Research Repository

MSAttNet: Multi-scale attention convolutional neural network for motor imagery classification

MSAttNet: Multi-Scale Attention Convolutional Neural Network for Motor Imagery Classification

Ruiyu Zhao[a], Ian Daly[b], Yixin Chen[a], Weijie Wu[a], Lifei Liu[a], Xingyu Wang[a], Andrzej Cichocki[c,d,e] and Jing Jin[a,f,g,1]

[a] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China*

[b] *the Brain-Computer Interfacing and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, Essex, U.K*

[c] *Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako-shi, 351-0198, Japan*

[d] *Systems Research Institute, Polish Academy of Sciences, Warsaw, 01-447, Poland*

[e] *Department of Informatics, Nicolaus Copernicus University, , Torun, 87-100, , Poland*

[f] *School of Math, East China University of Science and Technology, Shanghai, 200237, China*

[g] *The Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, , Shanghai, 200237, China*

ARTICLE INFO

Keywords:
Brain-Computer Interfaces, Motor Imagery, Convolutional Neural Network, Attention Convolution

ABSTRACT

Background: Convolutional neural networks (CNNs) are widely employed in motor imagery (MI) classification. However, due to cumbersome data collection experiments, and limited, noisy, and non-stationary EEG signals, small MI datasets present considerable challenges to the design of these decoding algorithms. **New method**: To capture more feature information from inadequately sized data, we propose a new method, a multi-scale attention convolutional neural network (MSAttNet). Our method includes three main components–a multi-band segmentation module, an attention spatial convolution module, and a multi-scale temporal convolution module. First, the multi-band segmentation module adopts a filter bank with overlapping frequency bands to enhance features in the frequency domain. Then, the attention spatial convolution module is used to adaptively adjust different convolutional kernel parameters according to the input through the attention mechanism to capture the features of different datasets. The outputs of the attention spatial convolution module are grouped to perform multi-scale temporal convolution. Finally, the output of the multi-scale temporal convolution module uses the bilinear pooling layer to extract temporal features and perform noise elimination. The extracted features are then classified. **Results**: We use four datasets, including *BCI Competition IV Dataset IIa*, *BCI Competition IV Dataset IIb*, the *OpenBMI* dataset and the *ECUST-MI* dataset, to test our proposed method. MSAttNet achieves accuracies of 78.20%, 84.52%, 75.94% and 78.60% in cross-session experiments, respectively. **Comparison with existing methods**: Compared with state-of-the-art algorithms, MSAttNet enhances the decoding performance of MI tasks. **Conclusion**: MSAttNet effectively addresses the challenges of MI-EEG datasets, improving decoding performance by robust feature extraction.

1. Introduction

Brain-Computer Interfaces (BCIs) construct a transmission link between brain activity and computing devices. They do so by attempting to decode information from brain signals Liang, Kuang, Wang, Yuan, Zhang and Sun (2023a). The electroencephalogram (EEG) is the dominant brain activity recording methodology used in the majority of current non-invasive BCI systems and contains limited, noisy, and non-stationary signal features Wang, Yao and Wang (2023). Numerous paradigms have been developed for EEG-based BCI systems based on decoding of specific neural events in the EEG. Examples of these neural events include the event-related potential (ERP) Jin, Xu, Daly, Zhao, Wang and Cichocki (2024), the steady-state visual evoked potential (SSVEP) Jin, Wang, Xu, Liu, Wang and Cichocki (2021) and Motor Imagery (MI) Arpaia, Esposito, Natalizio and Parvis (2022a). MI is a spontaneous BCI paradigm, requiring the BCI user to imagine the movement of body parts instead of perform actual action Padfield, Zabalza, Zhao, Masero and Ren (2019), and is one of the most popular BCI paradigms Duan, Li, Ji, Pang, Zheng, Lu, Li and Zhuang (2020); Barmpas, Panagakis, Bakas, Adamos, Laskaris and Zafeiriou (2023). However, the recognition of MI related brain activity is challenging due to the limited, noisy, and non-stationary of EEG signal properties, which makes design of effective decoding algorithms highly challenging Wang et al. (2023).

Over recent years, the most popular methods for extracting features of EEG activity related to motor imagery is the common spatial pattern (CSP) Müller-Gerking, Pfurtscheller and Flyvbjerg (1999) algorithm and variants of it Jin, Xiao, Daly, Miao, Wang and Cichocki (2020), such as the Filter-Bank CSP (FBCSP) algorithm Ang, Chin, Zhang and Guan (2008), which operates in the frequency-domain Arpaia et al. (2022a), Barmpas et al. (2023). These methods have achieved excellent results in binary classification by using classifiers, for example support vector machines (SVM) Hearst, Dumais, Osuna, Platt and Scholkopf (1998) or linear

*Corresponding author

(J. Jin)

jinjingat@gmail.com (J. Jin); (J. Jin)

ORCID(s):

[1] Email: jinjingat@gmail.com

discriminant analysis (LDA) Wu, Wu, Pal, Chen, Chen and Lin (2013). Furthermore, some methods also use channel selection to eliminate channels containing redundant information and noise in order to gain further improvements in single user decoding accuracies Jin, Miao, Daly, Zuo, Hu and Cichocki (2019), Xiao, Huang, Xu, Wang, Wang and Jin (2022).

MI classification networks are inspired by computer vision (CV). Various network structures, including convolutional neural networks (CNN) Schirrmeister, Springenberg, Fiederer, Glasstetter, Eggensperger, Tangermann, Hutter, Burgard and Ball (2017); Lawhern, Solon, Waytowich, Gordon, Hung and Lance (2018); Wang et al. (2023), recurrent neural networks (RNN) Arpaia, Esposito, Natalizio and Parvis (2022b) with sequential signal processing mechanisms, and Transformers with self-attention mechanisms Song, Zheng, Liu and Gao (2022), are increasingly applied to the challenge of MI classification and achieve good performance. Compared to methods such as CSP, the network automates feature extraction, reducing the dependence on manual design. Although emerging methods, particularly deep learning, have introduced novel structural designs and data processing techniques, conventional approaches still influence network structure design. For instance, ShallowNet utilizes neural network to achieve FBCSP-like log-variance calculations, aimed at decoding band power features Schirrmeister et al. (2017). However, the direct application of network architectures and mechanisms from CV and natural language processing (NLP) to MI signal recognition does not necessarily improve accuracy and may even decrease it Schirrmeister et al. (2017); Song et al. (2022).

Unlike the deep convolutional neural networks used in CV and NLP applications, the MI-EEG classification networks generally benefit from shallow networks, such as ShallowNet, which outperform deeper networks in MI classification Schirrmeister et al. (2017). In the EEGNet architecture, depthwise separable convolution is incorporated to achieve a compact structure Lawhern et al. (2018); Chollet (2017). However, lightweight CNNs are limited by their fewer convolutional layers and the processing of MI signals only in a single frequency band, which restricts both model scale and data feature representation, making them prone to performance bottlenecks Chen, Dai, Liu, Chen, Yuan and Liu (2020).

EEG signals convey information through various frequency components Ko, Jeon, Jeong and Suk (2021). During MI, task-specific event-related desynchronization (ERD) and event-related synchronization (ERS) events occur in sensory-motor rhythms within specific frequency bands and brain regions Wang et al. (2023). Generally, the main MI rhythms are localized in the μ band (7-13Hz) and β band (13-30Hz) McFarland, Miner, Vaughan and Wolpaw (2000). For instance, in FBCNet, a filter bank is applied to decompose MI data into distinct frequency bands, which are then processed through individual convolutional branches Mane, Chew, Chua, Ang, Robinson, Vinod, Lee and Guan (2021). This approach is also utilized in FBMSNet, and TSFCNet Liu, Yang, Yu, Wang and Wu (2022); Zhi, Yu, Yu, Gu and Yang (2023). Unlike the filter bank design, the Interactive Frequency Convolutional Neural Network (IFNet) model employs the concept of cross-frequency coupling to process MI data using two frequency bands Wang et al. (2023). Although γ band ($\geq$ 30Hz) generally cannot reach the scalp with sufficient integrity to be recorded via EEG with a good signal-to-noise ratio, making it difficult to use for MI activity classification Deng, Zhang, Yu, Liu and Sun (2021), high-frequency features have been shown to positively impact the generalization ability of network classification models for MI activities as network model performance continues to break through bottlenecks Liang et al. (2023a); Luo, Mao, Wang, Shi and Hei (2022). However, MI recognition is constrained by noise interference during EEG acquisition and limited dataset sizes, presenting significant challenges to MI classification attempts.

In contrast to the large-scale annotated datasets available for CV or NLP, the deficiency of publicly available datasets impedes the classification of MI tasks Liang et al. (2023a), Arpaia et al. (2022a), Deng, Dong, Socher, Li, Li and Fei-Fei (2009), Radford, Narasimhan, Salimans, Sutskever et al. (2018), Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick (2014). EEG data-gathering, as used for MI, is much more challenging, entailing long experimental sessions to record neural data from participants and high experimental costs. Furthermore, the EEG signal properties results in large degrees of variability between participants and within a participant across experimental sessions Liang et al. (2023a), Tangermann, Müller, Aertsen, Birbaumer, Braun, Brunner, Leeb, Mehring, Miller, Müller-Putz et al. (2012a). Consequently, the majority of MI datasets contain data from only a few participants. For example, the BCI Competition Datasets, that are widely used in BCI research, include EEG data from a relatively small number of participants. Furthermore, the feature distributions from these EEG signals are inconsistent across participants Arpaia et al. (2022a), Tangermann et al. (2012a), Blankertz, Muller, Krusienski, Schalk, Wolpaw, Schlogl, Pfurtscheller, Millan, Schroder and Birbaumer (2006). Few-shot learning may be applied to these small datasets with the aim of extracting additional feature information across participants to enhance classification results, but the outcomes encounter challenges, such as limitations in feature distributions, that are insurmountable compared to single-participant tasks.

