


Full Length Article

MSC-transformer-based 3D-attention with knowledge distillation for multi-action classification of separate lower limbs

Heng Yan^a, Zilu Wang^b, Junhua Li^{a,b,*} 

^a Laboratory for Brain-Bionic Intelligence and Computational Neuroscience, Wuyi University, 529020 Jiangmen, China

^b School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, UK

ARTICLE INFO

Keywords:

Brain computer interface
Lower limb
Motor imagery
Transformer
Knowledge distillation

ABSTRACT

Deep learning has been extensively applied to motor imagery (MI) classification using electroencephalogram (EEG). However, most existing deep learning models do not extract features from EEG using dimension-specific attention mechanisms based on the characteristics of each dimension (e.g., spatial dimension), while effectively integrate local and global features. Furthermore, implicit information generated by the models has been ignored, leading to underutilization of essential information of EEG. Although MI classification has been relatively thoroughly investigated, the exploration of classification including real movement (RM) and motor observation (MO) is very limited, especially for separate lower limbs. To address the above problems and limitations, we proposed a multi-scale separable convolutional Transformer-based filter-spatial-temporal attention model (MSC-T3AM) to classify multiple lower limb actions. In MSC-T3AM, spatial attention, filter and temporal attention modules are embedded to allocate appropriate attention to each dimension. Multi-scale separable convolutions (MSC) are separately applied after the projections of query, key, and value in self-attention module to improve computational efficiency and classification performance. Furthermore, knowledge distillation (KD) was utilized to help model learn suitable probability distribution. The comparison results demonstrated that MSC-T3AM with online KD achieved best performance in classification accuracy, exhibiting an elevation of 2 %-19 % compared to a few counterpart models. The visualization of features extracted by MSC-T3AM with online KD reiterated the superiority of the proposed model. The ablation results showed that filter and temporal attention modules contributed most for performance improvement (improved by 2.8 %), followed by spatial attention module (1.2 %) and MSC module (1 %). Our study also suggested that online KD was better than offline KD and the case without KD. The code of MSC-T3AM is available at: <https://github.com/BICN001/MSC-T3AM>.

1. Introduction

Brain-computer interface (BCI) has attracted widespread attention in various fields, such as medicine, neuroscience, entertainment, and engineering (Shih et al., 2012; Birbaumer, 2006; Nijholt, 2008). Motor imagery (MI) is one of the most commonly used BCI paradigms (Katona & Kovari, 2016). During MI, a particular action is imagined without real movements, which mainly activates the sensorimotor cortex. This brain region is also activated by real movement (RM) (Lotze & Halsband, 2006). This activation can also be observed during motor observation (MO) (Kraeutner et al., 2014). It is worth noting that the majority of such studies focus on the upper-limbs and consider both lower limbs together (Babiloni et al., 2001), leaving very limited exploration of the separate lower limbs (Tangemann et al., 2012). In order to recognize MI tasks,

physiological signals, such as electroencephalogram (EEG) (Liu et al., 2019), electrocardiogram (ECG) (Pinto et al., 2017), and functional near-infrared spectroscopy (fNIRS) (Ambi et al., 2020), are commonly employed. Among them, EEG is a good choice due to its high temporal resolution, relatively low cost and convenience in use.

Due to the noisy nature of EEG signals, feature extraction is required. For the EEG feature extraction, traditional methods, such as common spatial patterns (CSP) (Lotte & Guan, 2011) and filter bank common spatial patterns (FBCSP) (Ang et al., 2008), have been proposed to extract spatial features. Meanwhile, the methods for extracting spectral power features, such as continuous wavelet transform (CWT) (Bostanov, 2017), have been used. In addition, a method, named filter bank common spatio-spectral patterns (FBCSSP), was proposed to optimize spatial and spectral filters (Yuksel & Olmez, 2016). The extracted features are

* Corresponding author.

E-mail address: junhua.li@essex.ac.uk (J. Li).

<https://doi.org/10.1016/j.neunet.2025.107806>

Received 16 September 2024; Received in revised form 22 May 2025; Accepted 24 June 2025

Available online 25 June 2025

0893-6080/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

classified by a classifier, such as k-nearest neighbor (k-NN) (Md Isa et al., 2017), random forest (RF) (Steyrl et al., 2014), and support vector machine (SVM) (Ma et al., 2016) to recognize MI tasks. Similarly, Ramkumar et al. improved classification accuracy by combining CSP with a tunable optimized feed-forward neural network (FFNN) classifier (Ramkumar & Paulraj, 2025). These traditional methods are incapable of capturing highly abstract and high-level features, and also suffer from the limitations of manual feature engineering (Liu et al., 2020).

Deep learning solved the above problem and merged feature extraction and the classifier together to form a single model. The most commonly used deep learning models for EEG classification are convolutional neural networks (CNN) models, such as DeepConvNet (Schirmer et al., 2017) and EEGNet (Lawhern et al., 2017). Moreover, recurrent neural networks (RNN) models are often utilized for EEG classification, such as long short-term memory (LSTM) (Tortora et al., 2020). However, these deep learning models still have limitations. CNN usually extracts spatial correlations and local structures through local receptive fields and layer-by-layer convolution, but may not adequately capture global features when processing EEG signals (Xie et al., 2022). On the other hand, RNN is capable of processing time-series data for global features, but has limited ability to establish reliable long-term dependencies and capture local features (Xia et al., 2018).

A better option might be the attention-based model, which shifts attention to essential features and allocates a higher weight to critical information contained in EEG. LMDA-Net was proposed to extract filter-channel features efficiently through the depth attention module and the channel attention module (Miao et al., 2023). Another model, called ATCNet, combines multi-head self-attention with the temporal convolutional network to efficiently extract features from EEG signals (Altafheri et al., 2022). Although these models achieve good performance by shifting attention to a particular dimension, the information conveyed in the other dimensions is ignored. Since every dimension in EEG signals contains unique information, equal attention to each dimension may cause loss of such information. Therefore, different processing operations should be applied to different dimensions (Burle et al., 2015).

Another approach that has received significant attention is the Transformer model, which has made breakthroughs in natural language processing and computer vision. EEG-Transformer successively acquires the spatial-temporal features through channel-Transformer and time-Transformer (Song et al., 2021). Furthermore, a Transformer-based model, called EEG-Conformer, combines the advantages of CNN and Transformer to extract local and global feature representations (Song et al., 2022). However, Transformer model lacks the capability to handle computational complexity and memory consumption when dealing with sequential data (Yang et al., 2023). In addition, the existing Transformer model does not adequately extract the information from the other dimensions.

The deep learning models used for EEG classification utilize explicit features extracted from EEG, but they do not utilize implicit information, such as the probability distribution of the model outputs, which expresses the correlation between classes (Huang et al., 2022). To utilize such implicit information, knowledge distillation (KD) has been proposed (Hinton et al., 2015). KD makes a model (called as student model) to learn knowledge from the other model (called as teacher model) in training process.

KD is mainly categorized into response-based KD and feature-based KD due to different emphases on knowledge. Zhang et al. think the output probability distribution of the teacher model contains relationships between categories as knowledge (Zhang et al., 2020). In another study, Adriana et al. considered intermediate representations as knowledge and used them by directly matching the features of teacher model and student model (Adriana et al., 2015). In this study, we opt for response-based KD because it has less computational cost during the training process compared to feature-based KD (Zhao et al., 2022). In addition, intermediate-level features do not directly correspond to each

other in our proposed model, which focuses on relatively different information and has a different architecture compared to the CNN-based model (Liu et al., 2022). Moreover, due to the non-stationary characteristics of EEG signals, a suitable probability distribution that accepts implicit information is conducive to performance improvement. Specifically, KD is generally divided into offline KD and online KD (Gou et al., 2021). Offline KD transfers knowledge from pre-trained teacher model to student model during training. Online KD transfers knowledge while the teacher model and the student model are being trained simultaneously. We, in this study, used the output probability distribution of the teacher model as knowledge in the training process, which provides much more information than the ground-truth label (Hinton et al., 2015). However, it does not mean that can completely replace the ground-truth label.

