Spatio-Temporal Explanation of 3D-EEGNet for Motor Imagery EEG Classification Using Permutation and Saliency

Donghyun Park, *Graduate Student Member, IEEE*, Hoonseok Park, Sangyeon Kim, Sanghyun Choo, Sangwon Lee, Chang S. Nam, *Senior Member, IEEE*, and Jae-Yoon Jung

***Abstract*— Recently, convolutional neural network (CNN)-based classification models have shown good performance for motor imagery (MI) brain-computer interfaces (BCI) using electroencephalogram (EEG) in end-to-end learning. Although a few explainable artificial intelligence (XAI) techniques have been developed, it is still challenging to interpret the CNN models for EEG-based BCI classification effectively. In this research, we propose 3D-EEGNet as a 3D CNN model to improve both the explainability and performance of MI EEG classification. The proposed approach exhibited better performances on two MI EEG datasets than the existing EEGNet, which uses a 2D input shape. The MI classification accuracies are improved around 1.8% and 6.1% point in average on the datasets, respectively. The permutation-based XAI method is first applied for the reliable explanation of the 3D-EEGNet. Next, to find a faster XAI method for spatio-temporal explanation, we design a novel technique based on the normalized discounted cumulative gain (NDCG) for selecting the best among a few saliency-based methods due to their higher time complexity than the permutation-based method. Among the saliency-based methods, DeepLIFT was selected because the NDCG scores indicated its results are the most similar to the permutation-based results. Finally, the fast spatio-temporal explanation using DeepLIFT provides deeper understanding for the classification results of the 3D-EEGNet and the important properties in the MI EEG experiments.**

***Index Terms*— Brain–computer interfaces (BCI), motor imagery (MI), convolutional neural network (CNN), electroencephalogram (EEG), explainable artificial intelligence (XAI).**

Manuscript received 21 January 2023; revised 30 June 2023 and 10 August 2023; accepted 4 September 2023. Date of publication 7 November 2023; date of current version 16 November 2023. This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the High-Potential Individuals Global Training Program, under Grant 2020001560; and in part by the Artificial Intelligence Convergence Innovation Human Resources Development, Kyung Hee University, supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant RS-2022-00155911. *(Corresponding author: Jae-Yoon Jung.)*

Donghyun Park and Hoonseok Park are with the Department of Big Data Analytics, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Republic of Korea (e-mail: pdh@khu.ac.kr; hoonseok@khu.ac.kr).

Sangyeon Kim, Sanghyun Choo, and Chang S. Nam are with the Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: noizowl@gmail.com; schoo2@ncsu.edu; csnam@ncsu.edu).

Sangwon Lee is with the School of Industrial and Management Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: upcircle@korea.ac.kr).

Jae-Yoon Jung is with the Department of Industrial and Management Systems Engineering and the Department of Big Data Analytics, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Republic of Korea (e-mail: jyjung@khu.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3330922, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3330922

I. Introduction

MOTOR imagery (MI) refers to a mental action in which humans imagine movements of their body parts without any actual movements [1]. MI has been actively studied to identify the movement intentions of patients who have limited ability to perform physical movements (e.g., patients with cerebral palsy or strokes) [2], [3]. Electroencephalogram (EEG) is widely used in research on MI brain-computer interfaces (BCIs). EEG has a good time resolution on the order of milliseconds and can be collected more easily compared to invasive methods that measure signals from the surface of the brain directly. However, EEG involves many noisy signals and has a high data dimension. So, it is required to extract effective features that represent the characteristics of EEG signals [4]. Hence, many studies have adopted various feature extraction methods [5] for extracting suitable hand-crafted features and have developed various machine learning classifiers using the extracted features [6], [7], [8].

However, deep learning (DL) approaches such as convolutional neural networks (CNNs) can automatically extract useful features from the input through convolutional operation and can be trained through end-to-end learning in which raw data is directly used without complex feature engineering techniques [9]. Therefore, CNNs have made EEG classification models to extract various complex features more easily [10], [11], [12], [13], [14].

Nevertheless, CNNs are typically difficult to understand their internal mechanism due to their complex structures. To solve this limitation, many studies on EEG classification have applied explainable artificial intelligence (XAI) methods to understand how their CNN models work [15], [16], [17]. The CNN models for EEG classification generally have been explained in the spatial (e.g., which EEG channels were

important in the classification results) and temporal (e.g., which time intervals were important) dimensions [18].

In this study, 3D-EEGNet is proposed as an improved classification model for MI EEG. The 3D-EEGNet architecture extends the structure of EEGNet [14] for better explainability in EEG classification. First of all, the three-dimensional (3D) shape of input data preserves the spatial and temporal information of the EEG data. As a result, the experiment showed that the 3D shape could improve the accuracy of MI EEG classification as well as spatio-temporal explanatory power. By applying XAI methods to the 3D-EEGNet, it can be explained in terms of spatial (channel) and temporal (time) aspects.

In the meantime, to explain the 3D-EEGNet model in this research, two kinds of XAI methods, permutation and saliency-based methods, are considered. First, the *permutation method*, which is more accurate than the saliency-based ones, is used to interpret the model in spatial or temporal aspects. However, the permutation method is difficult to apply to inspect spatial and temporal features at the same time due to its high computational complexity. For that reason, *saliency-based methods*, which are much faster than the permutation method, are additionally adopted for the spatio-temporal explanation of 3D-EEGNet instead of the permutation method. Unfortunately, the saliency-based methods cannot guarantee their reliability [19]. Hence, in this study, a novel evaluation process is also proposed to choose the best XAI method for the spatio-temporal explanation of 3D-EEGNet. Specifically, the normalized discounted cumulative gain (NDCG) score [20] is used to compare the rank of important spatial and temporal features.

The main contribution of this study is two folds.

- **Improvement of explainability and classification performance of MI EEG classification**: The proposed 3D-EEGNet performs better in terms of the accuracy of MI classification, and also maintains the spatial and temporal properties of the original data in the 3D data shape for better explainability.
- **Novel evaluation process for selecting XAI methods based on rank**: An evaluation process is presented to choose the best XAI among a few saliency-based methods, specifically by using the NDCG score. Through this process, the computational cost for the spatio-temporal explanation can be reduced in consideration of reliability.

The remainder of this paper is structured as follows. In Section II, previous studies on MI EEG classification and XAI methods are described. In Section III, the framework for developing 3D-EEGNet and explaining its output is introduced. In Section IV, the detailed architecture of 3D-EEGNet is presented. The explanation methods for the 3D-EEGNet model are described in Section V. The experimental results are illustrated in Section VI with spatial and temporal explanations of the 3D-EEGNet. Finally, Section VII concludes this paper with future work.

II. Related Work

A. EEG Classification for Motor Imagery

The classification of MI EEG signals is difficult because of a few reasons such as limited spatial resolution and low signal-to-noise ratio (SNR) [17]. To handle these issues, conventional machine learning (ML)-based MI EEG classification models have employed common spatial pattern (CSP) features and its variants to mainly extract spatial information of MI EEG signals [21], [22]. However, such classifiers require comprehensive prior knowledge on EEG because the features are commonly extracted by hand-crafted ways which depend on experimental subjects and hardware settings.

