IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 April 2024, accepted 29 May 2024, date of publication 5 June 2024, date of current version 12 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3410036

RESEARCH ARTICLE

Multi-Scale Convolutional Attention and Riemannian Geometry Network for EEG-Based Motor Imagery Classification

BEN ZHOU[1], **LEI WANG**[2], **WENCHANG XU**[2], **AND CHENYU JIANG**[2]

[1]Shandong University of Traditional Chinese Medicine, Jinan 250355, China
[2]Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China

Corresponding author: Chenyu Jiang (jingcy@sibet.ac.cn)

This work was supported in part by the Department of Science and Technology of Shandong Provincial under Grant ZR2020QF018, in part by the Quancheng 5150 Project, in part by the Taishan Industrial Experts Program under Grant tscx202306125, and in part by the National Key Research and Development Program under Grant 2022YFC2406204.

ABSTRACT The electroencephalogram (EEG) is a non-invasive technique with high temporal resolution that has become the research frontier of brain-computer interface (BCI) systems. It is widely used in medical rehabilitation, gaming, and other industries. However, decoding EEG signals remains a challenging task. A network called MSCARNet, which combines multi-scale convolution and Riemannian geometry, was proposed for classifying motor imagery based on EEG. The network is supplemented by an attention mechanism and sliding window technique. The MSCARNet utilizes sliding windows to expand data dimensions and multiple convolution kernels to obtain spatial and temporal features. These features are then mapped to Riemannian space and undergo bilinear mapping and logarithmic operations for dimensionality reduction. This approach is beneficial in reducing the impact of noise and outliers and provides convenience for classification. Subject-dependent and subject-independent experiments were conducted using the BCI-IV-2a dataset to validate the effectiveness of the MSCARNet. The results show that the accuracy improved by approximately 4% compared to existing state-of-the-art methods. The hybrid network based on Riemannian space can effectively improve the accuracy of EEG motor imagery classification tasks without excessive preprocessing.

INDEX TERMS Electroencephalogram, deep learning, convolution neural network, motor imagery, Riemannian geometry.

I. INTRODUCTION

Brain-computer interface (BCI) has become a frontier technology and is widely used in industry, often directed at researching, mapping, assisting, augmenting, or repairing human cognitive or sensory-motor functions [1]. Motor imagery (MI) represents a key area of focus in the field of BCI, which is defined as a mental simulation of movement without any actual physical execution. A person performing motor imagery generates an electroencephalogram (EEG) signal of event-related potentials, which can be decoded to capture the person's intention. This technology has significant potential for applications in both medical and non-medical fields, including neurorehabilitation, neuroprosthetics and gaming [2].

EEG is used to record cognitive-behavioural changes in the brain and can be classified as invasive or non-invasive according to the recording method. Invasive methods can accurately capture the potential signals of the corresponding brain regions but require the implantation of a chip in conjunction with craniotomy, which is costly and risky. Non-invasive, portable, low-cost, low-risk, but more susceptible

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

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

to noise and low signal-to-noise ratio, resulting in delayed non-invasive research results and current limitations to classification of a few. The decoding of EEG signals enables the classification of simple intentions, such as with the left and right hands and feet. This technology offers considerable assistance to individuals with limb movement limitations, and its applications in various fields are promising. However, the algorithms used for decoding MI-EEG signals still need to be improved in terms of performance, generalization, and lightweight to be suitable for various industrial scenarios. The main challenge faced by decoding algorithms is to accurately identify human intentions from unstable EEG signals with low signal-to-noise ratio and various artifacts, including biological artifacts such as muscle movement, eye movement, and heart rate, as well as non biological artifacts such as electronic devices and environmental noise. This makes decoding EEG signals a challenging task.

Researchers generally solve the above problems through traditional machine learning (ML) or deep learning (DL) techniques. Among traditional machine learning algorithms that rely on manual feature extraction, the filter library common space pattern (FBCSP) [3] and its variants have the best performance in MI-EEG classification. Compared to traditional machine learning algorithms, deep learning algorithms can automatically extract potential features of signals, thereby compensating for the time-consuming and labor-intensive shortcomings of manual feature exploration. The DL algorithm has therefore been widely applied in various scenarios, including image [4], natural language processing [5], audio and video processing [6]. The number of papers in which researchers have used DL to classify MI tasks has increased rapidly over the past few years due to the excellent performance of DL algorithms in a variety of applications [7]. Different DL-based frameworks are used for MI classification, such as Recurrent Neural Network (RNN) [8], [9], Deep Belief Network (DBN) [10], Auto-encoder (AE) [11], Convolutional Neural Network (CNN) [12], [13], [14], [15] and hybrid DL model [12], [16], of which CNN and its mixture model are the most efficient. With the model based on Riemannian geometry gaining popularity in the field of image processing, Kim et al. [17] constructed a Riemannian classifier based on Fisher geometric minimum distance to the mean (FgMDM) for EEG signal classification in the post traumatic stress disorder (PTSD) resting state fMRI and achieved a classification accuracy of about 75%. Bakhshali et al. [18] classified the image speech by calculating the Riemannian distance of the correlation density (CSD) matrix of different channels of EEG signal and obtained an accuracy of about 90%. Chu et al. [19] used a Riemannian geometric framework containing Riemannian distances and Riemannian means to extract tangent space features from the spatial covariance matrix of a 6-categorical MI-EEG trial, and then used least squares to downscale the features, and finally inputting a linear discriminant analysis and a support vector machine classifier yielded a classification accuracy of about 80%, which validates the potential of Riemannian geometry for brain-computer interface classification tasks. However, these methods require a more professional background and a large number of experiments to find features manually, and the features obtained may not truly reflect the characteristics of the data, which significantly affects the actual effect of MI-based Brain-computer interface.

Huang and Gool [20] designed a learnable Riemann network (SPDNet) based on the symmetric positive-definite matrix for visual classification tasks, which solved the problem of complex manual feature extraction. Gao et al. [21] classified EEG-based motor imagery tasks through the hybrid of CNN and SPDNet. The results show that CNN has overwhelming advantages over pure CNN and pure Riemannian classifiers compared with the mixture model of Riemannian geometry. However, the design of its convolution layer has not fully captured enough features, resulting in a slight lack of accuracy. Inspired by it, this paper combined with sliding windows, multi-scale convolution, attention mechanism and Riemannian geometry proposes the MSCARNet for EEG-based motor imagery classification, to improve the classification performance.

