Journal of Neural Engineering

PAPER

Non–human primate epidural ECoG analysis using explainable deep learning technology

To cite this article: Hoseok Choi *et al* 2021 *J. Neural Eng.* **18** 066022

Journal of Neural Engineering

RECEIVED
29 January 2021

REVISED
4 October 2021

ACCEPTED FOR PUBLICATION
25 October 2021

PUBLISHED
25 November 2021

PAPER

Non-human primate epidural ECoG analysis using explainable deep learning technology

Hoseok Choi[1,2,5], Seokbeen Lim[2,5], Kyeongran Min[2,3], Kyoung-ha Ahn[4], Kyoung-Min Lee[4] and Dong Pyo Jang[2,*]

[1] Department of Neurology, University of California, San Francisco, CA, United States of America
[2] Department of Biomedical Engineering, Hanyang University, Seoul, Republic of Korea
[3] Samsung SDS Artificial Intelligence Research Center, Seoul, Republic of Korea
[4] Department of Neurology, Seoul National University Hospital, Seoul, Republic of Korea
[5] These authors contributed equally to this research.
[*] Author to whom any correspondence should be addressed.

E-mail: dongpjang@gmail.com

Keywords: brain-machine interface, epidural ECoG, deep learning, explainable artificial intelligence, bimanual

Supplementary material for this article is available online

Abstract

Objective. With the development in the field of neural networks, *explainable AI* (XAI), is being studied to ensure that artificial intelligence models can be explained. There are some attempts to apply neural networks to neuroscientific studies to explain neurophysiological information with high machine learning performances. However, most of those studies have simply visualized features extracted from XAI and seem to lack an active neuroscientific interpretation of those features. In this study, we have tried to actively explain the high-dimensional learning features contained in the neurophysiological information extracted from XAI, compared with the previously reported neuroscientific results. *Approach.* We designed a deep neural network classifier using 3D information (3D DNN) and a 3D class activation map (3D CAM) to visualize high-dimensional classification features. We used those tools to classify monkey electrocorticogram (ECoG) data obtained from the unimanual and bimanual movement experiment. *Main results.* The 3D DNN showed better classification accuracy than other machine learning techniques, such as 2D DNN. Unexpectedly, the activation weight in the 3D CAM analysis was high in the ipsilateral motor and somatosensory cortex regions, whereas the gamma-band power was activated in the contralateral areas during unimanual movement, which suggests that the brain signal acquired from the motor cortex contains information about both contralateral movement and ipsilateral movement. Moreover, the hand-movement classification system used critical temporal information at movement onset and offset when classifying bimanual movements. *Significance.* As far as we know, this is the first study to use high-dimensional neurophysiological information (spatial, spectral, and temporal) with the deep learning method, reconstruct those features, and explain how the neural network works. We expect that our methods can be widely applied and used in neuroscience and electrophysiology research from the point of view of the explainability of XAI as well as its performance.

1. Introduction

With the remarkable growth in the field of neural networks, deep neural network (DNN) research has produced an improved performance in fields such as image, speech, and text classification [1, 2]. As the performance of DNNs improves, understanding how they derive their high-performance results has become both interesting and important [3, 4]. Thus, a field called *explainable AI* (XAI) has recently been developed and studied to ensure that artificial intelligence models maintain a high level of accuracy and can be explained [3–5]. When the unknown black box in a DNN can be made transparent, users can trust the results of the model, and developers might find room to improve the model. XAI studies have shown the possibilities of understanding and interpreting artificial intelligence models by calculating features based on image classification, such as deconvolution and class activation map (CAM) [6–9]. For instance, a CAM can show the significant features of a DNN for classification or application in the form of heatmap images, making it intuitive and straightforward for users to understand [6].

Due to the high performance of DNNs, they have been applied to large-scale measured neural datasets to perform prediction and classification in the field of neuroscience [5]. For instance, some research groups have used DNNs to classify patterns in large-scale electroencephalography (EEG) data related to mental workload [10–12], motor imagery [13, 14], event-related potentials [15], emotion recognition [16], and seizure detection [17, 18]. In addition, due to the explainability of XAI, as well as its performance, some researchers have attempted to apply it to neuroscientific studies to explain neurophysiological information [19–21]. However, most of those studies have simply visualized the features extracted in the CAM without apparently trying to actively interpret the neuroscience involved in them. In addition, they manipulated only one- or two-dimensional data that are suitable for conventional speech- or image-based deep learning [10–15, 17, 18], even though EEG data contain high-dimensional (temporal, spectral, and spatial) information.

In this study, we designed a DNN classifier using 3D information from a neural signal and used it to classify monkey electrocorticogram (ECoG) data obtained during the unimanual and bimanual arm movement experiment.

We demonstrated that our 3D DNN model showed better arm movement classification performance than other methods, including the 2D DNN model. By using the 3D CAM, we figured out which features played a crucial role in the classification, and we interpreted the neuroscientific role of those spatial, temporal, and spectral features.

2. Methods

2.1. Subjects and surgical protocol

Two healthy, adult rhesus macaque monkeys (*Macaca mulatta*, denoted as M23 and M24) participated in the bimanual arm movement experiment, as described in detail in our previous study [22]. To implant the epidural ECoG electrode array (32-channel platinum array, Neuronexus, USA) in the brain, a licensed veterinarian performed the whole surgical process with appropriate sterility, anesthesia, monitoring, and recording for all measured physiological variables. Each animal was injected intramuscularly (I.M.) with atropine sulfate (0.08 mg kg^{-1}) 1 h before surgery to inhibit excessive salivation. After 30 min, the animal was sedated with tiletamine-zolazepam (Zoletil®, 10 mg kg^{-1}, I.M.), intubated, and placed under isoflurane anesthesia. Heart rate, blood pressure, body temperature, oxygen saturation, and respiratory rate were continuously monitored during surgery. Each monkey was fixed in a stereotaxic frame, and the craniotomy was conducted with a 2.5 cm radius on both hemispheres. The dura mater was not injured.

In each hemisphere of each monkey, the ECoG electrodes were placed at the dorsal premotor cortex (PMd), supplementary motor area (SMA), primary motor cortex (M1), primary somatosensory cortex (S1), and posterior parietal cortex (PPC), as shown in figure 1(A). The ground electrode was placed at the midline. After surgery, the animals were given adequate time to recover in their cages, consistent with the Guide for the Care the Use of Laboratory Animals. The temperature of the cages was preserved at 24 ± 4 °C, and the humidity was maintained at 50 ± 10%. The light was controlled in a 12 h light/dark cycle. The Seoul National University Hospital Animal Care and Use Committee (IACUC No. 13-0314) approved all experimental procedures.

2.2. Behavior task

Before each behavior task, the animal sat on a chair, and their head was fixed in place with the head holder. Then the cue buttons were set within the monkey's line of sight, and the ready buttons were placed in a natural and comfortable arm position, as shown in figure 1(B). The distance from the ready button to the cue button was approximately 30 cm, which is almost the maximum distance that the monkeys could reach out their arms. The task started when the monkey pushed the ready button which lit a green light. The time interval from the trial start to the cue onset was randomly set between 3 s and 7 s. Using the colored lights on the cue buttons, each trial was distinguished by three target movement types: left-arm movement, right-arm movement, and both-arms movement. The blue light indicated that the monkey moved their left arm, and the red light meant that the subject moved their right arm. If the blue and red lights were presented simultaneously, the animal moved both arms.

