scientific reports

Check for updates

OPEN

Feature-aware domain invariant representation learning for EEG motor imagery decoding

Jianxiu Li[1], Jiaxin Shi[1]✉, Pengda Yu[1], Xiaokai Yan[1] & Yuting Lin[2]

Electroencephalography (EEG)-based motor imagery (MI) is extensively utilized in clinical rehabilitation and virtual reality-based movement control. Decoding EEG-based MI signals is challenging because of the inherent spatio-temporal variability of the original signal representation, coupled with a low signal-to-noise ratio (SNR), which impedes the extraction of clean and robust features. To address this issue, we propose a multi-scale spatio-temporal domain-invariant representation learning method, termed MSDI. By decomposing the original signal into spatial and temporal components, the proposed method extracts invariant features at multiple scales from both components. To further constrain the representation to invariant domains, we introduce a feature-aware shift operation that resamples the representation based on its feature statistics and feature measure, thereby projecting the features into a domain-invariant space. We evaluate our proposed method via two publicly available datasets, BNCI2014-001 and BNCI2014-004, demonstrating state-of-the-art performance on both datasets. Furthermore, our method exhibits superior time efficiency and noise resistance.

Keywords EEG, Motor imagery, Domain-invariant, Representation learning

Electroencephalography (EEG) records spontaneous electrical activity in the brain, referred to as an electrogram[1]. EEG, a non-invasive technique used in brain–computer interface (BCI), offers the advantage of acquiring brain signals without the need for surgical intervention, which makes it widely applicable in areas such as post-stroke motor rehabilitation, wheelchair control, virtual reality, and the metaverse[2,3], as well as emotion detection[4,5] and fatigue detection[6]. However, in contrast to invasive BCIs, EEG signals exhibit a lower signal-to-noise ratio (SNR) and greater spatio-temporal variability, which constrain the development of EEG-based motor imagery (MI) signal decoding.

Most prior studies have concentrated on developing feature extraction methods to enhance the decoding performance of EEG-based MI signals[7–13]. For example, EEG Conformer[14] presents a hybrid architecture that combines Convolution and Transformer to capture both local spatiotemporal features and global temporal dependencies. EEG-TransNet[15] first extracts multimodal temporal features, such as the mean and variance, and then integrates a shared self-attention module to capture global dependencies across these dimensions, thereby enhancing the flexibility of signal segmentation for local fluctuations. Tensor-CSPNet[16] introduces a geometric deep learning framework for extracting spatio-temporal frequency patterns, while Graph-CSPNet[17] leverages Graph Neural Networks (GNNs) within the Symmetric Positive Definite (SPD) manifold space to capture complex relationships within EEG signals. Despite these advancements, these methods often rely on raw EEG-based MI signals for feature extraction. Deep learning algorithms, a recently preferred approach for feature extraction, effectively capture the nonlinear characteristics of EEG-based MI representations. Nevertheless, since the weights of these models are optimized solely based on task labels, noise, and spatiotemporal variability significantly affect decoding accuracy. Furthermore, these end-to-end deep learning methods may erroneously interpret irrelevant information as beneficial, severely limiting the robustness and performance of models in decoding EEG-based MI signals.

To address this challenge, we propose MSDI, a multi-scale spatio-temporal domain-invariant representation learning method. First, we transform the original signal into multiple representations across varying time and channel scales by applying temporal and spatial windows, allowing the model to focus on both fine-grained and coarse-grained features across time and channels. Next, we decouple temporal and spatial information to extract time-dependent and space-dependent features separately. This aims to minimize the impact of spatiotemporal variability and reduce the aliasing between temporal and spatial features. Based on the relative importance of features at different scales, we then adaptively fuse spatial and temporal features across multiple scales into a

[1]Inner Mongolia University, Huhhot 010021, China. [2]Lanzhou University, Lanzhou 730000, China. ✉email: 0221121528@mail.imu.edu.cn

unified representation. Acknowledging the inherent correlation between temporal and spatial information, we employ a Cross-Spatio-Temporal Attention (CST-Attn) mechanism to facilitate the cross-fusion of these features, followed by dynamic weighting of the fused multi-scale information to generate a novel signal representation. To further improve the robustness of the enhanced representation, we introduce a feature-aware shift operation. Inspired by work on domain generalization[18,19] and diffusion models[20], this operation maps a representation to a domain-invariant space by randomly resampling the representation based on its feature statistics (i.e., mean and variance) and incorporating Gaussian noise that is constrained by the feature measure. Finally, to extract features from the constructed MSDI representation and predict MI classification results, we leverage the strengths of Convolution and Transformer architectures to design a hybrid backbone for EEG-based MI decoding.

We evaluated our method on two publicly available datasets, BNCI2014-001 and BNCI2014-004, achieving state-of-the-art performance. The experimental results demonstrate that enhancing the original EEG signal representation improves the model's decoding accuracy and robustness. Additionally, transferring the representation to a domain-invariant space also positively impacts model performance, particularly leading to significant improvements for subjects with poor decoding performance. Furthermore, when decoding with limited time points, our model demonstrated more stable performance and greater time efficiency, consistently outperforming baseline models. In addition, when additional noise was introduced to the test data as interference, the model incorporating the feature-aware shift operation exhibited enhanced noise resistance and robustness, further confirming its superiority.

Our contributions are summarized as follows:

- We propose MSDI, a novel multi-scale spatio-temporal domain-invariant representation learning method designed to enhance EEG-based MI signal representation by reducing spatio-temporal variability.
- We introduce the feature-aware shift operation to map the feature space of MSDI to a domain-invariant space, enhancing robustness.
- Our method achieves classification accuracies of 81.06% and 89.42% on BNCI2014-001 and BNCI2014-004, respectively, while also demonstrating superior time efficiency and noise resistance.

Related works

EEG-based motor imagery signal decoding

Previous decoding methods have primarily focused on improving feature extraction from EEG signals[21–27]. Common Spatial Pattern (CSP), SSCSP[28], and FBCSP[29] use spatial filters to extract spatial features, maximizing signal variance across two tasks. However, these methods struggle to capture the nonlinear characteristics of EEG signals, particularly under low SNR conditions. Deep learning (DL) techniques have advanced the decoding of nonlinear features and mitigated the impact of low SNR. For example, EEGNet[10] uses compact convolutional neural networks (CNNs) for various BCI tasks. EEG Conformer[14] designs a hybrid architecture combining Convolution and Transformer to capture local spatio-temporal features and global temporal dependencies. EEG-TransNet[15] first extracts multimodal temporal features such as mean and variance, then incorporates a shared self-attention module to capture global dependencies across these feature dimensions, thereby offering enhanced flexibility in signal segmentation to capture local fluctuations. Tensor-CSPNet[16] introduces a geometric deep learning (DL) framework for extracting spatiotemporal-frequency patterns, while Graph-CSPNet[17] employs Graph Neural Networks (GNNs) in the SPD manifold space to capture complex relationships within EEG data. Similarly, MAMCNet[12] constructs frequency-domain convolutional blocks to merge optimized spatio-temporal feature maps from different frequency bands and converts these feature maps into Riemannian manifolds. While these methods design explicit feature extraction approaches, they do not explore enhancing the original EEG signal representation. In addition, DL-based methods also struggle with overfitting. In this work, we propose the feature-aware shift operation to generalize the features and expand the domain, transforming the representation into a domain-invariant space.