It is beneficial to apply KD to BCI decoding models. Some attempts have been done. For instance, KD can reduce computational load of a large-scale model by transferring knowledge from a teacher model in a sleep stage classification task (Liang et al., 2023). Other researchers have facilitated multi-step model training by KD in an emotion recognition task (Wang et al., 2021). In addition, Sakhavi et al. used KD to reduce calibration time and improve classification accuracy (Sakhavi & Guan, 2017). However, the majority of the existing studies have employed offline KD. Moreover, the effect of KD on the classification of movement-related EEG, particularly EEG associated with lower limb movements, remains largely unknown.

To address these challenges, we proposed multi-scale separable convolutional Transformer-based filter-spatial-temporal attention model and introduced KD to utilize explicit and implicit features in the training. Meanwhile, we reduced the computational cost of the MSC-Transformer blocks by adopting a multi-scale depth-wise convolution.

In the rest of this paper, we first introduced the details of the model in Section II. The comparison results were described in Section III, where we performed ablation experiments and compared the performance of the models. The feature visualization was presented in Section IV to explain the feature extraction of the model. Finally, conclusions and future work were drawn in Section V.

2. Method

In this study, we proposed a multi-scale separable convolutional Transformer-based attention model incorporating three-dimensional (i. e., filter, spatial, and temporal) attention module (MSC-T3AM, see Fig. 1 for the model architecture). By integrating 3D-attention and MSC-Transformer synergistically, our framework effectively captures the features of EEG signals from diverse dimensions while gaining local and global features. Additionally, we utilized knowledge distillation to learn the implicit information between categories, further enhancing the classification performance.

In the following sections, we detailed each block of the MSC-T3AM and elaborated on the offline and online KD in the process of training. As depicted in Fig. 1, the model architecture encompasses three principal components: the 3D-attention block, the MSC-Transformer blocks, and the classifier block. The details are as follows:

2.1. 3D-attention block of MSC-T3AM

2.1.1. Spatial attention module

Let $X^{(1,C,T)}$ represent the preprocessed EEG data sample, and let $W^{(F,1,C)}$ stand for an adaptive tensor. In this context, F , C , and T denote the number of filters, the number of EEG channels, and the number of EEG time points of a sample, respectively. The channel weighted tensor $X^{(F,C,T)}$ is calculated by the product of $X^{(1,C,T)}$ and $W^{(F,1,C)}$. After obtaining $X^{(F,C,T)}$, a spatial convolution is applied to the channel weighted tensor to extract features along the spatial dimension. The resulting output is fed into the filter and temporal attention modules, which dynamically

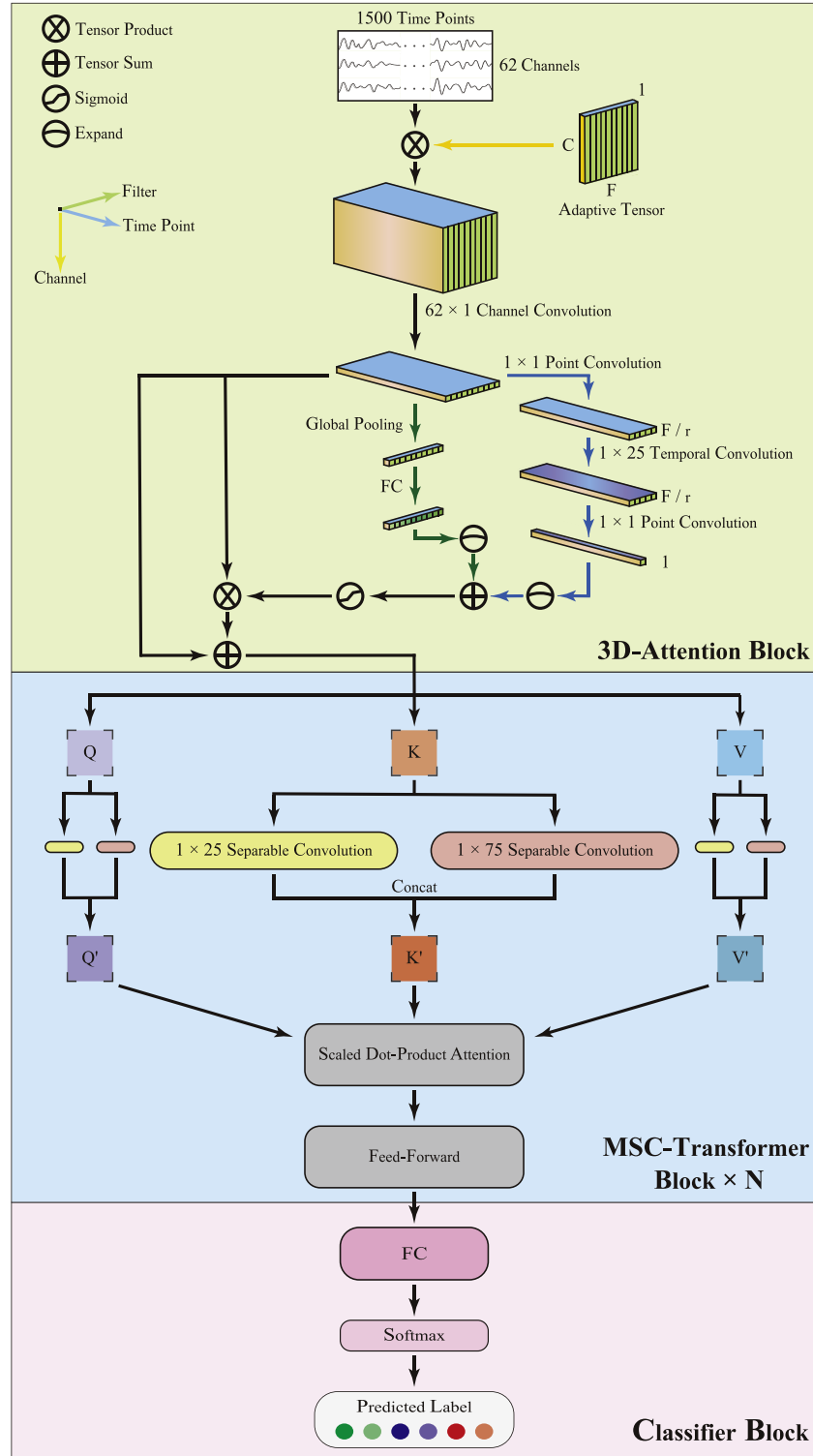


Fig. 1. The architecture of MSC-T3AM. MSC-T3AM comprises of 3D-attention block, MSC-Transformer block, and classifier block. 3D-attention block is designed to extract local features along the three dimensions (i.e., filter, spatial, and temporal). MSC-Transformer block is designed to get global feature.

regulate the weights of information across both the filter and temporal dimensions.

2.1.2. Filter attention module

In the filter attention module, global average pooling is applied to the filter dimension, yielding filter tensor denoted as $X^{(F,1,1)}$, which represents the global information within each filter. Subsequently, filter weights are computed across the filters based on the filter tensor through

a multi-layer perceptron (MLP). The hidden activation size is set to $X^{(F/r,1,1)}$, wherein the reduction ratio r affects the consumption of attention structure, we set r to 2.

2.1.3. Temporal attention module

As shown in the right part of the 3D-attention block in Fig. 1, the temporal weights are obtained through three layers of convolution. That is, the input tensor $X^{(F,1,T)}$ is first projected into the lower-dimensional

space by a point convolution layer (the output from this process is indicated by $X^{(F/r,1,T)}$). This is followed by a temporal convolution with a kernel of (1×25) , which is applied to extract the local temporal information. After that, one more point convolution is used to further reduce its dimensionality to obtain the temporal weight $X^{(1,1,T)}$.

The temporal weight obtained from the temporal attention module and the filter weight obtained from the filter attention module are expanded to match the size of the input $X^{(F,1,T)}$, and then are aggregated by adding them up. A sigmoid function is then applied to generate the combined filter and temporal attention. This combined filter and temporal attention are multiplied by the input $X^{(F,1,T)}$ and then added to itself. We designed such 3D-attention block to enable the model to shift the attention to dominant features while retaining the original input.