One of the three types appeared randomly on the cue buttons. For trials of left-arm and right-arm movement, the monkey was to move their target arm to push the cue button while pushing the ready button with their non-target arm (figure 1(B)). For the trials of both-arm movement, the monkey was to move both arms and push both cue buttons simultaneously (figure 1(B)). If the animal pushed and held the target cue buttons correctly for 1 s, the trial was finished, and the monkey received some water as a reward. The button and reward systems were operated by MATLAB (Mathworks, USA) and NI-PCI-6221 (National Instruments, USA). We recorded each monkey's epidural ECoG data and motion tracking

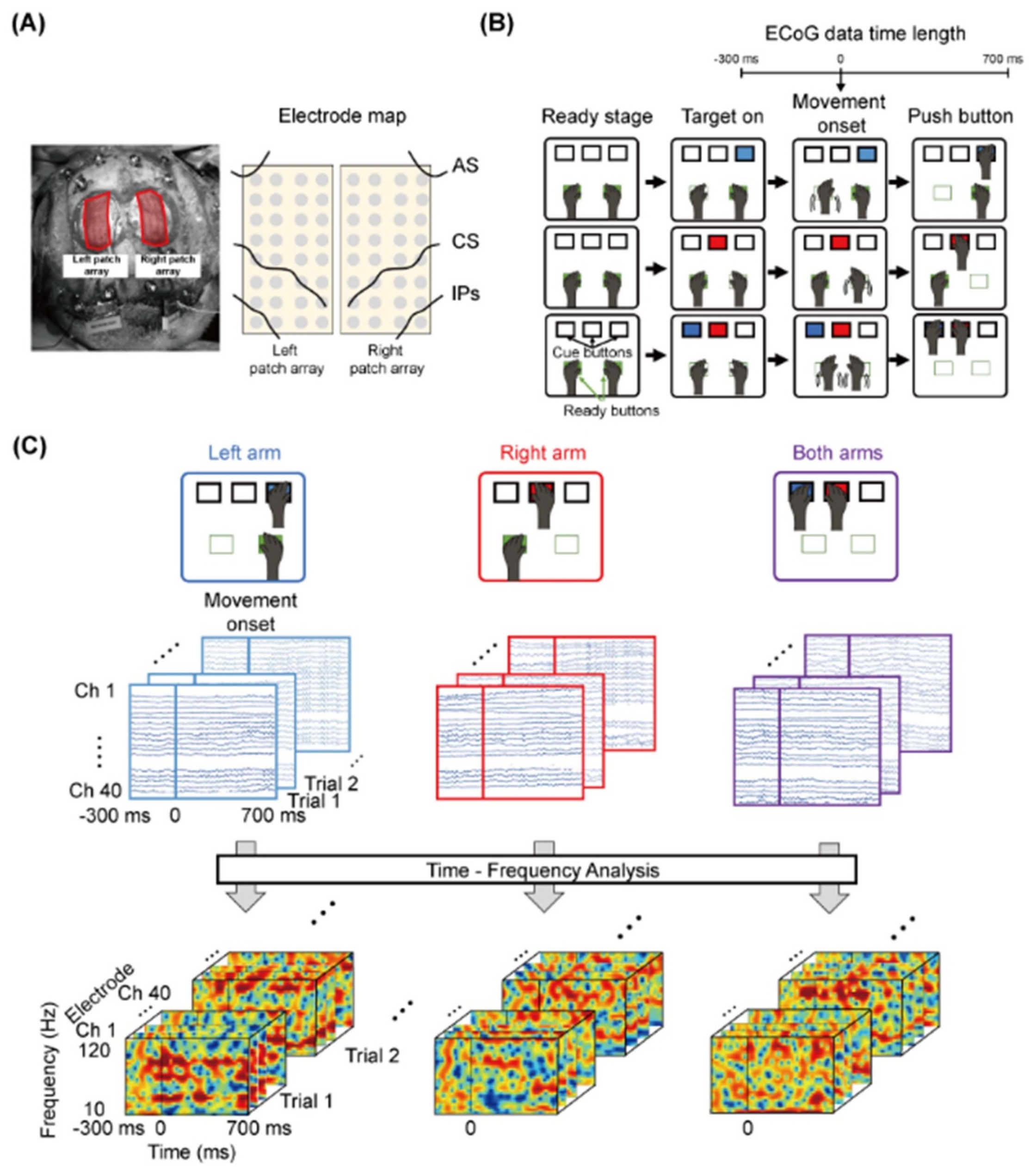


Figure 1. Positions of the ECoG electrode arrays, behavior protocol, and data processing. (A) In both hemispheres, the ECoG electrodes were inserted on the gyrus from the edge of the arcuate sulcus (AS) to the intraparietal sulcus (IP), passing the central sulcus (CS). The electrodes were placed on the primary motor cortex (M1), supplementary motor area (SMA), dorsal premotor cortex (PMd), primary somatosensory cortex (S1), and posterior parietal cortex (PPC). (B) Experimental protocol with three arm-movement tasks. The monkey conducted three types of arm-movement tasks, left arm, right arm, and both arms, while the ECoG signals were recorded. To start the behavior tasks, the monkey pressed and held the ready button, which turned on a green light. For each task, the colored cue buttons (left arm-blue, right arm-red) lit up randomly. Each session contained 200 trials, and each monkey completed one to four sessions per day for four months. (C) ECoG data preprocessing. For the data analysis, we used ECoG data from −300 ms to 700 ms for each movement onset. Using the spectrotemporal analysis method for each spatial information, we stacked the high-dimension feature data (spatial-spectral-temporal information) for each arm movement.

data for four months, one to four sessions per day, with each session containing 200 trials.

2.3. Data acquisition and preprocessing

Three weeks after the surgery, we started to record epidural ECoG signals while the monkeys conducted the behavior tasks. The ECoG signals were recorded using an EEG 1200 (Nihon Kohden, Japan) at a sampling frequency of 1 kHz per channel. For data preprocessing, we used the MATLAB and EEGLAB toolbox. The raw ECoG data were bandpass filtered in the range of 0.3–200 Hz and then notch filtered at 60, 120, and 180 Hz to remove the power noise.

As input data, we made spectrotemporal representation data, called *event-related spectral perturbation* (ERSP) data, using the EEGLAB ‘newtimef’ function wavelet transform, as shown in figure 1(C), from every electrode. This high-dimensional feature describes the spatial-spectral-temporal information of the signals from −300 ms to

Table 1. DenseNet structure. Modified structures for training the 2D ECoG feature map (DenseNet-2D) and the 3D ECoG feature map (DenseNet-3D).

Layers	DenseNet-161 ($k = 48$)	DenseNet-2D ($k = 48$)	DenseNet-3D ($k = 48$)
Convolution	7 × 7 conv, stride (2, 2)	7 × 7 conv, stride (2, 1)	7 × 7 × 7 conv, stride (2, 2, 1)
Pooling	3 × 3 max pool, stride (2, 2)	3 × 3 max pool, stride (2, 1)	3 × 3 max pool, stride (2, 1, 1)
Dense block (1)	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times 1\times 1\,\text{conv} \\ 3\times 3\times 3\,\text{conv} \end{bmatrix} \times 3$
Transition layer (1)	1 × 1 conv 2 × 2 average pool, stride (2, 2)	1 × 1 conv 1 × 1 average pool, stride (1, 1)	1 × 1 × 1 conv 1 × 1 × 1 average pool, stride (1, 1, 1)
Dense block (2)	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times 1\times 1\,\text{conv} \\ 3\times 3\times 3\,\text{conv} \end{bmatrix} \times 6$
Transition layer (2)	1 × 1 conv 2 × 2 average pool, stride (2, 2)	1 × 1 conv 1 × 1 average pool, stride (1, 1)	1 × 1 × 1 conv 1 × 1 × 1 average pool, stride (1, 1, 1)
Dense block (3)	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 36$	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times 1\times 1\,\text{conv} \\ 3\times 3\times 3\,\text{conv} \end{bmatrix} \times 6$
Transition layer (3)	1 × 1 conv 2 × 2 average pool, stride (2, 2)	1 × 1 conv 2 × 2 average pool, stride (2, 2)	1 × 1 × 1 conv 2 × 2 × 2 average pool, stride (2, 2, 2)
Dense block (4)	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1\times 1\,\text{conv} \\ 3\times 3\,\text{conv} \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times 1\times 1\,\text{conv} \\ 3\times 3\times 3\,\text{conv} \end{bmatrix} \times 4$
Classification layer	7 × 7 global average pooling	3 × 3 CAM_conv, stride (1, 1) 6 × 7 global average pooling	3 × 3 × 3 CAM_conv, stride (1, 1, 1) 6 × 7 × 10 global average pooling
	3 fully connected, softmax	3 fully connected, softmax	3 fully connected, softmax

700 ms of the arm movement time point with 100 bins. The spectral information was included from 10 to 120 Hz divided into 28 bins. To select the spatial information, the bad electrode was removed, interpolated, and rescaled based on the brain region where the electrode was located. In that way, we selected 40 spatial information from 64 electrode arrays. Thus, the input data dimension was (100 × 28 × 40) (100 temporal bins × 28 spectral bins × 40 spatial bins). We stacked the preprocessed data for every trial for four months (M23: 4426 trials, M24: 5029 trials) and randomly mixed the data to train the machine learning models (Total data size: 9455 × 100 × 28 × 40).

