A Transformer-Based Approach Combining Deep Learning Network and Spatial-Temporal Information for Raw EEG Classification

Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, *Senior Member, IEEE*, Huihui Zhou, and Yang Zhan

***Abstract*— The attention mechanism of the Transformer has the advantage of extracting feature correlation in the long-sequence data and visualizing the model. As time-series data, the spatial and temporal dependencies of the EEG signals between the time points and the different channels contain important information for accurate classification. So far, Transformer-based approaches have not been widely explored in motor-imagery EEG classification and visualization, especially lacking general models based on cross-individual validation. Taking advantage of the Transformer model and the spatial-temporal characteristics of the EEG signals, we designed Transformer-based models for classifications of motor imagery EEG based on the PhysioNet dataset. With 3s EEG data, our models obtained the best classification accuracy of 83.31%, 74.44%, and 64.22% on two-, three-, and four-class motor-imagery tasks in cross-individual validation, which outperformed other state-of-the-art models by 0.88%, 2.11%, and 1.06%. The inclusion of the positional embedding modules in the Transformer could improve the EEG classification performance. Furthermore, the visualization results of attention weights provided insights into the working mechanism of the Transformer-based networks during motor imagery tasks. The topography of the attention weights revealed a pattern of event-related desynchronization (ERD) which was consistent with the results from the spectral analysis of Mu and beta rhythm over the sensorimotor areas. Together, our deep learning methods not only provide novel and powerful tools for classifying and understanding EEG data but also have broad applications for brain-computer interface (BCI) systems.**

***Index Terms*— Motor imagery (MI), EEG classification, transformer, attention mechanism, CNN, visualization, brain–computer interface (BCI).**

Manuscript received 5 February 2022; revised 25 June 2022; accepted 13 July 2022. Date of publication 1 August 2022; date of current version 4 August 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701405, in part by the National Natural Science Foundation of China under Grant 62027804 and Grant 31800900, in part by the Shenzhen Science and Technology Innovation Commission under Grant JCYJ20180508152240368, in part by the Major Key Project of Peng Cheng Laboratory (PCL) under Grant PCL2021A13, in part by the Guangdong Provincial Key Laboratory of Brain Connectome and Behavior under Grant 2017B030301017, in part by the Shenzhen Basic Research Program under Grant JCYJ20200109114805984, and in part by the Shenzhen Key Laboratory of Translational Research for Brain Diseases under Grant ZDSYS20200828154800001. *(Jin Xie, Jie Zhang, and Jiayao Sun contributed equally to this work.) (Corresponding authors: Huihui Zhou; Yang Zhan.)*

Jin Xie and Liuni Qin are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China, also with the University of Chinese Academy of Sciences, Beijing 101408, China, and also with the Shenzhen Key Laboratory of Translational Research for Brain Diseases, Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions, Shenzhen 518055, China.

Jie Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China, also with the University of Chinese Academy of Sciences, Beijing 101408, China, and also with the Research Center for Artificial Intelligence, Peng Cheng Laboratory, Shenzhen 518066, China.

Jiayao Sun and Zheng Ma are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China, and also with the Shenzhen Key Laboratory of Translational Research for Brain Diseases, Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions, Shenzhen 518055, China.

Guanglin Li is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China, and also with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen 518055, China.

Huihui Zhou is with the Research Center for Artificial Intelligence, Peng Cheng Laboratory, Shenzhen 518066, China, and also with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China (e-mail: zhouhh@pcl.ac.cn).

Yang Zhan is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China, also with the Shenzhen Key Laboratory of Translational Research for Brain Diseases, Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions, Shenzhen 518055, China, and also with the CAS Key Laboratory of Brain Connectome and Manipulation, Shenzhen 518055, China (e-mail: yang.zhan@siat.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2022.3194600, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2022.3194600

I. Introduction

ELECTROENCEPHALOGRAM (EEG) reflects the activities from different neuron populations in the central nervous system (CNS). EEG has been widely used in neural engineering, neuroscience, and brain-computer interface (BCI) systems [1]–[3]. The Motor Imagery (MI) paradigm is commonly used in the electroencephalogram brain-computer interface (EEG-BCI) system [4]–[6], which requires subjects to imagine the movement of different parts of the body, rather than the actual movement. Therefore, the accurate classification of EEG signals from different MI tasks is important for the BCI system [5]. External devices can take advantage of the accurate classification to perform multiple tasks through the BCI system. Accurate classification is helpful for the rehabilitation and functional recovery of patients. However, EEG

signals have low spatial resolution, high temporal resolution, low signal-to-noise ratio, and large individual differences [2]. These characteristics pose a great challenge to the signal processing and accurate classification of motor imagery EEG in BCI systems.

Traditional machine learning methods for motor-imagery electroencephalogram (MI-EEG) classification usually included feature extraction and feature classification [6]. For example, Filter bank common spatial pattern (FBCSP) [7], Fast Fourier Transform (FFT) [8], and Wavelet Transform [8], [9], etc. for feature extraction, supervised learning methods of support vector machine (SVM) [10] and linear discriminant analysis (LDA) [11] for feature classification, or unsupervised learning method of K nearest neighbor (KNN) for feature classification [6], [12]. However, the useful information of EEG may be lost during feature extraction. With the abilities of automatic feature extraction and rich feature representation, the deep learning model can directly receive the preprocessed EEG data and establish an end-to-end model without feature extraction [2], [5]. Deep learning structures [2] of EEG classification have been applied in medical and neuroscience fields, mainly including convolution neural network (CNN) [5], [13], deep belief network (DBN) [14], recurrent neural network (RNN) [15], [16] and Hybrid CNN [17]. In the CNN model, the spatial kernel and the temporal kernel were respectively used to extract EEG information from different channels simultaneously and from the same channel at different times [5], [13]. Some studies chose Long short-term memory (LSTM) and a modified RNN to capture temporal information in EEG signals [15], [18], [19]. Some fusion models used CNN to extract spatial features, and then input these features into the RNN model to learn temporal information [20], [21]. Others have studied the combination of CNN and Multi-layer perceptron or Auto-encoder for EEG classification [22]. For the motor imagery EEG, CNN networks are widely used, while it has limitations in perceiving global dependence [2], [13]–[15], [23], [24].

Artificial attention in deep learning is proposed based on the attention mechanism in the brain, which helps to improve the flexibility and performance of the deep learning model. There are two commonly used self-attention mechanisms, namely the RNN-based attention mechanism, which calculates attention weights based on the hidden layer of the RNN, and the multi-head attention mechanism, which calculates the correlation between each pair of time points [25]. The RNN-based attention mechanism has been used for the EEG classification tasks [26]. The transformer uses multi-head attention instead of a recurrent layer or convolutional layer to extract information, such as BERT [27] and GPT-2 [28]. These methods improve the performance of multiple tasks in natural language processing (NLP) [25]. Recently, Transformer-based models have been developed for object detection [29], image classification [30], and protein engineering [31], suggesting their wide applicability. Time series such as EEG signals have long-range dependencies, which can be characterized by estimating the Long-Range Temporal Correlation (LRTC) [32], [33]. LRTC has been indeed observed in the EEG and becomes stronger during voluntary movement and motor imagery [32], [33]. The spatial linear and nonlinear dependencies have been observed in time-series EEG signals [34], [35], showing inter-channel correlation [36], [37]. Therefore, spatial and temporal dependencies inherent in the EEG signals can be extracted for the classification tasks. Learning long-range dependencies is a key challenge in many sequence transduction tasks [25]. Transformer is the first sequence transduction model relying entirely on an attention mechanism to draw long-range dependencies without using complex recurrent or convolutional neural networks [25]. The transformer also has better interpretability than the above-mentioned deep learning models [38].

