A Temporal Dependency Learning CNN With Attention Mechanism for MI-EEG Decoding

Xinzhi Ma, Weihai Chen, *Member, IEEE*, Zhongcai Pei, Jingmeng Liu, Bin Huang, *Member, IEEE*, and Jianer Chen

***Abstract*— Deep learning methods have been widely explored in motor imagery (MI)-based brain computer interface (BCI) systems to decode electroencephalography (EEG) signals. However, most studies fail to fully explore temporal dependencies among MI-related patterns generated in different stages during MI tasks, resulting in limited MI-EEG decoding performance. Apart from feature extraction, learning temporal dependencies is equally important to develop a subject-specific MI-based BCI because every subject has their own way of performing MI tasks. In this paper, a novel temporal dependency learning convolutional neural network (CNN) with attention mechanism is proposed to address MI-EEG decoding. The network first learns spatial and spectral information from multi-view EEG data via the spatial convolution block. Then, a series of non-overlapped time windows is employed to segment the output data, and the discriminative feature is further extracted from each time window to capture MI-related patterns generated in different stages. Furthermore, to explore temporal dependencies among discriminative features in different time windows, we design a temporal attention module that assigns different weights to features in various time windows and fuses them into more discriminative features. The experimental results on the BCI Competition IV-2a (BCIC-IV-2a) and OpenBMI datasets show that our proposed network outperforms the state-of-the-art algorithms and achieves the average accuracy of 79.48%, improved by 2.30% on the BCIC-IV-2a dataset. We demonstrate that learning temporal dependencies effectively improves MI-EEG decoding performance. The code is available at https://github.com/Ma-Xinzhi/LightConvNet**

Manuscript received 24 November 2022; revised 14 June 2023; accepted 19 July 2023. Date of publication 27 July 2023; date of current version 9 August 2023. This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant 2021C03050; in part by the Scientific Research Project of Agriculture and Social Development of Hangzhou under Grant 20212013B11; and in part by the National Nature Science Foundation under Grant U1909215, Grant 51975002, and Grant 51975029. *(Corresponding author: Weihai Chen.)*

Xinzhi Ma, Zhongcai Pei, and Jingmeng Liu are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China, and also with the Hangzhou Innovation Institute, Beihang University, Hangzhou 310052, China (e-mail: mxz1995@outlook.com; peizc@buaa.edu.cn; ljmbuaa110@163.com).

Weihai Chen is with the School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China (e-mail: whchen@buaa.edu.cn).

Bin Huang is with the Hangzhou Innovation Institute, Beihang University, Hangzhou 310052, China (e-mail: marshuangbin@buaa.edu.cn).

Jianer Chen is with the Department of Geriatric Rehabilitation, Third Affiliated Hospital, Zhejiang Chinese Medical University, Hangzhou 310009, China (e-mail: cje28@foxmail.com).

Digital Object Identifier 10.1109/TNSRE.2023.3299355

***Index Terms*— Brain–computer interface (BCI), motor imagery (MI), convolutional neural networks (CNNs), temporal dependency learning, attention mechanism.**

I. Introduction

BRAIN computer interface (BCI) systems provide new ways for users to communicate with computers by translating brain signals into useful commands in real time [1]. In recent years, BCI technologies have gradually played an efficient role in providing assistance and preventive care to people paralyzed by chronic neuromuscular disorders [2], [3], [4]. In BCI systems, electroencephalography (EEG) signals are widely used to record brain activity because of their minimal risk and the relative convenience of conducting studies [1].

Many types of neurophysiological patterns have been applied to EEG-based BCI systems, such as steady-state visual evoked potential (SSVEP) [5], P300 event-related potential (ERP) [6] and motor imagery (MI) [7]. Among these EEG measurements, MI has been gaining more attention because it allows users to generate corresponding signals actively without any external stimuli. MI can be seen as mental rehearsal of a motor act and results in event-related desynchronization or synchronization (ERD/ERS) by activating substantial related neurons within the sensorimotor area of cerebral cortex [8]. MI-based BCI systems are demonstrated as assistive tools for paralyzed patients, thereby facilitating motor rehabilitation [9] and external device control [7]. Therefore, decoding MI-EEG signals with high accuracy is critical to the application of MI-based BCI system. However, non-stationary MI-EEG signals with noise, individual differences of subjects, and the scarcity of training data make MI-EEG decoding much more intractable than it appears [10].

In earlier studies, researchers attempt to extract neurophysiological features from MI-EEG data and employ classical machine learning methods to classify them. Common spatial pattern (CSP) [11], which maximizes the variance difference between two classes of MI-EEG signals, is one of most commonly used features in MI-based BCI systems. A series of variant methods based on CSP has been proposed to improve the discrimination of extracted features. Filter bank common spatial pattern (FBCSP) [12] extracted CSP features from multiple frequency bands instead of a specific band. Sparse filter band common spatial pattern (SFBCSP) [13] was further proposed to optimize the spatial patterns by exploiting sparse

regression for automatic band selection. In recent years, Das and Pachori [14] proposed to use combination of multivariate iterative filtering (MIF) and CSP to automatically select optimal frequency bands. Jin et al. [15] designed a new feature selection method based on an improved objective function to select optimal features in the feature space used within CSP.

Besides CSP-based methods, some methods based on other feature extraction algorithms also performed well on MI-EEG decoding tasks. Bhaleraoa and Pachori [16] proposed a swarm decomposition (SWD) based classification framework to decompose and find significant oscillatory components related to the MI-EEG signal. In the Riemannian geometry based classification framework, the covariance structure of EEG data was exploited as the feature of interest [17], [18]. In terms of classifiers, linear discriminant analysis (LDA) and support vector machine (SVM) are commonly used to classify extracted features. Although these attempts have achieved good performance for MI-EEG decoding, they have a high dependency on handcrafted features. In general, above methods employ two-stage pipelines, namely, a handcrafted feature extractor and a classifier. Due to the definite subjectivity of handcrafted feature extractor, the global optimization of algorithm can not be carried out to further improve MI-EEG decoding performance.

Recently, many studies have introduced various end-to-end deep learning networks into MI-EEG decoding without the need for handcrafted features. Furthermore, many advanced deep learning architectures are designed to compensate for the defect of conventional BCI systems and improve MI-EEG decoding performance [19]. Among these deep learning architectures, convolutional neural network (CNN) has been widely applied in MI-EEG decoding because of its ability to learn potential information from the given dataset effectively [20], [21], [22]. Although these methods have outperformed classical machine learning methods, temporal or spatial features are not fully explored, and their improvements are limited.

More recently, some well-designed deep learning architectures have been proposed to fully utilize multi-domain information of MI-EEG signals and enhance decoding performance. Zhao et al. [23] presented a 3D representation for MI-EEG data, which preserved the spatial information of sampling electrodes, and designed a multi-branch 3D convolutional neural network for the new data representation. In 2021, a temporal-spectral-based squeeze-and-excitation feature fusion network (TS-SEFFNet) was proposed to fuse the temporal-spectral features and improve the MI-EEG decoding accuracy [24]. However, these networks use a fixed time window of MI-EEG signals to extract features without considering the importance of MI-related patterns in different time periods, resulting in limited decoding performance. Recent FBCNet [25] proposed a variance layer to effectively extract features from different time windows of the time series and achieved excellent performance on public datasets. Nevertheless, FBCNet investigates features in different time windows independently without exploring temporal dependencies among them, which is not enough for discriminative feature extraction.

During one MI task, every subject reacts to the corresponding movement cue and completes the task at their own pace. We find that temporal dependencies among MI-related patterns in different stages during MI tasks are essential to improve the MI-EEG decoding performance. The similar idea was applied to SSVEP detection, which introduced the temporally local weighting into the objective function [26]. Therefore, we aim to build bridges between features extracted from different time periods. In recent years, attention mechanism has been widely used to build generative models for natural language processing [27] and image recognition [28]. It effectively learns dependencies among context elements by assigning them attention weights which define a weighted sum over context representations. Considering the scarcity of EEG training data, inspired by the attention-based lightweight convolution [29], a novel temporal dependency learning CNN with attention mechanism is proposed for MI-EEG decoding in this paper. The proposed network first implements the spatial convolution to learn spatial and spectral information from multi-view EEG data, which is preprocessed with a filter bank. Then, we employ a series of non-overlapped time windows to segment the output time series. The discriminative feature from each time window is further extracted using a temporal variance layer to capture MI-related patterns in different stages during MI tasks. Moreover, we design a novel temporal attention module to further learn temporal dependencies among discriminative features from different time windows. The temporal attention module assigns different weights to features in various time windows according to their contribution to the final decoding performance, and fuses them into more discriminative features. Finally, the fused features are used for classification. We evaluate our proposed network on two public datasets and demonstrate its better performance for MI-EEG decoding compared with other state-of-the-art algorithms.