2.2. MSC-Transformer blocks of MSC-T3AM

We utilized three MSC-Transformer blocks to obtain global features. In this block, to reduce the computation of self-attention module and improve the classification performance, multi-scale separable convolution (MSC) which is employed to test the responses of long and short temporal sequences by setting different convolution kernel scales, while following the projections of query Q, key K, and value V, respectively. Q, K, and V are then entered into the Scaled-Dot-Product-Attention shown in Fig. 2, which is followed by a feed-forward layer.

2.3. Classifier block of MSC-T3AM

The output of MSC-Transformer blocks is input into the classification block, which consists of a fully-connected layer and softmax layer for label prediction. The prediction of MSC-T3AM is obtained from the fully connected layer, which represents the final output of the classification block.

2.4. Model training with knowledge distillation

We chose EEGNet as the teacher model for knowledge distillation (KD) during the training as shown in Fig. 3. This is because (1) EEGNet can be quickly fitted to help the student model generate suitable probability distribution, which is especially important for the online KD; (2) EEGNet has relatively better performance in the EEG classification so that the false positive rate could be lower to minimize the misleading

issue (Lawhern et al.); and (3) MSC-T3AM is a Transformer-based model that has advantages in global feature learning, but it does not have the property of spatial invariance that a convolutional architecture would have. Therefore, we intend to incorporate the benefits from both of them by using EEGNet as the teacher model in the model training with KD.

The loss of the teacher model was calculated by the difference between the teacher model's prediction and the ground-truth label in the case of online KD. However, the teacher model's loss has nothing to do with the backpropagation during the student model training in the case of offline KD. During the student model training (both offline and online KDs), the loss consists of two parts: the loss calculated by the difference between the student model's prediction and the ground-truth label and the loss calculated by the difference between the student model's prediction and the teacher model's prediction. Specifically, the former loss ensures that the student model is supervised by the ground-truth label, while the latter loss enforces the student model to utilize implicit information of the probability distribution of the teacher model's prediction.

2.5. Loss of knowledge distillation

In this study, we used cross-entropy (CE) loss and Kullback-Leibler (KL) divergence loss to optimize the student model in training. KL divergence loss is a standard response-based KD loss that minimizes the divergence between the probability distributions of the student model and the teacher model. The main distinction between CE loss and KL divergence loss is that CE loss focuses only on the difference in target category, while KL divergence loss emphasizes the difference across all categories.

The loss of the student model in KD (offline and online) comprises CE loss and KL divergence loss (See Eq. (1)). The coefficient α is set to adjust the ratio of CE loss (\mathcal{L}_{CE} , see Eq. (2)) to KL divergence loss (\mathcal{L}_{KL} , see Eq. (3)). For the teacher model, \mathcal{L}_{CE} is adopted in online KD. However, the loss is not included in offline KD because the teacher model is trained in advance before the training of the student model.

$$\mathcal{L} = \alpha \mathcal{L}_{CE} + (1 - \alpha) \mathcal{L}_{KL} \quad (1)$$

$$\mathcal{L}_{CE}(p, y) = - \sum_{i=1}^n y_i \log(p(x_i)) \quad (2)$$

$$\mathcal{L}_{KL}(p^s, p^t) = \sum_{i=1}^n p^t(x_i / T) \log \left(\frac{p^t(x_i / T)}{p^s(x_i / T)} \right) \quad (3)$$

We denote $x, y, n, f^t, f^s, p^t, p^s$, and T as the sample, the ground-truth label, the number of samples, the teacher model, the student model, the softened probability vector of the teacher model, the softened probability vector of the student model, and a hyperparameter that adjusts the effect of distillation, respectively.

To prevent \mathcal{L}_{KL} return negative numbers, we follow the official documentation of PyTorch to take logarithms of p^s , which does not affect the loss when comparing p^t (See Eq. (4)) and p^s (See Eq. (5)).

$$p^t(x_i / T) = \frac{\exp(f^t(x_i) / T)}{\sum_{j=0}^{n-1} \exp(f^t(x_j) / T)} \quad (4)$$

$$p^s(x_i / T) = \log \left(\frac{\exp(f^s(x_i) / T)}{\sum_{j=0}^{n-1} \exp(f^s(x_j) / T)} \right) \quad (5)$$

T as a hyperparameter affects the smoothing of probability distributions to make the probability distributions of the teacher model and student model closer in form, which causes the performance of learning implicit information. Referring to the existing literature, we have chosen $T = 1.5$ for the results of this paper (Hurlé et al., 2021). Moreover, we present Algorithm 1 to thoroughly illustrate the details of the training procedure for offline KD and online KD due to the differences between

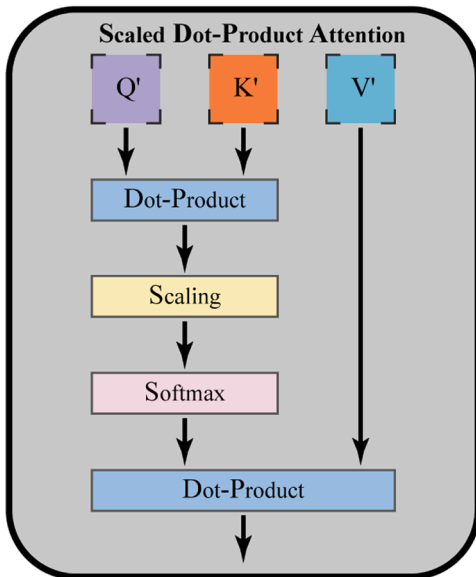


Fig. 2. Scaled dot-product attention for calculating Q, K, and V of each head in the multi-head attention.

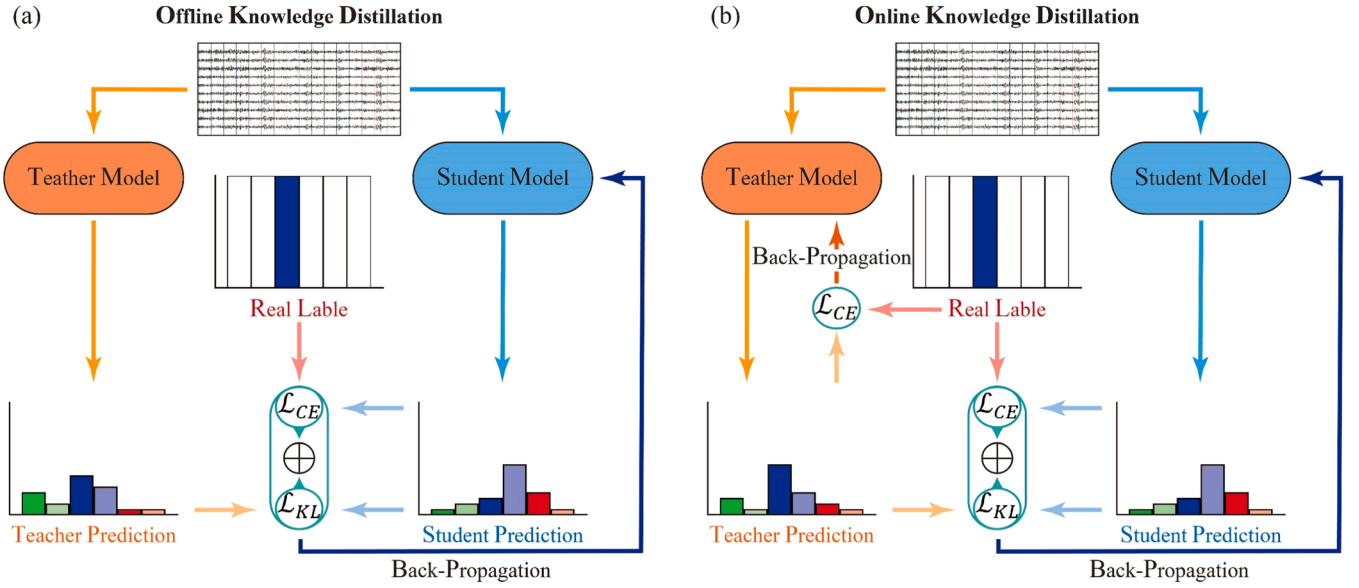


Fig. 3. Illustration of (a) offline and (b) online knowledge distillation (KD). The teacher model of KD participates in the student model's training, which differs between the training without KD and with KD. The left diagram shows offline KD, in which the teacher model is pre-trained and then provides guidance information for the training of the student model. The right diagram shows online KD, in which the teacher model is trained while providing guidance information for the training of the student model.

