



## OPEN ACCESS

## EDITED BY

Sang Wan Lee,  
Korea Advanced Institute of Science  
and Technology (KAIST), Republic of Korea

## REVIEWED BY

Young Ho Kang,  
Korea Advanced Institute of Science  
and Technology (KAIST), Republic of Korea  
Kun Chen,  
Wuhan University of Technology, China

## \*CORRESPONDENCE

Zhaoxiang Wang  
✉ wangzhaoxiang@zju.edu.cn  
Ruilin Li  
✉ RUILIN001@e.ntu.edu.sg

RECEIVED 01 June 2023

ACCEPTED 24 July 2023

PUBLISHED 17 August 2023

## CITATION

Cui J, Yuan L, Wang Z, Li R and Jiang T (2023)  
Towards best practice of interpreting deep  
learning models for EEG-based brain  
computer interfaces.  
*Front. Comput. Neurosci.* 17:1232925.  
doi: 10.3389/fncom.2023.1232925

## COPYRIGHT

© 2023 Cui, Yuan, Wang, Li and Jiang. This is an  
open-access article distributed under the terms  
of the [Creative Commons Attribution License  
\(CC BY\)](#). The use, distribution or reproduction  
in other forums is permitted, provided the  
original author(s) and the copyright owner(s)  
are credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted which  
does not comply with these terms.

# Towards best practice of interpreting deep learning models for EEG-based brain computer interfaces

Jian Cui<sup>1</sup>, Liqiang Yuan<sup>2</sup>, Zhaoxiang Wang<sup>1\*</sup>, Ruilin Li<sup>2\*</sup> and Tianzi Jiang<sup>1,3</sup>

<sup>1</sup>Research Center for Augmented Intelligence, Research Institute of Artificial Intelligence, Zhejiang Lab, Hangzhou, China, <sup>2</sup>School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore, <sup>3</sup>Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China

**Introduction:** As deep learning has achieved state-of-the-art performance for many tasks of EEG-based BCI, many efforts have been made in recent years trying to understand what have been learned by the models. This is commonly done by generating a heatmap indicating to which extent each pixel of the input contributes to the final classification for a trained model. Despite the wide use, it is not yet understood to which extent the obtained interpretation results can be trusted and how accurate they can reflect the model decisions.

**Methods:** We conduct studies to quantitatively evaluate seven different deep interpretation techniques across different models and datasets for EEG-based BCI.

**Results:** The results reveal the importance of selecting a proper interpretation technique as the initial step. In addition, we also find that the quality of the interpretation results is inconsistent for individual samples despite when a method with an overall good performance is used. Many factors, including model structure and dataset types, could potentially affect the quality of the interpretation results.

**Discussion:** Based on the observations, we propose a set of procedures that allow the interpretation results to be presented in an understandable and trusted way. We illustrate the usefulness of our method for EEG-based BCI with instances selected from different scenarios.

## KEYWORDS

brain-computer interface (BCI), convolutional neural network, deep learning interpretability, electroencephalography (EEG), layer-wise relevance propagation (LRP)

## 1. Introduction

A brain-computer interface (BCI) builds a direct communication pathway between the brain and external systems. Among the various neuroimaging techniques for BCIs, electroencephalography (EEG) is the most widely used method due to its noninvasiveness, affordability, and convenience. As one of the most powerful techniques to decode EEG signals, deep learning can automatically capture essential characteristics from a large volume of data by optimizing its parameters through back-propagation and stochastic gradient descent (SGD). It is reported that deep learning has achieved better performance than

conventional methods in many BCI domains such as identifying attentive mental state (Fahimi et al., 2019), movement-related cortical potential recognition (Lawhern et al., 2018), detection of driver drowsiness (Cui et al., 2021a), etc. Despite the success, deep learning has its major drawback of lacking transparency behind its behaviors, which could raise potential concerns of end users on the adoption of BCIs.

In recent years, many efforts have been made to interpret the decisions of deep learning models with application to image and text classification tasks. This is commonly done by generating a heatmap indicating to which extent each pixel of the input contributes to the final classification for a trained model. For the context of EEG-based BCI, the technique can reveal how different components dwelling locally in EEG, e.g., signals generated from different cortical sources, sensor noise, electromyography (EMG), eye movements and eye blinks activities, will affect the classification (Cui et al., 2021a,b, 2022). It is thus possible to know whether the model has learned neurologically meaningful features or the decisions are influenced largely by class-discriminative artifacts from the data, so that the process of improving the models towards better performance and reliability can be facilitated.

Deep learning interpretability has received wide attention in the field of EEG-based BCI (Sturm et al., 2016; Zhou et al., 2016; Bang et al., 2021; Cui et al., 2021a,b, 2022). Despite the wide use, it is neither well understood to which extent the obtained interpretation results can be trusted and how accurately they can reflect the model decisions, nor clearly explained in existing literature why a specific interpretation technique is chosen over others. These observations raise concern about biased conclusions that are made based on misinterpretation of the model decisions. In order to fill this research gap, we conduct a study to evaluate different deep interpretation techniques for EEG-based BCI and explore the best practice of utilizing the techniques. To summarize, the paper makes contributions in the following aspects:

- As far as we know, this is the first comprehensive evaluation of deep learning interpretation techniques across different models and datasets for EEG-based BCI. It provides insights into how seven well-known interpretation techniques, including saliency map (Simonyan et al., 2014), deconvolution (Zeiler and Fergus, 2014), guided backpropagation (Springenberg et al., 2014), gradient  $\times$  input (Shrikumar et al., 2016), integrated gradient (Sundararajan et al., 2017), LRP (Bach et al., 2015), and DeepLIFT (Shrikumar et al., 2016), behave under different conditions.
- Based on the evaluation results, we propose a set of procedures that allow sample-wise interpretation to be presented in an understandable and trusted way. We illustrate the usefulness of our method for EEG-based BCI with instances selected from different scenarios.
- We make the source codes that implement our method in this paper publicly available from Cuijancorbin (2022). This will allow other researchers from this field to conduct a quick test of their models or datasets in order to understand how the classification is influenced by different kinds of components in EEG signals.

In the following part of the paper, state-of-the-art interpretation techniques and their current applications to EEG signal

classification are reviewed in Section “2. Related work.” Datasets and models are prepared in Section “3. Preparation of datasets and models.” Deep learning interpretation techniques are selected and evaluated in Section “4. Evaluation of deep learning interpretability.” Sample-wise interpretation results are analyzed in Section “5. Proposed method for sample-wise interpretation,” which is followed by an extensive study of the applications in Section “6. Application scenarios.” The discussion is presented in Section “7. Discussion” and conclusions are made in Section “8. Conclusion.”

## 2. Related work

### 2.1. Deep learning interpretation techniques

In the field of deep learning interpretability, many techniques have been proposed to interpret deep learning models by generating a contribution map (alternatively called a “relevance” or “attribution” map (Ancona et al., 2017)). Each value in the contribution map indicates the importance of the corresponding pixel (or sampling point) of the input sample to the final decision of the model. Existing interpretation techniques majorly fall into two categories—backpropagation-based methods and perturbation-based methods.

Backpropagation-based methods generate the contribution map through a single or several forward and backward passes through the network. The saliency map method (Simonyan et al., 2014) uses a direct way to estimate the contribution map by calculating the absolute values of gradients back-propagated from the target output. It reflects how much the target output will change when the input is perturbed locally. Zeiler and Fergus (2014) proposed the deconvolution method, which modifies the back-propagation rule in the rectified linear units (ReLU) layer—the backward gradients are zeroed out if their values are negative. By combining these two approaches, Springenberg et al. (2014) proposed the guided backpropagation method which zeros out the gradients at the ReLU layer during back-propagation when either their values or values of inputs in the forward pass are negative. The gradient  $\times$  input method (Shrikumar et al., 2016) multiplies the (signed) partial derivatives with the input sample itself. Sundararajan et al. (2017) proposed the integrated gradient method, where the average gradient is computed by varying the input along a linear path from a baseline. Bach et al. (2015) proposed the layer-wise relevance propagation (LRP) method which redistributes the activation values at the target neuron to the neurons connected to it according to their contributions. The redistribution process continues layer by layer until the input layer is reached. Shrikumar et al. (2016) proposed the DeepLIFT method, which requires running twice forward passes with the input sample and the baseline. Similar to the LRP method, each neuron is assigned a contribution score in a top-down manner according to the difference of activations obtained from the twice forward passes.

The perturbation-based methods only focus on the change of output by perturbation of input, while treating the network as a black box. Specifically, such methods compute the difference in output when removing, masking, or altering the input sample. Zeiler and Fergus (2014) proposed the occlusion sensitivity method

which sweeps a “gray patch” to occlude different parts of an input image and observe how the prediction changes. Similar to the method, [Petsiuk et al. \(2018\)](#) proposed to use binary random masks to perturb the image and distribute the contribution scores among the pixels. [Zintgraf et al. \(2017\)](#) proposed the prediction difference analysis method. They calculated the difference of a prediction by marginalizing each feature (or pixel). [Fong and Vedaldi \(2017\)](#) proposed to use a soft mask with continuous values to preserve discriminative regions for classification. The soft mask is optimized with various regularizations to suppress artifacts. The method was further improved by [Yuan et al. \(2020\)](#) by using a discrete mask optimized with the generative adversarial network.

