



DESAnet: dual-branch spatial-temporal CNN with efficient channel and sparse attention networks for motor imagery classification

Bin Mao^{a,b}, Fei Wang^b, Yuanhang Zhou^a, Yufeng Deng^{a,b}, Jiahao Cheng^a, Sheng Guo^a,
Dezhong Yao^c, Peng Xu^c, X. San Liang^{b,d}, Tao Zhang^{a,b,*}

^a Mental Health Education Center and School of Science, Xihua University, Chengdu, China

^b Artificial Intelligence Department, Division of Frontier Research, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

^c MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China

^d Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai, China

ARTICLE INFO

Keywords:

Brain-computer interface
Motor imagery
Convolution neural network
Efficient channel attention
Sparse attention

ABSTRACT

Motor imagery brain-computer interfaces (MI-BCIs) play a vital role in neural interaction and clinical rehabilitation. However, their real-world deployment is often hindered by two major limitations: the relatively low decoding accuracy of MI tasks and the high computational cost of deep learning models. To address these challenges, we propose DESAnet—a lightweight end-to-end framework that incorporates a dual-branch spatial-temporal convolution module (DBSTCV), an efficient channel attention (ECA) module, and a sparse attention (SA) module. The dual branches capture multi-scale spatial and temporal structure in EEG, ECA adaptively emphasizes informative channels, and SA efficiently models long-range temporal dependencies with reduced overhead. We evaluate DESAnet on BCI Competition IV-2a (four-class) and IV-2b (two-class) under a subject-dependent setting. The model attains an average accuracy of 85.69% on 2a and 89.40% on 2b, outperforming representative deep-learning-based methods reported under comparable conditions. In addition, DESAnet maintains a low parameter count, compact model size, and few multiply-accumulate operations per forward pass, demonstrating favorable computational efficiency for resource-constrained or real-time MI-BCI use. Visualization with t-distributed stochastic neighbor embedding (t-SNE), attention weight maps, and class activation mapping (CAM) indicates clearer class separation and temporally focused responses, supporting the model's interpretability and design choices. DESAnet can effectively improve the classification performance of MI-BCI signals while ensuring computational efficiency. This study provides a potentially promising technical solution for advancing the practical application of MI-BCI, and may hold valuable reference significance for the development of medical rehabilitation and intelligent interaction systems in relevant scenarios.

1. Introduction

Brain-computer interfaces (BCIs) represent a transformative technology that establishes direct neural pathways between the human brain and external devices, bypassing traditional muscular outputs [1]. Among non-invasive modalities, electroencephalography (EEG) stands out for its portability, temporal resolution, and affordability, rendering it indispensable in clinical rehabilitation and assistive applications [2]. Within EEG-based systems, motor imagery BCIs (MI-BCIs)—which elicit neural activity through imagined limb movements—show strong potential in rehabilitation and assistive technologies [3]. However, reliable

decoding of MI signals remains challenging: EEG is usually noisy, easily affected by physiological artifacts, and exhibits large inter-subject and intra-subject variability [4]. Thus, devising robust, accurate decoding algorithms remains a pivotal imperative for advancing MI-BCI efficacy.

Conventional MI-BCI decoding pipelines predominantly rely on handcrafted feature engineering coupled with machine learning classifiers. Spatial filtering techniques, notably common spatial pattern (CSP) [5] and its variants [6,7], have long dominated spatial feature extraction. For example, Mishuhina et al. [8] employed CSP alongside time-frequency decomposition, followed by weighted regularization, to classify MI tasks. Luo et al. [9] use time-frequency decomposition and

* Corresponding author at: Mental Health Education Center and School of Science, Xihua University, Chengdu, China.

E-mail address: zhangtao1698@xhu.edu.cn (T. Zhang).

<https://doi.org/10.1016/j.bspc.2026.111458>

Received 4 February 2026; Received in revised form 11 August 2026; Accepted 10 September 2026

Available online 13 September 2026

1746-8094/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

multi-view feature extraction techniques to selectively extract key features from the manifold embedding space, combining the advantages of Euclidean space and Riemannian space. While these approaches adeptly capture motor-related desynchronization/synchronization (ERD/ERS) in mu and beta rhythms, they are encumbered by task-specific priors, demanding manual tuning that curtails generalization across sessions or subjects.

The advent of deep learning (DL) has revolutionized this landscape, including computer vision and natural language processing by automating hierarchical feature representation, obviating labor-intensive designs [10,11]. Architectures like convolutional neural networks (CNN) [12], recurrent neural networks (RNN) [13], and generative adversarial networks (GAN) [14] demonstrated prowess in distilling latent patterns from raw signals, minimizing reliance on domain expertise. Inspired by these advances, DL has been increasingly applied to MI-BCI classification tasks [15], showing strong potential in capturing temporal, spectral, and spatial patterns inherent in EEG. For example, Chen et al. [16] proposed an FBSF-TSCNN model for the MI-BCI classification task. The model first applies spatial filtering to attenuate irrelevant components and then extracts discriminative temporal and spatial representations through temporal and spatial convolutions, indicating the importance of multi-domain information for MI decoding. Amin et al. [17] and Li et al. [18] redefined the MI-BCI classification task as an image classification problem. They explored a CNN-based attention mechanism network and a long short-term memory-based feature fusion network, respectively, and further explored the potential of DL in feature extraction. Furthermore, our previous studies [19,20] further evidenced that multi-branch CNNs or temporal convolutional networks, tuned to band-specific dynamics and augmented with efficient channel attention (ECA), amplify representational depth in spatial-temporal realms. Together, these studies have shown that DL techniques can characterize important information in MI-BCI signals from multiple dimensions of time, frequency, and space. However, MI-BCI data is usually characterized by a small sample size and high temporal resolution, which not only provides fine-grained temporal information for signal analysis, but also places higher demands on the generalization ability of the model. Using only convolution operations in the temporal and spatial domains may hinder the neural network from capturing long-term dependencies and make it difficult to process global information.

Attention mechanisms have emerged as a potent antidote, adeptly distilling long-range interactions in sequential data across domains [21]. Presently, the prevalent attention frameworks mainly include global attention [22], local attention, and self-attention mechanisms [23]. Among them, the transformer architecture based on the self-attention mechanism [24] has exhibited exceptional performance in computer vision tasks [25]. The self-attention mechanism is also gradually gaining attention in the field of brain signal decoding and is widely used in recent research [26,27]. For example, Cai et al. [28] combined a masking mechanism with a transformer structure based on the self-attention mechanism for MI-BCI signal recognition. Zhao et al. [29] combined an EEGNet-like [30] CNN module with a transformer encoder. More recently, Qin et al. [31] proposed an M-FANet model based on spectral and spatial features for MI-BCI classification using multiple attention modules. In addition, sparse-attention mechanisms are gradually being explored in EEG decoding. Zhou et al. [32] proposed CSBrain, which employs structured sparse attention to model dependencies across temporal windows and anatomical brain regions. Cai et al. [33] developed EEG-CSANet, in which centralized sparse attention and sparse cross-attention are used to facilitate interactions among multiscale EEG feature branches. However, these sparse-attention strategies mainly rely on predefined temporal or spatial structures or are designed for multibranch feature fusion. Their sparse patterns and computational processes are closely associated with specific tokenization schemes or network architectures. Dense self-attention still requires pairwise interactions between all queries and keys, while existing structured sparse approaches may introduce additional architectural

complexity or task-specific structural priors. In this case, further consideration needs to be given to the computational efficiency, feature redundancy, and generalization capability of attention-based EEG models.

In this study, we propose a novel lightweight deep learning framework, termed DESAnet, which integrates three complementary modules to balance decoding performance and computational efficiency. The framework's core components include a dual-branch spatial-temporal convolution module (DBSTCV), an efficient channel attention module (ECA), and a sparse attention module (SA). The DBSTCV is designed to capture multi-scale spatial and temporal features from raw EEG signals—spatial features reflect neural activity differences across scalp channels, while temporal features capture dynamic changes in neural patterns during imagined movements. The ECA module further refines feature representation by adaptively highlighting information from channels that are most relevant to the target MI task, suppressing noise from irrelevant channels. Meanwhile, the SA module, by introducing the ProbSparse attention mechanism, dynamically selects dominant queries to reduce redundant attention computation while preserving informative long-range temporal dependencies. Compared with several representative deep-learning-based methods, DESAnet achieves competitive classification performance while maintaining relatively low computational complexity. The principal results and innovations of this study are as follows:

- (1) DESAnet, an efficacious MI-EEG decoder fusing spatial-temporal convolutions, channel attention, and sparse multi-head attention for holistic feature synthesis.
- (2) We demonstrate a favorable balance between decoding accuracy and model complexity through the integration of prob-sparse attention and ECA, enabling efficient long-range modeling.
- (3) On BCI IV-2a and IV-2b, DESAnet delivers competitive performance with markedly reduced complexity, and visualization analyses (t-SNE, attention maps, and CAM) support the rationality and interpretability of the model.

2. Methods and materials

2.1. Datasets description

Two publicly available benchmark datasets provided by Graz University, namely the BCI competition IV-2a (BCI IV-2a) [34] and BCI competition IV-2b (BCI IV-2b) [35], were used as the primary datasets to evaluate the proposed DESAnet framework. In addition, an extra validation was conducted on the larger OpenBMI dataset [36] to further assess the generalization capability of the proposed method, and the corresponding experimental settings and results are provided in [Supplementary Material B](#).

- (1) **BCI IV-2a (four-class):** The four-class dataset records the brain electrical signals of nine subjects. The acquisition device uses 22 Ag/AgCl electrodes and a sampling frequency of 250 Hz. The dataset includes four MI tasks: left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). Each subject participated in two experimental sessions, each comprising six runs with 48 trials per run (12 trials per class), totaling 288 trials per session. In our experiment, we utilize the data within 4 s from the start of the MI cue to the end as the original input for the model. The first session served as the training samples, and the second session served as the testing samples.
- (2) **BCI IV-2b (two-class):** The two-class dataset records the EEG signals of nine subjects using three bipolar electrodes (C3, Cz, and C4) with a sampling rate of 250 Hz. The dataset contained two MI tasks: left hand (class 1) and right hand (class 2). Each subject underwent five sessions, and each subject completed 120 trials (60 trials per class), with each trial lasting 4 s. In our experiment,

we utilize the data within 4 s from the beginning to the end of the MI cue as the original input for the model. The first three sessions served as the training samples, and the last two sessions served as the testing samples.

2.2. Data augmentation

For DL models, classification performance heavily depends on the amount of training data. Many previous studies have applied data augmentation techniques to generate synthetic training samples from the raw data, improving the capability of network feature representations [37]. Following [38], we use a conventional data segmentation and recombination technique to expand the training sample size. First, we assume that the training sample set from the same category $c \in \{1, \dots, M\}$ is X^c ($X^c \in R^{N \times C \times T}$), with corresponding label $Y^c \in N$, where M denotes the category, N denotes the number of training samples of category c , C denotes the channel count, and T denotes the time sample points. Then, the data X^c are uniformly partitioned into s time segments and denoted as $X_{index}^{c[s]}$, with c indicating the category index, and $index = \{N_1, \dots, N_s\}$, $|index| = s$ representing the segment index. Next, time segments $X_{index}^{c[s]}$ from different training samples within the same category are randomly selected and combined in their original temporal sequence to generate new augmented data X'_m and labels Y'_m , denoted by:

$$X'_m = \bigcup_{index=N_1}^{N_s} X_{index}^{c[s]}, Y'_m = Y^c \quad (1)$$

In each training iteration, the model acquires new samples from the augmented dataset to maintain the temporal consistency and label correspondence, thus improving the diversity of the data. Eventually, all the generated sets of augmented data and labels are concatenated together to form the final augmented dataset:

$$\tilde{X} = \text{concat}(X'_1, \dots, X'_m), \tilde{Y} = \text{concat}(Y'_1, \dots, Y'_m). \quad (2)$$

In the data augmentation procedure, each EEG trial is uniformly divided into s non-overlapping temporal segments, where s controls the segmentation granularity and each segment contains $\lfloor T/s \rfloor$ time samples. For each class, temporal segments are randomly selected from different trials within the same class and concatenated in their original temporal order to generate augmented samples. The augmentation is applied only to the training set, while the test sets remain unchanged. The number of augmented samples is set to be equal to the number of original training samples divided by the number of classes, ensuring balanced augmentation across classes.