DL-based end-to-end EEG classifiers have been used to overcome this problem because of their ability to automatically extract useful features. Particularly, CNNs have received the most attention owing to their advantages in effectively reflecting structural information of the input data. In EEG classification, DeepConvNet, ShallowConvNet [23], and EEGNet [14] are popular CNNs models. They commonly used minimally preprocessed EEG signals, and they used spatial and temporal convolutional layers to extract meaningful features of EEG. In this study, the architecture of EEGNet is used as the base structure of MI classification model to extract EEG features effectively.

B. XAI for EEG Classification

Most XAI methods have been introduced in the fields of computer vision and natural language process [24]. By all means, in BCI applications, XAI is also adopted to obtain transparency of CNNs models [25]. One of the early studies that addressed the interpretability of DL models on EEG signals investigated the application of layer-wise relevance propagation (LRP) [26], [27]. Studies applying LRP have generally provided heatmaps that represent the feature importance of each EEG channel in a single experimental trial [28], [29]. This visualization using heatmaps has been widely used to show their models work well compared with common methods such as CSP and FBCSP [30]. This helped to understand how the DL models exploit specific channels and time points for their classification tasks. In this context, as XAI methods have been enhanced, various methods such as DeepLIFT [31], SHAP [32], and Grad-CAM [33] have also been applied in the BCI applications to understand DL models for EEG [34], [35].

C. Model-Agnostic XAI: Permutation Method

Model-agnostic XAI methods are applicable to various DL models regardless of the type of models. The permutation method explains the models by arbitrarily removing or manipulating specific features. The method is not only straightforward, but also provides an excellent explanation of models in terms of the connection between the feature information and their performance. Therefore, in this study, the permutation method is used as a ground-truth method to explain the proposed MI EEG classification model. However, because of its high computational complexity of having to permuting features one by one, it cannot be used to explain the huge number of combinations of spatial and temporal features.

The permutation method for time-series data was introduced in [39]. The authors showed three permutation techniques, zero, swap, and inverse permutations. They showed which data points are important for the ML model's results. However, in EEG, there is no meaning to indicate certain time points

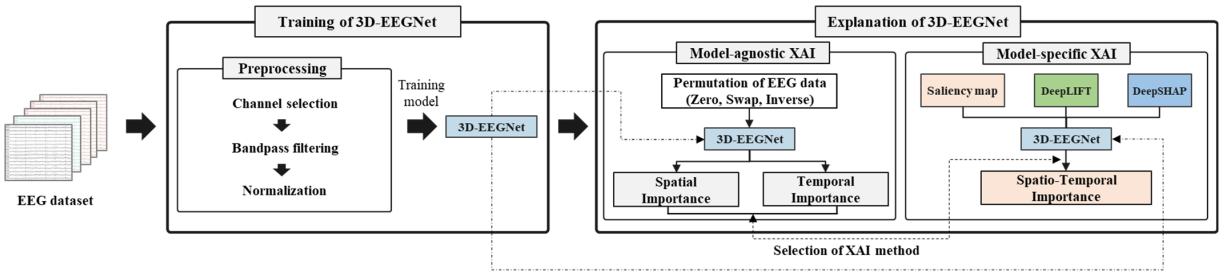


Fig. 1. Framework of training and explanation of 3D-EEGNet.

as important because users are only interested in important spatial and temporal information (channels and time intervals). Therefore, in this study, the three permutation techniques are applied to sets of time points in spatial and temporal dimensions.

D. Model-Specific XAI for CNN: Saliency-Based Methods

Model-specific XAI methods are dependent on the target algorithms. Among the methods, saliency-based methods are widely used to explain CNNs models [36]. The methods calculate the influence of the input on the prediction output. In this study, three saliency-based methods, saliency map, DeepLIFT, and DeepSHAP, are considered XAI for CNN models.

1) Saliency Map: This method uses the gradients of trained CNNs model to represent the importance of the input on the prediction result. For instance, the importances of input image pixels are calculated by multiplying an image with the weight vectors which are calculated using backpropagation [37]. The saliency map is created by summing the weight values of each image pixel. This method requires only a single backpropagation to obtain the importances of all the input pixels, so the calculation is simple and fast.

2) DeepLIFT: Deep Learning Important FeaTures (DeepLIFT) is designed to obtain the importance of input in the prediction of CNNs models. DeepLIFT uses multipliers that represent a slope describing how the outputs are changed when the inputs are different from reference data. In this study, DeepLIFT is also considered to explain the MI EEG classification model due to its good mathematical background and intuitive concept as well as its ease of application.

3) DeepSHAP: DeepSHAP (DeepLIFT + SHAP) is a method of approximating SHAP (SHapley Additive exPlanations) values using DeepLIFT. The calculation of the Shapley values is an expensive process because of the massive number of combinations of features. DeepLIFT makes it possible to obtain the Shapley values by approximating them using multipliers. In this study, DeepSHAP is also regarded as the explanation method because it is based on the Shapley values, which are mathematically proven to satisfy desirable properties for explaining models.

III. Framework

The proposed framework for developing and explaining a MI EEG classification model is shown in Fig. 1. In the first stage, an EEG dataset is preprocessed by channel selection, bandpass filtering, and normalization, and it is then used to train 3D-EEGNet for MI EEG classification. Several electrode channels located on the sensorimotor cortex are selected. The MI-related frequency bands, the alpha band (8-12 Hz), beta band (16-24 Hz), and gamma band (30-35 Hz) [38], are then extracted by finite impulse response (FIR) bandpass filtering [39]. Finally, the filtered data are normalized by the z-normalization for stable training. The preprocess EEG data is used as the input for training the proposed 3D-EEGNet model.

The next stage is the explanation of the trained 3D-EEGNet model. The XAI methods determine which brain areas (spatial importance) and time intervals in the EEG signal (temporal importance) are important to obtain the prediction results. As a model-agnostic XAI, the permutation method is used to obtain the spatial and temporal importance separately. To obtain spatial and temporal importance of EEG data, the data are permuted channel-wise and time-wise.

In contrast, three saliency-based methods, saliency map, DeepLIFT, and DeepSHAP are considered to obtain the spatio-temporal explanation for the 3D-EEGNet because they are much faster than the permutation method. As the saliency-based methods have been found to be unreliable in a recent study due to the unsatisfaction of input invariance [19], however, the best one is selected for faster, but reliable spatio-temporal explanation after comparison with the permutation method. In the comparison, the spatial and temporal importances are compared by the ranks of the important channels and time intervals because the scales of the XAI methods are different, making it difficult to compare their results directly.

The similarity of the ranks between the saliency-based methods and the permutation method is measured by the NDCG score, which is commonly used for evaluating recommendation systems. The NDCG score assigns more weights to higher ranks to focus on the high-rank items. The NDCG score is appropriate for the EEG XAI because users want to focus more on only important channels and time intervals rather than unimportant ones in terms of ranking.

The saliency-based method with the highest NDCG score is selected as the best XAI method. It is finally used to explain the 3D-EEGNet model in the spatio-temporal dimension so that we can understand the behavior of the 3D-EEGNet and the properties of the experimental subjects.

IV. 3D-EEGNet for EEG Classification

A. Design of 3D-EEGNet

The 3D-EEGNet is a variant of EEGNet, which is well-known as a compact CNNs model for EEG classification [14].