2.4. Deep learning analysis

The deep learning structure was modified from DenseNet [23], a convolutional neural network. The DenseNet structure includes a dense connectivity pattern that directly connects all layers to maximize the flow of information between layers in the model. This pattern, called *Dense block*, is characterized such that data delivered to the current layer concatenates all the data feature information from the previous layers. This pattern has several advantages. First, if the transmitted information is concatenated in the current layer, features of previous layers are used as is, so the information can flow to the next layer without being disturbed. The features of the previous layers can thus be reused. Second, because the layers are directly connected, the vanishing gradient problem can be improved, and it can minimize overfitting. Also, the pattern offers easy training with a few parameters [23].

To design a 3D DNN model to classify 3D ECoG data related to arm movement and use all the feature information, we added one dimension to the deep learning channel and resolved all temporal, spectral, and spatial information domains. Thus, the input data dimension for the 3D DNN was (9455 × 100 × 28 × 40 × 1). All 2D layers in the DenseNet-161 structure were changed into 3D layers (table 1). The parameters (such as stride) of some specific layers were adjusted according to the input data size. In addition, a convolution layer was created before the global average pooling layer for the 3D CAM analysis. For comparison, a 2D DNN model was also built using 2D layers instead of 3D layers. To verify and emphasize the performance, we compared our 3D DNN with other machine learning methods: a linear discriminant analysis (LDA), support vector machine (SVM), and 2D DNN. The classifier was trained by five-fold cross-validation. For five-fold cross-validation, the proportion of test data set was 20% of total data set so as not to overlap in each fold. Then, the remaining data were divided into the train data set and validation data by 8:2. The model training and analysis were conducted using the TensorFlow and Keras libraries in a Python environment. For training, an Nvidia Tesla P100 GPU 16 GB hardware device was used.

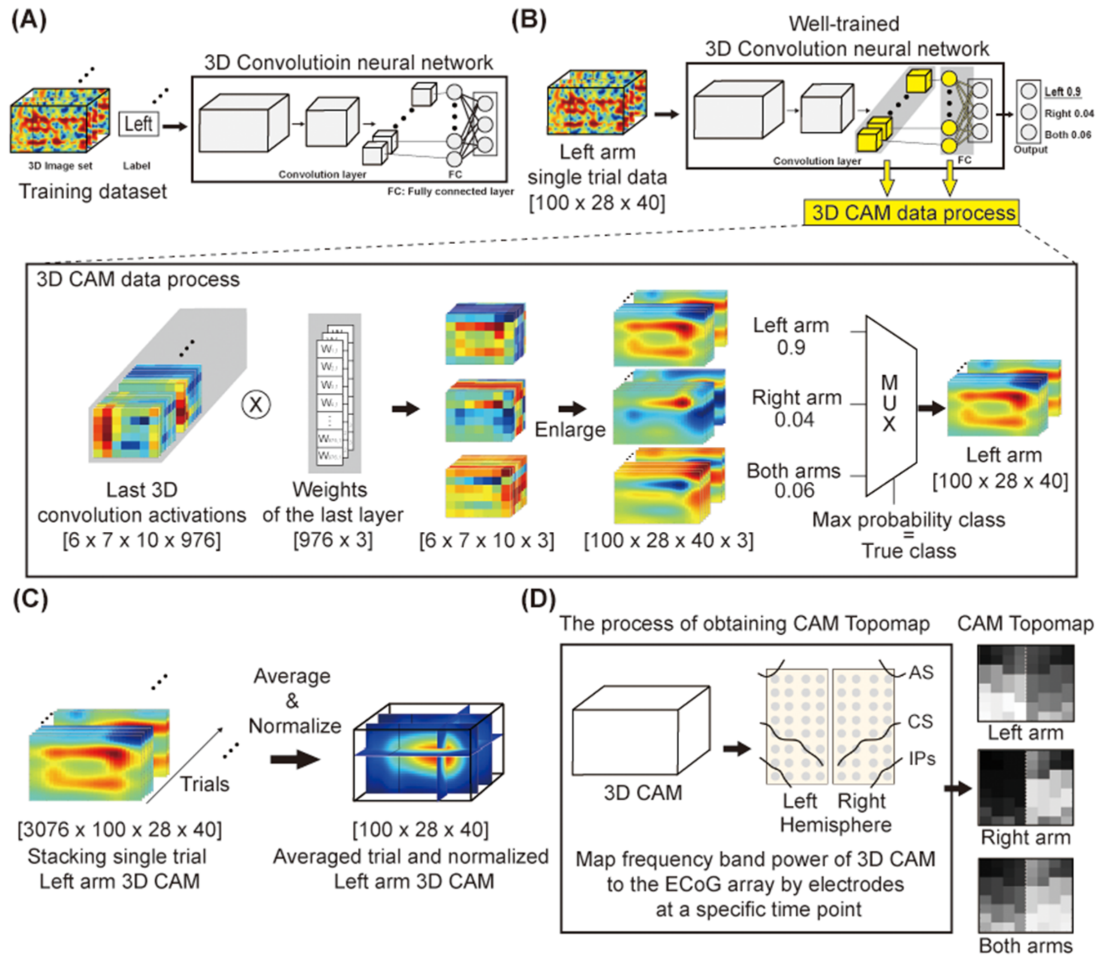


Figure 2. Schematic representation of the 3D class activation map using a 3D deep learning model. (A) Process of building the 3D CAM using a 3D deep learning model. We first trained a 3D convolutional neural network (using the DenseNet structure [23]) until we had excellent performance using preprocessed 3D ECoG data. (B) Using the trained model, which described in (A), we classified each arm movement from ECoG single-trial data. (Here shows the left arm trial example). After this arm movement type prediction, we extracted the last 3D convolution layer's activations and the last layer's weights from the model, then multiplied both to obtain the 3D CAM. The box content is the detailed process completed to get the 3D CAM. We multiplied the last 3D convolution activation map and the weights of the last layer when predicting input data. Then, it enlarged the volume (6 × 7 × 10) to the input image size (100 × 28 × 40). We obtained each 3D CAM corresponding to the max predictive class, which was equal to the true class. (C) To analyze highly effective feature maps when the model classified left arm movement, we calculated and stacked 3D CAMs of ECoG data for all trials. Then, the 3D CAM data were averaged by the trials and normalized. In the case of right arm movement and both-arm movement, the averaged trial and normalized 3D CAMs were obtained in the same way (figure S1 (available online at stacks.iop.org/JNE/18/066022/mmedia)). (D) To intuitively visualize 3D information, we converted the 3D CAM into a topological projected figure called CAM Topomap. The CAM Topomap was obtained by mapping the spectral information corresponding to specific temporal information in the electrode arrays.

2.5. Three-dimensional class activation map

A CAM is a heat map image that shows the most effective features when the designed DNN model classifies new input image data into a specific class [6]. A CAM can be obtained as the product of both the last convolutional layer before the global average pooling layer and the weight of the last layer among the components constituting the deep learning model [6], as shown in figures 2(B) and (C). We applied 3D CAM to view the 3D class activation feature for a 3D input data set. The process for obtaining the 3D CAM is the same as that for the 2D conventional CAM, but the calculation is performed by adding one more dimension. Figure 2(B) shows the process for obtaining the 3D CAM. If new input data were classified through the well-trained model, this model calculated the probability that those data would be sorted into each class. Then, the last 3D convolution layer (6 × 7 × 10 × 976) and the weight of the last layer (976 × 3) were extracted from the model and multiplied by two values to obtain 3D CAM. The 3D input image's characteristics abstracted by the convolution layers were stacked in the last 3D convolution layer. The weight of the last layer was the trained weight of the fully connected layer to classify the data in the last 3D convolution layer. The multiplication of those layers used matrix multiplication. The shape corresponding to the last 3D convolution layer's spatial-spectral-temporal information was changed to enable matrix multiplication (420 × 976). After performing the matrix multiplication (420 × 3), the result was transformed back into the same shape as the last 3D convolution layer (6 × 7 × 10 × 3). This 3D image was then resized to fit the input image size (100 × 28 × 40 × 3). The resized image showed the location of the features that highly influenced the classification of the input image in the form of a 3D heat map. To validate 3D CAM performance and to check these visualized features are actually matching the main feature of the input dataset, we conducted the performance test using the 2D image of the CIFAR-10 dataset [24], as shown in figure S4. Through this evaluation, we can generalize that 3D CAM can explain the input data's key features and it helps how humans can understand the machine works, which is related to the difference of the classes.