EEG-based Transformer models have been used for emotion recognition, classification of imagined speech, and sleep stage classification [39]–[43]. A few studies attempted to adopt Transformer models for motor imagery EEG. Tao *et al.* [44] employed a gated Transformer on the same PhysioNet dataset as used by this paper but they only conducted one multi-class classification. Song *et al.* [24] performed spatial filtering first on a different EEG motor imagery dataset and then subjected the data to a Transformer decoder. This study only had 9 subjects and used subject-specific models (within-individual training). Kostas *et al.* [45] used the same EEG dataset, while the model consisted of two stages: the first stage downsampled raw data using a stack of short-receptive field 1D convolutions, and the second stage used a transformer encoder to map data representations to some new sequence that embodies the target task. This study collapsed the spatial and temporal information in the Transformer. Another drawback of these studies is that they use within-subject training by combined EEG datasets and this approach has limited adaptability and robustness for different individuals [46]. These studies mainly rely on combined EEGs from multiple trials and they need frequency-domain features or require preprocessing of the EEG data. In this study, we propose an end-to-end Transformer framework that is capable of processing raw EEG data while retaining the spatiotemporal characteristics that are important for model visualization.

To address the above issues, we proposed a new deep learning structure for EEG classification based on the Transformer module, and analyzed the model behaviors for MI-EEG classification. The main contributions are as follows:

First, we designed a novel Transformer model for studying the brain-like neural mechanisms. Five categories of Transformer-based models, including spatial- Transformer (s-Trans), temporal-Transformer (t-Trans), spatial-CNN + Transformer (s-CTrans), temporal-CNN + Transformer (t-CTrans), and fusion-CNN + Transformer (f-CTrans), and systematically tested these models on the Physionet EEG Motor Movement/Imagery Dataset [47]. With 3s data, our models obtained the best accuracy of 83.31%, 74.44%, and 64.22% in two-, three-, and four-class classification tasks, respectively, which outperformed other state-of-the-art (SOTA) models.

Next, we explored three categories of Positional Embedding (PE) modules [25], relative positional encoding, channel correlation positional encoding, and learned positional encoding. Compared with the baseline model without positional

encoding, the accuracy of embedded positional coding was improved by 0.36% to 2.63%, which proved that positional embedding methods could improve the EEG classification ability.

Last, the weights of the multi-head attention module in the s-Trans model were visualized. The visualizations were plotted based on the EEG electrode locations. We found that the weights of the Transformer module corresponding to the sensorimotor areas [48]–[50] showed a pattern of event-related desynchronization (ERD). These data were consistent with the Mu and beta band rhythmic ERD [51], [52]. The visualization results demonstrated that Transformer-based methods can contribute to the understanding of network behavior for the classification tasks based on the EEG data.

II. Materials and Methods

A. Dataset and Preprocessing

We used the PhysioNet EEG Motor Movement/Imagery Dataset [47] containing 109 subjects with more than 1500 trials. The dataset was recorded using the BCI2000 system from 64 electrodes sampled at 160 Hz. Each subject performed 14 runs consisted of 2 baseline runs, 6 motor movement runs and 6 motor imagery runs. In this study, we focused on the motor imagery classification and selected the following runs:

- A baseline run for rest-state of opening eyes (O),
- Three task runs for motor imagery of left fist (L) against right fist (R),
- Three task runs for motor imagery of both fists against both feet (F).

Based on the above categories, data were arranged as three subsets: two-class of L/R, three-class of L/R/O, and four-class of L/R/O/F, respectively [5], [13]. For each subject, 21 trials were selected per class. Each trial lasted 8 seconds, with the first 2s for rest, the following 4s for motor imagery, and the last 2s for rest. 3s (480 samples) and 6s (960 samples) segments of EEG data were used to train and test our models. We used both 3s and 6s data for the classification. 3s data included the first 3s data from the motor imagery period, and 6s data included the entire motor imagery period as well as one second before and one second after the motor imagery period. We applied the Z-score normalization to preprocess the EEG data, and added the random noise to prevent over-fitting, as shown in the following formula:

$$X^* = \frac{X - \mu}{\delta} + \alpha N \tag{1}$$

X was the raw EEG data and X^* was the EEG data after preprocessing. μ was the mean value of data and δ was the standard deviation. N represented the random noise and α controlled the percent of random noise. The function of "np.random.randn" in Python was used to generate the random noise N with standard normal distribution. As the same with Wang *et al.* [5], the percent of random noise α was set to 0.01.

B. Model Architecture

1) Architecture Framework: Fig. 1 showed the overall framework of our Transformer-based approaches for EEG classification.

The framework demonstrates an end-to-end classification of the raw EEG data. The network consisted of Transformer modules as well as operations of Positional Embedding. We also designed methods that combined the CNN module and the Transformer module. CNN was included because of its good properties for feature representation [5]. In the implementation, we built a total of five Transformer-based models in which two models only relied on the Transformer without including the CNN and three models used network architecture of combined CNN and Transformer. After the CNN and the Transformer modules, we included a fully-connected layer.

2) Transformer Module: We adopted the network architecture of Transformer [25], which has achieved excellent performance in the translation quality of natural language processing (NLP). Like most competitive neural sequence transduction models, the Transformer module followed the encoder-decoder structure using stacked self-attention and point-wise, fully connected layers. The model multiplied the input vector with three different weight matrices to obtain the queries vector (Q), keys vector (K) and values vector (V). The "Scaled Dot-Product Attention" was shown in Fig. 2a, which computed the dot products of the queries with all keys, divided each by $\sqrt{d_k}$, and applied a softmax function to obtain the weights on the values, as shown in formula (2):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Multi-head attention consisted of several "Scaled Dot-Product Attention" layers, allowing the model to jointly focus on information from different representation subspaces at different locations [25]. The "Multi-Head Attention" was shown in Fig. 2b and formula (3):

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O \\ \text{Wherehead}_i &= \text{Attention}(Q, K, V) \end{aligned} \tag{3}$$

In this study, we employed $h = 8$ parallel attention layers (so-called 8 attention heads), and solely embedded the encoder part of Transformer into the EEG classification. As shown in Fig. 2c, the Transformer module had two submodules. The first submodule included a multi-head attention layer followed by a normalization layer. The second submodule included a position-wise fully connected feed-forward layer followed by a normalization layer. The residual connection was employed around each of the two submodules.