Algorithm 1

Model training.

Require: Sample pair $\{x_i, y_i\}_{i=1}^n$; Max number of epochs $e = 200$; The teacher model f^t and the student model f^s ; Hyperparameter T' .

For $epoch \in e$ **do**

 Softened probability vector p^t and p^s was calculated by using Eq. (4) and Eq. (5);

If online KD is True **then**

 The \mathcal{L}_{CE} of teacher model was calculated by using Eq. (2);

 The \mathcal{L}_{CE} and \mathcal{L}_{KL} of student model was calculated by using Eq. (2) and Eq. (3).

Else

 The \mathcal{L}_{CE} and \mathcal{L}_{KL} of student model was calculated by using Eq. (2) and Eq. (3).

End If

 The final loss of student model was calculated by using Eq. (1);

If online KD is True **then**

 Update f^t towards minimizing \mathcal{L}_{CE} of teacher model;

 Update f^s towards minimizing final loss of student model.

Else

 Update f^s towards minimizing final loss of student model.

End If

End For

offline KD and online KD.

3. Dataset and results

3.1. Dataset and experiment details

The dataset used in this study contains 28 subjects, who performed six actions using their lower limbs, which are MI of left lower limb (MI-L), MI of right lower limb (MI-R), RM of left lower limb (RM-L), RM of right lower limb (RM-R), MO of left lower limb (MO-L), and MO of right lower limb (MO-R). There are 72 trials (i.e., 72 samples) for each motor action (a total of $72 \times 6 = 432$ samples). 62 channels of EEG signals and two pairs of electrooculogram (EOG) signals were recorded and down-sampled to 250 Hz. Further information about the dataset can be found in Wang et al. (2022). The experiment protocol for this dataset recording was reviewed and approved by the Institutional Review Board of the National University of Singapore, and the Humanities, Science and Health, or Social Science Ethics Sub-Committee at the University of Essex. The consent form was signed by each subject before the experiment.

Our proposed model is implemented using the PyTorch library in Python 3.10, utilizing a NVIDIA GeForce GTX 1050 Ti. We trained all models using Adam optimizer with the learning rate of 0.01 in five-fold cross-validation scheme. Additionally, we set the batch size and epoch to 16 and 200, respectively. Meanwhile, we applied bandpass filter to narrow down the frequency band to be 8 Hz - 30 Hz and then normalized the data.

3.2. Results of model comparison

We compared the proposed MSC-T3AM with a few representative methods. One of them is EEG Conformer, which achieved good performance in EEG classification and is related to CNN and Transformer. Because ATCNet and LMDA-Net performed well through attention, we included them in the comparison. We also included FBCSP-SVM as a classic traditional approach for MI classification. Moreover, we included DeepConvNet and EEGNet, which exhibited remarkable results on many EEG datasets. Meanwhile, to show the effect of KD in the classification, we added our proposed model MSC-T3AM with KD (i.e., offline KD and online KD) to compare MSC-T3AM without KD.

The comparison results are shown in Fig. 4. All results were obtained by averaging across all subjects. The detailed classification accuracies for each subject are listed in Table 1, and the comparison results of the other evaluation metrics (averaged across all subjects) are shown in Table 2. From Fig. 4 and Table 2, we can see that MSC-T3AM with online KD is better than all the compared methods. According to the statistical results (using *t*-test), the better classification performance achieved by the proposed MSC-T3AM with online KD is statistically significant (all $p < 0.01$).

When we look into each category of the proposed model's classification, the average classification accuracy of MI is less than that of RM and MO (See the confusion matrix in Fig. 5). MI is relatively more likely to be misclassified as RM or MO while RM and MO are less misclassified as each other, which leads to a lower classification accuracy in MI.

3.3. Ablation study

To demonstrate the role of the key modules in MSC-T3AM, we conducted an ablation study. The results are shown in Table 3. A check mark indicates the inclusion of that spatial attention module, filter and temporal attention modules, or MSC module.

From Table 3, we can see that filter and temporal attention modules have the most significant impact on MSC-T3AM performance (2.8 % decrease in accuracy after removal), followed by MSC module (1.2 % decrease in accuracy after removal), and the spatial attention module (1 % decrease in accuracy after removal).

In the MSC module, the number of kernel and the size of the kernel can be adapted. We tested different numbers of kernels with different sizes. We adopted the kernel size (1×25), which was used in the temporal attention module, as the base kernel size. When more than one kernel was used, the kernel sizes were the multiple of this base kernel size, such as (1×50). The results are shown in Table 4. We found that the best performance was achieved when two kernels with the kernel sizes of (1×25) and (1×75) were used.

To assess the impact of the MSC-Transformer block and different training procedures, we explored the distances between category centers to investigate their effects. We did not use scatterplot because information could be lost when mapping a high-dimensional data into the low-dimensional scatterplot (Sips et al., 2009).

After obtaining the total samples comprising all subjects, we took the samples of each category and then averaged them to determine the centers of each category. Subsequently, we calculated the Euclidean

distance between category centers to measure the distance between the categories.

As depicted in Fig. 6, we can see that the distances between the category centers are nearer in the case without MSC-Transformer block (lighter blue) compared to the case with MSC-Transformer block (darker blue). It means that all category centers are more separated when MSC-Transformer block is adopted in MSC-T3AM. Specifically, the category centers of RM (RM-L and RM-R) are dramatically distant from the respective category centers of MI and MO in the case with MSC-Transformer block.

We introduce implicit information from CNN-based model through KD. The results, as shown in Fig. 7, reveal that both offline and online KD improve model performance. Moreover, since online KD can dynamically update both the teacher model and the implicit information from teacher model in real-time according to the task—aligning more closely with the practical applications of EEG signals, online KD outperforms offline KD in enhancing model performance.

In the offline KD, the teacher model has been trained before starting the training of the student model. Therefore, the teacher model is superior to the student model at the initial stage. The guidance from the teacher model can aid the training of the student model. In contrast, the teacher model and student model are simultaneously trained in the online KD. This ensures that the teacher model is progressively updated during training and helps the student model acquire probability distribution from the most recently updated teacher model. The advantage of the online KD is the dynamic training, in which the models can be adapted quickly to meet task requirements.

KD further increases the distance between category centers (see Fig. 8). In particular, the category centers of RM are far away from the category centers of MI and MO in the case with KD. Moreover, the distance between the category centers of MI-L and RM is farther than that of MI-R and RM in the case with offline KD. Additionally, the distances between the category centers of RM-L and MO-L are farther than that between RM-L and MO-R. These results indicate differences between the left and the right lower limb within the same motor category (e.g., MI), which reflects the feature representation is separable not only among motor categories but also between the left lower limb and the right lower limb.