The main contributions of this paper are summarized as follows:

- We propose an end-to-end deep learning architecture that effectively extracts discriminative features from different time periods of MI-EEG data and learns temporal dependencies among them. Particularly, a novel temporal attention module is designed to capture temporal dependencies between extracted features in different time periods and generate more discriminative features for MI-EEG decoding.
- To the best of our knowledge, this is the first deep learning study that focuses on temporal dependencies among discriminative features in different time periods during MI tasks. The proposed method indicates the potential of exploring temporal dependencies to improve MI-EEG decoding performance.
- We demonstrate that our proposed method performs better than the state-of-the-art algorithms on two benchmark public datasets. Furthermore, investigation via visualization of the learned features is carried out to

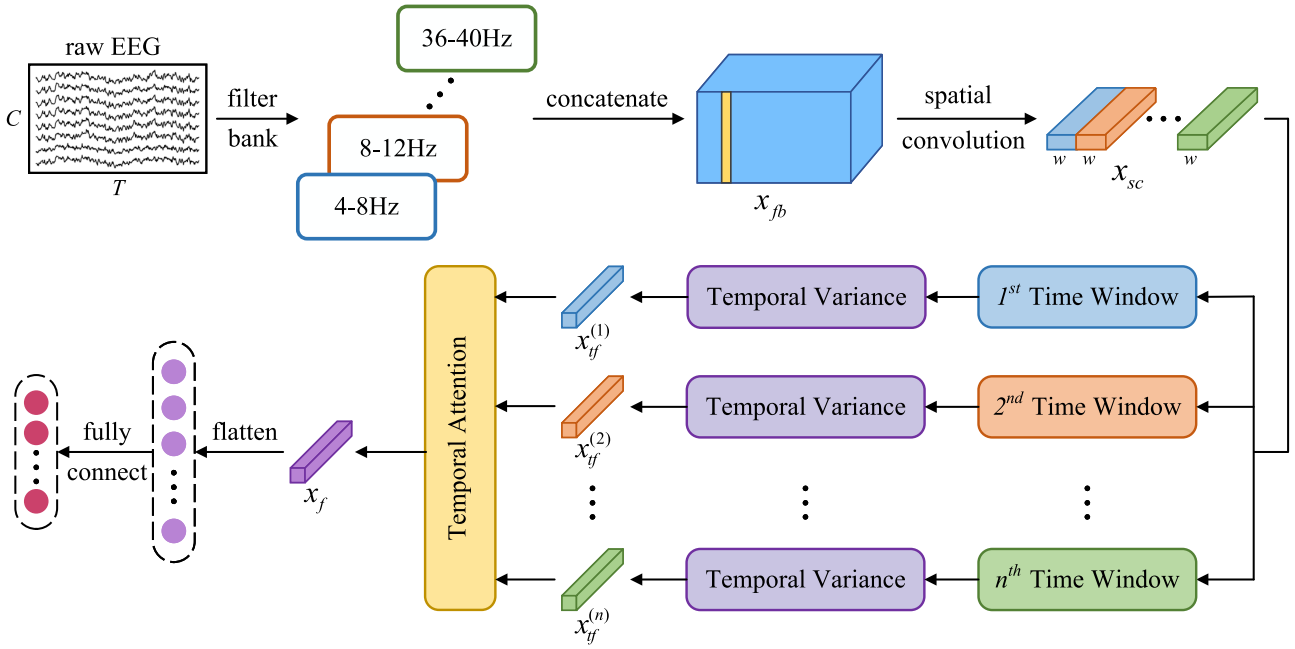


Fig. 1. Overall architecture of the proposed network for MI-EEG decoding, where C represents the number of EEG channels, T represents the number of sample points, and w represents the size of time window. The feature vectors that correspond to the same time window are represented with the same color as the time window.

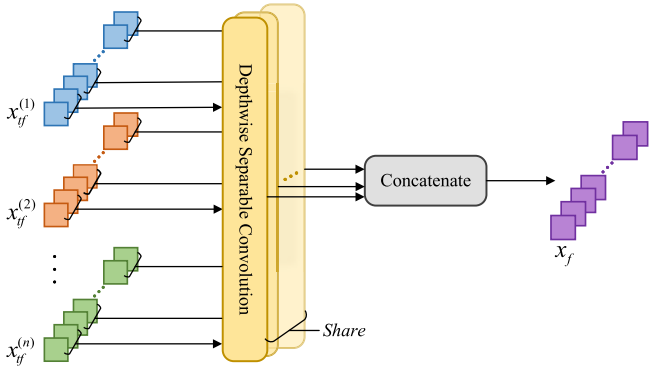


Fig. 2. Architecture of the proposed temporal attention module, where all depthwise separable convolutions share the same weight.

TABLE I
DETAILED ARCHITECTURE OF THE PROPOSED NETWORK

Layer	# filters	size	# params	Output
Input				(F, C, T)
Spatial Conv2d	m	(C, 1)	m * F * C	(m, 1, T)
BatchNorm			2 * m	(m, 1, T)
ELU			-	(m ,1, T)
Reshape			-	(m, T // w, w)
Temporal Variance			-	(m, T // w)
Reshape			-	(m // h, h, T // w)
Temporal Attention	h	T // w	h * (T // w)	(m // h, h, 1)
Flatten			-	(m // h * h)
Fully Connected Layer			N * (m // h * h)	N

F: number of filter bank, C: number of channels, T: number of sample points, m: number of spatial filters, w: time window size, h:number of attention heads, N: number of classes

interpret the superiority of the proposed method over others.

The remainder of this paper is organized as follows. Section II describes the input data representation and the structure of the proposed network. Section III elaborated the experiments. In Section IV and Section V, experimental results of the proposed method are presented and discussed, respectively. Finally, the paper is concluded in Section VI.

II. METHODOLOGY

In this section, we first describe the multi-view EEG data representation based on a bank of bandpass filters. Then, we present the proposed network for MI-EEG decoding in detail, including spatial and spectral information learning, temporal segmentation and feature extraction, temporal attention module, and classification. Finally, we introduce the training procedure of our proposed network. The overall architecture of our proposed network is illustrated in Fig. 1.

A. Multi-View EEG Data Representation

Suppose a single-trial raw EEG data represented as $x \in \mathbb{R}^{C\times T}$ and its corresponding label $y \in \{1, 2, \ldots, N_c\}$, where C, T, and N_c represent the number of EEG channels, sample points, and distinct classes, respectively.

ERD/ERS corresponds to variations in the synchrony of the underlying neuronal populations. The presence of ERD/ERS patterns is prominent in the alpha (8-13Hz) and beta (14-26Hz) frequency bands during MI tasks [30]. Many approaches [12], [25], [31] have employed a filter bank to decompose EEG signal into multiple narrow-band signals and achieved good performance on MI-EEG decoding. Therefore, in our research, we preprocess the raw EEG data $x \in \mathbb{R}^{C\times T}$ with a bank of bandpass filters $F = \{f_i\}_{i=1}^{N_b}$ and build a multi-view data representation $x_{fb} \in \mathbb{R}^{N_b\times C\times T}$ by concatenating the output filtered signals, where N_b represents the number of bandpass filters. Moreover, each view represents a filtered EEG signal in a specific narrow band.

The filter bank F can be composed of any number of filters with various frequency bands. In our study, following the FBCSP algorithm [12], we construct the filter bank using $N_b = 9$ filters with non-overlapping frequency bands, each of 4 Hz bandwidth, in range of 4-40 Hz (4-8, 8-12, ..., 36-40 Hz). The design of bandpass filters is based on the Chebyshev Type II filter with a transition bandwidth of 2 Hz and a stopband ripple of 30 dB [25].

B. Proposed Network Architecture

In this section, we present a novel temporal dependency learning CNN with the attention mechanism to decode MI-EEG signals. The proposed network consists of four parts, namely, spatial and spectral information learning, temporal segmentation and feature extraction, temporal attention module, and classification. The specifications of the proposed network are listed in Table I in detail.

1) Spatial and Spectral Information Learning: Given that EEG data is typically recorded by electrodes placed at different brain regions, the spatial information learning aims to encode spatial information from different channels of EEG data. We employ a spatial convolution layer with m output channels and the kernel size of $(C, 1)$ to work as m spatial filters over all channels. The convolution kernel size is set to the number of channels for integrating different information from all channels. In addition, as the input multi-view data represents filtered EEG signals in multiple frequency bands, the use of spatial convolution also contributes to the integration of spectral information from all frequency bands. A batch normalization layer [32] and an exponential linear unit (ELU) [33] are further adopted after the spatial convolution layer. Consequently, the whole process of spatial and spectral information learning outputs m time series $x_{sc} \in \mathbb{R}^{m\times 1\times T}$.

2) Temporal Segmentation and Feature Extraction: During MI tasks, EEG signal is typically recorded over a period of time to cover the whole process of task. EEG signals in different time periods present various MI-related patterns [31], [34]. Therefore, to capture MI-related patterns in different time periods, we employ n non-overlapped time windows of size w to segment the output x_{sc} along the time dimension and the discriminative feature is further extracted from each time window, where $n = \lfloor \frac{T}{w} \rfloor$. T represents the number of sample points and $\lfloor \cdot \rfloor$ means the rounding-down operation. Inspired by [25], due to the essential attribute of ERD/ERS, we employ a temporal variance layer to extract the discriminative feature from each time window. The temporal variance layer computes the variance of i^{th} time series in the k^{th} time window of size w along the time dimension as:

$$x_{tf}^{(k)}(i,1,1) = \frac{1}{w}\sum_{t=w\times k}^{(k+1)\times w-1}\left(x_{sc}(i,1,t) - \mu(i,1,k)\right)^2, \quad (1)$$

where $\mu(i,1,k)$ is the temporal mean of $x_{sc}(i,1,t)$ within the k^{th} time window.