3) Positional Embedding (PE) Module: In the field of NLP, the predecessors embedded the position coding in the Transformer model to increase the location information [25]. For the EEG data, we explored three categories of Positional Embedding (PE) modules, the relative positional encoding, the channel correlation positional encoding, and the learned positional encoding. The relative positional embedding method used the sine and cosine function to represent the relative position coding (formula (4) and (5), respectively). If we considered the channel position as pos, and the time points

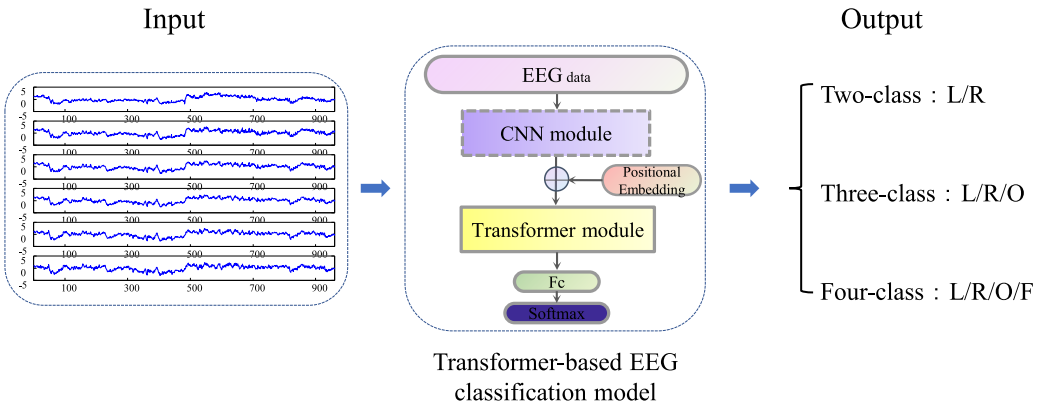

Fig. 1. Illustration of the proposed transformer-based deep learning framework for MI-EEG classification.

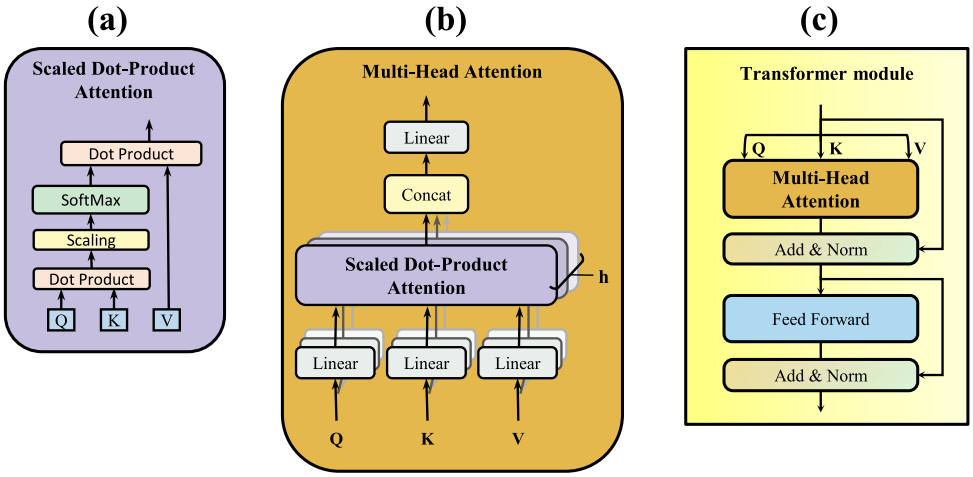

Fig. 2. The detailed structure of transformer module: (a) Scaled dot-product attention, (b) Multi-head attention, (c) Transformer module.

as i, positional encoding was described as follows,

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{4}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{5}$$

d represented the dimension of the vector. At each position of the electrode vector, the PE at the even and odd time points were described by the sine and cosine functions, respectively, and i was the index of the node in the electrode vector divided by 2. We conducted the inner product of the relative position coding of the position pos1 and pos2, and found that as the distance increased, the correlation between the two positions became smaller.

In the channel correlation positional embedding method, the Cz electrode was selected as the central electrode, and the cosine distance between other electrodes and the central electrode was calculated. $P_{central}$ represented the three-dimensional coordinate of the central electrode Cz, and P_k was the K_{th} position for the three-dimensional coordinate. As shown in the formula 6, the cosine distance from all electrodes to the central electrode Cz could be calculated. The sine and cosine operations were carried out by using the distance of sim_k instead of pos in the formula 4 and formula 5, and the resulting matrix was the position coding matrix of channel correlation positional encoding.

$$sim_k = \frac{P_{central} \cdot P_k}{\|P_{central}\| \|P_k\|} \tag{6}$$

In the learned positional embedding method, we embedded a trainable matrix of the same size to the inputs, and initialized the embedding matrix randomly. During training the model, the parameters of the position coding matrix were updated constantly by learning.

4) Spatial and Temporal Transformer Model: Transformer model considered the correlations between the data points in the sequence data. In order to consider correlations in both temporal and spatial dimensions in the EEG data, we arranged the input data for the Transformer modules in both spatial-wise and temporal-wise way. In the spatial-wise way (s-Trans, Fig. 3a), EEG data along the time axis from each channel were regarded as features, and the Transformer module calculated the correlations between different channels. In the temporal-wise way (t-Trans, Fig. 3b), EEG data along the channel axis at the same time point were regarded as features, and the model calculated the correlations between different time points.

