Speed Imagery EEG Classification with Spatial-temporal Feature Attention Deep Neural Networks

Xiaoqian Hao and Biao Sun, *Senior Member, IEEE*
School of Electrical and Information Engineering
Tianjin University
Tianjin 300072, China
Email: sunbiao@tju.edu.cn

***Abstract*—Decoding continuous brain intentions is a major challenge for the research and application of brain-computer interfaces (BCI). Neuronal activity has been experimentally observed through various brain activity measuring techniques, of which electroencephalography (EEG) is the most widely used as it is noninvasive, practical, and has high time resolution. Here we propose a spontaneous speed imagery BCI paradigm with an EEG signals decoding method, in which a spatial-temporal feature attention deep neural network is developed to decode the continuous brain intentions. The speed imagery EEG signals of 0 Hz, 0.5 Hz and 1 Hz of left-hand clenching by 11 healthy subjects are decoded in experiments. The results reveal that the proposed method has the advantages of good performance and high efficiency, which is of great significance for patient rehabilitation and consumer applications.**

***Index Terms*—speed imagery, electroencephalography, spatial-temporal features, deep neural networks**

I. Introduction

Decoding continuous neural intentions (such as speed and force) is a major challenge in the design and development of brain-computer interfaces (BCI) for motor imagery (MI) [1]. Most MI-BCI paradigms used single body movements (such as hands, feet, and tongue), which had limitations in obtaining flexible control commands [2]–[5]. Neural intentions can be detected through a variety of signals, among which electroencephalography (EEG) was widely used in MI-BCI due to its high time resolution, portability, and spatial resolution [6]. However, the MI-BCI based EEG signals are non-stationary, time-varying and individually diverse, which makes traditional machine learning methods difficult to achieve in decoding continuous neural activity EEG signals [7]–[9]. Recently, to improve the flexibility of control commands for MI-BCI, great efforts have been made toward improving the speed imagery paradigm. Yin *et al.* proposed an inspired speed imagery paradigm and used traditional machine learning for classification [10]. Fu *et al.* further explored the state of the brain arousal and classification methods for decoding speed imagery [11]. These studies have demonstrated the feasibility of the speed imagery paradigm.

In the classification stage, most methods relied on the hand-crafted features such as wavelet transform and common spatial pattern (CSP). Wu *et al.* [12] used CSP and linear discriminant analysis (LDA) for classification and obtained a convincing performance on public datasets. Zhang *et al.* used continuous wavelet transform to extract the time-frequency features of EEG signals, and proposed a convolutional neural network with automatic channel selection and squeeze-and-excitation Blocks (ACS-SE-CNN) for classification [13]. The recently proposed EEGNet adopted a lightweight convolutional neural network to combine feature extraction and classification into one pipeline, and obtained a state-of-the-art result [14]. Bang *et al.* used a three-dimensional convolutional neural network to extract spatial spectral features and achieved reliable results on multiple public datasets [15]. A recent emerging framework based on Transformer models has shown potential in Computer Vision, natural language processing, and other fields [16]–[19]. Despite the recent enthusiasm for this area, the study of developing Transformer models for EEG classification tasks remains almost untouched. Compared with convolutional neural networks (CNN), the Transformer models deploy multi-head attention to efficiently capture the spatial-temporal dependencies within the extracted features. Hence, it is possible to quantify the long-range spatial-temporal dependencies of speed imagery EEG signals.

In this paper, we propose a spontaneous speed imagery paradigm that reflects the actual state of MI-BCI. Moreover, we propose an end-to-end spatial-temporal feature attention network framework termed as STformer, for MI-EEG classification tasks. The framework integrates CNN and Transformer to extract spatial-temporal features from global and local perspectives, thus providing the capability for extracting effective information from EEG signals and classifying MI actions with high accuracy. The rest of this paper is organized as follows: Section II describes the details of the proposed method. Section III shows the experiments and results. Section IV concludes the paper.

