Cognitive Neurodynamics (2024) 18:3463–3476
https://doi.org/10.1007/s11571-024-10127-8

RESEARCH ARTICLE

MSHANet: a multi-scale residual network with hybrid attention for motor imagery EEG decoding

Mengfan Li[1,2,3] · Jundi Li[1,2,3] · Xiao Zheng[2,3,4] · Jiahao Ge[1,2,3] · Guizhi Xu[2,3,4]

Received: 29 December 2023 / Revised: 14 April 2024 / Accepted: 7 May 2024 / Published online: 21 May 2024

Abstract

EEG decoding plays a crucial role in the development of motor imagery brain-computer interface. Deep learning has great potential to automatically extract EEG features for end-to-end decoding. Currently, the deep learning is faced with the chanllenge of decoding from a large amount of time-variant EEG to retain a stable peroformance with different sessions. This study proposes a multi-scale residual network with hybrid attention (MSHANet) to decode four motor imagery classes. The MSHANet combines a multi-head attention and squeeze-and-excitation attention to hybridly focus on important information of the EEG features; and applies a multi-scale residual block to extracts rich EEG features, sharing part of the block parameters to extract common features. Compared with seven state-of-the-art methods, the MSHANet exhits the best accuracy on BCI Competition IV 2a with an accuracy of 83.18% for session- specific task and 80.09% for cross-session task. Thus, the proposed MSHANet decodes the time-varying EEG robustly and can save the time cost of MI-BCI, which is beneficial for long-term use.

Keywords Brain-computer interface · Multi-scale residual network · Hybrid attention · EEG decoding · Motor imagery

Introduction

A brain-computer interface (BCI) is a control system that enables direct communication between the human brain and a computer by monitoring the electroencephalogram (EEG) without the need for a peripheral nervous system (Gao et al. 2021). The electroencephalography (EEG) signal is a common physiological signal that can be used to diagnose conditions such as epilepsy (Kukker et al. 2021), depression (Shen et al. 2023), and stroke(Roesch et al. 2024). Motor imagery based on BCI (MI-BCI) requires people to imagine muscle movement without moving to control devices, which has important applications in medical rehabilitation (Mane et al. 2020) and control engineering (Duan et al. 2015; Li et al. 2023b). The enhancement of the performance of MI-BCI can help to improve the rehabilitation treatment (Ju et al. 2022; Chen et al. 2021) and improve the perception of human-computer interaction experience (Broccard et al. 2014).

The decoding of EEG signals is one of important components in MI-BCI (Craik et al. 2019; Xu et al. 2021). With the development of computer science, deep learning (DL) has gradually applied in decoding, showing excellent performance in MI-BCI with its advantages of automatic feature extraction and capturing complex features (Zhao et al. 2023). Commonly used methods in DL include convolutional neural networks (CNN) (Feng et al. 2023), recurrent neural networks (RNN) (Said et al. 2023), neural reinforcement learning (Kukker et al. 2023) and graph neural networks (GNN) (Sun et al. 2021). Traditional feature extraction methods include common space pattern (CSP), wavelet transform and hilbert-yellow transform (Jareda et al. 2019). DL has the feature of strong compatibility and can be combined with traditional algorithms, and scholars

✉ Mengfan Li
mfli@hebut.edu.cn

1 State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Science and Biomedical Engineering, Hebei University of Technology, Tianjin, China

2 Hebei Key Laboratory of Bioelectromagnetics and Neuroengineering, Tianjin, China

3 Tianjin Key Laboratory of Bioelectromagnetic Technology and Intelligent Health, Tianjin, China

4 School of Electrical Engineering, Hebei University of Technology, Tianjin, China

have used DL to decode traditional EEG features. Wang and Shajil used CNN classification with wavelet transform and CSP for feature extraction, which significantly improved the classification accuracy of MI (Wang et al. 2022; Shajil et al. 2020). Luo et al. proposed a new architecture for MI-BCI classification by extracting spatial frequency features via FBCSP and then inputting them into an RNN containing gated recurrent unit (GRU) by sliding window cropping (Luo et al. 2018). While traditional feature extraction methods lead to loss of feature information and manual extraction increases the workload, the use of DL can automatically extract deep EEG features that are relevant to the task. Li and Zhi proposed a network architecture to extract temporal and spatial features from pre-processed EEG to obtain highly discriminative representations of the features, which all achieved good results on public datasets (Li et al. 2022; Zhi et al. 2023). Compared to methods that require filtering and artifact elimination in the pre-processing, some scholars have proposed that DL-based end-to-end models can extract high-quality features from raw EEGs containing noisy interferences. This reduces the complexity of the decoding process by reducing the manually designed pre-processing process. Hauke et al. used an end-to-end CNN containing temporal and spatial filters to classify raw EEG and achieved good classification accuracy on a two-classified public dataset (Dose et al. 2018). Xie et al. proposed an end-to-end network combining CNNs and transformers for inputting raw EEGs, which has good classification performance in two, three, and four classified MI-BCI (Xie et al. 2022). While DL enables automatic feature extraction and performs well in decoding, the time-varying characteristics of EEG signals leads to feature redundancy. This is because the model is unable to adaptively focus on the important parts of the features.

Attentional mechanisms are becoming an effective means of addressing the reduction of redundant features in DL and have a wide range of applications in computing, natural language processing, and other fields, by virtue of their application of human perceptual patterns and attentional behaviors to neural networks to allow the network to perceive the important and unimportant parts of the data. Currently many researchers use different types of attention mechanisms such as temporal attention mechanism, spatial attention mechanism and channel attention mechanism in MI-BCI classification. Jia et al. proposed a multi-branch CNN feature extraction method combining channel attention mechanism with the classifier LightGBM for feature classification (Jia et al. 2023). Ma et al. designed a novel temporal attention module to capture the importance of features in different time windows, and verified that more discriminative features can be fused by this module (Ma et al. 2023). While attention mechanism can effectively improve the model classification performance, a hybrid attention mechanism can fuse more important feature information by combining different types of attention mechanisms. Wu et al. solved the problem of information inequality in time and channel for EEG by a module that fuses temporal attention mechanism and channel attention mechanism. The module can automatically learn the importance of different features with an average accuracy improvement of 3.45% (Wu et al. 2023). Li et al. proposed novel dual-attention module for MI adversarial networks, which can effectively reduce the domain differences between subjects by using the attention module to learn contextual information between different classes (Li et al. 2023a). Scholars have usually used networks combining a single-branch network and an attention mechanism to adaptively focus on important parts of the EEG, while processing multi-scale EEG signals does not consider feature information at different scales and suffers from the limitation of incomplete feature characterization. Therefore, multi-scale networks combined with attention mechanisms can help to extract abundant features.

Multi-scale network with multiple branches has been introduced to MI-BCI field since they have ability to capture local features and global features of different scale sizes by setting up convolutional kernels of different sizes. Roy proposed an effective convolutional neural network for multi-scale feature fusion to extract multi-scale features of different frequency bands in EEG, which improves the robustness of the model (Roy et al. 2022). Zhang and Ko introduced the multi-scale based on time-frequency features and spatial features under different branches and achieved competitive performance (Zhang et al. 2021; Ko et al. 2021). These studies have shown that the multi-scale structure can effectively cope with time-varying feature changes, improve the richness of feature extraction, and enhance the robustness of the model.