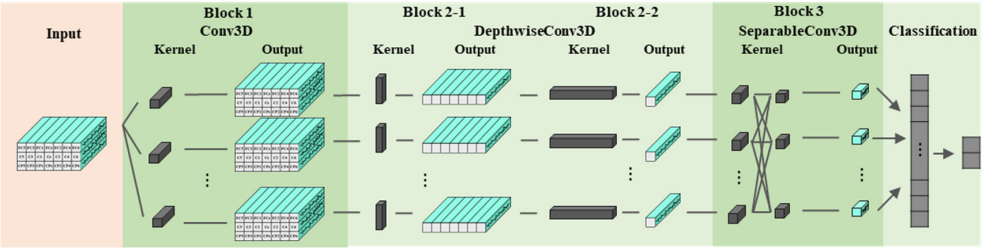


Fig. 2. Architecture of 3D-EEGNet.

The successful structure of EEGNet includes depthwise convolutional layers and separable convolutional layers, which enable to reduce the model's parameters and extract powerful features of EEG signals.

While the input of EEGNet has two dimensions (channels and time), the input of 3D-EEGNet is designed to have three dimensions so that the model can preserve the spatial information of channels. Although several studies have shown the positive effects of 3D input representation in EEG model developments [16], [40], the main motivation of the 3D input for 3D-EEGNet is the activated brain area when MI.

The 2D structure of the original EEGNet (herein, 2D-EEGNet) learns spatial and temporal features in two dimensions, respectively. However, the receptive fields of EEG itself can be considered the 2D structure as shown in Fig. 3. In other words, 2D-EEGNet uses a single layer (DepthwiseConv2D) for spatial learning of EEG signals, while 3D-EEGNet does two layers (Blocks 2-1 and 2-2 in DepthwiseConv3D layer) for vertical and horizontal spatial learning, respectively. Therefore, we think that the 3D structure of our proposed 3D-EEGNet might be beneficial to reflect the 2D spatial features and 1D temporal features.

Additionally, the convolutional filters of 3D-EEGNet are trained to capture the event-related desynchronization (ERD) of alpha (often called mu) and beta frequency bands which are measured in the primary sensorimotor area [41]. When humans imagine a kind of movement, ERD quantifies the task-related power of the frequency bands by showing a temporary decrease in the signal, which is observed at the onset of events. Note that the sensorimotor area where ERD is significant is related to the type of MI. For example, when movement of the right hand is imagined, ERD is known to be significant in the left sensorimotor area [42]. In the case of foot MI, the center of the sensorimotor area is activated.

B. Architecture of 3D-EEGNet

The architecture of 3D-EEGNet is shown in Figure 2, and the details of each layer are summarized in Table I. In block 1, the input data are shaped in (H, W, T), where H and W are the height and width in the channel dimension, respectively, and T is the length of the time dimension. In the first 3D convolutional layer, each temporal filter learns the temporal information of the specific frequency of MI with F_1 filters.

In block 2-1, (H, 1, 1) spatial filters in the depthwise convolutional layer learn the vertical (front-to-back of the head) information of the sensorimotor area. The number of filters, D, is set to W, which means that one filter only learns one vertical information, excluding the horizontal (ear-to-ear) information. The spatial filters are applied to each output given from the block 1, not to all the outputs (this is the depthwise convolution).

TABLE I
DESIGN OF THE 3D-EEGNET

Block	Layer	# filters	Size	Output	Act. ftn.	Options
1	Input			(H, W, T)		
	Reshape			(H, W, T, 1)		
	Conv3D	F_1	(1, 1, 13)	(H, W, T, F_1)	linear	mode=same
	BatchNorm			(H, W, T, F_1)		
2-1	DepthwiseConv3D	$D\times F_1$	(H, 1, 1)	(1, W, T, $F_1\times D$)	linear	mode=valid, depth=D
	BatchNorm			(1, W, T, $F_1\times D$)		
2-2	DepthwiseConv3D	$D\times F_1$	(1, W, 1)	(1, 1, T, $F_1\times D$)		
	BatchNorm			(1, 1, T, $F_1\times D$)		
	Activation			(1, 1, T, $F_1\times D$)	ELU	
	AveragePool3D		(1, 1, 4)	(1, 1, T/4, $F_1\times D$)		
	Dropout			(1, 1, T/4, $F_1\times D$)		p=0.5
3	SeparableConv3D	$D\times F_1$	(1, 1, 4)	(1, 1, T/4, $F_1\times D$)	linear	mode=same
	BatchNorm			(1, 1, T/4, $F_1\times D$)		
	Activation			(1, 1, T/4, $F_1\times D$)	ELU	
	AveragePool3D		(1, 1, 12)	(1, 1, T/48, $F_1\times D$)		
	Dropout			(1, 1, T/48, $F_1\times D$)		
4	Flatten			(T/48, $F_1\times D$)		
	Dense			N	softmax	max_norm =0.25

In block 2-2, (1, W, 1) spatial filters learn the horizontal features of the sensorimotor area by depthwise convolution. The number of filters is set to be the same as in the previous layer (block 2-1) because the output of block 2-1 has only one horizontal information (H becomes 1 in block 2-1). Ultimately, the output size of the block 2-2 becomes (1, 1, T/4, $F_1 \times D$), which means that the spatial information is summarized into a single value.

In block 3, two types of filters are learned by separable convolution: one of which summarizes the outputs from the previous layer, and the other reduces the dimension by using (1, 1, 4) feature maps. Subsequently, the outputs of the separable convolutional layer are flattened and connected to the dense layer for classification in block 4.

V. EXPLANATION OF 3D-EEGNET

A. Explanation Using Permutation

This section describes how important channels and time intervals in EEG data are recognized in the trained 3D-EEGNet model by using the permutation method.

1) Spatial Explanation: The spatial explanation of 3D-EEGNet is provided through important channels. Suppose an EEG dataset $\mathbf{X}$ for each subject includes I trials, J channels, and T times of EEG signals, denoted by $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_I\}$, where $\mathbf{X}_i = \{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^J\}$ is the i-th trial of EEG data, $\mathbf{x}_i^j = (x_{i,1}^j, \ldots, x_{i,T}^j)$ is the j-th channel vector of $\mathbf{X}_i$, and $x_{i,t}^j$ is the signal value of $\mathbf{x}_i^j$ at time t.

Three permutation techniques, *zero*, *swap*, and *inverse*, are applied to all channels one by one [43]. The *zero* permutation for spatial explanation replaces $\mathbf{x}_i^j$ with a zero vector of the same size as in (1). The *swap* permutation changes $\mathbf{x}_i^j$ in the reverse order as in (2). The *inverse* permutation inverts the values of $\mathbf{x}_i^j$ by subtracting each value from the maximum value of $\mathbf{x}_i^j$ as in (3).

$$f_{zero}(\mathbf{x}_i^j) = (0, \ldots, 0)\,, \tag{1}$$

$$f_{swap}(\mathbf{x}_i^j) = (x_{i,T}^j, \ldots, x_{i,1}^j), \tag{2}$$

$$f_{inverse}(\mathbf{x}_i^j) = (\max_t (x_{i,t}^j) - x_{i,1}^j, \ldots, \max_t (x_{i,t}^j) - x_{i,T}^j) \tag{3}$$

To determine the importance of the j-th channel, three types of the permuted data, $\mathbf{X}_{zero}^{-j}$, $\mathbf{X}_{swap}^{-j}$, and $\mathbf{X}_{inverse}^{-j}$, are prepared as follows.

$$\mathbf{X}_{type}^{-j} = \left\{\mathbf{X}^1, \ldots, f_{type}(\mathbf{X}^j), \ldots, \mathbf{X}^J\right\}, \tag{4}$$

$$\mathbf{X}^j = \left\{\mathbf{x}_1^j, \ldots, \mathbf{x}_I^j\right\},$$

$$f_{type}(\mathbf{X}^j) = \{f_{type}(\mathbf{x}_1^j), \ldots, f_{type}(\mathbf{x}_I^j)\}$$
$$\text{for } type \in \{zero, swap, inverse\}, \tag{5}$$

where $\mathbf{X}_{type}^{-j}$ is the permuted data for the j-th channel by replacing all the j-th channel data $\mathbf{X}^j$ with one of the three channel permutation functions, f_{zero}, f_{swap}, or $f_{inverse}$.