In this study, we obtained a 3D CAM for all trials in the left, right, and both-arm movement classes. Then, we stacked the 3D CAMs for all trials by each arm movement. After that, the 3D CAM data for each class were normalized and visualized. In that way, we tried to figure out which features were the most important in classification. We converted the 3D CAM into a class activation topomap for intuitive visualization, as shown in figure 2(E). The CAM Topomap was obtained by mapping the spectral information corresponding to specific temporal information in the 3D CAM to the 8 × 4 electrode array of each hemisphere, making it easy to understand how the spectral information from each region changed over time.

Visualization was performed using the graphical user interface design environment in MATLAB (Mathworks, USA). To confirm the 3D CAM data, the images were visualized in the spectrotemporal, spatiotemporal, and spatiospectral dimensions. Visualization, according to the spectral range, was performed to check the CAM Topomap for each arm movement.

3. Results

3.1. Decoding the accuracy of the 3D DNN

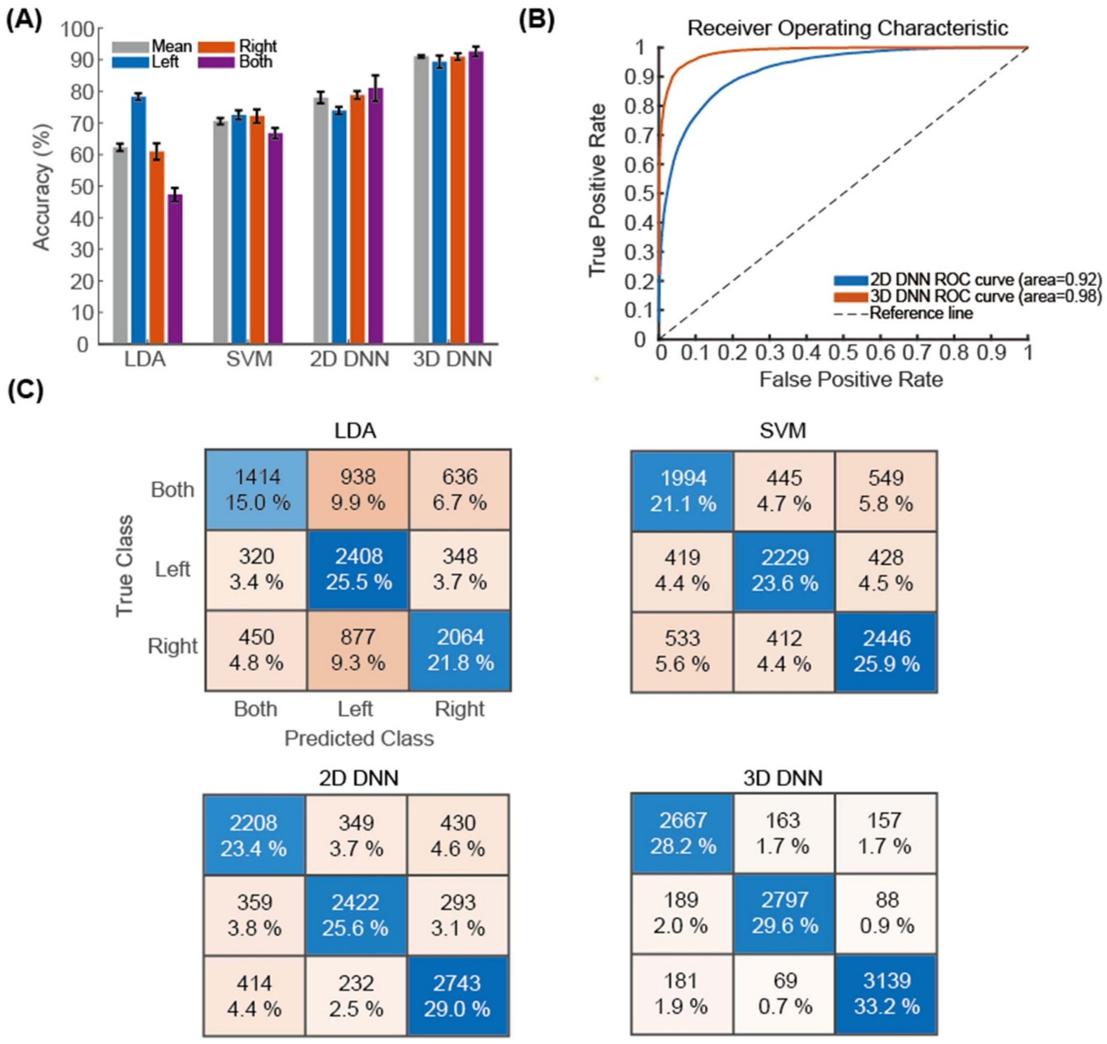
To evaluate the 3D convolutional neural network's performance, we trained a variety of machine learning models: LDA, SVM, DenseNet-2D, and DenseNet-3D. Table 1 shows the DenseNet structures applied in this paper (DenseNet-2D and DenseNet-3D). Those structures were modified to train the 2D and 3D ECoG feature maps based on DenseNet-161 [23]. We added one dimension to DenseNet-3D in the convolutional filter layer and the pooling layer. Furthermore, we conducted hyper-parameter tuning, such as strides. As shown in figure 3(A), the average classification accuracy of the DenseNet-3D was almost 90%, higher than that of the other models. In addition, the accuracy of individual classifications as unimanual or bimanual movements was higher than that of the other models. The results of the confusion matrix could be interpreted to indicate that the 3D DNN model has a maximum arm movement classification error of 2% or less, compared with the other models (figure 3(C)). The confusion matrix result for the 2D DNN model indicates that the arm movement classification error was between 2.5% and 4.6%, but its performance was worse than that of the 3D DNN model.

3.2. Analysis of the most effective classification feature in the 3D decoder

First of all, to evaluate the 3D CAM performance, we performed the simulation using the Densenet-3D structure based on the CIFAR-10 dataset [24]. CAM performance quantification was performed using the intersection over minimum (IoM) index among intersection over union [6] and IoM [25], which are commonly used in the field of object detection. As a result, the average accuracy was 86% (five-fold cross validation) was obtained in image classification. In addition, the average IoM in figure S6 was 0.8933. Through these simulation processes, we found that CAM can explain the input data's difference by class (Please refer to figure S4 for details).

We used the 3D CAM to analyze the major classification features in the 3D DNN when the well-trained model classified each arm movement. The 3D CAM configured the three dimensions (spatial-spectral-temporal information) such as 3D ECoG feature maps. To compare the 3D CAM pattern with that of frequency power activation, we conducted a spectrum power analysis using the ESRP and topomap reconstruction according to the arm movement task, as depicted in figures 4(A) and (B) for trial averaged results and figures S2 and S3 for example single trial results. All the ERSPs from electrodes in specific brain regions (M1, S1, PPC, PMd, and SMA) were averaged (figure 4(A)). For the spatial distribution analysis in the topomap, the temporal information was categorized into 'before movement' (−300 to 0 ms), 'just after movement onset' (0–100 ms), and 'during movement' (100–700 ms). In the before-movement state, the monkey was pressing and holding the green glowing ready button with both arms while waiting for the cue signal. In the movement onset state, the monkey started their arm movement to press the target button that was lit red or blue.