Representation learning

Representation learning is a subset of machine learning that aims to automatically discover optimal data representations to enhance task performance[30,31]. In the context of EEG decoding paradigms, representation learning has been instrumental in improving signal quality and decoding MI[32]. For instance, Xiang et al.[33] proposed a two-stream model to analyze temporal and spatial EEG representations, highlighting the importance of spatial and functional electrode connections. Li et al.[34] introduced CASCE, a model designed to enhance EEG signal representation and provide insights into functional neural activity during MI. Additionally, the SCDM employs spatial cross-modal generation and multi-scale temporal representation modules to adaptively learn latent temporal and spatial features within a unified representation space[35]. In contrast to these studies, our approach specifically aims to address the spatiotemporal variability of EEG signals and mitigate the impact of low SNR on decoding performance. Our method enhances EEG signal representation using a spatiotemporal decoupling strategy and increases robustness by transferring the representation to a domain-invariant space.

Methods

In this section. We introduce our proposed multi-scale spatio-temporal domain-invariant (MSDI) representation learning method, as shown in Fig. 1a. We first use MSDI to enhance the original signal representation to alleviate the influence of the variety of spatial and temporal information and make the representation more robust. To further enhance the generalization of the representation we proposed a feature-ware shift operation to shift the features of the enhanced representation to the invariant domain, as shown in Fig. 1b. Finally, we utilize a hybrid backbone, based on convolution and transformer architecture to extract the features and decode the EEG-based MI signal, as shown in Fig. 1c.

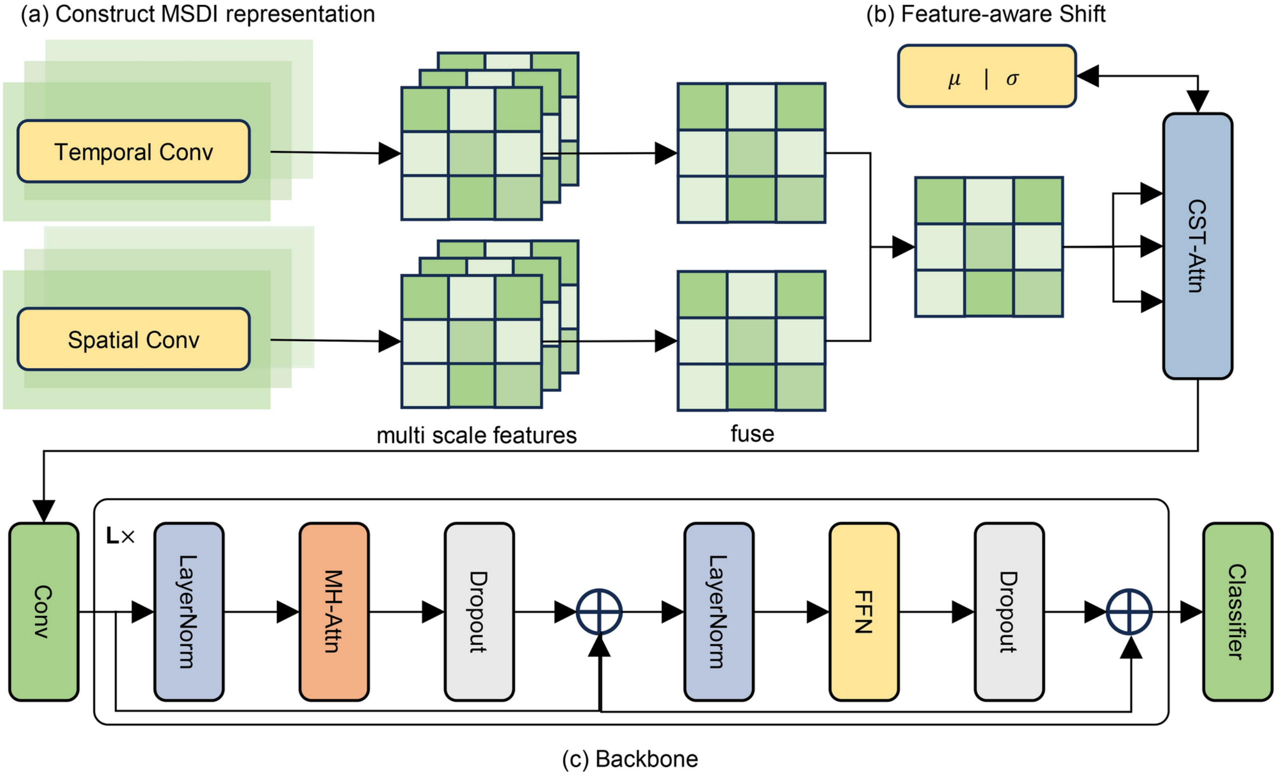


Fig. 1. The overall architecture of our proposed method. As shown in (**a**), the process begins with the construction of an MSDI representation. This involves parallel temporal and spatial convolutional pathways, which extract multi-scale features from the input EEG data. These features are subsequently fused to create a comprehensive representation. Subsequently, as shown in (**b**), a Feature-aware Shift mechanism is employed. This module performs resampling from noise constrained by feature measure to further enhance the domain-invariant representation and generalization ability of MSDI. Finally, (**c**) depicts the backbone network for decoding the EEG signals.

Multi-scale spatio-temporal domain-invariant representation

Multi-scale spatial and temporal feature extraction

Given EEG data $\mathscr{D} \in \mathbb{R}^{C\times T}$, where C denotes the number of channels and T the number of time points, we initially utilize two convolutional groups, denoted as $\text{Conv}^{spatial}_{\mathscr{S}i}$ and $\text{Conv}^{temporal}_{\mathscr{S}i}$, to extract multi-scale spatial and temporal features from the signal $\mathscr{D}$. To decouple spatial and temporal features, we employ spatial and temporal convolutions, each moving exclusively along a single dimension (i.e., the spatial or temporal dimension), leveraging the translational invariance of convolution to mitigate aliasing effects between the channel and time domains.

$$\mathscr{D}'_S = \text{Concat}\left(\text{Conv}^{spatial}_{\mathscr{S}i}\left(D\right)\right), \quad \text{for} \quad i = 1, \dots, \mathscr{S}_S \tag{1}$$

where $\mathscr{D}'_S \in \mathbb{R}^{\mathscr{S}_S\times C\times T}$ is used to represent multi-scale spatial features at different scales $\mathscr{S}_S$. We then transpose the signal D to $D^T \in \mathbb{R}^{T\times C}$, and use temporal convolution to extract multi-scale temporal features:

$$\mathscr{D}'^{T}_T = \text{Concat}\left(\text{Conv}^{temporal}_{\mathscr{S}j}\left(D^T\right)\right), \quad \text{for} \quad j = 1, \dots, \mathscr{S}_T \tag{2}$$

where $\mathscr{D}'^{T}_T \in \mathbb{R}^{\mathscr{S}_T\times T\times C}$ is used to represent multi-scale temporal features at different scales $\mathscr{S}_T$.

Weighted aggregation and refinement of multi-scale features

We use linear layer weight $W_S = \text{Linear}(\mathscr{D}'_S)$, $W_T = \text{Linear}(\mathscr{D}'_T)$, and softmax function to dynamically obtain the weight of every scale of spatial and temporal features:

$$\sigma^{spatial}_i = \frac{e^{W_{Si}}}{\sum_{j=1}^{\mathscr{S}_S} e^{W_{Sj}}}, \quad \text{for} \quad i = 1, \dots, \mathscr{S}_S \tag{3}$$

$$\sigma^{temporal}_i = \frac{e^{W_{Ti}}}{\sum_{j=1}^{\mathscr{S}_T} e^{W_{Tj}}}, \quad \text{for} \quad i = 1, \dots, \mathscr{S}_T \tag{4}$$

Then, we use these dynamic weights to weigh the multi-scale features, and extract independent features of spatial and temporal dimensions:

$$F_S = \mathrm{Conv}_S \left(\sum_{i=1}^{s_S} \sigma_i^{spatial} \times \mathscr{D}'_{Si} \right) \tag{5}$$

$$F_T = \mathrm{Conv}_T \left(\sum_{i=1}^{s_T} \sigma_i^{temporal} \times \mathscr{D}'_{Ti} \right) \tag{6}$$

where Conv_S and Conv_T only extract spatial features and temporal features, respectively.