These interpretation techniques have been previously evaluated on both real ([Hooker et al., 2018](#)) and synthetic ([Tjoa and Guan, 2020](#)) image datasets, as well as synthetic time-series datasets ([Ismail et al., 2020](#)). However, they have not yet been systematically evaluated on EEG datasets. In this paper, we design metrics to evaluate how accurately these techniques can interpret the deep learning models designed for EEG-based BCI.

## 2.2. Deep learning interpretability for EEG-based BCI

For EEG-based BCI, deep learning interpretability can reveal how different factors contained in EEG influence the model decisions. For example, [Bang et al. \(2021\)](#) compared sample-wise interpretation by the LRP method between two subjects and analyzed the potential reasons that lead to the worse performance of one of them. The LRP method was also used by [Sturm et al. \(2016\)](#) to analyze the deep learning model designed for a motor imagery task. They attributed the factors leading to the wrong classification to artifacts from visual activity and eye movements, which dwell in EEG channels from occipital and frontal regions. [Özdenizci et al. \(2020\)](#) proposed to use an adversarial inference approach to learn stable features from EEG across different subjects. By interpreting the results with the LRP method, they showed their proposed method allowed the model to focus on neurophysiological features in EEG while being less affected by artifacts from occipital electrodes. [Cui et al. \(2021a\)](#) used the Class Activation Map (CAM) method ([Zhou et al., 2016](#)) to analyze individual classifications of single-channel EEG signals collected from a sustained driving task. They found the model had learned to identify neurophysiological features, such as Alpha spindles and Theta bursts, as well as features resulting from electromyography (EMG) activities, as evidence to distinguish between drowsy and alert EEG signals. In another work, [Cui et al. \(2022\)](#) proposed a novel interpretation technique by taking advantage of hidden states output by the long short-term memory (LSTM) layer to interpret the CNN-LSTM model designed for driver drowsiness recognition from single-channel EEG. The same group of authors recently reported a novel interpretation technique ([Cui et al., 2022](#)) based on a combination of the CAM method ([Zhou et al., 2016](#)) and the CNN-Fixation methods ([Mopuri et al., 2018](#)) for multi-channel EEG signal classification and discovered stable features across different subjects for the task of driver drowsiness recognition. With the interpretation technique, they also analyzed the reasons behind some wrongly classified samples.

Despite the progress, it is yet not understood to which extent the interpretation results can be trusted and how accurately they can reflect the model decisions. It is also not well explained in existing work why a specific interpretation technique is chosen over others. This research gap motivates us to conduct quantitative evaluations and comparisons of these interpretation techniques on deep learning models designed for mental state recognition from EEG signals.

## 3. Preparation of datasets and models

### 3.1. Dataset

We selected three public datasets belonging to active, reactive and passive BCI domains, respectively, for this study. These datasets have been widely used for the development of deep learning models.

#### 3.1.1. Dataset 1: sensory motor rhythm (SMR)

Motor imagery (MI) is an active BCI paradigm that decodes commands of users when they are imagining the movements of body parts ([De Vries and Mulder, 2007](#)). It is reflected in EEG as desynchronization of sensorimotor rhythm (SMR) over the corresponding sensorimotor cortex areas. The EEG dataset comes from [BCI Competition IV \(2008\)](#) Dataset 2A. It consists of EEG data collected from 9 subjects conducting four different motor imagery tasks, which are the imagination of moving left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). There are two sessions of the experiment conducted on different days for each subject. Each session consists of 6 runs separated by short breaks and each run consists of 48 trails (12 for each imaginary task). Therefore, there are in total 288 trials of a session for each subject.

The EEG data were collected from 22 channels with a sampling rate of 250 Hz. They were bandpass filtered to 0.5 Hz–100 Hz, and further processed by a 50 Hz notch filter to suppress line noise. We followed the practice in [Lawhern et al. \(2018\)](#) to down-sample the signals to 128 Hz and extracted the EEG samples for each trail from 0.5 to 2.5 s after the cue appeared. The dimension of each sample is therefore 22 (channel) × 254 (sample points).

#### 3.1.2. Dataset 2: feedback error-related negativity (ERN)

Feedback error-related negativity (ERN) refers to the amplitude change of EEG, featured as a negative error component and a positive component after a subject receives erroneous visual feedback ([Margaux et al., 2012](#)). In the experiment, ERN was induced by a P300 speller task, which is a passive BCI paradigm for selecting items displayed on the screen by detecting P300 response from EEG signals. The experiment consists of five sessions. Each of the first four sessions contains 12 tasks of 5-letter word spelling, while the last session contains 20 tasks. For each subject there are 4 (sessions) × 12(tasks) × 5(letters) + 1(session) × 20 (tasks) × 5(letters) = 340 trails. 26 subjects (13 men) aged between 20 and 37 participated in the experiment. Their EEG data were recorded at 600 Hz by 56 passive Ag/AgCl EEG sensors (VSM-CTF compatible system) placed according to a 10–20 system. The authors down-sampled the EEG signals to 200 Hz and divided

them into training data (from 16 subjects) and testing data (from 10 subjects). The dataset has been made public from the “BCI Challenge” hosted by Kaggle (B Challenge @ Ner 2015, 2015).

We selected 32 channels out of 56 channels in our study by following the practice described in Margaux et al. (2012). Next, we processed the data according to the steps used in Lawhern et al. (2018) by band-pass filtering the signals to 1–40 Hz and extracting a sample from each trail from 0 to 1.25 s after the feedback was displayed. In this way, each sample has dimensions of 32 (channels)  $\times$  200 (sampling points).

### 3.1.3. Dataset 3: driver drowsiness recognition

This dataset was built with EEG data collected from a sustained-driving experiment (Cao et al., 2019). The subjects were required to drive a car in a virtual experiment and respond quickly to randomly introduced lane-departure events that drifted the car away from the center of the road. Their reaction time was recorded to reflect their level of drowsiness. A total of 27 subjects (aged from 22 to 28) participated in the experiment. The EEG signals were recorded at 600 Hz with 30 electrodes, and band-pass filtered to 1–50 Hz followed by artifact rejection.

We use a pre-processed version of the dataset described in Cui et al. (2021a) for this study. Specifically, the EEG data were down-sampled to 128 Hz. A 3-s length sample prior to each car deviation event was extracted. The dimension of each sample is 30 (channel)  $\times$  384 (sample points). The samples were labeled with “alert” and “drowsy” according to their corresponding global and local reaction time, which were defined in Wei et al. (2018). The samples were further balanced for each subject and class. The final dataset contains 2022 samples in total from 11 different subjects.

## 3.2. Models and implementation

In this study, we select two benchmark deep learning models for the test. The first model is a shallow CNN named “EEGNet,” which was proposed by Lawhern et al. (2018). The model was tested on different active and reactive BCI paradigms (Lawhern et al., 2018). The second model named “InterpretableCNN” was proposed by Cui et al. (2022) in their recent work for driver drowsiness recognition. The model has a compact structure with only seven layers. The selection of the model for this study is motivated by its superior performance over both conventional methods and other state-of-the-art deep learning models including EEGNet on the passive BCI domain (Cui et al., 2022).

The cross-subject paradigm was carried out on the three paradigms, in order to encourage the models to derive stable EEG features across different subjects. For Dataset 1, we followed the procedures described in BCI Competition IV (2008) by splitting the data collected from the first and second sessions into training and testing sets. For each time, the data from one subject collected from the second session are used as the test set, while the data of all the other subjects collected in the first session are used as the training set. The process was iterated until every subject served once as the test subject. For Dataset 2, we followed the original division of the dataset (B Challenge @ Ner 2015, 2015) by using the data from 16 subjects as the training set and the data from the other 10 subjects as the testing set.

For Dataset 3, we followed the practice in Cui et al. (2022) by conducting a leave-one-subject-out cross-subject evaluation on the models.

We set batch size as 50 and used default parameters of the Adam method (Kingma and Ba, 2014) ( $\eta = 0.001$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) for optimization. Considering Dataset 2 contains imbalanced samples, we applied the weights of 1 and 0.41 (which is inverse proportion of the training data) to the “error” and “correct” classes, respectively, in the loss function. Considering the neural networks are stochastic, we repeated the evaluation of each model on each test subject 10 times. In each evaluation, we randomized the network parameters and trained the models from 1 to 50 epochs, and selected the best epoch for the three datasets, respectively.

## 4. Evaluation of deep learning interpretability

### 4.1. Interpretation techniques

#### 4.1.1. Statement of the problem

Formally, suppose an EEG sample  $X \in R^{N \times T}$  ( $N$  is the number of channels and  $T$  is the sample length) is predicted with label  $c$ . The task is to generate a contribution map  $R_c \in R^{N \times T}$ , which assigns a score  $R_c(i, j)$  ( $1 \leq i \leq N$ ,  $1 \leq j \leq T$ ) to each sampling point  $X(i, j)$ , indicating its contribution to the classification.

By averaging  $R_c$  over the temporal dimension, we can obtain a mean contribution map  $\bar{R}_c \in R^N$ , reflecting the average contribution of each EEG channel to the final classification. By interpolating  $\bar{R}_c$  over the whole scalp area, we can obtain a topographic map that reveals the source of signals that contain important features.  $\bar{R}_c$  has been widely used in existing work (Özdenizci et al., 2020; Bang et al., 2021; Cui et al., 2022) to interpret the classification results. In this paper, we name  $\bar{R}_c$  as “channel contribution map”.  $R_c$  is called “contribution map” or alternatively “sample contribution map,” referring to the map generated for the whole sample.