II. Methods

Taking full account of the spatial-temporal dependence relationship of EEG signals can help us build a suitable neural network. Studies have shown that CNN combined with

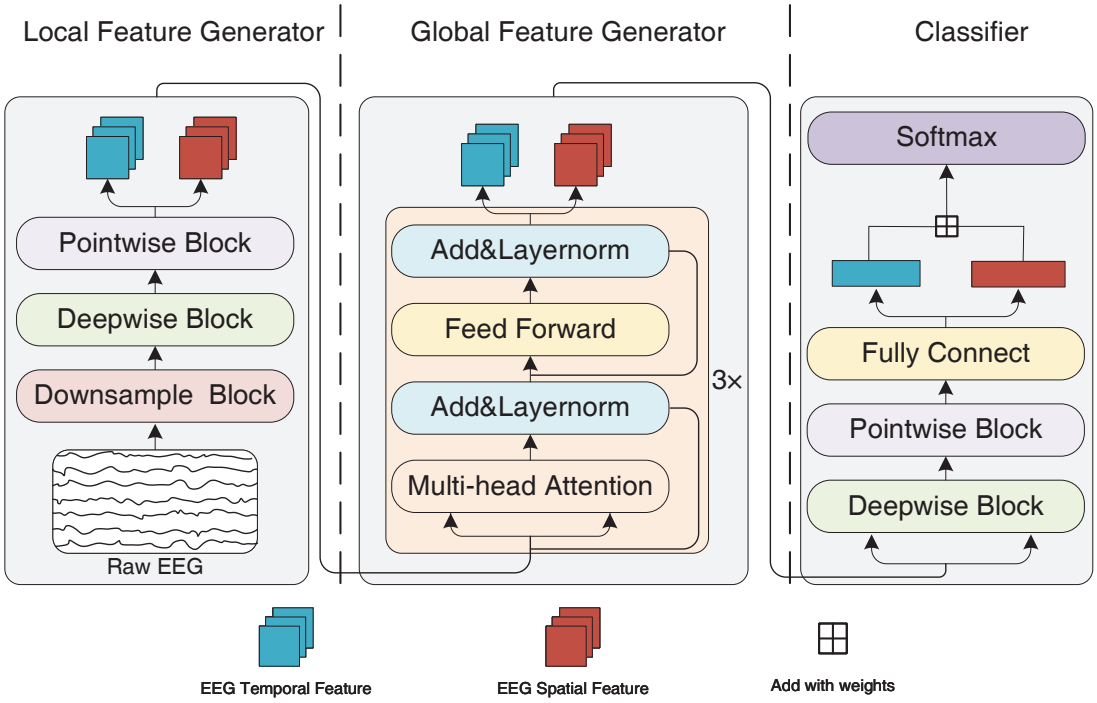

Fig. 1. Architecture of STformer with spatial-temporal attention modules.

Transformer can effectively improve network performance and make the model more interpretable [20]. Therefore, the STformer contains a local feature generator (LFG), a global feature generator (GFG) and a classifier. the architecture of STformer is shown in Fig. 1.

A. Local Feature Generator

The LFG module contains a downsample block, a deepwise block, and a pointwise block. Every block implements a one-dimensional convolution layer with both the stride size as 1 and ELU as the activation function. Different kernel size can obtain richer local feature information and reduce the amount of computation. In particular, pooling layer of downsample block and deepwise block is a MaxPooling layer with a kernel size of 2, while the pointwise block uses a AvgPooling layer with a kernel size of 4. Moreover, all blocks have a batch normalization layer and in order to prevent overfitting, dropout rate is set to 0.25. EEG spatial-temporal features are processed in parallel, and the module output is dimensionally transformed to obtain local spatial-temporal features. The input data is a $N \times C \times T$ three-dimensional tensor, where N is the number of samples, C represents the number of channels, and T represents the time sampling points of each sample. The convolution kernel performs convolution operations on the input data to obtain the local spatial-temporal features, and then applies nonlinear activation to obtain the output mapping. The feature of the k-th convolution layer is

$$\boldsymbol{H}_{i,j}^{k} = f\left(\left(\boldsymbol{W}^{k} * \boldsymbol{X}\right)_{i,j} + \boldsymbol{B}_{k}\right), \tag{1}$$

where $\boldsymbol{X}$ is the input of the network, $\boldsymbol{W}^{k}$ represents the weight matrix, $\boldsymbol{B}_{k}$ represents the bias value, and $f(\cdot)$ represents the nonlinear activation function

$$\mathrm{ELU}(x) = \begin{cases} x, x > 0, \\ \alpha\left(e^{x} - 1\right), x \leq 0. \end{cases} \tag{2}$$

B. Global Feature Generator

The GFG module aims to recalibrate the features learned by LFG for improving its performance. It stacks three identical blocks (tearmed as Transformer block) to generate the final features, where each block includes a normalization (LN) layer, a multi-head attention (MHA) layer and a feed-forward operation (FF), as shown in Fig. 1.