The availability of publicly accessible EEG motor imagery datasets from a multitude of organisations has significantly contributed to the advancement of related research. The datasets provided by various BCI competitions have gained considerable popularity for research purposes, with the BCI Competition IV series released in 2008 representing a significant milestone. Furthermore, datasets such as BCI-IV-2a and BCI-IV-2b have emerged as highly cited datasets in current research. These datasets collect EEG data from multiple subjects engaged in various motor imagery tasks and provide a benchmark for the evaluation of classification models, with BCI-IV-2a representing a significant opportunity for improvement in classification accuracy.

The main contributions of this paper are as follows: (1) A novel hybrid deep learning method based on Riemannian geometry for motor imagery decoding was proposed. (2) The proposed method along with the state-of-the-art method achieved the best results in the classification of the BCI-IV-2a dataset. (3) The key features acquired at each layer of the model were visualized to enhance the interpretability of the model.

The structure of this article is as follows. Section II provides an overview of the proposed method details. Section II-A introduces the experimental dataset, specific process, and experimental results, and compares the proposed method with the benchmark model. Section III discusses the meaning of the research results and summarizes the contributions and shortcomings of this study.

II. METHODOLOGY

The proposed MSCARNet model consists of three main blocks: sliding window block, multi-scale convolutional attention block and Riemannian geometry embedding block as shown in Figure 1.

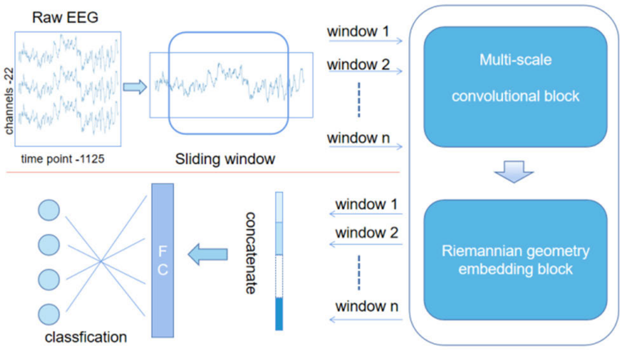

FIGURE 1. Components of the proposed MSCARNet model. Firstly, the EEG signal is segmented into n sliding windows and then the spatio-temporal features are extracted by inputting multi-scale convolutional block and Riemannian geometry embedding block successively, and then the features of these n windows are spliced and inputted into the Full Connection (FC) layer for classification.

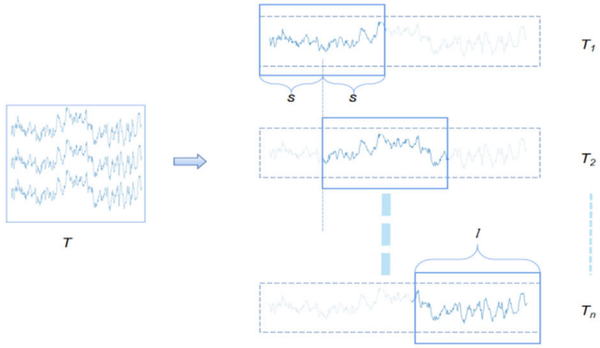

FIGURE 2. Sliding window.The raw EEG signal (T) is segmented into n windows (T1,T2 . . . Tn), each of length l, with a sliding step of s = l÷2.

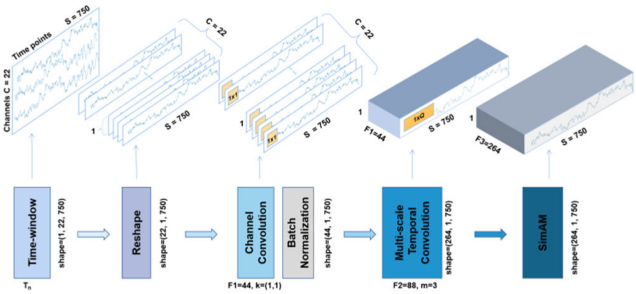

FIGURE 3. Multi-scale convolutional attention block. This block successively passes the signal from each window through channel convolution (filters = $F1$ = 44, kernel size = [1,1]) and temporal convolution, where the temporal convolution has m = 3 convolution kernels of different sizes Q = (15,35,55) and filters = $F2$ = 88, finally these features are spliced into SimAM which can acquire 3D attention.

The EEG signal is segmented into $n$parts using a sliding window to address the issue of limited data size. These $n$time windows are then fed into a multi-scale convolutional attention block, which extracts spatio-temporal features of the EEG signal and incorporates an attention mechanism. The obtained features are then mapped into Riemannian geometry to minimize the impact of noise and extreme values on the signal. This embedding reduces the number of parameters and computational costs through bilinear mapping operation. The decoding of MI-EEG is accomplished by concatenating the outputs of Riemannian geometry embedding blocks with different time windows and inputting them into the fully connected layer for classification. The MSCARNet model will be described in detail in the following sections.

A. SLIDING WINDOW

The Sliding window method can effectively expand the amount of EEG signal data, making it suitable for more complex networks [22], [23]. By setting the window length $l$and step size s, the start and end positions of the time window in the original signal Tcan be obtained, and its starting position index is $s \times (n\text{-}1)$, the ending position index is $s \times (n\text{-}1)+1$, where $n = 1,2,3\ldots$, is the index of the window, which divides the original EEG signal Tinto $T1$, $T2,\ldots$, Tn. Here, we set the step size s to half the window length, that is, $s = l \div 2$, to expand the data while reducing the overall computational workload. Figure 2 describes the principle of sliding windows, and the effects of different window numbers will be shown in the third chapter.

Sliding window was used for all extracted EEG signals and it was verified through extensive experiments that the model works best with a window number of 2. The number and size of sliding windows remain constant regardless of the number of training sessions.

B. MULTI-SCALE CONVOLUTIONAL ATTENTION BLOCK

The temporal windows ($T1$, $T2$, etc.) of section II A are input into a convolutional module with an attention mechanism in order to extract the spatial and temporal features of the samples. This module includes a channel convolutional layer, $m$temporal convolutional layers with different convolutional kernel sizes, and a 3D attention layer. The approximate framework is shown in Figure 3.