As a result, the application of time window and temporal variance layer generates discriminative features $\{x_{tf}^{(1)}, x_{tf}^{(2)}, \cdots, x_{tf}^{(n)}\}$ in different time windows, where $x_{tf}^{(i)} \in \mathbb{R}^{m\times 1\times 1}$. Then, all features are passed through the logarithmic computation and fed into the temporal attention module for further feature extraction.

3) Temporal Attention Module: The aforementioned features in different time windows are independent of one another. Considering that every subject has their own way of performing MI tasks, we find that temporal dependencies among features in different time windows are essential to improve MI-EEG decoding performance. Hence, as shown in Fig. 2, a novel temporal attention module is designed to address this issue. The temporal attention module captures the importance of features in different windows with a set of weights and aggregates them across the time dimension. Instead of performing a single attention, multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions by incorporating multiple attention heads [27]. Therefore, we utilize a depthwise separable convolution [35] with h output channels to work as h parallel attention heads, which performs a convolution independently over every channel. Moreover, the weight of depthwise separable convolution is normalized along the time dimension using a softmax operation.

As illustrated in Fig. 2, all features $\{x_{tf}^{(1)}, x_{tf}^{(2)}, \cdots, x_{tf}^{(n)}\}$ are first divided into $\frac{m}{h}$ groups along the channel dimension, each group with h channels. Then, the features in the same groups are fed into the above-mentioned depthwise separable convolution. To perform the multi-head attention, we share the weight of depthwise separable convolution with different groups. Finally, computed features in different groups are concatenated to generate the output feature $x_f \in \mathbb{R}^{m\times 1\times 1}$. In summary, the temporal attention module computes for the output channel c, as follows:

$$x_f(c,1,1) = \sum_{j=1}^{n} W((c \bmod h),1,j)x_{tf}^{(j)}(c,1,1), \quad (2)$$

where $W \in \mathbb{R}^{h\times 1\times n}$ represents the weight of depthwise separable convolution and 'mod' means the modulus operation. The output feature x_f is then used for the classification.

4) Classification: Based on the above feature extraction, the classification is designed to provide the final decoding result. The feature x_f is first flattened into a 1-D feature vector. The vector is then fed into a fully connected layer. Finally, the label with the max value of output is considered as the final result.

C. Training Procedure for the Proposed Network

Early stopping is a common procedure to improve the training effect. This method has been widely adopted in the EEG processing area [21], [24], [25]. Therefore, in our study, we adopt a training procedure with early stopping during the training phase. The original training data is split into training and validation sets. The proposed network is first trained only using the training set, and the validation set accuracy is monitored. When the validation set accuracy does not increase within the early stopping patience, the first stage of training stops. After meeting the early stopping criteria, the network parameters with the best validation set accuracy are saved. Starting from the parameters saved in the first stage of training,

the proposed network continues to be trained on the original training data, which combines the training and validation sets. The second stage of training ends when the number of training epochs reaches up to the specified maximum. To avoid the case of non-convergence, in our experiment, the maximum number of training epochs is limited to 500 and 200 for the first and second stages of training, respectively. Moreover, the early stopping patience is set to 50 consecutive training epochs, within which the accuracy on the validation set does not increase.

In terms of network optimization, the model is trained by minimizing the cross-entropy loss. Adam optimizer [36] at default settings is adopted as the optimization method. The initial value of the learning rate is set to 0.001. The learning rate is decreased with a decay rate of 0.6, when the training loss has no improvement for 20 consecutive epochs.

III. Experiments

A. Evaluation Datasets

To demonstrate the effectiveness of our proposed network, we evaluate it on two public MI-EEG datasets, namely, BCI Competition IV 2a (BCIC-IV-2a) [37] and Korea University EEG dataset (OpenBMI) [38].

1) BCIC-IV-2a Dataset: The BCIC-IV-2a dataset contains EEG data from nine healthy subjects. This BCI paradigm is cue based, including four different MI tasks (left hand, right hand, feet, and tongue). The EEG signals were recorded at a sampling rate of 250 Hz by 22 Ag/AgCl electrodes and bandpass filtered between 0.5 and 100 Hz. Each subject has two sessions, namely, the training session and the evaluation session, recorded on different days. There are 288 trials (72 for each class) in each session, and each trial has 4 s duration.

2) OpenBMI Dataset: In the MI paradigm of the OpenBMI dataset, 54 healthy subjects perform two different MI tasks (left hand, right hand). Each subject has two sessions, and each class has 100 trials per session. In other words, there are 200 trials per session in total. Each trial also has 4 s duration. The EEG signals were originally collected using 62 Ag/AgCl electrodes at a sampling rate of 1000 Hz. In our research, as it is done in the original work [38], we selected 20 electrodes in the motor cortex region for MI-EEG decoding (FC-5/3/1/2/4/6, C-5/3/1/z/2/4/6, and CP-5/3/1/z/2/4/6). Furthermore, we performed a downsampling process using a factor of 4 to obtain signals at a sampling rate of 250 Hz, the same as the BCIC-IV-2a dataset. Note that the downsampling process is necessary, making it able to use a unified hyperparameter configuration for the proposed model.

For both datasets, in our study, EEG data of each trial is extracted using the same time window [0, 4] seconds relative to the cue onset. We take each trial as a sample, and each sample is represented as a 2D matrix of *channels* $\times$ *sample points*.

B. Evaluation Baselines

We compare the proposed network with four baseline models, one CSP-based algorithm FBCSP [12]; two classical CNN architectures, namely, Deep ConvNet [21] and EEGNet-8,2 [20]; and one state-of-the-art deep learning approach FBCNet [25]. All these models are retested on two datasets in our experiment.

1) FBCSP: FBCSP was developed to employ the original CSP algorithm on each sub-band of EEG signals and extract distinguishable EEG features from multiple frequency bands. In our experiment, we decomposed the raw EEG signal using multiple bandpass filters and selected four most discriminative CSP filters from each band. The raw EEG signal was then filtered using selected CSP filters, and the log variance of the filtered signal was extracted as the feature for classification. Subsequently, a support vector machine (SVM) was trained to classify the extracted feature. Finally, the SVM classifier with optimal parameters was used for testing.

2) Deep ConvNet: As a deep learning model, the design of the Deep ConvNet was based on classical CNN architectures and proven to be effective for decoding MI-EEG signals. In our experiment, the Deep ConvNet was configured in the optimal way, as recommended in [21].

3) EEGNet-8,2: EEGNet-8,2 was proposed as a compact deep learning model to address different BCI paradigms. In our experiment, we reproduced EEGNet-8,2 to perform a fair comparison and implemented the model following the description in the original publication [20]. Furthermore, raw EEG data was resampled at a sampling rate of 128 Hz before decoding.

4) FBCNet: Influenced by FBCSP, FBCNet employed a multi-view data representation and captured discriminative features from multiple frequency bands of the EEG signals. The model achieved state-of-the-art performance on several public MI-EEG datasets. In our experiment, the multi-view data representation of raw EEG data was constructed in the same way as in [25]. FBCNet was implemented under optimal conditions, as recommended in the original paper.

It's worth nothing that FBCSP, FBCNet, and our proposed method all employed the multi-view EEG data representation and used the same filter bank to preprocess EEG data. In our experiment, the filter bank consists of 9 narrow-band bandpass Chebyshev Type II filters, each has 4 Hz bandwidth, spanning from 4 to 40 Hz (4-8, 8-12, ..., 36-40 Hz) with transition bandwidth of 2 Hz and stopband ripple of 30 dB [25].

C. Experimental Methods

To demonstrate our proposed network as a generalized MI-EEG decoding model, we conducted experiments for the subject-specific analysis in the session-dependent and session-independent settings on two benchmark datasets (Fig. 3). The subject-specific analysis referred to training and test datasets from the same subject. Accuracy and F1-score were used to evaluate decoding performance of all considered methods.