Subsequently, the three permuted data for the j-th channel, $\mathbf{X}_{zero}^{-j}$, $\mathbf{X}_{swap}^{-j}$, and $\mathbf{X}_{inverse}^{-j}$, are used to evaluate the change in the accuracy of the MI classification model. If a specific feature is important, the prediction performance would be significantly decreased when the corresponding permutated data is given as input of the model. The accuracy change for the j-th channel permutation, denoted by Δ_{type}^j, is measured by the difference between the accuracy of the original model, $acc(\mathbf{X})$, and the accuracy of the model trained by the j-th channel permuted data, $acc(\mathbf{X}_{type}^{-j})$.

$$\Delta_{type}^j = acc\,(\mathbf{X}) - acc(\mathbf{X}_{type}^{-j}) \tag{6}$$

Finally, the permutation importance of the j-th channel, denoted by I_{pt}^j, is evaluated by the min-max normalization of the average of three accuracy changes for the j-th channel, Δ_{zero}^j, Δ_{swap}^j, and $\Delta_{inverse}^j$.

$$\Delta^j = (\Delta_{zero}^j + \Delta_{swap}^j + \Delta_{inverse}^j)/3 \tag{7}$$

$$I_{pt}^j = \frac{\Delta^j - min(\Delta^j)}{max(\Delta^j) - min(\Delta^j)} \tag{8}$$

2) Temporal Explanation: The temporal explanation of 3D-EEGNet is performed by identifying the important time intervals in EEG signals. To evaluate the importance of a time interval, one can extract a subsequence $\tilde{\mathbf{x}}_i^{j,k}$ from the j-th channel vector of the i-th trial, $\mathbf{x}_i^j = (x_{i,1}^j, \ldots, x_{i,T}^j)$, as follows.

$$\tilde{\mathbf{x}}_i^{j,k} = \left(x_{i,t(k)+1}^j, \ldots, x_{i,t(k)+s}^k\right), \tag{9}$$

where $\tilde{\mathbf{x}}_i^{j,k}$ is the k-th time interval in $\mathbf{x}_i^j$ for k= 1 to K ($\ll T$), s is the size of a time interval, and $t\,(k) = s\,(k-1)$.

The goal of the temporal explanation is to evaluate the importance of K time intervals across all I trials. Each time interval $\tilde{\mathbf{x}}_i^{j,k}$ can also be permuted by the three types of permutation in the same way as the channel permutation.

Similar to (1)-(3), three types of permutations for $\tilde{\mathbf{x}}_i^{j,k}$, $f_{type}(\tilde{\mathbf{x}}_i^{j,k})$ for $type \in \{zero, swap, inverse\}$, can be applied to obtain three time-interval permutation data for the k-th time interval, $\mathbf{X}_{type}^{-k}$, similarly to (4)-(5). By using the permutation data, the change in the accuracy for the time interval, Δ_{type}^k, is calculated, and the importance of the k-th time interval, I_{pt}^k, can also be evaluated similarly to (6)-(8).

B. Explanation Using Saliency

The saliency-based methods are applied to the trained 3D-EEGNet. The test data is given to the model as inputs, and the saliency-based methods generate the gradients (Saliency map), attribution scores (DeepLIFT), or SHAP values (DeepSHAP) with respect to the inputs. When the saliency-based methods are applied to EEG signals, one signal point is considered as a single feature. In other words, they generate the importance of each signal point.

1) Spatial Explanation: Spatial explanation can be conducted by identifying important channels using signal point importance. The importance of a single signal point $x_{i,t}^j$ is denoted as $v_{i,t}^j$. The importance of the j-th channel, also known as attribution, α^j, can be calculated by the summation of $v_{i,t}^j$'s in all the trials and time as follows.

$$\boldsymbol{\alpha} = \left(\alpha^1, \ldots, \alpha^j, \ldots, \alpha^J\right), \tag{10}$$

$$\alpha^j = \sum_{i=1}^{I}\sum_{t=1}^{T} v_{i,t}^j, \tag{11}$$

where $v_{i,t}^j$ is the importance of channel j on trial i at time t.

Finally, the importance of the j-th channel, I_{sal}^j, calculated by a saliency-based method *sal*, is obtained by normalizing the attribution value into a relative value between 0 and 1. It is performed to obtain relative values for each saliency-based method to compare the importance among the methods.

$$I_{sal}^j = \frac{\alpha^j - \min\left(\alpha^j\right)}{\max(\alpha^j) - \min(\alpha^j)}$$
$$\text{for } sal \in \{SaliencyMap, DeepLIFT, DeepSHAP\}. \tag{12}$$

2) Temporal Explanation: The temporal explanation of 3D-EEGNet is conducted in a similar way to the spatial explanation. The EEG signals were divided into K time intervals as (9), and the importances of the time intervals are determined.

Similar to (10)-(12), the importance of the k-th time interval is obtained by the attribution of the k-th time interval, α^k,

calculated by the summation of $v_{i,t}^k$'s in all the trials and time. The importance of the k-th time interval, I_{sal}^k, calculated by a saliency-based method, is also obtained by transforming the attribution value through the min-max normalization.

C. Selection of the Best Saliency-Based Method

This section describes the selection process of the best saliency-based method. In specific, using the NDCG score [20], the results of three saliency-based methods are compared with that of the permutation method to investigate their consistencies.

The importance of the j-th channel calculated by the permutation method, I_{pt}^j, is compared with importances of three saliency-based methods, I_{sal}^j's. The ideal DCG (IDCG) is calculated using I_{pt}^j because the I_{pt}^j is considered the ground-truth. According to the magnitude of I_{pt}^j, the j-th channel is ranked as r_{pt}^j, and the spatial IDCG is calculated as $IDCG^{spt} = \sum_{j=1}^{J} \left(I_{pt}^j / \log_2 \left(r_{pt}^j + 1 \right) \right)$. Similarly, according to the I_{sal}^j obtained from a saliency-based method, the j-th channel is ranked as r_{sal}^j. The spatial DCG of the saliency-based method is calculated as $DCG_{sal}^{spt} = \sum_{j=1}^{J} \left(I_{pt}^j / \log_2 \left(r_{sal}^j + 1 \right) \right)$. Finally, the spatial NDCG score representing the similarity of the ranks between the permutation method and the saliency-based method in terms of spatial explanation is calculated as $NDCG_{sal}^{spt} = DCG_{sal}^{spt} / IDCG^{spt}$.