4. Visualization

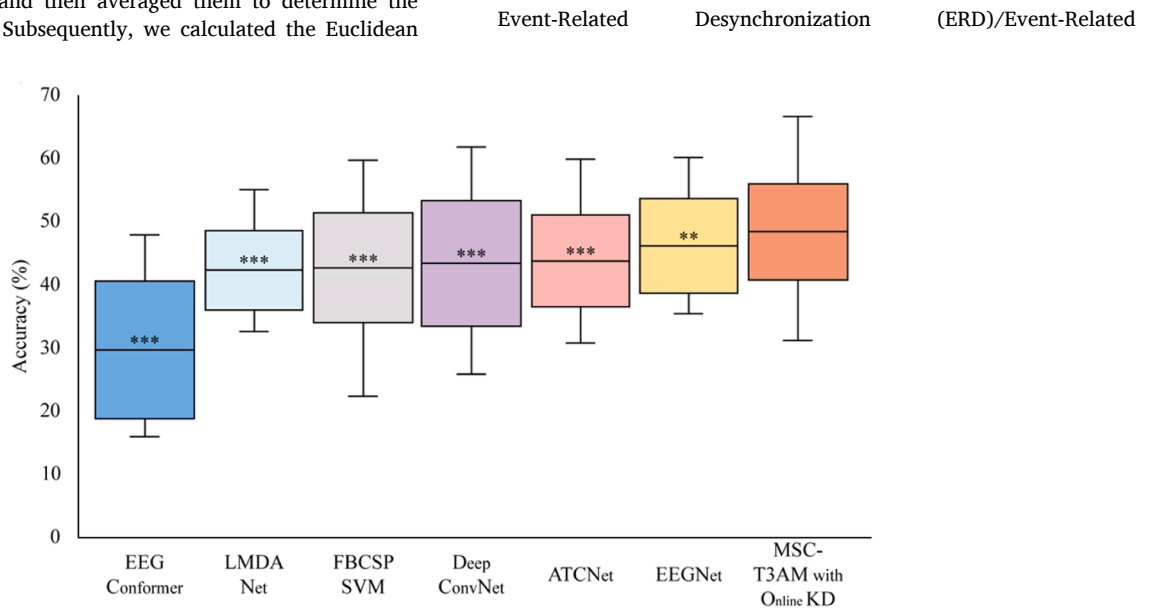


Fig. 4. Classification accuracies of the proposed model and the compared models. The box boundaries represent the first and third quartiles, with the central horizontal line denoting the median. Whiskers indicate the minimum and maximum values excluding any outliers. Asterisks denote the statistical significance levels (** $p < 0.01$ and *** $p < 0.001$).

Table 1

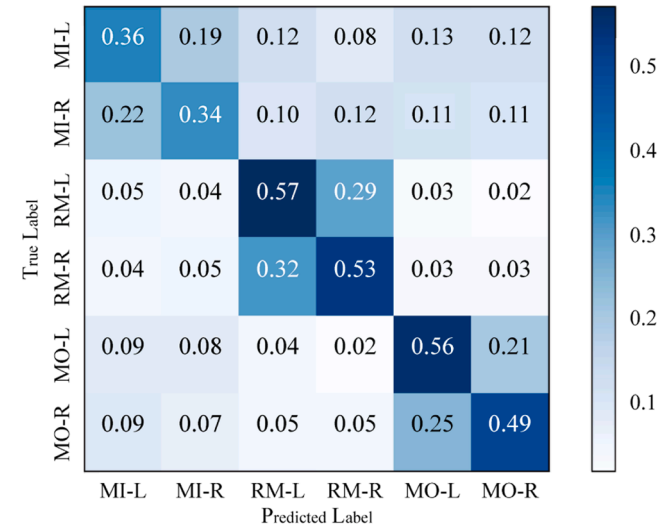
The results of model comparison in terms of accuracy.

Method Subject	EEG Conformer	LMDA- Net	FBCSP- SVM	Deep ConvNet	ATCNet	EEGNet	MSC- T3AM	MSC- T3AM with Offline KD	MSC- T3AM with Online KD
S1	36.98 %	51.62 %	54.87 %	58.81 %	54.39 %	60.18 %	54.16 %	54.18 %	57.89 %
S2	22.24 %	37.96 %	49.98 %	35.42 %	40.52 %	44.90 %	44.90 %	47.70 %	44.92 %
S3	42.81 %	34.92 %	38.63 %	38.66 %	41.67 %	44.20 %	42.82 %	42.12 %	41.18 %
S4	45.07 %	55.09 %	46.29 %	59.48 %	55.79 %	56.02 %	58.78 %	58.31 %	60.16 %
S5	16.66 %	32.66 %	43.74 %	32.88 %	33.78 %	35.89 %	40.96 %	40.76 %	40.28 %
S6	16.19 %	39.55 %	36.80 %	41.20 %	40.26 %	41.19 %	45.58 %	45.80 %	47.65 %
S7	37.70 %	42.12 %	33.34 %	41.67 %	40.73 %	41.43 %	44.20 %	44.21 %	43.73 %
S8	24.70 %	43.06 %	41.89 %	44.42 %	39.34 %	46.53 %	43.06 %	41.20 %	44.43 %
S9	27.70 %	40.48 %	35.42 %	42.10 %	42.59 %	44.41 %	47.44 %	47.89 %	48.84 %
S10	29.13 %	46.06 %	46.77 %	42.57 %	46.07 %	47.23 %	47.92 %	51.38 %	51.37 %
S11	36.78 %	43.28 %	45.62 %	43.52 %	46.76 %	44.89 %	46.08 %	46.07 %	46.29 %
S12	34.69 %	38.43 %	50.22 %	48.83 %	43.05 %	42.58 %	44.89 %	44.43 %	49.28 %
S13	17.35 %	35.41 %	22.44 %	29.40 %	34.48 %	35.43 %	42.12 %	42.59 %	42.12 %
S14	47.22 %	52.54 %	46.51 %	46.30 %	54.17 %	56.01 %	56.71 %	55.33 %	55.07 %
S15	37.18 %	49.75 %	47.45 %	49.29 %	48.38 %	54.84 %	53.22 %	52.30 %	51.61 %
S16	30.28 %	48.16 %	48.61 %	58.82 %	59.97 %	60.19 %	60.20 %	64.82 %	66.68 %
S17	30.51 %	41.21 %	37.50 %	35.16 %	39.11 %	49.54 %	44.91 %	44.68 %	42.36 %
S18	16.43 %	34.25 %	24.55 %	28.24 %	30.80 %	36.12 %	41.43 %	41.19 %	42.58 %
S19	15.96 %	39.36 %	27.31 %	25.92 %	40.54 %	36.33 %	28.48 %	29.40 %	31.25 %
S20	26.38 %	46.29 %	48.65 %	48.85 %	45.37 %	46.76 %	46.31 %	44.68 %	47.23 %
S21	17.34 %	33.09 %	39.14 %	33.33 %	33.09 %	37.72 %	39.82 %	40.98 %	39.58 %
S22	47.92 %	46.53 %	51.16 %	49.31 %	44.89 %	51.14 %	53.46 %	53.23 %	52.99 %
S23	31.23 %	35.43 %	47.22 %	35.65 %	38.43 %	38.42 %	43.29 %	45.15 %	47.69 %
S24	16.66 %	35.86 %	39.35 %	42.34 %	40.50 %	41.42 %	46.74 %	48.83 %	47.21 %
S25	43.71 %	50.45 %	59.74 %	61.82 %	55.32 %	59.04 %	60.88 %	63.87 %	63.67 %
S26	22.40 %	42.11 %	43.04 %	55.54 %	50.01 %	46.76 %	47.46 %	48.85 %	49.52 %
S27	43.76 %	46.52 %	47.70 %	51.36 %	44.67 %	48.84 %	53.92 %	50.93 %	51.84 %
S28	16.43 %	43.27 %	42.59 %	35.17 %	42.11 %	46.31 %	48.61 %	47.20 %	48.14 %
Mean	29.70 %	42.34 %	42.73 %	43.43 %	43.81 %	46.23 %	47.44 %	47.79 %	48.41 %
STD	10.90 %	6.28 %	8.67 %	9.92 %	7.30 %	7.81 %	7.03 %	7.37 %	7.58 %

Table 2

The results of model comparison in terms of precision, recall, F1 score, specificity, and kappa.

Method Metrics	EEG Conformer	LMDA- Net	FBCSP- SVM	Deep ConvNet	ATCNet	EEGNet	MSC- T3AM	MSC- T3AM with Offline KD	MSC- T3AM with Online KD
Precision	32.42 %	42.17 %	42.59 %	43.31 %	43.58 %	45.93 %	47.12 %	47.46 %	48.09 %
Recall	29.72 %	42.34 %	42.64 %	43.44 %	43.81 %	46.23 %	47.44 %	47.79 %	48.42 %
Specificity	85.94 %	88.46 %	88.53 %	88.69 %	88.76 %	89.25 %	89.48 %	89.55 %	89.68 %
F1 Score	0.3101	0.4225	0.4261	0.4337	0.4369	0.4608	0.4728	0.4762	0.4825
Kappa	0.1566	0.3081	0.3117	0.3212	0.3258	0.3547	0.3693	0.3735	0.3810

**Fig. 5.** Confusion matrix of the MSC-T3AM with online KD. The numbers are the accuracies averaged across all subjects.**Table 3**

The effect of each module of MSC-T3AM on classification performance.