The MHA uses self-attention mechanism to quantify the inter-dependence within spatial-temporal features. Firstly, we set the channel matrix as position coding, which expands the module's ability to focus on different positions, thereby improving the ability to learn global dependence. Next, H heads are used to focus on different feature positions, which richens the feature focus space. Therefore, the attention weights generated by each subspace increase the importance of each subspace, and these representations are concatenated in series to produce a better overall representation, which enhances the accuracy of classification.

The input of the GFG module, denoted as $\boldsymbol{X} \in \mathbb{R}^{(C+1)\times T}$ [1], is split into H sub-matrices $\boldsymbol{X}^{h} \in \mathbb{R}^{(C+1)\times \frac{T}{H}}, h =$

[1]Note that the GFG module has two inputs: the temporal feature matrix and the spatial feature matrix. For simplicity, we will denote them both by $\boldsymbol{X}$.

$1, 2, \ldots, H$. Three proxies are calculated using the corresponding transformation matrices as $\boldsymbol{Q}^h = \boldsymbol{W}^{\text{query}}\boldsymbol{X}^h$, $\boldsymbol{K}^h = \boldsymbol{W}^{\text{key}}\boldsymbol{X}^h$, $\boldsymbol{V}^h = \boldsymbol{W}^{\text{value}}\boldsymbol{X}^h$. After that, each head calculates the attention as

$$\text{Att}^h(\boldsymbol{Q}^h, \boldsymbol{K}^h, \boldsymbol{V}^h) = \text{softmax}\left(\frac{\boldsymbol{Q}^h {\boldsymbol{K}^h}^{\text{T}}}{\sqrt{T}}\right)\boldsymbol{V}^h. \quad (3)$$

Finally, all the head representations are concatenated together to produce the final output

$$\text{MHA}(\boldsymbol{X}) = \text{Concat}\left(\text{Att}^1, \text{Att}^2, \ldots, \text{Att}^H\right). \quad (4)$$

The Transformer block has two Add & Layernorm layers to utilize the lower-layer features by propagating them to the higher layers. Additionally, the normalization operation speeds up the training process. The outputs of the multi-head attention layer are fed into a feed-forward layer. In order to break the non-linearity in the model and consider the interactions among latent dimensions, the GFG module employs ELU activation function. Both the temporal feature matrix and the spatial feature matrix are processed by the GFG module and obtained the corresponding outputs as $\boldsymbol{X}^{\text{t}}$ and $\boldsymbol{X}^{\text{s}}$.

C. Classifier

The classifier includes a deepwise block (DW), a pointwise block (PW), and a fully connected (FC) layer. Finally, the classification is performed through Softmax function. Among them, the structure of the deepwise block and pointwise block is the same as the LFG module. Before the Softmax function, the temporal feature matrix and the spatial feature matrix are mixed as

$$\boldsymbol{X}^{\text{mix}} = \boldsymbol{W}^{\text{s}}(\text{FC}\circ\text{PW}\circ\text{DW}(\boldsymbol{X}^{\text{s}})) + \boldsymbol{W}^{\text{t}}(\text{FC}\circ\text{PW}\circ\text{DW}(\boldsymbol{X}^{\text{t}})), \quad (5)$$

where $\boldsymbol{W}^{\text{s}}$ and $\boldsymbol{W}^{\text{t}}$ are trainable variables. Finally, the mixed feature $\boldsymbol{X}^{\text{mix}}$ is used as the input of Softmax for category prediction.

Since STformer is applied to classification task, the classification cross-entropy loss function is used, and the batch size is 50. In each block, the dropout operation is used to avoid overfitting [21] and the rate is set to 0.25. Additionally, the network adopts Adam optimizer with a learning rate of 0.001 for training.

III. Experiments and Results

A. Experimental protocol and setup

Subjects for this study were eleven right-handed students (five males and six females, mean age: 24.4 years, standard deviation (STD): 3.1 years, range: 21-30 years) recruited from Tianjin University. None of them had any history of neurological diseases. All EEG recording procedures are approved by the China Rehabilitation Research Center Ethics Committee (No. CRRC-IEC-RF-SC-005-01). The experimental scheme of the research is shown in Fig. 2. The speed imagery task includes three different left-hand clenching frequencies of 0 Hz, 0.5 Hz and 1 Hz. There are 15 blocks per subject and each block has 60 trials. One trail includes 7 s speed imagery and 12 s rest.