Considering the advantage of attention and multi-scale structure, this study proposes a multi-scale residual network with hybrid attention (MSHANet) to decode EEG. By combining hybrid attention block and multi-scale block, MSHANet can adaptively adjust the importance of features in different scales, which helps to improve the robustness of the model. Firstly, a novel hybrid attentional block consisting of a multi-head attentional that captures different levels of information and an SE attentional that adaptively adjusts the weights of the channels is proposed, which can effectively extract the important information of the feature. Secondly, a multi-scale residual block consists of temporal convolution networks of different scale sizes in parallel to extract rich features. It can effectively adapt to individual differences and improve the robustness of the model. Finally, the method of sharing some of the network parameters of the multi-scale residual block is adopted, which can

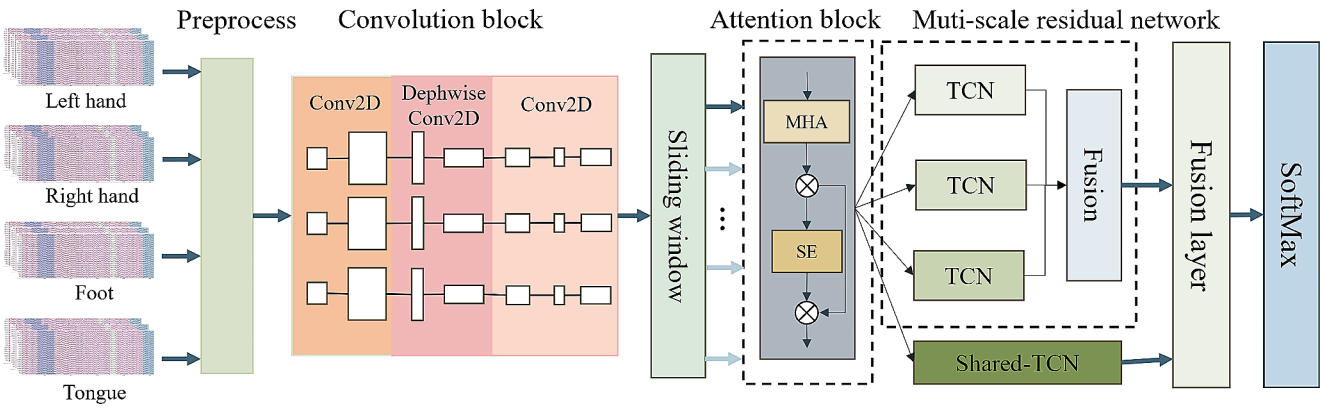


Fig. 1 The architecture of BCI decoding model based on multi-scale network

Table 1 MSHANet pseudocode

Algorithm:	MSHANet
input:	Training samples is a portion of N subject set $\{(E_j, Y_j)\}^N_{j=1}$.
output:	Classification result R.
1:	for j in N do
2:	E_j is normalized to X_j.
3:	Feed (X_j, Y_j) to convolution block to get features $F(X_j)$.
4:	for m in n_window do
5:	Feed sub-features $F(X_j^m)$to hybrid attention block to get the refined feature $R(F(X_j^m))$.
6:	Feed $R(F(X_j^m))$ to shared-TCN to get the common features $C(R(F(X_j^m)))$.
7:	for k in num do
8:	Feed $R(F(X_j^m))$ to TCN branch with kernel size k to get features $T(R(F(X_j^{mk})))$.
9:	Feed $T(R(F(X_j^{mk})))$ into fusion layer to obtain features $T(R(F(X_j^m)))$.
10:	Integration of $C(R(F(X_j^m)))$ and $T(R(F(X_j^m)))$ features to get fusion features $W(R(F(X_j^m)))$.
11:	Add $W(R(F(X_j^m)))$ into Softmax to get R.
12:	end

extract the common features between different branches and reduce the number of parameters. By comparing with some of the state-of-the-art baseline models, MSHANet robustly decodes time-varying signals to reduce the time cost of MI-BCI, facilitating the translation to practical applications.

Method

The MSHANet is consisted of a convolution block, a sliding window, a hybrid attention block, and a multi-scale residual block, as shown in Fig. 1. The EEG data is fed into the convolution block consisting of two temporal filters and one spatial filter. The EEG features are extracted into a compact convolutional neural network and then are fed separately into a multi-branch network. Each branch owns a hybrid attention block which is a combination of a multi-head attention mechanism and an SE attention mechanism and it will focus on the important information in features. Then, the multi-scale residual block consisting of temporal convolutional networks (TCN) with different convolutional kernel sizes further extract deeper and richer information. In addition, a separate TCN exists to share network parameters across multiple branches to extract common features in different time windows. Finally, the high-dimensional features fused from different branches are input to the softmax layer for classification. Table 1 shows the pseudocode for MSHANet.

Convolution block

The pre-processed EEG data for each subject is represented as $\{(X_j, Y_j)\}^N_{j=1}$, and $X_j=\{X^i_j\}^{|Xj|}_{i=1}$ represents samples from subject j, where i denotes the i th trial. $Y_j=\{y^i_j\}^{|Yj|}_{i=1}$ is the corresponding truth labels. As shown in Fig. 2, the convolution block contains three EEGNet-based modules: two temporal filters and a spatial filter. The X_j is first fed into the first temporal filter, which contains a 2D convolution and a batch normalization layer. The convolution kernel size F_1 is set to K_1 ($K_1=64$), which is 1/4 of the data sampling rate (256 Hz) and captures frequency information at 4 Hz and above. The spatial filter consists of Depthwise Conv2D and a pooling layer, then batch normalization and activation functions are applied along the feature map dimensions. An average pooling layer is used to reduce the spatial dimensions of the inputs and a regularization technique dropout is added to prevent overfitting. The second temporal filter differs from EEGNet (Lawhern et al. 2018) by utilizing a 2D convolution with F_2 convolutional kernels of size $(1, K_2)$ ($K_2=64$). After convolution block to obtain $F(X_j)$, the data still retains time-related information and the size becomes $(F2, 1, T)$, where T denotes time points after convolutional transformation. Table 2 gives the detailed parameters of the convolution block for building.

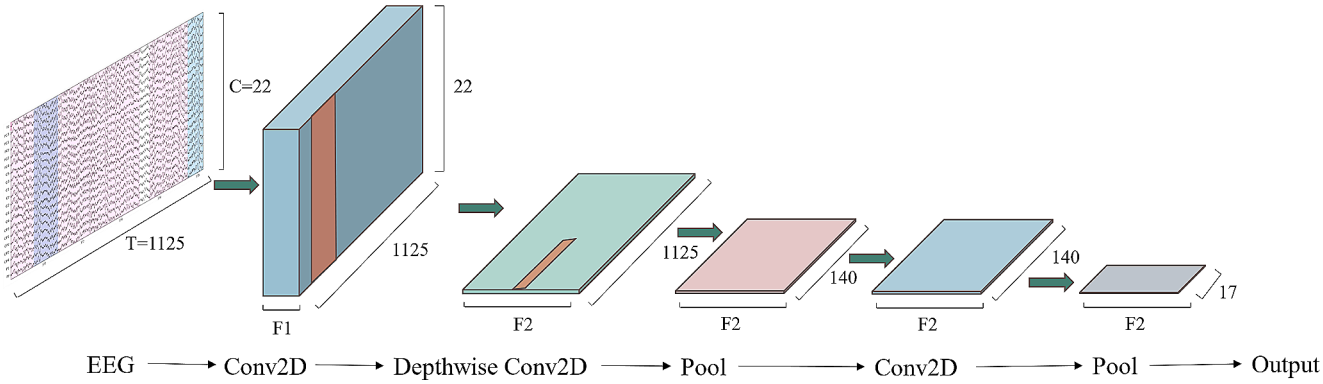


Fig. 2 The architecture of convolution block

Table 2 Parameter settings for convolution block

Layer	Filters	size	params	Output
Input				(C, T)
Reshape				$(T, C, 1)$
Conv2D-1	F_1	$(K_1, 1)$	$64 * F_1$	(T, C, F_1)
BatchNorm			$2 * F_1$	(T, C, F_1)
DepthwiseCon2D	$D * F_1$	$(1, C)$	$C * D * F_1$	$(T, 1, D * F_1)$
BatchNorm			$2 * D * F_1$	$(T, 1, D * F_1)$
Activation				$(T, 1, D * F_1)$
AveragePool2D		(8, 1)		$(T // 8, 1, D * F_1)$
Dropout				$(T // 8, 1, D * F_1)$
Conv2D-2	F_2	$(K_2, 1)$	$16 * D * F_1 + F_2 * (D * F_1)$	$(T // 8, 1, F_2)$
BatchNorm			$2 * F_2$	$(T // 8, 1, F_2)$
Activation				$(T // 8, 1, F_2)$
AveragePool2D		(8, 1)		$(T // 64, 1, F_2)$
Dropout				$(T // 64, 1, F_2)$

Sliding window block

A sliding window block (Altaheri et al. 2022) divides the feature into sub-features that serve as inputs for different branches. The block with window size w and sliding step t_w is acted on the features, and X_j is divided into n windows. Equation (1) represents the calculation between F_2, t_w and n.

$$n = F_2 - t_w + 1, F_2 > t_w \geq 1 \tag{1}$$

The data after the sliding window is defined as $F(X_j^m)$, where m denotes the number of windows. The EEG has a data dimension of (17, 1, 32) after convolution block, and is divided into three (15, 1, 32) sized sub-feature matrices after the sliding window. Each sub-feature matrix serves as an input to a branch, and each branch contains a hybrid attention block and a multi-scale residual block.