To address the issue of insufficient individual EEG data, Transfer Learning has been introduced, aiming to enhance model generalization performance by learning discriminative signal features across different subjects Wu, Jiang and Peng (2022). However, the feature distributions derived from these EEG signals exhibit significant inconsistency across subjects. EEG combined with other modal signals enriches the brain activity data captured at the same moment, for example, by introducing functional near-infrared spectroscopy (fNIRS), which enhances the spatiotemporal resolution Wang, Yuan, Zhang, Wan, Li and Xu (2025);

Xu, Zhou, Yang, Li, Li, Bezerianos and Wang (2023); Li, Sun, Wan, Yuan, Jung and Wang (2025). However, it still faces challenges such as limited individual data and other issues. Furthermore, some researchers have adopted large model frameworks, such as EEGPT Wang, Liu, He, Xu, Ma and Li (2024) and LaBraM Jiang, Zhao and Lu (2024), for cross-dataset task recognition by aggregating EEG data from diverse tasks. However, their performance currently falls short of State-of-the-Art (SOTA) models. Moreover, designing effective decoding algorithms for these small and highly variable datasets is a considerable challenge Liang et al. (2023a), Wang et al. (2023).

With these challenges in mind, in this paper, we propose a novel CNN, the multi-scale attention convolutional neural network (MSAttNet), to improve MI classification performance. Our MSAttNet model includes three main components: a multi-band segmentation module, an attention spatial convolution module, and a multi-scale temporal convolution module. First, the multi-band segmentation module adopts filter banks with overlapping frequency bands to increase the number of channels in the MI signal and enhance band features in the frequency domain. Then, the attention spatial convolution module is used to adaptively adjust different convolutional kernel parameters according to the input through the attention mechanism to capture the important features of the EEG data. The outputs of the attention spatial convolution module are grouped to perform multi-scale temporal convolution. Finally, the multi-scale temporal convolution module uses a variance operation to extract temporal features and to perform noise elimination. The extracted features are then classified. The main contributions of this work are summarized as follows:

- We propose a new method, a multi-scale attention convolutional neural network (MSAttNet). Our method achieves state-of-the-art performance across multiple datasets, recording an accuracy of 78.20% on the IV2a dataset, 84.52% on the IV2b dataset, 75.94% on the OpenBMI dataset, and 78.60% on the ECUST-MI dataset. All other metrics—including Precision, Recall, F1-score, and Kappa coefficient—also reach state-of-the-art levels, demonstrating robustness to inter-subject variability and differing channel configurations.

- We employ a multi-band segmentation module. By setting overlapping frequency band divisions (4-16Hz, 12-24Hz, 20-36Hz, 32-44Hz, 40-100Hz) and utilizing optimized convolutional kernel sizes (63, 31, 15, 7, 3) in the multi-scale temporal convolution, effective frequency-domain feature extraction is achieved.

- We construct an attention spatial convolution module, which automatically selects the most suitable convolutional kernel weights without manual parameter adjustment, adapting to different subjects across various datasets.

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 presents our proposed method. Section 4 presents the experimental results. Section 5 presents a discussion of the results. Finally, Section 6 describes the conclusions from our work.

2. Method

Our proposed network, MSAttNet, is constructed with the intention of effectively extracting spectro-spatial features from the EEG by diverse convolutional kernels based on an attention mechanism. This allows the model to avoid the need for manual parameter adjustment for different MI signals for different participants and datasets. The MSAttNet includes the multi-band segmentation module, the attention spatial convolution module, and the multi-scale temporal convolution module, and the classifier module, shown in Figure 1. The parameter of the MSAttNet structure is listed in Table 1.

2.1. Multi-Band Segmentation Module

A single trial of raw MI-EEG signals are denoted as $(x_i, y_i), i = 1, 2, \cdots, n$, where $x_i \in \mathbb{R}^{C\times T}$, $y_i \in \{1, 2, \cdots, N_c\}$, with x_i representing MI-EEG data, y_i the trial labels, C the number of channels, T the number of sample points, and N_c the number of motor imagery tasks. So, (X, Y) is input n trials of MI-EEG data, where $X = [x_1, x_2, \cdots, x_n]^T$ and $Y = [y_1, y_2, \cdots, y_n]^T$.

In FBCSP and its variants, the filter bank uses a frequency band from 4 to 40 Hz to construct 9 filters (4-8, 8-12, . . . , 36-40 Hz) Ang et al. (2008). A multi-scale convolutional transformer model uses 4-60Hz and 4-120Hz frequency bands Ahn, Lee, Jeong and Lee (2022). To capture more feature information, the MI-EEG signals $(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{T})$ are then transformed by the multi-band segmentation module, where $\boldsymbol{B}$ represents batch trials. The multi-band segmentation using a filter bank reconstructs single-frequency data into various frequency bands for data augmentation, enhancing feature extraction efficiency Wang et al. (2023). The filter bank $F = \{f_i\}_{i=1}^{N_b}$ uses N_b fifth-order Butterworth filters $f_i \in \mathbb{R}^{(N_b\times C)\times T}$ in overlapping bands, spanning 4-16Hz, 12-24Hz, 20-36Hz, 32-44Hz, and 40-100Hz, as shown in the Figure 1. The filtered MI data output becomes:

$$X_{\mathrm{FB}} = \mathcal{F}_{\mathrm{FB}}(X), \quad X_{FB} \in \mathbb{R}^{(N_b\times C)\times T} \tag{1}$$

with $N_b = 5$ narrow-band temporal filters. The channel dimension of the MI data increases from $(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{T})$ to $(\boldsymbol{B}, 5\boldsymbol{C}, \boldsymbol{T})$.

2.2. Attention Spatial Convolution Module

Generally, the first two layers of the network structure habitually use temporal or spatial convolutional layers to realize the convolution of channels and sample points, such as EEGNet, and EEGsym Lawhern et al. (2018); Pérez-Velasco, Santamaría-Vázquez, Martínez-Cagigal, Marcos-Martínez and Hornero (2022). In our proposed network, the first layer of the structure discards the temporal or spatial

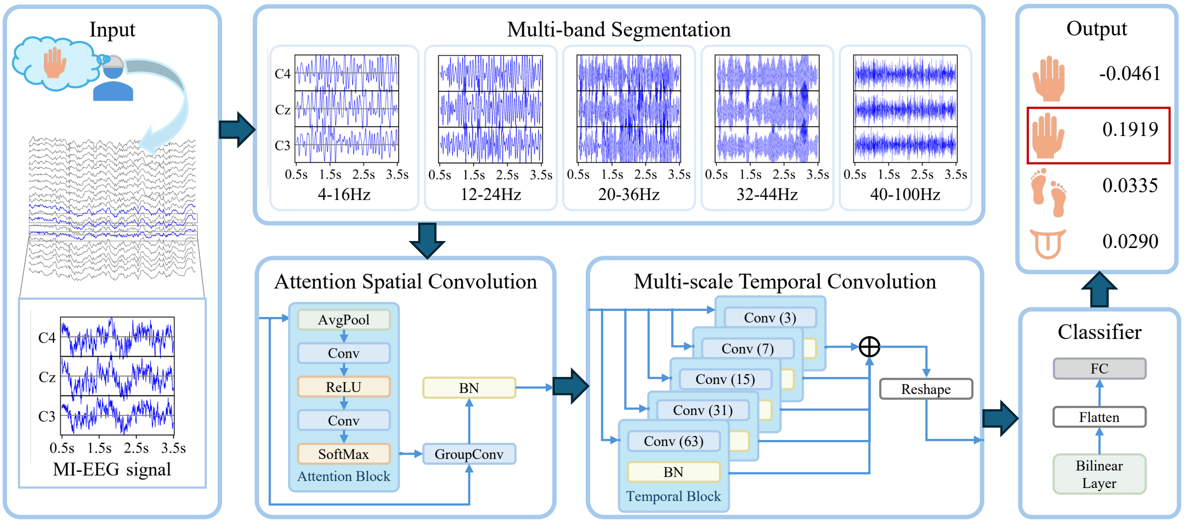


Figure 1: The structure of MSAttNet

Table 1
The parameters of our proposed method MSAttNet

Module	Layer	Kernel Size	Output
Multi-band Segmentation	Input Filter Bank		(B, C, T) $(B, 5C, T)$
Attention Spatial Convolution	Attention Block Group Convolution BatchNorm	(1)	(B, K) $(B, 5C_1, T)$ $(B, 5C_1, T)$
Multi-scale Temporal Convolution	Depth-wise Conv1D BatchNorm1D Sum Reshape	$(63), (31), (15), (7), (3)$	$5(B, C_1, T)$ $5(B, C_1, T)$ (B, C_1, T) $(B, C_1, N_p, T/N_p)$
Classifier	Bilinear Pooling Flatten Fully connected		(B, C_1, N_p) $(B, C_1 N_p)$ (B, N_c)

B is the batch size of the EEG trials, C is the number of the EEG channels, T is the number of EEG sample points, K is the number of convolutional kernels in the attention spatial convolution module, C_1 denotes the number of channels output from the Attention Convolution, and N_p is the patch size in the bilinear Pooling layer.

convolutional layer, and attention spatial convolution is substituted to capture features that are common across different datasets. The attention spatial convolution module uses K parallel convolutional kernels instead of the single kernel, continuously adjusting the weights of each kernel to select the optimal kernel size for feature extraction. The attention spatial convolution module consists of the attention block, the group convolution layers, and batch normalization layers. The attention block calculates the attention weights for each kernel in the group convolutions. It includes an average pooling layer, two convolution layers, a ReLU activation layer, and a SoftMax activation layer. The calculation process is as follows.

