotal for achieving a superior classification accuracy in EEG signal analysis.

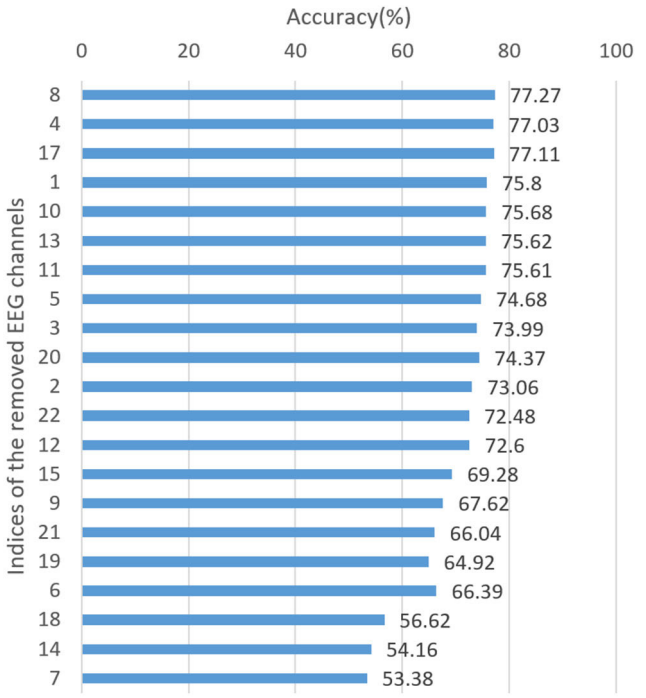


FIGURE 9. Relationship between the indices of the removed EEG channels and accuracy. The vertical axis represents continued removal based on the previous removals, from top to bottom.

2) TBTSCN

The TBTSCN significantly enhances the model's capability by capturing features across various branches, which is instrumental in boosting the overall performance. As illustrated in Figure 8, the transition from the yellow to green line, which indicates a substantial improvement in accuracy, is attributed solely to the inclusion of TBTSCN. Table 3 further quantifies this impact, showing a 36.35% increase in average accuracy upon integrating TBTSCN, thereby underscoring its vital role in the precision of the model.

3) TRANSFORMER ENCODER

Incorporating the Transformer Encoder into the TBTSCTnet significantly improves the model's understanding of EEG data by capturing global dependencies. As shown in Figure 8, the addition of the Transformer Encoder (indicated by the red line) enhances the accuracy of the model, and the average accuracy is enhanced by 2.36%. The impact of this component is evident from the results, with a marked increase in performance, highlighting the ability of the encoder to extract relevant features over long sequences, which is crucial for EEG signal classification.

F. TBTSCNET WEIGHT VISUALIZATION TO EXPLAIN THE EFFECT OF THE NUMBER OF EEG CHANNELS ON WHOLE PERFORMANCE

To discuss the effect of the number of EEG channels on overall performance, this study proposes a method that

combines weight visualization and recursive channel elimination to identify the most valuable EEG channels quickly and rationally. This approach avoids the traditional large-scale combinatorial search for identifying EEG channels. Specifically, the convolutional kernels in the spatial convolution module of the TBTSCN are visualized. Note that the height of these convolutional kernels is equal to the number of EEG channels, as shown in the upper part of Figure 10. Then, the columns of the convolutional kernels are summed, as depicted in the lower part of Figure 10, to determine the importance of each EEG channel to the classification results. Figure 10 already marks the actual brain location corresponding to each channel. The importance of EEG channels in descending order is: [16, 7, 14, 18, 6, 19, 21, 9, 15, 12, 22, 2, 20, 3, 5, 11, 13, 10, 1, 17, 4, 8].

Next, the recursive channel elimination was implemented. The least important channels are successively removed, and the final classification accuracy is recorded after each removal, as shown in Figure 9. It can be observed that removing the 8th, 4th, and 17th channels does not significantly decrease the classification accuracy. However, after removing the first channel, the accuracy drops by 1.3%. Interestingly, retaining only the top 9 channels from the sorted list still results in a final classification accuracy exceeding 72%.

The above study indicates that it is not necessary to use data from all 22 electrodes to achieve high accuracy; using data from 19 electrodes can attain the same level of accuracy. If there are cost constraints on the sampling equipment, even selecting only the top 9 electrodes for sampling and analysis is viable. This study demonstrates the optimal method for selecting channels to balance accuracy and cost.