Hybrid attention block

The hybrid attention block contains the multi-head attention (MHA) (Tao et al. 2018) and squeeze and excitation (SE) attention (Hu et al. 2018). The three important parameters obtained from the input features in the attention mechanism, value (V) is the vector representing the input features. Query (Q) and Key (K) are the feature vectors for computing attention weights. MHA in MSHANet adopts multiple independent attention heads to learn different features in the input sequence in parallel. The input features are first processed with multiple sets of self-attention, and then the k-head results are stitched together and linearly transformed to obtain the final output.

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_k) W^o \tag{2}$$

$$head_i = Attention1(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where $k = 2$ and $head_i$ is the output of the i th, $W_i^Q \in R^{d \times d_k}$, $W_i^K \in R^{d \times d_k}$, $W_i^V \in R^{d \times d_v}$, $W_i^O \in R^{d \times d_v}$. The input features are mapped to the hidden unit dimensions by a linear transformation before attention computation, where hidden size is set to 8. The structure of one head is shown underneath the Fig. 3.

The SE attention mechanism is added after the MHA to further enhance the feature representation and selection and remove the redundant features present after the MHA. Adaptive selection and weighting of feature channels to capture key information in features. SE are divided into squeeze, excitation, and scale. Firstly, the input features $\widehat{X} \in \mathbb{R}^{H \times W \times C}$ are squeezed into a feature vector, where H, W, C are the feature matrices' height, width, and channels. As shown in Eq. (4):

$$z_c = squeeze\left(\widehat{X}_c\right) = \frac{1}{H \times W} \sum_i^H \sum_j^W x_c(i, j) \tag{4}$$

where $\widehat{X}_c \in \mathbb{R}^{H\times W}, c \in [1,2,3,\ldots,C]$. $x_c(i,j)$ represents the data point in $\widehat{X}_c$. $squeeze()$ represents feature squeezes along spatial dimensions, which squeezes each two-dimensional feature $\widehat{X}_c$ into z_c. The weights of each channel are learnt in excitation in the order of the fully connected layers and ReLU. The scaling part uses the learned weight vectors for each channel of the input feature map to get the final output.

The attention mechanism can strengthen the feature expression ability of the model. The hybrid attention block can combine the advantages of MHA and SE, not only focusing on the feature relationship between different heads, but also focusing on the importance of channels. The input feature map $F(X_j^m)$ passes through MHA as $F_m(X_j^m)$. And then, $F_m(X_j^m)$ is multiplied with the result $R(F(X_j^m))$after inputting SE for the refined feature map of the hybrid attention mechanism.

$$F_m\left(X_j^m\right) = MHA\left(F\left(X_j^m\right)\right) \tag{5}$$

$$R\left(F\left(X_j^i\right)\right) = SE\left(F_m\left(X_j^i\right)\right) \otimes F_m\left(X_j^i\right) \tag{6}$$

where $F(X_j^m)$ is the input feature map, $F_m(X_j^m)$ is the result of MHA, and $R(F(X_j^m))$ is the output of SE and hybrid attention block. The structure of the hybrid attention block is shown in Fig. 3.

Multi-scale residual block

The multi-scale residual block extracts feature-rich and robust features from EEG through convolution kernels at different scales, as shown in Fig. 4. Inspired by the inception structure, multi-scale features are acquired through multiple convolutional layers with different convolutional kernel sizes, and feature fusion is carried out using averaging to synthesize the multi-scale feature information. Smaller convolutional kernels are utilized to capture local information over a short period of time and larger convolutional kernels to capture global information over a long period of time, resulting in better perception of features at different scales.

Each branch consists of TCN branches of different scale sizes. The TCN (Bai et al. 2018) is a residual network based on 1D convolution that produces an output of the same length as the input. The receptive field is expanded using causal convolution and expansion convolution. The receptive field depends on L, filter size and expansion coefficient.

$$F(s) = (x *_d f)(s) = \sum_{(i=0)}^{(k-1)} f(i) \cdot x_{(s-d\cdot i)}. \tag{7}$$

where d is the dilation factor, k is the kernel size. Corrected linear unit (ReLU) is used after each convolution to normalize the weights to be used in the convolution. 1×1 convolution is used to keep the input and output having the same shape.

In addition, to extract the shared features among different sub-feature matrices, MSHANet performs parameter sharing by selecting a shared-TCN. Each sub-feature matrix after sliding window is mapped in the same feature space, and shared-TCN transfers information between different branches and extracts common information in different branches.

$$C\left(R\left(F\left(X_j^m\right)\right)\right) = f\left(\sum\nolimits_k {w_k}^n \cdot R\left(F\left(X_j^m\right)\right)\right. \tag{8}$$