The session-dependent experiment was conducted using a 10-fold cross validation to evaluate decoding performance for each subject, with 9 folds being used for training and 1 fold for testing, as illustrated in Fig. 3(a). All folds were constructed by a sequential and class-balanced allocation of trials, and this allocation was maintained constant during the entire evaluation process. The average decoding performance was calculated from 10 evaluations and used to represent the

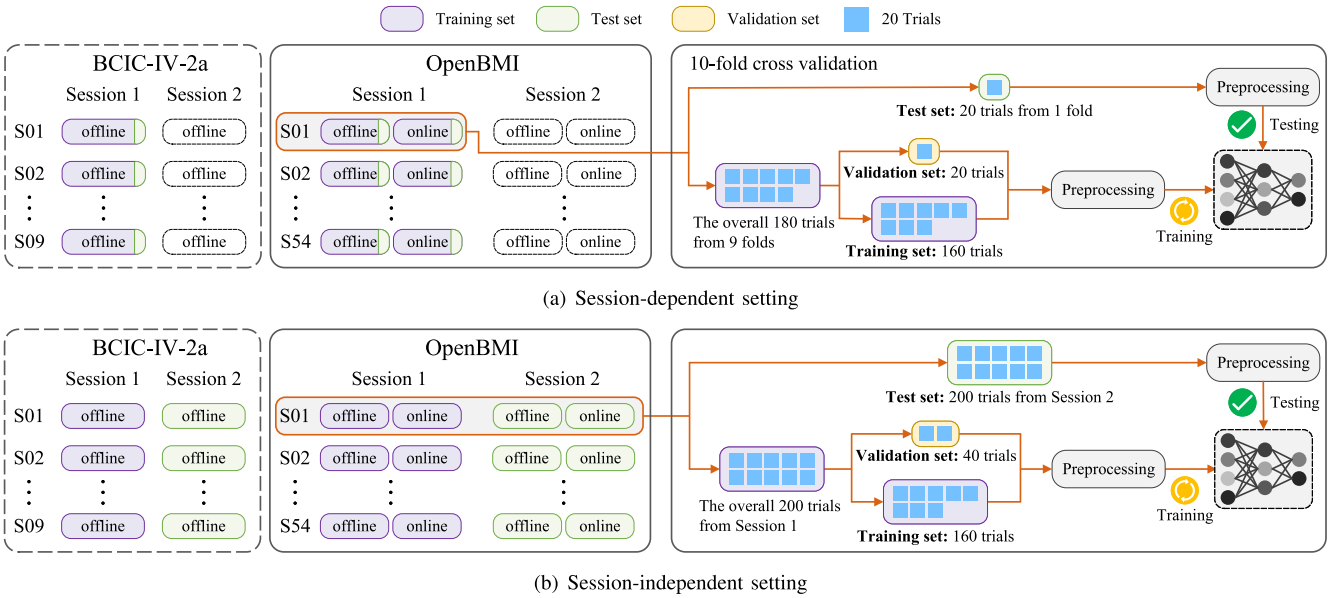


Fig. 3. Illustration of the experimental methods for subject-specific analysis in (a) session-dependent and (b) session-independent settings.

overall performance for the given subject. For the BCIC-IV-2a and OpenBMI datasets, we only used the data from Session 1 in the session-dependent setting to avoid the influence of cross-session variability.

In the session-independent setting, we conducted a cross-session experiment to understand the influence of cross-session variability on the decoding performance. Fig. 3(b) depicts an example of how we divided the training and test sets in the session-independent setting. For the BCIC-IV-2a and OpenBMI datasets, the entire data from Session 1 for the given subject was used for training, and the resulting model was tested on the entire data from Session 2.

In the default configuration of our proposed network, we set the number of spatial filters m to 64 and the number of attention heads h to 8. Given that the BCIC-IV-2a and OpenBMI datasets contain trials of 4 s duration at a sampling rate of 250 Hz, the time window size w was set to 250, which was equal to sample points in the duration of time 1 s. We carried out one-way repeated measures analysis of variance (ANOVA) with Bonferroni correction to analyze significant differences in decoding performance between our proposed network and all baseline methods [39]. We further conducted experiments for different hyperparameter combinations to evaluate the effect of different values of m, h, and w. The proposed deep learning network was implemented with Pytorch [40], and the computation for the deep learning model was implemented on NVIDIA GeForce GTX 1060 with 3 GB memory.

IV. Results

Our proposed method is a more competitive model in the presence of effective feature extraction with the novel temporal attention module. In this section, we present the results and evaluations of our proposed method to validate its effectiveness. The decoding performance of each evaluation is presented as average accuracy and F1-score with standard deviation (Accuracy ± Std and F1-score ± Std).

A. Overall Performance Comparison

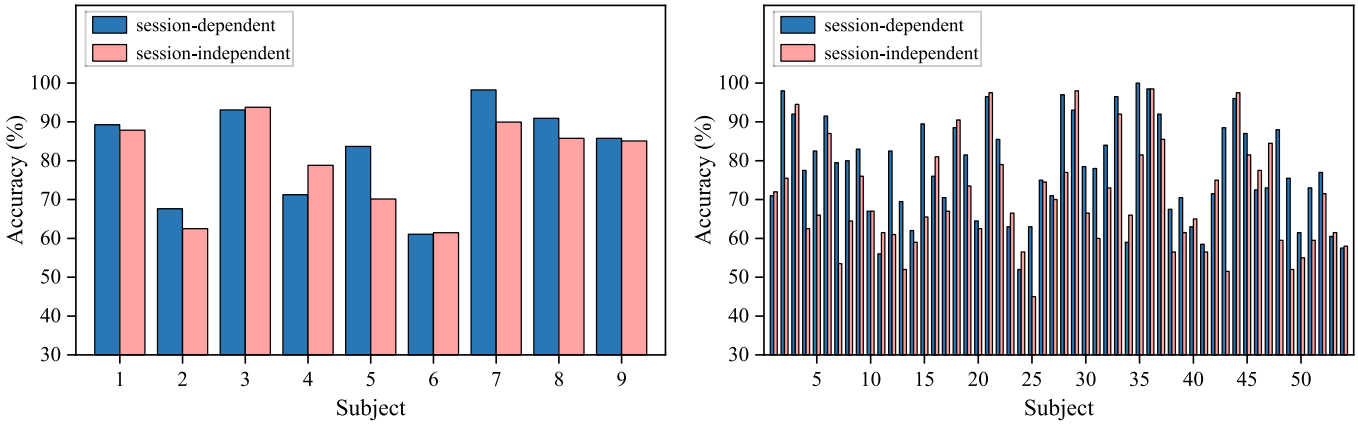
Table II presents the complete decoding results of our proposed method and four baseline methods across all subjects in the session-dependent and session-independent settings on both BCIC-IV-2a and OpenBMI datasets. As shown in Table II, our method achieves the best performance in terms of average accuracy and F1-score on both BCIC-IV-2a and OpenBMI datasets. The significant decoding performance improvement can be seen between our proposed method and other baseline methods on the BCIC-IV-2a dataset. In the session-dependent setting, our method reaches the highest average accuracy and F1-score of 82.32% and 82.03% on the BCIC-IV-2a dataset, which are significantly better than those of other baseline methods ($p < 0.05$). Meanwhile, in the session-independent setting, the average accuracy improvement of our method is significant ($p < 0.05$) on the BCIC-IV-2a dataset. Moreover, the average accuracy and F1-score of our method is 4.29% and 4.42% higher than that of FBCNet in the session-independent setting on the BCIC-IV-2a dataset, indicating its ability in extracting discriminative features under the use of temporal attention module. Considering the OpenBMI dataset, significant differences are seen between our method and other baseline methods in terms of average accuracy and F1-score in the session-dependent setting, $p < 0.05$. However, in the session-independent setting, the overall decoding performance of FBCNet is close to our method on the OpenBMI dataset. Furthermore, the significant performance improvement of our method compared to FBCNet is not found in the session-independent setting on the OpenBMI dataset.

To further explore the influence of cross-session variability, we compare the decoding performance of our proposed network in different settings on both BCIC-IV-2a and OpenBMI datasets. As illustrated in Fig. 4, the overall performance in the session-independent setting is slightly worse than the session-dependent setting. Even though our proposed network still suffers from cross-session variability,

TABLE II
COMPARATIVE DECODING PERFORMANCE (AVERAGE ACCURACY ± STD AND F1-SOCRE± STD) IN % IN THE SESSION-DEPENDENT AND SESSION-INDEPENDENT SETTINGS ON THEBCIC-IV-2A AND OPEMBMI DATASETS

Dataset	Method	Session-dependent		Session-independent	
		Accuracy	F1-score	Accuracy	F1-score
BCIC-IV-2a	FBCSP	74.16 ± 13.50	73.60 ± 13.80	66.51 ± 14.04	64.51 ± 15.98
	Deep ConvNet	71.10 ± 11.51	70.42 ± 11.36	70.87 ± 12.90	70.59 ± 13.00
	EEGNet-8,2	73.07 ± 11.54	72.68 ± 11.77	72.76 ± 11.19	72.68 ± 11.35
	FBCNet	79.25 ± 13.21	78.90 ± 13.48	75.19 ± 13.10	74.58 ± 13.66
	Our method	**82.32 ± 12.72***	**82.03 ± 12.92***	**79.48 ± 12.01***	**79.00 ± 12.58**
OpenBMI	FBCSP	64.44 ± 17.12	63.59 ± 17.77	59.46 ± 14.73	53.57 ± 23.20
	Deep ConvNet	63.94 ± 11.21	58.97 ± 14.77	59.62 ± 8.99	52.35 ± 21.37
	EEGNet-8,2	66.98 ± 12.29	66.54 ± 12.48	59.72 ± 8.62	56.27 ± 17.80
	FBCNet	75.77 ± 14.10	75.48 ± 14.27	68.45 ± 13.94	67.83 ± 15.36
	Our method	**77.52 ± 12.92***	**77.21 ± 13.20***	**70.43 ± 13.64**	**68.63 ± 18.62**

The best numerical values are highlighted in boldface, and * represents the performance value which is significantly higher than all comparison pairs, $p < 0.05$.