Firstly, through the Reshape layer, the time window $Tn$is dimensionally rearranged to change the original data dimension from $(1, C, S)$ to $(C, 1, S)$, where $C = 22$ is the number of channels and $S = 750$ is the number of sampling points for the time window to match the input size of channel convolution. After converting dimensions, using $F1 = 44$ convolution kernels with the size of 1×1 improves the data dimension and obtains the spatial characteristics of EEG signals through the Channel Conversion layer, it is worth noting that too many filters will result in more noise, while too few will result in insufficient feature extraction and 44 is a compromise number used in this study, which is experimentally proven to be more effective than 64 or 24. Combined with the training of the following modules, it can minimize the influence of outliers on decoding. Through our experiments, this is smaller than using the convolution kernel with the size of $(C, 1)$ directly on the original data, and the model effect is better. As far as we know, no researcher has done such work. After obtaining spatial features, use the Batch Normalization (BN) layer to

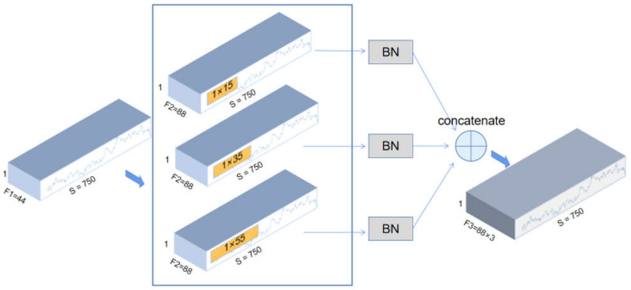

FIGURE 4. Multi-scale temporal convolution module. This module passes the spatial features obtained earlier through $m = 3$ convolutional layers of different sizes, each with 88 filters, extracts the temporal features and inputs them into the Batch Normalization (BN) layer, and finally stitches them together into a feature matrix of (88 × m, 1, 750).

further stabilize the data distribution and accelerate model training. Next, input the feature map into the Multi-scale temporal convolution layer. Multi-scale convolution has been proven to effectively extract sample information and improve model accuracy in different fields [23], [24]. This layer is detailed in Figure 4.

Using m convolutional kernels of different sizes to further extract the spatial and temporal features of EEG signals at different frequencies and input them into BN layers for concatenation. Although too many convolutional kernels or too large a convolutional kernel size may capture more features, it is also difficult to train the model due to the surge in parameters and computational complexity. The optimal $m = 3$ obtained in this model experiment by comparing the effect of single and multiple convolutional kernels on classification results which shown in section III G, and the convolutional kernel sizes are respectively 1 × 15, 1 × 35, 1 × 55. After obtaining the spatio-temporal features of EEG signals, we used the SimAM [25] module. This module is proposed for visual classification tasks, which is lighter and can calculate 3D weights compared to attention mechanism modules such as Convolutional Block Attention Module (CBAM) [26]. This is suitable for multi-channel EEG signals. Based on the spatial suppression mechanism proposed by Webb et al. [27], an algorithm is designed to extract key neurons as follows:

$$e_t(w_t, b_t, y, x_i) = (y_t, -\hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x}_i)^2 \quad (1)$$

$\hat{x}\ i = w_t\ x_{\ i} + b_t$ and $\hat{t} = w_t t + b_t$ are transformed from x_i and t, where t and x_i are the target neurons and other neurons in a single channel of the input feature $X \in R^{C\times H\times W}$ $(C, H, W$ are channel, height and width of feature). $t$is an index of the channel, and $M$is the number of neurons on that channel. w_t and b_t are weight and bias. y is the natural number bias of the whole formula including y_0 and y_t. The calculation of (1) effectively distinguishes the target neurons from the other neurons in the same channel. Furthermore, a regularizer is added into (1) and binary labels (1 and -1) are applied to y_t and y_o. Finally, the energy function changes to:

$$\mathrm{e_t}(w_t, b_t, \mathrm{y}, x_i) = \frac{1}{M-1} \times \sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2 \quad (2)$$

w_t and b_t can be easily get by (3) and (4):

$$w_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \quad (3)$$

$$b_t = -\frac{1}{2}(t+\mu_t)w_t \quad (4)$$

$$\mu_t = \frac{1}{M-1}\sum\nolimits_{i=1}^{M-1} x_i \quad (5)$$

$$\sigma_t^2 = \frac{1}{M-1}\sum_{i}^{M-1}(x_i - \mu_t)^2 \quad (6)$$

Eqn. (5) calculates the mean of all neurons except $t$in one channel, while (6) is variance. All neurons in a single channel are assumed to follow the same distribution and are reused for all neurons on that channel [28] to reduce the computational cost and avoid repeated calculation of μ and σ at each position. Therefore, the minimum energy can be calculated using the following:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (7)$$

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_{i,}$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2 \quad (8)$$

Eqn (7) determines the importance of each neuron by $1/e_{\mathrm{t}}^*$, and the lower the energy e_{t}^*, the more distinct it is from other neurons and more important for signal processing.

At this point, we obtained a multidimensional EEG feature map with attention weights.

C. RIEMANNIAN GEOMETRY EMBEDDING BLOCK

The mapping of data to a Riemann space can be effectively employed for a number of purposes, including classification, smoothing, extrapolation, and averaging [21], [29]. These operations can be locally approximated by Euclidean spaces via their tangent spaces. The SPDNet [20] can effectively embed Riemannian geometry into a deep learning model, we implement the BiMap layer and LogEig layer from SPDNet on our model. As shown in Figure 5.

First, reduce the dimension of the three-dimensional spatial and temporal feature map X_a (F3, 1, 750) obtained in Section II B to a two-dimensional matrix ($F3$, 750), and obtain the spatial Covariance matrix X_b of the feature through the following function:

$$x_b = (x_a \cdot x_a^T) \div (S-1) \quad (9)$$

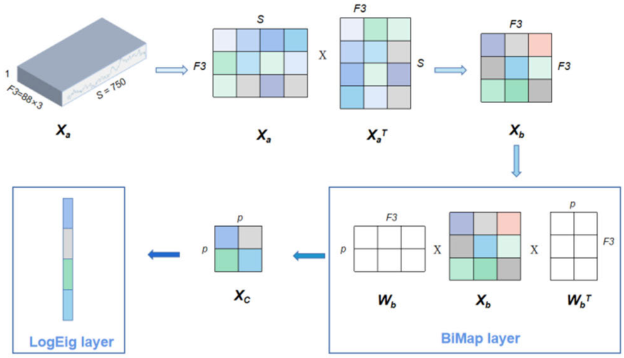

FIGURE 5. Riemannian geometry embedding block. This block first transforms the acquired spatio-temporal feature matrix into a symmetric positive definite (SPD) matrix X_b and inputs it into the BiMap layer, which not only reduces the feature dimensions but also extracts the key information of the Riemannian space, and finally flattens the features X_c by LogEig and activates them through logarithmic operations. $p = 32$ is height of W_b which is the weight matrix that need to be trained.

where X_a^T is the transposition of X_a, $S$is the time points of the sample, "X" in Figure 5 is the matrix matmul product.