Figure 4(A) depicts the spectrotemporal ERSP for a representative electrode in five brain regions of the left and right hemispheres. In the left arm movement task, gamma band activation was noticed in the contralateral regions (right hemisphere) and vice versa during the right arm movement task, as we expected. In addition, the gamma band power increased in both hemispheres during the bimanual task. In the ERSP topomap, the mid gamma band was visualized as the frequency band for each time range (figure 4(B)) because the main activation features for



LDA

True Class	Both	Left	Right
Both	1414 15.0 %	938 9.9 %	636 6.7 %
Left	320 3.4 %	2408 25.5 %	348 3.7 %
Right	450 4.8 %	877 9.3 %	2064 21.8 %

SVM

1994 21.1 %	445 4.7 %	549 5.8 %
419 4.4 %	2229 23.6 %	428 4.5 %
533 5.6 %	412 4.4 %	2446 25.9 %

2D DNN

2208 23.4 %	349 3.7 %	430 4.6 %
359 3.8 %	2422 25.6 %	293 3.1 %
414 4.4 %	232 2.5 %	2743 29.0 %

3D DNN

2667 28.2 %	163 1.7 %	157 1.7 %
189 2.0 %	2797 29.6 %	88 0.9 %
181 1.9 %	69 0.7 %	3139 33.2 %

Figure 3. Performance comparison of the machine learning models. (A) Performance graph showing the classification accuracy of the 4 classifiers. The gray bar indicates the mean accuracy for each arm movement classification. (B) Receiver operating characteristic curve comparing the 2D DNN and 3D DNN. (C) Confusion matrixes showing the classification accuracy of the 4 classifiers.

arm movement classification as shown in figures 4(A) and (C) were distributed in the mid gamma band range (50–90 Hz). As shown in figure 4(B), the ERSP Topomap confirmed, in the case of one-hand movement, that the gamma band power in the contralateral brain hemisphere region increased in the 'just after movement onset' (0–100 ms) and 'during movement' times. Both hemispheres were involved in bimanual movement.

In contrast to the frequency power activation, the CAM analysis indicates that the features used for classification showed different patterns according to the movement tasks, as shown in figure 4(C). In the case of left arm movement, the CAM model highly weighted the gamma band of the left M1, S1, and PPC for classification. For the right arm movement class, the most weighted area was right M1, S1, and PPC, which are also ipsilateral areas, as same as described in the left-arm movement class weight result. In the both-arm movement task, the model used a broad band, including the gamma and beta bands of the right S1 and PPC at the specific temporal points related to movement onset and offset.

4. Discussion

In this study, we designed a DNN classifier using 3D information from a neural signal, figured out which information was used for the classification, and used biological knowledge to explain why those features were chosen. The 3D CAM allowed us to confirm which information—electrodes, frequency bands, and time points—made the most effective classification features for ECoG data collected during the arm movement tasks. As far as we know, we are the first case to extract features through a 3D CAM using spatial-spectral-temporal information from a neural signal rather than using its own signal or just an image. In terms of XAI, our work is a good precedent because we have shown that neurophysiology data can be interpreted by comparing it with the biological background.

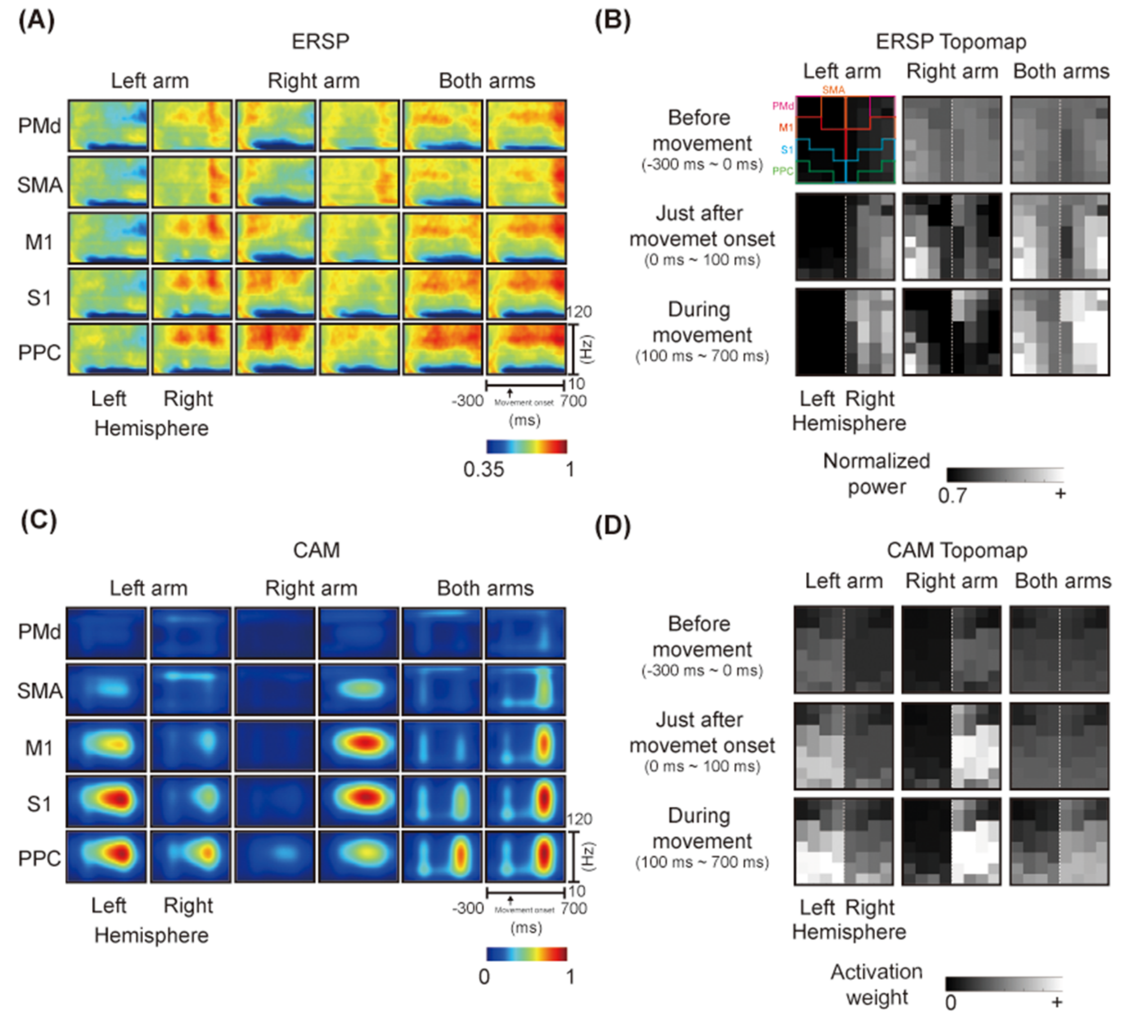


Figure 4. 3D CAM and CAM Topomap. (A) Spectrotemporal ERSP from the 3D ERSP. (B) ERSP Topomap. These maps were averaged by the middle gamma frequency range from 50 Hz to 90 Hz for three temporal states (before movement state, just after movement onset state, and during movement state). (C) CAM for each electrode from the 3D CAM. (D) CAM Topomap.

4.1. The advantages of 3D CAM

As depicted in figure 3, our 3D DNN model showed the best arm movement classification performance of the tested methods, including a 2D DNN model. In general, as more information is used during the classification, the results become more accurate. To use all information, we converted arm-movement ECoG data into spatial-spectral-temporal 3D data and we applied it to a 3D DNN model rather than a typical 2D DNN model. That allowed us to train our model using all data features with minimal information loss, and thus our model showed better performance than the 2D DNN model trained with the spectrotemporal data typically used in the neuroscience field.

We used the 3D CAM to identify which spatial-spectral-temporal features were highly weighted in the classification of a well-trained 3D DNN model. We found that the model we trained classified arm movements by using M1 and PPC for spatial features, the gamma band (mainly mid-gamma-band) for spectral information, and the movement onset and during motion windows for temporal information showing the consistence with previous study results [22, 26–43].

4.2. The importance of 3D CAM ipsilateral information in classifying arm movements

One interesting thing in our results is the strong weight given to the ipsilateral region when classifying a hand movement. Figure 5 shows that the ECoG signal's gamma-band power was activated in the contralateral motor cortex region during unimanual movement. However, in the 3D CAM weight, as confirmed through the CAM Topomap, the activation weight increased in the ipsilateral motor and somatosensory cortex regions. A left arm movement produced large weight distributions in the left hemisphere S1 and PPC, and a right arm movement produced large weight distributions in the right hemisphere M1 and S1. Thus, we interpret the 3D DNN model deemed ipsilateral information to be important for arm movement classification when classifying one-hand movement.