(a) BCIC-IV-2a (b) OpenBMI

Fig. 4. Subject specific decoding accuracy comparison in the session-dependent and session-independent settings on the BCIC-IV-2a and OpenBMI datasets.

it achieves satisfactory decoding performance on both public datasets.

B. Evaluation of Temporal Attention Module

Our proposed network is based on an important module, namely temporal attention module. The temporal attention module assigns different weights to discriminative features in different time windows according to their contribution to the final classification results and fuses them into more discriminative features. To evaluate its effectiveness, we conducted ablation studies on the proposed network. We also employed an ANOVA with Bonferroni correction for statistical analysis to analyze significant differences in decoding performance between the proposed network with and without temporal attention module.

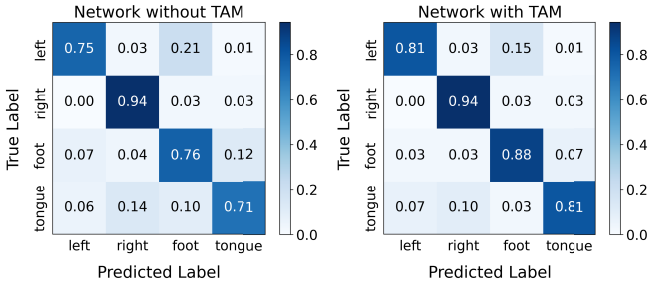
As shown in Table III, we compare the decoding performance of the proposed network with and without temporal attention module. It can be observed that the addition of temporal attention module greatly improves the performance of the proposed network in terms of average accuracy and F1-score in the session-dependent and session-independent settings on both BCIC-IV-2a and OpenBMI datasets. Sepcifically, the significant improvement can be seen between the proposed network with and without temporal attention module in the session-independent setting on the BCIC-IV-2a dataset, $p < 0.05$. As for the OpenBMI dataset, our method achieves significantly better performance than the method without temporal attention module in the session-dependent setting ($p < 0.05$). Considering that the decoding performance is more sensitive to these two different networks in the session-independent setting, we further analyze confusion matrices on the representative Subject 8 of the BCIC-IV-2a dataset and Subject 2 of the OpenBMI dataset. Fig. 5 illustrates the confusion matrices of the network with and without temporal attention module on the BCIC-IV-2a and OpenBMI datasets. According to Fig. 5(a), the temporal attention module greatly improves the decoding performance on foot and tongue MI tasks for the BCIC-IV-2a dataset. As for the OpenBMI dataset, the left-hand MI task is identified with higher accuracy because of the temporal attention module, as shown in Fig. 5(b). Consequently, the above results indicate

TABLE III

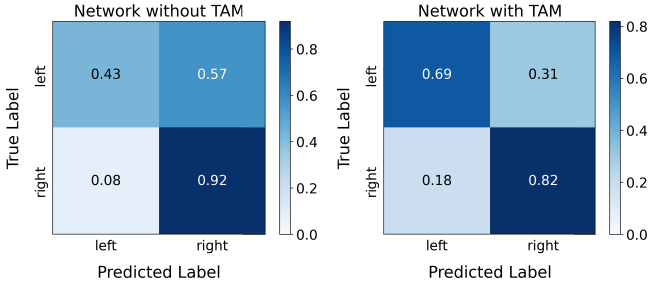
DECODING PERFORMANCE (AVERAGE ACCURACY ± STD AND F1-SOCRE± STD) IN % COMPARISON BETWEEN OUR METHOD AND OUR METHOD WITHOUT TEMPORALATTENTION MODULE (TAM)

Dataset	Method	Session-dependent		Session-independent	
		Accuracy	F1-score	Accuracy	F1-score
BCIC-IV-2a	Our method-w/o TAM	79.65 ± 11.34	79.21 ± 11.59	73.88 ± 13.00	73.38 ± 13.48
	Our method	82.32 ± 12.72	82.03 ± 12.92	79.48 ± 12.01*	79.00 ± 12.58*
OpenBMI	Our method-w/o TAM	75.53 ± 13.49	75.28 ± 13.63	66.45 ± 13.89	65.51 ± 17.09
	Our method	77.52 ± 12.92*	77.21 ± 13.20*	70.43 ± 13.64*	68.63 ± 18.62

* represents the performance value which is significantly higher than our method without TAM, $p < 0.05$.



(a) BCIC-IV-2a



(b) OpenBMI

Fig. 5. Confusion matrices of the proposed network with and without temporal attention module (TAM) on (a) the BCIC-IV-2a dataset (subject 8) and (b) the OpenBMI dataset (subject 2) in the session-independent setting.

that the temporal attention module is beneficial for extracting discriminative features and improving the MI-EEG decoding accuracy.

In addition, the time window size w and the number of attention heads h also play important roles in the temporal attention module to extract and fuse features, as presented in (2). The parameter w is defined as sample points over a specific time period and utilized to segment the time series into different temporal segments along the time dimension. Discriminative features are then extracted from different temporal segments and inputted into the temporal attention module. In other words, different values of w capture different information from the time series, resulting in different performances of the temporal attention module. Meanwhile, the temporal attention module utilizes multi-head attention to capture temporal dependencies. The parameter h is defined as the number of attention heads accordingly. Each attention head operates independently and learns to focus on different parts of the input. To provide more quantitative information for each parameter, we further examined the effect of w and h on the overall decoding performance. In our experiment,

TABLE IV

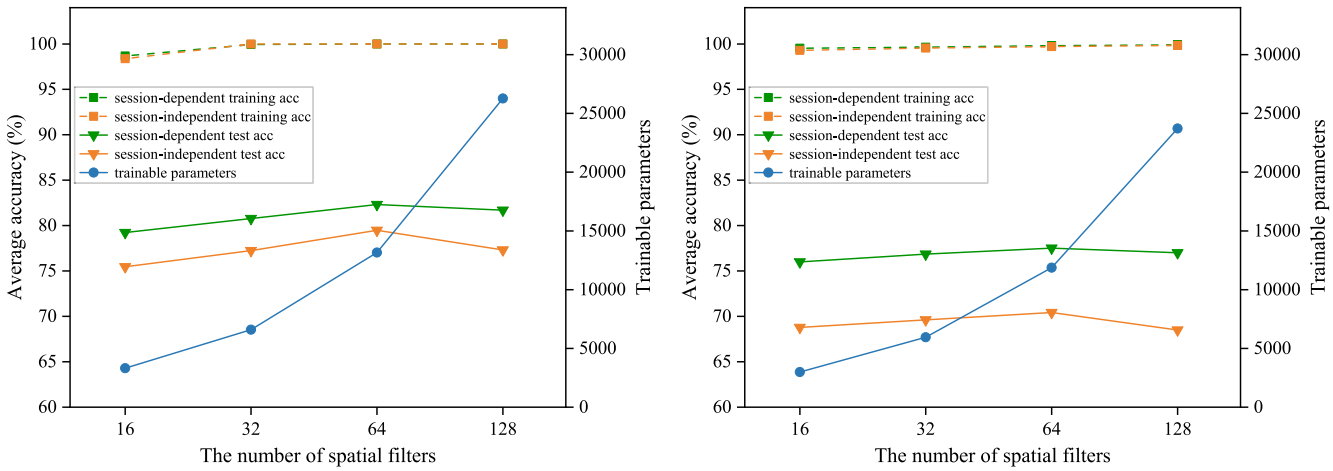
DECODING PERFORMANCE (AVERAGE ACCURACY ± STD AND F1-SOCRE± STD) IN % WITH DIFFERENT TIME WINDOW SIZES (w) AND DIFFERENT NUMBERS OFATTENTION HEADS (h). THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE

Dataset	w	h	Session-dependent		Session-independent	
			Accuracy	F1-score	Accuracy	F1-score
BCIC-IV-2a	125 (0.5 s)	4	81.24 ± 12.59	80.74 ± 13.18	77.70 ± 12.96	77.04 ± 13.75
		8	82.05 ± 12.50	81.70 ± 12.87	78.36 ± 12.59	77.83 ± 13.22
		16	81.98 ± 12.21	81.60 ± 12.54	77.82 ± 13.02	77.07 ± 13.94
	250 (1 s)	4	81.39 ± 13.12	81.07 ± 13.33	77.82 ± 12.10	77.28 ± 12.64
		8	**82.32 ± 12.72**	**82.03 ± 12.92**	**79.48 ± 12.01**	**79.00 ± 12.58**
		16	81.40 ± 13.18	81.00 ± 13.46	77.74 ± 11.82	77.10 ± 12.54
	500 (2 s)	4	79.97 ± 11.86	79.54 ± 12.07	75.23 ± 12.32	74.50 ± 13.25
		8	79.20 ± 12.51	78.87 ± 12.78	76.31 ± 12.25	75.67 ± 13.09
		16	80.09 ± 12.04	79.76 ± 12.22	75.50 ± 12.35	74.68 ± 13.46
OpenBMI	125 (0.5 s)	4	77.61 ± 13.47	77.37 ± 13.65	70.01 ± 13.54	67.52 ± 20.15
		8	**77.69 ± 13.24**	**77.44 ± 13.42**	**70.50 ± 13.30**	**69.63 ± 16.26**
		16	77.31 ± 13.50	77.06 ± 13.68	70.30 ± 13.40	68.83 ± 18.56
	250 (1 s)	4	76.94 ± 13.65	76.67 ± 13.84	70.33 ± 13.54	69.25 ± 17.24
		8	77.52 ± 12.92	77.21 ± 13.20	70.43 ± 13.64	68.63 ± 18.62
		16	77.02 ± 13.64	76.75 ± 13.86	70.11 ± 13.75	68.23 ± 19.79
	500 (2 s)	4	76.19 ± 13.62	75.86 ± 13.85	69.44 ± 13.77	68.89 ± 17.13
		8	75.37 ± 13.77	75.07 ± 13.98	69.62 ± 13.15	68.69 ± 17.00
		16	75.38 ± 13.68	75.08 ± 13.88	69.53 ± 13.15	68.45 ± 17.31