Feature fusion
We finally fuse the spatial and temporal features:

$$W = W_f \left(F_S + F_T \right) \tag{7}$$

where W_f is a single linear layer. Then we utilize multi-head attention to further mix the features:

$$\mathscr{D}_{\mathrm{Enhance}} = \text{CST-Attn} \left(W, W, W \right) + \mathscr{D} \tag{8}$$

where $\mathscr{D}_{\mathrm{Enhance}}$ is the enhanced representation. Cross-Spatio-Temporal Attention (CST-Attn) is a multi-head attention to fuse the spatial and temporal features.

Feature-aware shift operation

Our goal is to enhance the robustness of the representation by shifting it to a wider, more domain-invariant space. This involves increasing the diversity of the feature statistics, effectively broadening their distribution.

Shift to domain-invariant space
We model the distribution of the $\mathscr{D}_{\mathrm{Enhance}}$ as a multivariate Gaussian distribution for simplicity:

$$\mathscr{D}_{\mathrm{Enhance}} \sim \mathscr{N} \left(\mu, \sigma^2 \right) \tag{9}$$

We use $\mu, \sigma^2 \in \mathbb{R}^{Batch \times 1}$ to represent the feature mean and variance:

$$\mu = \frac{1}{C \times T} \sum_{c=1}^{C} \sum_{t=1}^{T} d_{c,t} \tag{10}$$

$$\sigma^2 = \frac{1}{C \times T} \sum_{c=1}^{C} \sum_{t=1}^{T} \left(d_{c,t} - \mu \right)^2 \tag{11}$$

where $d \in \mathbb{R}^{1 \times C \times T}$ is a single batch of the enhanced representation. For a batch of data, we can calculate the variance of μ and σ:

$$\sigma_\mu^2 = \frac{1}{\mathrm{Batch}} \sum_{b=1}^{\mathrm{Batch}} \left(\mu_b - \mu' \right)^2 \tag{12}$$

$$\sigma_\sigma^2 = \frac{1}{\mathrm{Batch}} \sum_{b=1}^{\mathrm{Batch}} \left(\sigma_b - \sigma' \right)^2 \tag{13}$$

where $\mu' = \frac{1}{Batch} \sum_{b=1}^{Batch} \mu_b$, $\sigma' = \frac{1}{Batch} \sum_{b=1}^{Batch} \sigma_b$. Next, let ε_μ and ε_σ are both random noise sampling from the standard normal distribution and regularized by the original feature measure and the offset. The mean and standard deviation for resampling are given by $\mu_r = \mu + \varepsilon_\mu \sigma_\mu$ and $\sigma_r = \sigma + \varepsilon_\sigma \sigma_\sigma$. The processed data is as follows:

$$\mathscr{D}_{\mathrm{Enhance}} = \sigma_r \left(\frac{\mathscr{D}_{\mathrm{Enhance}} - \mu}{\sigma} \right) + \mu_r \tag{14}$$

Feature-adaptive noise injection and shift
To align the spatial scale of the shifted representation with that of the original representation, we use the feature mean of the original representation to constrain the sampling of the standard normal distribution:

$$\varepsilon_\mu = \alpha \cdot \mathrm{mean}_{Batch} \left(\mu \right) \cdot \mathrm{n}_1 + \gamma_1 \tag{15}$$

$$\varepsilon_\sigma = \beta \cdot \mathrm{mean}_{Batch} \left(\sigma \right) \cdot \mathrm{n}_2 + \gamma_2 \tag{16}$$

where α and β are hyperparameters used to control the magnitude of noise sampling. γ_1 and γ_2 serve as offsets to allow the model to explore a broader range of features, further enhancing diversity.

Decoding backbone

Inspired by Conformer[14], we designed a hybrid architecture to decode the enhanced EEG signal with MSDI. We leverage a hybrid backbone combining convolutional embedding for local feature extraction, which captures fine-grained information reflecting localized neuronal activity, and a Transformer encoder for capturing long-range dependencies, essential for understanding the interactions and synchronization between different brain regions reflected in the EEG signals.

Convolutional embedding

The Convolutional Embedding module aims to extract local temporal and spatial features from the enhanced EEG signal representation $\mathscr{D}_{\text{Enhance}}$:

$$\text{Embedding} = AvgPool\left(C^{spatial}\left(C^{temporal}\left(\mathscr{D}_{\text{Enhance}}\right)\right)\right) \tag{17}$$

where $C^{spatial}(\cdot)$ and $C^{temporal}(\cdot)$ represent two convolutional layers with different kernel sizes. Specifically, $C^{spatial}(\cdot)$ is a spatial convolution with kernel size (*C*, 1) and $C^{temporal}(\cdot)$ is a temporal convolution with kernel size $(1, F_t)$, where *C* is the number of channels, and F_t is the temporal filter length. $AvgPool(\cdot)$ is an average pooling layer with kernel size $(1, P_t)$ and stride $(1, S_t)$, where P_t is the pooling length and S_t is the stride.

The output of the embedding is then projected into a higher-dimensional space:

$$\text{Embedding} = C^{\text{proj}}\left(\text{Embedding}\right) \tag{18}$$

where $C^{\text{proj}}(\cdot)$ is a 1×1 convolution operation with weights.

Transformer block

The Transformer Block aims to capture long-range dependencies. It consists of a stack of *N* identical Transformer Encoder blocks, each comprising a Multi-Head Self-Attention (MSA) layer and a Feed-Forward Network (FFN). Each Transformer Encoder block can be represented as:

$$Z_{l+1} = \text{FFN}\left(\text{MSA}\left(LN\left(Z_l\right)\right)\right) + Z_l \tag{19}$$

where Z_l is the input to the *l*-th block, $LN(\cdot)$ denotes Layer Normalization, and the residual connection adds the input Z_l to the output of the FFN.

For each input embedding Z_l, three linear transformations are performed to obtain the query *Q*, key *K*, and value *V* matrices:

$$Q = Z_l W^Q, \quad K = Z_l W^K, \quad V = Z_l W^V \tag{20}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{F_n \times F_n}$ are learnable weight matrices. The queries, keys, and values are then split into *H* heads:

$$Q_h = QW_h^Q, \quad K_h = KW_h^K, \quad V_h = VW_h^V \tag{21}$$

where $h \in \{1, \ldots, H\}$ and $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{F_n \times d_k}$ are learnable weight matrices with $d_k = \frac{F_n}{H}$. The attention for each head is calculated as:

$$A_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) \tag{22}$$

where the scaling factor $\sqrt{d_k}$ prevents the dot products from becoming too large. A dropout layer Dropout with probability p_d is applied to the attention weights:

$$A_h = \text{Dropout}\left(A_h\right) \tag{23}$$

The output of each head is then computed as:

$$O_h = A_h V_h \tag{24}$$

Finally, the outputs of all heads are concatenated and linearly projected:

$$\text{MSA}\left(Z_l\right) = [O_1, \ldots, O_H]W^O \tag{25}$$

where $W^O \in \mathbb{R}^{F_n \times F_n}$ is a learnable weight matrix, and $[O_1, \ldots, O_H]$ denotes the concatenation of the output of each head.

The FFN consists of two linear transformations:

$$\mathrm{FFN}\left(X\right)=W_2\sigma\left(W_1X+b_1\right)+b_2 \tag{26}$$

where $\sigma(\cdot)$ is the GELU activation function.