Mean \pm STD	Spatial Attention	Filter and Temporal Attention	MSC
39.44 \pm 9.36 %			
43.45 \pm 9.50 %	✓		
44.13 \pm 8.93 %		✓	
41.38 \pm 9.29 %			✓
46.41 \pm 6.57 %		✓	✓
44.62 \pm 6.97 %	✓		✓
46.29 \pm 7.02 %	✓	✓	
47.44 \pm 7.03 %	✓	✓	✓

Table 4

The effect of different kernel settings of the MSC module on classification performance.

	One Kernel	Two Kernels	Three Kernels
Kernel Size	(1, 25)	(1, 25) (1, 75)	(1, 25) (1, 50) (1, 75)
Mean \pm STD	45.31 \pm 6.94 %	47.44 \pm 7.03 %	46.39 \pm 7.81 %

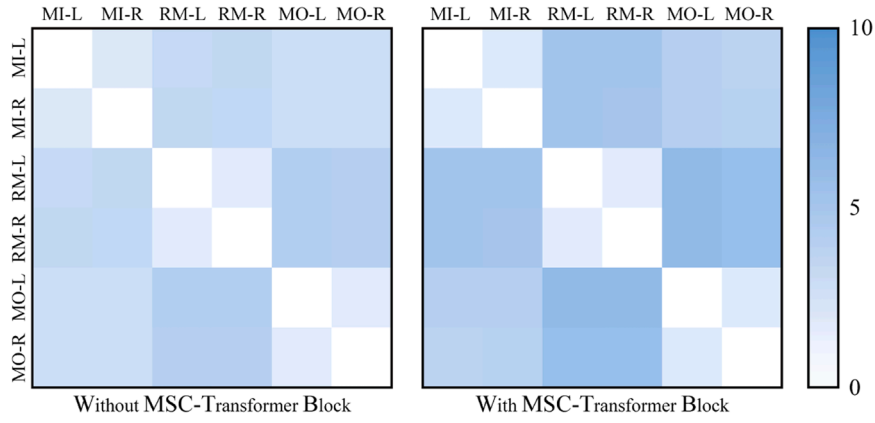


Fig. 6. The distances between classes in the cases without MSC-Transformer block and with MSC-Transformer block. A lighter color indicates the closer distance between the classes while a darker color indicates the more separate between the classes.

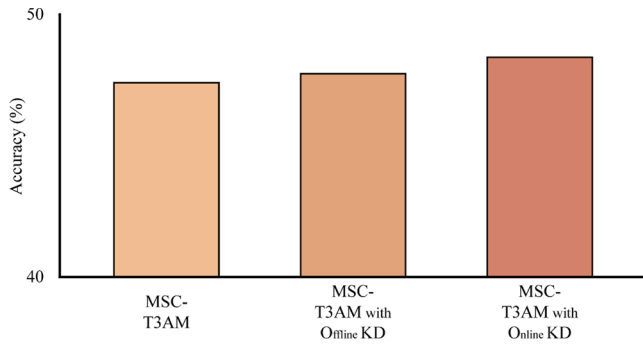


Fig. 7. Classification accuracies of the cases without KD, with offline KD, and with online KD.

Synchronization (ERS) can be interpreted as activation/deactivation of brain regions to identify the regions associated with motor tasks (Pfurtscheller, 2001). If a machine learning model can focus on the relevant brain regions to extract features, it should be good for the classification.

To identify the relevant brain regions associated with six motor actions (i.e., motor imagery for separate left or right lower limb, real movement for separate left or right lower limb, and movement observation for separate left or right lower limb), we followed the EEG mapping approach proposed by Z. Miao et al. (Miao et al., 2023) to draw brain topographies. We illustrated the relevant brain regions using the case of MSC-T3AM with online KD because its classification

performance is better than that of MSC-T3AM without KD and with offline KD.

We used all samples that were correctly classified by MSC-T3AM with online KD at the confidence level of 95 % (Subject 16 and Subject 24 were used for example subjects in the illustration) as a sample pool. After compiling these samples, we sorted them separately for each of the six motor actions, and only the top 20 % samples with the highest confidence level were retained. Then, top 10 maximum values in the temporal dimension for each top sample were averaged and then drawn brain topographies as shown in Fig. 9.

Fig. 9 shows that the relevant brain regions are positioned around the midline and these regions are not identical for different motor categories (i.e., MI, RM, and MO). Specifically, the brain regions associated with RM tend towards the frontal area compared to that of MI. On the other hand, the brain regions associated with MO tend towards the occipital area compared to that of MI. These results are consistent with the findings in Batula et al. (2017); Taube et al. (2015). These observations also help explain why RM and MO are less misclassified from each other (see the confusion matrix in Fig. 5) and why the category centers of RM and MO are furthest away from each other (see Figs. 6 and 8).

We explored the features extracted by the spatial attention module of MSC-T3AM with online KD. These features extracted were then averaged across all samples of a subject and subsequently visualized to demonstrate the spatial distribution (See the right column of Fig. 10). Furthermore, we also visualized the pre-processed EEG data, which was averaged across all samples of a subject, in the left column of Fig. 10. We observed that the spatial attention module of MSC-T3AM with online KD effectively focuses attention on the central regions of the brain,

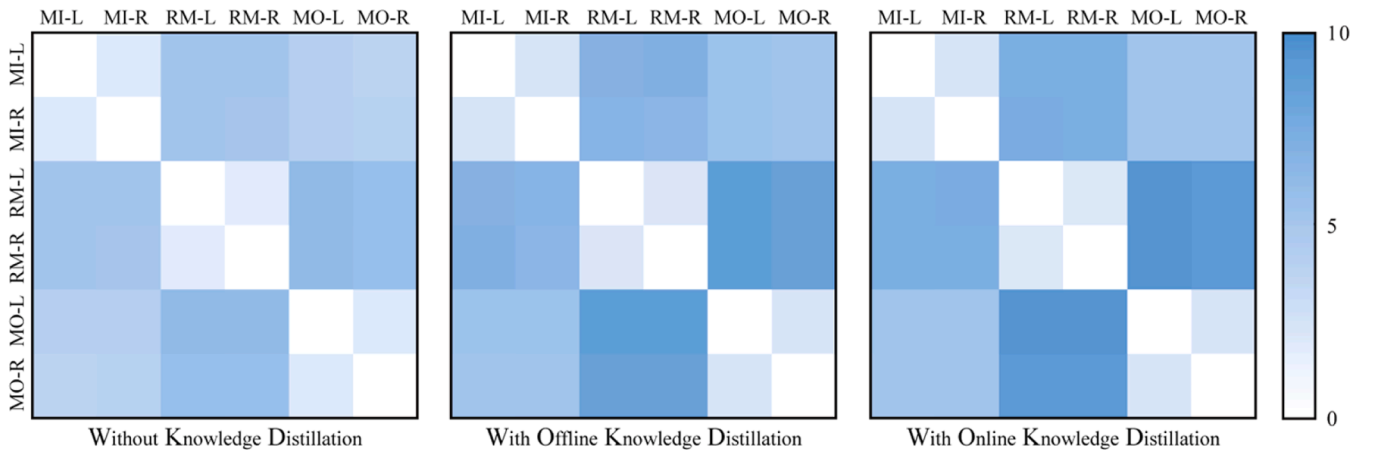


Fig. 8. The distances between classes in the cases without KD, with offline KD, and online KD from the left subplot to the right subplot, respectively.