2.3. Data Preprocessing

In this work, the raw MI-BCI signals $X_{input} \in R^{C \times T}$ are used without conventional band-pass filtering or artifact removal procedures. Here, C denotes the channel count, and T denotes the sample points. The output generated by the model is $Y \in R^{N_c}$, where N_c denotes the categories. Before being fed into the proposed model, the input signals are standardized using z-score normalization to reduce differences in signal scale and improve training stability, as defined by:

$$X_{output} = \frac{X_{input} - \mu}{\sqrt{\sigma^2}} \quad (3)$$

where X_{input} and X_{output} denote the input original data and the normalized data, respectively, while μ and σ^2 represent the mean and standard deviation of the input data across channel dimension C .

2.4. DESAnet framework

As shown in Fig. 1, the DESAnet model primarily comprises three

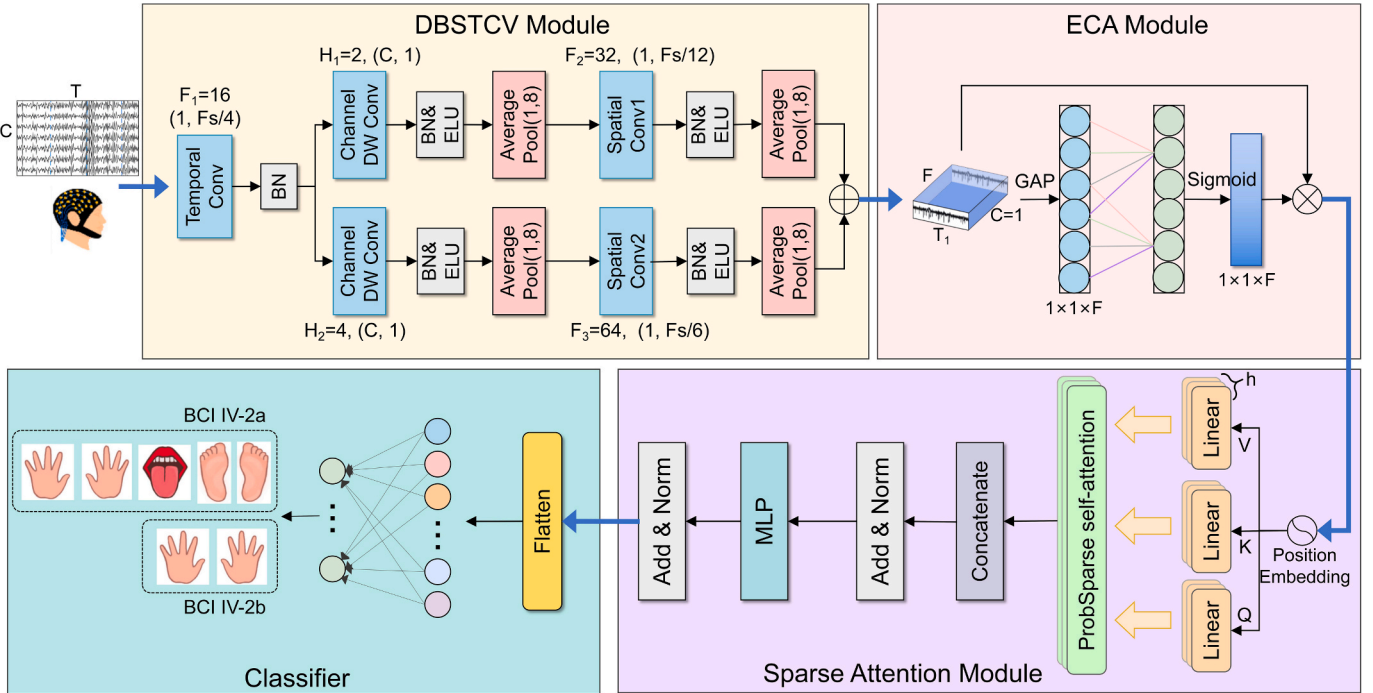


Fig. 1. The overall framework of the proposed DESAnet. In the DBSTCV module, the main parameters F_1 , F_2 , F_3 , and H_1 , H_2 represent the number of convolution kernels and the multiplicative relationship between the number of kernels in the current layer and the number of channels in the previous layer, respectively. The parameter F_s denotes the sampling rate of the input data. In the ECA module, the parameter F represents the number of channels of the input feature maps in that layer. In the Sparse Attention module, the parameter h refers to the number of attention heads, and “Add & Norm” indicates the operation of adding two inputs followed by the layer normalization. The symbol “ $\oplus$ ” between the DBSTCV module and the ECA module denotes element-wise addition, whereas the symbol “ $\otimes$ ” between the ECA module and the SA module denotes element-wise multiplication.

core modules: the DBSTCV module, the ECA module, and the SA module. The detailed structure and function of each module in DESAnet will be further described in the following sections.

2.4.1. Dual-branch spatial-temporal convolutional module

To effectively represent temporal and spatial information, a dual-branch spatial-temporal convolution (DBSTCV) module was designed. The module has two main stages (see Fig. 1), and the detailed parameter settings are shown in Table 1. The first stage performs a temporal convolution using M_1 filters of kernel size $(1, K_E)$, where K_E represents the filter length along the time axis. K_E was configured to a quarter of the sampling rate (62 for BCI IV-2a and IV-2b). This enables the filters to capture temporal information related to frequencies above 4 Hz. This operation produces M_1 temporal feature maps.

The second stage contains two parallel branches with the same structure but different temporal kernels to achieve multi-scale encoding. In the first layer of each branch, a depthwise convolution (DW) with kernel size $(C, 1)$ learns spatial filters per temporal feature map, where C denotes the number of EEG channels. The number of DW kernels is $H_b \times M_1$ for the branch $b \in \{1, 2\}$, producing $H_b M_1$ feature maps; we use $H_1 = H_2 = 2$ based on validation. The DW is followed by an average pooling layer with a kernel size of $(1, 8)$ and a stride of $(1, 8)$. This reduces the dimensionality of the feature maps and lowers the sampling rate of the signal to around 32 Hz. The second layer of the branch is a separable convolution with F filters of kernel size $(1, K_b)$ for the branch $b \in \{1, 2\}$. Different temporal kernel sizes are used in the two branches to characterize complementary higher-order spatial-temporal information. Because K_1 and K_2 operate on the downsampled latent feature maps rather than directly on the original EEG sequence, they are not intended to correspond to explicit physiological frequency bands in the same manner as K_E . Experimentally, we set $K_1 = 21$ and $K_2 = 48$ (For details, please see Table S1 in the supplementary materials section A), corresponding to nominal temporal spans of approximately 0.67 s and 1.54 s on the downsampled sequence, respectively, so that the two branches

aggregate relatively short-duration and long-duration temporal contexts. The two separable convolutions produce F_1 and F_2 feature maps, respectively. Then, we use a second average pooling layer with a kernel size of $(1, 8)$ and a stride of $(1, 8)$, further reducing the sampling rate to about 4 Hz. Both convolutional layers of the branches are accompanied by batch normalization (BN) [39] to improve training efficiency and exponential linear unit (ELU) activation [40] to introduce nonlinearity. Finally, the spatial-temporal sequences in the two branches are connected by concatenation.

The DBSTCV module ultimately generated a temporal sequence $X_i \in R^{F \times T_1}$, where T_1 time vectors of dimension $F = F_1 + F_2 = H_1 \times M_1 + H_2 \times M_1 = 64$ represent the temporal information. The duration of the temporal sequence is specified by

$$T_1 = \lfloor T/64 \rfloor \quad (4)$$

where $\lfloor \cdot \rfloor$ denotes the floor (rounding-down) operation, T denotes the time points in the original EEG data.

2.4.2. Efficient channel attention module

To emphasize informative channels without introducing heavy parameters, we adopt an ECA [41] module that reweights the feature maps produced by the previous module. The module employs adaptive convolution to adjust the kernel size based on the input feature maps dynamically, strengthening the model's effectiveness in learning meaningful representations. The adaptive kernel size is determined by the following function:

$$k = \varphi(F) = \left\lfloor \frac{\log_2(F)}{\gamma} + b \right\rfloor \quad (5)$$

where k represents the filter size, F denotes the number of the feature map channel, and $\gamma = 2$ and $b = 1$ are utilized to adjust the ratio between features and filters. After the adaptive one-dimensional convolution is complete, the sigmoid activation function is utilized to obtain the

Table 1
The detailed network parameters of the DBSTCV module.

Block	Layer	Filters	Kernel Size	Options	Outputs
Input	—	—	—	—	$(1, C, T)$
TemporalConv block	Conv2D	M_1	$(1, K_E)$	padding = same	(M_1, C, T)
	BN	—	—		(M_1, C, T)
Spatial-TemporaConv block1	Depthwise2D	$H_1 * M_1$	$(C, 1)$	padding = valid group = M_1 max_norm = 5.	$(H_1 * M_1, 1, T)$
	BN	—	—	—	$(H_1 * M_1, 1, T)$
	Activation	—	—	function = ELU	$(H_1 * M_1, 1, T)$
	AveragePooling2D	—	$(1, 8)$	—	$(H_1 * M_1, 1, T//8)$
	Dropout	—	—	dropRate = 0.1	$(H_1 * M_1, 1, T//8)$
	SeparableConv2D	F_1	$(1, K_1)$	padding = same	$(F_1, 1, T//8)$
	BN	—	—	—	$(F_1, 1, T//8)$
	Activation	—	—	function = ELU	$(F_1, 1, T//8)$
	AveragePooling2D	—	$(1, 8)$	—	$(F_1, 1, T//64)$
	Dropout	—	—	dropRate = 0.1	$(F_1, 1, T//64)$
	Dropout	—	—	—	$(F_1, 1, T//64)$
Spatial-TemporalConv block2	Depthwise2D	$H_2 * M_1$	$(C, 1)$	padding = valid group = M_1 max_norm = 5.	$(H_2 * M_1, 1, T)$
	BN	—	—	—	$(H_2 * M_1, 1, T)$
	Activation	—	—	function = ELU	$(H_2 * M_1, 1, T)$
	AveragePooling2D	—	$(1, 8)$	—	$(H_2 * M_1, 1, T//8)$
	Dropout	—	—	dropRate = 0.1	$(H_2 * M_1, 1, T//8)$
	SeparableConv2D	F_2	$(1, K_2)$	padding = same	$(F_2, 1, T//8)$
	BN	—	—	—	$(F_2, 1, T//8)$
	Activation	—	—	function = ELU	$(F_2, 1, T//8)$
	AveragePooling2D	—	$(1, 8)$	—	$(F_2, 1, T//64)$
	Dropout	—	—	dropRate = 0.1	$(F_2, 1, T//64)$
	Dropout	—	—	—	$(F_2, 1, T//64)$

Note: “2D” refers to the use of 2-dimensional operations, BN stands for batch normalization, ELU refers to the exponential linear unit activation, C denotes the number of EEG channels, and T denotes the number of time points, “//” denotes floor division, i.e., $\lfloor T/64 \rfloor$.

channel weights in each feature map as follows:

$$\omega = \text{sigmoid}(F_1 D_k(y)) \quad (6)$$

where y represents the result obtained by applying global average pooling to the input feature map of this layer, $F_1 D_k$ denotes a one-dimensional convolution of kernel size k . The weights ω generated by the ECA module are fused with the raw input feature map vectors to mitigate overfitting. Finally, the weighted outcome produces a feature vector of size (F, T_1) .