Similar to spatial importance, for the comparison of temporal importance, I_{pt}^k is compared with the temporal importances of three saliency-based methods, I_{sal}^k's. The IDCG is calculated using I_{pt}^k because the I_{pt}^k is considered the ground truth. According to the I_{pt}^k obtained from a saliency-based method, the k-th time interval is ranked as r_{pt}^k, and the temporal IDCG is calculated as $IDCG^{tpr} = \sum_{k=1}^{K} \left(I_{pt}^k / \log_2 \left(r_{pt}^k + 1 \right) \right)$. And the temporal DCG of the saliency-based method is calculated as $DCG_{sal}^{tpr} = \sum_{k=1}^{K} \left(I_{pt}^k / \log_2 \left(r_{sal}^k + 1 \right) \right)$. Finally, the temporal NDCG score of the saliency-based method is calculated as $NDCG_{sal}^{tpr} = DCG_{sal}^{tpr} / IDCG^{tpr}$.

VI. Experiments

A. Datasets

In the experiment, the BCI Competition III-IVa (BCIC) dataset [44] and GigaDB dataset [45] were used. The BCIC dataset was collected from five subjects (aa, al, av, aw, and ay). The subjects were instructed to imagine movements of their right hand and right foot according to visual cues. Each subject conducted 280 trials (right hand: 140, right foot: 140) and the visual cues appeared for 3.5 seconds. The cues were randomly given by periods of 1.75 to 2.25 seconds for the subjects to relax. The EEG signals were collected from 118 electrode channels of the extended 10/20 system [46] at 100 Hz.

GigaDB involves the MI EEG of 52 subjects collected with 64 electrode channels at 512 Hz. Each subject conducted 100 or 120 trials (right hand, left hand) and the cue appeared for 3 seconds. In this paper, only the five subjects (s01-s05) were used to evaluate the performance of 3D-EEGNet.

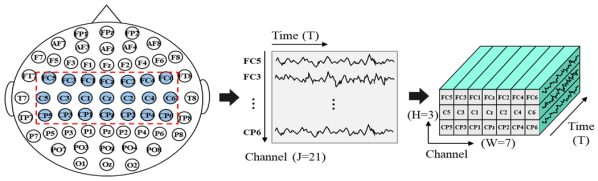


Fig. 3. Transforming the input shape into 3D.

To consider only the channels located on the sensorimotor cortex of the brain, 21 channels (FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, and CP6) were selected among the many channels, as depicted in Fig. 3. It helps to prevent the other channels from influencing the model with noisy signals.

The frequency bands, alpha (8-12 Hz), beta (16-24 Hz), and gamma (30-40 Hz) bands, are known to be significant to MI analysis [47]. Therefore, to involve the above frequency bands, the range of 8–40 Hz was extracted by using band-pass filtering. Then, the EEG signals were normalized to assign the mean and standard deviation of the EEG signals as 0 and 1, respectively.

The original shape of the dataset is represented as (trials I, channels $J = 21$, time length T). To train the 3D-EEGNet, the input shape is transformed to (I, $W = 3$, $H = 7$, T), where the channel dimension is converted to two dimensions. The (3,7) channel shape was designed based on the actual location of those channels on the sensorimotor cortex, as shown in Fig. 3.

B. Training of 3D-EEGNet

The hyperparameters of the 3D-EEGNet were set to consider the properties of MI EEG. The number of temporal filters in the first 3D convolutional layer (F_1) was set to 24 to match the number of frequencies mainly related to the MI (alpha band: 8-12 Hz, beta band: 16-24 Hz, gamma band: 30-40 Hz).

The length of the temporal filters was set depending on the EEG sampling rate. The 8 Hz EEG in the alpha band was the longest frequency signal. This means that the 8 Hz signal is shown every 12.5 signal points in the 100 Hz EEG (BCIC). Therefore, the length of temporal filters was set to 13 because the filter length should be an integer. In the result, the length 13 temporal filters can involve the alpha, beta, and gamma bands altogether. Likewise, in the GigaDB, the length of the temporal filters was set to 64 (512 Hz / 8).

The number of pointwise filters in the separable convolutional layer was set to $D \times F_1$, which means learning the representation as much as the number of inputs, as suggested in [14]. Batch normalization was applied after the convolutional layers, and dropout was applied after blocks 2 and 3 to prevent overfitting. The ELU [48] was used as the activation function before applying average pooling.

The 3D-EEGNet model was trained on each subject's data. The 10-fold cross-validation was used to validate the models' performances objectively, and the early stopping technique was used to prevent overfitting. The batch size was 16, and Adam optimizer [49] was used. The learning rate of each subject differed as {aa:1e-3, al:1e-3, av:5e-4, aw:5e-2, ay:5e-2 in BCIC; s01-05: 1e-3 in GigaDB} to optimize the model training.

TABLE II
ACCURACY COMPARISON OF EEGNET AND 3D-EEGNET

(A) BCIC

Subj.	EEGNet	3D-EEGNet
	Mean ± Std.	Mean ± Std.
aa	0.9250 ± 0.0489	0.9250 ± 0.0460
al	0.9614 ± 0.0439	0.9679 ± 0.0594
av	0.8071 ± 0.0828	0.8357 ± 0.0656
aw	0.9107 ± 0.0636	0.9393 ± 0.0447
ay	0.9214 ± 0.0499	0.9457 ± 0.0405
avg.	**0.9051 ± 0.0578**	**0.9227 ± 0.0512**

(B) GIGADB

Subj.	EEGNet	3D-EEGNet
	Mean ± Std.	Mean ± Std.
s01	0.6418 ± 0.0575	0.7368 ± 0.1067
s02	0.6361 ± 0.1159	0.6061 ± 0.0934
s03	0.7474 ± 0.1110	0.9242 ± 0.0645
s04	0.7374 ± 0.0845	0.8032 ± 0.0829
s05	0.9950 ± 0.0158	0.9950 ± 0.0158
avg.	**0.7515 ± 0.0769**	**0.8131 ± 0.0727**