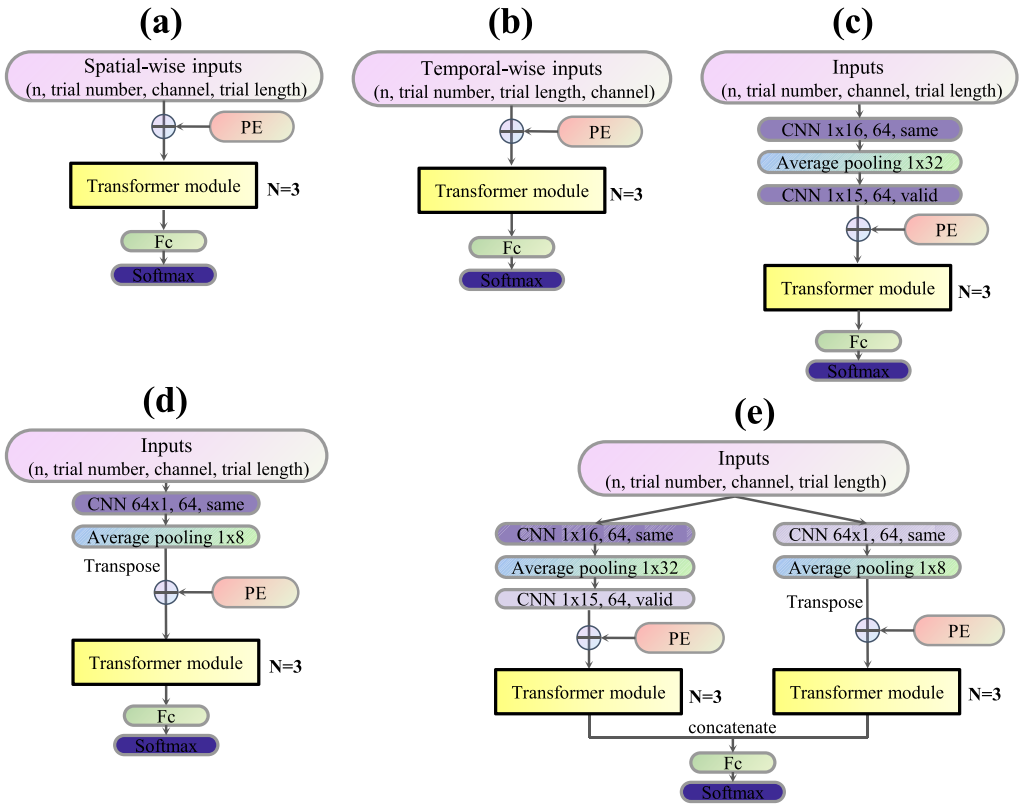

Fig. 3. The detailed architecture of EEG classification models based on Transformer. (a) spatial-transformer (s-Trans) model, (b) temporal-Transformer (t-Trans) model, (c) spatial-CNN + Transformer (s-CTrans) model, (d) temporal-CNN + Transformer (t-CTrans) model, (e) fusion-CNN + Transformer (f-CTrans) model.

As shown in Fig. 3(a-b), the spatial-wise or temporal-wise EEG data were embedded with the position coding and then fed into the Transformer modules. The features obtained by Transformer modules were fed into the fully connected layers for EEG classification. We explored the influence of the number of Transformer modules on the classification results. The number of Transformer modules was tested from 1 to 6. When the number of 3 was chosen, the classification achieved the best results. We therefore included three Transformer modules in our models.

5) Transformer-Based Model Combining With Convolutional Neural Network (CNN + Transformer Model): Convolutional Neural Network (CNN) has been used for generalized feature learning and dimension reduction for the EEG data [5], we designed a fusion model which combined the CNN and the Transformer module. CNN implementation also took the spatial and temporal representation into consideration. The CNN performed the feature extraction and these features were fed into the multi-head attention layer of the Transformer.

In the spatial implementation of the CNN + Transformer model (s-CTrans, Fig. 3c), the CNN module included two convolutional layers and one average pooling layer. In the first convolutional layer, we used 64 kernels with the size of 1×16 (channel $\times$ time points) to extract EEG temporal information, and adopted the SAME padding. The average pooling layer had the pooling size of 1×32. The second convolutional layer used 64 kernels with the size of 1×15, and adopted the VALID padding.

In the temporal implementation of the CNN + Transformer model (t-Ctrans, Fig 3d), the CNN module included one convolutional layer and one average pooling layer. The convolutional layer used 64 kernels with the size of 64×1(channel $\times$ time points) to extract EEG spatial information, and adopted the SAME padding. The average pooling layer had the pooling size of 1×8. After the average pooling layer, we transposed the features.

For both s-CTrans and t-CTrans models, features obtained by the CNN module were embedded with the position coding, and then passed through the Transformer modules and then through fully connected layers for EEG classification, as shown in Fig. 3c-d.

The fusion CNN + Transformer model (f-CTrans) dealt with the spatial and temporal information in parallel (Fig. 3e). After the CNN and the Transformer processing, the two outputs from the two streams were combined. The combined features were fed into the fully connected layers for EEG classification.

C. Training Setup

1) Training Parameter Settings: Empirically, the number of head in each multi-head attention layer was set to 8 [25]. The dropout rate was set to 0.3. The parameter of the position-wise fully connected feed-forward layer with a ReLU activation was set to 512. The weight attenuation was 0.0001. All the models used the Adam optimizer. The training epoch was set to 50. We used three GeForce GTX 1080Ti GPUs to train our

TABLE I
PARAMETER SETTINGS FOR OUR TRANSFORMER-BASED EEG CLASSIFICATION MODELS

Parameters	s-Trans	t-Trans	s-CTrans	t-CTrans	f-CTrans
Learning rate	0.0007	0.0005	0.0007	0.0007	0.001
Momentum	0.9, 0.999	0.9, 0.95	0.9, 0.95	0.9, 0.95	0.9, 0.95
d_k	64	64	128	128	128
d_v	64	64	128	128	128
Decay milestones	5,15,30,45	5,15,30	10,20,30	10,30	10,30
Decay rate	0.1	0.1	0.05	0.1	0.1

models, and constructed these deep learning models on the Pytorch platform. In the spatial-Transformer model, the data points across all time points in a given channel were used as features. For the Positional Embedding, the dimension of d in formula 4 and formula 5 was 480 (3s) or 960 (6s), depending on the temporal length of inputs. In the temporal- Transformer model, the data points across all the channels at one given time point were used as features. The dimension of d was equal to the number of channels. For the CNN + Transformer model, the dimension of d was set to 64. The remaining parameters were shown in Table I.

2) Model Training Techniques: There were two training techniques for training and testing EEG classification models, the within-individual training and cross-individual training [22]. EEG data were usually recorded in multiple sessions for one individual. During the within-individual training, multiple-session data was divided into the training set and test set. The model was tested on new sessions, but training and test sessions belonged to the same individual. The within-individual training technique could give higher accuracy. During the cross-individual training, the individuals were divided into training individuals and test individuals. Then, the model was tested on new individuals. The cross-individual training technique involved information transfer between different individuals. Although cross-individual approach was challenging, the models evaluated by this approach were more robust and generalized [22]. Thus, we used the cross-individual training technique for training and testing our proposed Transformer-based models.

In this study, we adopted the 5-fold cross-validation to test the model performance. The data samples were from all the trials across multiple sessions. During training, the individuals were randomly split into 5 subsets. One of the 5 subsets was selected for testing, and the remaining 4 subsets were used for training. We repeated this process for 5 rounds to obtain 5 accuracies. The classification result was the average of 5 repetitions.

3) Performance Metric: We used the classification accuracy and the confusion matrix to evaluate the classification performance. The accuracy was the ratio of correctly recognized test samples to all test samples (test subjects × trials/subject). The confusion matrix was a performance measurement technique in the binary and multiclass classification problems, which represented the counts from the predicted and actual values broken down by each class.

D. Visualization

The visualizations of attention weights have been widely used in the NLP area [53]. In the multi-head attention module, visualization could show how the model allocated weights in different nodes of input, which was helpful to understand the working mechanism of the Transformer [38]. Since in our spatial-Transformer model the length of the output features from the attention layer equals the number of the channels, this allowed a direct comparison of the EEG topographical analysis to the value of the weights according to the electrode locations on the brain. We then used the s-Trans model to perform the visualizations for the two-class classification.