Classifier
The output of the Transformer Block is first flattened into a vector $z\in\mathbb{R}^{B\times T'F_n}$:

$$z=\mathrm{Flatten}\left(Z_N\right) \tag{27}$$

This flattened vector is then fed into a Fully Connected layer:

$$FC=\sigma\left(L_2\left(\mathrm{Dropout}\left(\sigma\left(L_1\left(z\right)\right)\right)\right)\right) \tag{28}$$

where $L_1(\cdot)$ and $L_2(\cdot)$ represent two linear layers.

The output *FC* is then fed to a final linear layer, $L_f(\cdot)$, followed by a LogSoftmax function to obtain the class probabilities:

$$\hat{y}=\mathscr{L}\left(L_f\left(FC\right)\right) \tag{29}$$

where $L_f(\cdot)$ is a linear transformation with weights $W_f\in\mathbb{R}^{F_{fc2}\times N_c}$ and bias $b_f\in\mathbb{R}^{N_c}$, and N_c is the number of classes. $\mathscr{L}(\cdot)$ is a LogSoftmax function, defined as:

$$\mathscr{L}(x_i)=\log\left(\frac{\exp(x_i)}{\sum_{j=1}^{N_c}\exp(x_j)}\right) \tag{30}$$

where x_i is the *i*-th element of the input vector *x*. The output $\hat{y}\in\mathbb{R}^{B\times N_c}$ represents the log probabilities of each class for each sample in the batch.

Experiment

Datasets

We evaluated our model on two publicly available MI datasets, as shown in Table 1.

BNCI 2014-001
The BNCI 2014-001 MI dataset[36] contains EEG data from nine subjects performing four MI tasks: imagining the movement of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). Each subject participated in two sessions on separate days, including six runs of 48 trials, totaling 288 trials per session. During each trial, subjects fixated on a cross for 2 s, followed by a 1.25-s cue indicating the MI task. Subjects continued the MI task until the cross disappeared at 6 s, without receiving feedback. EEG signals were recorded using 22 Ag/AgCl electrodes, spaced 3.5 cm apart, in a monopolar configuration with the left mastoid as reference and the right mastoid as ground. The data were sampled at 250 Hz, bandpass-filtered between 0.5 and 100 Hz, and notch-filtered at 50 Hz to reduce line noise.

BNCI 2014-004
The BNCI 2014-004 dataset[37] contains EEG data from nine right-handed subjects with normal or corrected vision. Each subject participated in five sessions: the first two sessions were training without feedback, and the last three included feedback. EEG recordings were taken from three bipolar electrodes (C3, Cz, and C4) at 250 Hz, bandpass-filtered between 0.5 and 100 Hz, with a 50 Hz notch filter. The ground electrode was positioned at Fz. Additionally, EOG data were recorded using three monopolar electrodes. The screening paradigm involved two MI tasks: left hand (class 1) and right hand (class 2). Each subject completed two screening sessions without feedback, each consisting of six runs with ten trials per run, totaling 120 trials per session and 240 trials per subject. Each trial began with a fixation cross and an acoustic warning, followed by a 1.25-s visual cue and 4 s of MI. In the three feedback sessions, four runs with smiley feedback were recorded per session, each run consisting of 20 trials per MI class. The feedback paradigm involved a centered gray smiley at trial start, a warning beep at second 2, and a visual cue from seconds 3 to 7.5, followed by a blank screen with a randomized interval of 1.0–2.0 s.

Feature	BNCI 2014-001	BNCI 2014-004
Subjects	9	9
Tasks	Left hand, right hand, both feet, tongue	Left hand, right hand
Sessions	2 Sessions	5 Sessions
Sampling rate	250 Hz	250 Hz

Table 1. Comparison of BNCI 2014-001 and BNCI 2014-004 datasets.

Environment configuration

In order to test the stability of our proposed models and to ensure the fairness of the experiments, all models are trained and evaluated on a 12th Gen Intel(R) Core(TM) i7-12700H CPU, NVIDIA GeForce RTX 4090 GPU 24GB, 64GB RAM device. We chose to use a Common Average Reference (CAR) filter to eliminate the common noise and signal components across all channels, thereby improving the SNR. The CAR filter works by subtracting the average value of all channels, making the signal of each channel more independent:

$$V_i' = V_i - \frac{1}{N}\sum_{j=1}^{N} V_j \tag{31}$$

where V_i represents the original signal of the i-th channel, N is the number of channels, and $\sum_{j=1}^{N} V_j$ is the sum of the signals across all channels.

Classification performance

In this section, we evaluate our proposed method and compare it with other classic or recent methods. We all use the Hold-out data split method to evaluate these models. Our method not only achieves state-of-the-art classification performance but also demonstrates stability when just using limited time points for training.

Achieve state-of-the-art performance

Our proposed method demonstrates superior classification performance. As shown in Table 2, the proposed method achieves an average accuracy of 81.06% on the BNCI2014-001 dataset. This represents a significant improvement compared to established methods such as FBCSP[29], SSCSP[28], DRDA[38], DRDW[39], Conformer[14], MI-CAT[40] and EEG-TransNet[15]. Notably, the proposed method consistently outperforms competing methods in most of the nine subjects. Furthermore, gains are particularly pronounced for subjects A4, and A6.

On the BNCI2014-004 dataset, the proposed method maintains high accuracy. Table 3 illustrates an average accuracy of 89.42%. While the performance margin narrows against Conformer and EEG-TransNet on this dataset, the proposed method still delivers the highest overall accuracy and excels in subject-specific performance, notably in subjects B1, B2, and B5. The consistently high performance across both datasets suggests the robustness and generalization of the proposed method for EEG signal classification.

Timepoint efficiency

Table 4 presents a comparative analysis of time point efficiency, evaluating model performance with varying percentages of available time points on the BNCI2014-001 dataset. The results demonstrate the proposed method's superior data efficiency compared to Conformer and EEG-TransNet, as shown in Fig. 2. Specifically, when using only 25% of the time points, the proposed method achieves an average accuracy of 67.34%, significantly outperforming Conformer (62.45%) and EEG-TransNet (55.08%). This advantage is maintained at 50% of the time points, with the proposed method achieving 72.42% accuracy, while Conformer and EEG-TransNet reach 71.65% and 60.02%, respectively.

Our proposed method demonstrates superior robustness and stability, evidenced by consistently lower standard deviations across subjects and data percentages. Notably, at 25% of the time points, it achieves a standard deviation of 0.0858, outperforming Conformer (0.1100) and EEG-TransNet (0.1101). This indicates enhanced resilience to training data variability and effective generalization from limited subsets. Its higher accuracy and lower standard deviation across time point percentages underscore its efficiency in feature extraction, making it highly suitable for real-time BCI applications.

Ablation study

Influence on MSDI

The ablation studies, detailed in Tables 5 and 6, reveal the significant impact of MSDI on classification accuracy. Firstly, the absence of MSDI, represented by the first row in both tables, results in markedly lower performance across both datasets. Specifically, on BNCI2014-001, the average accuracy without MSDI is 73.46%, whereas

Model	A1 (%)	A2 (%)	A3 (%)	A4 (%)	A5 (%)	A6 (%)	A7 (%)	A8 (%)	A9 (%)	Avg. (%)
FBCSP[29]	77.80	62.13	84.69	69.82	53.46	51.39	92.69	80.90	87.19	73.34
SSCSP[28]	76.74	58.68	81.25	57.64	38.54	48.26	76.39	79.17	78.82	66.17
DRDA[38]	83.19	55.14	87.43	75.28	62.29	57.15	86.18	83.61	82.00	74.70
DRDW[39]	83.29	63.97	90.30	76.94	**69.34**	60.08	89.31	82.35	82.81	77.60
Conformer[14]	87.15	57.29	88.89	74.65	59.38	47.99	86.81	81.94	85.42	74.39
MI-CAT[40]	90.62	54.51	91.32	72.57	63.19	62.85	87.15	85.07	84.03	76.74
EEG-TransNet[15]	86.46	48.26	87.15	69.79	52.78	57.29	86.81	80.56	84.38	72.61
Ours	**90.63**	**64.24**	**92.01**	**79.51**	66.67	**64.24**	**93.75**	**89.24**	**89.24**	**81.06**

Table 2. Classification performance on BNCI2014-001 dataset. The proposed method achieves superior accuracy compared to existing methods, demonstrating its effectiveness in EEG-based MI decoding. The bold values indicate the highest performance achieved in each respective column.