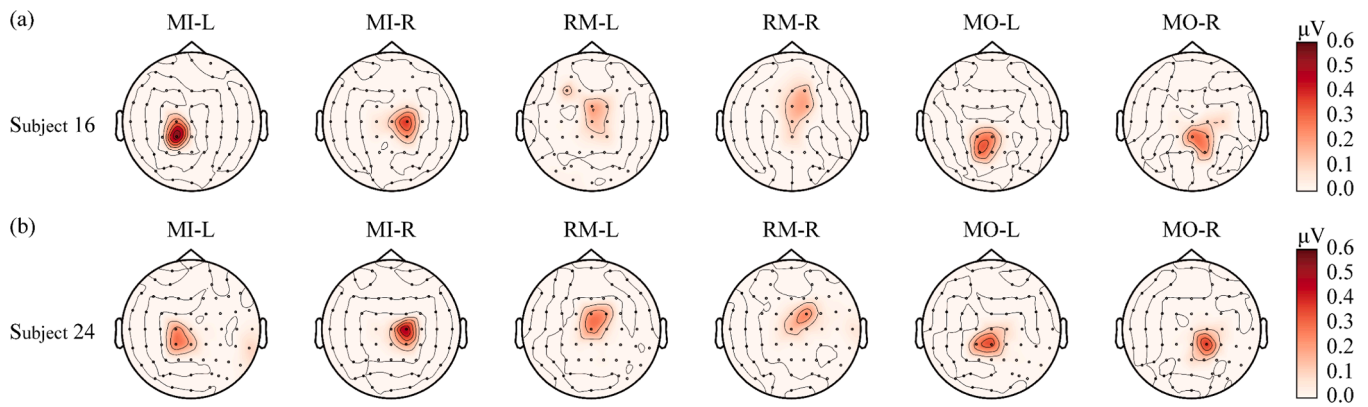


Fig. 9. Comparison of the spatial distributions among different motor actions for (a) subject 16 and (b) subject 24. The relevant brain regions for MI, RM, and MO are not identical.

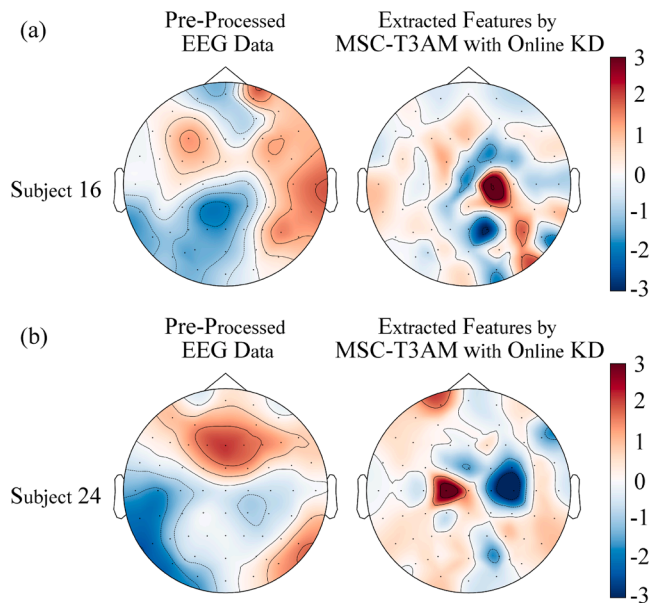


Fig. 10. Topographies of the pre-processed EEG data and the features extracted by the spatial attention module of the proposed MSC-T3AM with online KD for (a) subject 16 and (b) subject 24. The focus of the spatial attention module covers the brain region associated with motor actions of the lower limbs.

which are the relevant brain regions for lower limb motor actions. It reflects the capability of MSC-T3AM with online KD for extracting the most important spatial features.

5. Conclusion and future work

We proposed a multi-scale separable convolutional Transformer-based 3D-attention model with knowledge distillation for classifying six motor actions of the separate lower limbs (including motor imagery for separate left or right lower limb, real movement for separate left or right lower limb, and movement observation for separate left or right lower limb). The proposed model was compared to the state-of-the-art methods. The comparison results demonstrated that our proposed model with online knowledge distillation achieved the best performance in terms of classification accuracy, exhibiting an elevation of 2 %–19 %. We further investigated the role of each crucial module of our proposed model by an ablation study. Moreover, we are also assessing the impact of the MSC-Transformer block and knowledge distillation by measuring the distances for category centers. These results demonstrated that those modules and methods played a positive role in the improvement of the

classification performance. In brief, we used a 3D-attention block to extract features from the filter, spatial, and temporal dimensions, which makes the full utilization of the information contained in EEG signals. We developed the MSC-Transformer block to capture the global features. In addition, multi-scale separable convolutions are used to reduce the complexity of self-attention computation while facilitating the classification. In this study, we investigated the relevant brain regions in different lower limb motor categories through topographies for different motor actions. The observations are consistent with the findings of the other studies. In order to reveal whether the proposed model captures the spatial information well, we visualized features extracted by the spatial attention module of the proposed model. The topographies show that the attention weights assigned to the brain region close to the midline are higher, which matches the neuroscience findings that the motor cortex controlling the lower limbs is close to the midline of the brain.

In this study, we proposed an effective model for multiple motor actions classification. We also demonstrated the advantages of the proposed model. This study can provide informative results for future studies in EEG classification and has the potential to inform the development of rehabilitation training for the human motor system. In future work, we plan to validate and address two limitations: (1) the probability of the target category of the teacher model may interfere with the training of the student model, and (2) the challenge posed by the fixed temperature parameter for learning flexibility. The former limitation may affect KD performance. Meanwhile, the latter limitation becomes particularly evident when the probability distributions of the teacher model and the student model diverge significantly, which requires a higher-than-usual temperature.

CRedit authorship contribution statement

Heng Yan: Software, Methodology, Writing – original draft, Formal analysis. **Zilu Wang:** Data curation. **Junhua Li:** Conceptualization, Supervision, Methodology, Writing – review & editing, Project administration, 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.

References

- Shih, J. J., Krusienski, D. J., & Wolpaw, J. R. (2012). Brain-computer interfaces in medicine. *Mayo Clinic Proceedings*, 87, 268–279. <https://doi.org/10.1016/j.mayocp.2011.12.008>