2.4.3. Sparse attention module

Self-attention can model long-range dependencies in EEG sequences, but dense self-attention requires pairwise interactions between all queries and keys, resulting in quadratic computational and memory complexity [24]. To overcome this limitation and reduce computational overhead, we adopt the ProbSparse attention mechanism proposed in Informer [42]. Unlike general sparse attention methods [43,44], ProbSparse attention evaluates the importance of query-key pairs and only selects dominant queries for attention computation. In this work, we further adapt this mechanism to EEG sequence modeling tasks, where temporal dependencies are long-range but sparse in nature. Therefore, the proposed Sparse Attention module is built upon ProbSparse attention and integrated into our network to enhance long-sequence modeling efficiency while maintaining discriminative feature extraction capability. The sparse attention (SA) module follows a transformer-style self-attention block, but replaces the dense attention with multi-probSparse attention. As shown in Fig. 2, each probSparse attention layer with the sparse attention module comprises three parts: query (Q), key (K), and value (V). The following are the realization details of the mechanism.

Given an encoded sequence X_e with position information, the query, key, and value sequence vectors are computed via a linear transformation as follows:

$$Q = W_Q LN(X_e) \in R^{T_Q \times d}, W_Q \in R^{d \times T_Q} \quad (7)$$

$$K = W_K LN(X_e) \in R^{T_K \times d}, W_K \in R^{d \times T_K} \quad (8)$$

$$V = W_V LN(X_e) \in R^{T_V \times d}, W_V \in R^{d \times T_V} \quad (9)$$

where LN denotes the layer normalization [45], W_Q , W_K , and W_V denote the weight matrix. The query, key, and value sequences are divided into h subsequences, with one subsequence from each query, key, and value forming an attention head. The process yields h attention heads, with $h = \text{heads} = 5$ empirically.

Given that the dot product of each query and key may not be entirely valid [42], which implies that certain elements could be significantly correlated with only a subset of others, it is essential to prioritize values that display unique or distinctive attributes to enhance computational efficiency. Consequently, the probSparse self-attention adopts a probabilistic perspective towards self-attention, with the experimental procedure outlined as follows.

Initially, we obtain K_{sample} by the probSparse self-attention samples K , and then calculate the M value of each K_{sample} to assess the sparsity of Q :

$$M(q_i, K) = \ln \sum_{j=1}^{L_K} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{q_i k_j^T}{\sqrt{d}} \quad (10)$$

where the first term represents the Log-Sum-Exp (LSE) of q_i across all K , and the second term corresponds to their arithmetic mean value.

Based on this, the sparse evaluation formula is obtained as the probSparse self-attention formula, namely:

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (11)$$

where Q' represents a sparse matrix of the identical dimensionality as q , which includes the highest-ranked u queries determined by the sparse evaluation $M(q, K)$, where a sampling factor determines the size of u . This design allows the probSparse self-attention to compute dot product operations for each query-key pair only once. Also, as proved in the

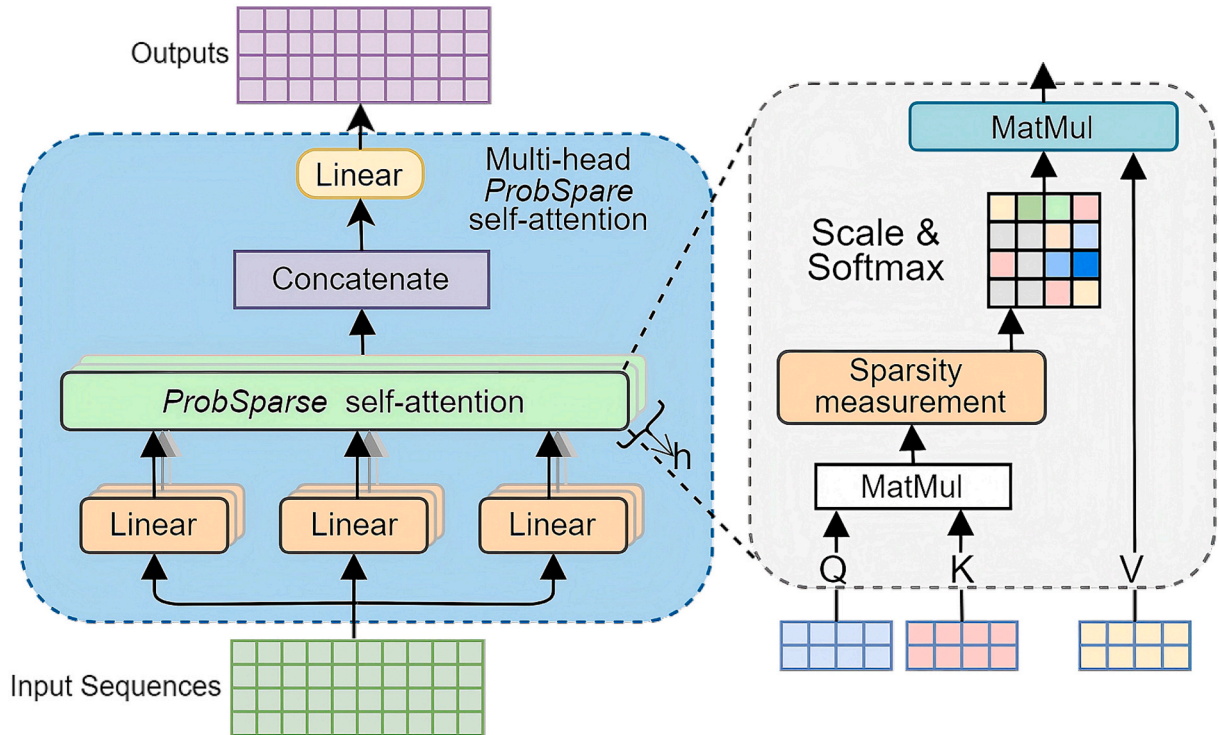


Fig. 2. Multi-head probSparse self-attention. The parameter h represents the number of heads in the multi-head probSparse self-attention. The parameters Q , K , and V represent the query, key, and value of the probSparse self-attention mechanism, respectively.

article [42], its sparse evaluation carries out the computation of upper bounds, ensuring that the computation's time and space complexity remains $O(L\log L)$.

Finally, the probsparse self-attention is computed by concatenating all attention heads vectors, followed by their integration with the input window X_e using residual concatenation, as shown below:

$$X'_e = \text{concat}(A_1, \dots, A_h) + X_e \in R^{T_1 \times d} \quad (12)$$

where T denotes the time points in the original EEG data.

2.5. Evaluation metrics

The proposed DESAnet model uses accuracy (Acc), kappa score ($Kappa$), and standard deviation (Std) as evaluation metrics. Accuracy is defined as:

$$Acc = \frac{A_{pre}}{N_{all}} \quad (13)$$

where N_{all} indicates the number of the entire sample, A_{pre} denotes the number of correctly classified samples.

The formula of the kappa score is as follows:

$$Kappa = \frac{r_o - r_e}{1 - r_e} \quad (14)$$

where r_o denotes the actual accuracy of the model, and r_e denotes the expected accuracy under random conditions.

The formula of the standard deviation is as follows:

$$Std = \sqrt{\frac{1}{M} \sum_{i=1}^M (Acc_i - \mu)^2} \quad (15)$$

where M is the subject number, Acc_i denotes the accuracy of subject i , and μ is the mean accuracy of all subjects.

3. Results

All experiments were conducted on an NVIDIA GeForce RTX 3070 GPU with 8 GB of memory, utilizing the PyTorch framework. In the experimental period, we trained the model using the adaptive moment estimation (Adam) optimizer and the categorical cross-entropy loss function, where the learning rate was set to 0.0006, and the batch size was set to 32. The training process lasts for 2500 epochs. If no improvement is observed after 30 consecutive iterations, we will terminate the training early to avoid overfitting. These hyperparameters were carefully chosen through extensive experimentation to optimize the models' generalization performance.

The proposed model is evaluated using a subject-dependent approach. Specifically, we use the same training and testing data as the original competition: for the BCI IV-2a dataset, the trials data in session 1 are used for training, and the trials data in session 2 are used for testing; for the BCI IV-2b dataset, trials data from the first three

sessions are used for training, while the last two sessions are used for testing.

3.1. Experimental results and comparisons

Tables 2 and 3 show the results of DESAnet compared with the representative deep-learning-based methods on the four-class and two-class datasets, respectively, including the EEGNet [30], EEGConformer [26], ATCNet [27], CTNet [29], BTGECNN [46], MAE-EEG-Transformer [28], and EISATC-Fusion [47]. To further evaluate whether the observed differences are statistically meaningful, subject-wise paired statistical tests were conducted. Paired-samples t-tests were used when the paired differences satisfied the normality assumption, while Wilcoxon signed-rank tests were applied otherwise.

On the BCI IV-2a four-class dataset, DESAnet achieved an average accuracy of 85.69% and a mean kappa value of 0.8092. Compared with EEGNet, EEGConformer, BTGECNN, CTNet, and ATCNet, DESAnet improved the average accuracy by 12.54%, 7.03%, 5.23%, 3.17%, and 0.31%, respectively. The statistical results show that DESAnet significantly outperformed EEGNet, EEGConformer, and BTGECNN, while the improvements over ATCNet and CTNet were not statistically significant. On the BCI IV-2b two-class dataset, DESAnet achieved an average accuracy of 89.40% and a mean kappa value of 0.7879, improving the accuracy over EEGNet, EEGConformer, MAE-EEG-Transformer, CTNet, and EISATC-Fusion by 5.52%, 4.77%, 8.03%, 0.91%, and 1.82%, respectively. Significant improvements were observed over EEGNet, MAE-EEG-Transformer, and EISATC-Fusion, whereas the differences compared with EEGConformer and CTNet were not statistically significant. To further evaluate the generalization capability of DESAnet on a larger subject population, we additionally validated the proposed method on the OpenBMI dataset (54 subjects). The results showed that DESAnet achieved an average accuracy of 72.91% and a kappa value of 0.458, outperforming all compared baseline methods (please see Table S2 in the supplementary materials section B). These results indicate that DESAnet achieves competitive performance among the compared representative methods.

In addition, DESAnet showed lower standard deviations on both datasets, with 8.27% on BCI IV-2a and 7.88% on BCI IV-2b, suggesting more stable performance across subjects. Nevertheless, slight variations can still be observed for several subjects in the BCI IV-2a dataset, such as A01–A05 and A08–A09, which may be attributed to inter-subject EEG variability. To further analyze class-wise performance, confusion matrices were added for both datasets, as shown in Figs. 3 and 4. The strong diagonal dominance in these matrices indicates that DESAnet can effectively distinguish different MI classes and provides a more detailed view of class-wise classification performance.