Model	B1 (%)	B2 (%)	B3 (%)	B4 (%)	B5 (%)	B6 (%)	B7 (%)	B8 (%)	A9 (%)	Avg. (%)
FBCSP[29]	70.00	60.36	60.94	97.50	93.12	80.63	78.13	92.50	86.88	80.01
SSCSP[28]	65.00	56.79	54.06	95.63	74.69	79.06	80.00	87.81	82.81	75.09
DRDA[38]	81.37	62.86	63.63	95.94	93.56	88.19	85.00	95.25	90.00	83.98
DRDW[39]	84.66	66.57	68.04	96.78	94.32	82.61	88.47	93.96	90.10	85.06
Conformer[14]	77.81	70.71	85.31	**98.44**	97.50	87.19	92.19	94.69	91.56	88.38
MI-CAT[40]	**86.11**	65.97	61.46	98.26	93.75	89.24	86.11	**95.69**	90.97	85.26
EEG-TransNet[15]	79.06	70.71	**87.81**	**98.44**	96.88	**91.56**	91.88	**95.63**	90.00	89.11
Ours	81.56	**71.07**	85.63	97.81	**98.75**	89.06	**93.44**	94.69	**92.81**	**89.42**

Table 3. Classification performance on BNCI2014-004 dataset. The proposed method exhibits competitive performance, achieving the highest average accuracy and demonstrating strong generalization capability. The bold values indicate the highest performance achieved in each respective column.

Model	Percentage	Avg.	Std.
Conformer[14]	25	0.6245	0.1100
	50	0.7165	0.1493
EEG-TransNet[15]	25	0.5508	0.1101
	50	0.6002	0.1465
Ours	25	**0.6734**	**0.0858**
	50	**0.7242**	**0.1291**

Table 4. Time point efficiency of our method. The performance was evaluated on the BNCI2014-001 dataset. The bold values indicate the highest performance achieved in each respective column.

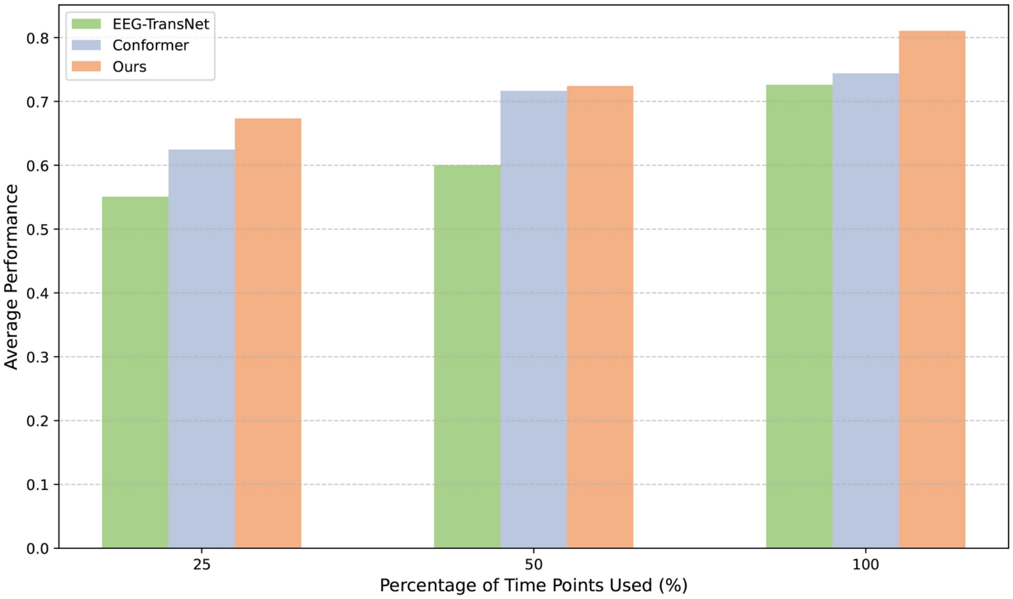


Fig. 2. Time point efficiency of Conformer, EEG-TransNet, and our method.

incorporating MSDI with spatial and temporal windows of [1, 3, 5, 7] boosts the accuracy to 79.55% (without feature-aware shift operation). A similar trend is observed on BNCI2014-004, where the accuracy increases from 85.30 to 88.54%. The pronounced disparity underscores the criticality of capturing multi-scale information across both spatial and temporal dimensions for effective EEG signal classification. By integrating features from diverse scales, MSDI enables the model to learn richer and more discriminative representations of the underlying neural activity.

Furthermore, the selection of spatial and temporal window sizes within MSDI influences the model's effectiveness. On BNCI2014-001, reducing the spatial windows from [1, 3, 5, 7] to [1, 3, 5] while keeping temporal windows constant at [1, 3, 5, 7] maintains a high average accuracy of 81.06%, but further reduction to [3] for both

MSDI												
Spatial	**Temporal**	**Feature-aware shift**	**A1 (%)**	**A2 (%)**	**A3 (%)**	**A4 (%)**	**A5 (%)**	**A6 (%)**	**A7 (%)**	**A8 (%)**	**A9 (%)**	**Avg. (%)**
			85.42	60.42	87.85	72.92	59.38	57.99	68.06	82.64	86.46	73.46
[1, 3, 5, 7]	[1, 3, 5, 7]		90.63	62.85	92.01	76.39	62.15	62.50	93.40	88.54	87.50	79.55
[1, 3, 5, 7]	[1, 3, 5, 7]	✓	90.63	64.24	92.01	**79.51**	66.67	64.24	93.75	**89.24**	**89.24**	**81.06**
[1, 3, 5]	[1, 3, 5, 7]	✓	**90.69**	63.89	**92.36**	78.13	**69.79**	62.85	**94.44**	88.54	88.89	**81.06**
[1, 3, 5]	[3, 7]	✓	89.93	64.24	93.06	78.47	67.36	63.54	**94.44**	88.54	86.81	80.71
[3]	[3]	✓	90.28	**65.28**	92.71	77.78	66.32	**65.63**	93.40	87.15	88.19	80.75

Table 5. Ablation study on BNCI2014-001 dataset. The bold values indicate the highest performance achieved in each respective column.

MSDI												
Spatial	**Temporal**	**Feature-aware shift**	**B1 (%)**	**B2 (%)**	**B3 (%)**	**B4 (%)**	**B5 (%)**	**B6 (%)**	**B7 (%)**	**B8 (%)**	**B9 (%)**	**Avg. (%)**
			79.06	63.93	81.56	97.50	96.56	84.69	85.63	92.19	86.56	85.30
[3]	[3]	✓	**82.50**	70.00	81.88	97.81	97.81	**91.25**	89.69	94.69	91.25	88.54
[1, 3]	[1, 3]	✓	79.69	**73.93**	**85.94**	**98.44**	**98.75**	87.19	92.81	**95.00**	91.88	89.29
[1, 3]	[1, 3, 5, 7]	✓	81.56	71.07	85.63	97.81	**98.75**	89.06	**93.44**	94.69	**92.81**	**89.42**

Table 6. Ablation study on BNCI2014-004 dataset. The bold values indicate the highest performance achieved in each respective column.

spatial and temporal windows slightly decreases performance to 80.75%. This suggests that capturing multi-scale information, particularly with a broader range of spatial scales, is beneficial for this dataset. Conversely, on BNCI2014-004, using spatial windows of [1, 3] and temporal windows of [1, 3, 5, 7], yields the highest accuracy (89.42%). These findings demonstrate the critical role of MSDI in enhancing classification performance, highlighting the importance of adaptable multi-scale feature extraction. In addition, while increasing the scale can enhance model performance, it concurrently increases the computational overhead of model inference, necessitating a careful balance between performance and computational efficiency.