First, compression along the temporal dimension T is performed using the average pooling layer to capture the global information of each channel:

$$x_{\mathrm{ap}}^{c} = F_{\mathrm{AP}}\left(X_{\mathrm{FB}}^{(i,c)}\right) = \frac{1}{T}\sum_{i=1}^{T} X_{\mathrm{FB}}^{(i,c)}, \quad c = 1, 2, \dots, C_b \quad (2)$$

The channels can be expressed as $X_{\mathrm{AP}} = \left[x_{\mathrm{ap}}^{1}, x_{\mathrm{ap}}^{2}, \dots, x_{\mathrm{ap}}^{c}\right]^{T} \in \mathbb{R}^{C_b \times 1}$, where $C_b = N_b \times C$. A two-layer convolution layer is applied to X_{AP}, using activation functions to obtain the attention weights W_{GC}:

$$W_{\mathrm{GC}} = \mathcal{F}_{\sigma} \circ \mathcal{F}_{\mathrm{IC}} \circ \mathcal{F}_{\delta} \circ \mathcal{F}_{\mathrm{RC}}(X_{\mathrm{AP}}) \quad (3)$$

where $\mathcal{F}_{\mathrm{RC}}$ represents the dimensionality reduction convolution layer, $\mathcal{F}_{\delta}$ denotes the ReLU layer, $\mathcal{F}_{\mathrm{IC}}$ represents the dimensionality expansion convolution layer, and $\mathcal{F}_{\sigma}$ denotes the Softmax layer.

The attention weights W_{GC} are then applied to each convolutional kernel as follows:

$$X_{\mathrm{GC}}^{(g)}(C_g) = \mathcal{F}_{\mathrm{GC}}\left(W_{\mathrm{GC}} X_{\mathrm{FB}}^{C_b}\right) \tag{4}$$

where $C_g = \dfrac{C_b}{N_g}$ is the number of output channels after dividing the input channels C_b into N_g groups, and $X_{\mathrm{GC}}^{(g)}(C_g), g = 1, 2, \dots, N_g$ is the output signal of each group processed by the dynamic weighted convolution.

Additionally, considering that the number of input channels is too small and significant spatial information between channels is lost, channel expansion is introduced in Eq. (2),(3), and (4). This expands the low-dimensional compressed representation of the original MI-EEG signal into a higher-dimensional space, enhancing the information dimensions processed by subsequent layers. Finally, $X_{\mathrm{GC}}^{(g)}$ is processed through the batch normalization layer to obtain $X_{\mathrm{BN}}^{(g)}$, which is then fed into the multi-scale temporal convolution module.

2.3. Multi-scale Temporal Convolution Module

In the Multi-scale Temporal Convolution Module, we employ a multi-branch architecture where the number of branches corresponds to the number of frequency band divisions. Each signal group is processed by convolutional layers with distinct kernel sizes: specifically, 4-16 Hz band utilizes a kernel size of 63, 12-24 Hz employs kernel size 31, 20-36 Hz uses kernel size 15, 32-44 Hz applies kernel size 7, and 40-100 Hz operates with kernel size 3. Each convolutional layer is succeeded by a Batch Normalization layer, collectively constituting the temporal block as expressed by:

$$X_{\mathrm{TB}}^{(g)} = \mathcal{F}_{\mathrm{BN}}^{(g)} \circ \mathcal{F}_{\mathrm{C}}^{(g)}\left(X_{\mathrm{BN}}^{(g)}\right) \tag{5}$$

Subsequently, the $X_{\mathrm{TB}}^{(g)}$ signals from each group are concatenated along the channel dimension, transforming from 5 groups of (B, C_1, T) into a single group $X_{\mathrm{S}} = (B, C_1, T)$. Finally, to align with the signal dimensions of the Bilinear layer, X_{S} is reshaped into $X_{\mathrm{R}} = \left(B, C_1, N_p, T/N_p\right)$ and fed into the classifier module.

2.4. Classification Module

Most networks use pooling layers at the end of a CNN for feature extraction and dimensionality reduction. However, the features derived from the first two modules typically contain substantial intra-class variance and high noise levels. Considering that MI-EEG activity exhibits distinct spectral power, variance computation—which reflects the spectral power of a given time series—is a more suitable choice for representing EEG temporal characteristics. Therefore, we apply a bilinear layer to enhance temporal features and reduce noise. This layer is defined as:

$$X_V = F_V\left(X_R(t)\right) = \frac{1}{T}\sum_{t=0}^{T-1}\left(X_R(t) - \bar{X}_R(t)\right)^2 \tag{6}$$

Algorithm 1 MSAttNet

```
Require: EEG trials $X \in \mathbb{R}^{B\times C\times T}$
Ensure: Predicted labels $\hat{Y} \in \mathbb{R}^{B}$
// Multi-band Segmentation
Define $F \leftarrow [4,16],[12,24],[20,36],[32,44],[40,100]$
for each $f_l, f_h$ in $F$ do                                  ▷ Filter Bank
    $X_{\mathrm{FB}} \leftarrow X_{\mathrm{FB}} \cup \mathrm{Butterworth}(X, \mathrm{band} = (f_l, f_h)) \in \mathbb{R}^{B\times 5C\times T}$      ▷ Eq. (1)
end for
// Attention Spatial Convolution
$X_{\mathrm{AP}} \leftarrow \mathrm{AvgPool1D}(X_{\mathrm{FB}}) \in \mathbb{R}^{B\times 5C\times 1}$      ▷ Eq. (2)
$W_{\mathrm{GC}} \leftarrow \mathrm{SoftMax}(\mathrm{Conv1D}(\mathrm{ReLU}(\mathrm{Conv1D}(X_{\mathrm{AP}})))) \in \mathbb{R}^{K}$      ▷ Eq. (3)
$X_{\mathrm{GC}} \leftarrow \mathrm{GroupConv1D}(W_{\mathrm{GC}} X_{\mathrm{FB}}) \in \mathbb{R}^{B\times 5C_1\times T}$      ▷ Eq. (4)
$X_{\mathrm{BN}} \leftarrow \mathrm{BatchNorm}(X_{\mathrm{GC}})$
// Multi-scale Temporal Convolution
Define $G \leftarrow (63, 31, 15, 7, 3)$
for each $g$ in $G$ do                  ▷ Process each frequency band
    $X_{\mathrm{TB}}^{(g)} \leftarrow \mathrm{DepthwiseConv1D}(X_{\mathrm{BN}}^{(g)}, \mathrm{kernel} = g)$
    $X_{\mathrm{TB}}^{(g)} \leftarrow \mathrm{BatchNorm1D}(X_{\mathrm{TB}}^{(g)})$      ▷ Eq. (5)
end for
$X_{\mathrm{S}} \leftarrow \sum_g X_{\mathrm{TB}}^{(g)} \in \mathbb{R}^{B\times C_1\times T}$
$X_{\mathrm{R}} \leftarrow \mathrm{reshape}(X_{\mathrm{S}}, [B, C_1, N_{\mathrm{p}}, T/N_{\mathrm{p}}])$
// Classifier Module
$X_{\mathrm{V}} \leftarrow \mathrm{BilinearPool}(X_{\mathrm{R}}) \in \mathbb{R}^{B\times C_1\times N_{\mathrm{p}}}$      ▷ Eq. (6)
$X_{\mathrm{flat}} \leftarrow \mathrm{Flatten}(X_{\mathrm{V}}) \in \mathbb{R}^{B\times(C_1\cdot N_{\mathrm{p}})}$
$\hat{Y} \leftarrow \mathrm{FC}(X_{\mathrm{flat}}) \in \mathbb{R}^{B\times N_{\mathrm{c}}}$      ▷ Eq. (7)
    retuen $\hat{Y}$
```

where $X_R(t)$ is the signal processed by the first two modules, $\bar{X}_R(t)$ is the mean of $X_R(t)$, and t represents the sampling time point.

The resulting signal is then passed through a dropout layer for regularization and subsequently flattened using $\mathcal{F}_{FL}$. Finally, the output is passed to a fully connected layer $\mathcal{F}_{FC}$ to produce the final result:

$$Y = \mathcal{F}_{FC} \circ \mathcal{F}_{FL}(X_V) \tag{7}$$

The procedure of the proposed MSAttNet method is illustrated in Algorithm 1, where each step corresponds to its relevant formula locations.

3. Experiments

3.1. Experimental setting

We use four datasets, including The BCI Competition IV Dataset IIa (IV2a) dataset Tangermann, Müller, Aertsen, Birbaumer, Braun, Brunner, Leeb, Mehring, Miller, Müller-Putz et al. (2012b), the BCI Competition IV Dataset IIb (IV2b) dataset Tangermann et al. (2012b), the OpenBMI dataset Lee, Kwon, Kim, Kim, Lee, Williamson, Fazli and Lee (2019), and the East China University of Science and Technology Motor Imagery (ECUST-MI) dataset, to evaluate our proposed method.