#### 4.1.2. Interpretation techniques and implementation

We select seven widely used interpretation techniques for the test and they are the saliency map (Simonyan et al., 2014), deconvolution (Zeiler and Fergus, 2014), guided backpropagation (Springenberg et al., 2014), gradient  $\times$  input (Shrikumar et al., 2016), integrated gradient (Sundararajan et al., 2017), LRP (Bach et al., 2015), and DeepLIFT (Shrikumar et al., 2016). In comparison to the other methods, the selected ones have the following advantages to be implemented for EEG-based BCI:

- The methods can be applied to deep learning models with different structures.
- The selected interpretation techniques are free of adjustable parameters to be fine-tuned.
- The selected methods are computationally efficient.
- The methods are free of parameters to be randomized for initialization so that the results are reproducible.



Suppose a sample  $x$  is input into a DNN network and the final activation score  $S_c(x)$  is obtained for class  $C$ . Each layer of the DNN model performs linear transformations  $z_j = \sum_i w_{ji}x_i + b_j$  followed by a non-linear mapping  $x_j = f(z_j)$ . The methods of the saliency map, gradient  $\times$  input, and integrated gradients can be, by definition, implemented as a function of the partial derivatives of the target output with respect to each input feature. The methods of LRP and DeepLIFT can also be implemented in a similar way by applying the chain rule for gradients when the instant gradient  $f'(z)$  at each non-linearity is replaced with a function  $g$  that depends on the method (Ancona et al., 2017). We further reformulate the methods of deconvolution and guided back-propagation in the same framework and the contribution score for a pixel  $x_i$  of each method is listed in **Table 1**.

The selected interpretation methods were implemented for EEGNet and InterpretableCNN with the Pytorch library. We used the input sample with zero entries as baselines (Ancona et al., 2017) for integrated gradient and DeepLIFT. We computed the average of gradients generated from the path between the baseline and the input with 100 steps for the integrated gradient method. We implemented the LRP with  $\epsilon$ -rule and DeepLIFT with rescale-rule by modifying the gradient flow in the non-linear activation layers (the ReLU activation layer for InterpretableCNN and three ELU activation layers for EEGNet) with the method proposed by Ancona et al. (2017). LRP is equivalent to gradient  $\times$  input for InterpretableCNN as the model only has ReLU activation for the non-linear layer (Ancona et al., 2017). In order to remove the influence of other samples from the same batch on the interpretation results, we made the batch normalization layers behave linearly by fixing the parameters of batch mean and standard deviation during backpropagation. The parameters were obtained from an additional forward pass of the tested batch data.

## 4.2. Evaluation metrics

### 4.2.1. Sensitivity test

Our first test was inspired by the sensitivity- $n$  test proposed by Ancona et al. (2017). In the method, they randomly perturbed  $n$  pixels (by setting their values to zeros) from the input sample and observed a change in output. Ideally, the sum contribution of the  $n$  points is proportional to the change of the model output score of the predicted class. They varied  $n$  from 1 pixel to about 80% pixels and calculated Pearson correlation coefficient (PCC)  $r$  for each  $n$  as the quality metric of the contribution map.

Our test is different in the aspect that we perturbed the input sample locally in small patches with fixed length  $n$  instead of  $n$  random points from different parts of the sample. In this way, we can evaluate the local accuracy of the contribution map while introducing less high-frequency noise to EEG signals. We limit the patch size  $n$  to 0.1–0.5 of the sample length. That is,  $n$  is 25–127, 20–100 and 38–192, respectively, for Datasets 1, 2 and 3 with sample length of 254, 200, and 384, correspond to 0.45–2.27%, 0.31–1.56%, and 0.33–1.67% of the total sampling points of a sample. The size of  $n$  is chosen empirically so that it will not significantly drift the sample away from its original distribution, at

TABLE 1 Implementation of the interpretation techniques.

Methods	Contribution $R_i^C(x)$
Gradient* Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$
Integrated Gradient	$(x_i - \bar{x}_i) \cdot \int_{\alpha=0}^1 \frac{\partial S_c(\tilde{x})}{\partial \tilde{x}_i} \Big _{\tilde{x}=\bar{x}_i+\alpha(x-\bar{x}_i)} d\alpha$
LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, g = \frac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x}_i) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$
Saliency Map	$\frac{\partial S_c(x)}{\partial x_i}$
Deconvolution	$\frac{\partial^g S_c(x)}{\partial x_i}, g = f'(z)(f'(z) > 0)$
Guided back-propagation	$\frac{\partial^g S_c(x)}{\partial x_i}, g = f'(z)(f'(z) > 0)(f(z) > 0)$

the same time allowing the performance of different methods to be distinguishable. For each  $n$ , we randomly perturbed the sample 100 times and calculate the correlation coefficient as a quality metric of the contribution map. For each perturbation, the channel to be perturbed and position of the patch are randomly selected, so that the perturbation could occur on any part of the sample, resembling the experiment conducted in Ancona et al. (2017).

The channel contribution map was evaluated with a single correlation coefficient, which is obtained by perturbing each channel once.

### 4.2.2. Deletion test

The deletion test proposed in Petsiuk et al. (2018) is used in this study. In this test, we ranked the sampling points of the input sample in descending order according to their scores in the contribution map. By varying  $n$  from 1%, 2%, ..., 100% of the sample size, we calculated the probabilities of the predicted class when the first  $n$  points with the highest values in the contribution map were removed from the sample by setting their values to zeros. A sharp drop of the probability on the predicted class, or alternatively a small area under the probability curve (as a function of  $n$ ) is indicative of a high-quality contribution map.

We performed the deletion test for the channel contribution map in a similar way—each time we removed a channel by setting them to zeros and calculated the probability of the predicted class.

## 4.3. Test settings

Dataset 1 contains EEG data collected from 9 subjects. Each subject has 288 training samples and 288 testing samples with balanced 4 classes. We randomly selected 25 samples for each class from the test samples of each subject. In this way, we have in total 9 (subjects)  $\times$  4 (classes)  $\times$  25 (samples) = 900 samples for evaluating the interpretation results.

Dataset 2 contains EEG data collected from 16 training subjects and 10 testing subjects. Each subject has samples with unbalanced labels. We selected 100 samples from each test subject and thus have in total 10 (subjects)  $\times$  100 (samples) = 1,000 samples for evaluation. For the test subjects 1–6, 9, and 10, we randomly selected 50 samples of each class from each subject. For subjects

7 and 8 with less than 50 samples for the class of error feedback, we used all the samples from this class and randomly selected the rest samples from the class of correct feedback.

Dataset 3 contains EEG data collected from 11 subjects. We randomly selected 50 samples of each class from each subject, and thus have in total  $11 (\text{subjects}) \times 2 (\text{classes}) \times 50 (\text{samples}) = 1,100$  samples for evaluation.

For each sample, we generated a contribution map with random values as a baseline. The corresponding channel contribution map is obtained by averaging the sample contribution map over the temporal dimension.

## 4.4. Results

### 4.4.1. Performance of different interpretation techniques

As can be seen from the results displayed in **Figure 1**, the interpretation techniques fall into two groups by their performance in the tests. The first group of methods, consisting of gradient  $\times$  input, DeepLIFT, integrated gradient, and LRP, have similar and better performance than the baseline method, which uses randomly generated contribution maps, while the second group of methods consisting of the saliency map, deconvolution, and guided backpropagation fail to outperform the baseline method in most conditions. Specifically, in the sensitivity tests the medians of correlation coefficients of the first group of methods range from 0.4 to 1 under different conditions, while the medians of the second group of methods are mostly around 0 under different conditions, which fail to outperform the baseline method. In the deletion tests, the first group of methods reaches a low probability on the predicted class (below 0.2 under all the conditions) when a portion of less than 0.1 of the total data are removed under different conditions, indicating a small portion of features that contribute most to the classification have been successfully localized. However, the second group of methods mostly have large AUC (Area under the Curve), indicating the most important features learned by the deep learning models for classification are not accurately localized.

In order to understand why the interpretation techniques fall into two groups that have distinct performances on both EEGNet and InterpretableCNN models, we conduct a simple analysis to explain the phenomena from a mathematical perspective. We consider the case when the DNN model only contains a linear layer so that the non-linear activation shown in **Table 1** becomes  $f(z_j) = z_j$  and  $f'(z_j) = 1$ . The contribution score of  $x_i$  for integrated gradient, LRP, and DeepLIFT are the same as  $x_i \cdot \frac{\partial S_c(x)}{\partial x_i} = w_{ji}x_i$ , while the deconvolution method becomes the same as the saliency map method, which is  $\frac{\partial S_c(x)}{\partial x_i} = w_{ji}$ . The guided back-propagation is also the same as the saliency map method on cases when  $f(z) > 0$ . However, the second group of methods will generate a constant contribution score  $w_{ji}$  regardless of the inputs. This will be problematic for both the sensitivity test and the deletion test—removal of pixels from the input will not cause a change or drop in the activation score  $S_c(x)$  accordingly. By contrast, the activation score will change with the contribution score  $w_{ji}x_i$  in a correct manner when  $x_i$  is perturbed. Our simple case could partially explain why the selected methods could fall into two groups and their difference in performance. In the actual cases where DNN

models with more complex structures are involved, the results follow a similar pattern as that observed in the simple linear case.