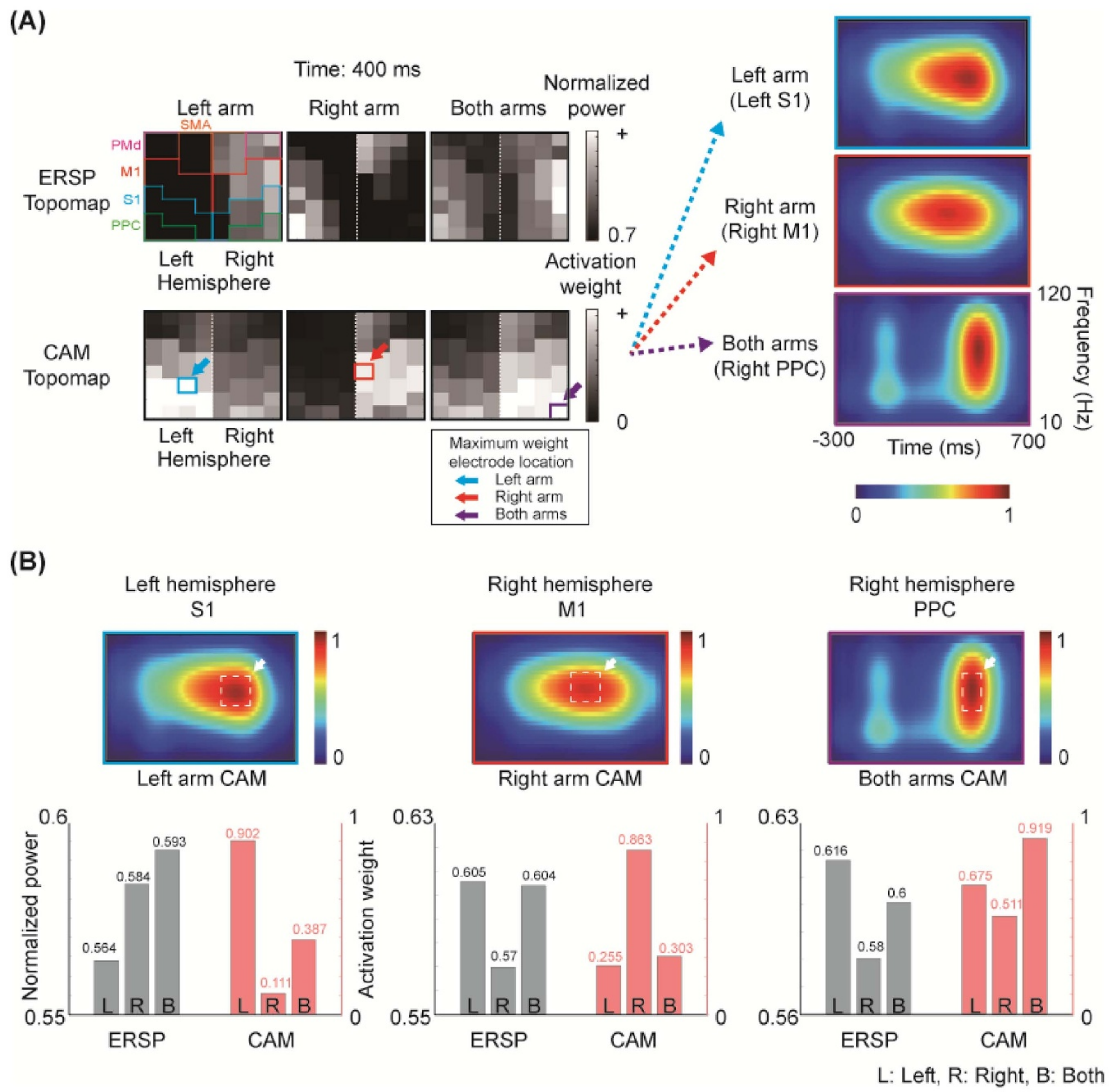


Figure 5. Comparison of ERSP and CAM in an example electrode. (A) ECoG spectral power topomap and CAM Topomap for each arm movement 400 ms after movement onset. At this temporal point, the 3D CAM weight was maximum. The ECoG spectral power and 3D CAM weight were averaged in the middle gamma range: 50–90 Hz. In the case of a unimanual movement, the ECoG spectral power represented a contralateral distribution. However, the CAM weight emphasized the ipsilateral distribution (left arm movement: left S1; right arm movement: right M1). In the case of bimanual movement, the ECoG spectral power represented a bilateral distribution, but the CAM weight emphasized the right hemisphere, especially the right S1 and PPC. The red arrow indicates the electrode area with the maximum CAM weight for each arm movement. The spectrotemporal CAM corresponds to the arrow color (blue: left arm movement, red: right arm movement; purple: both-arm movement) on the CAM Topomap for each arm movement. About 0 ms is movement onset. These maps are jet scale. When the 3D deep learning machine classified a one-arm movement, the machine used the gamma frequency range during movement (−100 ms to 500 ms). But for two-arm movements, the model used a broad frequency range at two dominant time points (−100 ms and 400 ms). (B) Results in the CAM Topomap and CAM around the maximum weight. There are bar graphs of the relationship between the 3D CAM weight and ERSP power below the CAM Topo map. The gray bar represents the ERSP power, and the pink bar represents the CAM weight, which is the average weight of the area corresponding to 90% of the maximum value in the spectrotemporal CAM in figure (A). The white arrows and dashed boxes indicated the averaged area. The ERSP power was calculated by taking the average value of the same area as the 3D CAM weight.

The general understanding of the central nervous system for motor control is that most neurons that originate in the motor cortex cross to the other side during their descent in the brainstem. Therefore, intersected neural activity controls the contralateral region. According to the principle of the cortical control of movement, previous researchers mainly interpreted brain signals from the contralateral hemisphere when subjects performed an arm-movement task [26–29]. However, non-crossed neurons still remain; the motor tract is not entirely contralateral, so both ipsilateral and contralateral neurons should be considered. Previously, several studies have demonstrated the involvement of both ipsilateral and contralateral cortex in the movement [30–32, 41–43]. According to a couple of previous studies, some M1 neurons show ipsilateral and bimanual-related activity [33], and some parietal neurons represent ipsilateral arm movements [44]. In the neuroimaging study, when the subjects moved the ipsilateral finger, the ipsilateral sensory and motor cortex of the finger was modulated which significantly related to the pattern induced by the contralateral finger [41]. Recently, two studies showed why ipsilaterally-related activity does not cause contralateral motor output, while M1 activity has been correlated with both contralateral and ipsilateral limb movements [42, 43]. They explained that activity related to each arm occupies a distinct subspace, enabling muscle-activity decoders to naturally ignore signals related to the other arm [42, 43]. From a machine learning point of view, Bundy found that ipsilateral arm movement kinematics could be decoded by ipsilateral signals [32], and Mooshagian found that a linear SVM classifier trained on the neural activity of parietal neurons could decode ipsilateral arm movements [45]. These results suggest that brain signals acquired in the motor cortex contain information about both contralateral and ipsilateral movement.

Unlike the one-handed exercise, the two-handed movement allows interacting both hemispheres of the brain to move the two hands together concurrently, so the brain regions involved are different from those used in one-handed motions [22]. Many studies have reported that the posterior parietal cortex plays a role in regulating the movement of both hands [45–49]. Some representative studies of bimanual coordination in primates analyzed the electrical activity of PPC nerves during bimanual movement. Kermadi *et al* (2000) electrically measured the neural activity of individual neurons in the PPC and showed that more neurons were activated when performing two-hand movements [46]. Mooshagian *et al* (2018) measured neural signals in the parietal reach region (PRR) of the PPC during ten different movement patterns using one and both hands. They suggested that the neural activity occurring in the PRR during bimanual movements was not a linear sum of that from one-handed movements but a pattern specific to two-handed movements that played a role in bimanual coordination [45]. Some other clinical research showed that subjects who were damaged in that area were limited in moving both hands during the bimanual movement task. Those results suggested that the PPC makes an important contribution to two-handed adjustments. In figure 5, the CAM Topomap for two-handed movement, unlike the one-handed CAM Topomap, shows the maximum weight distribution in the right hemisphere S1 and PPC, which contain the PRR. In other words, the ECoG power of the PRR was a significant feature in the two-hand motion classification, which is consistent with what E. Mooshagian described. We also verified that bimanual coordination could be confirmed in the PPC using ECoG, which has characteristics different from those in the neural spike signal. Therefore, our 3D DNN model's features are consistent with previously published neuroscience knowledge.