The IV2a dataset includes left hand, right hand, both feet, and tongue. The MI-EEG data of 9 healthy participants are recorded by 22 electrodes sampled at 250 Hz. For each participant, the first session is the training set, and the second session is the test set. Each session contains 288 trials with 72 trials per class.

The IV2b dataset consists of left hand and right hand. The MI-EEG data is recorded from 9 healthy participants with 3 electrodes sampled at 250 Hz. For each participant, three sessions are training set, and the last two sessions are test set. Each session contains 120 trials.

The OpenBMI dataset contains EEG signals of 62 channels recorded from 54 healthy subjects. The experiments involve balanced left-hand and right-hand MI tasks, and 100 trials for both the training and testing phases. Each trial lasts 4 seconds, and the data are downsampled to 250 Hz. All trials on the IV2a dataset, the IV2b dataset, and the OpenBMI dataset use data from 0.5 to 3.5 seconds.

The ECUST-MI dataset is approved by the East China University of Science and Technology (ECUST-2022-054). Eleven healthy participants (9 males, and 2 females), aged between 25 and 28 years, participated in the experiments. One subject was excluded due to insufficient trials. The MI-EEG data are recorded from 16 electrodes (FC5, FC1, FCz, FC2, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP1, CP2, and CP6) and sampled at 600 Hz. Each subject is required to execute two MI tasks, which are the left hand and right hand. The tasks have 4 sessions, the interval of which is 5 minutes, and each session has 30 trials. The first two sessions are training sets, and the last two sessions are test sets. At the beginning of each trial, a cross in the center of the screen appears for 2s and reminds the subject to concentrate on MI task. In 2s, a left or right arrow appears for 3s and reminds the subject to image the left or right-hand motor. Then a blank interface of 2s in a solid color guides the subjects to rest for 10s and prepare for the next trial. All trials of the ECUST-MI datasets use 3s from the beginning cue of MI tasks.

We employ cross-entropy loss together with the Adam optimizer to update all model parameters during training Kingma and Ba (2014). The learning rate of Adam optimizer is 0.001. The number of training epochs is set to 1000 for all models Wang et al. (2023). Therefore, a batch size of 256 to reduce the influence of the mini-batch size with the increased error is used for all models Ioffe and Szegedy (2015); Wu and He (2018). Increasing the batch size within the same epoch reduces training time and promotes stable gradient descent, thereby enhancing network training effectiveness. Note that, when the number of trials in the training set is less than the desired batch size, the training batch size is set to the number of trials. We use a NVIDIA RTX4090 GPU and an Intel i9-13900KF processor, and 32GB of RAM for training and testing.

3.2. Comparison with State-of-the-Art Methods

We compare MSAttNet with 12 models, including ShallowNet Schirrmeister et al. (2017), DeepNet Schirrmeister et al. (2017), EEGNet Lawhern et al. (2018), FBCNet Mane et al. (2021), EEGConformer Song et al. (2022), FBMSNet Liu et al. (2022), IFNet Wang et al. (2023), TSFCNet Zhi et al. (2023), ADFCNN Tao, Wang, Wong, Jia, Li, Chen, Chen and Wan (2023), EISATCFusion Liang, Cao, Wang, Zhang and Wu (2024), EEGSimpleConv El Ouahidi, Gripon, Pasdeloup, Bouallegue, Farrugia and Lioi (2024), and DMSACNN Liu, Xing, Yang, Yu, Xiao, Wang and Wu (2025), on four datasets, including the IV2a dataset, the IV2b dataset, the OpenBMI dataset, and the ECUST-MI dataset, to evaluate the effectiveness of our proposed method. The experimental evaluation metrics adopted include accuracy (Acc.), precision (Prec.), recall (Rec.), F1-score (F1), and Cohen's kappa coefficient (Kappa) Liang, Yu, Liu, Wang, Liu and Dong (2023b). The p-value (p) of t-test is used to examine whether the proposed method achieves a statistically significant improvement in accuracy. The results are shown in Table 2. To ensure comparability, all models adopt their original hyperparameters except for necessary input/output adjustments. Input trials are standardized to 3s segments, and the final fully connected layer dimensions are modified to match dataset-specific class counts.

As shown in Table 2, the proposed MSAttNet achieves statistically significant improvements over other state-of-the-art (SOTA) methods across all four datasets. It attains an accuracy of 78.20% on the IV2a dataset, 84.52% on IV2b, 75.94% on OpenBMI, and 78.60% on the ECUST-MI dataset. The method also achieves SOTA performance in all four metrics, including precision, recall, F1-score, and Kappa. Although some networks perform similarly to ours on the IV2b dataset, our proposed method exhibits less sensitivity to inter-subject variability across different datasets, maintaining stable optimal performance even in scenarios with a large number of subjects and channels, such as on the OpenBMI dataset. This advancement is attributed to the network's optimized lightweight hierarchical architecture, where the filter banks are better aligned with the neurophysiological characteristics of motor imagery tasks. The synergistic integration of these components enables efficient extraction of discriminative spatial-spectral features while maintaining computational efficiency, significantly outperforming traditional methods in motor imagery EEG feature representation.

4. Discussion

4.1. Different Bands

The proposed methods, MSAttNet, employs different frequency bands, which can have varying impacts on the network's recognition performance. We first consider the setting without high frequencies (4-16Hz, 12-24Hz, 20-36Hz, 32-40Hz) and without overlapping frequency bands (4-12Hz, 12-20Hz, 20-32Hz, 32-40Hz, 40-100Hz). We reference existing frequency band segmentation methods, including the single-band setting (4-40Hz) used in EEGNet Lawhern et al. (2018) and the filter bank methods used in

Table 2

Comparisons of our proposed method, MSAttNet, and 12 models, on four datasets, including the IV2a, IV2b, OpenBMI, and ECUST-MI dataset. The experimental evaluation metrics include accuracy (Acc., %), precision (Prec., %), recall (Rec., %), F1-score (F1, %), kappa, and p-value (p). The *, **, and *** indicate that the accuracies of MSAttNet are significantly higher than the compared methods with $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

Method	Acc.	Prec.	Rec.	F1	Kappa	p
ShallowNet	63.04	63.09	63.04	62.68	0.5072	***
DeepNet	55.79	56.22	55.79	54.80	0.4105	***
EEGNet	60.92	61.53	60.92	60.77	0.4789	***
FBCNet	74.69	75.27	74.69	74.40	0.6626	*
EEGConformer	57.14	57.28	57.14	56.86	0.4285	***
FBMSNet	73.80	74.32	73.80	73.36	0.6507	**
IFNet	74.46	75.26	74.46	74.22	0.6595	*
TSFCNet	71.64	72.34	71.64	71.41	0.6219	*
ADFCNN	66.36	66.74	66.36	65.89	0.5514	**
EISATCFusion	63.81	64.26	63.81	63.51	0.5175	***
EEGSimpleConv	65.08	67.59	65.08	63.88	0.5345	**
DMSACNN	74.42	74.62	74.42	74.00	0.6589	*
MSAttNet	**78.20**	**78.69**	**78.20**	**78.02**	**0.7094**	-

(a) The results on the IV2a dataset

Method	Acc.	Prec.	Rec.	F1	Kappa	p
ShallowNet	81.27	81.74	81.26	81.10	0.6253	*
DeepNet	81.46	82.06	81.46	81.10	0.6293	***
EEGNet	78.33	78.77	78.33	78.24	0.5667	0.06
FBCNet	80.40	80.80	80.40	80.24	0.6080	***
EEGConformer	80.73	81.38	80.73	80.55	0.6147	0.19
FBMSNet	80.35	80.46	80.35	80.30	0.6070	***
IFNet	81.21	81.71	81.21	81.10	0.6242	***
TSFCNet	80.31	80.68	80.31	80.01	0.6062	***
ADFCNN	82.45	83.35	82.45	82.25	0.6489	0.49
EISATCFusion	81.13	81.81	81.13	80.97	0.6225	0.15
EEGSimpleConv	80.21	80.44	80.21	80.16	0.6043	**
DMSACNN	80.07	80.75	80.07	79.92	0.6015	***
MSAttNet	**84.52**	**84.81**	**84.52**	**84.44**	**0.6905**	-

(b) The results on the IV2b dataset

Method	Acc.	Prec.	Rec.	F1	Kappa	p
ShallowNet	65.98	66.24	65.98	65.11	0.3196	***
DeepNet	68.91	70.72	68.91	67.51	0.3781	***
EEGNet	70.31	70.77	70.31	69.96	0.4063	**
FBCNet	72.46	73.47	72.46	71.56	0.4493	***
EEGConformer	71.17	71.69	71.17	70.73	0.4233	*
FBMSNet	70.76	71.59	70.76	69.97	0.4152	***
IFNet	74.67	75.92	74.67	73.90	0.4933	*
TSFCNet	74.62	75.11	74.62	74.28	0.4924	*
ADFCNN	72.31	73.04	72.31	71.84	0.4461	*
EISATCFusion	70.74	71.12	70.74	70.57	0.4148	**
EEGSimpleConv	66.24	67.47	66.24	65.26	0.3248	***
DMSACNN	72.65	73.86	72.65	71.97	0.4530	***
MSAttNet	**75.94**	**76.50**	**75.94**	**75.68**	**0.5189**	-