G. COMPARING DIFFERENT TRANSFORMER ENCODER SCHEMES

The number of heads in the multi-head attention mechanism of the transformer model is a crucial parameter that facilitates the learning of different aspects of features. We also evaluated the impact of varying the number of heads on the model performance. As shown on the left side of Figure 11, we experimented with five different numbers of heads ranging from 1 to 20. On the vertical axis, there is no clear relationship between the number of heads and the impact on the results, with no significant difference in the distribution across different subjects. The average accuracy remains moderately fluctuating, with only a 1.93% range on dataset 2a. On dataset 2a, the average accuracy of using ten heads is 1.28% higher than that of using only one head. Overall, variations in the number of heads did not significantly enhance feature learning.

Depth is also a key factor that influences the fitting capability of a model, particularly in traditional models. As shown on the right side of Figure 11, we explored the impact of depth on accuracy by incrementally increasing the depth of self-attention layers from one to six. It can be observed that for dataset 2a, the highest accuracy is only 1.97% higher than the lowest accuracy, with the difference being statistically

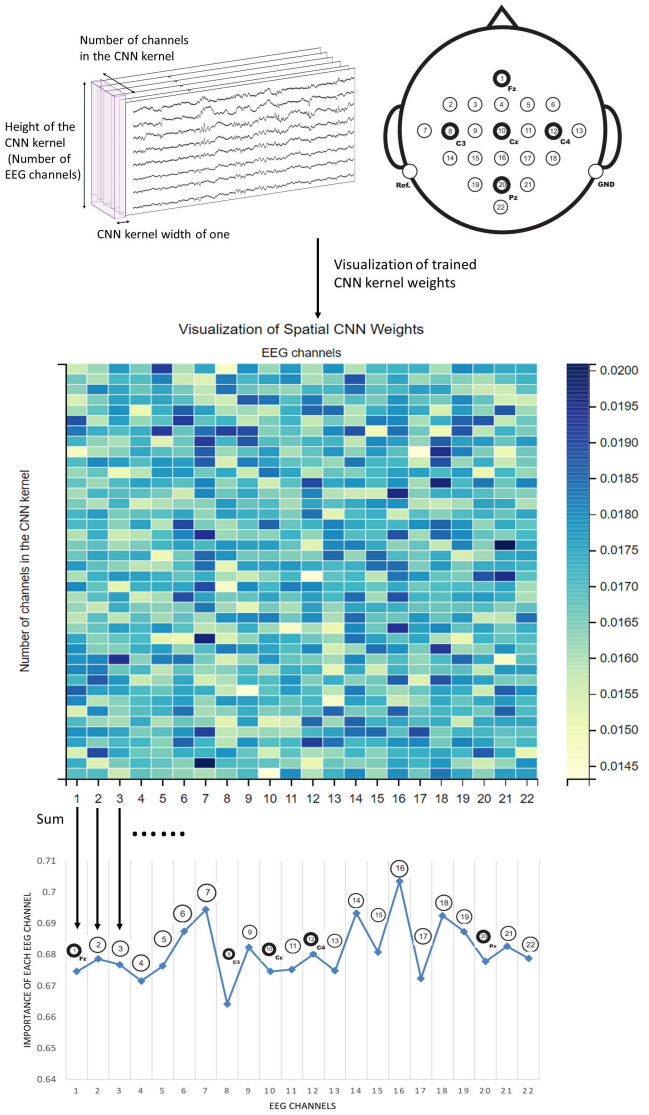


FIGURE 10. This figure consists of three parts. The top section shows a schematic diagram of the convolutional kernels in the spatial CNN and the distribution of electrodes on the brain. The middle section displays a heatmap of the convolutional kernel weights. The bottom section indicates the importance of each EEG channel to the classification results.