EEG data from both BCIC-IV-2a and OpenBMI datasets is at a sampling rate of 250 Hz and each trial has 4 s duration. Therefore, the parameter search for w and h was carried out in the set $\{125, 250, 500\}$ and $\{4, 8, 16\}$, respectively. The chosen set $\{125, 250, 500\}$ for w represents sample points in the duration of time 0.5 s, 1 s, and 2 s on both datasets, respectively. Moreover, the selected number of time windows is set to $\{2, 4, 8\}$. Table IV summarizes different decoding performances when different values of the parameters w and h are configured in the temporal attention module. It can be observed that w has a greater influence on the final decoding performances than h, especially on the BCIC-IV-2a dataset. Given that ERD/ERS patterns associated with MI have a duration of time 0.5 s to 1 s, our model benefits more from the proper value of w [25]. The overall performance of the time window size $w = 250$ and $w = 125$ outperforms other sizes in the session-independent and session-dependent settings on the BCIC-IV-2a and OpenBMI datasets, respectively. Moreover, an extremely large time window size ($w = 500$) reduces decoding accuracy. It indicates that the proposed network is not able to fully explore temporal dependencies under the use of a large time window size. When the time window size w is configured, the number of attention heads h has a noticeable impact on the final decoding performance in the session-independent setting on the BCIC-IV-2a dataset. However, on the OpenBMI dataset, only subtle differences are seen among different values of h. To sum up, the parameter h value of 8 contributes to the best decoding performance on both BCIC-IV-2a and OpenBMI datasets.

C. Evaluation of Spatial Information Learning and Parameter Sensitivity

The number of spatial filters m introduced in the spatial information learning is a hyperparameter used to balance the capacity and computational efficiency of the spatial



(a) BCIC-IV-2a (b) OpenBMI

Fig. 6. Influence of the number of spatial filters m on the number of trainable parameters, training accuracy, and test accuracy on the BCIC-IV-2a and OpenBMI datasets.

TABLE V
DECODING PERFORMANCE (AVERAGE ACCURACY ± STD AND F1-SOCRE± STD) IN % AND TRAINABLE PARAMETERS WITH DIFFERENT NUMBERS OF SPATIAL FILTERS(m). THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLDFACE

Dataset	m	Trainable parameters	Session-dependent		Session-independent	
			Accuracy	F1-score	Accuracy	F1-score
BCIC-IV-2a	16	3316	79.23 ± 15.36	78.66 ± 15.87	75.46 ± 14.63	74.89 ± 15.16
	32	6596	80.78 ± 14.30	80.40 ± 14.60	77.24 ± 13.55	76.59 ± 14.43
	64	13156	**82.32 ± 12.72**	**82.03 ± 12.92**	**79.48 ± 12.01**	**79.00 ± 12.58**
	128	26276	81.70 ± 11.65	81.36 ± 12.02	77.32 ± 13.20	76.58 ± 14.11
OpenBMI	16	2994	76.00 ± 14.22	75.73 ± 14.41	68.78 ± 14.33	67.74 ± 17.76
	32	5954	76.85 ± 13.94	76.59 ± 14.14	69.61 ± 13.16	68.54 ± 15.70
	64	11874	**77.52 ± 12.92**	**77.21 ± 13.20**	**70.43 ± 13.64**	**68.63 ± 18.62**
	128	23714	77.00 ± 13.00	76.75 ± 13.18	68.50 ± 13.58	66.85 ± 17.40

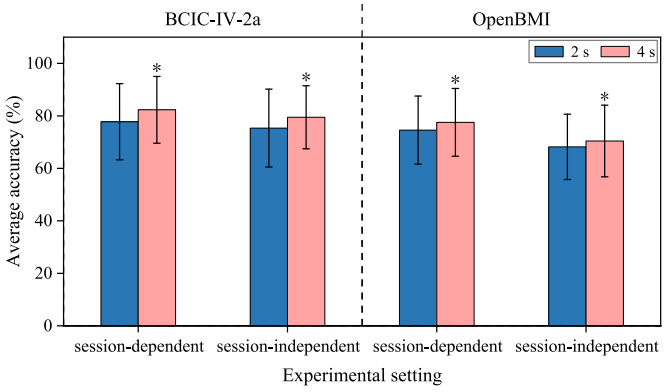


Fig. 7. Results of decoding EEG signals extracted from time windows [0.5s, 2.5s] (2 s) and [0s, 4s] (4 s) on the BCIC-IV-2a and OpenBMI datasets, where * represents the significant difference between different EEG signals for decoding, $p < 0.05$.

convolution in the network. It controls the spatial and spectral learning capacity of our model. In addition, as presented in Table I, it determines the total number of trainable parameters to a great extent. To investigate the trade-off between decoding performance and computational efficiency, we examined the effect of different values of m in the set {16, 32, 64, 128}.

Table V presents the comparison of decoding performance with different values of m. It can be seen that the decoding performance of the proposed network is fluctuant to the value of m, and the value 64 results in the best performance on both BCIC-IV-2a and OpenBMI datasets. In addition, the parameter m plays a critial role in determining the total number of trainable parameters of the proposed network, which can be used to reduce energy consumption with less number of operations and model size [41]. To further analyze the parameter sensitivity of the proposed network, we explore the relation between the total number of trainable parameters and the decoding performance by increasing the number of spatial filters m. As illustrated in Fig. 6, the parameter curves show that the number of trainable parameters is approximately linearly proportional to the number of spatial filters. However, not only do incremental spatial filters not improve the decoding performance monotonically, but they also add complexity to the entire model, which makes the model less cost-effective. The same phenomenon can be observed on both BCIC-IV-2a and OpenBMI datasets. This is probably due to the heavy overfitting problem. To sum up, $m = 64$ strikes a good balance between decoding performance and computational efficiency.

D. Comparison of Decoding EEG Signals Extracted From Different Time Periods

Decoding EEG signals involves analyzing and interpreting the brain's electrical activity recorded over a specific duration. The choice of time period for decoding can impact the information extracted from the EEG signal. To explore the influence of the choice of time period, we further conducted experiments on EEG signals extracted from different time periods. Compared with the previous study (time window [0s, 4s]), we used a time window [0.5, 2.5] seconds post cue onset (the same window which was used in [20] and [42]) to extract EEG signals for decoding.

Fig. 7 illustrates the results of decoding EEG signals extracted from different time periods on both BCIC-IV-2a and OpenBMI datasets. It can be seen that our proposed network achieves better performance when decoding EEG signals extracted from the time window [0s, 4s] than [0.5s, 2.5s].

TABLE VI
TIME CONSUMPTION OF TRAINING (T_{train}) AND PREDICTION (T_{test}) IN SECONDS PER EPOCHAND THE NUMBER OF TRAINABLE PARAMETERS OF DIFFERENT DEEP LEARNING METHODS IN THE SESSION-INDEPENDENTSETTING

Dataset	Method	Trainable parameters	$T_{train}(s)$	$T_{test}(s)$
BCIC-IV-2a	Deep ConvNet	283254	0.356	0.123
	EEGNet-8,2	1716	0.094	0.043
	FBCNet	11812	0.352	0.258
	Our method	13156	0.242	0.222
OpenBMI	Deep ConvNet	282004	0.243	0.083
	EEGNet-8,2	1426	0.064	0.031
	FBCNet	8930	0.257	0.183
	Our method	11874	0.169	0.134

Moreover, in a paired t-test, there are significant differences in the average accuracy between the time window [0s, 4s] and [0.5s, 2.5s], $p < 0.05$. Although decoding short time periods allows for capturing rapid changes in brain activity, short time periods may not capture long-term trends or sustained brain activity. Our proposed network is designed to capture temporal dependencies between features in different time periods. Therefore, EEG signals over long time periods are more suitable for our proposed network.

V. DISCUSSION

A. Analysis of the Proposed Method

Recently, deep learning has become popular in BCI due to its capability of effectively learning the brain activity patterns from EEG data. A deep learning model is expected to have as fewer trainable parameters as possible to ensure its robust generalizability because of the scarcity of EEG training data. To further evaluate the feasibility of our model, we compare the number of model parameters and time computation of model training and testing with baseline deep learning models. The results are listed in Table VI. Note that the measurement is carried out in the session-independent setting for ease of representation. The training time is defined as the duration time of each training epoch, whereas the test time is defined as the duration time for classifying all test trials from another session.