Influence on feature-aware shift operation

The feature-aware shift operation demonstrates a consistent positive impact on classification performance. Comparing the second row (MSDI without feature-aware shift) to the subsequent rows (MSDI with feature-aware shift) in both Tables 5 and 6, the inclusion of the feature-aware shift operation consistently improves or maintains accuracy across different MSDI configurations. Notably, on BNCI2014-001, with spatial and temporal windows of [1, 3, 5, 7], the addition of the feature-aware shift increases the average accuracy from 79.55 to 81.06%. This improvement is observed across most individual subjects, indicating the general applicability of this operation. Concurrently, it demonstrates that the feature-aware shift operation enhances model robustness and enables MSDI to acquire more discriminative representations.

Analysis of noise sensitivity

To further demonstrate the robustness of the model, we evaluated its performance on the BNCI2014001 dataset by introducing varying levels of noise into the test set. As shown in Fig. 3, the feature-aware shift operation strengthens the domain-invariant representation of signals, consequently enhancing the model's resilience to noise. Specifically, as the noise level increases, the advantages of the feature-aware shift operation become more pronounced.

Decoding visualization

Feature distribution visualization

To better understand the effect of the feature-aware shift operation on the distribution of MSDI features, we visualize the enhanced MSDI representation $\mathscr{D}_{\text{Enhance}}$ after applying the operation and compare it with the MSDI representation without applying the operation, as shown in Fig. 4. It can be seen that both the original and enhanced features approximate a Gaussian distribution. The main difference lies in the feature distribution curve after applying the feature-aware shift operation, which widens with increasing variance. At the same time, the distribution of all channels changes to be more uniform. These indicate an increase in feature diversity and the original features are shifted into a more invariant space, which improves the generalization performance of the model.

t-SNE visualization

To further analyze the effectiveness of the proposed method in learning discriminative features, t-distributed Stochastic Neighbor Embedding (t-SNE) is employed for dimensionality reduction and visualization. T-SNE is a non-linear technique that maps high-dimensional data points to a two- or three-dimensional space while

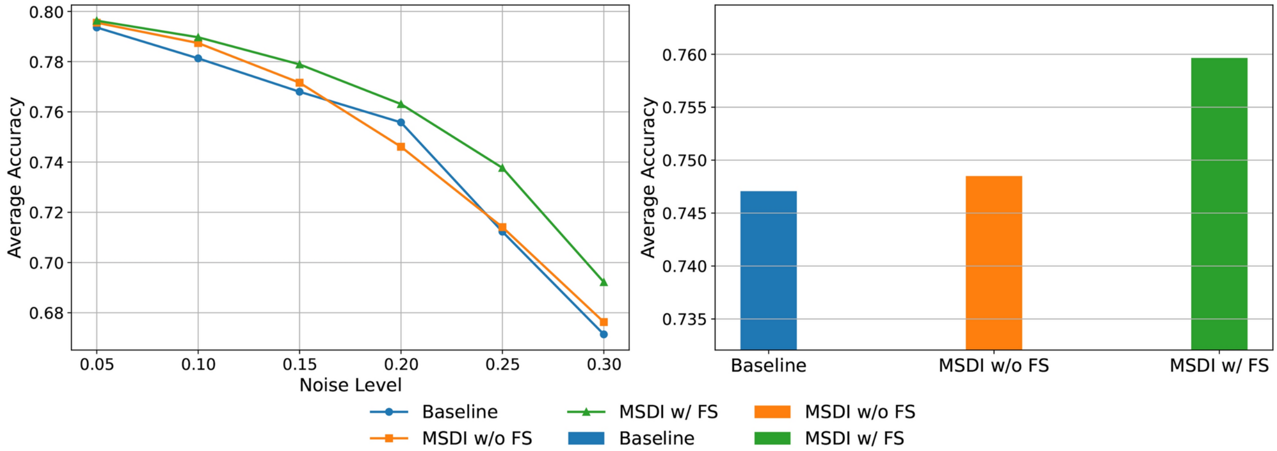
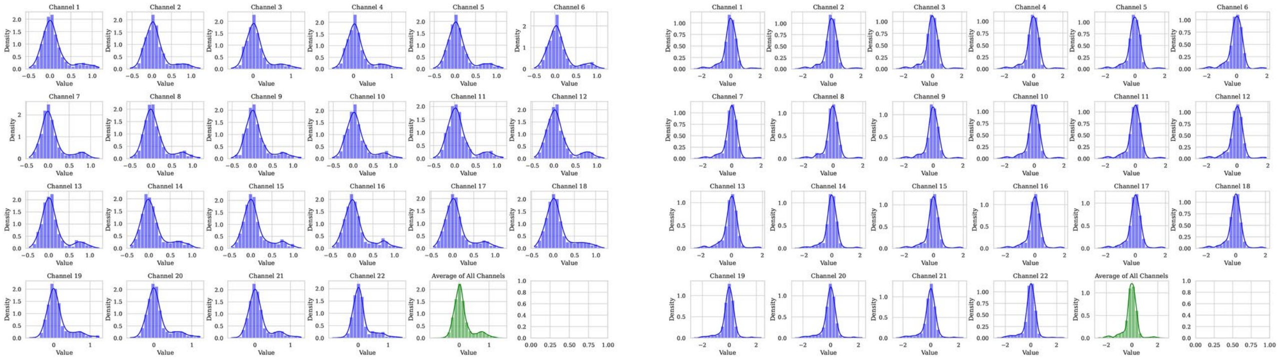


Fig. 3. The average classification accuracy under different noise levels. "FS" means the feature-aware shift operation.



(a) MSDI w/o FS

(b) MSDI w/ FS

Fig. 4. MSDI feature distribution visualization on the BNCI2014-001 dataset for subject 5. 'FS' refers to the feature-aware shift operation. By observing the horizontal axes in (**a**) and (**b**), it can be noted that (**b**), which applies FS, exhibits a wider range.

preserving local neighborhood structures, allowing for the visualization of clusters and relationships within the data[41].

Figure 5 presents the t-SNE visualization of the learned features for all nine subjects in the BNCI2014-001 dataset. Distinct clustering is evident for the four MI classes (left hand, right hand, feet, and tongue) across most subjects. For instance, subjects 1, 3, 7, and 9 show a clear separation between all four classes, indicating that the model has learned highly discriminative features for these subjects. In subjects 2, 4, and 8, while some overlap exists, particularly between feet and tongue movements, the left and right-hand classes remain well-separated. Subjects 5 and 6 exhibit more complex patterns, yet discernable clusters are still present, suggesting that the model captures some class-specific information even in these challenging cases. The well-defined clusters observed in the majority of subjects demonstrate the ability of the proposed method to extract features that effectively differentiate between the various MI tasks. This visual evidence corroborates the quantitative results presented in the classification performance tables, further supporting the efficacy of the proposed method in learning robust and discriminative representations for EEG-based MI classification.