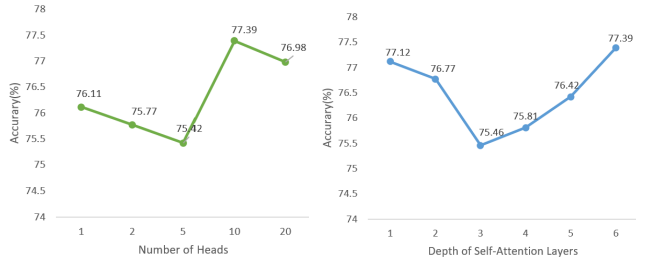


FIGURE 11. Accuracy Trends Across Varying Number of Heads and Depth of Self-Attention Layers on Transformer Encoder Module.

insignificant. However, the number of parameters in the transformer dramatically increases with depth, significantly increasing the training cost of the model. Hence, it can be

concluded that the TBTSCTnet is insensitive to the depth of self-attention.

V. LIMITATIONS AND FUTURE WORK

The proposed TBTSCTnet model, while demonstrating significant improvements in MI-EEG classification accuracy, is not without its limitations. One primary limitation is the model's dependence on large-scale data for training, which may not always be feasible in real-world scenarios due to the scarcity of labeled EEG data. Additionally, the computational complexity of the model, particularly the Transformer Encoder module, may pose challenges for deployment in resource-constrained environments. The model's performance variability across different subjects also indicates a need for further refinement to enhance its robustness and generalization capabilities.

Future work will focus on several key areas to address these limitations. Firstly, optimizing the model architecture to reduce computational demands without compromising accuracy will be crucial for practical applications. This could involve investigating alternative lightweight Transformer architectures or hybrid models that balance performance and efficiency. Secondly, improving the model's adaptability to individual differences in EEG signals through personalized training strategies or domain adaptation techniques will be a significant focus, aiming to enhance the generalization of the TBTSCTnet model across diverse user populations.

VI. CONCLUSION

In conclusion, the TBTSCTnet significantly advances MI-EEG classification by addressing feature extraction and overfitting challenges. By integrating Data Augmentation, Three-Branch Temporal-Spatial Convolutions, and Transformer encoder modules, TBTSCTnet captures both local and global features, enhancing classification performance. Extensive experiments on public MI-EEG datasets demonstrate the model's robustness and adaptability, achieving higher accuracy than existing methods.

The weight visualization study shows that fewer EEG channels do not significantly impact performance, allowing for an optimal selection of channels to balance accuracy and cost. Additionally, experiments with different Transformer encoder schemes reveal minimal impact from varying the number of heads and the depth of self-attention layers, indicating the model's insensitivity to these parameters. The ablation study confirms the importance of each module, particularly the Data Augmentation and Three-Branch Convolutional Network, in enhancing overall accuracy.

Future work will focus on optimizing the model for real-world applications and improving its adaptability to individual EEG signal variations.

REFERENCES

[1] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain–computer interfaces," *Trends Cognit. Sci.*, vol. 25, no. 8, pp. 671–684, Aug. 2021.

[2] S. Samejima, A. Khorasani, V. Ranganathan, J. Nakahara, N. M. Tolley, A. Boissenin, V. Shalchyan, M. R. Daliri, J. R. Smith, and C. T. Moritz, "Brain-computer-spinal interface restores upper limb function after spinal cord injury," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1233–1242, 2021.

[3] R. Yu, Z. Zhou, S. Wu, X. Gao, and G. Bin, "MRASleepNet: A multi-resolution attention network for sleep stage classification using single-channel EEG," *J. Neural Eng.*, vol. 19, no. 6, Dec. 2022, Art. no. 066025.

[4] X. Shen, X. Zhang, Y. Huang, S. Chen, Z. Yu, and Y. Wang, "Intermediate sensory feedback assisted multi-step neural decoding for reinforcement learning based brain-machine interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2834–2844, 2022.

[5] A. Echtioui, W. Zouch, M. Ghorbel, C. Mhiri, and H. Hamam, "Multi-class motor imagery EEG classification using convolution neural network," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 591–595.

[6] J. Hwang, S. Park, and J. Chi, "Improving multi-class motor imagery EEG classification using overlapping sliding window and deep learning model," *Electronics*, vol. 12, no. 5, p. 1186, Mar. 2023.

[7] Z. Wang, L. Cao, Z. Zhang, X. Gong, Y. Sun, and H. Wang, "Short time Fourier transformation and deep neural networks for motor imagery brain computer interface recognition," *Concurrency Comput., Pract. Exper.*, vol. 30, no. 23, Dec. 2018, Art. no. e4413.