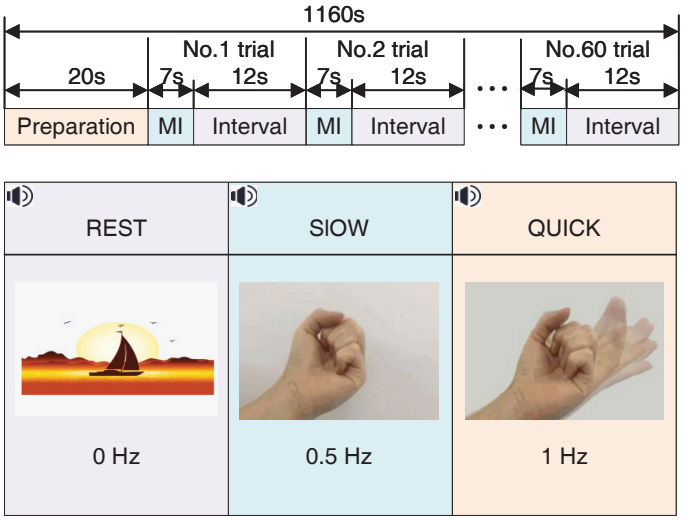

Fig. 2. Experimental scheme.

The EEG recording device is equipped with the Neuroscan system which has 34 Ag/AgCl scalp electrodes throughout the sensorimotor and supplementary motor. Out of these 34 electrodes, two are defined as reference electrodes, and two are used for monitoring eye movements. The EEG recording device are arranged according to the international standard 10/20 system and the sampling frequency of the data is 1000 Hz. The preprocessing process of the EEG signals include baseline correction, artifact rejection, filter, and epoch. These operations are carried out in Python MNE package. Specifically, in order to make the subsequent data analysis conveniently and reduce the need for computing resources, the EEG data is down-sampled to 256 Hz. Considering the actual situation, we use 6-45 Hz and 0-4 s of the data information for follow-up analysis.

B. Results

In this work, the model is programed in a Python 3.8 environment on an Intel(R) Xeon(R) W-3223 3.50-GHz CPU with 64 GB of RAM and an Nvidia GTX 3080Ti GPU. PyTorch is used for designing and testing the proposed network. During the experiments, the classifiers are trained and tested separately for each subject. The performance of the proposed method was evaluated on accuracy and area under curve (AUC) metrics. The AUC is a measure of classification performance that removes the effect of the accuracy of random classification. Two baseline methods including ACS-SE-CNN [13] and EEGNet [14] are chosen for performance comparison with the proposed STformer method. 10-fold cross validation is used, i.e., the trials are randomly partitioned into 10 equal-sized parts, of which 9 parts are selected as the training data and 1 part is kept as the validation data. The cross-validation process is repeated 10 times, with each of the 10 parts used

exactly once as the validation data. After the repetitions, the 10 results are averaged to produce the final estimate.

TABLE I
CLASSIFICATION ACCURACY (%), AUC (%) AND STANDARD DEVIATION (STD) RESULTS FOR ACS-SE-CNN, EEGNET, STFORMER, SFORMER AND TFORMER

Subject	Accuracy/AUC % (mean ± std)				
	ACS-SE-CNN	*EEGNet*	*STformer*	*Sformer*	*Tformer*
Sub.1	86.0±5.6	85.9±3.3	**92.8±5.0**	91.7±4.5	88.5±4.5
Sub.2	66.7±6.4	87.1±7.2	**95.7±5.6**	91.5±6.5	91.7±5.3
Sub.3	70.0±6.9	84.7±5.2	**96.7±4.4**	96.7±5.0	90.0±7.3
Sub.4	67.1±5.3	87.0±2.1	**91.8±3.6**	91.7±2.7	90.6±2.8
Sub.5	76.1±5.4	74.5±4.1	**82.9±4.9**	80.0±5.8	78.3±4.5
Sub.6	87.0±4.4	86.8±4.1	**91.9±4.0**	91.8±3.4	88.9±4.3
Sub.7	73.6±6.2	66.2±5.1	75.0±4.7	**77.1±3.4**	66.7±2.4
Sub.8	75.9±3.9	85.2±5.6	**95.5±5.3**	93.2±5.0	82.5±4.1
Sub.9	67.5±6.8	79.1±5.3	82.5±4.4	77.5±5.3	**90.9±4.7**
Sub.10	77.7±4.7	90.8±1.8	**91.0±2.6**	89.9±3.7	86.9±1.9
Sub.11	**78.7±4.8**	69.7±3.8	68.1±3.3	70.8±2.7	70.8±3.5
Average	75.1±5.5	81.5 ±4.3	**87.6±4.3**	86.5±4.4	84.2±4.2
AUC	83.6±4.9	91.4±4.5	**98.7±4.4**	97.8±4.5	97.6±4.3
Public	81.3±6.1	82.6±4.9	**88.1±4.4**	87.2±4.6	83.6±4.8

Note: Public represents the accuracy of the corresponding framework on the public dataset BCI competition IV 2b.