In the "Scaled Dot-Product Attention" (Fig. 2a), the input consisted of queries (Q) and keys (K) of dimension d_k, and values (V) of dimension d_v. In our Transformer-based models, we set $d_k = d_v = 64$, which was the same size as EEG channel numbers. The attention weight matrix in this paper was the Attention(Q, K, V) in the formula (2). We visualized the high-dimensional features from the last Transformer module in which more representative capability of the EEG should be extracted. The size of the attention weight matrix was 64 × 64. The diagonal of the attention weight matrix represented the degree of activation of each electrode. Then we mapped the diagonal of the attention weight matrix into the head topography using MNE-Python based on the international 10-10 EEG standard. The values were then normalized to the range of [0, 1].

The visualization and the statistical analysis used the data from one-fold test result of 5-fold cross-validation. The brain areas corresponding to the sensorimotor cortex had been demonstrated to exhibit ERD during motor imagery [54]–[56]. Similar to other literature, we chose 9 electrodes from the left hemisphere of the sensorimotor area including: FC5, FC3, FC1, C5, C3, C1, CP5, CP3, CP1, and 9 electrodes from the right hemisphere [50], [57]–[59], including: FC2, FC4, FC6, C2, C4, C6, CP2, CP4, CP6. We averaged the attention weights for the above electrodes in each hemisphere and performed a Wilcoxon signed-rank test between the left and the right hemispheres across all the test subjects. We presented averaged attention weights for all the test subjects using 64 electrodes on the head topography. Because 8 heads were included in the multi-head attention module, we obtained 8 head topographies of attention weights per class.

III. RESULTS

A. Performance of Transformer-Based EEG Classification Models

1) The Classification Accuracies: We summarized the results for the classification accuracy from our five model implementations in Table II. Considering the application comparability between models, our Transformer-based models are compared with some recent representative baselines [5], [13], [45], [60]–[62] using the same dataset of the PhysioNet and similar cross-individual training method, as shown in

TABLE II
ACCURACY (%) COMPARISON BETWEEN OUR MODELS AND OTHER SOTA MODELS IN THE PHYSIONET DATASET FOR CROSS-INDIVIDUAL CLASSIFICATION

Models	3s			>= 4s		
	L/R	L/R/O	L/R/O/F	L/R	L/R/O	L/R/O/F
Our s-Trans	81.11	70.25	59.35	87.46	75.41	64.04
Our t-Trans	80.77	70.31	58.21	86.10	75.24	62.15
Our s-CTrans	**83.31**	72.88	63.25	87.80	77.09	68.10
Our t-CTrans	82.56	72.87	63.48	87.80	**78.98**	**68.54**
Our f-CTrans	82.95	**74.44**	**64.22**	87.26	78.44	67.96
CNN (2018) [5]	80.38	69.82	58.58	**87.98**	76.61	65.73
EEGNet (2020) [13]	82.43	72.33	63.16	--	--	--
EEGNet Fusion (2020) [60]	--	--	--	83.80	--	--
DG-CRAM (2020) [61]	74.71	--	--	--	--	--
MAML-CNN (2021) [62]	80.60	--	--	--	--	--
BENDR (2021) [45]	--	--	--	86.70	--	--

TABLE II. Brief descriptions of these baseline models were given as follows: CNN [5] is based on a shallow CNN with two convolutional layers. EEGNet [13] proposes a general-purpose CNN model with a single EEGNet architecture. EEGNet Fusion [60] is also based on CNN with a three-branch EEGNet architecture. DG-CRAM [61] is a graph convolutional recurrent attention model with combined CNN and RNN. MAML-CNN [62] proposes a CNN-based model with model-agnostic meta-learning. BENDR [45] is a Transformer-based model with two stages.

Using 3s data, the best accuracies from our models were 83.31%, 74.44%, and 64.22% for two-, three-, and four-class classifications, respectively. Our results performed better than the baseline models in all three classifications. Among these models (>=4s) in TABLE II, except EEGNet Fusion [60] on 4s data, other models all used 6s data. Using 6s data, the best accuracies of our models were 87.80%, 78.98%, and 68.54%. Therefore, inclusion of the EEG data with longer period produced higher classification accuracy. When compared with the baseline models, our methods were better for the three- and four-class classifications. In case of three- and four-class classifications, for 3s data our f-CTrans performed best and for 6s data our t-CTrans model performed best. Therefore, our Transformer-based classification methods had great classification ability.

2) The Training Curves: The classification accuracy training curves of EEG classification models in different motor imagery tasks were plotted against the training epochs with 3s data (Fig. 4). The classification accuracies gradually increased to stable levels as the training epochs increased. After adding CNN, the accuracy training curves converged faster, which was consistent with the higher accuracy. And the training curves of two-class classification converged better than that of three- and four-class classifications, which was consistent with the higher accuracy of two-class in compared with three- and four-class.

3) The Confusion Matrix Results: In the study, the confusion matrix was used to observe the correct classification and misclassification of each category. The number on the diagonal of the confusion matrix represented the number of correctly classified samples in each category, otherwise the off-diagonal of the matrix represented the number of misclassifications.

TABLE III
CLASSIFICATION RESULTS OF SPATIAL-TRANSFORMER MODEL USING DIFFERENT POSITIONAL EMBEDDING METHODS

Methods	480 (3s)			960 (6s)		
	L/R	L/R/O	L/R/O/F	L/R	L/R/O	L/R/O/F
relative PE	81.11%	**70.25%**	59.35%	**87.46%**	75.41%	64.04%
Channel correlation PE	**81.49%**	69.48%	**59.47%**	87.14%	75.26%	64.05%
learned PE	81.47%	70.02%	59.08%	87.07%	**75.52%**	**64.06%**
No PE	81.13%	68.25%	57.23%	86.83%	73.15%	61.43%

As shown in Fig. S1, the correct classification number of all models was considerably larger than the misclassification number, indicating the effectiveness of the Transformer model.

4) The Results of Positional Embedding Methods: To understand how the Positional Embedding (PE) contributed to the classification, we compared the classification results using the three PE methods in the s-Trans model. Compared to the model without the PE, the three PE methods had better classification results (Table III). For 3s and 6s data, different PE methods had different accuracies, though the differences were modest. It should be noted that the learned positional embedding method required training and had more training parameters. Therefore, inclusion of the PE methods in our models increased the classification accuracy.

B. Visualization of Transformer-Based EEG Classification Models

1) t-SNE Visualization: We applied the t-distributed Stochastic Neighbor Embedding (t-SNE) method to visualize the high-dimensional features in the fully-connected layer of our s-CTrans model. In the two-class classification task (Fig. 5a), the two clusters were well separated. In the three-class classification task (Fig. 5b), the left fist and right fist clusters were separated, but the cluster of opening eyes were distributed between the other two clusters. In the four-class classification task (Fig. 5c), a large portion of samples in the cluster of both feet and opening eyes overlapped with the left and right fist clusters.