(c) The results on the OpenBMI dataset

Method	Acc.	Prec.	Rec.	F1	Kappa	p
ShallowNet	65.79	67.21	65.79	64.74	0.3159	***
DeepNet	67.47	71.25	67.42	65.11	0.3486	**
EEGNet	67.68	68.37	67.66	67.28	0.3533	*
FBCNet	71.67	72.24	71.67	71.38	0.4332	**
EEGConformer	70.87	71.10	70.87	70.75	0.4174	*
FBMSNet	69.64	70.29	69.66	69.20	0.3930	***
IFNet	75.23	75.94	75.26	75.02	0.5049	*
TSFCNet	74.06	74.90	74.05	73.77	0.4809	***
ADFCNN	71.55	73.13	71.51	71.05	0.4305	*
EISATCFusion	66.83	67.03	66.85	66.72	0.3369	*
EEGSimpleConv	65.10	67.86	65.11	62.02	0.3022	**
DMSACNN	73.37	74.10	73.39	73.01	0.4676	**
MSAttNet	**78.60**	**79.04**	**78.60**	**78.50**	**0.5721**	-

(d) The results on the ECUST-MI dataset

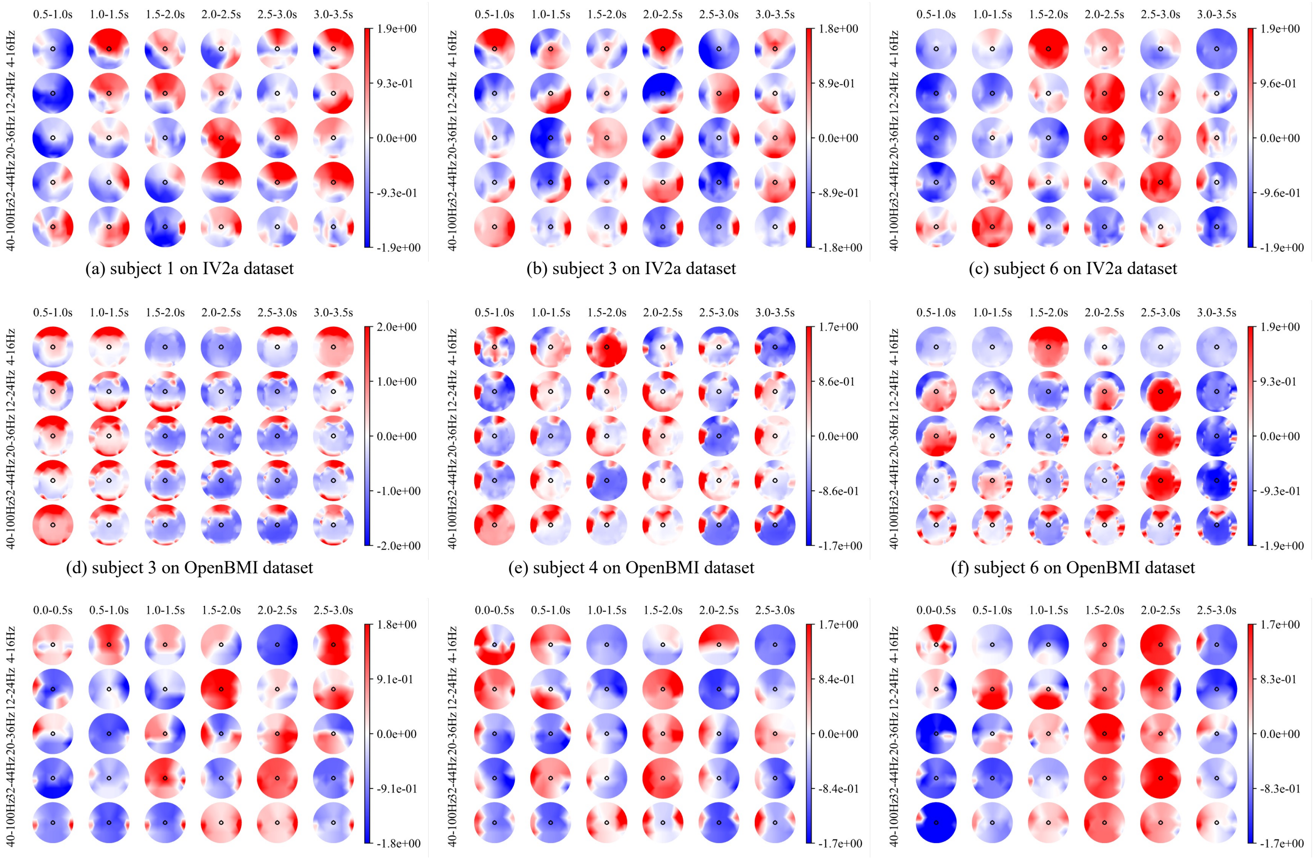
FBCNet Mane et al. (2021), which divide the bands into (4-8Hz, 8-12Hz, 12-16Hz, 16-20Hz, 20-24Hz, 24-28Hz, 28-32Hz, 32-36Hz, 36-40Hz). We introduce frequency band settings mentioned in IFNet Wang et al. (2023), including the (4-16Hz, 16-40Hz), (4-8Hz, 8-16Hz, 16-40Hz) and (4-8Hz, 8-16Hz, 16-30Hz, 30-40Hz). Finally, we consider the impact of starting the segmentation from higher frequencies on the network's performance, so we design three reverse filter bank methods, including (16-40Hz, 4-16Hz), (40-100Hz, 32-44Hz, 20-36Hz, 12-24Hz, 4-16Hz) and (36-40Hz, 32-36Hz, 28-32Hz, 24-28Hz, 20-24Hz, 16-20Hz, 12-16Hz, 8-12Hz, 4-8Hz). The other settings of the network remain unchanged.

As shown in Table 3, in the proposed network, the accuracy and Kappa exhibit an oscillating downward trend across the four datasets when using different frequency band divisions. We also observe that reversing the filter bank division—(16-40Hz, 4-16Hz) versus (4-16Hz, 16-40Hz)—yields comparable performance, but the two methods of 9-band division produce significantly different results. When using a network without the 40-100Hz frequency band, the performance is significantly lower than that of MSAttNet, and we find that overlapping frequency bands help improve the network's recognition performance. This set of experiments demonstrates the effectiveness of the specific frequency division filter banks we employ.

Based on the cortical topography maps presented in Figure 2 for different subjects from IV2a, OpenBMI, and ECUST-MI datasets, distinct MI time periods across frequency bands are observed. This finding aligns with the frequency plot in Figure 1, which reveals differential activation frequencies at distinct time points in the subject. This phenomenon likely reflects collaborative interactions among brain regions through oscillatory coupling mechanisms.

Table 3
The comparison of accuracy (%, Acc) and Kappa results between different frequency bands using MSAttNet on four datasets.

Dataset	IV2a		IV2b		OpenBMI		ECUST-MI	
Metric	Acc.	Kappa	Acc.	Kappa	Acc.	Kappa	Acc.	Kappa
4-16, 12-24, 20-36, 32-40	74.34	0.6579	82.52	0.6505	69.68	0.3935	66.68	0.3336
4-12, 12-20, 20-32, 32-40, 40-100	77.04	0.6939	83.48	0.6696	72.92	0.4583	73.07	0.4610
4-40	73.69	0.6492	82.73	0.6547	68.50	0.3700	67.18	0.3436
4-16, 16-40	75.85	0.6780	82.91	0.6582	69.40	0.3880	65.15	0.3031
4-8, 8-16, 16-40	72.18	0.6291	81.11	0.6222	66.00	0.3200	64.50	0.2899
4-8, 8-16, 16-30, 30-40	73.46	0.6461	82.12	0.6424	67.13	0.3426	63.11	0.2624
4-8, 8-12, 12-16, 16-20, 20-24, 24-28, 28-32, 32-36, 36-40	65.70	0.5427	80.85	0.6171	65.58	0.3117	62.63	0.2525
16-40, 4-16	76.04	0.6806	81.86	0.6372	67.71	0.3543	67.94	0.3588
40-36, 36-32, 32-28, 28-24, 24-20, 20-16, 16-12, 12-8, 8-4	72.69	0.6358	81.25	0.6250	65.96	0.3193	64.95	0.2995
4-8, 8-12, 12-16, 16-20, 20-24, 24-28, 28-32, 32-36, 36-40	64.16	0.5221	79.35	0.5870	60.83	0.2167	58.10	0.1604
4-16, 12-24, 20-36, 32-44, 40-100 (MSAttNet)	78.20	0.7094	84.52	0.6905	75.94	0.5189	78.60	0.5721



(a) subject 1 on IV2a dataset

(b) subject 3 on IV2a dataset

(c) subject 6 on IV2a dataset

(d) subject 3 on OpenBMI dataset

(e) subject 4 on OpenBMI dataset

(f) subject 6 on OpenBMI dataset

(g) subject 5 on ECUST-MI dataset

(h) subject 9 on ECUST-MI dataset

(i) subject 10 on ECUST-MI dataset

Figure 2: Topographic Maps from different subject on three datasets over different times and frequencies.