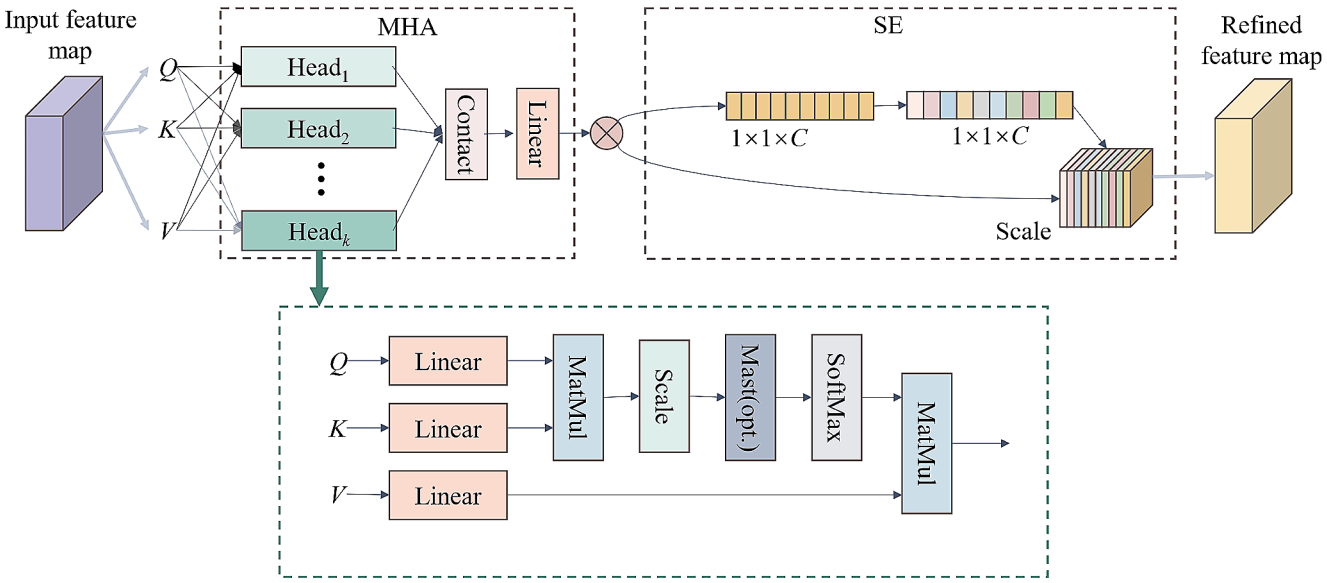


Fig. 3 The architecture of hybrid attention block

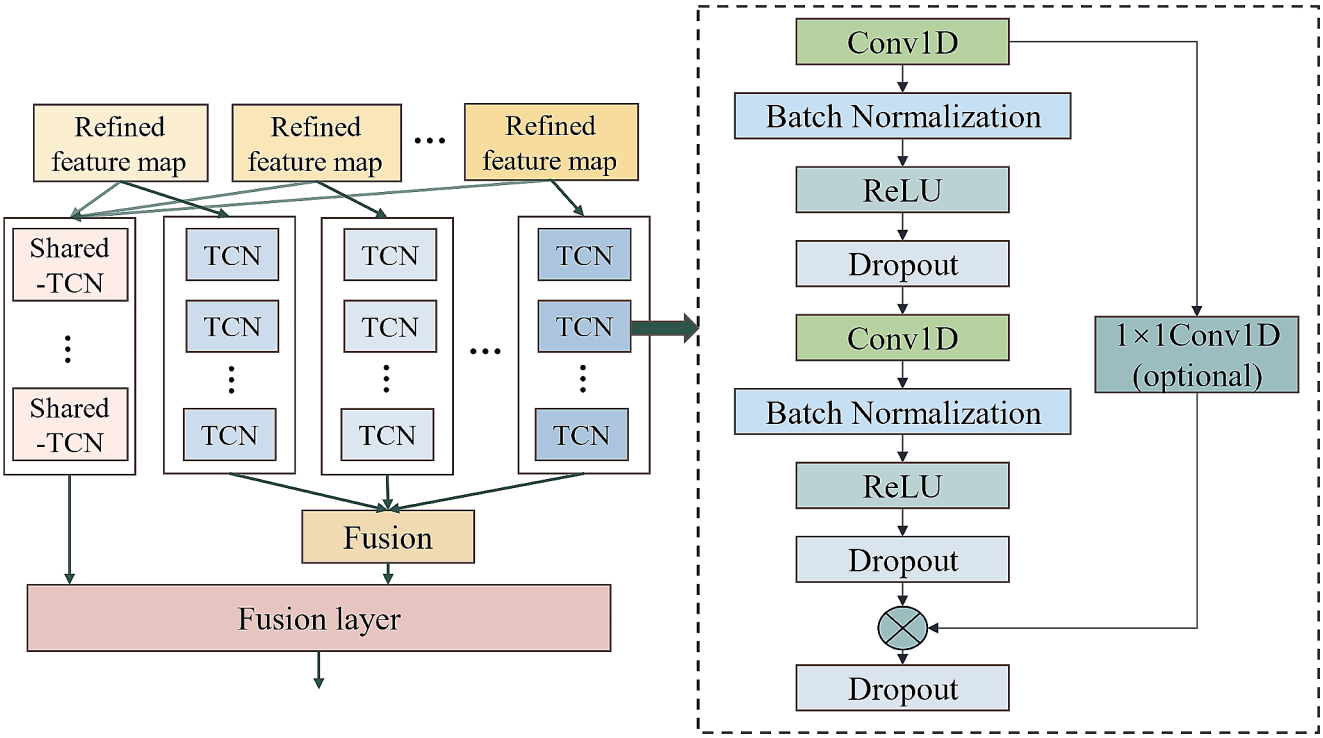


Fig. 4 The architecture of multi-scale residual block

Table 3 Model parameters of multi-scale residual block

Branch	Layer	filters	size	Output
TCN 1 ($d=2$)	Input			(C, T)
	Conv1D	F_3	(2, 1)	(C, F_3)
	BatchNorm			(C, F_3)
	Activation			(C, F_3)
	Dropout			(C, F_3)
	Conv1D	F_3	(2, 1)	(C, F_3)
	BatchNorm			(C, F_3)
	Activation			(C, F_3)
	Dropout		(16, 1)	(C, F_3)
	Residual-2	F_3	(2, 1)	(C, F_3)
TCN 2 ($d=3$)	Residual-1	F_3	(4, 1)	(C, F_3)
	Residual-2	F_3	(4, 1)	(C, F_3)
	Residual-3	F_3	(4, 1)	(C, F_3)
TCN 3 ($d=2$)	Residual-1	F_3	(6, 1)	(C, F_3)
	Residual-2	F_3	(6, 1)	(C, F_3)
Shared-TCN ($d=2$)	Residual-1	F_4	(4, 1)	(C, F_4)
	Residual-2	F_4	(4, 1)	(C, F_4)

*C and T are the number of channels and time points, respectively. $F_3=16$ and $F_4=32$ correspond to the number of temporal filters for different sub-feature matrixes. d is the number of residual block stacks. The dropout rate is set to 0.2, 0.1, 0.2 and 0.2, and activation function is set to ReLU.

where $R(F(X_j^m))$ denotes the sub-feature of the jth subject in the mth sliding window, w denotes the weight, and k denotes the indexed variable. Table 3 demonstrates the specific parameter settings for the multi-scale residual block.

Dataset and pre-processing

Dataset description

This study evaluated the performance on BCI Competition 2a Dataset that contains 25-channels EEG data with a sampling rate of 250 Hz (Brunner et al. 2008). The experiment utilizes a 25-channel EEG acquisition equipment, capturing data at a sample rate of 250 Hz. The experimental paradigm consisted of four different motor imagery tasks, left hand, right hand, feet, and tongue motor imagery. The dataset has a total of nine subjects doing two sessions on different days and each session contains 288 trials (72 per class).

Preprocessing

The data of a single trial and label are represented as $E_j \in \mathbf{R}^{C\times T}$ and $Y_j \in \{1, 2, 3, 4\}$, where C denotes channel number and T denotes time point number. The 4.5s EEG of motor imagery are selected and three ophthalmoscope channels are removed, yielding $C=22$ and $T=1125$.

The trials from the two sessions of each subject are mixed and 288 are randomly taken as the training set and

the remaining 288 as the test set. All trials are standardized to the function:

$$X = (E - mean(E))/std(E), i = 1,2,\ldots,C \quad (9)$$

Control algorithms

This study compares MSHANet with EEGNet, EEGTC-Net, ShallowNet, DeepConvNet, EEGNeX, TCNFusion, ATCNet to demonstrate its performance. EEGNet is a compact convolutional neural network to extract temporal and spatial features, achieving a smaller number of parameters and computational complexity, which gives better results in EEG feature extraction (Lawhern et al. 2018). EEGTC-Net introduces TCNs based on EEGNet to further utilize temporal information with good generalization (Ingolfsson et al. 2020). ShallowNet is a shallow convolutional neural network designed and implemented for EEG with good classification performance under multiple EEG paradigms (Schirrmeister et al. 2017). DeepConvNet is a deep convolutional neural network model that consists of a combination of multiple convolutional, pooling, and fully-connected layers for extracting more complex features in EEG (Schirrmeister et al. 2017). EEGNeX is an improvement on EEGNet (Chen et al. 2024). TCNetFusion is an improved network based on the EEGTCNet (Musallam et al. 2021). ATCNet is a multi-branch convolutional neural network model based on the attention mechanism, which combines the attention mechanism to improve the perceptual ability of the model and achieve superior performance in EEG classification (Altaheri et al. 2022). Both the proposed method and the comparison method use Adam as the optimizer, the batchsize is set to 64, the learning rate is set to 5e-4, and the number of iterations is fine-tuned with the different methods. The proposed MSHANet and other comparison model is implemented in tensorflow 2.7 by using an Intel Xeon Platinum 8255 C CPU (2.50 GHz) and an NVIDIA RTX 3080 GPU with 10 GB RAM.

Evaluation metrics

Accuracy and *kappa* are applied to evaluate the performance. *Kappa* is applicable to measure the multiclassification model, usually the higher the *kappa* value, the better the model classification performance, it is calculated as Eq. (10) and Eq. (11):

$$kappa = (Accuracy - p_e)/(1 - p_e) \quad (10)$$

$$p_e = \left(\sum_{i=1}^{N_c} n_{:r} * n_{r:}\right)/M^2 \quad (11)$$

Receiver operating characteristic (ROC) curve presents one category as a positive category and the remaining three categories as negative categories. The area under the curve (AUC) is used to judge how good the classification is, the larger the AUC the greater the sensitivity and specificity of the model, the better the model performance. The sample standard deviation (*std*) is a statistic used to measure the degree of dispersion of a set of data. It is calculated as shown in Eq. (12):

$$std = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}} \quad (12)$$

where x_i denotes the observations of the sample, $\overline{x}$ is the mean value and n denotes the number of samples.

Result

Ablation

In this study, ablation experiments are used to verify the validity of the hybrid attention block and shared-TCN. Two models are built by adding different types of attention mechanisms at the same location of the model, namely, MSANet-MHA, MSANet-SE. MSNet represents a multiscale residual model containing the hybrid attention block. MSNet represents a multi-scale residual model containing hybrid attention blocks. S*j* denotes subject *j*. Fig. 5 uses a topographic map to show more clearly the effect of different types of attention mechanisms on the classification performance of the model. The classification accuracy of MSANet-SE for S2 is less that of MSANet-MHA and MSNet by 5.55% and 6.94%, while the average accuracies are less by 1.04% and 2.28%, respectively. MSANet-MHA 's classification accuracy at S8 is lower than MSNet's by 1.73%, and although it is higher than MSNet's by 2.09% at S1, its average accuracies for multiple subjects are still lower than MSNet's. Therefore, the hybrid attention block is more effective in capturing more information than MHA and SE can capture more feature information and improve the classification performance of the model compared to MHA and SE.