X_b $(F3, F3)$ is then input to the BiMap layer, which utilizes a bilinear mapping operation to reduce the dimensionality of the features and reduce the computational complexity of the network training process. Implement through the following function: $X_c = W_b{\cdot}X_b{\cdot}W_b^T$, where W_b is the weight matrix that need to be trained of size $(p, F3)$, by comparing 8, 16, and 64 experiments, $p = 32$ is the optimal setting, too large a value leads to long computation time while too small a value leads to lack of performance, W_b^T is the transposition of W_b, X_c (32, 32) is EEG features embedded in Riemannian manifold. Finally, the LogEig layer performs a matrix logarithmic operation on the SPD matrix features and flattens the matrix to a vector of size (1024,) for classification. More bilinear mapping and LogEig layers related backpropagation details can be found in Huang and Gool [20].

D. CLASSIFICATION

In sections 2.2 and 2.3, spatial and temporal features with a size of (1024,) were extracted from each window. In this section, we concatenated the features of each window (in this paper, 2 windows and 1024 features each window, totally 2048 features) and feed them into the Full Connection layer to calculate the probability of the four categories for MI-EEG classification.

III. EXPERIMENT AND RESULT

In order to test the effectiveness of the proposed method, we used Intel @ Xeon (R) Silver 4210R CPU @ 2.40GHz x40 and Nvidia GeForce RTX 3090 GPU on the Ubuntu 20.04 system to train and verify the proposed model based on pytorch 1.8 framework.

A. DATASET DESCRIPTION AND PREPROCESSING

BCI Competition IV-2a (BCI-IV-2a) dataset [30] is used to evaluate the proposed model. It is also a popular dataset for many researchers due to its challenges. This dataset collected MI EEG data from nine subjects in two sessions, both of which contained the same experimental content but were captured on separate dates. Each session contained four MI tasks: left hand, right hand, feet, and tongue, and each task was performed 72 times for a total of 288 trials. A total of 576 trials were performed in two sessions per subject, for a total of 5184 trials for nine subjects. The EEG signals in the dataset consist of 22 channels, for which the publisher has applied band-pass filtering from 0.5-100 Hz as well as a trap filter at 50 Hz.

In this paper, we scrutinized the data to make sure there were no bad channels before the experiment, then raw EEG signals was feed into the model without more preprocessing to maximize the retention of real data. Data from 1.5-6s after the start of whole 576 trial were used to train and test the model, which included 1125 time points for one sample.

B. PERFORMANCE METRICS

Accuracy and Kappa score are considered important indicators in MI-EEG signal classification, which is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} TP_i \div l_i}{n} \tag{10}$$

where ACC is accuracy, $n$indicates the number of classes, TP_i is the abbreviation for true positive which means the number of correctly predicted samples in class i, and l_i is the number of samples in class i.

$$k\_score = \frac{1}{n}\sum_{a=1}^{n}\frac{P_a - P_e}{1 - P_e} \tag{11}$$

where k_score is Kappa score, P_e is the expected percentage chance of agreement, P_a is the actual percentage of agreement, and $n$is the number of classes.

Standard deviation measures the degree of dispersion of a set of data, this study shows the stability of the model by calculating the standard deviation of the average classification accuracy of all subjects. Wilcoxon test was performed to investigate the effect of the proposed method on decoding accuracy while significant differences ($p < 0.05$) were observed.

For the 4-class MI-EEG classification problem, the probabilistic chance level is given by $100/4 = 25\%$, which is only valid for an enormous number of trials. However, for finite MI experiments, the level of theoretical chance in terms of statistics is crucial for assessing decoding performance and can be obtained using binomial cumulative distributions (analytical method). The significant chance level of the analytical method was calculated by:

$$chance\_level = binom(1 - \alpha, n, c) \times 100 \div n \tag{12}$$

where $binom$is the function in SciPy, $a = 0.05$ is the confidence level, $n = 5184$ or 2592 is the total number of samples for subject-dependent or subject-independent experiment and $c = 4$ is the number of category. Eventually, the significant

TABLE 1. The architecture of the proposed MSCARNet, where *n* is the number of windows, *l* is the length of each window, *s* is the stride of sliding, *F* is the filters of convolution layer, *m* is the number of different temporal convolution which has various kernel sizes,and *p* is the height of the weight matrix.

OPERATION	PARAMETERS	NUMBER OF KERNELS	KERNEL SIZE
SLIDING WINDOW	n=2, l=750,s=375	-	-
CHANNEL CONVOLUTION	$F1$=44	44	(1,1)
MULTI-SCALE TEMPORAL CONVOLUTION	m=3 $F2$=88	88 88 88	(1,15) (1,35) (1,55)
RIEMANNIAN GEOMETRY EMBEDDING	p=32	-	-
CLASSIFICATION	Class=4	-	-

TABLE 2. Configuration of subject-dependent and subject-independent classification. Detailed description of the training set and test set segmentation for different classification experiments. LOSO means leave one subject out.

	SUBJECT-DEPENDENT CLASSIFICATION	SUBJECT-INDEPENDENT CLASSIFICATION
SESSIONS USED	SESSION 1 AND SESSION 2	SESSION 1 ONLY
NUMBER OF TRIALS PER SUBJECT USED	576	288
VALIDATION TYPE	5-FOLD CROSS-VALIDATION	LOSO
NUMBER OF TRAINING TRIALS	ABOUT 460	2304
NUMBER OF TESTING TRIALS	ABOUT 116	288

chance level of subject-dependent experiment for decoding accuracy was 25.98% while 26.39% of subject-independent experiment.