TABLE III
ABLATION STUDY ON THE ACCURACY OF 3D-EEGNET MODEL

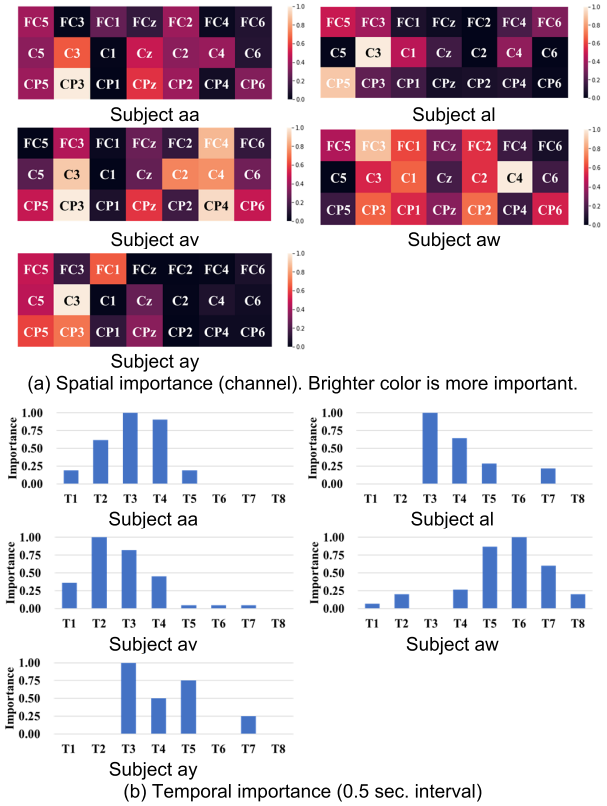
(A) BCIC

Subj.	Current model	Ablated models					
		Size of filters in Block 1		# of filters in Block 1		# of filters in Block 2	
	(13, 24, 7)	(**8**, 24, 7)	(**16**, 24, 7)	(13, **20**, 7)	(13, **28**, 7)	(13, 24, **5**)	(13, 24, **9**)
aa	0.9250	0.8964	0.9143	0.9250	0.9107	0.9286*	0.9107
al	0.9679	0.9643	0.9714	0.9679	0.9750	0.9750	0.9786*
av	0.8357	0.8250	0.8393*	0.8357	0.8179	0.8464	0.8429
aw	0.9393*	0.8750	0.9357	0.9250	0.9214	0.9179	0.8750
ay	0.9357*	0.9107	0.9214	0.9143	0.9214	0.9143	0.9179
avg.	**0.9207***	**0.8943**	**0.9164**	**0.9136**	**0.9093**	**0.9164**	**0.9050**

(B) GIGADB

Subj.	Current model	Ablated models					
		Size of filters in Block 1		# of filters in Block 1		# of filters in Block 2	
	(64, 24, 2)	(**32**, 24, 2)	(**128**, 24, 2)	(64, **20**, 2)	(64, **28**, 2)	(64, 24, **1**)	(64, 24, **3**)
s01	0.7368*	0.7010	0.7005	0.7351	0.6557	0.7018	0.6760
s02	0.6061	0.5857	0.6113	0.6155	0.6160	0.6618*	0.6007
s03	0.9242	0.8881	0.9447	0.8828	0.9571*	0.8878	0.8789
s04	0.8032*	0.7781	0.7573	0.7526	0.7731	0.7931	0.7828
s05	0.9950*	0.9950*	0.9950*	0.9950*	0.9900	0.9950*	0.9900
avg.	**0.8131***	**0.7895**	**0.8017**	**0.7962**	**0.7983**	**0.8079**	**0.7346**

* The best hyperparameter set for each subject or average.



(a) Spatial importance (channel). Brighter color is more important.

(b) Temporal importance (0.5 sec. interval)

Fig. 4. Spatial and temporal importance by the permutation method.

TABLE IV
TOP-3 IMPORTANT CHANNELS AND TIME INTERVALS OBTAINED FROM THE PERMUTATION METHOD

Subj.	Spatial importance			Temporal importance		
	# 1	# 2	# 3	# 1	# 2	# 3
aa	CP3	C3	CPz	T3	T4	T2
al	C3	CP5	FC5	T3	T4	T5
av	CP3	CP4	C3	T2	T3	T4
aw	C4	FC3	CP2	T6	T5	T7
ay	C3	CP3	FC1	T3	T5	T4

C. Classification Performances of 3D-EEGNet

The classification accuracies are presented in Table II. The accuracy of each subject was the averaged value of the 10-fold cross validation results. The mean and standard deviation of the 10 test accuracies is represented. The proposed 3D-EEGNet showed better performances in the two datasets compared to the original EEGNet. In the BCIC, 3D-EEGNet showed 1.76% higher accuracy with lower variance in average on the all subjects. In the GigaDB, 3D-EEGNet improved the performance by 6.16% with lower variance.

The hyperparameters of the 3D-EEGNet model were verified by ablation studies, as shown in Table III. Three hyperparameters, the size and the number of temporal convolutional filters in Block 1, and the number of spatial filters in Block 2 were tested. The current 3D-EEGNet model showed the best performances on average for the accuracy of the all subjects in the two datasets (See Supplementary materials Table S.I for the ablation results of 2D-EEGNet).

D. Spatial and Temporal Explanation Using Permutation

In this section, the explanation results for the BCIC dataset are represented (see Supplementary materials Fig. S1 and Table S.II for the results of GigaDB dataset). Three types of permutation, zero, swap, and inverse, were applied to investigate the importance of the channels and time intervals for five subjects. The importances of channels and time intervals are shown in Figs. 4a and 4b, respectively. In Table IV, the top-3 important channels and time intervals are presented. For temporal explanation, EEG signals were divided into eight time-intervals with 0.5 second interval.

As summarized in Table IV, the important channels of five subjects were slightly different. For four subjects, aa, al, av, and ay, channels C3 and CP3, which are on the left area of the sensorimotor cortex, are shown as important channels. This means that the 3D-EEGNet model uses the EEG signals of the left area of the sensorimotor cortex, which is known to be activated when the right-hand MI is conducted.

TABLE V
NDCG SCORES OF SALIENCY-BASED METHODS

(A) CHANNEL IMPORTANCE

Subj.	IDCG	DCG			NDCG		
	Permutation	Saliency	DeepLIFT	DeepSHAP	Saliency	DeepLIFT	DeepSHAP
aa	3.0300	2.9715	2.9788	2.9975	0.9807	0.9831	**0.9893**
al	2.7304	2.6557	2.5784	2.3930	**0.9726**	0.9443	0.8764
av	4.0294	3.8183	3.9020	3.8772	0.9476	**0.9684**	0.9622
aw	3.9965	3.8922	3.9023	3.7746	0.9739	**0.9764**	0.9445
ay	2.7499	2.7303	2.7387	2.6649	0.9929	**0.9960**	0.9691
				avg.	0.9735	**0.9736**	0.9483

(B) TEMPORAL IMPORTANCE

Subj.	IDCG	DCG			NDCG		
	Permutation	Saliency	DeepLIFT	DeepSHAP	Saliency	DeepLIFT	DeepSHAP
aa	2.0361	2.0009	1.9951	1.9951	**0.9827**	0.9799	0.9799
al	1.6407	1.5089	1.6314	1.1583	0.9197	**0.9943**	0.7060
av	1.9490	1.8509	1.8509	1.8509	**0.9497**	**0.9497**	**0.9497**
aw	2.1325	2.1259	2.1259	1.5649	**0.9969**	**0.9969**	0.7339
ay	1.8309	1.7276	1.7386	1.8199	0.9436	0.9496	**0.9940**
				avg.	0.9585	**0.9741**	0.8727

In the temporal importance, time intervals T3 (0.5 – 1.0 sec.) and T4 (1.0 – 1.5 sec.) were found as the important ones for the four subjects. This means that the EEG signals for 1 second after 0.5 second from the visual cue were important for MI classification. This result corresponds to the prior knowledge that ERD is observed immediately after initiating motor imagery. Therefore, it is expected that the 3D-EEGNet successfully learned the related features for MI classification.

Meanwhile, the subject aw showed different results with the other subjects. In this subject, the channel C4, which is located on the right area of the sensorimotor cortex, was found to be the most important channel, and time intervals T6 (2.0 - 2.5 sec.) and T5 (1.5 - 2.0 sec.) to be the most important. The reason of this different result was investigated in the ERD/ERS map (see Appendix A). The ERD of aw cannot be found in the left-side channels differently from the other subjects. Thus, it can be inferred that the 3D-EEGNet learned different important features in the right-side channels (e.g., C4) in the late time intervals (see the results for subject aw in Fig. 4).