TABLE II
CLASSIFICATION ACCURACY (%) , AUC(%)AND STANDARD DEVIATION (STD) RESULTS FOR STFORMER W/O GFG, STFORMER W/O GFG AND FULL STFORMER.

Module	Accuracy % (mean ± std)	AUC % (mean ± std)
STformer w/o GFG	83.6±4.2	95.2±4.3
STformer w/o LFG	85.4±4.3	96.1±4.5
STformer	87.6±4.3	98.7±4.4

In order to evaluate the effectiveness of the speed imagery datasets and the classification performance of the proposed STformer, we first compare it with ACS-SE-CNN and EEGNet on datasets of all 11 subjects. The experimental results are shown in TABLE I. We observe that STformer outperforms the other two baseline methods, where the average accuracies of STformer, ACS-SE-CNN, and EEGNet are 87.6%, 81.5%, and 75.1% respectively. The results indicate that STformer provides a 12.5% improvement with respect to ACS-SE-CNN, and a 6.1% improvement with respect to EEGNet in terms of average accuracy. Additionally, the average AUC of STformer, ACS-SE-CNN, and EEGNet are 98.7%, 83.6%, and 91.4%, respectively. To further demonstrate the effectiveness of STformer, we validate it on the BCI competition IV 2b dataset [22], the above results show that STformer has significantly better performance and more stable data adaptability than baseline methods.

To further evaluate the contribution of spatial-temporal features to classification, we conduct a feature ablation study, where we only keep the spatial features (termed as *Sformer*) or the temporal features (termed as *Tformer*), and reapply the model to predict trials in the validation data. Classification results for this ablation study on the 11 subjects are shown in Table I. It can be found that the classification accuracy obtained by the Sformer and the Tformer are 86.5% and 84.2%, respectively. Besides, the AUC values of the Sformer and the Tformer are 97.8% and 97.6%, respectively. Evidently, the combination of spatial-temporal features can further improve the classification performance.

Moreover, to verify the effectiveness of the STformer, we consider to remove LFG (termed as STformer w/o LFG) and GFG (termed as STformer w/o GFG) respectively and compare with STformer of all 11 subjects. The average accuracies and AUCs are shown in TABLE II. We observe that removing any of the two modules decreases overall performance. Therefore, the combination of the two modules is necessary to extract more comprehensive spatial-temporal features and improve classification performance.

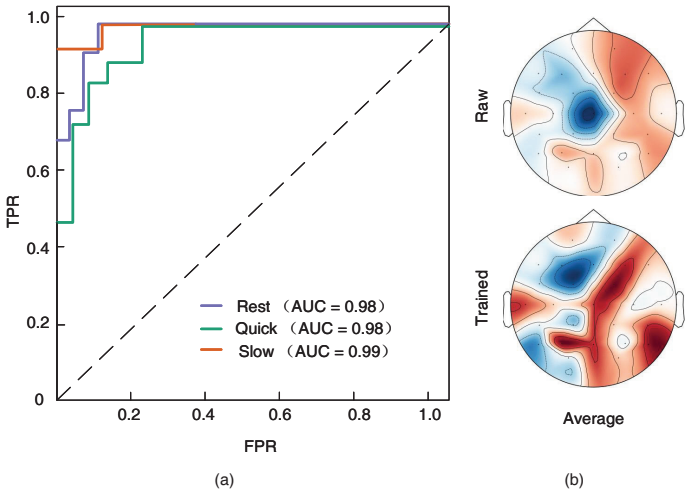

Fig. 3. (a) Classification results for the three clenching speeds. (b) Brain activation maps for raw data and trained data.

The speed imagery classification performance of the three clenching speeds is shown in Fig. 3. We observe that the speed imagery datasets have an excellent ability to capture continuous spontaneous neural intentions. The AUCs of the three classes are Slow=0.99, Rest=0.98, and Quick=0.98, respectively. In addition, from the average activation, the brain activation is more obvious than the raw EEG data. The sensorimotor area and supplementary sports area have a more obvious characterization. Importantly, it can be explained that spatial-temporal feature information can significantly improve the ability of the BCI to capture neural activity.