C. EXPERIMENT SETUP

Table 1 shows the proposed MSCARNet used in the experiment. In the training process, the early stop method is adopted with 100 epochs.

Two different cross-validation forms are utilized for the performance evaluation: subject-dependent classification and subject-independent classification. As illustrate in Table 2, subject-dependent classification, for each subject, we mixed and shuffled the data from two sessions and implemented a 5-fold cross-validation experiment using Scikit-Learn package, these samples are split into 5 folds, then pick one as the test set and the rest as the training set each time, with about 460 trials for training and 116 for testing and iterate 5 times. In subject-independent classification, Leave One Subject Out (LOSO) experiment was utilized, for each subject, we used this subject's first session for testing and other 8 subjects' first session for training, with 2304 trails for training and 288 for testing.

D. PERFORMANCE EVALUATION

Experiments are conducted for each subject to test the classification performance of the proposed MSCARNet, Figure 6 and Figure 7 illustrate separately the accuracy and Kappa score in subject-dependent classification and subject-independent classification in test set.

In subject-dependent classification, most subjects get an acceptable result except subjects 2, 5 and 6, which is a common problem for most research, the proposed MSCARNet reached 82.66% and 0.7483 on average accuracy and Kappa scores of nine subjects, and the standard deviation of accuracy and Kappa scores is 10.02 and 10.36. In subject-independent classification, most models cannot achieve a good performance, the proposed model obtains 61.24% and 0.5552 on average accuracy and Kappa scores of nine subjects, and the standard deviation of accuracy and Kappa scores is 10.0 and 9.47.

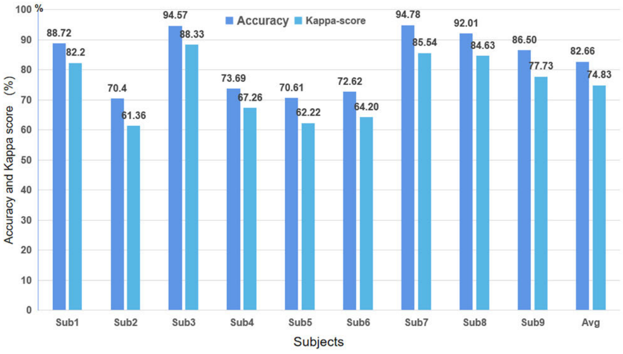

FIGURE 6. Each subject's average accuracy and kappa score in subject-dependent classification of test set. Using 5-fold cross-validation experiments. Avg means grand-average accuracy of all subjects.

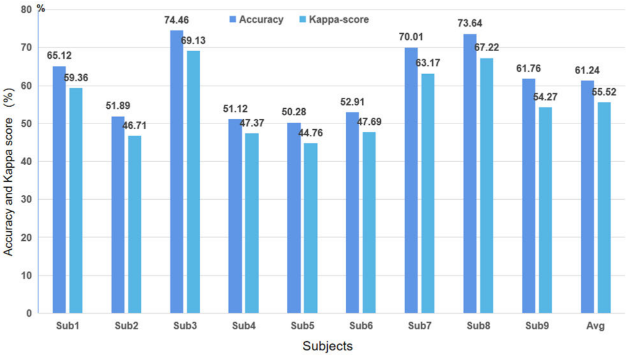

FIGURE 7. Each subject's accuracy and kappa score in subject-independent classification of test set. Using LOSO experiments. Avg means average accuracy of all subjects.

To demonstrate that the proposed model is not biased toward categorization, Figure 8 shows the confusion matrix

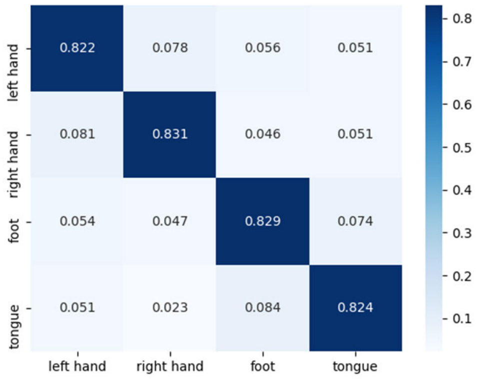

FIGURE 8. Confusion matrix for the results of the test set for all subjects of 5-fold cross-validation experiment.

TABLE 3. Average accuracy, Kappa score and standard deviation (Std.) comparison of subject-dependent classification for the proposed model with other models of all subjects. P-value is calculated to compared with MSCARNet (p < 0.05). Best results are bolded.

Models	Accuracy (Std.)	Kappa scores (Std.)	P-value
SPDNet	55.28% (11.84)	0.4912 (12.54)	0.002
EEGNet	75.29% (10.44)	0.7013 (10.88)	0.007
FBCNet	76.81% (11.77)	0.6978 (12.64)	0.012
MI-EEGNet	76.21% (11.31)	0.6945 (11.99)	0.011
EEG-TCNet	78.11% (10.61)	0.7060 (11.14)	0.026
EEGNeX	76.90% (11.43)	0.7024 (12.14)	0.022
KMDA	78.05% (10.49)	0.7055 (10.94)	0.037
CRGNet	78.62% (10.90)	0.7094 (11.08)	0.081
MSCARNet (proposed)	**82.66% (10.02)**	**0.7483 (10.36)**	-

for the results of the test set for all subjects of 5-fold cross-validation (subject-dependent) experiment.

E. COMPARING TO OTHER RESEARCH

The results of classification are compared with some representative models in MI-EEG classification without data preprocessing and the methods is reproduced manually, including SPDNet [20], EEGNet [14], FBCNet [31], MI-EEGNet [32], EEG-TCNet [33], EEGNeX [34], KMDA [35], CRGNet [21]. Table 3 shows the comparison of subject-dependent classification and based on the Wilcoxon test, the increase in average accuracy is statistically significant (p< 0.05) compared to all methods except CRGNet.

The experimental results of subject-dependent show that MSCARNet reaches the best average accuracy and Kappa score, the accuracy is approximately 4% higher and the Kappa score is about 0.04 higher than the state-of-the-art method, otherwise the lowest standard deviation among subjects indicating the proposed model is more robust.