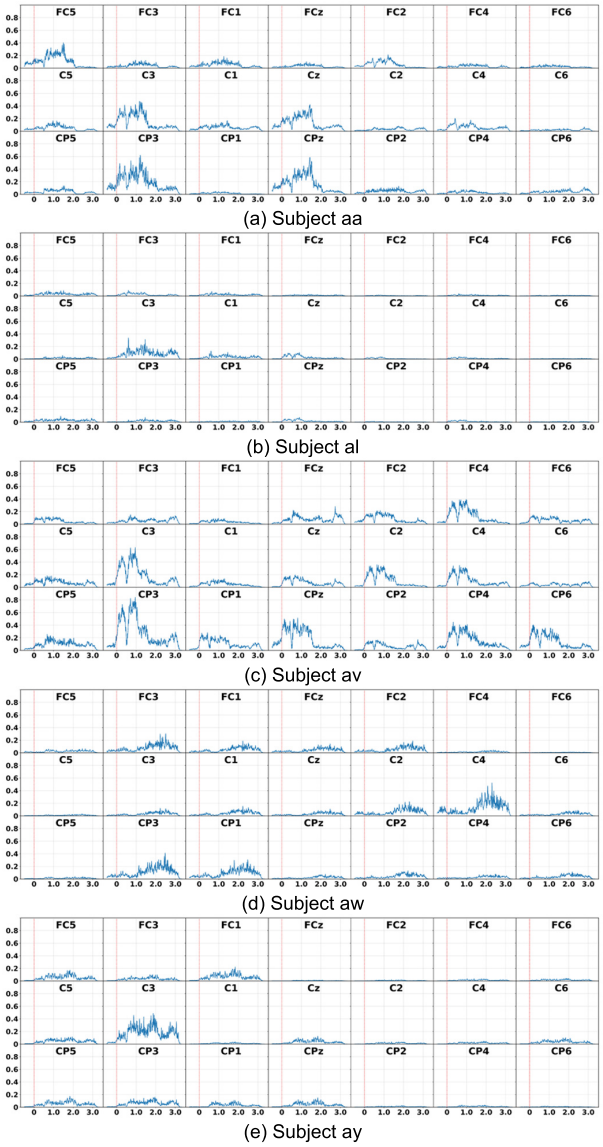


Fig. 5. Spatio-temporal importances using DeepLIFT; the vertical red dot line at time 0 represents when the visual cue was given.

E. The Best Saliency-Based Method Selection and Spatio-Temporal Explanation

To select the best saliency-based method for spatio-temporal explanation, the spatial and temporal importances of saliency map, DeepLIFT, and DeepSHAP were compared with those of the permutation method. Their obtained importances of channels and time intervals were transformed to ranks to compare with one another by the NDCG score.

The NDCG scores calculated for the BCIC dataset are summarized in Table V. (see Supplementary materials Table S.III for the NDCG scores for the GigaDB dataset). In the tables the NDCG scores represent how close the spatial and temporal importances of the saliency-based method yielded to those of the permutation method. Among the three saliency-based methods, DeepLIFT showed the highest average NDCG scores of 0.9736 in the spatial dimension and 0.9741 in the temporal dimension, respectively. This means that DeepLIFT provides the most similar ranking results compared with the permutation method, which is known to provide reliable results than saliency-based methods despite high computation cost. Therefore, DeepLIFT was selected as the most proper spatio-temporal explanation method for the trained 3D-EEGNet model.

As a result, using DeepLIFT, the 3D-EEGNet model is explained in the spatio-temporal dimension. In Fig. 5, the spatio-temporal importances obtained by DeepLIFT are depicted. The higher value represents the more important. The important channels and time intervals can be recognized simultaneously in 0.01-second time resolution.

In Fig. 5, for the subjects aa and av, the importances of channels CP3, CPz, and C3 (right area of the sensorimotor cortex) from 0 s to 1.5 s are found to be high. This result is consistent with the permutation results where the channels CP3, C3, and CPz and the time intervals T2, T3, and T4 (0.5-1.5 seconds) were important (see Table IV). Specifically, the importances of subject av are widespread over many left-side channels like CP4 and CP4, as well. From this, it can

be inferred that the subject av did not concentrate on the experiment and sometimes imagined left-hand movements.

For the subjects al and ay, the importances are smaller than those of other subjects, and the importance is concentrated on just one channel, C3. This means that the 3D-EEGNet model easily classified the MI by using the features of the only one channel C3. This is consistent with the result that the 3D-EEGNet of the subjects al and ay showed the highest accuracies (see Table III). In particular, the subject al, which had the best accuracy among five subjects, showed very low importance for all the other channels except C3.

The subject aw is the peculiar subject who gives different explanation. While the important channels of the other subjects are found in the left side of the sensorimotor cortex C3, the most important channel of the subject aw is the channel C4 which is on the right-side of the sensorimotor cortex. The reason was investigated in the ERD/ERS maps as shown in Appendix A. In the ERD/ERS maps of the other subjects, ERD was found just after the visual cue (represented red dot line) in left side of the sensorimotor cortex, C3. In contrast, for the subject aw, however, no ERD was not found after the visual cue in the left-side channel C3 (see Appendix A(d)). Instead, ERD of the subject aw appeared in the right-side channels like C4 after the cue. From that result, it can be inferred that the subject aw imagined the left-hand movement even though the visual cue was the right-hand movement.

VII. Conclusion

In this paper, we proposed a framework for developing a 3D shape of EEGNet model for MI EEG classification and explaining the models in spatial and temporal aspects. The 3D-EEGNet preserves spatial information of EEG to improve the MI classification accuracy significantly as well as effective explanation. The 3D-EEGNet exhibited better classification accuracies in the two MI EEG datasets.

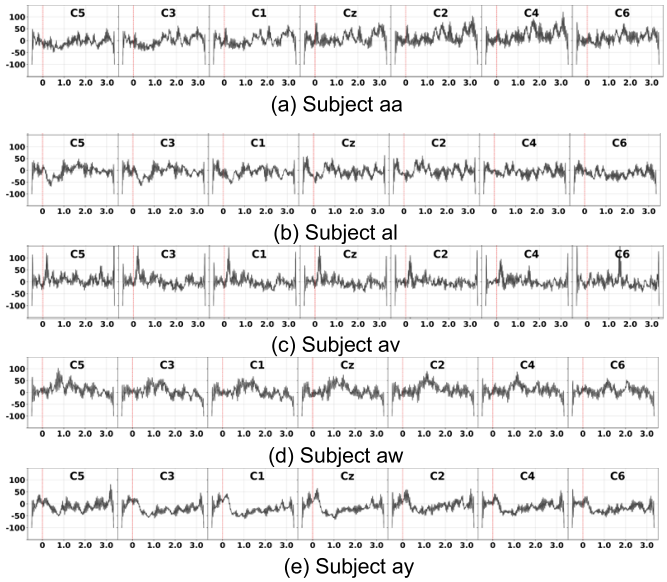
Using the permutation method, the trained 3D-EEGNet could also be explained in spatial and temporal dimensions to indicate which channels and time intervals are important in the classification. Furthermore, in the two experimental datasets, spatio-temporal explanations were provided using the DeepLIFT, which had been selected among the three saliency-based methods because of its highest NDCG score in comparison with the permutation method.

In the experimental results of the BCIC dataset, channels C3 and CP3 were important in the BCI classification and the time period for 1 second after 0.5 second from the visual cue were also important. In the case of the GigaDB dataset, channels C4, C2, and FC2 were important and the time period for 0.5 second after 1 second from the visual cue were important. The spatio-temporal explanations could indicate more detailed features that primarily affected the BCI classification.