As can be seen from Fig. 6, by comparing the model MSHANet with the addition of shared-TCN to the model MSNet without the addition of shared-TCN, MSHANet outperforms MSNet in terms of classification accuracy in most subjects. Classification accuracy improves for more than two-thirds of the subjects, with an average classification accuracy increase of 0.81%. Among them, S1, S4 and S9 show better classification accuracy improvement, with 3.48%, 2.43% and 1.04% respectively. The improvement in

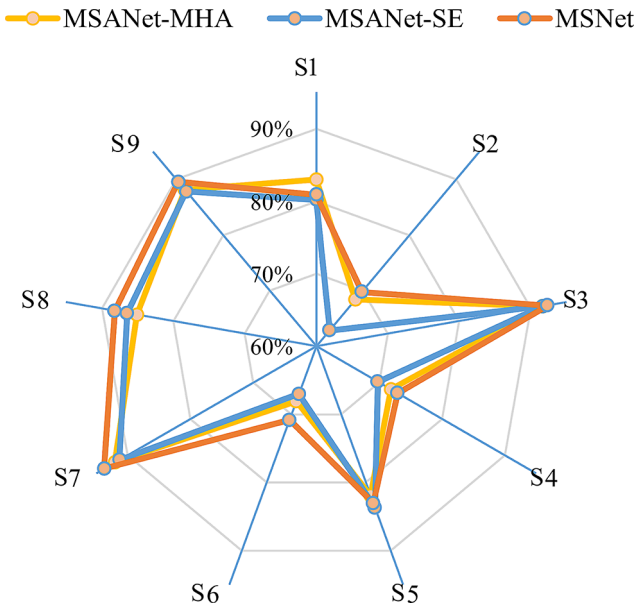


Fig. 5 Comparison of attention and hybrid attention mechanisms in terms of classification performance

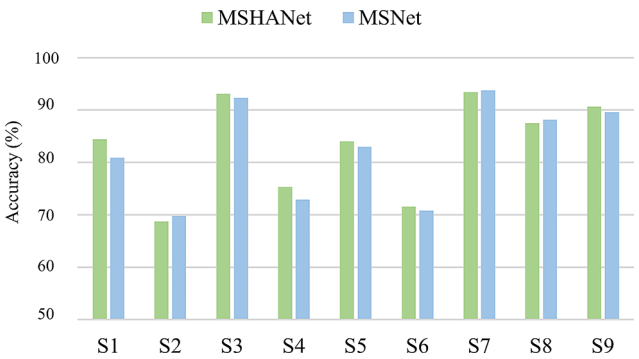


Fig. 6 Classification performance comparison of shared and no-shared TCNs

classification performance is achieved without increasing the number of parameters, which indicates that the shared TCN can effectively extract common features under different time windows and retain the invariant features when time changes.

Performance comparison of session- specific

Accuracy and *kappa* of MSHANet

Four-classification accuracy is used to measure the decoding accuracy of MSHANet and seven models in motor imagery EEG, as shown in Fig. 7. MSHANet outperforms the baseline with a higher average accuracy of nine subjects than the other seven methods. MSHANet shows an improvement of more than 15% compared to DeepConvNet, ShallowNet, and EEGNeX. Compared to EEGNet,

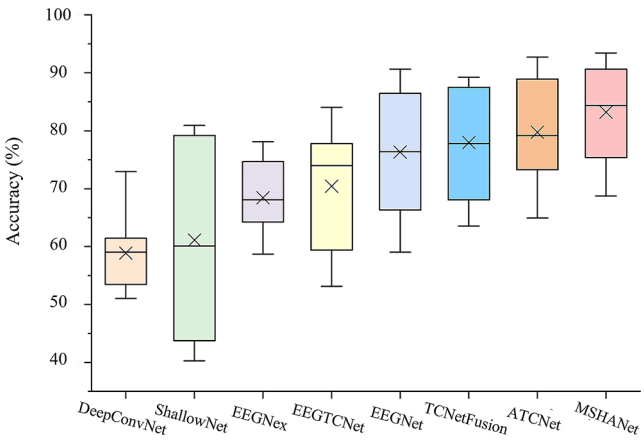


Fig. 7 Comparison of the accuracy of MSHANet and seven baseline model

EEGTCNet and TCNet-Fusion there is an improvement of 12.77%, 6.87% and 5.21% respectively, and it can be seen from the box plots that MSHANet are more concentrated. The degree of concentration in ATCNet is comparable to that of MSHANet. However, the median of MSHANet is significantly higher than that of ATCNet, and the minimum and maximum observations also outperform ATCNet. Therefore, the decoding model can achieve good classification performance for time-varying EEG in four-classification MI decoding. As can be seen from Table 4, six subjects have accuracy rates higher than 84%. Both *kappa* and accuracy are optimal among all subjects, which indicates that the model has good classification performance. In addition, among the baselines with classification accuracies higher than 70%, the standard deviation of the model is 14.29% lower compared to ATCNet and TCNetFusion, and 25% lower compared to EEGNet. This demonstrates that the model is robust and can maintain relatively high accuracy for different subjects. Therefore, MSHANet is a robust model that achieves high accuracy across multiple subjects.

Feature distribution

To visualize how MSHANet feature extraction performance, Fig. 8 used t-Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction technique, to show the feature distribution of the eight methods after going through the feature extraction network by embedding high-dimensional data in a two- or three- dimensional space. Cyan indicates the left hand, magenta the right hand, black the foot, and red the tongue. In each subfigure, the feature effects of the different methods can be seen for the four types of motor imagery. Firstly, in terms of sample density, after ShallowNet, DeepConvNet and EEGNeX extraction, the features of all four types are sparse and cannot be completely distinguished. While EEGNet, EEGTCNet,

Table 4 Classification accuracy and *kappa* of different subjects under different models.

Models	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average ± Std
Deep ConvNet	52.08	51.04	54.51	60.76	72.92	59.03	64.58	61.46	53.47	58.87 ± 0.07
	0.37	0.35	0.40	0.47	0.64	0.46	0.53	0.49	0.39	0.45 ± 0.09
Shallow ConvNet	60.07	43.75	79.86	48.26	41.32	40.28	76.39	79.17	80.90	61.11 ± 0.18
	0.47	0.26	0.73	0.31	0.22	0.20	0.68	0.72	0.75	0.48 ± 0.24
EEGNeX	68.06	59.03	70.83	64.24	75.69	58.68	74.65	66.67	78.12	68.44 ± 0.07
	0.57	0.46	0.61	0.52	0.68	0.45	0.66	0.55	0.71	0.58 ± 0.09
EEGTCNet	73.96	54.17	84.03	59.38	70.14	53.12	84.03	77.78	77.08	70.41 ± 0.12
	0.65	0.39	0.79	0.46	0.60	0.37	0.79	0.70	0.69	0.61 ± 0.16
EEGNet	76.39	61.46	88.54	66.32	74.31	59.03	90.62	83.68	86.46	76.31 ± 0.12
	0.68	0.49	0.85	0.55	0.66	0.46	0.87	0.78	0.82	0.68 ± 0.16
TCNet-Fusion	77.78	64.93	87.50	68.06	76.39	63.54	89.24	86.46	87.85	77.97 ± 0.10
	0.70	0.53	0.83	0.57	0.69	0.51	0.86	0.82	0.84	0.71 ± 0.14
ATCNet	77.78	64.93	88.89	73.26	79.17	65.28	92.71	85.76	89.93	79.75 ± 0.10
	0.70	0.54	0.85	0.64	0.72	0.54	0.90	0.81	0.87	0.73 ± 0.14
MSHANet (Proposed)	**84.38**	**68.75**	**93.06**	**75.35**	**84.03**	**71.53**	**93.40**	**87.50**	**90.62**	**83.18 ± 0.09**
	0.79	**0.58**	**0.91**	**0.67**	**0.79**	**0.62**	**0.91**	**0.83**	**0.88**	**0.78 ± 0.12**

*The top of each column in the table indicates accuracy, the bottom indicates *kappa*. "std" stands for sample standard deviation