4.3. The most important timing and frequency bands when classifying arm movements

As we would normally expect from a physiological analysis, the most important indicator in the time domain was during the task period. These results can be explained by Pfurtscheller and da Silva 1999 and Neuper and Pfurtscheller 2010 [34, 50]. They showed that event-related power during the start, middle, and end times of the exercise is a key function of brain-machine interface. In our research, we found that the 3D CAM weight was maximized during movement, especially 400 ms after the movement onset, and we also confirmed the weight of unimanual movement expressed in the ipsilateral distribution during movement (time range: -100 ms to 500 ms, figure 4). As shown in figure 5, 3D CAM weights for the left and right arm movements, which were averaged in the frequency range of 50 Hz–90 Hz at 400 ms, showed a mainly ipsilateral distribution. Thus, temporal information provides important insights for classifying hand movements during movement execution. However, for both-arm movements, the weights on movement onset and offset were higher, which we assume is related to somatosensory feedback.

In the spectrotemporal CAM of one-hand movement, the well-trained 3D decoder used the gamma frequency range during movement. On the other hand, the decoder used a broad frequency range with a focus on two main points, before ($\fallingdotseq -100$ ms) and after movement ($\fallingdotseq 400$ ms), in the spectrotemporal CAM of two-hand movement. We interpret that to indicate that the decoder determined that unimanual or bimanual movement–related signal features in the gamma frequency band were important throughout the movement.

In previous reports, gamma bands played an important role in the planning and execution of hand or finger movements [35, 36, 51]. In particular, an increase in the amplitude of γ oscillations during movement execution, a process described as *movement-related* γ *synchronization*, made a great contribution to the classification of arm movements [34, 37–40].

When both arms move, both hemispheres communicate the movement information through the corpus callosum [45, 47, 52]. Seltzer and Pandya *et al* reported that arm movement information was shared through the posterior corpus callosum [52]. Eliassen *et al* asked patients with corpus callosum resection to draw symmetric or asymmetric figures. Before resection surgery, the patients drew good symmetrical figures but were not good at asymmetrical figures. After surgery, the patient whose posterior corpus callosum was resected showed a significant decline in the ability to make symmetrical drawings using both hands [47]. These results were interpreted to indicate that each hemisphere shares information about two-handed coordination through the corpus callosum, especially the posterior corpus callosum. Unlike the gamma frequency caused by local activity in the cortex, the process of transmitting information while interchanging signals between cortex regions or hemispheres occur in a relatively low-frequency band, so we interpreted our results to indicate that the 3D DNN model used both gamma and low-frequency features to classify both-hand movements.

4.4. Correlation between neural signals and the machine's weighting scheme

To confirm how the machine increased classification weights, we checked the ERSP data corresponding to the region accounting for about 90% of the CAM's largest weight value. The bar graphs in figure 5(B) show the results in the left S1, right M1, and right PPC. From the left S1 and right M1 areas, which were the most prominent areas for each left and right unimanual movement class, results were contrary to our expectations. The machine gave more weight at the lowest ERSP power, not the highest ERSP power for classification. On contrary, in the right PPC, where the largest weight was placed in the bimanual movement class, the relationship between the ERSP and weight did not indicate that the machine used the low ERSP power for classification, unlike the one-arm movement results. However, in this case, because the CAM weight of the bimanual movement used broad frequency bands, as shown in figure 5(A), we had to check the ERSP data and CAM weight of the corresponding frequency band to confirm that the results used the relatively low ERSP power for classification (figure S5(B)). Figure S5 illustrates a case in which the CAM weight increased at a relatively high ERSP power for arm movement classification, which could infer that machine learning models can select features using the general perspective that posits relatively strong values as features of classification and also selects relatively weak values as features of a classification. In that way, we confirmed that the model classified the arm movement data not only by increasing the weight for the relatively high ERSP power for each class but also by increasing the weight for the relatively low ERSP power as a characteristic of the classification.

4.5. Limitations of CAM

CAM is a method of finding and showing the characteristics that most emphasize the differences between groups to be distinguished. Thus, the CAM result is different depending on which group is to be classified because the criteria for classification differ. In fact, this is a feature that all machine learning techniques related to classification have.

In this analysis, we analyzed the brain signal, and basically, there are similar signals in each class (or called noise). So, in CAM, there is a possibility to see the difference between each class's feature rather than the exact corresponding feature of each class. We verified this issue in two ways: first with a simple cuboid simulation to make this special situation, and then tested with our brain signal data. (figures S6 and S7). In this test, we could check the CAM result is different depending on the classifier even for the same class. Thus, to interpret CAM from an XAI perspective, we should have a good understanding of the characteristic of the input data and should be careful to decide the number of classes or consider the correlation between classes.

Another limitation of CAM, although this is also a limitation of machine learning, is that the result is depending on input data. In other words, the result of CAM changes depending on what the machine receives as input. Using better input data not only can distinguish a class well (better performance) but also the feature will be expressed well in the CAM. Due to the recent development of deep learning, even if low-dimensional data is served, sometimes classified through high-dimensional analysis. But the impact that directly input high-level data is totally different. In this study as well, it was not easy to analyze the high-level correlations and coherences, which are more complicated features than our simple inputs. Also, these high-level features kind of coherence values were not used as major features for classifying. Therefore, when determining input data, it is necessary to consider whether the characteristic is sufficient to distinguish each class.

5. Conclusion

In this study, we measured the ECoG signals in two monkeys and converted them into spatial-spectral-temporal information that we used as input data for a 3D DNN model. We found that specific temporal, spectral, and spatial information played an important role in classifying our data, thanks to our 3D activation information and 3D machine learning model. Furthermore, the results of our system showed better accuracy than other machine learning techniques (2D DNN, LDA, and SVM). In the 3D CAM, the ipsilateral region and mid gamma frequency range also played an important role in classifying hand movements. We expect that our method can be widely used in neuroscience electrophysiology research from the point of view of XAI and can be improved the brain-machine interface or machine learning performance.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This research was supported by the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2016M3C7A1904987). This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2019M3C7A1031278), the Postdoctoral Fellowship Program of the NRF (NRF-2021R1A6A3A14045108), the Technology Innovation Program funded By the Ministry of Trade, Industry & Energy (MOTIE) (Alchemist Project, 20012461), and the Brain Convergence Research Program of the NRF funded by MSIT (NRF-2021M3E5D2A01021156).

ORCID iDs

Hoseok Choi https://orcid.org/0000-0001-5930-3688
Seokbeen Lim https://orcid.org/0000-0002-4142-3478

References

[1] Craik A, He Y T and Contreras-Vidal J L 2019 Deep learning for electroencephalogram (EEG) classification tasks: a review *J. Neural. Eng.* **16** 031001

[2] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

[3] Adadi A and Berrada M 2018 Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) *IEEE Access* **6** 52138–60

[4] Turek D M Explainable artificial intelligence (XAI) *Defense Advanced Research Projects Agency (DARPA)* (available at: www.darpa.mil/program/explainable-artificial-intelligence) (Accessed 18 August 2020)

[5] Fellous J M, Sapiro G, Rossi A, Mayberg H and Ferrante M 2019 Explainable artificial intelligence for neuroscience: behavioral neurostimulation *Front. Neurosci.* **13** 1346

[6] Bolei Zhou A K, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2921–9 (arXiv:1512.04150)

[7] Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K and Muller K R 2010 How to explain individual classification decisions *J. Mach. Learn. Res.* **11** 1803–31 <Go to ISI>://WOS: 000282522400002

[8] Zeiler M D and Fergus R 2013 Visualizing and understanding convolutional networks (arXiv)

[9] Sundararajan M, Taly, A and Yan Q 2017 Axiomatic attribution for deep networks (arXiv)

[10] Yin Z and Zhang J H 2017 Cross-session classification of mental workload levels using EEG and an adaptive deep learning model *Biomed. Signal Process.* **33** 30–47

[11] Mao Z, Huang Y and Hajinoroozi M 2015 Prediction of driver's drowsy and alert states from EEG signals with deep learning *2015 IEEE 6th Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* pp 493–6

[12] Jiao Z C, Gao X B, Wang Y, Li J and Xu H J 2018 Deep convolutional neural networks for mental load classification based on EEG data *Pattern Recogn.* **76** 582–95