The study justifies the usefulness of the gradient  $\times$  input, DeepLIFT, integrated gradient, and LRP methods for interpreting deep learning models designed for classifying EEG signals, while the methods of saliency map, deconvolution, and guided backpropagation may not be suitable for interpreting deep learning models designed for EEG signal classification. The observation raises concerns about potential misinterpretation of the model decisions in existing literature work, e.g., (Borra et al., 2020), where any of these methods are used.

In the following part of this paper, we focus on methods of gradient  $\times$  input, DeepLIFT, integrated gradient, and LRP. We use them alternatively for interpreting sample-wise classification results.

### 4.4.2. Quality of individual interpretation

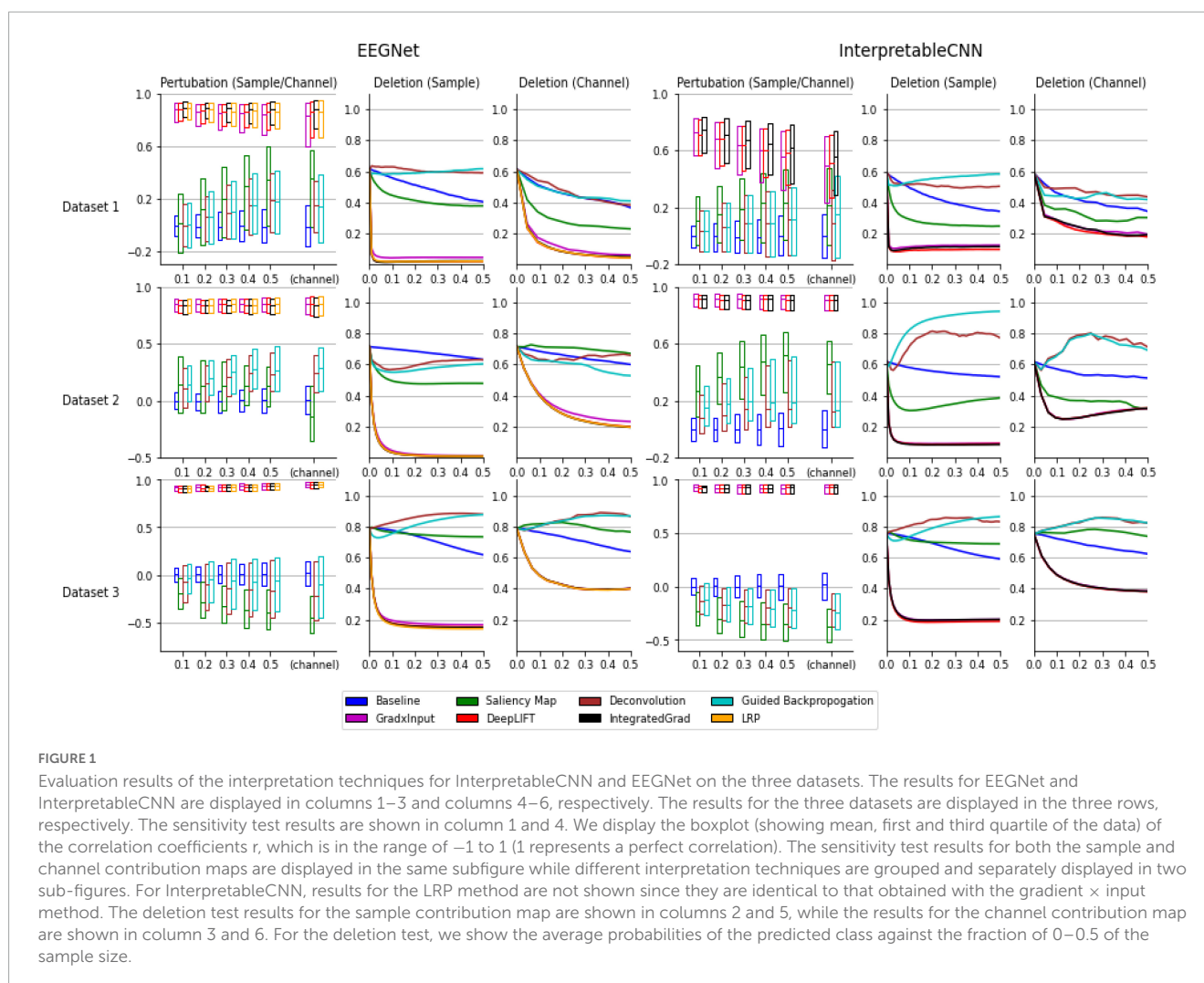
Despite when we focus only on interpretation results produced with the best available methods of gradient  $\times$  input, DeepLIFT, integrated gradient, and LRP, there still exists a large variation in the quality of individual samples under certain circumstances. As can be seen in **Figure 1**, the correlation coefficients have a wide range for Dataset 1 and it increase along with the size of the perturbation patch. When size reaches 0.5 of the sample length, the correlation coefficients fall in the range of 0.6 to 1 for EEGNet and 0.2 to 0.8 for InterpretableCNN, while in comparison the range falls stably within 0.8 to 1 for Dataset 3 under different conditions. The results reveal that the interpretation results generated for some samples are not meaningfully correlated locally with the model outputs. In addition, the model structure also has an impact on the quality of interpretation results. Specifically, for Dataset 2 the correlation coefficients for EEGNet have a narrower range than those of InterpretableCNN. The apparent difference in performance for different models can also be observed from the tests on Dataset 1.

To summarize, both the dataset type and model structure have an impact on the quality of interpretation results, while this problem may not be perfectly solved by changing the interpretation technique. The individual interpretation results should therefore be cautiously treated since they could be uninformative or even misleading for many samples. We further discuss how the individual interpretation results can be presented in an understandable and trusted way in Section “5. Proposed method for sample-wise interpretation.”

## 5. Proposed method for sample-wise interpretation

### 5.1. Enhancement of visualization

The contribution maps are commonly visualized as heatmaps (Zhou et al., 2016; Zou et al., 2018) or colormaps (Cui et al., 2021a,b, 2022) after normalization. However, the colormaps produced in a such way tend to be too noisy to be interpretable for the backpropagation-based methods investigated in this paper. We show a concrete sample in **Figure 2A**, in order to illustrate the problem we meet. The sample is obtained from Dataset 3. It is predicted correctly with the label of “drowsy” by InterpretableCNN



with a probability of 1, indicating that features strongly correlated with the drowsiness have been identified by the model. The sample contains apparent alpha spindles in its first half part (around  $0$ – $1.5$  s), which were found to be a strong signal of drowsiness (Cui et al., 2021a). The contribution map is obtained with the  $\text{grad} \times \text{input}$  method, and it is visualized in Figure 2A as a colormap directly after normalization. However, it is difficult to observe any meaningful pattern from the contribution map, as it is corrupted by the heavy high-frequency noise. We would expect to observe distinguishable features from FCZ, CZ, CPZ, and FT7 channels, which are highlighted in the topographic map.

In order to enhance the visualization, we propose to conduct two additional steps consisting of thresholding and smoothing after normalization. Specifically, we manually set a threshold and remove the unnecessary information from the normalized contribution map below this threshold, in order to reduce the abundant information contained in the sample. After that, we conduct smoothing by moving an average window in order to make the features distinguishable. Coming back to the sample shown in Figure 2A, we performed the proposed processing steps with the thresholds of 2 and 1 for the sample and channel contribution maps, respectively, and the smoothing window with a size of 5. As can be seen in Figure 2B, the visualization is apparently improved –

the alpha spindle features are clearly visible from the channels of FCZ, CZ, CPZ, and FT7, which is consistent with the information revealed from the channel contribution map.

## 5.2. Generation of sample-wise evaluation

As has been discussed in 4.4.1. “Performance of different interpretation techniques”, there exists a large variation in the quality of interpretation results for individual samples even when the best available interpretation technique is used. It is therefore important to conduct sample-wise evaluation and present the results along with the interpretation since the contribution maps themselves cannot reflect how accurately the model decisions are interpreted. In this way, the interpretation results of low accuracies can be excluded so that misinterpretation of the model decisions can be to a large extent avoided.