TABLE 4. Average accuracy, kappa score and standard deviation (Std.) comparison of subject-independent classification for the proposed model with other models of all subjects. P-value is calculated to compared with MSCARNet (p < 0.05). The best results are bolded.

Models	Accuracy (Std.)	Kappa score (Std.)	P-value
SPDNet	38.94% (9.21)	0.3184 (9.51)	0.001
EEGNet	58.62% (10.08)	0.5115 (9.88)	0.034
EEG-TCNet	52.83% (11.24)	0.4726 (10.78)	0.021
EGNeX	57.67% (10.76)	0.5047 (10.64)	0.029
KMDA	56.77% (10.92)	0.4992 (11.07)	0.017
CRGNet	58.74% (10.44)	0.5133 (10.33)	0.045
MSCARNet (proposed)	**61.24% (10.00)**	**0.5552 (9.47)**	-

Table 4 presents the comparison of subject-independent classification.

As shown in Table 4, most models do not get a good result in subject-independent classification, the proposed MSCARNet reaches the best metrics except for standard deviation of accuracy, and the accuracy and Kappa score is approximately 0.03 higher than state-of-the-art method.

Overall, the proposed model demonstrated an improvement in MI-EEG decoding compared to other model, though still stuck in inefficient of subject-independent classification.

F. ABLATION ANALYSIS

In this subsection, we evaluate the effectiveness of each block in the MSCARNet model. The blocks were removed prior to the training and validation operations. Table 5 shows the effect of removing a block in the MSCARNet model on the average accuracy and Kappa score of subject-dependent classification using the BCI-IV-2a dataset.

The results showed sliding window block increased the grand-average average accuracy by 8.82%, Multi-scale convolution by 17.39%, SimAM by 2.93% and Riemannian embedding by 12.74%. When removing the multiscale convolution and the Riemannian embedding layer leaving only the channel convolution and single scale temporal convolution the grand-average accuracy drops by 34.75%. Each block has a contribution to classification, especially the multi-scale convolution block and Riemannian embedding block, those two blocks play a crucial role in the MSCARNet.

G. COMPARISON OF DIFFERENT PARAMETERS

1) NUMBER OF TIME-WINDOWS

As we set the sliding stride s half of the window length l, it's simple to obtain the s and l by setting the number of windows. In this part we compared 5 different numbers of

TABLE 5. Contribution of each block in the MSCARNet to the performance of subject-dependent classification using the BCI-IV-2a dataset. "-" represents the MSCARNet without any blocks removed.

Removed block	Accuracy (%)	Kappa score
-	82.66	0.7483
Sliding window	73.84	0.6672
Multi-scale convolution	65.27	0.5716
SimAM	79.73	0.7109
Riemannian embedding	69.92	0.6233
Sliding window + SimAM	71.07	0.6404
Multi-scale convolution + Riemannian embedding	47.91	0.4398

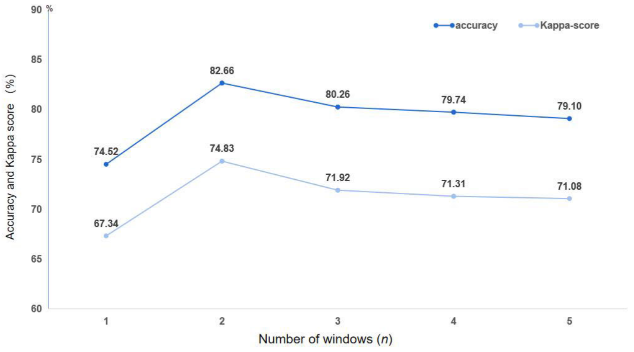

FIGURE 9. Grand-average accuracy and kappa score of different number of windows on subject-dependent classification.

windows, which are 1, 2, 3, 4, 5, to avoid the parameters of the model quantity skyrocketing, we adjusted the parameters of the Riemannian embedding block accordingly which may influence the results but controlled the parameters similarly. Figure 9 shows the performance of the MSCARNet with different number of windows.

The results show 2 time-windows reach the best performance with 750 time points per window, the accuracy decreased by increasing of number from 2, thus we don't attempt more windows. The number of windows employed corresponded to the length of the windows, which varied in size. This enabled the extraction of signal features over different time periods. The optimal results were obtained with two windows, which overlapped the data from the intermediate time periods. This suggests that the motor imagery features was implied approximately one second after the onset of the cue.

2) SCALE OF TEMPORAL CONVOLUTION

For the convolution layer, the number and size of the kernel determine the effectiveness of the network. In this section, we set 5 single kernel sizes which are (1,15), (1,35), (1,55), (1,75), (1,100), and their combination to evaluate the advantage size over the MSCARNet. Shown in Figure 10.

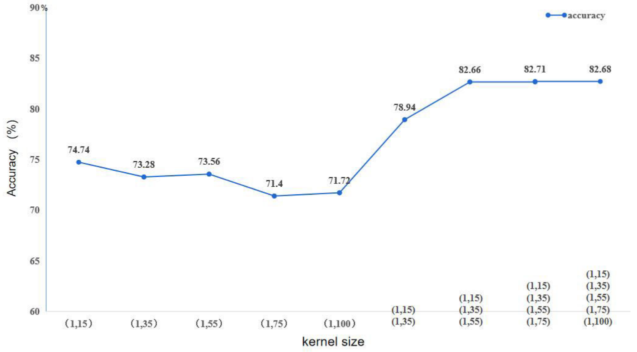

FIGURE 10. Grand-average accuracy of different kernel sizes in subject-dependent classification on the MSCARNet. A comparison of single and multi scale convolution kernels has been performed.

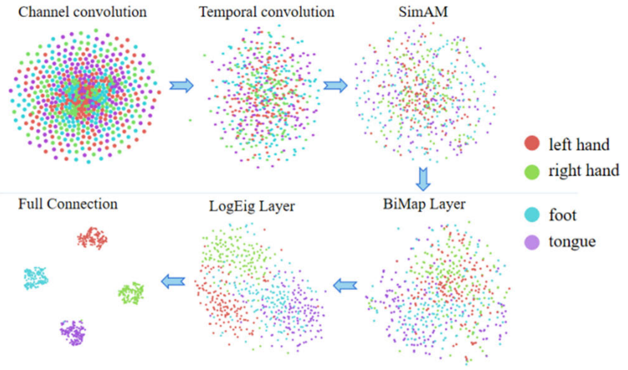

FIGURE 11. Visualization of features output by each layer based on T-SNE method to validate the influence of layers. The order of presentation is consistent with the order in which the features are output in the model.