3.2. Resource consumption

To assess computational efficiency, we report multiply-accumulate operations (MACs) per forward pass in addition to accuracy, trainable parameters count, and model size. All MACs were recalculated using a batch size of 1 under identical input dimensions and the same profiling

Table 2

Subject-dependent four-class classification performance (accuracy (%) and kappa) comparison between the proposed DESAnet model and representative deep-learning-based methods on the BCI IV-2a dataset.

Models	A01	A02	A03	A04	A05	A06	A07	A08	A09	Average	Std	Kappa
EEGNet	84.72	57.64	89.58	61.11	70.49	57.99	81.25	80.90	74.65	73.15**	12.01	0.6420
EEGConformer	88.19	61.46	93.40	78.13	52.08	65.28	92.36	88.19	88.89	78.66*	15.30	0.7155
ATCNet	88.47	70.49	97.56	80.97	82.97	73.56	93.10	90.28	91.00	85.38	9.07	0.8050
CTNet	90.97	73.61	96.53	84.72	77.08	64.24	86.11	84.38	85.07	82.52	9.61	0.7670
BTGECNN	87.90	67.49	93.41	71.49	83.70	60.93	90.61	83.76	84.85	80.46*	11.16	0.7410
DESAnet	89.24	72.92	96.88	83.33	81.25	75.69	96.18	86.81	88.89	85.69	8.27	0.8092

Note: The Bold values indicate the best results among the compared methods. Statistical significance between DESAnet and each compared method is marked by * $p < 0.05$ and ** $p < 0.01$, using paired-samples t-tests or Wilcoxon signed-rank tests according to the normality of paired differences.

Table 3

Subject-dependent two-class classification performance (accuracy (%) and kappa) comparison between the proposed DESAnet model and representative deep-learning-based methods on the BCI IV-2b dataset.

Models	B01	B02	B03	B04	B05	B06	B07	B08	B09	Average	Std	Kappa
EEGNet	69.56	68.79	86.00	96.13	89.50	77.81	88.13	92.88	86.13	83.88**	9.75	0.6776
EEGConformer	82.50	65.71	63.75	98.44	86.56	90.31	87.81	94.38	92.19	84.63	12.18	0.6926
MAE-EEG-Transformer	81.87	61.43	59.69	96.25	88.44	85.94	80.00	90.00	88.75	81.37*	12.70	0.6275
CTNet	78.75	71.07	84.38	97.19	97.81	87.81	94.06	94.69	90.63	88.49	9.03	0.7697
EISATC-Fusion	75.00	72.86	86.56	96.88	97.81	84.38	94.06	93.75	86.88	87.58*	9.07	0.7515
DESAnet	83.44	73.93	88.13	98.44	97.19	87.50	95.00	95.00	85.94	89.40	7.88	0.7879

Note: The Bold values indicate the best results among the compared methods. Statistical significance between DESAnet and each compared method is marked by * $p < 0.05$ and ** $p < 0.01$, using paired-samples t-tests or Wilcoxon signed-rank tests according to the normality of paired differences.

**Fig. 3.** Confusion matrix for the classification results of the BCI IV-2a dataset.

procedure. Table 4 summarizes the performance and computational complexity of DESAnet and several representative baseline models (EEGNet, EEGConformer, ATCNet, CTNet, and EISTAC-Fusion) on the BCI IV-2a and IV-2b datasets. On BCI IV-2a, DESAnet achieved an accuracy of 85.69% with 14,799 trainable parameters, a model size of 9.8 KB, and 2.8 M MACs. Its computational cost was approximately 22.8, 14.2, and 12.8 times lower than those of EEGConformer, ATCNet, and

CTNet, respectively. On BCI IV-2b, DESAnet achieved an accuracy of 89.40% with 13,056 trainable parameters, a model size of 9.8 KB, and 1.6 M MACs, requiring approximately 9.8, 5.8, and 8.1 times fewer MACs than EEGConformer, CTNet, and EISATC-Fusion, respectively. Although EEGNet contains fewer parameters, DESAnet achieved substantially higher classification accuracy while requiring fewer MACs under the revised single-sample profiling setting. These results

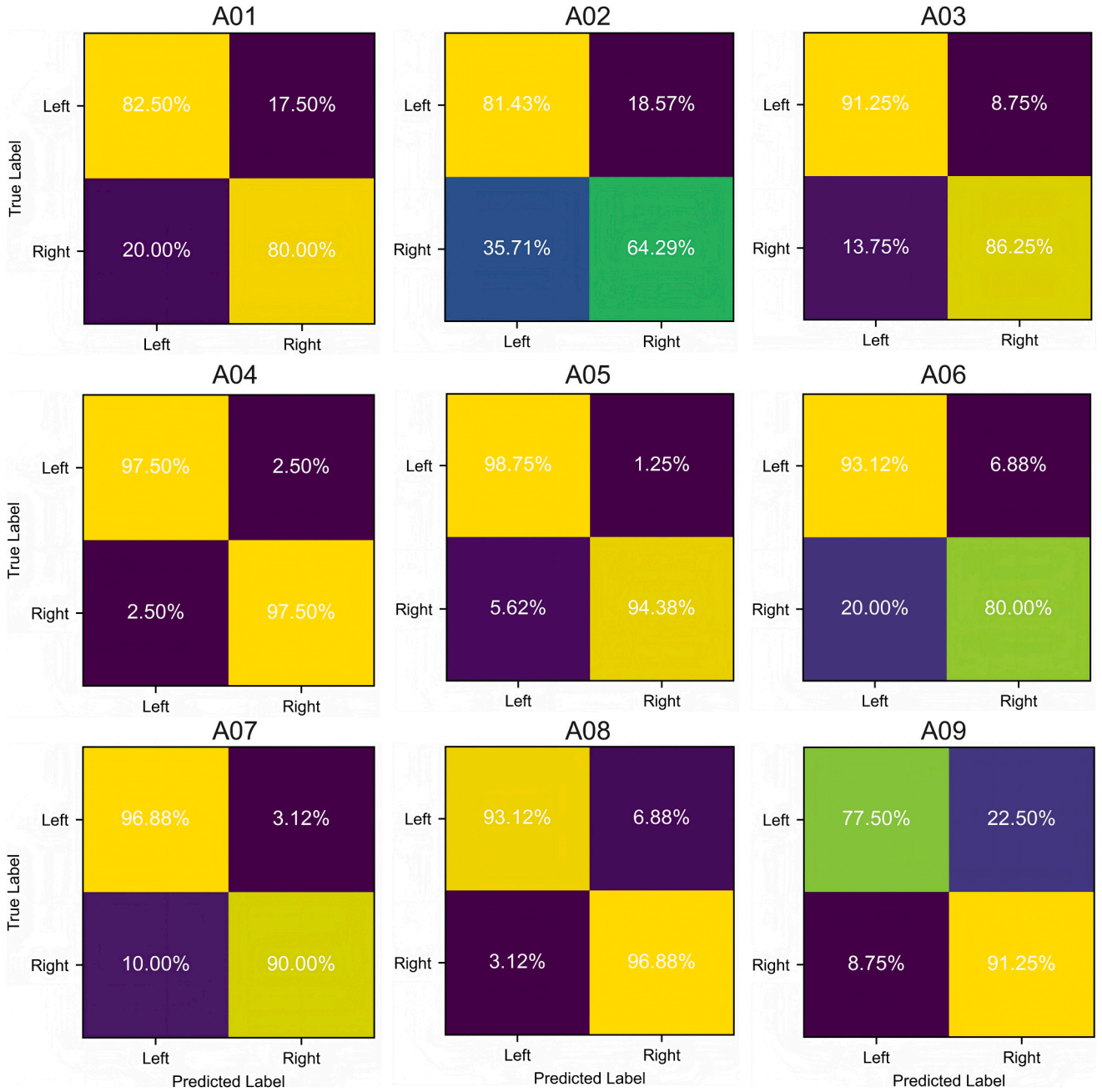


Fig. 4. Confusion matrix for the classification results of the BCI IV-2b dataset.

demonstrate that DESAnet provides competitive classification performance with a favorable balance between decoding accuracy and computational efficiency.

3.3. Ablation study

Table 5 presents the ablation results of the individual components of DESAnet on the BCI IV-2a and BCI IV-2b datasets. To ensure a fair comparison, all module-level ablation experiments were conducted under the same data augmentation strategy. In addition, to evaluate the contribution of data augmentation itself, we further performed an independent ablation experiment by disabling data augmentation and training the model using only the original training samples. Prior to statistical testing, the normality of the paired differences was examined. When the normality assumption was satisfied, paired-samples *t*-tests were used; otherwise, Wilcoxon signed-rank tests were applied. The

results show that data augmentation has a significant positive effect on decoding performance. Removing data augmentation decreased the average accuracy by 8.03% on BCI IV-2a and 8.02% on BCI IV-2b. For BCI IV-2a, the paired differences did not satisfy the normality assumption; therefore, the Wilcoxon signed-rank test was applied, showing that removing data augmentation significantly reduced both accuracy ($W = 0.0, p = 0.0039$) and kappa ($t(8) = 5.725, p = 0.0004$). For BCI IV-2b, the paired-samples *t*-test demonstrated a significant accuracy degradation ($t(8) = 3.372, p = 0.0098$), while the kappa difference was not statistically significant ($t(8) = 1.263, p = 0.2421$). These results indicate that data augmentation effectively improves the generalization capability of DESAnet under limited MI-EEG training samples.

The module-level ablation results further demonstrate the contributions of ECA and SA. Removing the ECA module decreased the average accuracy by 2.09% on BCI IV-2a and 2.95% on BCI IV-2b, while reducing the kappa values by 2.78% and 5.04%, respectively. Removing

Table 4

Evaluation of the efficacy of the DESAnet model on BCI IV-2a and BCI IV-2b datasets.

Datasets	Models	Parameters	Model size (KB)	MACs (M)	Accuracy (%)
BCI IV-2a	EEGNet	3444	3.4	12.5	73.15
	EEGConformer	789,816	789.6	63.9	78.66
	ATCNet	115,172	106.0	39.7	85.38
	CTNet	152,684	148.7	35.9	82.52
	Ours (DESAnet)	14,799	9.8	2.8	85.69
BCI IV-2b	EEGNet	2146	2.1	1.9	83.88
	EEGConformer	759,416	759.2	15.6	84.63
	CTNet	150,722	146.7	9.3	88.49
	EISATC-Fusion	24,674	24.1	13.0	87.58
	Ours (DESAnet)	13,056	9.8	1.6	89.40

Table 5

Contribution of each module in the DESAnet model to the performance of MI classification using the BCI IV-2a and BCI IV-2b datasets.