To keep the consistency of the paper, we use the sensitivity and deletion tests as described in Section “4. Evaluation of deep learning interpretability.” B for the evaluation. The sensitivity test is conducted on the original contribution maps to reflect the best correlation achieved between the perturbed batches and the



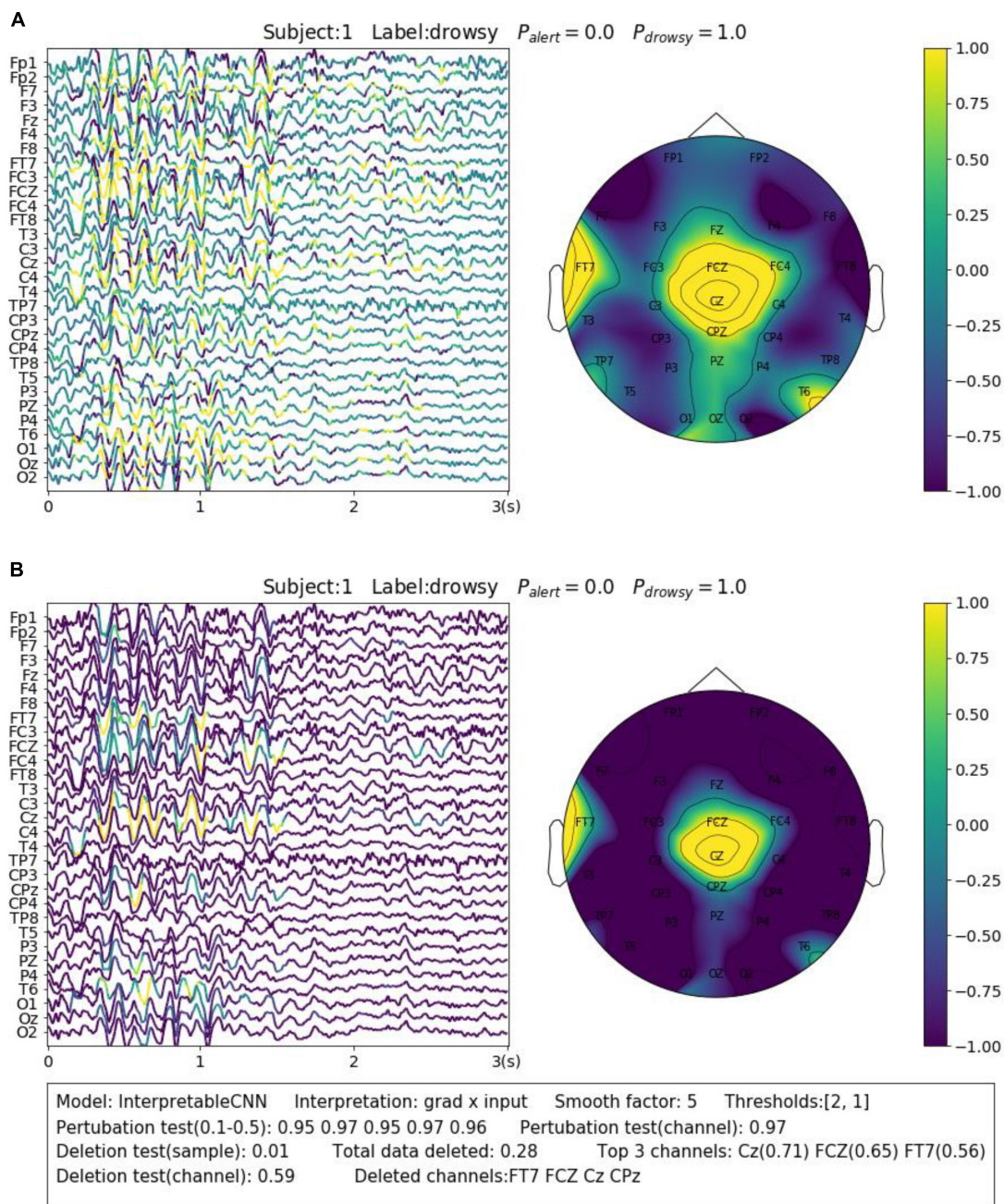


FIGURE 2

Comparison of the contribution maps visualized (A) simply after normalization and (B) after processing with the proposed pipeline for a sample selected from Dataset 3. The subject ID, ground truth label, probabilities of both classes are shown on the top of each sub-figures. The contribution scores are converted to colors overlaid on the signals. An evaluation report is generated and attached below the instance in panel (B). In the report, the deep learning model, the interpretation technique, the smoothing window size and the thresholds (for the sample and channel contribution maps) are displayed in the first line. The perturbation testing results for the sample and channel contribution maps are displayed in the second line. The deletion test results for the sample and channel contribution maps are displayed in the third and fourth lines, respectively. The first item in the third line shows the new probability after deleting the highlighted parts of the signal. The second item shows the total amount (portion) of data deleted. The third item shows the top 3 channels that contain the most of the deleted data and the amount (portion) of deleted data from them. In the fourth line, the first item shows new probability after deletion of the channels listed in the second item.

model output. The deletion test is conducted on the processed contribution map—we remove the highlighted areas (for the sample contribution map) or the channels (for the channel contribution map) and report the probability output by the model on the predicted class. In this way, the deletion test results can be directly related to what is observed from the displayed colormap. We

generate the evaluation report in the text box under each figure. The sample for illustration is shown in **Figure 2B**.

We display another two samples in **Figure 3** for the purpose of illustrating the importance of sample-wise evaluation. By observation of merely the interpretation results of the sample shown in **Figure 3B**, we may draw the conclusion that the features



recorded at around 0.75–1 s from the CP3 channel contribute most to the wrong classification results. However, the evaluation results show that the local regions of the sample contribution map do not correlate well with the model output, and removal of channel CP3 will not actually cause the prediction probability to drop. Without the evaluation, it will be easy to draw biased conclusions from the misleading interpretation results. Despite the fact that some factors that could potentially influence the quality of interpretation can be observed from the obtained results, e.g., the deep learning model structures and the different types of EEG features, it is yet not fully understood what actually lead to the failure of interpretation for some samples, e.g., the one in **Figure 3B**. We leave further investigation on this topic to future work.

## 6. Application scenarios

In this section, we extensively explore how we can benefit from interpreting deep learning models with the method proposed in this paper for different EEG-based BCIs. The applications are explored in two scenarios. In the first scenario, we visualize the neurophysiological features learned by the models from EEG for different datasets, which is an important step of model validation. In the second scenario, we show the advantages of using deep learning interpretability to discover different types of noise and artifacts in the datasets and discuss how the classification accuracy can be potentially improved based on the findings.

### 6.1. Visualization of neurophysiological features

Deriving insights into what the model has learned from the data is an initial step of model validation. The interpretation results allow us to know whether neurophysiological features have been learned from the data to distinguish different mental states. We have selected three representative samples from the three datasets, respectively, for the purpose of illustration.

For the SMR dataset (Dataset 1), the samples were collected while the subjects were imaging movements of different body parts. This reflects in EEG as event-related desynchronization (ERD) during imagination and event-related synchronization (ERS) after imagination of sensorimotor rhythm (SMR) or Mu rhythm (8–13 Hz) over the corresponding sensorimotor cortex areas (Pfurtscheller and Da Silva, 1999). A representative sample is shown in **Figure 4A**. The sample is correctly predicted with the label of “tongue movement” by EEGNet with a probability of 0.95. From the channel contribution map, we can observe that the model has found important features from P1 and POZ channels, which are closest to the sensorimotor area of the tongue (Maetzawa, 2017). From the sample contribution map, it can be observed that the model has recognized the decreased amplitude sensorimotor rhythm (or ERD) at around 0–0.4 s from POz channel, as well as an instant burst of Mu spindles at around 0.4–0.6 s as evidence for the prediction. The amplitude change of SMR reflects neuron activities resulted from the tongue imagination task in the corresponding sensorimotor area.

Feedback related negativity (FRN, or feedback ERN) refers to a specific kind of evoked responses produced by the brain when negative feedbacks are received from external stimuli. It is featured by a negativity peaking of EEG signals around 250 ms after feedback is presented (Miltner et al., 1997). For Dataset 2, FRN occurs when the subject receives an error prediction of the letter displayed on the screen. As it can be seen in the example shown in **Figure 4B**, the model has identified the typical FRN feature, which has a negativity peaking followed by a positive peaking [see Figure 7 in Margaux et al. (2012)], at 0.25–0.5 s after the feedback is presented. The sample is predicted with the “error” class with a probability of 0.96. Removal of the highlighted areas (taking up 0.02 of total data) will cause the probability to drop from 0.96 to 0.33.

For Dataset 3, Alpha spindles, which are characterized by an arrow frequency peak within the alpha band and a low-frequency modulation envelope resulting in the typical “waxing and waning” of the alpha rhythm (Simon et al., 2011), are the most notable features in EEG associated with drowsiness (Cui et al., 2022). A typical sample is shown in **Figure 2B**. The sample is predicted correctly by InterpretableCNN with a probability of 1.0. The Alpha spindle features have been identified from several episodes of the signal majorly in the central cortical areas. Removal of the features (taking up 0.28 of total data) will cause the probability to drop from 1.0 to 0.28.

### 6.2. Discriminating different noises and artifacts

Electroencephalography recording is highly susceptible to various forms and sources of noise and artifacts. The sensor noise contained in EEG is one of the major reasons that affect the model decisions. The interpretation results allow us to understand how different kinds of sensor noise impact the model decisions. We selected two samples from Dataset 3 for illustration. The first sample shown in **Figure 5A** contains apparent sensor noise in the TP7 channel. The model falsely identified several local regions of the sensor noise as evidence to support a decision. The significant amplitude changes of the signal in the TP7 channel could be caused by loose conduct between the sensor and the skin. Such kinds of sensor noise that seriously affect the model’s decision should be cleaned from the data in the pre-processing phase. The second sample shown in **Figure 5B** contains heavy high-frequency noise in several EEG channels, e.g., CP4. The model identifies several regions from the noise areas as evidence for classification. We have observed many samples containing similar noise in the dataset that are correctly classified with alert labels, which indicates that this kind of noise could have a strong relationship with the alert state. The noise is caused by electromyography (EMG) activities resulting from the tension of scalp muscles. They usually dominate the wakeful EEG signals (Britton et al., 2016) and become the most apparent feature of alertness (Cui et al., 2021a,b, 2022), while the cortical source Beta activities with very low amplitude (Da Silva et al., 1976) in wakeful EEG are not as easy to be distinguished.