Grad-CAM visualization

Gradient-weighted Class Activation Mapping (Grad-CAM) provides visual explanations of the model's decision-making process by highlighting the regions of the input that contribute most significantly to the predicted class[42]. Grad-CAM utilizes the gradients flowing into the final convolutional layer to produce a coarse localization map, indicating the important regions for a particular prediction. Figure 6 depicts the Grad-CAM visualizations for each subject (A1-A9) across the four MI tasks (left hand, right hand, feet, and tongue) on the BNCI2014-001 dataset. Analysis of these visualizations reveals task-specific activation patterns. In most subjects, the visualizations show that the model has effectively learned to focus on distinct brain regions.

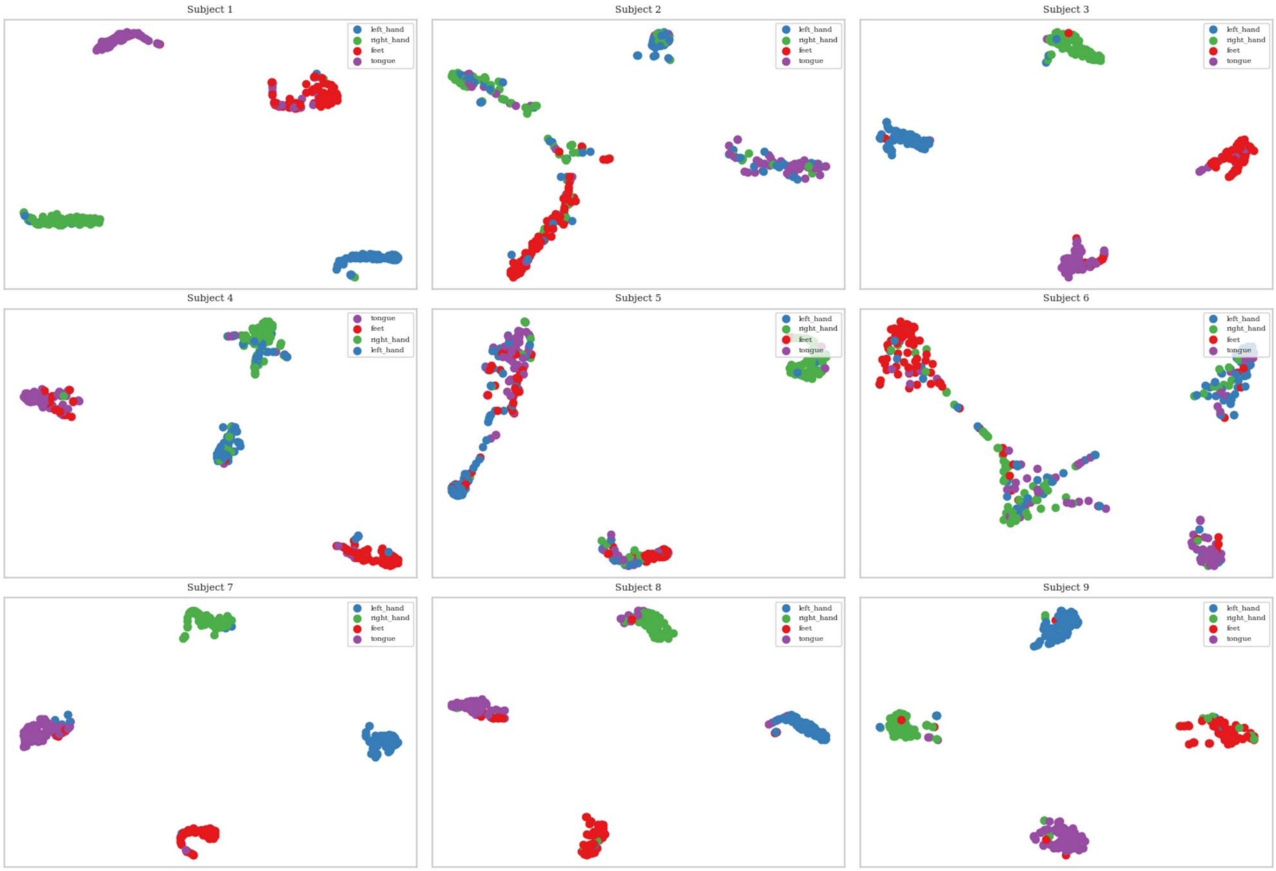


Fig. 5. The t-SNE visualization of our proposed MSDI on the BNCI2014-001 dataset.

Conclusion

In this paper, we proposed a novel multi-scale spatio-temporal domain-invariant representation learning method (MSDI) to enhance EEG-based motor imagery signal decoding. MSDI effectively captures robust features by decoupling and adaptively fusing multi-scale temporal and spatial information and mapping the representation to a domain-invariant space using a feature-aware shift operation. Our method achieved state-of-the-art performance on two public datasets, with classification accuracies of 81.06% and 89.42% on BNCI2014-001 and BNCI2014-004, respectively. This highlights the importance of robust signal representation in improving decoding accuracy and stability. Furthermore, our method demonstrates time efficiency and noise resistance, paving the way for more reliable brain-computer interfaces.

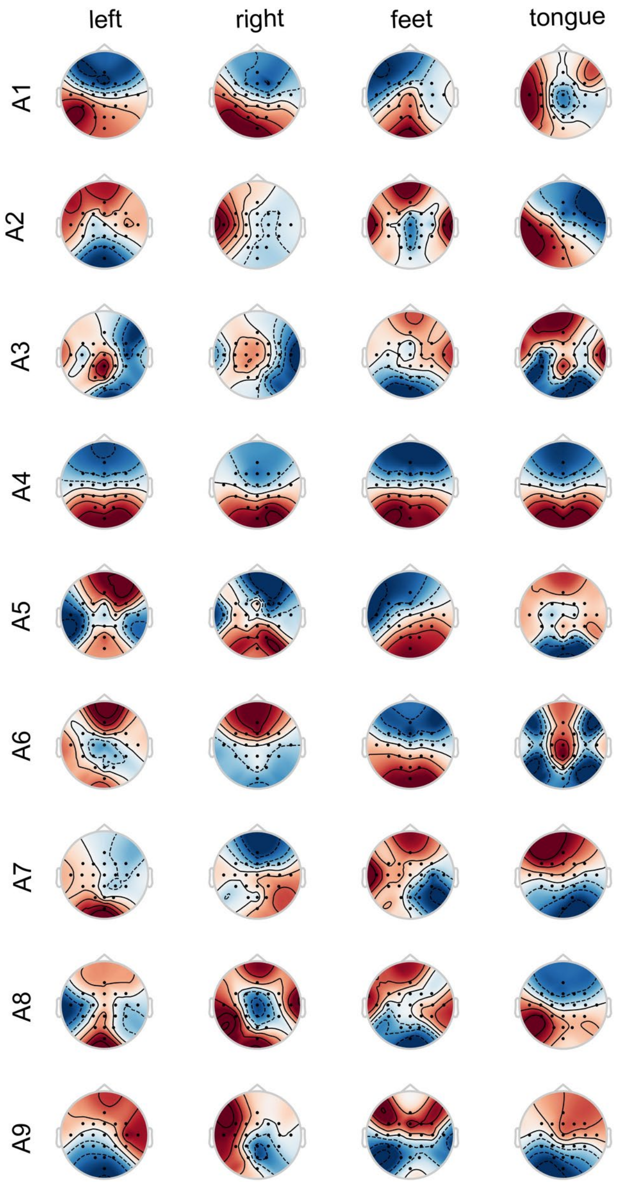


Fig. 6. The Grad-CAM visualization of our proposed MSDI on the BNCI2014-001 dataset.

Data availability

All data can be accessed on the website https://www.bbci.de/competition/iv/. The code is available upon request from the corresponding author.