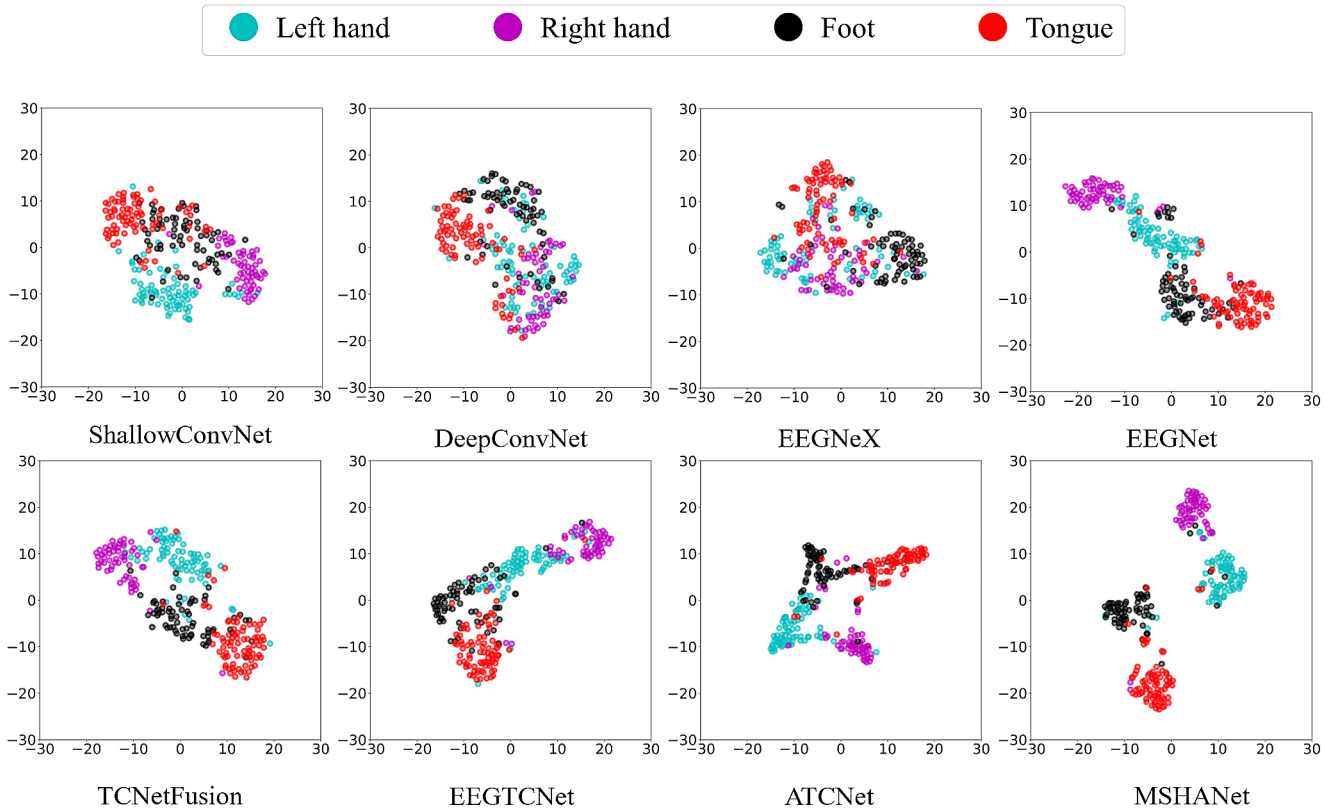


Fig. 8 Distribution of features of the eight models for four types of motor imagery

TCNet-Fusion, ATCNet and MSHANet extracted features are more concentrated, with the right hand features being the densest. In contrast, ATCNet and MSHANet have more aggregated features for all the four types. From the observation of category distribution, EEGNet, EEGTCNet, TCNet-Fusion and ATCNet have overlapping regions in the four types of features without a clear demarcation line. As for the MSHANet, it performs well and is clearly distinguishable in the feature space. In addition, in each subgraph, the left hand features are close to the right hand features, which indicates that the left hand features are more correlated with the right hand features. Therefore, MSHANet performs

well in multiple perspectives, which means that MSHANet is capable of extracting EEG features with differentiability from time-varying data.

Performance of MSHANet

Confusion matrix of MSHANet

Figure 9 shows the mean confusion matrix for 9 subjects in MSHANet and the confusion matrices under 3 subjects, S1, S5, and S9, respectively. The confusion matrices include the predictions for the categories of left-handed, right-handed, foot, and tongue, and the number of correct classifications is indicated on the diagonal. Due to individual differences, there are large variations between the confusion matrices of different subjects, but the overall level remained high. The precision for the four categories for all subjects is 81.58%, 84.28%, 85% and 81.97%, respectively. The highest accuracy is found for the classification of foot, which may be explained by the fact that subjects imagined foot more clearly and produced more distinctive motor imagery EEG features. In the classification of the four motor imagery types, the darker the diagonal of the confusion matrix, the greater the number of samples correctly classified. Among them, the precision of S3 for both right and left hand classification reach over 94%, and the precision of foot and tongue is over 90%. S7 and S8 had excellent classification results with precision of over 86% for all four categories, which signified that the model can accurately identify the motor imagery types.

ROC curves of MSHANet

The performance of MSHANet on different categories can be visualised from the ROC curves. The overall performance of the classifier is evaluated by the AUC, and the closer to 1 indicates the better performance of the classifier. As can be seen from Fig. 10, the average AUC area of the four

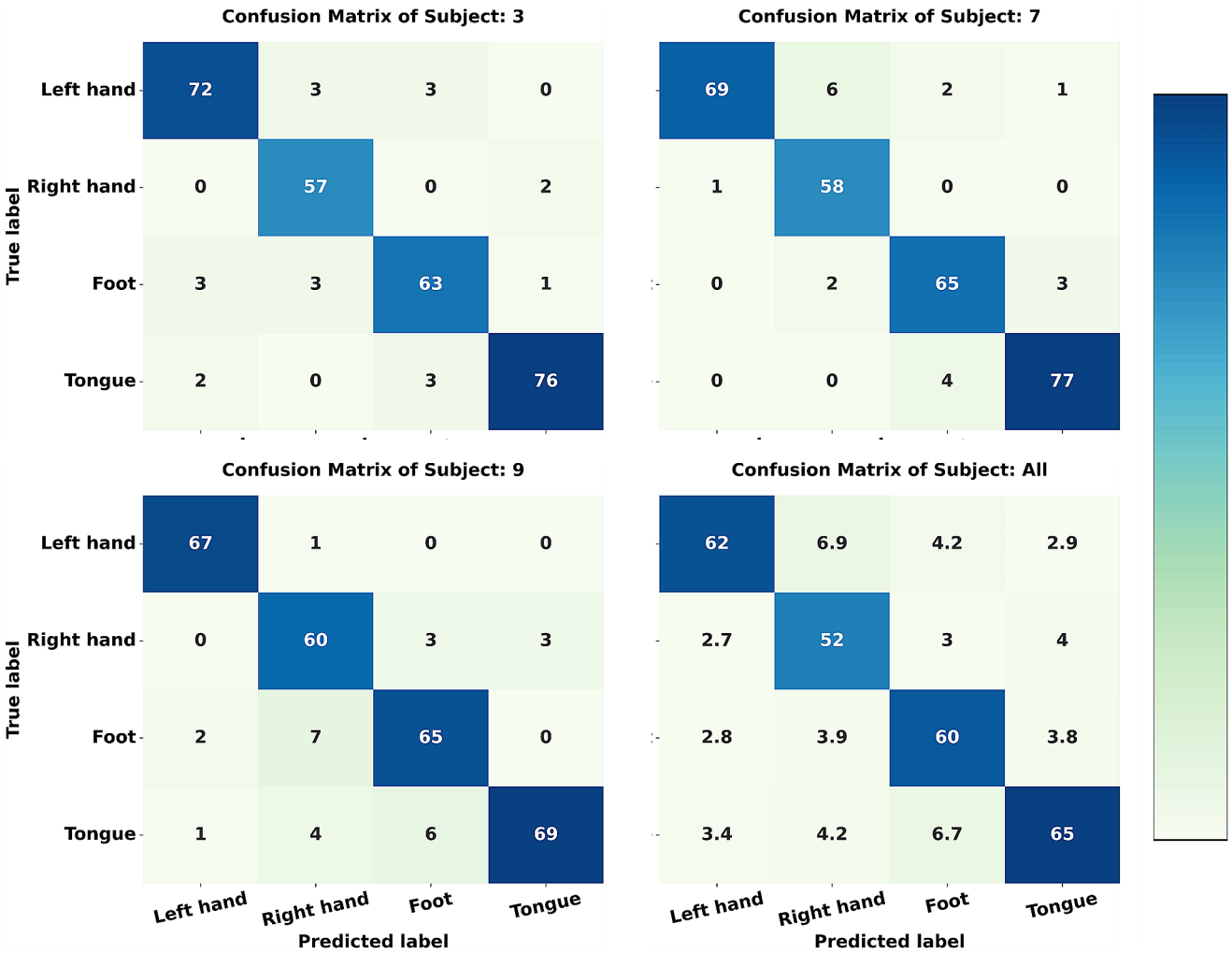
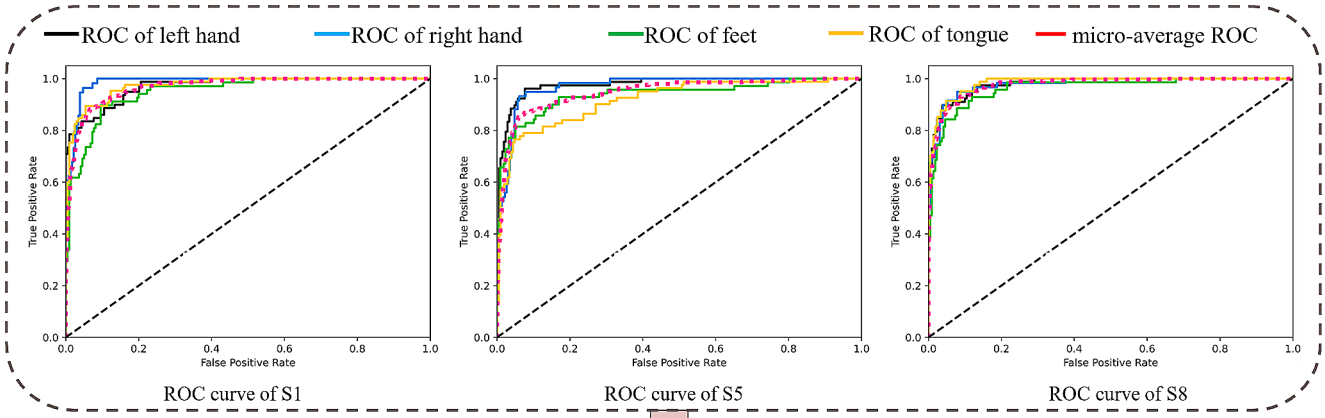