Eye blinks and movements are another common source of artifacts in EEG signals. The interpretation results allow us to identify samples that contain such kinds of artifacts and understand how they affect the model decision. For Dataset 2, we find there are

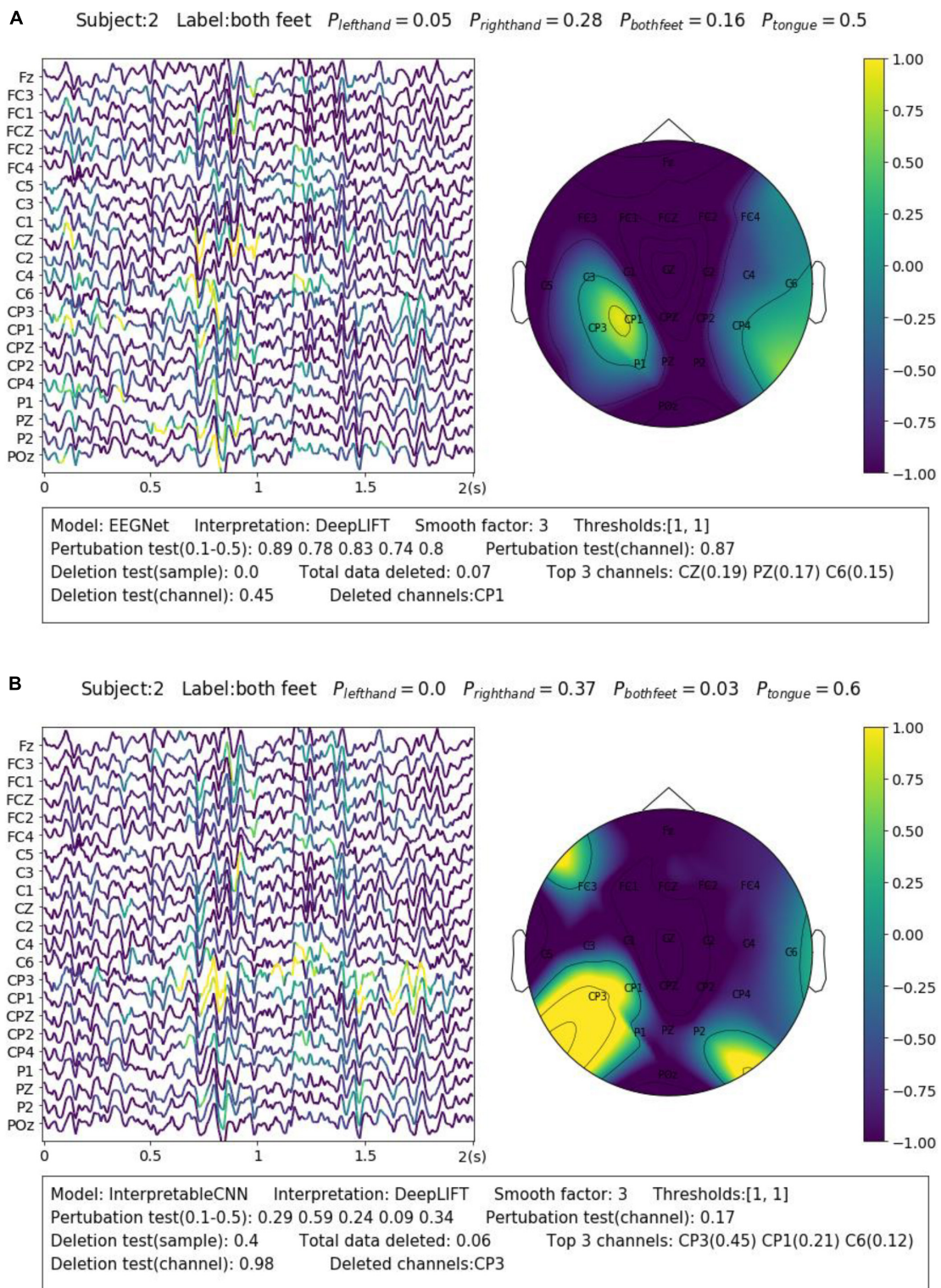


FIGURE 3

Comparison of the interpretation results on a sample classified by (A) EEGNet and (B) InterpretableCNN. The contribution maps generated for the two samples display similar patterns, while the qualities of the interpretation results vary greatly as revealed by the evaluation report. For the sample shown in panel (A), the sample contribution map is well correlated with the model output with the correlation coefficients in the range of 0.74–0.89. Removal of the highlighted regions (taking up 0.07 of total data) of the first sample will cause the probability drop from 0.5 to 0.0, while removal of the channel (CP1) with the highest contribution will cause the probability drop slightly from 0.5 to 0.45. However, for second sample shown in panel (B), the local regions of the sample contribution map do not correlate well with the model output as the correlation coefficients fall in the range of 0.09–0.59. Removal of the most important channel (CP3) for the second sample will cause the probability, on the contrary, increase from 0.6 to 0.98. Therefore, the evaluation results allow us to confidently reject the interpretation shown in panel (B).



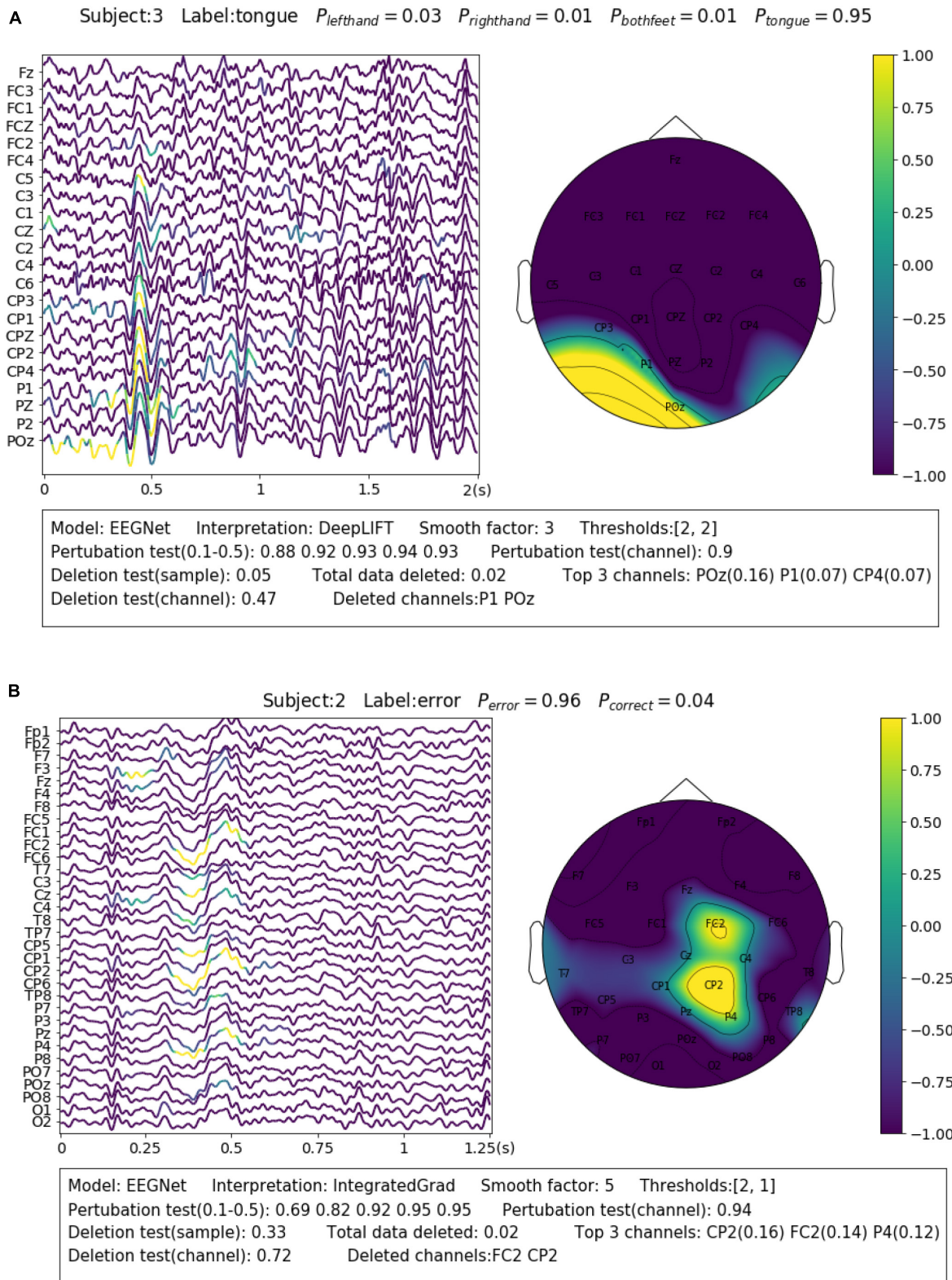
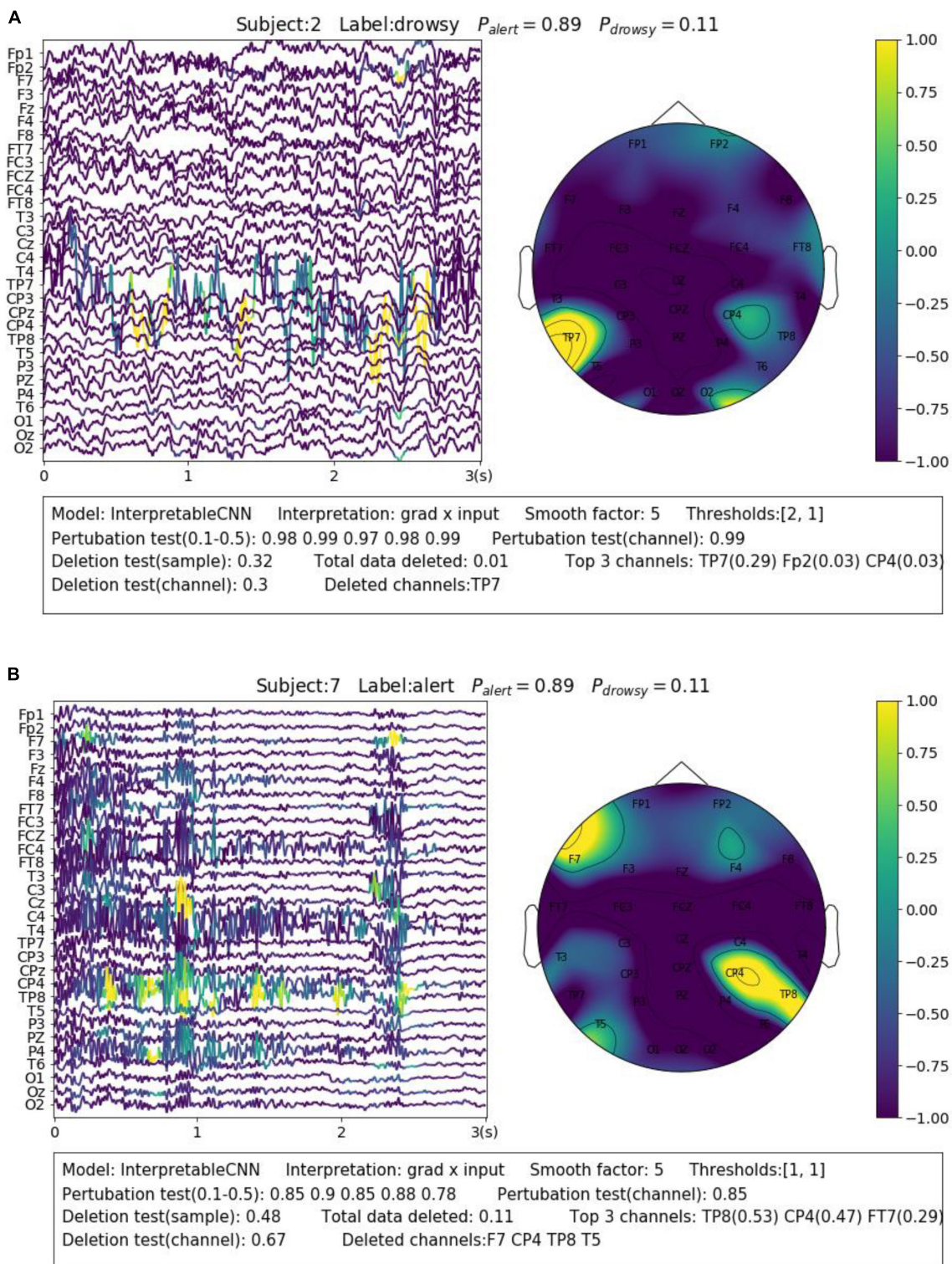