Remove block	BCI IV-2a		BCI IV-2b	
	ACC%	Kappa	ACC%	Kappa
None (DESAnet)	85.69	80.92	89.40	78.79
w/o Data augmentation	77.66**	70.22**	81.38**	75.38
ECA	83.60**	78.14**	86.45*	73.75*
SA	83.80*	78.39	86.79*	74.45*
SA + ECA	80.14**	73.33**	83.75**	71.14**

Note: The Bold values indicate the best results among the compared methods. Statistical significance is marked by * $p < 0.05$ and ** $p < 0.01$, using paired-samples t-tests or Wilcoxon signed-rank tests according to the normality of paired differences.

the SA module resulted in accuracy decreases of 1.89% and 2.61% on BCI IV-2a and BCI IV-2b, respectively. When both ECA and SA were removed, the average accuracy decreased by 5.55% on BCI IV-2a and 5.65% on BCI IV-2b. The joint removal of ECA and SA caused statistically significant degradation in both accuracy and kappa on BCI IV-2a ($t(8) = 5.507, p = 0.0006$; $t(8) = 4.961, p = 0.0011$) and BCI IV-2b ($t(8) = 4.624, p = 0.0017$; $W = 0.0, p = 0.0039$), respectively. These results indicate that ECA and SA complement the DBSTCV backbone and jointly contribute to improved decoding performance.

Table 6 further evaluates the effects of the DBSTCV module and the SA module on classification performance and model complexity under the same data augmentation setting. Compared with EEGNet-16,2, replacing the module with DBSTCV improved the average accuracy by 6.79% and the kappa score by 9.05%, while increasing the module parameters from 4452 to 10,820. This result indicates that the DBSTCV module provides stronger spatial-temporal representation capability through its dual-branch design. In addition, compared with the Transformer block, the SA module improved the average accuracy by 2.05% and the kappa score by 2.94%, while reducing the number of module

Table 6

Impact of the DBSTCV and SA modules on model alignment performance and complexity on the BCI IV-2a dataset.

Method	Average	Kappa	Block Parameters
EEGNet-16,2	76.43**	68.57**	4452
DBSTCV	83.22	77.62	10,820
EEG + Transformer	83.64	77.98	5085
EEG + SA	85.69	80.92	3975

Note: The Bold values indicate the best results among the compared methods. Statistical significance is marked by * $p < 0.05$ and ** $p < 0.01$, using paired-samples t-tests or Wilcoxon signed-rank tests according to the normality of paired differences.

parameters from 5085 to 3975. These results suggest that SA offers a more parameter-efficient strategy for modeling long-range temporal dependencies and improves the accuracy-efficiency trade-off of DESAnet.

3.4. Parameter sensitivity

We assessed how key hyperparameters affect DESAnet by fine-tuning the DBSTCV and SA modules under a subject-dependent cross-session validation protocol. For DBSTCV, the number of output channels in the DW convolution layers was varied while fixing the input channels at $M_1 = 16$. Specifically, we compared ($H_1, H_2 \in (2, 4)$), as shown in Fig. 5. The experimental results show that when increasing H_1 and H_2 from 2 to 4, the four-class classification accuracy decreases from 85.92% to 83.10%, while the two-class classification accuracy declines from 89.40% to 88.47%, with both reductions being less than 3%. However, the model parameters exhibit a doubling trend, increasing sharply from 14 k to 28 k for the four-class task and from 11 k to 22 k for the two-class task. Balancing accuracy and model compactness, we therefore adopt $H_1=H_2 = 2$ as the default configuration.

To the parameter settings of the SA module, as shown in Fig. 6, we use a controlled variable experiment strategy conducted on the two datasets. The head amount for the SA module was empirically set to 5, and then the impact of different factor parameters on experimental performance was tested, with the factor value ranging from 1 to 8. Different factor values affect the selection of queries, and not all queries exhibit high correlations with their corresponding keys. Overall, the BCI IV-2a dataset maintains performance between approximately 83% and 85% across all factor parameters, while the BCI IV-2b results remained between 88% and 90%, with minimal fluctuations. The experimental results show that when the factor is set to 5, both datasets achieve favorable performance.

3.5. Feature visualization

In this work, we employ three complementary interpretability techniques to analyze the representations learned by the DESAnet model: the t-SNE feature visualization [48], attention score [49], and class activation mapping (CAM) within the SA module.

Fig. 7 presents the t-SNE feature distributions of subjects A01, A03, and A07 on the four-class BCI IV-2a dataset. These subjects were selected according to the subject-wise classification results of DESAnet and correspond to the three highest-performing subjects, with accuracies of 89.24%, 96.88%, and 96.18%, respectively. As shown in Fig. 7 (a), the raw feature distributions are relatively dispersed, with substantial overlap and indistinct boundaries among the four classes. After processing by the DBSTCV and ECA modules, as shown in Fig. 7(b) and (c), the feature distributions become progressively more structured, and the clustering characteristics are enhanced through spatial-temporal feature extraction and channel-wise feature recalibration, although the degree of class separation varies across subjects. Fig. 7(d) shows that the ProbSparse attention mechanism in the SA module further captures long-range temporal dependencies, resulting in improved within-class compactness and clearer between-class separation. The consistent stage-wise refinement observed across the three subjects suggests that DESAnet progressively enhances the discriminability of MI-EEG representations despite inter-subject variability.

Fig. 8 shows the raw attention matrix distributions of different head numbers under different factor parameters for subject 3 in the four-class task. The green region corresponds to high attention weight values, characterizing critical feature regions; the red region corresponds to low weight values, indicating suppressed non-critical regions. The heatmap in Fig. 8 illustrates the impact of the factor parameter on the sparsity of attention and the feature-capturing capability of the SA module. When factor = 3 (the first row of Fig. 8), the attentional distribution is too sparse, and the weights of most regions are close to zero, which leads to

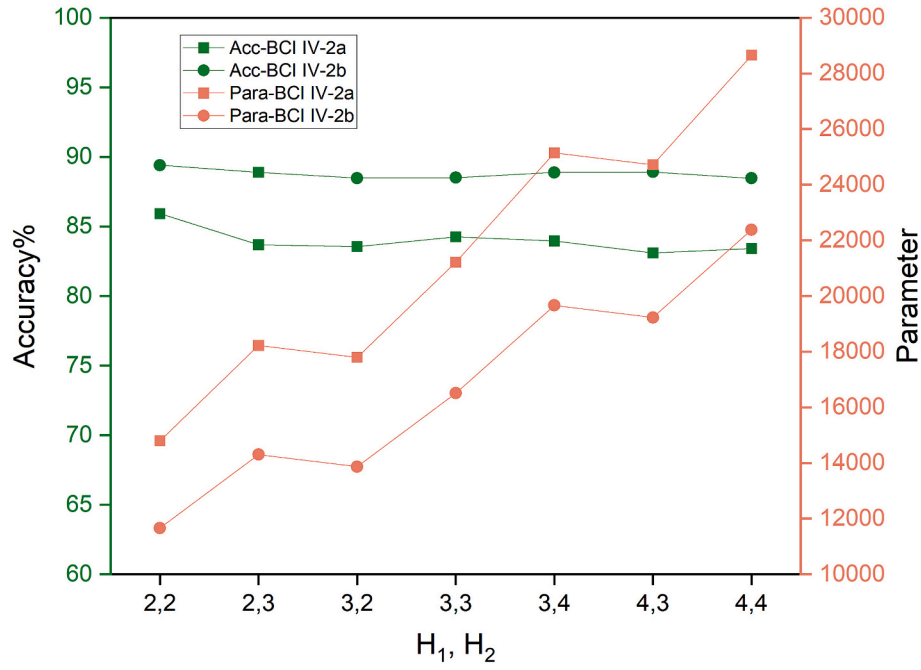


Fig. 5. The impact of the number of output channels in the DW convolutional layer of the DBSTCV module on the performance and parameter count for the BCI IV-2a and BCI IV-2b datasets.

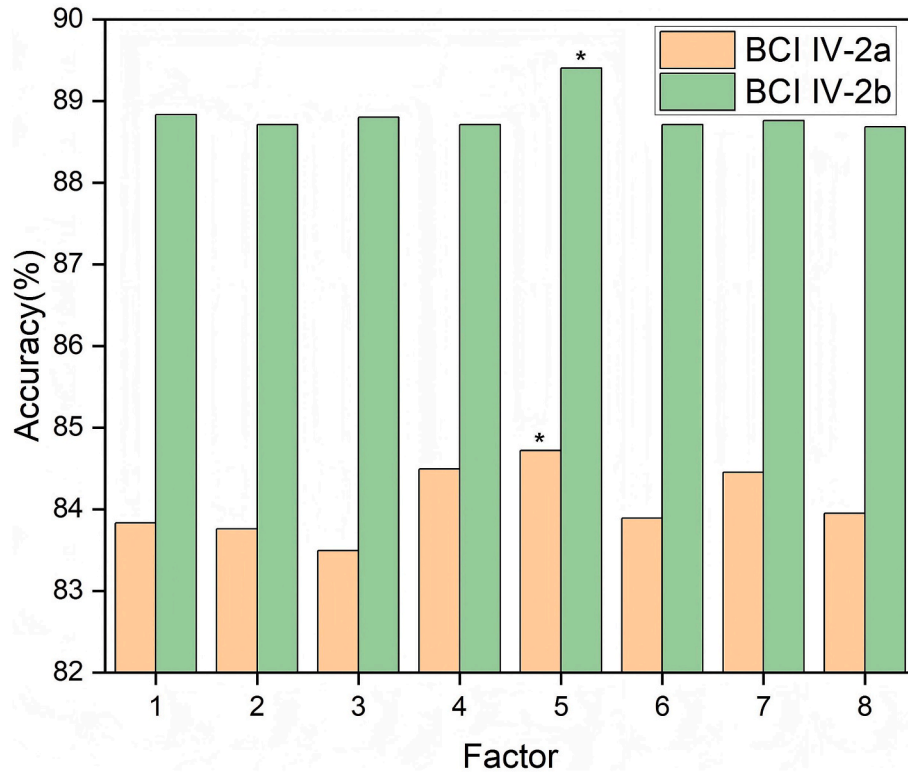


Fig. 6. Average classification accuracy of various parameter factors on the BCI IV-2a and BCI IV-2b datasets. The vertical axis denotes the average accuracy across all subjects, the horizontal axis represents the parameter factor, with values ranging from 1 to 8, and * indicates the best result with different parameters for each dataset.

incomplete global dependency capture; when factor = 7 (the third row of Fig. 8), the attentional distribution tends to be dense, and almost all regions receive high weights, which may introduce the interference of noise; and when factor = 5 (the second row of Fig. 8), the attentional distribution presents a clear and balanced feature.