Received: 2 December 2024; Accepted: 19 March 2025
Published online: 27 March 2025

References

1. Altaheri, H. et al. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Comput. Appl.* **35**, 14681–14722 (2023).
2. Mokienko, O., Chernikova, L., Frolov, A. & Bobrov, P. Motor imagery and its practical application. *Neurosci. Behav. Physiol.* **44**, 483–489 (2014).
3. Khan, M. A., Das, R., Iversen, H. K. & Puthusserypady, S. Review on motor imagery based BCI systems for upper limb post-stroke neurorehabilitation: From designing to application. *Comput. Biol. Med.* **123**, 103843 (2020).
4. Pang, M. et al. Multi-scale masked autoencoders for cross-session emotion recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* **32**, 1637–1646. https://doi.org/10.1109/TNSRE.2024.3389037 (2024).
5. Chen, C. *et al.* Comprehensive multisource learning network for cross-subject multimodal emotion recognition. In *IEEE Transactions on Emerging Topics in Computational Intelligence* 1–16 (2024). https://doi.org/10.1109/TETCI.2024.3406422.
6. Chen, C. et al. Self-attentive channel-connectivity capsule network for EEG-based driving fatigue detection. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 3152–3162. https://doi.org/10.1109/TNSRE.2023.3299156 (2023).
7. Decety, J. The neurophysiological basis of motor imagery. *Behav. Brain Res.* **77**, 45–52 (1996).

8. Gwon, D. et al. Review of public motor imagery and execution datasets in brain-computer interfaces. *Front. Hum. Neurosci.* **17**, 1134869 (2023).
9. Schirrmeister, R. T. et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **38**, 5391–5420 (2017).
10. Lawhern, V. J. et al. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **15**, 056013 (2018).
11. Zhou, B., Wu, X., Lv, Z., Zhang, L. & Guo, X. A fully automated trial selection method for optimization of motor imagery based brain–computer interface. *PLoS ONE* **11**, e0162657 (2016).
12. Lu, B., Huang, X., Chen, J., Fu, R. & Wen, G. Manifold attention-enhanced multi-domain convolutional network for decoding motor imagery intention. *Knowl. Based Syst.* **296**, 111904. https://doi.org/10.1016/j.knosys.2024.111904 (2024).
13. Liu, D., Zhang, J., Wu, H., Liu, S. & Long, J. Multi-source transfer learning for EEG classification based on domain adversarial neural network. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 218–228. https://doi.org/10.1109/TNSRE.2022.3219418 (2023).
14. Song, Y., Zheng, Q., Liu, B. & Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 710–719 (2022).
15. Ma, X., Chen, W., Pei, Z., Zhang, Y. & Chen, J. Attention-based convolutional neural network with multi-modal temporal information fusion for motor imagery EEG decoding. *Comput. Biol. Med.* 108504 (2024).
16. Ju, C. & Guan, C. Tensor-cspnet: A novel geometric deep learning framework for motor imagery classification. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 10955–10969. https://doi.org/10.1109/TNNLS.2022.3172108 (2023).
17. Ce, J. & Cuntai, G. Graph neural networks on SPD manifolds for motor imagery classification: A perspective from the time-frequency analysis. *IEEE Trans. Neural Netw. Learn. Syst.* https://doi.org/10.1109/TNNLS.2023.3307470 (2023).
18. Li, X. *et al.* Uncertainty modeling for out-of-distribution generalization. arXiv:2202.03958 (2022).
19. Wang, F., Han, Z., Gong, Y. & Yin, Y. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7151–7160 (2022).
20. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10850–10869 (2023).
21. Hong, X. et al. Dynamic joint domain adaptation network for motor imagery classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 556–565 (2021).
22. Zhong, X.-C. et al. A deep domain adaptation framework with correlation alignment for EEG-based motor imagery classification. *Comput. Biol. Med.* **163**, 107235 (2023).
23. Liu, W., Guo, C. & Gao, C. A cross-session motor imagery classification method based on Riemannian geometry and deep domain adaptation. *Expert Syst. Appl.* **237**, 121612 (2024).
24. Yin, K., Lim, E. Y. & Lee, S.-W. GITGAN: Generative inter-subject transfer for EEG motor imagery analysis. *Pattern Recogn.* **146**, 110015 (2024).
25. Han, J., Gu, X. & Lo, B. Semi-supervised contrastive learning for generalizable motor imagery EEG classification. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* 1–4 (IEEE, 2021).
26. Zhao, R. et al. A mutli-scale spatial-temporal convolutional neural network with contrastive learning for motor imagery EEG classification. *Med. Novel Technol. Dev.* **17**, 100215 (2023).
27. Li, W. et al. Self-supervised contrastive learning for EEG-based cross-subject motor imagery recognition. *J. Neural Eng.* **21**, 026038 (2024).
28. Arvaneh, M., Guan, C., Ang, K. K. & Quek, H. C. Spatially sparsed common spatial pattern to improve BCI performance. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2412–2415 (2011). https://doi.org/10.1109/ICASSP.2011.5946970.
29. Ang, K. K., Chin, Z. Y., Zhang, H. & Guan, C. Filter bank common spatial pattern (FBCSP) in brain–computer interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* 2390–2397 (2008). https://doi.org/10.1109/IJCNN.2008.4634130.
30. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50 (2013).
31. Zhong, G., Wang, L.-N., Ling, X. & Dong, J. An overview on data representation learning: From traditional feature learning to recent deep learning. *J. Finance Data Sci.* **2**, 265–278. https://doi.org/10.1016/j.jfds.2017.05.001 (2016).
32. Kommineni, A., Avramidis, K., Leahy, R. & Narayanan, S. Knowledge-guided EEG representation learning. arXiv:2403.03222 (2024).
33. Xiang, T.-Y. et al. Learning EEG motor characteristics via temporal-spatial representations. *IEEE Trans. Emerg. Top. Comput. Intell.* https://doi.org/10.1109/TETCI.2024.3425328 (2024).
34. Wang, J. & Li, M. CASCE: A contrastive representation learning framework for motor imagery EEG-based unilateral upper limb decoding. *IEEE Trans. Instrum. Meas.* https://doi.org/10.1109/TIM.2024.3500057 (2024).
35. Li, Y. & Wang, S. SCDM: Unified representation learning for EEG-to-fNIRS cross-modal generation in MI-BCIs. arXiv:2407.04736 (2024).
36. Tangermann, M. et al. Review of the BCI competition IV. *Front. Neurosci.* **6**, 55. https://doi.org/10.3389/fnins.2012.00055 (2012).
37. Leeb, R. et al. Brain–computer communication: Motivation, aim, and impact of exploring a virtual apartment. *IEEE Trans. Neural Syst. Rehabil. Eng.* **15**, 473–482. https://doi.org/10.1109/TNSRE.2007.906956 (2007).
38. Zhao, H., Zheng, Q., Ma, K., Li, H. & Zheng, Y. Deep representation-based domain adaptation for nonstationary EEG classification. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 535–545 (2020).
39. She, Q. et al. Improved domain adaptation network based on Wasserstein distance for motor imagery EEG classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 1137–1148 (2023).
40. Zhang, D., Li, H. & Xie, J. MI-CAT: A transformer-based domain adaptation network for motor imagery classification. *Neural Netw.* **165**, 451–462. https://doi.org/10.1016/j.neunet.2023.06.005 (2023).
41. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
42. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74.

Acknowledgements

This work was supported by the Inner Mongolia Natural Science Foundation project of China (No.2023QN06002) and the Inner Mongolia University high-level talent research project (JJ230408).

Author contributions

J.L. took the lead in writing the manuscript. J.S. provided the idea, designed the MSDI model, and conducted all the experiments. P.Y., X.Y., and Y.L. supervised the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.S.

Reprints and permissions information is available at www.nature.com/reprints.