[13] Tabar Y R and Halici U 2017 A novel deep learning approach for classification of EEG motor imagery signals *J. Neural. Eng.* **14** 016003

[14] Wang Z J, Cao L, Zhang Z, Gong X L, Sun Y R and Wang H R 2018 Short time Fourier transformation and deep neural networks for motor imagery brain computer interface recognition *Concurr. Comp-Pract.* E **30** e4413

[15] Liu M F, Wu W, Gu Z H, Yu Z L, Qi F F and Li Y Q 2018 Deep learning based on Batch Normalization for P300 signal detection *Neurocomputing* **275** 288–97

[16] Cho J and Hwang H 2020 Spatio-temporal representation of an electoencephalogram for emotion recognition using a three-dimensional convolutional neural network *Sensors* **20** 3491

[17] Ullah I, Hussain M, Qazi E U and Aboalsamh H 2018 An automated system for epilepsy detection using EEG brain signals based on deep learning approach *Expert. Syst. Appl.* **107** 61–71

[18] Tsiouris K M, Pezoulas V C, Zervakis M, Konitsiotis S, Koutsouris D D and Fotiadis D I 2018 A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals *Comput. Biol. Med.* **99** 24–37

[19] Jonas S, Rossetti A O, Oddo M, Jenni S, Favaro P and Zubler F 2019 EEG-based outcome prediction after cardiac arrest with convolutional neural networks: performance and visualization of discriminative features *Hum. Brain Mapp.* **40** 4606–17

[20] Aslan Z and Akin M 2020 Automatic detection of schizophrenia by applying deep learning over spectrogram images of EEG signals *Trait Signal* **37** 235–44

[21] Li Y, Yang H, Li J, Chen D and Du M 2020 EEG-based intention recognition with deep recurrent-convolution neural network: performance and channel selection by Grad-CAM *Neurocomputing* **415** 225–33

[22] Choi H, Lee J, Park J, Lee S, Ahn K-H, Kim I Y, Lee K-M and Jang D P 2018 Improved prediction of bimanual movements by a two-staged (effector- then-trajectory) decoder with epidural ECoG in nonhuman primates *J. Neural. Eng.* **15** 016011

[23] Gao Huang Z L, van der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 4700–8 (arXiv:1608.06993)

[24] Krizhevsky A 2009 *Learning Multiple Layers of Features from Tiny Images* University of Toronto, 05 August 2009

[25] Zeng Z, Zhou Y, Jenkins O C and Desingh K 2018 Semantic mapping with simultaneous object detection and localization (arXiv)

[26] Chao Z C, Nagasaka Y and Fujii N 2010 Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys *Front. Neuroeng.* **3** 3

[27] Shimoda K, Nagasaka Y, Chao Z C and Fujii N 2012 Decoding continuous three-dimensional hand trajectories from epidural electrocorticographic signals in Japanese macaques *J. Neural. Eng.* **9** 036015

[28] Chen C, Shin D, Watanabe H, Nakanishi Y, Kambara H, Yoshimura N, Nambu A, Isa T, Nishimura Y and Koike Y 2013 Prediction of hand trajectory from electrocorticography signals in primary motor cortex *PLoS One* **8** e83534

[29] Farrokhi B and Erfanian A 2018 A piecewise probabilistic regression model to decode hand movement trajectories from epidural and subdural ECoG signals *J. Neural. Eng.* **15** 036020

[30] Farrokhi B and Erfanian A 2020 A state-based probabilistic method for decoding hand position during movement from ECoG signals in non-human primate *J. Neural. Eng.* **17** 026042

[31] Ganguly K, Secundo L, Ranade G, Orsborn A, Chang E F, Dimitrov D F, Wallis J D, Barbaro N M, Knight R T and Carmena J M 2009 Cortical representation of ipsilateral arm movements in monkey and man *J. Neurosci.* **29** 12948–56

[32] Bundy D T, Szrama N, Pahwa M and Leuthardt E C 2018 Unilateral, 3D arm movement kinematics are encoded in ipsilateral human cortex *J. Neurosci.* **38** 10042–56

[33] Donchin O, Gribova A, Steinberg O, Mitz A R, Bergman H and Vaadia E 2002 Single-unit activity related to bimanual

arm movements in the primary and supplementary motor cortices *J. Neurophysiol.* **88** 3498–517

[34] Pfurtscheller G and Da Silva F H L 1999 Event-related EEG/MEG synchronization and desynchronization: basic principles *Clin. Neurophysiol.* **110** 1842–57

[35] Pfurtscheller G, Graimann B, Huggins J E, Levine S P and Schuh L A 2003 Spatiotemporal patterns of beta desynchronization and gamma synchronization in corticographic data during self-paced movement *Clin. Neurophysiol.* **114** 1226–36

[36] Muthukumaraswamy S D 2010 Functional properties of human primary motor cortex gamma oscillations *J. Neurophysiol.* **104** 2873–85

[37] Pfurtscheller G and Andrew C 1999 Event-related changes of band power and coherence: methodology and interpretation *J. Clin. Neurophysiol.* **16** 512–9

[38] Pfurtscheller G and Aranibar A 1979 Evaluation of event-related desynchronization (Erd) preceding and following voluntary self-paced movement *Electroen Clin. Neuro* **46** 138–46

[39] Han Yuan A D, Gururajan A and Bin H 2008 Cortical imaging of event-related (de)synchronization during online control of brain-computer interface using minimum-norm estimates in frequency domain *IEEE Trans. Neural. Syst. Rehabil. Eng.* **16** 425–31

[40] Cheyne D, Bells S, Ferrari P, Gaetz W and Bostan A C 2008 Self-paced movements induce high-frequency gamma oscillations in primary motor cortex *Neuroimage* **42** 332–42

[41] Diedrichsen J, Wiestler T and Krakauer J W 2013 Two distinct ipsilateral cortical representations for individuated finger movements *Cereb. Cortex* **23** 1362–77 ⩽Go to ISI⩾://WOS:000318649100010

[42] Ames K C and Churchland M M 2019 Motor cortex signals for each arm are mixed across hemispheres and neurons yet partitioned within the population response *Elife* **8** 1–36 ⩽Go to ISI⩾://WOS: 000489621000001

[43] Heming E A, Cross K P, Takei T, Cook D J and Scott S H 2019 Independent representations of ipsilateral and contralateral limbs in primary motor cortex *Elife* **8** 1–26 ⩽Go to ISI⩾://WOS: 000494356800001

[44] Chang S W C, Dickinson A R and Snyder L H 2008 Limb-specific representation for reaching in the posterior parietal cortex *J. Neurosci.* **28** 6128–40

[45] Mooshagian E, Wang C G, Holmes C D and Snyder L H 2018 Single units in the posterior parietal cortex encode patterns of bimanual coordination *Cereb. Cortex* **28** 1549–67

[46] Kermadi I, Liu Y and Rouiller E M 2000 Do bimanual motor actions involve the dorsal premotor (PMd), cingulate (CMA) and posterior parietal (PPC) cortices? Comparison with primary and supplementary motor cortical areas *Somatosens Mot. Res.* **17** 255–71 <Go to ISI>://WOS:000088923800005

[47] Eliassen J C, Baynes K and Gazzaniga M S 1999 Direction information coordinated via the posterior third of the corpus callosum during bimanual movements *Exp. Brain Res.* **128** 573–7

[48] Halsband U, Schmitt J, Weyers M, Binkofski F, Grutzner G and Freund H J 2001 Recognition and imitation of pantomimed motor acts after unilateral parietal and premotor lesions: a perspective on apraxia *Neuropsychologia* **39** 200–16

[49] Serrien D J, Nirkko A C, Lovblad K O and Wiesendanger M 2001 Damage to the parietal lobe impairs bimanual coordination *Neuroreport* **12** 2721–4

[50] Neuper C and Pfurtscheller G 2010 Neurofeedback training for BCI control *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction* ed B Graimann, G Pfurtscheller and B Allison (Berlin: Springer) pp 65–78

[51] Crone N E *et al* 1998 Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis—I. Alpha and beta event-related desynchronization *Brain* **121** 2271–99

[52] Seltzer B and Pandya D N 1983 The distribution of posterior parietal fibers in the corpus-callosum of the rhesus-monkey *Exp. Brain Res.* **49** 147–50 <Go to ISI>://WOS:A1983QA012 00016