To further examine the spatial interpretability and neurophysiological relevance of DESAnet, Fig. 9 presents the subject-averaged raw scalp topographies and the corresponding CAM-based topographic maps for the four MI classes on the BCI IV-2a dataset: left hand, right hand, feet, and tongue. Compared with the raw topographies, the CAM maps

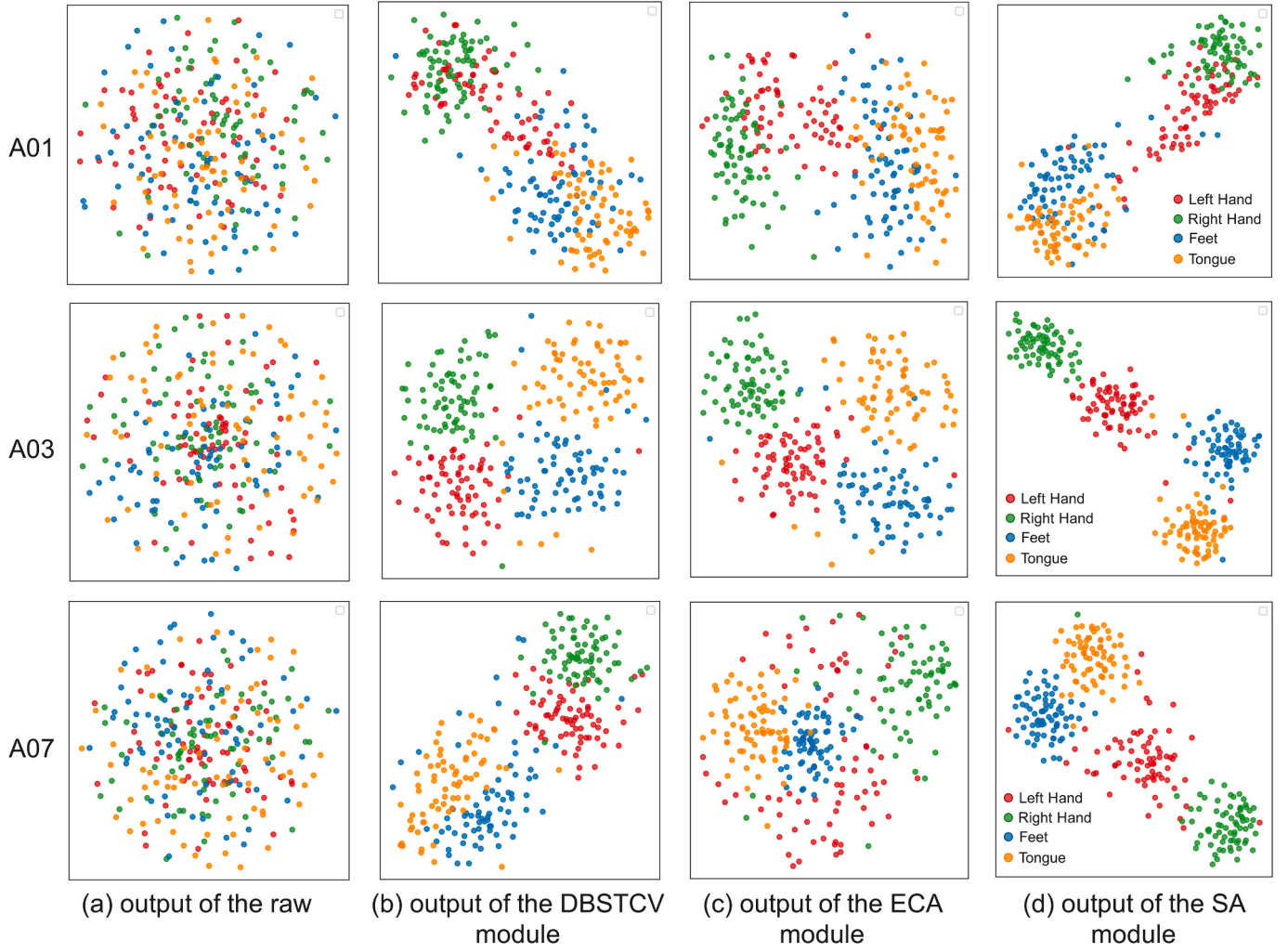


Fig. 7. The t-SNE visualization for subjects A01, A03, and A07 on the BCI IV-2a dataset. (a) Output of the raw. (b) Output of the DBSTCV module. (c) Output of the ECA module. (d) Output of the SA module.

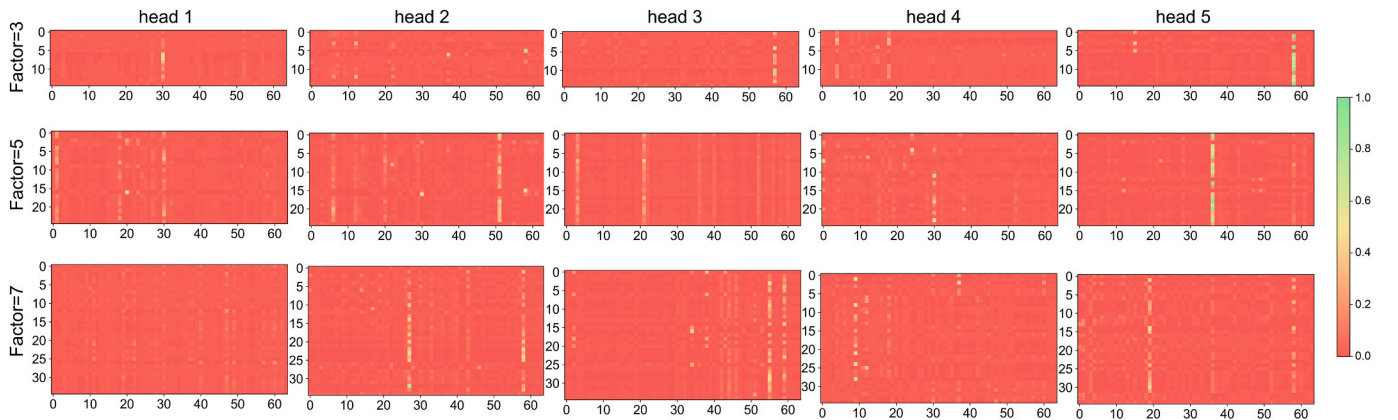


Fig. 8. Visualization of attentional score maps from 5 heads of different parameter factor values in the SA module for subject 3 on the BCI IV-2a dataset.

exhibit more localized and class-dependent spatial contribution patterns. The left-hand and right-hand conditions show distinguishable lateralized distributions over central and lateral scalp regions, suggesting that DESAnet preserves class-dependent hemispheric information associated with unilateral hand imagery. The feet condition exhibits a relatively pronounced midline-central contribution, which is broadly compatible with the medial organization of lower-limb sensorimotor

representations. In contrast, the tongue condition shows a more distributed bilateral pattern, potentially reflecting the broader spatial organization of orofacial sensorimotor processing.

4. Discussion

This study proposed a lightweight yet high-performance framework

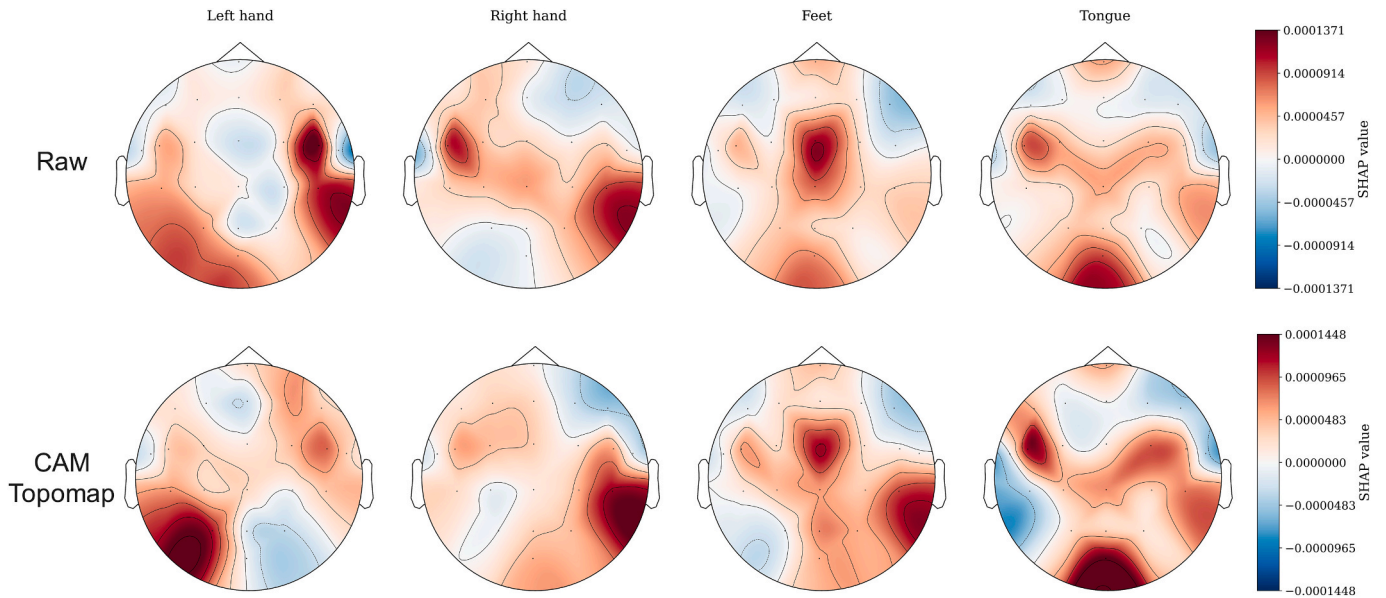


Fig. 9. Subject-averaged raw scalp topographies and CAM-based topographic maps for the four MI classes on the BCI IV-2a dataset. The first row shows the raw topographic distributions, and the second row shows the corresponding CAM topomaps learned by DESAnet. Warmer colors indicate greater positive contributions to the classification decision, while cooler colors indicate lower or negative contributions.

for MI-BCI classification, termed DESAnet. The framework integrates a DBSTCV module to extract local spatial–temporal features from raw EEG signals, an ECA module to adaptively emphasize discriminative channels, and an SA module to efficiently capture long-range temporal dependencies. The overall architecture is designed in a task-driven manner for MI-EEG decoding under limited training data and strong inter-subject variability, where each component addresses a complementary aspect of representation learning, forming a progressive pipeline from local feature extraction to global dependency modeling.

Compared with existing models such as EEGNet, ATCNet, EEGConformer, and CTNet, the essential difference of DESAnet lies in its hierarchical and efficiency-oriented design strategy. EEGNet mainly focuses on compact convolutional feature extraction without explicit modeling of long-range temporal dependencies. ATCNet introduces attention-enhanced temporal modeling but still relies on relatively dense feature processing. EEGConformer and CTNet integrate transformer-style attention mechanisms; however, they typically incur higher computational costs due to dense self-attention or hybrid architectures. In contrast, DESAnet first compresses and refines EEG representations through dual-branch spatial–temporal convolution, then applies lightweight channel recalibration via ECA, and finally performs sparse attention on compact temporal sequences. This design enables efficient global modeling while avoiding unnecessary computation on redundant raw representations. Experimental results on the BCI IV-2a and BCI IV-2b datasets further support the effectiveness of this design. DESAnet achieves competitive classification performance while maintaining favorable resource efficiency. These results, supported by ablation studies, parameter sensitivity analyses, and feature visualizations, highlight DESAnet's potential to facilitate practical MI-BCI deployment in resource-constrained environments.

The enhanced classification performance of DESAnet can be attributed to its synergistic modular design, which effectively captures the inherent spatial–temporal dynamics of EEG signals in MI tasks. EEG data from MI paradigms exhibit non-stationary characteristics, with discriminative features distributed across spatial channels (reflecting motor cortex topography) and temporal sequences (encoding event-related desynchronization/synchronization patterns) [50]. The DBSTCV module, through its parallel branches, enables concurrent processing of local spatial filters and dilated temporal convolutions, allowing for multi-scale representation learning without excessive

parameter proliferation. This is evidenced by the ablation results, where substituting a standard EEGNet backbone with DBSTCV yielded a 6.79% accuracy gain on the four-class task, albeit—a trade-off justified by the module's role in refining multi-band feature extraction. Such improvements align with prior observations that parallel convolutional architectures can mitigate information loss in hierarchical EEG processing [19,51], yet DESAnet's implementation ensures that this enhancement does not compromise overall compactness.