We firstly test the accuracy of the single kernel with sizes (1,15), (1,35), (1,55), (1,75) and (1,100) in subject-dependent classification on the proposed model, the results indicate that a bigger size of the kernel may not own better performance, size of (1,15) win the top score. A substantial body of research has demonstrated that characteristics of low-frequency signals are more pertinent to motion imagery. This finding aligns with this study, which demonstrates that a small convolutional kernel is capable of extracting features of lower-frequency signals that are more conducive to classification. Then we adopt multi-scale kernel in subject-dependent classification which are

$$
\begin{aligned}
&(1, 15) + (1, 35),\\
&(1, 15) + (1, 35) + (1, 55),\\
&(1, 15) + (1, 35) + (1, 55) + (1, 75),\\
&(1, 15) + (1, 35) + (1, 55) + (1, 75) + (1, 100).
\end{aligned}
$$

It seems that accuracy increased by number of kernel scale indeed, but accuracy among 3, 4 and 5 kernel sizes barely noticeable difference. However bigger size means more parameters and calculations, so the kernel size of (1,15) + (1,35) + (1,55) were implemented in our proposed model.

H. FEATURE VISUALIZATION

The T-SNE method was used to show the output features of each layer, Figure 11 visualized the output features of each layer according to the input order of the features to show the effect of each layer.

Figure 11 shows that the SimAM layer can expand the variability of acquired spatio-temporal features, while the BiMap and LogEig layers have initially categorized features, confirming the effectiveness of these layers.

IV. DISCUSSION AND CONCLUSION

The proposed method improves the classification accuracy of the BCI-IV-2a dataset by approximately 4%. This paper extends the work of Gao et al. [21] by proposing a novel attention and Riemannian geometry-based convolutional network (MSCARNet) for MI-EEG classification. MSCARNet is composed of three main blocks: the sliding window block, which splits raw EEG signals into several samples of equal length; the multi-scale temporal convolutional block, which extracts high-level spatial and temporal features from time-windows; and the Riemannian geometry embedding block, which maps the features between Euclidean and Riemannian manifolds. These blocks have been shown to significantly improve the performance of MSCARNet in MI-EEG classification through ablation analysis.

The analysis indicates that a larger number of time windows does not necessarily result in better classification. Further investigation is needed to determine the optimal size and number of time windows. Multi-scale convolutional modules are more effective than single-scale ones, but an excessive number of convolutional sizes does not improve classification quality. The modules' effectiveness and the model's overall efficiency are fully verified through feature visualization and statistical analysis.

The proposed model, MSCARNet, achieves an accuracy of 82.66% and 57.01% in subject-dependent and subject-independent classification, respectively, using the BCI-IV-2a dataset. This demonstrates the model's potential to decode MI-EEG signals with minimal preprocessing, particularly when working with datasets of limited size and complexity. However, the study still has some issues that require further investigation. Firstly, the embedding of Riemannian geometry leads to a significant improvement in model classification, but also results in a non-negligible increase in training cost, including computation and parameter count. Secondly, the LOSO experiments yielded unsatisfactory results, indicating that the model fails to fully capture individual differences. To address the aforementioned issues, future work can focus on the following directions: utilizing a pooling layer to decrease the feature dimensions of the input Riemannian geometry, optimizing the computational process of the BiMap layer, and implementing methods such as jump linking to capture more information and solve the problem of excessive individual differences.

REFERENCES

[1] M. L. Martini, E. K. Oermann, N. L. Opie, F. Panov, T. Oxley, and K. Yaeger, "Sensor modalities for brain-computer interface technology: A comprehensive literature review," *Neurosurgery*, vol. 86, no. 2, pp. E108–E117, Feb. 2020, doi: 10.1093/neuros/nyz286.

[2] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, p. 1423, Mar. 2019, doi: 10.3390/s19061423.

[3] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, Mar. 2012, doi: 10.3389/fnins.2012.00039.

[4] S. Liu, Q. Zhang, and L. Huang, "Graphic image classification method based on an attention mechanism and fusion of multilevel and multiscale deep features," *Comput. Commun.*, vol. 209, pp. 230–238, Sep. 2023, doi: 10.1016/j.comcom.2023.07.001.

[5] M. J. Islam, R. Datta, and A. Iqbal, "Actual rating calculation of the zoom cloud meetings app using user reviews on Google play store with sentiment annotation of BERT and hybridization of RNN and LSTM," *Expert Syst. Appl.*, vol. 223, Aug. 2023, Art. no. 119919, doi: 10.1016/j.eswa.2023.119919.

[6] Y. Zou, W. Min, H. Zhao, and Q. Han, "A novel framework for crowd counting using video and audio," *Comput. Electr. Eng.*, vol. 109, Aug. 2023, Art. no. 108754, doi: 10.1016/j.compeleceng.2023.108754.

[7] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal, "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14681–14722, Jul. 2023, doi: 10.1007/s00521-021-06352-5.

[8] S. Kumar, R. Sharma, and A. Sharma, "OPTICAL+: A frequency-based deep learning scheme for recognizing brain wave signals," *PeerJ Comput. Sci.*, vol. 7, p. e375, Feb. 2021, doi: 10.7717/peerj-cs.375.

[9] T.-J. Luo, C.-L. Zhou, and F. Chao, "Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network," *BMC Bioinf.*, vol. 19, no. 1, p. 344, Sep. 2018, doi: 10.1186/s12859-018-2365-1.

[10] J. Xu, H. Zheng, J. Wang, D. Li, and X. Fang, "Recognition of EEG signal motor imagery intention based on deep multi-view feature learning," *Sensors*, vol. 20, no. 12, p. 3496, Jun. 2020, doi: 10.3390/s20123496.

[11] A. Hassanpour, M. Moradikia, H. Adeli, S. R. Khayami, and P. Shamsinejadbabaki, "A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals," *Expert Syst.*, vol. 36, no. 6, Dec. 2019, Art. no. e12494, doi: 10.1111/exsy.12494.

[12] D. Zhang, K. Chen, D. Jian, and L. Yao, "Motor imagery classification via temporal attention cues of graph embedded EEG signals," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 9, pp. 2570–2579, Sep. 2020, doi: 10.1109/JBHI.2020.2967128.