2) Attention Weight Visualization: We analyzed the attention weights from the multi-head attention layer and visualized the weights in a way similar to the EEG topography. The brain areas corresponding to the sensorimotor cortex have been demonstrated to display ERD during MI tasks [54]–[56]. Similar to other ERD studies using spectral analysis of the EEG [50], [57]–[59], we focused on the sensorimotor areas for both hemispheres (left: FC5, FC3, FC1, C5, C3, C1, CP5, CP3, CP1; right: FC2, FC4, FC6, C2, C4, C6, CP2, CP4, CP6; as shown in the left subgraph of Fig. 6a). We visualized all the attention weights for 8 heads in our Transformer model (as shown in the Fig. S2). In the Fig. 6, we presented a typical result of one head (head5). Interestingly, our visualization results showed the same patterns as observed in the ERD with one hemisphere showing reduced response compared to the other one (Fig 6a). We performed statistical analysis for the averaged attention weights between the left and right hemispheres in the 21 subjects from the test datasets.

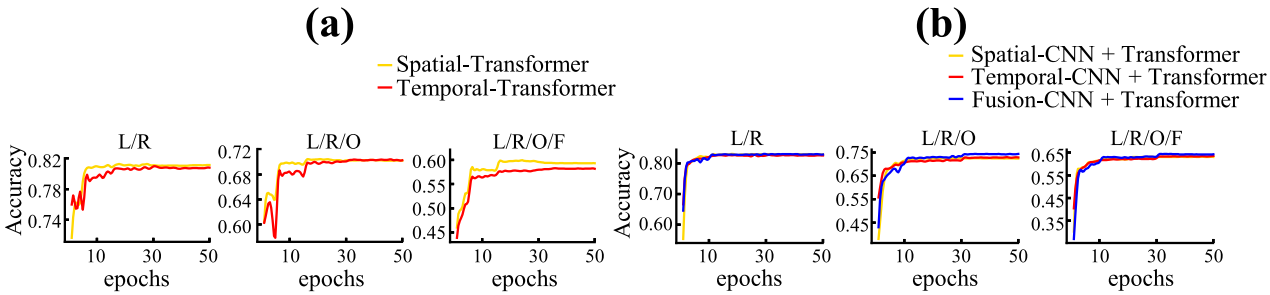

Fig. 4. The accuracy training curves of EEG classification models in three motor imagery tasks. (a) The s-Trans and t-Trans models, (b) s-CTrans, t-CTrans, and f-CTrans models.

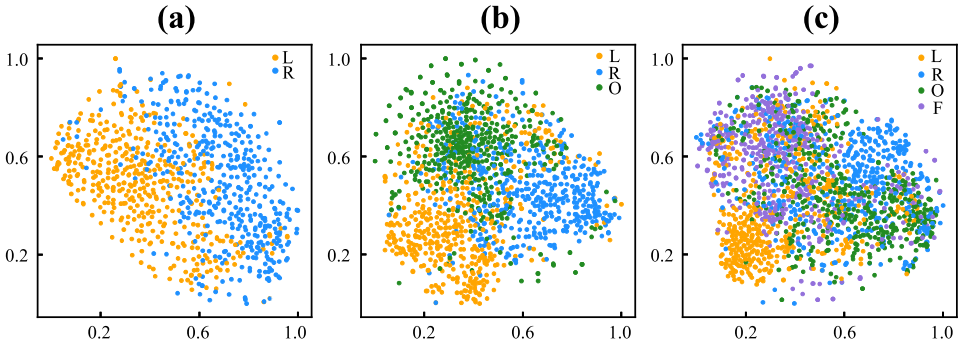

Fig. 5. The t-SNE visualization for the high-dimensional features from the fully connected layer in the s-CTrans model: (a) the two-class classification of motor imagery of left (L) and right (R) fist, (b) the three-class classification of motor imagery of left (L) and right (R) fist, and opening eyes (O), (c) the four-class classification of motor imagery of left (L) and right (R) fist, opening eyes (O), and motor imagery of both feet (F).

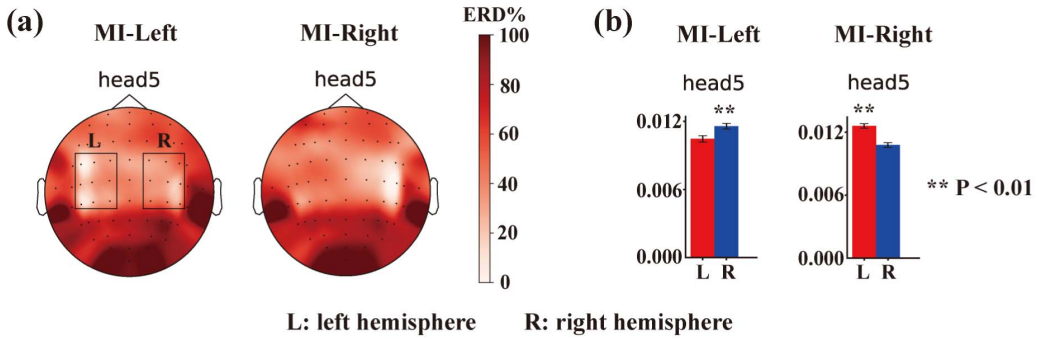

Fig. 6. Visualization of attention weights in the Transformer module. The head5 was plotted. (a) The head EEG topography based on attention weights. (b) The statistical analysis for the averaged attention weights between the left and right hemispheres. Motor imagery of left fist (MI-Left), motor imagery of right fist (MI-Right).

The statistical analysis results of all the 8 heads are detailed in the Fig. S3. During the motor imagery of left fist (as shown in the left subgraph of Fig. 6b), the average attention weights of the right hemisphere were significantly greater ($P < 0.01$, Wilcoxon signed rank test) than those of the left hemisphere in head5. During the motor imagery of right fist (as shown in the right subgraph of Fig. 6b), the average attention weights of the left hemisphere were significantly greater ($P < 0.01$, Wilcoxon signed rank test) than those of the right hemisphere in head5. Consistent with the mu and beta band analysis in previous ERD studies [56], [63], our attention weight results also showed contralateral enhancement of the ERD. These data suggested that our Transformer model can disclose movement-related rhythmic patterns during motor imagery tasks. Based on the visualization analysis for the Transformer-based model in motor imagery tasks, we found a brain-like neural mechanism.

IV. Discussion

In this work, we develop classification methods incorporating Transformer models for motor imagery EEG datasets. Transformer has been successfully applied in the natural language processing. By constructing input feature vectors in both spatial and temporal ways, Transformer models have the ability to extract dependency between different EEG channels and different time points. Transformer models alone have good performance, however, when combined with the CNN, the fusion models can improve the performance. This is probably due to that CNN is good for feature extraction. Positional Embedding (PE) added before the Transformer processing can

further improve the model accuracy. The overall improvement of our Transformer-based model accuracy over the baseline models on the PhysioNet datasets with cross-individual validation demonstrate that Transformer implementation is a good candidate when considering deep-network methods for EEG classification.