Complementing DBSTCV, the ECA and SA modules introduce lightweight attention mechanisms that refine feature salience without the computational burden of full self-attention. The ECA module, by applying adaptive channel-wise recalibration via a 1D convolution with kernel size $k = 5$, emphasizes informative EEG channels (e.g., those over contralateral motor areas) while suppressing noise-dominated ones. This design was supported by ablation results, in which removing ECA reduced accuracy by 2.09% and 2.95% on the four- and two-class tasks, respectively (see Table 5), and by further comparisons with no channel attention, and other common attention blocks (for example, SE [52] and CBAM [53]), where ECA achieved the highest accuracy on both BCI IV-2a (85.69%) and BCI IV-2b (89.40%), significantly outperformed all alternatives ($p < 0.05$), and maintained the lowest computational overhead among the attention-based variants (Please see Table S3 in the Supplementary Material Section C). Similarly, the SA module, leveraging a probabilistic sparse attention scheme with a factor of 5, efficiently models global temporal correlations by focusing queries on highly relevant keys, reducing quadratic complexity to near-linear. Its ablation resulted in 1.89% and 2.61% drops, respectively, and comparative analysis against a Transformer baseline showed equivalent efficiency with 2.05% higher accuracy and 22% fewer parameters. This efficiency stems from the SA's sparsity-inducing mechanism, which prunes irrelevant temporal dependencies, a strategy resonant with recent advances in sparse transformers for sequential data [42]. Collectively, these modules elevate baseline DBSTCV performance by 5.55% in accuracy (cumulative from ECA and SA additions), demonstrating their incremental yet indispensable roles in a holistic pipeline.

In comparison to the representative deep-learning-based methods, DESAnet's results position it as a competitive yet efficient alternative for MI-EEG decoding. On BCI IV-2a, DESAnet's 85.69% accuracy and 80.92% kappa surpass EEGNet (73.15% accuracy [30]), EEGConformer (78.66% [26]), ATCNet (85.38% [27]), CTNet (82.52% [29]),

BTGECNN (81.52% [46]), and EISATC-Fusion (80.46% [47]), with notably lower standard deviations (8.27% vs. 9–16% in many baselines). For BCI IV-2b, the 89.40% accuracy and 78.79% kappa similarly outperform EEGConformer (84.63%), CTNet (88.49%), and others. Additionally, the additional evaluation on the larger OpenBMI dataset further supports the robustness and scalability of DESAnet across a larger and more diverse subject population. (More details see [supplementary materials section B](#)). These gains are particularly salient in the context of resource metrics: DESAnet's 2.8 M MACs and 9.8 KB size represent over 20-fold reduction in operations compared to EEGConformer (63.9 M MACs) and a 10-fold parameter saving relative to ATCNet (106.0 KB), enabling deployment on edge devices like wearables. Unlike transformer-heavy models (e.g., EEGConformer or MAE-EEG-Transformer), which excel in capturing long-range dependencies but at prohibitive costs [42], DESAnet balances expressiveness and frugality through hybrid convolutions and sparse attention. This efficiency edge is crucial for real-world MI-BCIs, where battery life and latency constrain model complexity [54]. To further assess cross-subject generalization, leave-one-subject-out experiments were conducted (Please see [Supplementary Material section D](#)), in which DESAnet achieved accuracies of 71.79% (Kappa = 0.624) and 80.47% (Kappa = 0.609) on BCI IV-2a and BCI IV-2b, respectively, outperforming the compared baseline methods. Nevertheless, the performance decline relative to the subject-dependent setting indicates that inter-subject variability [55] remains a major challenge for MI-EEG decoding.

Parameter sensitivity analyses further validate DESAnet's design robustness, revealing optimal configurations that prioritize compactness over marginal gains. For DBSTCV, increasing output channels from ($H_1 = H_2 = 2$) to (4) doubled parameters (14 k–28 k for four-class) but yielded only <3% accuracy erosion (85.92%–83.10%), suggesting diminishing returns beyond minimal channel expansions—a finding consistent with DW convolutions' efficiency in EEGNet variants [56]. For SA, varying the sparsity factor from 1 to 8 maintained stable performance (83–85% on IV-2a, 88–90% on IV-2b), with factor = 5 achieving peak balance by avoiding over-sparsity (factor = 3, incomplete dependencies) or over-density (factor = 7, noise amplification). These insights, derived from controlled cross-session validations, inform hyperparameter tuning guidelines for similar architectures, emphasizing empirical trade-offs in low-data regimes typical of EEG [57].

Feature visualizations provide complementary evidence for the interpretability and neurophysiological plausibility of DESAnet [58]. The progressive clustering observed in [Fig. 7](#) indicates that the DBSTCV, ECA, and SA modules successively improve within-class compactness and between-class separability. The attention matrices in [Fig. 8](#) further show that factor = 5 achieves a balanced temporal weighting pattern, avoiding the incomplete dependency coverage observed at factor = 3 and the diffuse, potentially noise-sensitive weighting observed at factor = 7. The subject-averaged CAM topographies in [Fig. 9](#) provide an additional spatial perspective on the learned representations. The distinguishable lateralized distributions for left- and right-hand imagery suggest that the model retains hemispheric information relevant to unilateral motor imagery. The stronger midline-central contribution observed for feet imagery is broadly compatible with the medial representation of lower-limb movements, whereas the more distributed bilateral pattern observed for tongue imagery may reflect the broader organization of orofacial sensorimotor processing. Although DESAnet does not explicitly perform frequency-domain modeling, these spatial attribution patterns are compatible with the established organization of motor-imagery-related sensorimotor activity [26,50,58]. Thus, the CAM results provide qualitative neurophysiological support for the class-discriminative spatial representations learned by DESAnet.

Despite these strengths, several limitations remain. First, although the leave-one-subject-out results provide initial evidence of subject-independent generalization, the lower accuracy than in the subject-dependent setting indicates that inter-subject variability remains unresolved. In addition, the experiments were conducted offline, and cross-

dataset transfer and real-time robustness were not evaluated. Future work will therefore focus on domain adaptation and online validation [59]. Second, although the t-SNE representations, attention matrices, and CAM topographies improve interpretability by revealing feature separability, temporal weighting, and class-dependent spatial relevance, these analyses remain qualitative. In particular, scalp-level CAM attribution does not establish a direct correspondence between the learned features and their underlying cortical generators [60]. Future studies will combine temporal attribution analysis, frequency-specific interpretation, EEG source localization, and multimodal neuroimaging evidence to more rigorously relate the representations learned by DESAnet to established neural signatures of motor imagery, improving neurophysiological interpretability and clinical confidence [61].

5. Conclusion

In this study, we propose a novel DL architecture called DESAnet, which is designed for MI-BCI classification. The architecture integrates the DBSTCV module, ECA module, and SA module, and introduces a prospars attention mechanism to capture and fuse multi-scale temporal and spatial features efficiently. Detailed experiments show that our proposed DESAnet model performs well on both four-class and two-class classification datasets. T-SNE visualization and sparse attention heatmaps further confirm that the model improves intra-cluster compactness and inter-cluster separability, enhancing the interpretability and applicability of the model. Overall, DESAnet achieves an optimal balance between accuracy, robustness, and computational efficiency. While further validation across larger and more diverse datasets is warranted, our findings suggest that the proposed framework offers a promising direction for advancing practical MI-BCI applications.

Data and code availability statement

The code supporting the results of this study is available at <https://github.com/maobin007/EEG-DESAnet>, and the data is openly available at <https://www.bbc.de/competition/iv/#datasets>.

CRediT authorship contribution statement

Bin Mao: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation. **Fei Wang:** Supervision, Methodology. **Yuanhang Zhou:** Visualization, Investigation. **Yufeng Deng:** Investigation. **Jiahao Cheng:** Visualization. **Sheng Guo:** Software. **Dezhong Yao:** Methodology. **Peng Xu:** Conceptualization. **X. San Liang:** Writing – review & editing. **Tao Zhang:** Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (Grant SML2023SP203).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bspc.2026.111458>.

Data availability

I. The code supporting the results of this study is available at <https://github.com/maobin007/EEG-DESAnet>, and the data is openly available at <https://www.bbci.de/competition/iv/#datasets>.

II. The data is openly available at <https://www.bbci.de/competition/iv/#datasets>. (Reference data) (The code supporting the results of this study is available at <https://github.com/maobin007/EEG-DESAnet>.)