According to Table VI, the training time of our proposed model is 0.242 s and 0.169 s on the BCIC-IV-2a dataset and the OpenBMI dataset, respectively. Although our proposed model has more parameters than the similar model FBCNet, it consumes less time when training. Except for the compact model EEGNet, which has the least trainable parameters, the test time difference among other three models is not significant. The compactness of our proposed model highly relies on the use of temporal attention module, which has fewer parameters and efficiently fuses features in different time windows into more discriminative features. In addition, the use of time window and temporal variance layer greatly reduces the feature vector dimension by extracting relevant features from different time windows.

To give insight into the training process of the proposed network, we demonstrate the trend of training loss and test accuracy during training, as shown in Fig. 8. Due to our training procedure with early stopping, when the first stage of training stops, the proposed network continues to be trained on the original training data which combines the training and validation sets. Therefore, the training loss has a sudden increase after meeting the early stopping criteria. In addition, it can be observed that the whole training process is stable under the use of temporal attention module. These findings suggest that the proposed model achieves an appropriate trade-off between complexity and decoding performance.

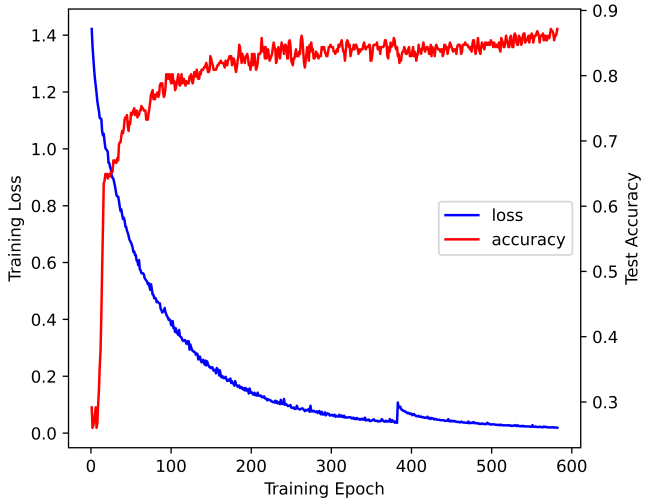


Fig. 8. Training loss and test accuracy during training of the proposed network.

B. Comparison of Different Decoding Methods

Our proposed model aims to focus on features in different time periods during MI tasks by incorporating a novel temporal attention module. The results in Table II strongly support the efficacy of our model, wherein our model has achieved the best decoding performance across two public datasets. To further validate the performance of our model, we compare the decoding results reported by recent studies on the BCIC-IV-2a dataset. Notably, for the OpenBMI dataset, most studies [43], [44] reported their MI-EEG decoding results using different evaluation methods, making it difficult for a fair comparison. Therefore, the performance comparison on the OpenBMI dataset is not provided in our study.

Table VII presents the decoding results reported by recent studies on the BCIC-IV-2a dataset. All these studies train their models on Session 1 and test on Session 2 for each subject, namely the session-independent setting in our experiment. As shown in Table VII, our proposed model achieves the best decoding performance among these models on the BCIC-IV-2a dataset. In contrast to the models in [22] and [45], our model directly learns spatial and spectral information from the multi-view EEG data without the need to initially process raw EEG signals with spatial filters based on FBCSP, hence gaining the accuracy increase of 5.02% and 7.48%, respectively. Studies [24], [46], and [47] concentrate on exploring different domain information to extract features with deep and complex network architectures. However, they all

TABLE VII
COMPARISON WITH EXISTING METHODS ON THE BCIC-IV-2A DATASET

Dataset	Method	Year	Average accuracy(%)
BCIC-IV-2a	C2CM [22]	2018	74.46
	CP-MixedNet [46]	2019	74.60
	FBSF-TSCNN [45]	2020	72.00
	MI-EEGNET [47]	2021	74.61
	TS-SEFFNet [24]	2021	74.71
	MI-BMInet [48]	2022	77.18
	Our method	**2022**	**79.48**

The proposed method is highlighted in boldface.

ignore the variability of features in different time periods and use a fixed time window to extract features, resulting in limited decoding performance. In addition, it spends a longer time training those models because they all have a larger amount of parameters. Recent MI-BMInet [48] proposed an automatic EEG channel selection method based on spatial filters to select the most relevant EEG channels for MI-EEG decoding. By contrast, our method emphasizes the importance of temporal dependencies between features in different time periods and achieves 2.3% higher than MI-BMInet in terms of average accuracy. These evidences imply that learning temporal dependencies among features in different time periods can help our network capture discriminative features and improve MI-EEG decoding performance.

C. Effectiveness of Temporal Attention Module

The temporal attention module plays the most important role in our proposed network. It is designed to capture temporal dependencies among features in different time periods because every subject has their own way of performing MI tasks. The experimental results in Table III have confirmed the effectiveness of the temporal attention module. To further investigate the role of temporal attention module in extracting features, the t-SNE [49] method is used to visualize the learned features. Fig. 9 illustrates the t-SNE projection of the learned features with and without the temporal attention module on the BCIC-IV-2a dataset. The extracted features with the temporal attention module are clustered toward a more compact form in each class. By comparison, the extracted features without the temporal attention module appear to be dispersed throughout the projection space. This finding suggests that the temporal attention module contributes to enhancing the discrimination of the learned features among different classes.

D. Influence of Hyperparameter Selection

Our proposed model has three important hyperparameters, the number of spatial filters (m), the time window size (w), and the number of attention heads (h). The selection of these three hyperparameters can significantly impact the decoding performance and generalization ability of our model. Among these three hyperparameters, the value of m plays the most important role in our model. According to Table I, it can be found that the parameter m has a great impact on the total number of trainable parameters. In addition, the parameter m

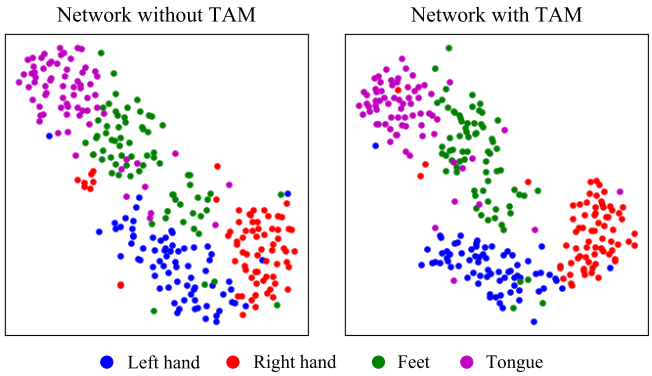


Fig. 9. Features extracted by our proposed network with and without temporal attention module (TAM) for Subject 3 of the BCIC-IV-2a dataset in the session-independent setting. The extracted features are projected into the two-dimensional embedding space by the t-SNE method.

impacts the model's capacity to learn complex patterns and representations. As shown in Table V, $m = 64$ strikes the best balance between decoding performance and computational efficiency in the set $\{16, 32, 64, 128\}$. Moreover, too large value of m leads to reduced accuracy for MI-EEG decoding, which is probably due to the heavy overfitting problem, as illustrated in Fig. 6.

The parameters w and h influence the temporal dependency learning ability of the temporal attention module. The parameter w is defined as sample points over a period of time and utilized to segment the time series along the time dimension. Discriminative features are extracted from different temporal segments and then inputted into the temporal attention module. The selection of w is to select a proper time period in which the model is able to extract discriminative features and build bridges between them. In our experiment, sample points in the time period 0.5 s, 1 s and 2 s were chosen for the parameter w. Meanwhile, the temporal attention module utilizes multi-head attention to weigh the importance of different features and the parameter h is defined as the number of attention heads accordingly. Different attention heads operate independently and focus on different parts of the input, enabling the temporal attention module to capture diverse relationships and dependencies. Compared with the parameter h, as shown in Table IV, the parameter w has a greater impact on final decoding results than h. It indicates that the model's ability to effectively attend to and capture the relevant information is more critical than the exact positions in the sequence that receive high attention weights. In other words, the model's understanding of the content and its ability to extract the relevant information are more important than the specific position of the tokens in the sequence. To sum up, proper selection of these hyperparameters through experimentation and validation is necessary to achieve the best possible results for MI-EEG decoding.

E. Visualization of Learned Features

To further investigate the capacity of different deep learning methods to extract highly discriminative features, we employ the t-SNE method to visualize different learned features. The t-SNE method is applied to the input of the last fully connected

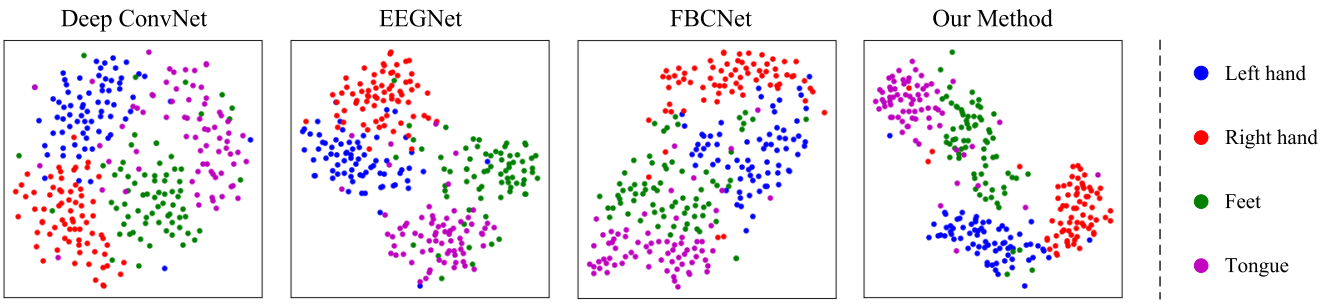


Fig. 10. Visualization of features learned by different deep learning methods on the BCIC-IV-2a dataset using two-dimensional t-SNE projection. The learned features are picked from Subject 3 in the session-independent setting for visualization purposes.

layer in all different trained models. The learned features are further projected into the two-dimensional embedding space. According to Fig. 10, compared with Deep ConvNet, EEGNet, and FBCNet, the features learned by our method appear to be more compact in each class and can be easily classified. By contrast, the features learned by other methods are relatively ambiguous, leading to reduced decoding accuracy. This finding indicates that our method extracts the most discriminative features among these deep learning methods and consequently achieves the best MI-EEG decoding performance.