FIGURE 4 Visualization of the learned neurophysiological features from two samples. The first sample (A) is from aset 3 and the second sample (B) is from Dataset 2.

many samples, similar to the case shown in **Figure 6A**, containing the eye blink artifacts while they are identified by the model as evidence of the “correct” class. The eye blinks could be a sign of

relief after the subjects stare at the screen with a high degree of concentration in the P300 task. Such class-discriminative artifacts should be removed from the dataset and the model is expected

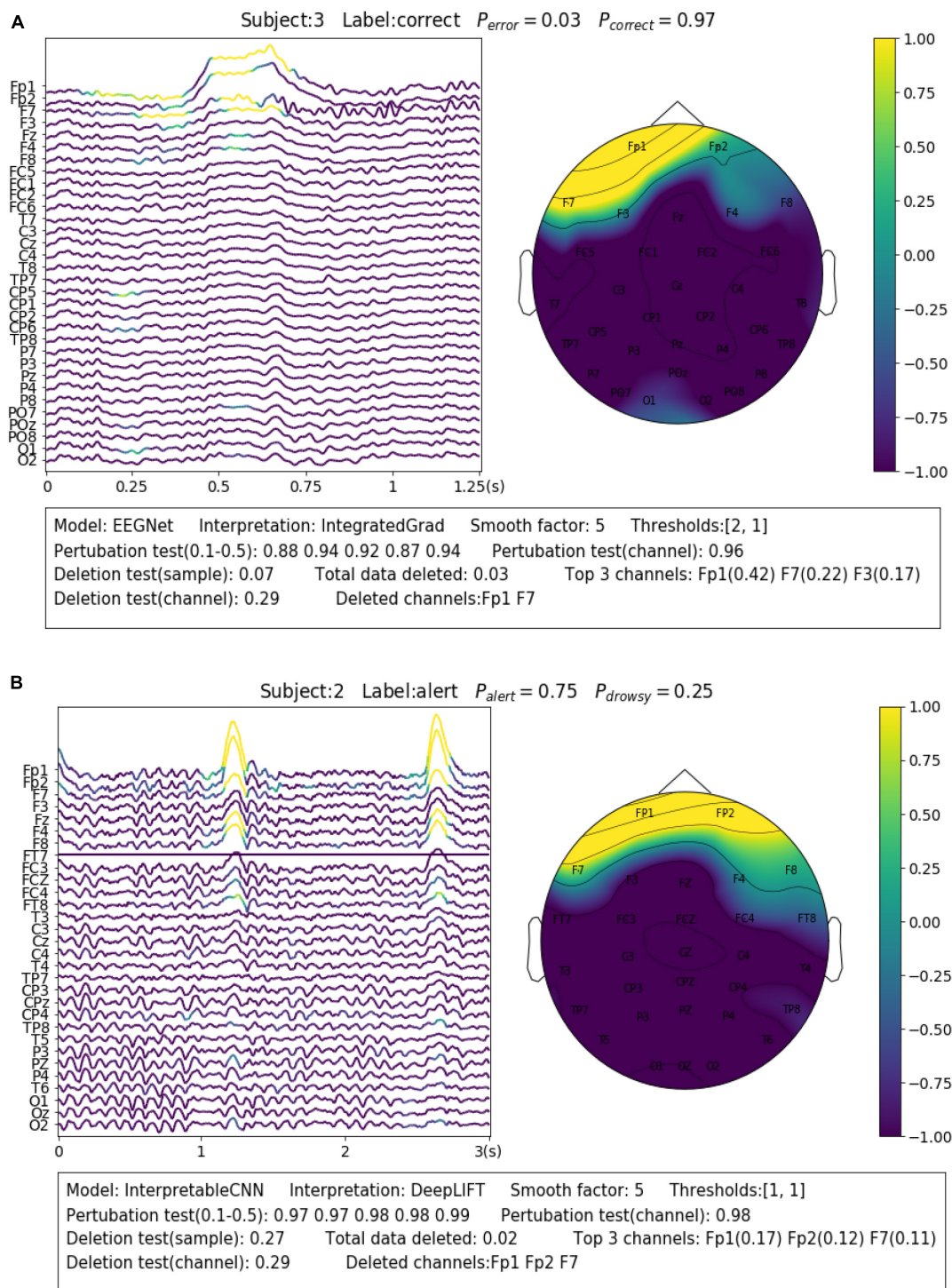


**FIGURE 5** Visualization of the interpretation results for selected samples containing different types of noise. **(A)** Apparent sensor noise in TP7 channel. **(B)** Heavy high-frequency noise in multiple channels.

to learn EEG features (e.g., ERP) generated from cortical sources instead. However, the case is on the contrary for Dataset 3, where the eye blink and movement features are overly cleaned. The rapid eye blinks, as reflected in EEG a short-term pulse in the frontal

channels (**Figure 6B**), are indicators of the alert and wakeful state. They have been found by deep learning models as important features for classification (Cui et al., 2022). However, the eye blink and movement features are overly cleaned in the pre-processing





**FIGURE 6** Visualization of the interpretation results for selected samples containing different types of noise. **(A)** The sample contains eye blinks that could be a sign of relief after the subjects stare at the screen with high concentration in the P300 task. Such class-discriminative artifacts should be removed from the dataset. **(B)** The sample contains eye blinks that are indicators of the alert and wakeful state.

phase for Dataset 3, which causes the prediction accuracy to drop around 0.3 for Subject 2 in our test. We show in **Figure 6B** an uncleaned sample from Subject 2 and it can be seen that the model has made the right prediction based on such kind of features. Removal of the features will cause the model to make the wrong prediction.

The observations above lead to the conclusion that different types of noise should be treated differently rather than indiscriminately removed from EEG signals. The noise and artifacts defined in one scenario could become important features in another scenario. Deep learning interpretability provides us with the advantage of understanding how they impact the

model decisions so that we can take proper strategies accordingly to deal with them.

## 7. Discussion

In this paper, we investigated the topic of applying deep learning interpretability to EEG signal recognition. Despite the wide application of deep learning, there are yet no guidelines or recommendations on how to interpret the results, what methods should be used, and how accurately they can reflect the model decisions for EEG-based BCI. In order to fill this research gap, we conducted quantitative studies to evaluate existing interpretation techniques and explored the best practice of interpreting deep learning designed for EEG signal recognition.