4.2. Different Kernal Size on Multi-Scale Temporal Convolution Module

We select different kernel sizes to evaluate the performance of the multi-scale temporal convolution module. First, three kernel sizes—31, 63, and 127—are tested. For example, when the size 31 is chosen, all kernels in the network are uniformly set to 31. Additionally, we design a configuration scheme starting with sizes 31 and 127. Taking 31 as an example, the kernel sizes in Table 1 are modified to (31, 15, 7, 3, 1). The experimental results are presented in Table 4.

We find that whether choosing all convolutional kernels to be 63 or starting with 63 in MSAttNet, both perform better than using 31 or 127. This may be because grouped convolutions at different frequencies are suited to different kernel sizes. Additionally, comparisons reveal that when all convolutional kernels use the same size, their performance

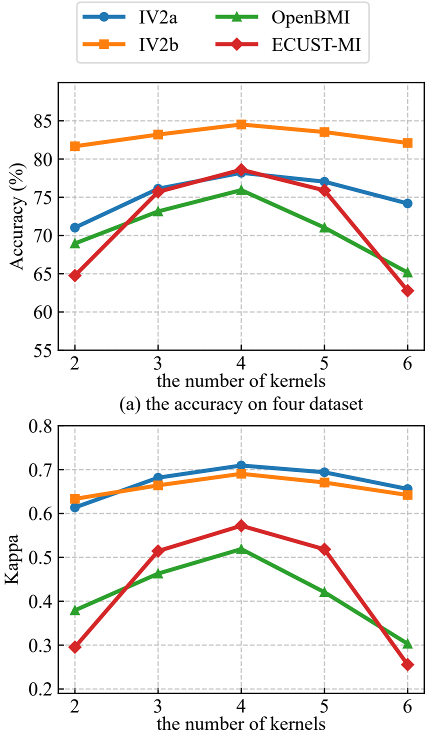
Table 4
The comparison of accuracy (%, Acc) and Kappa results between different kernal size using MSAttNet on four datasets.

Dataset	IV2a		IV2b		OpenBMI		ECUST-MI	
Metric	Acc.	Kappa	Acc.	Kappa	Acc.	Kappa	Acc.	Kappa
All31	74.92	0.6656	82.51	0.6502	68.08	0.3617	65.93	0.3189
All63	76.00	0.6800	82.46	0.6491	70.83	0.4167	69.56	0.3919
All127	72.42	0.6322	81.38	0.6276	64.03	0.2806	59.06	0.1809
Begin31	74.38	0.6584	82.87	0.6573	69.40	0.3880	70.16	0.4033
Begin127	72.61	0.6348	81.45	0.6291	66.06	0.3211	61.25	0.2253
MSAttNet	78.20	0.7094	84.52	0.6905	75.94	0.5189	78.60	0.5721

on the OpenBMI dataset is inferior to settings with varying kernel sizes. This indicates that when using Filter Banks in combination with the multi-scale temporal convolution module, selecting different kernel sizes can more effectively extract the time-frequency features of the signal.

4.3. The parameters of the attention spatial convolution module

The hyperparameter K determines the number of convolution kernels controls attention weights. We set convolution kernels K to 2, 3, 4, 5, and 6.



(a) the accuracy on four dataset

(b) the Kappa on four dataset

Figure 3: The comparison of accuracy (%, Acc) and Kappa results between different kernels of the attention spatial convolution module using MSAttNet on four datasets.

As shown in Figure 3, selecting 4 kernels for the attention spatial convolution module consistently yields optimal performance across different datasets. Both accuracy and Kappa coefficients gradually increase when the kernel count rises from 2 to 4, but gradually decrease beyond this point. We also observe that performance fluctuations vary across datasets due to differences in acquisition channels, subject numbers, and original EEG frequencies. Notably, the ECUST-MI dataset contains the fewest trials per subject among the four datasets, indicating it is more susceptible to variations in trial counts.

4.4. Different Pooling Layer

To investigate the impact of different pooling layers on the final results of MSAttNet, we select commonly used average pooling layer, max pooling layer, and the bilinear pooling layer employed in MSAttNet for evaluation.

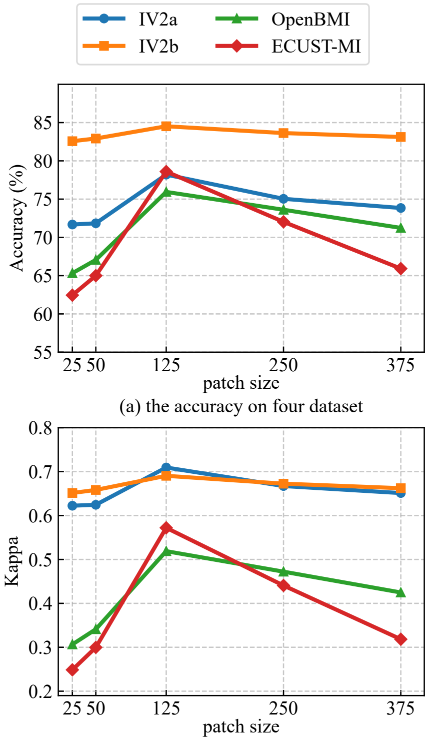
We demonstrate that employing the bilinear pooling layer effectively integrates spatial information between global and local features, achieving optimal performance with 78.2% accuracy and 0.7094 Kappa on the IV2a dataset, 84.52% accuracy and 0.6905 Kappa on the IV2b dataset, 75.94% accuracy and 0.5189 Kappa on the OpenBMI dataset, and 78.60% accuracy and 0.5721 Kappa on the ECUST-MI dataset. The max pooling layer consistently outperforms the average pooling layer across multiple datasets, likely because averaging local features generates uniform values across regions, thereby weakening the network's feature extraction capability. Given the shallow architecture and numerous network branches amplifying outlier effects, max pooling effectively discriminates between different local regions while accentuating outlier influence. Through bilinear pooling application, the network significantly enhances outlier handling capacity while simultaneously balancing local and global feature representation, ultimately achieving superior performance.

Furthermore, we investigate the impact of using different patch sizes (375, 250, 125, 50, and 25) in the bilinear pooling layer on the results, as shown in Figure 4.

In Figure 4, we observe that the model achieves optimal performance when the patch size approaches 125. If the vectors input to the bilinear pooling layer are regarded as features extracted from MI signals by the convolutional neural network, the patch size can be understood as sampling points. When the sampling points are too few, short samples cannot provide effective classifiable features through local bilinear transformations. Moreover, when the sampling points are too large, although more global information is

Table 5
The comparison of accuracy (%, Acc) and Kappa results between different pooling layer using MSAttNet on three datasets.

Dataset	IV2a		IV2b		OpenBMI		ECUST-MI	
Metric	Acc.	Kappa	Acc.	Kappa	Acc.	Kappa	Acc.	Kappa
AvgPool	32.29	0.0972	58.80	0.1760	58.43	0.1685	61.78	0.2362
MaxPool	62.50	0.5000	81.56	0.6313	61.76	0.2352	61.44	0.2289
MSAttNet	78.20	0.7094	84.52	0.6905	75.94	0.5189	78.60	0.5721



(a) the accuracy on four dataset

(b) the Kappa on four dataset

Figure 4: The comparison of accuracy (%, Acc) and Kappa results between different patch size of the bilinear pooling layer using MSAttNet on four datasets.

considered, the reduction in local information leads to a decline in performance at the local level.

4.5. Ablation Study

The four modules or layers of MSAttNet need to be discussed. These modules include the multi-band segmentation module, the attention spatial convolution module (ASC), and the depthwise convolution layer, and the bilinear pooling layer. Within the multi-band segmentation module, the filter bank is replaced with a bandpass filter of 4-100Hz, and the number of network branches is adjusted to one. The depthwise convolution layer use normal convolution layer to connect the filter bank. The attention spatial convolution module, and the bilinear pooling layer are removed.

In the ablation experiments, it demonstrates that all modules of the proposed MSAttNet are useful. When the multi-band segmentation module is ablated, the results are consistent with those in the 4-40Hz range in Table 3, indicating that 4-40Hz cannot meet the requirements of multi-branch usage. As shown in Table 4, the multi-band segmentation module divides into multiple branches, but only branches with the same kernel size perform similarly. When the depthwise convolution layer is ablated, despite the presence of the multiband segmentation module, a single branch cannot extract feature information from multiple frequency bands. This proves that multi-band division requires multiple different convolutional kernels, and the absence of any condition leads to performance degradation. The ablation of the attention spatial convolution layer results in the least metric decline. The ablation of the bilinear pooling layer causes the largest performance drop, indicating that the pooling structure is essential in the network. Moreover, as seen in Table 5, even replacing it with other pooling layers cannot achieve the same effect as the bilinear pooling layer.

4.6. Parameters of different networks

We evaluate four key metrics across three datasets. The number of parameters (Paras) reflects model size in thousands (K) and spatial complexity through trainable weights. Training time (TT, seconds, s) represents the time required to train on the training set for an individual model. Inf time (millisecond, ms) denotes the per-trial inference duration for the different depth model on the test set.

From the evaluation metrics of SOTA methods in Table 2 and the results such as Parameters in the Table 7, it can be observed that although the Paras, Training Time, and Inference Time of the proposed MSAttNet are higher than those of ADFCNN and EEGSimpleConv, the accuracy of these networks is significantly lower than that of MSAttNet. Compared to networks using nine frequency bands, such as FBCNet and FBMSNet, the training time and inference time of MSAttNet decrease, while the accuracy and other metrics improve, demonstrating the effectiveness of the proposed frequency band division and the network structure design coordinated with the frequency bands. Meanwhile, compared to networks using a single frequency band, it achieves comprehensive improvements across all metrics.