Fig. 9 Confusion matrices for S3, S7, S9 and all subjects



AUC of subject	S1	S5	S8
left hand	0.97	0.98	0.98
right hand	0.99	0.97	0.98
feet	0.95	0.94	0.96
tongue	0.97	0.92	0.98
Micro-average	0.97	0.95	0.98

Fig. 10 ROC curves for S1, S5 and S8

Table 5 Classification accuracy and *kappa* of different subjects under different models.

Models	S1	S2	S3	S4	S5	S6	S7	S8	S9	Average ± Std
TCNet	78.82	61.11	88.89	64.93	72.22	59.38	85.76	79.51	78.47	74.34 ± 0.09
Fusion	0.74	0.49	0.87	0.59	0.68	0.53	0.76	0.74	0.75	0.66 ± 0.12
ATCNet	80.21	61.81	89.93	69.44	75.69	64.93	82.29	80.56	81.60	76.27 ± 0.11
	0.72	0.48	0.85	0.53	0.63	0.46	0.81	0.73	0.71	0.68 ± 0.14
MSHANet	**81.94**	**68.40**	**92.01**	**75.69**	**76.74**	**67.36**	**88.54**	**83.68**	**86.46**	**80.09 ± 0.09**
(Proposed)	**0.76**	**0.58**	**0.89**	**0.68**	**0.69**	**0.56**	**0.85**	**0.78**	**0.82**	**0.73 ± 0.12**

*The top of each column in the table indicates *accuracy*, the bottom indicates *kappa*. "std" stands for sample standard deviation

categories of S1,S5 and S8 are all higher than 0.95, which means that MSHANet has a better classification effect. And the AUC of the left and right hand motor imagery of the three subjects is larger than that of the other two categories, which means that MSHANet performs better on hand motor imagery. The AUC of right-handed motor imagery of S1 is larger than the other three categories of motor imagery, so MSHANet performs better for right-handed motor imagery of S1. Secondly, looking at the ROC curves of the four categories of the three subjects in the figure, the closer to the upper left corner, the better the classification effect is. The ROC curves of the left hand and right hand motor imagery in S5 are closer to the upper left corner than those of the foot and the tongue, which means that it is easier to differentiate between the left hand and the right hand in the EEG characteristics of S5. From the ROC curve of S8, it can be seen that MSHANet performed very well in all four categories, so the average ROC curve is closer to the upper left corner than the rest of the subjects. Therefore, MSHANet is a robust model for different subject and different categories.

MSHANet-performance of cross session

Experiments are executed across session using one session as the training dataset and another session as the test dataset from BCI competition IV 2a. TCNet-Fusion and ATCNet, which perform better in the session- specific task, are selected for comparison, as shown in Table 5. Accuracy and *kappa* are used to evaluate performance after fine-tuning. MSHANet outperformed TCNet-Fusion and ATCNet by 5.75% and 3.82%, respectively, in average accuracy across all subjects with the least variance. Five subjects have classification accuracies higher than 80%, with S3 having the highest of 92.01%.

Kappa is higher than the two sota methods in all subjects with low variance. This indicates that MSHANet is still able to maintain robust decoding across sessions, and time

variability has less effect on it. Compared with session- specific task, the decoding performance of MSHANet across session remains stable. And the cross-session MSHANet proves its stability and robustness in decoding time-varying data.

Discussion

In this study, MSHANet achieves robust decoding in time-varying data by combining a hybrid attention mechanism and a multi-scale residual block. At first, the features in each branch are used to focus on the important information of the EEG features using a hybrid attention block that combines MHA and SE. MHA focuses on the information in multiple dimensions of the EEG features through different heads, which effectively solves the problem of overconcentration of self-attention mechanism. The combined SE effectively focuses on the important channels in the multi-channel EEG, and MSHANet can optimize the extraction of the important information of the EEG with smaller computational resources. After fully obtaining the important information in the EEG, a multi-scale residual block is used to further extract more detailed features. The local feature kernels extracted by different convolutional kernel sizes are fused with global features to maximize exploitation, and combined with the residual network in the deep part of the decoding model to effectively solve the network degradation. In addition, the parameters of the shared partial branches effectively extract common information and reduce the number of parameters by linking the information between different branches.

Recently, several researchers have investigated MI decoding in terms of attentional mechanisms and multi-scale architectures to evaluate model performance on BCI competition IV 2a. On the one hand, through the attention mechanism the model can be constrained to focus on the features that are in focus, Zhang proposed a self-attention based CNN (TFCSP) that automatically extracts spatial-temporal information from EEG, obtaining a classification accuracy of 79.28% (Zhang et al. 2023). This is the same as the conclusion obtained in this study through ablation experiments which found that attention plays a key role in improving the robustness of decoding, and therefore MSHANet improves with the addition of the hybrid attention block. On the other hand, multi-scale architecture is an important factor that cannot be ignored. Altuwaijri combined a multi-branch convolutional neural network with SE attention for motion image decoding, and achieved 82.87% and 96.8% accuracy on BCI Competition IV2a and high Gamma datasets, respectively (Altuwaijri et al. 2022). Compared with them, the present study provides a greater improvement in extracting rich EEG features at different scale sizes and fusing local and global information. In cross-session experiments, Liu used a multi-scale time-periodic convolutional neural network with center loss (FBMSNet) to improve intra-class compactness and inter-class separability with an accuracy of 79.17% (Liu et al. 2022). The difference with other methods is that MSHANet can decode more detailed features through the combination of the attention mechanism and the multi-scale residual block, which effectively enhances the robustness of MI-BCI.

However, there is still room for improvement in this study. Firstly, the current model has a high number of parameters, which leads to higher computational costs. In the future, increasing model pruning to achieve lightweight network structures will help accelerate the development of BCI in practical applications. Secondly, the model has been validated on a single dataset only, which may result in the degree of generalization of the model not being validated. Future validation on real rehabilitation systems will be conducted to improve the reliability and usefulness of the method.