For the classifications of motor-imagery tasks based on EEG data, some studies [24], [46], [64] used within-individual method to train subject-specific models. During within-individual validation, samples from different trials of the same subject are split to training set and test set, and data from different channels are usually combined to obtain more samples. Though the within-individual method yielded higher accuracy, the trained model cannot be generalized to different subjects. Various ways of collapsing data also destroyed the inherent neural representations within the EEG. In this study, we performed the cross-individual validation on the classification of motor imagery task to train with a global model, which can be generalized to different subjects with good adaptability and robustness, essential for the applications of our proposed methods for other EEG datasets with more general purpose.

Previous Transformer-based methods for EEG applications focus on the classification results and did not perform visualization of the Transformer layers. Our model structure, the ability to process single-trial raw EEG data, and the model implementations, the positional embedding (PE) module for EEG visualization differed significantly from existing Transformer-based models. Here we considered both the spatial and the temporal information of the single-trial EEG data to regain the physiological features inherent in the EEG. In addition, we designed the PE modules to retain the weight structures corresponding to the spatial-temporal information of the EEG to facilitate visualization.

We provided an approach to visualize the Transformer multi-head attention layer. Mapping the attentional matrix into the topographical EEG representation was based on the actual electrode position. In this work, the visualization did not consider the inter-channel dependencies but solely relied on the diagonal value of the attentional matrix for individual channels. The visualization method allows direct comparison of the attentional weights to the surface EEG activities over different brain areas. We found that the topography of the attention weights over the sensorimotor areas [48]–[50] showed a pattern of ERD, which is consistent with the previous results using spectral methods. These data suggested that when fully trained, the attentional network can acquire a pattern similar to the rhythmic activity changes during the motor imagery tasks.

In this study, we proposed five categories of Transformer-based models so that the Transformer models can be flexibly adapted to different EEG scenarios. Taken together, for 3s data the f-CTrans model performed best and for 6s data the t-CTrans model performed best. With 3s data, the f-CTrans model performed best in three- and four-class classification tasks, while its performance of all classification tasks outperformed all baseline models. Shorter input data improved the processing efficiency for the BCI system. While the multi-class classification with shorter data is more challenging, f-CTrans model showed better performance, indicating its robustness. The f-CTrans model did not perform the best in the 2-class classification task, which is the limitation we need to further overcome. In the future, we will further improve the effectiveness of the fusion by optimizing the structure of the fusion model or improving the combination method of spatial-temporal information.

Nevertheless, one of the advantages of our Transformer-based model is the ability to extract features from large datasets. The performance of our model will be improved with more EEG data included. Our Transformer models can be further optimized in two aspects. First, temporal features extracted from the EEG may have different time-scales. Further work can construct multi-scale attention model and test the model performance. Second, the visualization results indicate that some attentional heads may not contribute to the neural-mechanism based representation. Removal of these heads can reduce the computation load and improve the model robustness.

V. Conclusion

In the present study, we discussed the application of the Transformer model in motor imagery EEG classification. Five categories of Transformer-based models were designed including spatial-Transformer model, temporal-Transformer model, spatial-CNN + Transformer model, temporal-CNN + Transformer model, and fusion-CNN + Transformer model. For the 3s data, the highest accuracy of two-, three-, and four-class classifications consistently outperformed other SOTA models. For three- and four-class classifications, the fusion model had the best performance. Our results showed that Transformer models provided good performance for EEG classification during motor imagery tasks, and can be applied in other classification tasks such as disease diagnosis and brain-computer interface control tasks based on EEG data.

References

[1] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, Feb. 2012.

[2] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, Jun. 2019, Art. no. 031001.

[3] Y. He, D. Eguren, J. M. Azorín, R. G. Grossman, T. P. Luu, and J. L. Contreras-Vidal, "Brain–machine interfaces for controlling lower-limb powered robotic systems," *J. Neural Eng.*, vol. 15, no. 2, pp. 1–15, Apr. 2018.

[4] K. A. Condori, E. C. Urquizo, and D. A. Diaz, "Embedded brain machine interface based on motor imagery paradigm to control prosthetic hand," in *Proc. IEEE ANDESCON*, Oct. 2016, pp. 1–4.

[5] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Syst. Appl.*, vol. 114, pp. 532–542, Dec. 2018.

[6] F. Farooq, N. Rashid, A. Farooq, M. Ahmed, A. Zeb, and J. Iqbal, "Motor imagery based multivariate EEG signal classification for brain controlled interface applications," in *Proc. 7th Int. Conf. Mechatronics Eng. (ICOM)*, Oct. 2019, pp. 1–6.

[7] P. S. Lopez, H. K. Iversen, and S. Puthusserypady, "An efficient multi-class MI based BCI scheme using statistical fusion techniques of classifiers," in *Proc. TENCON IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 378–382.

[8] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," *ISRN Neurosci.*, vol. 2014, Feb. 2014, Art. no. 730218.

[9] S. Taran and V. Bajaj, "Motor imagery tasks-based EEG signals classification using tunable-Q wavelet transform," *Neural Comput. Appl.*, vol. 31, no. 11, pp. 6925–6932, Nov. 2019.

[10] D. Planelles, E. Hortal, Á. Costa, A. Úbeda, E. Iáez, and J. Azorín, "Evaluating classifiers to detect arm movement intention from EEG signals," *Sensors*, vol. 14, no. 10, pp. 18172–18186, Oct. 2014.

[11] S. Bhattacharyya, A. Khasnobish, A. Konar, D. N. Tibarewala, and A. K. Nagar, "Performance analysis of left/right hand movement classification from EEG signal by intelligent algorithms," in *Proc. IEEE Symp. Comput. Intell., Cogn. Algorithms, Mind, Brain (CCMB)*, Apr. 2011, pp. 1–8.

[12] S. Bhattacharyya, A. Khasnobish, S. Chatterjee, A. Konar, and D. Tibarewala, "Performance analysis of LDA, QDA and KNN algorithms in left-right limb movement classification from EEG data," in *Proc. Int. Conf. Syst. Med. Biol. Syst. (ICSMB)*, Dec. 2010, pp. 126–131.

[13] X. Wang, M. Hersche, B. Tömekce, B. Kaya, M. Magno, and L. Benini, "An accurate EEGNet-based motor-imagery brain–computer interface for low-power edge computing," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2020, pp. 1–6.

[14] H. Y. Xu and N. P. Konstantinos, "Affective states classification using EEG and semi-supervised deep learning approaches," in *Proc. IEEE 18th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2016, pp. 1–6.

[15] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "LSTM-based EEG classification in motor imagery tasks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2086–2095, Nov. 2018.

[16] E. Bresch, U. Großekathöfer, and G. Garcia-Molina, "Recurrent deep neural networks for real-time sleep stage classification from single channel EEG," *Frontiers Comput. Neurosci.*, vol. 12, no. 85, pp. 1–12, Oct. 2018.

[17] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, pp. 1–11, Feb. 2017.