5. conclusion

We propose a novel convolutional neural network (CNN), the Multi-scale Attention Convolutional Neural Network (MSAttNet), to improve motor imagery (MI) classification performance. This architecture dynamically adapts to inter-subject neurophysiological patterns through its attention spatial convolutional module, eliminating the need

Table 6
The comparisons of the ablation study with the Proposed Method, MSAttNet, for the method's four modules, including the multi-band segmentation (MS) module, the attention spatial convolution module (ASC), and the depthwise convolution (DC) layer, and the bilinear pooling (BP) layer, on four dataset. The experimental evaluation metrics include accuracy (Acc., %), precision (Prec., %), recall (Rec., %), F1-score (F1, %), kappa, and p-value (p). The *, **, and *** indicate that the accuracies of MSAttNet are significantly higher than the compared methods with $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

Method	Acc.	Prec.	Rec.	F1	Kappa	p
w/o MS	73.69	74.11	73.69	73.13	0.6492	*
w/o ASC	75.00	76.25	75.00	74.65	0.6667	**
w/o DC	69.75	68.59	69.75	68.43	0.5967	**
w/o BP	28.59	28.88	28.59	28.61	0.0478	***
MSAttNet	78.20	78.69	78.20	78.02	0.7094	-

(a) The results on IV2a dataset

Method	Acc.	Prec.	Rec.	F1	Kappa	p
w/o MS	82.73	83.04	82.73	82.68	0.6547	***
w/o ASC	81.48	82.20	81.48	81.27	0.6297	***
w/o DC	80.09	80.29	80.09	79.96	0.6018	***
w/o BP	55.17	55.20	55.17	55.08	0.1035	***
MSAttNet	84.52	84.81	84.52	84.44	0.6905	-

(b) The results on IV2b dataset

Method	Acc.	Prec.	Rec.	F1	Kappa	p
w/o MS	68.50	70.41	68.50	66.98	0.3700	***
w/o ASC	74.56	75.42	74.56	73.96	0.4911	*
w/o DC	65.83	66.89	65.83	65.08	0.3167	***
w/o BP	57.11	57.16	57.11	57.02	0.1422	***
MSAttNet	75.94	76.50	75.94	75.68	0.5189	-

(d) The results on OpenBMI dataset

Method	Acc.	Prec.	Rec.	F1	Kappa	p
w/o MS	67.18	68.25	67.18	66.80	0.3436	***
w/o ASC	74.03	74.96	74.05	73.58	0.4810	*
w/o DC	65.11	66.59	65.02	63.77	0.3009	***
w/o BP	55.36	55.55	55.33	54.89	0.1065	***
MSAttNet	78.60	79.04	78.60	78.50	0.5721	-

(d) The results on ECUST-MI dataset

Table 7
The comparison of the number of parameters (K), training time (TT, s) and inference time (IT, ms) of networks on four datasets.

Model	IV2a			IV2b			OpenBMI			ECUST-MI		
	Paras	TT	IT	Paras	TT	IT	Paras	TT	IT	Paras	TT	IT
ShallowConvNet	43.36	96.14	34.47	9.44	24.34	10.94	103.84	208.05	44.89	30.24	19.41	6.27
DeepConvNet	280.45	71.51	24.50	266.98	24.34	9.09	303.85	142.00	28.88	275.10	15.15	5.91
EEGNet	2.62	38.24	11.75	1.52	16.24	7.52	2.46	98.22	20.59	1.73	10.28	5.12
FBCNet	10.66	235.46	145.06	3.46	43.70	15.59	20.45	270.37	105.24	7.20	34.32	13.90
EEGConformer	615.49	79.74	19.21	585.03	39.71	12.03	679.43	122.67	28.82	605.83	22.82	7.85
FBMSNet	15.08	223.89	96.36	7.88	53.91	16.38	24.87	466.25	95.62	11.62	35.44	12.83
IFNet	10.88	121.35	44.91	7.68	26.60	9.12	15.23	107.25	33.86	9.35	11.59	5.60
TSFCNet	43.98	167.85	73.76	8.27	45.19	15.91	114.47	304.97	93.13	31.67	32.98	13.45
ADFCNN	4.03	50.02	14.91	2.67	22.26	7.86	6.91	117.51	24.34	3.60	13.29	5.91
EISATCFusion	25.86	48.44	14.84	24.42	31.25	10.56	26.31	109.41	21.99	24.84	18.20	6.87
EEGSimpleConv	29.44	29.95	11.82	25.67	17.72	8.37	36.99	151.00	33.37	28.16	8.70	4.52
DMSACNN	27.72	92.91	25.17	14.64	29.83	9.45	21.72	237.79	51.28	16.20	17.50	6.84
MSAttNet	43.21	110.08	43.27	15.68	46.00	14.84	113.97	171.47	49.04	32.60	23.14	10.27

for subject-specific parameter calibration while effectively capturing discriminative spectral-spatial features across frequency bands. The integration of overlapping filter banks and multi-scale temporal convolutions ensures comprehensive coverage of rhythm dynamics and transient neural events characteristic of MI tasks.

In frequency band selection, we demonstrate the effectiveness of partitioning bands at 4-16Hz, 12-24Hz, 20-36Hz, 32-44Hz, and 40-100Hz. Furthermore, ablation studies involving the removal of the multi-scale temporal convolution structure confirm that the number of frequency bands should match the number of branches in the multi-scale architecture. As shown in Figure 2, features vary across different time segments and frequency bands, allowing complementary integration through the carefully designed network architecture. We conduct detailed parameter comparisons across different structural components and perform ablation experiments to verify the effectiveness of the proposed architecture.

Experimental validation across four benchmark datasets demonstrates that this framework outperforms 12 state-of-the-art (SOTA) methods. Our method achieves SOTA performance across multiple datasets, recording accuracies of 78.20% on the IV2a dataset, 84.52% on the IV2b dataset, 75.94% on the OpenBMI dataset, and 78.60% on the ECUST-MI dataset.

The existing frequency band division mainly relies on the settings of the Filter Bank, and networks such as OctConv network can automatically separate high-frequency and low-frequency signals for images Chen, Fan, Xu, Yan, Kalantidis, Rohrbach, Yan and Feng (2019). We will combine dynamic spatial convolution and OctConv to design a network for the automatic division of frequency bands, and at the same time to adjust the weight of different frequency bands to achieve further improvements in accuracy. We also incorporate viable methodologies from other EEG tasks, such as SSVEP Deng, Li, Zhang, Zheng, Liu, Ding, Wang and Gao (2025), emotion recognition Ye, Jing, Wang, Li, Liu, Yan, Zhang and Gao (2023); Gao, Liu, Zhang, Wang, Chang, Ouyang, Liu and Li (2025), and fatigue detection Li, Zhang, Liu, Lin, Zhang, Tang and Gao (2023), into MSAttNet to further enhance recognition efficiency. In transfer learning, we experiment with various cross-subject approaches to reduce individual variability and mitigate the impact of limited single-subject data Zhang, Li, Chang, Liu, Qin, Xie, Wang, Gao and Wu (2025); Lin, Li, Wang, Bai, Cui, Yu, Gao and Zhang (2024). We introduce multiple modalities, such as EOG Tang, Li, Zhang, Deng, Liu, Zheng, Chang, Zhao, Wang, Zuo et al. (2024) and fNIRS Xu et al. (2023), and design variants of MSAttNet to improve recognition performance. Ultimately, this enables MSAttNet to demonstrate utility beyond MI-EEG tasks.

References

Ahn, H.J., Lee, D.H., Jeong, J.H., Lee, S.W., 2022. Multiscale convolutional transformer for eeg classification of mental imagery in different modalities. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 646–656.

Ang, K.K., Chin, Z.Y., Zhang, H., Guan, C., 2008. Filter bank common spatial pattern (FBCSP) in brain-computer interface, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE. pp. 2390–2397.

Arpaia, P., Esposito, A., Natalizio, A., Parvis, M., 2022a. How to successfully classify EEG in motor imagery BCI: a metrological analysis of the state of the art. Journal of Neural Engineering 19, 031002.

Arpaia, P., Esposito, A., Natalizio, A., Parvis, M., 2022b. How to successfully classify EEG in motor imagery BCI: a metrological analysis of the state of the art. Journal of Neural Engineering 19, 031002.

Barmpas, K., Panagakis, Y., Bakas, S., Adamos, D.A., Laskaris, N., Zafeiriou, S., 2023. Improving generalization of CNN-based motor-imagery EEG decoders via dynamic convolutions. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 1997–2005.

Blankertz, B., Muller, K.R., Krusienski, D.J., Schalk, G., Wolpaw, J.R., Schlogl, A., Pfurtscheller, G., Millan, J.R., Schroder, M., Birbaumer, N., 2006. The bci competition iii: Validating alternative approaches to actual bci problems. IEEE transactions on neural systems and rehabilitation engineering 14, 153–159.

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z., 2020. Dynamic convolution: Attention over convolution kernels, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11030–11039.

Chen, Y., Fan, H., Xu, B., Yan, Z., Kalantidis, Y., Rohrbach, M., Yan, S., Feng, J., 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3435–3444.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Deng, L., Li, P., Zhang, H., Zheng, Q., Liu, S., Ding, X., Wang, M., Gao, D., 2025. Tsmnet: A comprehensive network based on spatio-temporal representations for ssvep classification. Biomedical Signal Processing and Control 105, 107554.