- Birbaumer, N. (2006). Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology*, 43, 517–532. <https://doi.org/10.1111/j.1469-8986.2006.00456.x>
- Nijholt, A. (2008). BCI for games: A 'state of the art' survey. *Entertainment Computing - ICEC 2008*, 225–228. https://doi.org/10.1007/978-3-540-89222-9_29
- Katona, J., & Kovari, A. (2016). A brain-computer interface project applied in computer engineering. *IEEE Transactions on Education*, 59, 319–326. <https://doi.org/10.1109/TE.2016.2558163>
- Lotze, M., & Halsband, U. (2006). Motor imagery. *Journal of Physiology-Paris*, 99, 386–395. <https://doi.org/10.1016/j.jphysparis.2006.03.012>
- Krautner, S., Gionfriddo, A., Bardouille, T., & Boe, S. (2014). Motor imagery-based brain activity parallels that of motor execution: Evidence from magnetic source imaging of cortical oscillations. *Brain Research*, 1588, 81–91. <https://doi.org/10.1016/j.brainres.2014.09.001>
- Babiloni, F., et al. (2001). Recognition of imagined hand movements with low resolution surface laplacian and linear classifiers. *Medical Engineering & Physics*, 23, 323–328. [https://doi.org/10.1016/S1350-4533\(01\)00049-2](https://doi.org/10.1016/S1350-4533(01)00049-2)
- Tangemann, M., et al. (2012). Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6, 55. <https://doi.org/10.3389/fnins.2012.00055>
- Liu, Y.-H., Lin, L.-F., Chou, C.-W., Chang, Y., Hsiao, Y.-T., & Hsu, W.-C. (2019). Analysis of electroencephalography event-related desynchronization and synchronization induced by lower-limb stepping motor imagery. *Journal of Medical and Biological Engineering*, 39, 54–69. <https://doi.org/10.1007/s40846-018-0379-9>
- Pinto, T. P., Ramos, M. M. R., Lemos, T., Vargas, C. D., & Imbiriba, L. A. (2017). Is heart rate variability affected by distinct motor imagery strategies? *Physiology & Behavior*, 177, 189–195. <https://doi.org/10.1016/j.physbeh.2017.05.004>
- Ambi, J., et al. (2020). Investigation of deep convolutional neural network for classification of motor imagery fNIRS signals for BCI applications. *Biomedical Signal Processing and Control*, 62, Article 102133. <https://doi.org/10.1016/j.bspc.2020.102133>
- Lotte, F., & Guan, C. (2011). Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58, 355–362. <https://doi.org/10.1109/TBME.2010.2082539>
- Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *2008 IEEE International Joint Conference on Neural Networks* (pp. 2390–2397). <https://doi.org/10.1109/IJCNN.2008.4634130>
- Bostanov, V. (2017). BCI competition 2003-data sets ib and iib: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Transactions on Biomedical Engineering*, 51, 1057–1061. <https://doi.org/10.1109/TBME.2004.826702>
- Yüksel, A., & Olmez, T. (2016). Filter bank common spatio-spectral patterns for motor imagery classification. In *Information technology in bio-and medical informatics: 7th international conference, ITBAM 2016, Porto, Portugal, September 5–8, 2016, Proceedings* (pp. 69–84). https://doi.org/10.1007/978-3-319-43949-5_5_7
- Md Isa, N. E., Amir, A., Ilyas, M. Z., & Razalli, M. S. (2017). The performance analysis of K-nearest neighbors (K-NN) algorithm for motor imagery classification based on EEG signal. In *MATEC web of conferences*, Article 01024. <https://doi.org/10.1051/mateconf/201714001024>
- Steyrl, D., Scherer, R., Förstner, O., & Müller-Putz, G. R. (2014). Motor imagery brain-computer interfaces: Random forests vs regularized LDA-non-linear beats linear. In *Proceedings of the 6th international brain-computer interface conference* (pp. 241–244). https://doi.org/10.3217/978-3-85125-378-8_61
- Ma, Y., Ding, X., She, Q., Luo, Z., Potter, T., & Zhang, Y. (2016). Classification of motor imagery EEG signals with support vector machines and particle swarm optimization. *Computational and Mathematical Methods in Medicine*, 2016, Article 4941235. <https://doi.org/10.1155/2016/4941235>
- Ramkumar, E., & Paulraj, M. (2025). Optimized FFNN with multichannel CSP-ICA framework of EEG signal for BCI. *Computer Methods in Biomechanics and Biomedical Engineering*, 28(1), 61–78. <https://doi.org/10.1080/10255842.2024.2319701>
- Liu, X., Shen, Y., Liu, J., Yang, J., Xiong, P., & Lin, F. (2020). Parallel spatial-Temporal self-attention CNN-based motor imagery classification for BCI. *Frontiers in neuroscience*, 14, Article 587520. <https://doi.org/10.3389/fnins.2020.587520>
- Schirmmeister, R. T., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38, 5391–5420. <https://doi.org/10.1002/hbm.23730>
- V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, and B.J. Lance. EEGNet: A compact convolutional network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15, 056013. <https://doi.org/10.1088/1741-2552/aa8e8c>
- Tortora, S., Ghidoni, S., Chisari, C., Micera, S., & Artoni, F. (2020). Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network. *Journal of Neural Engineering*, 17, Article 046011. <https://doi.org/10.1088/1741-2552/ab9842>
- Xie, J., et al. (2022). A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30, 2126–2136. <https://doi.org/10.1109/TNSRE.2022.3194600>
- Xia, W., Zhu, W., Liao, B., Chen, M., Cai, L., & Huang, L. (2018). Novel architecture for long short-term memory used in question classification. *Neurocomputing*, 299, 20–31. <https://doi.org/10.1016/j.neucom.2018.03.020>
- Miao, Z., Zhao, M., Zhang, X., & Ming, D. (2023). LMDA-Net: A lightweight multi-dimensional attention network for general EEG-based brain-computer interfaces and interpretability. *NeuroImage*, 276, Article 120209. <https://doi.org/10.1016/j.neuroimage.2023.120209>
- Altaheri, H., Muhammad, G., & Alsulaiman, M. (2022). Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19, 2249–2258. <https://doi.org/10.1109/TII.2022.3197419>
- Burle, L., Spieser, R., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology*, 97, 210–220. <https://doi.org/10.1016/j.ijpsycho.2015.05.004>
- Song, Y., Jia, X., Yang, L., & Xie, L. (2021). Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv Preprint* <http://arxiv.org/abs/2106.11170>
- Song, Y., Zheng, Q., Liu, B., & Gao, X. (2022). EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 710–719. <https://doi.org/10.1109/TNSRE.2022.3230250>
- Yang, H., Zhao, M., Yuan, L., Yu, Y., Li, Z., & Gu, M. (2023). Memory-efficient Transformer-based network model for traveling Salesman problem. *Neural Networks*, 161, 589–597. <https://doi.org/10.1016/j.neunet.2023.02.014>
- Huang, T., You, S., Wang, F., Qian, C., & Xu, C. (2022). Knowledge distillation from a stronger teacher. *Advances in neural information processing systems* (pp. 33716–33727). <https://arxiv.org/abs/2205.10536>
- G. Hinton, O. Vinyals, and J. Dean. (2015). Distilling the knowledge in a neural network. *arXiv Preprint* <http://arxiv.org/abs/1503.02531>
- Zhang, Z., et al. (2020). Distilling knowledge from well-informed soft labels for neural relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 9620–9627). <https://doi.org/10.1609/aaai.v34i05.6509>
- Adriana, R., Nicolas, B., Ebrahimi, K. S., Antoine, C., Carlo, G., & Yoshua, B. (2015). Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*. <https://arxiv.org/abs/1412.6550>
- Zhao, Q. Cui, Song, R., Qiu, Y., & Liang, J. (2022). Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 11953–11962). <https://arxiv.org/abs/2203.08679>
- Liu, Y., Cao, J., Li, B., Hu, W., Ding, J., and Li, L. (2022). Cross-architecture knowledge distillation. In *Proceedings of the Asian conference on computer vision* (pp. 3396–3411). <https://arxiv.org/abs/2207.05273>
- Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129, 1789–1819. <https://doi.org/10.1007/s11263-021-01453-z>
- Liang, H., Liu, Y., Wang, H., & Jia, Z. (2023). Teacher assistant-based knowledge distillation extracting multi-level features on single channel sleep EEG. In *Proceedings of the thirty-second international joint conference on artificial intelligence* (pp. 3948–3956). <https://doi.org/10.24963/ijcai.2023/439>
- Wang, Z., Gu, T., Zhu, Y., Li, D., Yang, H., & Du, W. (2021). FLDNet: Frame-level distilling neural network for EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2533–2544. <https://doi.org/10.1109/JBHI.2021.3049119>
- Sakhavi, S., & Guan, C. (2017). Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI. In *2017 8th international IEEE/EMBS conference on neural engineering (NER)* (pp. 588–591). <https://doi.org/10.1109/NER.2017.8008420>
- Hurlé, J. N., Lattner, S., & Richard, G. (2021). Darkgan: Exploiting knowledge distillation for comprehensible au-dio synthesis with gans. *International Society for Music Information Retrieval*. <https://arxiv.org/abs/2108.01216>
- Wang, Z., Li, J., Daly, I., & Li, J. (2022). Machine learning for multi-action classification of lower limbs for BCI. In *2022 international conference on computing, electronics & communications engineering* (pp. 84–89). <https://doi.org/10.1109/icCECE55162.2022.9875092>
- Sips, M., Neubert, B., Lewis, J. P., & Hanrahan, P. (2009). Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28, 831–838. <https://doi.org/10.1111/j.1467-8659.2009.01467.x>
- Pfurtscheller, G. (2001). Functional brain imaging based on ERD/ERS. *Vision Research*, 41, 1257–1260. [https://doi.org/10.1016/S0042-6989\(00\)00235-2](https://doi.org/10.1016/S0042-6989(00)00235-2)
- Batula, M., Mark, J. A., Kim, Y. E., & Ayaz, H. (2017). Comparison of brain activation during motor imagery and motor movement using fNIRS. *Computational Intelligence and Neuroscience*, 2017, 1–12. <https://doi.org/10.1155/2017/5491296>
- Taube, W., Mouthon, M., Leukel, C., Hoogewoud, H.-M., Annoni, J.-M., & Keller, M. (2015). Brain activity during observation and motor imagery of different balance tasks: An fMRI study. *Cortex*, 64, 102–114. <https://doi.org/10.1016/j.cortex.2014.09.022>