[18] T.-J. Luo, C.-L. Zhou, and F. Chao, "Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network," *BMC Bioinf.*, vol. 19, no. 1, pp. 344–361, Sep. 2018.

[19] J. Thomas, T. Maszczyk, N. Sinha, T. Kluge, and J. Dauwels, "Deep learning-based classification for brain-computer interfaces," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 234–239.

[20] W. Qiao and X. Bi, "Deep spatial-temporal neural network for classification of EEG-based motor imagery," in *Proc. Int. Conf. Artif. Intell. Comput. Sci.*, Jul. 2019, pp. 265–272.

[21] X. Shi, T. Wang, L. Wang, H. Liu, and N. Yan, "Hybrid convolutional recurrent neural networks outperform CNN and RNN in task-state EEG detection for Parkinson's disease," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 939–944.

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

[23] W. Fadel, C. Kollod, M. Wahdow, Y. Ibrahim, and I. Ulbert, "Multi-class classification of motor imagery EEG signals using image-based deep recurrent convolutional neural network," in *Proc. 8th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Apr. 2020, pp. 193–196.

[24] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for EEG decoding," 2021, *arXiv:2106.11170*.

[25] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.

[26] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using a deep attention-based LSTM network," *IEEE Sensors J.*, vol. 20, no. 6, pp. 3113–3122, Mar. 2020.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 1–9, 2019.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2020, pp. 213–229.

[30] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[31] R. Rao *et al.*, "Evaluating protein transfer learning with TAPE," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 9689–9701.

[32] M. Wairagkar, Y. Hayashi, and S. J. Nasuto, "Modeling the ongoing dynamics of short and long-range temporal correlations in broadband EEG during movement," *Frontiers Syst. Neurosci.*, vol. 13, no. 66, pp. 1–16, Nov. 2019.

[33] M. Wairagkar, Y. Hayashi, and S. J. Nasuto, "Dynamics of long-range temporal correlations in broadband EEG during different motor execution and imagery tasks," *Frontiers Neurosci.*, vol. 15, May 2021, Art. no. 660032.

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

[35] H. Mahrous and R. Ward, "Block sparse compressed sensing of electroencephalogram (EEG) signals by exploiting linear and non-linear dependencies," *Sensors*, vol. 16, no. 2, pp. 201–216, Feb. 2016.

[36] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, and G. Chen, "A channel-fused dense convolutional network for EEG-based emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 945–954, Dec. 2021.

[37] Y. Varatharajah *et al.*, "EEG-GRAPH: A factor-graph-based model for capturing spatial, temporal, and observational relationships in electroencephalograms," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5372–5381.

[38] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics Syst. Demonstrations*, Jul. 2019, pp. 37–42.

[39] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4359–4368, Mar. 2022.

[40] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D. K. Kim, "Deep learning application to clinical decision support system in sleep stage classification," *J. Personalized Med.*, vol. 12, no. 2, pp. 136–148, Feb. 2022.

[41] S. Bagchi and D. R. Bathula, "EEG-ConvTransformer for single-trial EEG based visual stimulus classification," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108757.

[42] Y.-E. Lee and S.-H. Lee, "EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech," in *Proc. 10th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2022, pp. 1–4.

[43] W. Qu *et al.*, "A residual based attention model for EEG based sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2833–2843, Oct. 2020.

[44] Y. Tao *et al.*, "Gated transformer for decoding human brain EEG signals," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 125–130.

[45] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Human Neurosci.*, vol. 15, Jun. 2021, Art. no. 653659.

[46] W. Q. Huang, W. W. Chang, G. H. Yan, Z. F. Yang, H. Luo, and H. Y. Pei, "EEG-based motor imagery classification using convolutional neural networks with local reparameterization trick," *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115968.

[47] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[48] C. A. S. Filho, R. Attux, and G. Castellano, "EEG sensorimotor rhythms' variation and functional connectivity measures during motor imagery: Linear relations and classification approaches," *PeerJ*, vol. 5, p. e3983, Nov. 2017.

[49] O. Özdenizci and D. Erdoğmus, "On the use of generative deep neural networks to synthesize artificial multichannel EEG signals," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2021, pp. 427–430.

[50] A. Prasad, "Feature extraction and classification for motor imagery in EEG signals," Ph.D. dissertation, Kauno Technologijos Universitetas, Kaunas, Lithuania, 2016, pp. 1–57.

[51] Z. U. Asi, M. Sultan, U. Muneer, A. Abbas, and S. Ilyas, "Classification of non-discriminant ERD/ERS comprising motor imagery electroencephalography signals," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 346–375, Jan. 2020.

[52] J.-R. Duann and J.-C. Chiou, "A comparison of independent event-related desynchronization responses in motor-related brain areas to movement execution, movement imagery, and movement observation," *PLoS ONE*, vol. 11, no. 9, Sep. 2016, Art. no. e0162546.

[53] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 49–72, Mar. 2019.

[54] Z. Fang *et al.*, "Learning regional attention convolutional neural network for motion intention recognition based on EEG data," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2021, pp. 1570–1576.

[55] Z.-C. Tang, C. Li, J.-F. Wu, P.-C. Liu, and S.-W. Cheng, "Classification of EEG-based single-trial motor imagery tasks using a B-CSP method for BCI," *Frontiers Inf. Technol. Electron. Eng.*, vol. 20, no. 8, pp. 1087–1098, Aug. 2019.

[56] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. C. Jun, "EEG datasets for motor imagery brain–computer interface," *Gigascience*, vol. 6, no. 7, pp. 1–8, May 2017.

[57] V. Asanza, E. Pelaez, and F. Loayza, "EEG signal clustering for motor and imaginary motor tasks on hands and feet," in *Proc. IEEE 2nd Ecuador Tech. Chapters Meeting (ETCM)*, Oct. 2017, pp. 1–5.

[58] O. Diana and A. Mihaela, "Comparison of classifiers and statistical analysis for EEG signals used in brain computer interface motor task paradigm," *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 1, pp. 8–12, Jan. 2015.

[59] O. Eva, R. Aldea, and A. Lazar, "Detection and classification of Mu rhythm for motor movement/imagery dataset," *Bull. Polytech. Inst. Jassy*, vol. 60, pp. 36–44, Jun. 2014.

[60] K. Roots, Y. Muhammad, and N. Muhammad, "Fusion convolutional neural network for cross-subject EEG motor imagery classification," *Computers*, vol. 9, no. 3, pp. 72–80, Sep. 2020.

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

[62] D. Li, P. Ortega, X. Wei, and A. Faisal, "Model-agnostic meta-learning for EEG motor imagery decoding in brain-computer-interfacing," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2021, pp. 527–530.

[63] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. L. Da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, May 2006.

[64] X. Lun, Z. Yu, T. Chen, F. Wang, and Y. Hou, "A simplified CNN classification method for MI-EEG via the electrode pairs signals," *Frontiers Human Neurosci.*, vol. 14, pp. 1–14, Sep. 2020.