References

- [1] D. Jaipriya, K. Sriharipriya, Brain computer interface-based signal processing techniques for feature extraction and classification of motor imagery using EEG: a literature review, *Biomed. Mater. Devices* 2 (2024) 601–613, <https://doi.org/10.1007/s44174-023-00082-z>.
- [2] S. Kumar, A. Sharma, Advances in non-invasive EEG-based brain-computer interfaces: signal acquisition, processing, emerging approaches, and applications, *Signal Process. Strateg.* (2025) 281–310, <https://doi.org/10.1016/B978-0-323-95437-2.00014-8>.
- [3] A. Al-Saegh, S.A. Dawwd, J.M. Abdul-Jabbar, Deep learning for motor imagery EEG-based classification: a review, *Biomed. Signal Process. Control* 63 (2021) 102172, <https://doi.org/10.1016/j.bspc.2020.102172>.
- [4] M. Xu, F. He, T.-P. Jung, X. Gu, D. Ming, Current challenges for the practical application of electroencephalography-based brain–computer interfaces, *Engineering* 7 (2021) 1710–1712, <https://doi.org/10.1016/j.eng.2021.09.011>.
- [5] H. Wang, W. Zheng, Local temporal common spatial patterns for robust single-trial EEG classification, *IEEE Trans. Neural Syst. Rehabil. Eng.* 16 (2008) 131–139, <https://doi.org/10.1109/TNSRE.2007.914468>.
- [6] K.K. Ang, Z.Y. Chin, C. Wang, C. Guan, H. Zhang, Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b, *Front. Neurosci.* 6 (2012) 39, <https://doi.org/10.3389/fnins.2012.00039>.
- [7] Y. Zhang, G. Zhou, J. Jin, X. Wang, A. Cichocki, Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface, *J. Neurosci. Methods* 255 (2015) 85–91, <https://doi.org/10.1016/j.jneumeth.2015.08.004>.
- [8] V. Mishuhina, X. Jiang, Complex common spatial patterns on time-frequency decomposed EEG for brain-computer interface, *Pattern Recogn.* 115 (2021) 107918, <https://doi.org/10.1016/j.patcog.2021.107918>.
- [9] T.-J. Luo, Selective multi-view time–frequency decomposed spatial feature matrix for motor imagery EEG classification, *Expert Syst. Appl.* 247 (2024) 123239, <https://doi.org/10.1016/j.eswa.2024.123239>.
- [10] V. Sorin, Y. Barash, E. Konen, E. Klang, Deep learning for natural language processing in radiology—fundamentals and a systematic review, *J. Am. Coll. Radiol.* 17 (2020) 639–648, <https://doi.org/10.1016/j.jacr.2019.12.026>.
- [11] I.H. Sarker, Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions, *SN Comput. Sci.* 2 (2021) 1–20, <https://doi.org/10.1007/s42979-021-00815-1>.
- [12] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, *IEEE Trans. Neural Networks Learn. Syst.* 33 (2021) 6999–7019, <https://doi.org/10.1109/TNNLS.2021.3084827>.
- [13] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans. Signal Process.* 45 (1997) 2673–2681, <https://doi.org/10.1109/78.650093>.
- [14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Commun. ACM* 63 (2020) 139–144, <https://doi.org/10.1145/3422622>.
- [15] H. Altaheri, G. Muhammad, M. Alsulaiman, S.U. Amin, G.A. Altuwaijri, W. Abdul, M.A. Bencherif, M. Faisal, Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review, *Neural Comput. Appl.* 35 (2023) 14681–14722, <https://doi.org/10.1007/s00521-021-06352-5>.
- [16] J. Chen, Z. Yu, Z. Gu, Y. Li, Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces, *IEEE Trans. Neural Syst. Rehabil. Eng.* 28 (2020) 2356–2366, <https://doi.org/10.1109/TNSRE.2020.3023417>.
- [17] S.U. Amin, H. Altaheri, G. Muhammad, W. Abdul, M. Alsulaiman, Attention-inception and long-short-term memory-based electroencephalography classification for motor imagery tasks in rehabilitation, *IEEE Trans. Ind. Inf.* 18 (2021) 5412–5421, <https://doi.org/10.1109/TII.2021.3132340>.
- [18] H. Li, M. Ding, R. Zhang, C. Xiu, Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network, *Biomed. Signal Process. Control* 72 (2022) 103342, <https://doi.org/10.1016/j.bspc.2021.103342>.
- [19] H. Jia, S. Yu, S. Yin, L. Liu, C. Yi, K. Xue, F. Li, D. Yao, P. Xu, T. Zhang, A model combining multi branch spectral-temporal CNN, efficient channel attention, and LightGBM for MI-BCI classification, *IEEE Trans. Neural Syst. Rehabil. Eng.* 31 (2023) 1311–1320, <https://doi.org/10.1109/TNSRE.2023.3243992>.
- [20] S. Yu, Z. Wang, F. Wang, K. Chen, D. Yao, P. Xu, Y. Zhang, H. Wang, T. Zhang, Multiclass classification of motor imagery tasks based on multi-branch convolutional neural network and temporal convolutional network model, *Cereb. Cortex* 34 (2024) bhad511, <https://doi.org/10.1093/cercor/bhad511>.
- [21] V. Mnih, N. Heess, A. Graves, Recurrent models of visual attention, *Adv. Neural Inform. Process. Syst.* 27 (2014). 10.48550/arXiv.1406.6247.
- [22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *International conference on machine learning, PMLR* (2015) 2048–2057, <https://dl.acm.org/doi/10.5555/3045118.3045336>.
- [23] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, *Neurocomputing* 452 (2021) 48–62, <https://doi.org/10.1016/j.neucom.2021.03.091>.
- [24] A. Vaswani, Attention is all you need, *Adv. Neural Inform. Process. Syst.* 30 (2017). 10.48550/arXiv.1706.03762.
- [25] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, A survey on vision transformer, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (2022) 87–110, <https://doi.org/10.1109/TPAMI.2022.3152247>.
- [26] Y. Song, Q. Zheng, B. Liu, X. Gao, EEG conformer: convolutional transformer for EEG decoding and visualization, *IEEE Trans. Neural Syst. Rehabil. Eng.* 31 (2022) 710–719, <https://doi.org/10.1109/TNSRE.2022.3230250>.
- [27] H. Altaheri, G. Muhammad, M. Alsulaiman, Physics-informed attention temporal convolutional network for EEG-based motor imagery classification, *IEEE Trans. Ind. Inf.* 19 (2022) 2249–2258, <https://doi.org/10.1109/TII.2022.3197419>.
- [28] M. Cai, Y. Zeng, MAE-EEG-Transformer: a transformer-based approach combining masked autoencoder and cross-individual data augmentation pre-training for EEG classification, *Biomed. Signal Process. Control* 94 (2024) 106131, <https://doi.org/10.1016/j.bspc.2024.106131>.
- [29] W. Zhao, X. Jiang, B. Zhang, S. Xiao, S. Weng, CTNet: a convolutional transformer network for EEG-based motor imagery classification, *Sci. Rep.* 14 (2024) 20237, <https://doi.org/10.1038/s41598-024-71118-7>.
- [30] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, *J. Neural Eng.* 15 (2018) 056013, <https://doi.org/10.1088/1741-2552/aace8c>.
- [31] Y. Qin, B. Yang, S. Ke, P. Liu, F. Rong, X. Xia, M-FANet: Multi-feature attention convolutional neural network for motor imagery decoding, *IEEE Trans. Neural Syst. Rehabil. Eng.* 32 (2024) 401–411, <https://doi.org/10.1109/TNSRE.2024.3351863>.
- [32] Y. Zhou, J. Wu, Z. Ren, Z. Yao, W. Lu, K. Peng, Q. Zheng, C. Song, W. Ouyang, C. Gou, Csbrian: a cross-scale spatiotemporal brain foundation model for eeg decoding, *Adv. Neural Inform. Process. Syst.* 38 (2026) 87150–87195. 10.48550/arXiv.2506.23075.
- [33] X. Cai, S. Ma, L. Cao, J. Li, T. Liu, Y. Dong, Fusion of Multiscale Features Via Centralized Sparse-attention Network for EEG Decoding, *arXiv preprint arXiv: 2512.18689*, (2025). 10.48550/arXiv.2512.18689.
- [34] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, G. Pfurtscheller, BCI Competition 2008–Graz data set a, Institute for knowledge discovery (laboratory of brain-computer interfaces), *Graz Univ. Technol.* 16 (2008) 1–6, <https://dx.doi.org/10.21227/katb-zv89>.
- [35] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, G. Pfurtscheller, BCI competition 2008–graz data set B, *Graz Univ. Technol. Austria* 16 (2008) 1–6.
- [36] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, S.-W. Lee, EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy, *GigaScience* 8 (2019) giz002.
- [37] R.T. Schirrmester, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggenberger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, *Hum. Brain Mapp.* 38 (2017) 5391–5420, <https://doi.org/10.1002/hbm.23730>.
- [38] X. Ma, W. Chen, Z. Pei, Y. Zhang, J. Chen, Attention-based convolutional neural network with multi-modal temporal information fusion for motor imagery EEG decoding, *Comput. Biol. Med.* 175 (2024) 108504, <https://doi.org/10.1016/j.combiomed.2024.108504>.
- [39] S. Ioffe, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167*, (2015). doi:10.48550/arXiv.1502.03167.
- [40] D.-A. Clevert, Fast and accurate deep network learning by exponential linear units (elus), *arXiv preprint arXiv:1511.07289*, (2015). 10.48550/arXiv.1511.07289.
- [41] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11534–11542. doi:10.48550/arXiv.1910.03151.
- [42] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: beyond efficient transformer for long sequence time-series forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 11106–11115, <https://doi.org/10.1609/aaai.v35i12.17325>.
- [43] H. Wu, Y. Xiao, Y. Chen, C. Chen, R. Dong, D. Lin, DF-CoopNet: Cooperative perception via local feature enhancement and global sparse attention, *Adv. Eng. Inf.* 71 (2026) 104282, <https://doi.org/10.1016/j.aei.2025.104282>.
- [44] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, *arXiv preprint arXiv:1904.10509*, (2019). 10.48550/arXiv.1904.10509.
- [45] J. Lei Ba, J.R. Kiros, G.E. Hinton, Layer normalization, *ArXiv e-prints*, (2016) arXiv: 1607.06450. 10.48550/arXiv.1607.06450.
- [46] J. Shi, J. Tang, Z. Lu, R. Zhang, J. Yang, Q. Guo, D. Zhang, A brain topography graph embedded convolutional neural network for EEG-based motor imagery classification, *Biomed. Signal Process. Control* 95 (2024) 106401, <https://doi.org/10.1016/j.bspc.2024.106401>.
- [47] G. Liang, D. Cao, J. Wang, Z. Zhang, Y. Wu, EISATC-fusion: inception self-attention temporal convolutional network fusion for motor imagery EEG decoding, *IEEE Trans. Neural Syst. Rehabil. Eng.* 32 (2024) 1535–1545, <https://doi.org/10.1109/TNSRE.2024.3382226>.
- [48] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (2008) 2579–2605. <https://www.jmlr.org/papers/v9/vandermaaten08a.html>.

- [49] Z. Guo, H. Kamigaito, T. Watanabe, Attention score is not all you need for token importance indicator in kv cache reduction: Value also matters, arXiv preprint arXiv:2406.12335, (2024). 10.48550/arXiv.2406.12335.
- [50] A. Jiang, J. Shang, X. Liu, Y. Tang, H.K. Kwan, Y. Zhu, Efficient CSP algorithm with spatio-temporal filtering for motor imagery classification, *IEEE Trans. Neural Syst. Rehabil. Eng.* 28 (2020) 1006–1016, <https://doi.org/10.1109/TNSRE.2020.2979464>.
- [51] J. Wang, Y. Luo, L. Wang, L. Zhang, X. Kang, MTSAN-MI: multiscale temporal-spatial convolutional self-attention network for motor imagery classification, in: *International Conference on Neural Information Processing*, Springer, 2023, pp. 338–349, https://doi.org/10.1007/978-981-99-8138-0_27.
- [52] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141, <https://doi.org/10.1109/CVPR.2018.00745>.
- [53] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cham: convolutional block attention module, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.
- [54] L. Mei, T.M. Ingolfsson, C. Cioflan, V. Kartsch, A. Cossettini, X. Wang, L. Benini, An ultra-low power wearable BMI system with continual learning capabilities, *IEEE Trans. Biomed. Circ. Syst.* 19 (2024) 511–522, <https://doi.org/10.1109/TBCAS.2024.3457522>.
- [55] Y. Chen, R. Yang, M. Huang, Z. Wang, X. Liu, Single-source to single-target cross-subject motor imagery classification based on multisubdomain adaptation network, *IEEE Trans. Neural Syst. Rehabil. Eng.* 30 (2022) 1992–2002, <https://doi.org/10.1109/TNSRE.2022.3191869>.
- [56] S. Xiong, L. Wang, G. Xia, J. Deng, MBCNN-EATCFNet: a multi-branch neural network with efficient attention mechanism for decoding EEG-based motor imagery, *Rob. Auton. Syst.* 185 (2025) 104899, <https://doi.org/10.1016/j.robot.2024.104899>.
- [57] J. Leon, J.J. Escobar, A. Ortiz, J. Ortega, J. Gonzalez, P. Martín-Smith, J.Q. Gan, M. Damas, Deep learning for EEG-based Motor Imagery classification: accuracy-cost trade-off, *PLoS One* 15 (2020) e0234178, <https://doi.org/10.1371/journal.pone.0234178>.
- [58] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, D. Yang, EANet: Iterative edge attention network for medical image segmentation, *Pattern Recogn.* 127 (2022) 108636, <https://doi.org/10.1016/j.patcog.2022.108636>.
- [59] Y. Ma, W. Zhao, M. Meng, Q. Zhang, Q. She, J. Zhang, Cross-subject emotion recognition based on domain similarity of EEG signal transfer learning, *IEEE Trans. Neural Syst. Rehabil. Eng.* 31 (2023) 936–943, <https://doi.org/10.1109/TNSRE.2023.3236687>.
- [60] I. Sturm, S. Lapuschkin, W. Samek, K.-R. Müller, Interpretable deep neural networks for single-trial EEG classification, *J. Neurosci. Methods* 274 (2016) 141–145, <https://doi.org/10.1016/j.jneumeth.2016.10.008>.
- [61] S. Næss, G. Haldnes, E. Hagen, D.J. Hagler Jr, A.M. Dale, G.T. Einevoll, T.V. Ness, Biophysically detailed forward modeling of the neural origin of EEG and MEG signals, *Neuroimage* 225 (2021) 117467, <https://doi.org/10.1016/j.neuroimage.2020.117467>.