[13] G. A. Altuwaijri, G. Muhammad, H. Altaheri, and M. Alsulaiman, "A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification," *Diagnostics*, vol. 12, no. 4, p. 995, Apr. 2022, doi: 10.3390/diagnostics12040995.

[14] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013, doi: 10.1088/1741-2552/aace8c.

[15] D. Li, J. Xu, J. Wang, X. Fang, and Y. Ji, "A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of EEG signals decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2615–2626, Dec. 2020, doi: 10.1109/TNSRE.2020.3037326.

[16] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Gener. Comput. Syst.*, vol. 101, pp. 542–554, Dec. 2019, doi: 10.1016/j.future.2019.06.027.

[17] Y.-W. Kim, S. Kim, M. Shim, M. J. Jin, H. Jeon, S.-H. Lee, and C.-H. Im, ''Riemannian classifier enhances the accuracy of machine-learning-based diagnosis of PTSD using resting EEG,'' *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 102, Aug. 2020, Art. no. 109960, doi: 10.1016/j.pnpbp.2020.109960.

[18] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, ''EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density,'' *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101899, doi: 10.1016/j.bspc.2020.101899.

[19] Y. Chu, X. Zhao, Y. Zou, W. Xu, G. Song, J. Han, and Y. Zhao, ''Decoding multiclass motor imagery EEG from the same upper limb by combining Riemannian geometry features and partial least squares regression,'' *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 046029, doi: 10.1088/1741-2552/aba7cd.

[20] Z. Huang and L. Van Gool, ''A Riemannian network for SPD matrix learning,'' in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, doi: 10.1609/aaai.v31i1.10866.

[21] C. Gao, W. Liu, and X. Yang, ''Convolutional neural network and Riemannian geometry hybrid approach for motor imagery classification,'' *Neurocomputing*, vol. 507, pp. 180–190, Oct. 2022, doi: 10.1016/j.neucom.2022.08.024.

[22] N. Singh Malan and S. Sharma, ''Time window and frequency band optimization using regularized neighbourhood component analysis for multi-view motor imagery EEG classification,'' *Biomed. Signal Process. Control*, vol. 67, May 2021, Art. no. 102550, doi: 10.1016/j.bspc.2021.102550.

[23] H. Altaheri, G. Muhammad, and M. Alsulaiman, ''Physics-informed attention temporal convolutional network for EEG-based motor imagery classification,'' *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2249–2258, Feb. 2023, doi: 10.1109/TII.2022.3197419.

[24] S. Jiang, D. Li, and Y. Zhang, ''A deep neural network based on multi-model and multi-scale for arrhythmia classification,'' *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 105060, doi: 10.1016/j.bspc.2023.105060.

[25] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, ''SimAM: A simple, parameter-free attention module for convolutional neural networks,'' in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 11863–11874. Accessed: Dec. 8, 2023. [Online]. Available: https://proceedings.mlr.press/v139/yang21o.html

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, ''CBAM: Convolutional block attention module,'' in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[27] B. S. Webb, N. T. Dhruv, S. G. Solomon, C. Tailby, and P. Lennie, ''Early and late mechanisms of surround suppression in striate cortex of macaque,'' *J. Neurosci.*, vol. 25, no. 50, pp. 11666–11675, Dec. 2005, doi: 10.1523/jneurosci.3414-05.2005.

[28] B. Hariharan, J. Malik, and D. Ramanan, ''Discriminative decorrelation for clustering and classification,'' in *Computer Vision—ECCV 2012* (Lecture Notes in Computer Science), A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 459–472, doi: 10.1007/978-3-642-33765-9_33.

[29] F. Yger, M. Berar, and F. Lotte, ''Riemannian approaches in brain–computer interfaces: A review,'' *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1753–1762, Oct. 2017, doi: 10.1109/TNSRE.2016.2627016.

[30] R. Leeb and C. Brunner. (2008). *BCI Competition 2008*. Accessed: Jul. 17, 2023. [Online]. Available: https://www.semanticscholar.org/paper/BCI-Competition-2008-%7B-Graz-data-set-B-Leeb-Brunner/9031aac7e9b6adae909ce22fa35fa74ec52a52ec

[31] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, ''A multi-view CNN with novel variance layer for motor imagery brain computer interface,'' in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2950–2953, doi: 10.1109/EMBC44109.2020.9175874.

[32] M. Riyad, M. Khalil, and A. Adib, ''MI-EEGNET: A novel convolutional neural network for motor imagery classification,'' *J. Neurosci. Methods*, vol. 353, Apr. 2021, Art. no. 109037, doi: 10.1016/j.jneumeth.2020.109037.

[33] T. Mar Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, ''EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain-machine interfaces,'' May 2020, *arXiv:2006.00622*.

[34] X. Chen, X. Teng, H. Chen, Y. Pan, and P. Geyer, ''Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX,'' *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105475, doi: 10.1016/j.bspc.2023.105475.

[35] Q. Jiang, Y. Zhang, and K. Zheng, ''Motor imagery classification via kernel-based domain adaptation on an SPD manifold,'' *Brain Sci.*, vol. 12, no. 5, p. 659, May 2022, doi: 10.3390/brainsci12050659.

BEN ZHOU was born in 1998. He received the bachelor's degree in software engineering from Huangshan University. He is currently pursuing the master's degree with Shandong University of Traditional Chinese Medicine. His research interests include signal processing and artificial intelligence.

LEI WANG received the Ph.D. degree in computer applied technology from Harbin Institute of Technology, in 2013. He is currently a Professor with Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, and the Master's Tutor with the University of Science and Technology of China. His research interests include medical decision support systems for healthcare, optimization algorithms, collaborative decision-making technology based on big data analysis, and software development technology for biomedical information management systems.

WENCHANG XU was born in 1993. She received the master's degree in mechanical electronic engineering from the University of Chinese Academy of Sciences, in 2018. Her main research interests include optimization algorithms, collaborative decision-making technology based on big data analysis, and the application of deep learning on medical.

CHENYU JIANG received the joint Ph.D. degree in medicine from Tsinghua University and Peking Union Medical College. He is currently a Professor with Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. His research interests include spectral analysis methods and medical testing instruments, organizational spectral detection technology, and equipment and medical spectral diagnosis.