IV. CONCLUSION

This paper proposes a spontaneous speed imagery BCI paradigm and a decoding method for EEG signal classification. The proposed model relies on extracting the spatial-temporal features from EEG signals using a LFG module and a GFG module. We capture the spatial-temporal dependencies among the extracted features by using the two modules. Furthermore, we evaluated the effectiveness of the proposed method using an experimental dataset with 11 subjects. Results show that our method has high classification accuracy and robustness.

REFERENCES

[1] A. Schwarz, J. Pereira, R. Kobler, and G. R. Müller-Putz, "Unimanual and bimanual reach-and-grasp actions can be decoded from human eeg," *IEEE transactions on biomedical engineering*, vol. 67, no. 6, pp. 1684–1695, 2019.

[2] P. Ofner, A. Schwarz, J. Pereira, D. Wyss, R. Wildburger, and G. R. Müller-Putz, "Attempted arm and hand movements can be decoded from low-frequency eeg from persons with spinal cord injury," *Scientific reports*, vol. 9, no. 1, pp. 1–15, 2019.

[3] N. Shajil, S. Mohan, P. Srinivasan, J. Arivudaiyanambi, and A. A. Murrugesan, "Multiclass classification of spatially filtered motor imagery eeg signals using convolutional neural network for bci based applications," *Journal of Medical and Biological Engineering*, vol. 40, no. 5, pp. 663–672, 2020.

[4] Q. Ai, A. Chen, K. Chen, Q. Liu, T. Zhou, S. Xin, and Z. Ji, "Feature extraction of four-class motor imagery eeg signals based on functional brain network," *Journal of neural engineering*, vol. 16, no. 2, p. 026032, 2019.

[5] B. Sun, H. Zhang, Z. Wu, Y. Zhang, and T. Li, "Adaptive spatiotemporal graph convolutional networks for motor imagery classification," *IEEE Signal Processing Letters*, vol. 28, pp. 219–223, 2021.

[6] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[7] S. Siuly and Y. Li, "Improving the separability of motor imagery eeg signals using a cross correlation-based least square support vector machine for brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 4, pp. 526–538, 2012.

[8] M. Miao, H. Zeng, A. Wang, C. Zhao, and F. Liu, "Discriminative spatial-frequency-temporal feature extraction and classification of motor imagery eeg: An sparse regression and weighted naïve bayesian classifier-based approach," *Journal of neuroscience methods*, vol. 278, pp. 13–24, 2017.

[9] V. Peterson, D. Wyser, O. Lambercy, R. Spies, and R. Gassert, "A penalized time-frequency band feature selection and classification procedure for improved motor intention decoding in multichannel eeg," *Journal of neural engineering*, vol. 16, no. 1, p. 016019, 2019.

[10] X. Yin, B. Xu, C. Jiang, Y. Fu, Z. Wang, H. Li, and G. Shi, "A hybrid bci based on eeg and fnirs signals improves the performance of decoding motor imagery of both force and speed of hand clenching," *Journal of neural engineering*, vol. 12, no. 3, p. 036004, 2015.

[11] Y. Fu, X. Xiong, C. Jiang, B. Xu, Y. Li, and H. Li, "Imagined hand clenching force and speed modulate brain activity and are classified by nirs combined with eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1641–1652, 2016.

[12] S.-L. Wu, C.-W. Wu, N. R. Pal, C.-Y. Chen, S.-A. Chen, and C.-T. Lin, "Common spatial pattern and linear discriminant analysis for motor imagery classification," in *2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*. IEEE, 2013, pp. 146–151.

[13] H. Zhang, X. Zhao, Z. Wu, B. Sun, and T. Li, "Motor imagery recognition with automatic eeg channel selection and deep learning," *Journal of Neural Engineering*, vol. 18, no. 1, p. 016004, 2021.

[14] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[15] J.-S. Bang, M.-H. Lee, S. Fazli, C. Guan, and S.-W. Lee, "Spatio-spectral feature representation for motor imagery classification using convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[21] P. Baldi and P. J. Sadowski, "Understanding dropout," *Advances in neural information processing systems*, vol. 26, pp. 2814–2822, 2013.

[22] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz *et al.*, "Review of the bci competition iv," *Frontiers in neuroscience*, vol. 6, p. 55, 2012.