[8] S. Belgacem, A. Echtioui, R. Khemakhem, W. Zouch, M. Ghorbel, I. Kammoun, and A. B. Hamida, "Deep learning models for classification of motor imagery EEG signals," in *Proc. 6th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2022, pp. 1–4.

[9] F. Hassan, S. F. Hussain, and S. M. Qaisar, "Fusion of multivariate EEG signals for schizophrenia detection using CNN and machine learning techniques," *Inf. Fusion*, vol. 92, pp. 466–478, Apr. 2023.

[10] C. Zhang, Y.-K. Kim, and A. Eskandarian, "EEG-inception: An accurate and robust end-to-end neural network for EEG-based motor imagery classification," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046014.

[11] Y. Wen, W. He, and Y. Zhang, "A new attention-based 3D densely connected cross-stage-partial network for motor imagery classification in BCI," *J. Neural Eng.*, vol. 19, no. 5, Sep. 2022, Art. no. 056026.

[12] S. Bagchi and D. R. Bathula, "EEG-ConvTransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108757.

[13] J. Kalafatovich, M. Lee, and S.-W. Lee, "Decoding visual recognition of objects from EEG signals based on attention-driven convolutional neural network," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2985–2990.

[14] O. George, R. Smith, P. Madiraju, N. Yahyasoltani, and S. I. Ahamed, "Data augmentation strategies for EEG-based motor imagery decoding," *Heliyon*, vol. 8, no. 8, Aug. 2022, Art. no. e10240.

[15] Z. Zhang, F. Duan, J. Solé-Casals, J. Dinarès-Ferran, A. Cichocki, Z. Yang, and Z. Sun, "A novel deep learning approach with data augmentation to classify motor imagery signals," *IEEE Access*, vol. 7, pp. 15945–15954, 2019.

[16] Y. Pei, Z. Luo, Y. Yan, H. Yan, J. Jiang, W. Li, L. Xie, and E. Yin, "Data augmentation: Using channel-level recombination to improve classification performance for motor imagery EEG," *Frontiers Hum. Neurosci.*, vol. 15, Mar. 2021, Art. no. 645952.

[17] B. Li, B. Yang, C. Guan, and C. Hu, "Three-class motor imagery classification based on FBCSP combined with voting mechanism," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, Jun. 2019, pp. 1–4.

[18] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[19] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.

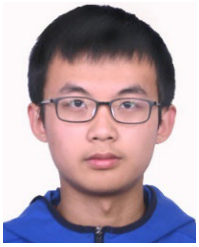
[20] Q. Zheng, F. Zhu, and P.-A. Heng, "Robust support matrix machine for single trial EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 551–562, Mar. 2018.

[21] B. Yang, C. Fan, C. Guan, X. Gu, and M. Zheng, "A framework on optimization strategy for EEG motor imagery recognition," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 774–777.

[22] G. A. Altuwaijri, G. Muhammad, H. Altaheri, and M. Alsulaiman, "A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification," *Diagnostics*, vol. 12, no. 4, p. 995, Apr. 2022, doi: 10.3390/diagnostics12040995.
[23] A. E. Voinas, R. Das, M. A. Khan, I. Brunner, and S. Puthusserypady, "Motor imagery EEG signal classification for stroke survivors rehabilitation," in *Proc. 10th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2022, pp. 1–5.
[24] J. Chen, Z. Yu, Z. Gu, and Y. Li, "Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2356–2366, Nov. 2020.
[25] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[26] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.
[27] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.
[28] S. Ioffe and C. J. A. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[29] D.-A. Clevert, T. Unterthiner, and S. J. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *Proc. Int. Conf. Learn. Represent.*, vol. abs/1511.07289, 2016.
[30] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.
[31] Y. LeCun, Y. Bengio, and G. J. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[32] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, Apr. 2012.
[33] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 8, pp. 2289–2298, Aug. 2013.
[34] Q. Zheng, F. Zhu, J. Qin, B. Chen, and P.-A. Heng, "Sparse support matrix machine," *Pattern Recognit.*, vol. 76, pp. 715–726, Apr. 2018.

WEIMING CHEN is currently pursuing the bachelor's degree with Jilin University. His research interests include deep learning and algorithm for analysis and processing of electroencephalogram signals.

YIQING LUO is currently pursuing the bachelor's degree with Jilin University.



JIE WANG is currently pursuing the bachelor's degree with Jilin University.