Conclusion

In this study, the MSHANet is proposed to decode four types of MI by combining the hybrid attention block and the multi-scale residual block.To solve the robust decoding in time-varying signals, MHA and SE attention are combined into a hybrid attention block, which adaptively focuses on the important information in EEG. Then connecting the multi-scale residual block containing shared parameters contributes to the extraction of richer EEG information, information sharing among different branches and reduction of network parameters. MSHANet achieves high accuracy even cross session, demonstrating the superiority of the method. The result indicates that the hybrid attention combined with multi-scale can robustly decode time-varying EEG, which is significant in saving time cost and promoting BCI.

Acknowledgements This work is supported in part by the Natural Science Foundation of Hebei Province (Grant Nos. F2021202003), the Technology Nova of Hebei University of Technology (Grant Nos. JBKYXX2007), the National Natural Science Foundation of China (62176090), and the Key Research and Development Foundation of Hebei (21372002D).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Altaheri H, Muhammad G, Alsulaiman M (2022) Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. IEEE Trans Industr Inf 19(2):2249–2258

Altuwaijri GA, Muhammad G, Altaheri H et al (2022) A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for eeg-based motor imagery signals classification. Diagnostics 12:995

Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv Preprint arXiv :180301271

Broccard FD, Mullen T, Chi YM et al (2014) Closed-loop brain–machine–body interfaces for noninvasive rehabilitation of movement disorders. Ann Biomed Eng 42:1573–1593

Brunner C, Leeb R, Müller-Putz G et al (2008) BCI Competition 2008–Graz data set A. Institute for Knowledge Discovery (Laboratory of Brain-Computer interfaces). Graz Univ Technol 16:1–6

Chen C, Yu X, Belkacem AN et al (2021) EEG-based anxious states classification using affective BCI-based closed neurofeedback system. J Med Biol Eng 41:155–164

Chen X, Teng X, Chen H et al (2024) Toward reliable signals decoding for electroencephalogram: a benchmark study to EEGNeX. Biomed Signal Process Control 87: 105475

Craik A, He Y, Contreras-Vidal JL (2019) Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 16:031001

Dose H, Møller JS, Iversen HK et al (2018) An end-to-end deep learning approach to MI-EEG signal classification for BCIs. Expert Syst Appl 114:532–542

Duan F, Lin D, Li W et al (2015) Design of a multimodal EEG-based hybrid BCI system with visual servo module. IEEE Trans Auton Ment Dev 7:332–341

Feng X, Cong P, Dong L et al (2023) Channel attention convolutional aggregation network based on video-level features for EEG emotion recognition. Cogn Neurodyn: 1–19

Gao X, Wang Y, Chen X et al (2021) Interface, interaction, and intelligence in generalized brain–computer interfaces. Trends Cogn Sci 25:671–684

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. Proc IEEE Conf Comput Vis Pattern Recognit 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

Ingolfsson TM, Hersche M, Wang X et al (2020) EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE: 2958–2965. https://doi.org/10.1109/SMC42975.2020.9283028

Jareda MK, Sharma R, Kukker A (2019) EEG signal based seizure classification using wavelet transform. 2019 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE: 537–539

Jia H, Yu S, Yin S et al (2023) A model combining Multi Branch spectral-temporal CNN, efficient Channel attention, and LightGBM for MI-BCI classification. IEEE Trans Neural Syst Rehabilitation Eng 31:1311–1320

Ju J, Feleke AG, Luo L et al (2022) Recognition of drivers' hard and soft braking intentions based on hybrid brain-computer interfaces. Cyborg Bionic Syst. https://doi.org/10.34133/2022/9847652

Ko W, Jeon E, Jeong S et al (2021) Multi-scale neural network for EEG representation learning in BCI. IEEE Comput Intell Mag 16:31–45

Kukker A, Sharma R (2021) A genetic algorithm assisted fuzzy Q-learning epileptic seizure classifier. Comput Electr Eng 92:107154

Kukker A, Sharma R, Mishra O et al (2023) Epileptic seizure classification using fuzzy lattices and Neural Reinforcement Learning. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization: 1–9. https://doi.org/10.1080/21681163.2023.2290361

Lawhern VJ, Solon AJ, Waytowich NR et al (2018) EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. J Neural Eng 15:056013

Li H, Ding M, Zhang R et al (2022) Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. Biomed Signal Process Control 72:103342

Li H, Zhang D, Xie J (2023a) MI-DABAN: a dual-attention-based adversarial network for motor imagery classification. Comput Biol Med 152:106420

Li M, Wei R, Zhang Z et al (2023b) CVT-based asynchronous BCI for brain-controlled robot navigation. Cyborg Bionic Syst 4:0024. https://doi.org/10.34133/cbsystems.0024

Liu K, Yang M, Yu Z et al (2022) FBMSNet: a filter-Bank Multi-scale convolutional neural network for EEG-Based motor imagery decoding. IEEE Trans Biomed Eng 70:436–445

Luo T, Zhou C, Chao F (2018) Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. BMC Bioinform 19:1–18

Ma X, Chen W, Pei Z et al (2023) A temporal dependency learning CNN with attention mechanism for MI-EEG decoding. IEEE Trans Neural Syst Rehabilitation Eng 31:3188–3200

Mane R, Chouhan T, Guan C (2020) BCI for stroke rehabilitation: motor and beyond. J Neural Eng 17:041001

Musallam YK, AlFassam NI, Muhammad G et al (2021) Electroencephalography-based motor imagery classification using temporal convolutional network fusion. Biomed Signal Process Control 69:102826

Roesch J, Vetter D, Baldassarre A et al (2024) Individualized treatment of motor stroke: a perspective on open-loop, closed-loop and adaptive closed-loop brain state-dependent TMS. Clin Neurophysiol 158:204–211

Roy AM (2022) An efficient multi-scale CNN model with intrinsic feature integration for motor imagery EEG subject classification in brain-machine interfaces. Biomed Signal Process Control 74:103496

Said A, Göker H (2023) Spectral analysis and Bi-LSTM deep network-based approach in detection of mild cognitive impairment from electroencephalography signals. Cogn Neurodyn 1–18. https://doi.org/10.1007/s11571-023-10010-y

Schirrmeister RT, Springenberg JT, Fiederer LDJ et al (2017) Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 38:5391–5420

Shajil N, Mohan S, Srinivasan P et al (2020) Multiclass classification of spatially filtered motor imagery EEG signals using convolutional neural network for BCI based applications. J Med Biol Eng 40:663–672

Shen J, Zhan Y, Liang H et al (2023) Depression recognition from EEG signals using an adaptive channel fusion method via improved focal loss. IEEE J Biomed Health Inf 27:3234–3245

Sun B, Zhang H, Wu Z et al (2021) Adaptive spatiotemporal graph convolutional networks for motor imagery classification. IEEE Signal Process Lett 28:219–223

Tao C, Gao S, Shang M et al (2018) Get the point of my Utterance! Learning towards effective responses with multi-head attention mechanism. IJCAI:4418–4424

Wang C, Wu Y, Wang C et al (2022) MI-EEG classification using Shannon complex wavelet and convolutional neural networks. Appl Soft Comput 130:109685

Wu R, Jin J, Daly I et al (2023) Classification of motor imagery based on multi-scale feature extraction and the channeltemporal

attention module. IEEE Trans Neural Syst Rehabilitation Eng 31:3075–3085

Xie J, Zhang J, Sun J et al (2022) A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification. IEEE Trans Neural Syst Rehabilitation Eng 30:2126–2136

Xu L, Xu M, Jung TP et al (2021) Review of brain encoding and decoding mechanisms for EEG-based brain–computer interface. Cogn Neurodyn 15:569–584

Zhang G, Luo J, Han L et al (2021) A dynamic multi-scale network for EEG signal classification. Front Neurosci 14:578255

Zhang R, Liu G, Wen Y et al (2023) Self-attention-based convolutional neural network and time-frequency common spatial pattern for enhanced motor imagery classification. J Neurosci Methods 398:109953

Zhao J, Shi Y, Liu W et al (2023) A hybrid method fusing frequency recognition with attention detection to enhance an asynchronous brain-computer interface. IEEE Trans Neural Syst Rehabilitation Eng

Zhi H, Yu Z, Yu T et al (2023) A Multi-Domain Convolutional Neural Network for EEG-Based Motor Imagery Decoding. IEEE Trans. Neural Syst. Rehabilitation Eng, 2023