Several studies showed their EEG models work well using XAI methods. To the best of our knowledge, however, comparison of the XAI methods has not been tried even though there is the specific best explanation method for their model. In this paper, we compared the XAI methods to find the best method for the proposed 3D-EEGNet using NDCG. We think that the proposed method helps engineers to experiment and choose the best fitted method to their neural networks model.

There are still a few limitations in this study. The ablation study on the elements of 3D-EEGNet such as the size of convolutional filters, the number of filters or types of activation functions is necessary to identify their effects to the model. Other saliency-based methods such as LRP or gradient $*$ input [50] could also be used as candidates for spatio-temporal explanation for 3D-EEGNet.

Appendix A
ERD/ERS Map of the Mu Band for Five Subjects of the BCIC Dataset



References

[1] M. Lotze and U. Halsband, "Motor imagery," *J. Physiol. Pairs*, vol. 99, pp. 386–395, Feb. 2006.

[2] V. K. Benzy, A. P. Vinod, R. Subasree, S. Alladi, and K. Raghavendra, "Motor imagery hand movement direction decoding using brain computer interface to aid stroke recovery and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 3051–3062, Dec. 2020.

[3] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, p. 1423, Mar. 2019.

[4] M. D. Nunez et al., "Electroencephalography (EEG): Neurophysics, experimental methods, and signal processing," in *Handbook of Neuroimaging Data Analysis*, vol. 1, 1st ed. Boca Raton, FL, USA: CRC Press, 2016, ch. 7, Sec. 7, pp. 175–197.

[5] M. Chavez, F. Grosselin, A. Bussalb, F. De Vico Fallani, and X. Navarro-Sune, "Surrogate-based artifact removal from single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 540–550, Mar. 2018.

[6] Y. Pei et al., "A tensor-based frequency features combination method for brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 465–475, 2022.

[7] D. Iacoviello, A. Petracca, M. Spezialetti, and G. Placidi, "A classification algorithm for electroencephalography signals by self-induced emotional stimuli," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3171–3180, Dec. 2016.

[8] X. Hou, Y. Liu, O. Sourina, Y. R. E. Tan, L. Wang, and W. Mueller-Wittig, "EEG based stress monitoring," in *Proc. IEEE Int. Conf. Syst. Man, Cybern.*, Hong Kong, Oct. 2015, pp. 3110–3115.

[9] M. Aljalal, R. Djemal, K. AlSharabi, and S. Ibrahim, "Feature extraction of EEG based motor imagery using CSP based on logarithmic band power, entropy and energy," in *Proc. 1st Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Riyadh, Saudi Arabia, Apr. 2018, pp. 1–6.

[10] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Comput. Biol. Med.*, vol. 100, pp. 270–278, Sep. 2018.

[11] J.-H. Cho, J.-H. Jeong, and S.-W. Lee, "NeuroGrasp: Real-time EEG classification of high-level motor imagery tasks using a dual-stage deep learning framework," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13279–13292, Dec. 2022.

[12] 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, Nov. 2016, Art. no. 016003.

[13] T. Uktveris and V. Jusas, "Application of convolutional neural networks to four-class motor imagery classification problem," *Inf. Technol. Control*, vol. 46, no. 2, pp. 260–273, Jun. 2017.

[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," *J. Neural Eng.*, vol. 15, no. 5, Jul. 2018, Art. no. 056013.

[15] M. Dai, D. Zheng, R. Na, S. Wang, and S. Zhang, "EEG classification of motor imagery using a novel deep learning framework," *Sensors*, vol. 19, no. 3, p. 551, Jan. 2019.

[16] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019.

[17] 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.

[18] M. Riyad, M. Khalil, and A. Adib, "A novel multi-scale convolutional neural network for motor imagery classification," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102747.

[19] P.-J. Kindermans et al., "The (Un)reliability of saliency methods," 2017, *arXiv:1711.00867*.

[20] Y. Wang, L. Wang, Y. Li, D. He, and T. Y. Liu, "A theoretical analysis of NDCG type ranking measures," in *Proc. 26th Annu. Conf. Learn. Theory*, Princeton, NJ, USA, 2013, pp. 25–54.

[21] K. Keng Ang, Z. Yang Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Hong Kong, Jun. 2008, pp. 2390–2397.

[22] J. Jiang, C. Wang, J. Wu, W. Qin, M. Xu, and E. Yin, "Temporal combination pattern optimization based on feature selection method for motor imagery BCIs," *Frontiers Hum. Neurosci.*, vol. 14, p. 231, Jun. 2020.

[23] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Aug. 2018.

[25] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, Aug. 2019, Art. no. 051001.

[26] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[27] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial EEG classification," *J. Neurosci. Methods*, vol. 274, pp. 141–145, Dec. 2016.

[28] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27074–27085, 2020.

[29] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, no. 5, pp. 1234–1241, Mar. 2020.

[30] X. Deng, B. Zhang, N. Yu, K. Liu, and K. Sun, "Advanced TSGL-EEGNet for motor imagery EEG-based brain-computer interfaces," *IEEE Access*, vol. 9, pp. 25118–25130, 2021.

[31] P. Greenside and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, Sydney, NSW, Australia, 2017, pp. 3145–3153.

[32] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.

[34] B. Tripathi and R. K. Sharma, "EEG-based emotion classification in financial trading using deep learning: Effects of risk control measures," *Sensors*, vol. 23, no. 7, p. 3474, Mar. 2023.

[35] Y. Li, H. Yang, J. Li, D. Chen, and M. Du, "EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by grad-CAM," *Neurocomputing*, vol. 415, pp. 225–233, Nov. 2020.

[36] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Process.*, vol. 93, no. 6, pp. 1401–1407, Jun. 2013.

[37] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[38] C. Stippich, H. Ochmann, and K. Sartor, "Somatotopic mapping of the human primary sensorimotor cortex during motor imagery and motor execution by functional magnetic resonance imaging," *Neurosci. Lett.*, vol. 331, no. 1, pp. 50–54, Oct. 2002.

[39] F. Mintzer and B. Liu, "Practical design rules for optimum FIR bandpass digital filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 204–206, Apr. 1979.

[40] 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 Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 3038–3049, Jul. 2022.

[41] G. Pfurtscheller, "Spatiotemporal ERD/ERS patterns during voluntary movement and motor imagery," *Suppl. Clin. Neurophysiol.*, vol. 53, pp. 196–198, May 2000.

[42] J. Schröder, F. Wenz, L. R. Schad, K. Baudendistel, and M. V. Knopp, "Sensorimotor cortex and supplementary motor area changes in schizophrenia: A study with functional magnetic resonance imaging," *Brit. J. Psychiatry*, vol. 167, no. 2, pp. 197–201, Aug. 1995.

[43] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of XAI methods on time series," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4197–4201.

[44] B. Blankertz et al., "The BCI competition III: Validating alternative approaches to actual BCI problems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, Jun. 2006.

[45] 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, Jul. 2017, Art. no. gix034.

[46] M. Vanputten, "Extended BSI for continuous EEG monitoring in carotid endarterectomy," *Clin. Neurophysiol.*, vol. 117, no. 12, pp. 2661–2666, Dec. 2006.

[47] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102172.

[48] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.