F. Limitations and Future Directions

Although the proposed network achieves promising decoding results, there are still several limitations. Firstly, the filter bank and EEG electrods are selected manually, which may leads to the suboptimal decoding performance. Many studies have focused on different selection methods to enhance the decoding performance [50], [51], [52]. Thus, in the future study, we will focus on incorporating adaptively-selecting method into our proposed network. Secondly, the designed temporal attention module is based on depthwise convolution, however, the transformer module has been widely used to extract global dependency [53], [54]. Therefore, the transformer module will be considered in our future work to improve temporal dependency learning. Thirdly, we can explore our proposed network on other EEG measurements, such as SSVEP and ERP. Finally, although subject-specific experiments shows the superiority of the proposed method, the proposed method can not be used for cross-subject tasks directly. Therefore, the transfer learning [44], [55] will be explored to improve the generalization capability of our model in the future.

VI. Conclusion

This paper proposes a novel temporal dependency learning CNN architecture with attention mechanism to decode MI-EEG signals. In this architecture, a novel temporal attention module is designed to capture temporal dependencies among discriminative features in different time windows. Moreover, the temporal attention module assigns different weights to features in different time windows and fuses them into features with high discrimination. Experiments are conducted on two public MI-EEG datasets to evaluate the effectiveness of the proposed method. The experimental results reveal that our method achieves significantly better decoding performance than other compared methods. With interpretability analysis, we demonstrate that the improved performance is driven by efficiently capturing temporal dependencies among discriminative features in different time windows. This finding indicates that learning temporal dependencies can be regarded as a potential approach to improve the performance of MI-EEG based BCI systems.

References

[1] D. McFarland and J. Wolpaw, "Eeg-based brain–computer interfaces," *Current Opinion Biomed. Eng.*, vol. 4, pp. 194–200, Dec. 2017.

[2] R. Mane, T. Chouhan, and C. Guan, "BCI for stroke rehabilitation: Motor and beyond," *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 041001.

[3] D. Camargo-Vargas, M. Callejas-Cuervo, and S. Mazzoleni, "Brain–computer interfaces systems for upper and lower limb rehabilitation: A systematic review," *Sensors*, vol. 21, no. 13, p. 4312, Jun. 2021.

[4] I. Lazarou, S. Nikolopoulos, P. C. Petrantonakis, I. Kompatsiaris, and M. Tsolaki, "EEG-based brain–computer interfaces for communication and rehabilitation of people with motor impairment: *A novel approach of the 21st century*," *Frontiers Human Neurosci.*, vol. 12, p. 14, Jan. 2018.

[5] J. Faller, G. Müller-Putz, D. Schmalstieg, and G. Pfurtscheller, "An application framework for controlling an avatar in a desktop-based virtual environment via a software SSVEP brain–computer interface," *Presence, Teleoperators Virtual Environ.*, vol. 19, no. 1, pp. 25–34, Feb. 2010.

[6] Y. Yu, Y. Liu, E. Yin, J. Jiang, Z. Zhou, and D. Hu, "An asynchronous hybrid spelling approach based on EEG–EOG signals for Chinese character input," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1292–1302, Jun. 2019.

[7] Y. Yu, Y. Liu, J. Jiang, E. Yin, Z. Zhou, and D. Hu, "An asynchronous control paradigm based on sequential motor imagery and its application in wheelchair navigation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 12, pp. 2367–2375, Dec. 2018.

[8] G. Pfurtscheller and C. Neuper, "Motor imagery activates primary sensorimotor area in humans," *Neurosci. Lett.*, vol. 239, nos. 2–3, pp. 65–68, Dec. 1997.

[9] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 392–401, Apr. 2017.

[10] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, 2007.

[11] H. Ramoser, J. M´uller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[12] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in brain–computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 2390–2397.

[13] Y. Zhang, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface," *J. Neurosci. Methods*, vol. 255, pp. 85–91, Nov. 2015.

[14] K. Das and R. B. Pachori, "Electroencephalogram based motor imagery brain computer interface using multivariate iterative filtering and spatial filtering," *IEEE Trans. Cogn. Develop. Syst.*, early access, Oct. 14, 2022, doi: 10.1109/TCDS.2022.3214081.

[15] J. Jin, R. Xiao, I. Daly, Y. Miao, X. Wang, and A. Cichocki, "Internal feature selection method of CSP based on L1-norm and Dempster–Shafer theory," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4814–4825, Nov. 2021.

[16] S. Bhalerao and R. Pachori, "Automatic detection of motor imagery EEG signals using swarm decomposition for robust BCI systems," in *Human-Machine Interface Technology Advancements and Applications*. CRC Press, Oct. 2022.

[17] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[18] P. Gaur, A. Chowdhury, K. McCreadie, R. B. Pachori, and H. Wang, "Logistic regression with tangent space-based cross-subject learning for enhancing motor imagery classification," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 3, pp. 1188–1197, Sep. 2022.

[19] F. Lotte et al., "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 031005.

[20] 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.

[21] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

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

[23] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019.

[24] Y. Li, L. Guo, Y. Liu, J. Liu, and F. Meng, "A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1534–1545, 2021.

[25] R. Mane et al., "FBCNet: A multi-view convolutional neural network for brain–computer interface," 2021, *arXiv:2104.01233*.

[26] J. Jin, Z. Wang, R. Xu, C. Liu, X. Wang, and A. Cichocki, "Robust similarity measurement based on a novel time filter for SSVEPs detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 14, 2021, doi: 10.1109/TNNLS.2021.3118468.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[28] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[29] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," 2019, *arXiv:1901.10430*.

[30] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.

[31] V. Mishuhina and X. Jiang, "Complex common spatial patterns on time-frequency decomposed EEG for brain–computer interface," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107918.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[33] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[34] 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.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[37] M. Tangermann et al., "Review of the BCI competition IV," *Frontiers Neurosci.*, vol. 6, p. 55, Jul. 2012.

[38] M.-H. Lee et al., "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, May 2019, Art. no. giz002.

[39] P. Autthasan et al., "MIN2Net: End-to-end multi-task learning for subject-independent motor imagery EEG classification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 6, pp. 2105–2118, Jun. 2022.

[40] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[41] N. Phukan, S. Mohine, A. Mondal, M. S. Manikandan, and R. B. Pachori, "Convolutional neural network-based human activity recognition for edge fitness and context-aware health monitoring devices," *IEEE Sensors J.*, vol. 22, no. 22, pp. 21816–21826, Nov. 2022.

[42] F. Lotte, "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces," *Proc. IEEE*, vol. 103, no. 6, pp. 871–890, Jun. 2015.

[43] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2020.

[44] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

[45] J. Chen, Z. Yu, Z. Gu, and Y. Li, "Deep temporal–spatial feature learning for motor imagery-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2356–2366, Nov. 2020.

[46] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019.

[47] 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.

[48] X. Wang, M. Hersche, M. Magno, and L. Benini, "MI-BMInet: An efficient convolutional neural network for motor imagery brain–machine interfaces with EEG channel selection," 2022, *arXiv:2203.14592*.

[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008.

[50] P. Gaur, K. McCreadie, R. B. Pachori, H. Wang, and G. Prasad, "An automatic subject specific channel selection method for enhancing motor imagery classification in EEG-BCI using correlation," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102574.

[51] S. V. Bhalerao and R. B. Pachori, "Sparse spectrum based swarm decomposition for robust nonstationary signal analysis with application to sleep apnea detection from EEG," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103792.

[52] S. Bhalerao, S. Ainwad, and R. Pachori, "FBSE based automated classification of motor imagery EEG signals in brain–computer interface," in *Handbook of Neural Engineering* (Handbook of Neural Engineering), vol. 2. Elsevier, Oct. 2022.

[53] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023.

[54] H.-J. Ahn, D.-H. Lee, J.-H. Jeong, and S.-W. Lee, "Multiscale convolutional transformer for EEG classification of mental imagery in different modalities," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 646–656, 2023.

[55] F. Mattioli, C. Porcaro, and G. Baldassarre, "A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain–computer interface," *J. Neural Eng.*, vol. 18, no. 6, Dec. 2021, Art. no. 066053.