Deng, X., Zhang, B., Yu, N., Liu, K., Sun, K., 2021. Advanced TSGL-EEGNet for motor imagery EEG-based brain-computer interfaces. IEEE access 9, 25118–25130.

Duan, L., Li, J., Ji, H., Pang, Z., Zheng, X., Lu, R., Li, M., Zhuang, J., 2020. Zero-shot learning for eeg classification in motor imagery-based bci system. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, 2411–2419.

El Ouahidi, Y., Gripon, V., Pasdeloup, B., Bouallegue, G., Farrugia, N., Lioi, G., 2024. A Strong and Simple Deep Learning Baseline for BCI Motor Imagery decoding. IEEE Transactions on Neural Systems and Rehabilitation Engineering .

Gao, D., Liu, M., Zhang, H., Wang, M., Chang, H., Ouyang, G., Liu, S., Li, P., 2025. A multi-domain constraint learning system inspired by adaptive cognitive graphs for emotion recognition. Neural Networks 188, 107457.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications 13, 18–28.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, pmlr. pp. 448–456.

Jiang, W.B., Zhao, L.M., Lu, B.L., 2024. Large brain model for learning generic representations with tremendous eeg data in bci. arXiv preprint arXiv:2405.18765 .

Jin, J., Miao, Y., Daly, I., Zuo, C., Hu, D., Cichocki, A., 2019. Correlation-based channel selection and regularized feature optimization for MI-based BCI. Neural Networks 118, 262–270.

Jin, J., Wang, Z., Xu, R., Liu, C., Wang, X., Cichocki, A., 2021. Robust similarity measurement based on a novel time filter for SSVEPs detection. IEEE Transactions on Neural Networks and Learning Systems 34, 4096–4105.

Jin, J., Xiao, R., Daly, I., Miao, Y., Wang, X., Cichocki, A., 2020. Internal feature selection method of csp based on l1-norm and dempster–shafer theory. IEEE transactions on neural networks and learning systems 32, 4814–4825.

Jin, J., Xu, R., Daly, I., Zhao, X., Wang, X., Cichocki, A., 2024. MOCNN: A multiscale deep convolutional neural network for ERP-based brain-computer interfaces. IEEE Transactions on Cybernetics .

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Ko, W., Jeon, E., Jeong, S., Suk, H.I., 2021. Multi-scale neural network for eeg representation learning in bci. IEEE Computational Intelligence Magazine 16, 31–45.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of neural engineering 15, 056013.

Lee, M.H., Kwon, O.Y., Kim, Y.J., Kim, H.K., Lee, Y.E., Williamson, J., Fazli, S., Lee, S.W., 2019. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. GigaScience 8, giz002.

Li, P., Zhang, Y., Liu, S., Lin, L., Zhang, H., Tang, T., Gao, D., 2023. An eeg-based brain cognitive dynamic recognition network for representations of brain fatigue. Applied Soft Computing 146, 110613.

Li, Y., Sun, Y., Wan, F., Yuan, Z., Jung, T.P., Wang, H., 2025. Metanirs: A general decoding framework for fnirs based motor execution/imagery. Neural Networks , 107873.

Liang, G., Cao, D., Wang, J., Zhang, Z., Wu, Y., 2024. EISATC-fusion: Inception self-attention temporal convolutional network fusion for motor imagery EEG decoding. IEEE Transactions on Neural Systems and Rehabilitation Engineering .

Liang, S., Kuang, S., Wang, D., Yuan, Z., Zhang, H., Sun, L., 2023a. An auxiliary synthesis framework for enhancing eeg-based classification with limited data. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 2120–2131.

Liang, T., Yu, X., Liu, X., Wang, H., Liu, X., Dong, B., 2023b. EEG-CDILNet: a lightweight and accurate CNN network using circular dilated convolution for motor imagery classification. Journal of Neural Engineering 20, 046031.

Lin, L., Li, P., Wang, Q., Bai, B., Cui, R., Yu, Z., Gao, D., Zhang, Y., 2024. An eeg-based cross-subject interpretable cnn for game player expertise level classification. Expert Systems with Applications 237, 121658.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13, Springer. pp. 740–755.

Liu, K., Xing, X., Yang, T., Yu, Z., Xiao, B., Wang, G., Wu, W., 2025. DMSACNN: Deep Multiscale Attentional Convolutional Neural Network for EEG-Based Motor Decoding. IEEE Journal of Biomedical and Health Informatics .

Liu, K., Yang, M., Yu, Z., Wang, G., Wu, W., 2022. FBMSNet: A filter-bank multi-scale convolutional neural network for EEG-based motor imagery decoding. IEEE Transactions on Biomedical Engineering 70, 436–445.

Luo, J., Mao, Q., Wang, Y., Shi, Z., Hei, X., 2022. Algorithm contest of calibration-free motor imagery bci in the bci controlled robot contest in world robot contest 2021: A survey. Brain Science Advances 8, 127–141.

Mane, R., Chew, E., Chua, K., Ang, K.K., Robinson, N., Vinod, A.P., Lee, S.W., Guan, C., 2021. FBCNet: A multi-view convolutional neural network for brain-computer interface. arXiv preprint arXiv:2104.01233 .

McFarland, D.J., Miner, L.A., Vaughan, T.M., Wolpaw, J.R., 2000. Mu and beta rhythm topographies during motor imagery and actual movements. Brain topography 12, 177–186.

Müller-Gerking, J., Pfurtscheller, G., Flyvbjerg, H., 1999. Designing optimal spatial filters for single-trial EEG classification in a movement task. Clinical neurophysiology 110, 787–798.

Padfield, N., Zabalza, J., Zhao, H., Masero, V., Ren, J., 2019. Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges. Sensors 19, 1423.

Pérez-Velasco, S., Santamaría-Vázquez, E., Martínez-Cagigal, V., Marcos-Martínez, D., Hornero, R., 2022. Eegsym: Overcoming inter-subject variability in motor imagery based bcis with deep learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering 30, 1766–1775.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. Human brain mapping 38, 5391–5420.

Song, Y., Zheng, Q., Liu, B., Gao, X., 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 710–719.

Tang, M., Li, P., Zhang, H., Deng, L., Liu, S., Zheng, Q., Chang, H., Zhao, C., Wang, M., Zuo, G., et al., 2024. Hms-tenet: a hierarchical multi-scale topological enhanced network based on eeg and eog for driver vigilance estimation. Biomedical Technology 8, 92–103.

Tangermann, M., Müller, K.R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K.J., Müller-Putz, G.R., et al., 2012a. Review of the BCI competition IV. Frontiers in neuroscience 6, 55.

Tangermann, M., Müller, K.R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K.J., Müller-Putz, G.R., et al., 2012b. Review of the BCI competition IV. Frontiers in neuroscience 6, 55.

Tao, W., Wang, Z., Wong, C.M., Jia, Z., Li, C., Chen, X., Chen, C.P., Wan, F., 2023. ADFCNN: attention-based dual-scale fusion convolutional neural network for motor imagery brain–computer interface. IEEE Transactions on Neural Systems and Rehabilitation Engineering 32, 154–165.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., Li, H., 2024. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. Advances in Neural Information Processing Systems 37, 39249–39280.

Wang, H., Yuan, Z., Zhang, H., Wan, F., Li, Y., Xu, T., 2025. Hybrid eeg-fnirs decoding with dynamic graph convolutional-capsule networks for motor imagery/execution. Biomedical Signal Processing and Control 104, 107570.

Wang, J., Yao, L., Wang, Y., 2023. IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 1900–1911.

Wu, D., Jiang, X., Peng, R., 2022. Transfer learning for motor imagery based brain–computer interfaces: A tutorial. Neural Networks 153, 235–253.

Wu, S.L., Wu, C.W., Pal, N.R., Chen, C.Y., Chen, S.A., Lin, C.T., 2013. Common spatial pattern and linear discriminant analysis for motor imagery classification, in: 2013 IEEE symposium on computational intelligence, cognitive algorithms, mind, and brain (Ccmb), IEEE. pp. 146–151.

Wu, Y., He, K., 2018. Group normalization, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Xiao, R., Huang, Y., Xu, R., Wang, B., Wang, X., Jin, J., 2022. Coefficient-of-variation-based channel selection with a new testing framework for MI-based BCI. Cognitive Neurodynamics , 1–13.

Xu, T., Zhou, Z., Yang, Y., Li, Y., Li, J., Bezerianos, A., Wang, H., 2023. Motor imagery decoding enhancement based on hybrid eeg-fnirs signals. IEEE Access 11, 65277–65288.

Ye, Z., Jing, Y., Wang, Q., Li, P., Liu, Z., Yan, M., Zhang, Y., Gao, D., 2023. Emotion recognition based on convolutional gated recurrent units with attention. Connection Science 35, 2289833.

Zhang, H., Li, P., Chang, H., Liu, S., Qin, Y., Xie, J., Wang, M., Gao, D., Wu, D., 2025. A coupling of common-private topological patterns learning approach for cross-subject emotion recognition. Biomedical Signal Processing and Control 105, 107550.

Zhi, H., Yu, Z., Yu, T., Gu, Z., Yang, J., 2023. A multi-domain convolutional neural network for EEG-based motor imagery decoding. IEEE Transactions on Neural Systems and Rehabilitation Engineering 31, 3988–3998.