We investigated seven well-known interpretation techniques, including saliency map (Simonyan et al., 2014), deconvolution (Zeiler and Fergus, 2014), guided backpropagation (Springenberg et al., 2014), gradient  $\times$  input (Shrikumar et al., 2016), integrated gradient (Sundararajan et al., 2017), LRP (Bach et al., 2015), and DeepLIFT (Shrikumar et al., 2016), behave under different conditions. Our initial observation is obvious that the interpretation techniques fall into two groups by their performance in the tests. The first group of methods, consisting of gradient  $\times$  input, DeepLIFT, integrated gradient, and LRP, have similar and better performance than the baseline method, which uses randomly generated contribution maps, while the second group of methods consisting of the saliency map, deconvolution, and guided backpropagation fails to outperform the baseline method in most conditions. This is curious, we try our best to understand their differences underneath the behavior. By definition, all the methods require a forward pass and a back pass to generate a gradient. They are different in handling the non-linear layer of the network.

The results revealed the importance of selecting a proper interpretation technique in the first step. Existing interpretation techniques, e.g., the saliency map method, despite being widely used in existing work for interpreting learned EEG patterns (Borra et al., 2020), could actually fail to outperform baseline with randomly generated contribution maps in certain circumstances. In addition, we also find that the quality of the interpretation results is inconsistent for individual samples despite when a method with overall good performance is used. Many factors, including the model structure and types of features in the samples, could potentially affect the quality of the interpretation results. It is therefore recommended to conduct a sample-wise evaluation to validate the results. By far as we know these findings have not yet raised wide awareness in work interpreting deep learning for EEG-based BCI, e.g., (Sturm et al., 2016; Özdenizci et al., 2020).

In order to make the interpretation results understandable, we proposed a few processing steps that can effectively enhance the visualization. Furthermore, we extensively used deep learning interpretability to explore how different types of noise and artifacts in the datasets can affect the model decisions. We show the benefits with examples. The method allows us to conclude that different types of noise and artifacts should be treated differently based on a comprehensive understanding of the overall pattern

learned from the dataset. By far as we know, this has not yet been realized in existing studies. The noise and artifacts defined in one scenario could become important features in another scenario. Deep learning interpretability could be potentially used as a powerful tool to discover the patterns underlying a dataset so that a proper strategy can be specifically designed in the pre-processing pipeline.

## 8. Conclusion

In this paper, we explored the best practice of applying deep learning interpretability to EEG-based BCI. Firstly, we surveyed existing deep learning interpretation techniques and shortlisted seven of them that can be applied to deep learning models with different structures. We designed evaluation metrics and tested them with two benchmark deep-learning models on three different EEG datasets.

The results show that the interpretation techniques of gradient  $\times$  input, DeepLIFT, integrated gradient, and LRP, have similar and better performance than the baseline method, which uses randomly generated contribution maps, while the methods consisting of the saliency map, deconvolution, and guided backpropagation fail to outperform the baseline method in most conditions. The obtained results reveal the importance of selecting a proper interpretation technique for illustration. In addition, we also find that the quality of the interpretation results is inconsistent for individual samples despite when a method with overall good performance is used. Many factors, including the model structure and types of features in the samples, could potentially affect the quality of the interpretation results.

Based on the obtained results, we proposed a set of procedures that allow the interpretation results to be presented in an understandable and trusted way. We extensively used our proposed method to explore how different types of noise and artifacts in the datasets can affect the model decisions and used examples to illustrate how deep learning interpretability can benefit EEG-based BCI. Our work illustrates a promising direction of using deep learning interpretability to discover meaningful patterns from complex EEG signals.

## Data availability statement

The dataset analyzed in this study are publicly available. This data can be found here: <http://www.bbci.de/competition/iv/#dataset2a>, <https://www.kaggle.com/c/inria-bci-challenge>, and [https://figshare.com/articles/dataset/EEG\\_driver\\_drowsiness\\_dataset/14273687](https://figshare.com/articles/dataset/EEG_driver_drowsiness_dataset/14273687).

## Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## Author contributions

JC, LY, ZW, and RL designed the algorithm and/or interpreted the data and performed the experiments and analyzed the data. JC, LY, ZW, and TJ drafted the manuscript. LY, ZW, and TJ revised the manuscript critically for important intellectual content. All authors approved the final version of the manuscript to be published and agreed to be accountable for all aspects of the manuscript.

## Funding

This work was partially supported by Key Research Project of Zhejiang Lab, China (No. 2022KI0AC02) and Youth Foundation Project of Zhejiang Lab, China (No. K2023KI0AA01).

## References

- Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv [Preprint]*. arXiv:1711.06104.
- B Challenge @ Ner 2015 (2015). Available online at: <https://www.kaggle.com/c/inria-bci-challenge> (accessed August 8, 2023).
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10, e0130140. doi: 10.1371/journal.pone.0130140
- Bang, J.-S., Lee, M., Fazli, S., Guan, C., and Lee, S. (2021). "Spatio-spectral feature representation for motor imagery classification using convolutional neural networks," in *Proceedings of the IEEE transactions on neural networks and learning systems*, (Piscataway, NJ). doi: 10.1109/TNNLS.2020.3048385
- BCI Competition IV (2008). Available online at: <http://www.bbci.de/competition/iv/#dataset2a> (accessed August 8, 2023).
- Borra, D., Fantozzi, S., and Magosso, E. (2020). Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination. *Neural Netw.* 129, 55–74. doi: 10.1016/j.neunet.2020.05.032
- Britton, J. W., Frey, L. C., Hopp, J. L., Korb, P., Koubeissi, M. Z., Lievens, W. E., et al. (2016). *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*. Chicago, IL: American Epilepsy Society.
- Cao, Z., Chuang, C., King, J., and Lin, C. (2019). Multi-channel EEG recordings during a sustained-attention driving task. *Sci. Data* 6, 1–8. doi: 10.1038/s41597-019-0027-4
- Cui, J., Lan, Z., Liu, Y., Li, R., Li, F., Sourina, O., et al. (2021a). A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods* 202, 173–184. doi: 10.1016/j.ymeth.2021.04.017
- Cui, J., Lan, Z., Sourina, O., and Müller-Wittig, W. (2022). "EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network," in *Proceedings of the IEEE transactions on neural networks and learning systems*, (Piscataway, NJ). doi: 10.1109/TNNLS.2022.3147208
- Cui, J., Lan, Z., Zheng, T., Liu, Y., Sourina, O., Wang, L., et al. (2021b). "Subject-independent drowsiness recognition from single-channel EEG with an interpretable CNN-LSTM model," in *Proceedings of the 2021 International Conference on Cyberworlds (CW)*, (Caen). doi: 10.1109/CW52790.2021.00041
- Cuijiancorbin (2022). *Towards-best-practice-of-interpreting-deep-learning-models-for-EEG-based-BCI*. Available online at: <https://github.com/cuijiancorbin/Towards-Best-Practice-of-Interpreting-Deep-Learning-Models-for-EEG-based-BCI/pulse> (accessed July 4, 2022).
- Da Silva, F. L., van Rotterdam, A., Barts, P., van Heusden, E., and Burr, W. (1976). "Models of neuronal populations: The basic mechanism of rhythmicity," in *Perspectives of brain research, progress in brain research*, Vol. 45, eds M. A. Corner and D. F. Swaab (Amsterdam: Elsevier), 281–308. doi: 10.1016/S0079-6123(08)60995-4
- De Vries, S., and Mulder, T. (2007). Motor imagery and stroke rehabilitation: A critical discussion. *J. Rehabil. Med.* 39, 5–13. doi: 10.2340/16501977-0020
- Fahimi, F., Zhang, Z., Goh, W. B., Lee, T., Ang, K. K., and Guan, C. (2019). Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *J. Neural Eng.* 16:026007. doi: 10.1088/1741-2552/aaf3f6
- Fong, R. C., and Vedaldi, A. (2017). "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE international conference on computer vision*, (Piscataway, NJ). doi: 10.1109/ICCV.2017.371
- Hooker, S., Erhan, D., Kindermans, P., and Kim, B. (2018). A benchmark for interpretability methods in deep neural networks. *arXiv [Preprint]*. arXiv:1806.10758.
- Ismail, A. A., Gunady, M., Corrada Bravo, H., and Feizi, S. (2020). Benchmarking deep learning interpretability in time series predictions. *arXiv [Preprint]*. arXiv:2010.13924.
- Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv [Preprint]*. arXiv:1412.6980.
- Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c
- Maewaza, H. (2017). Cortical mechanisms of tongue sensorimotor functions in humans: A review of the magnetoencephalography approach. *Front. Hum. Neurosci.* 11:134. doi: 10.3389/fnhum.2017.00134
- Margaux, P., Maby, E., Daligault, S., Bertrand, O., and Mattout, J. (2012). Objective and subjective evaluation of online error correction during P300-based spelling. *Adv. Hum. Comput. Interact.* 2012:13. doi: 10.1155/2012/578295
- Miltner, W. H., Braun, C. H., and Coles, M. G. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *J. Cogn. Neurosci.* 9, 788–798. doi: 10.1162/jocn.1997.9.6.788
- Mopuri, K. R., Garg, U., and Babu, R. V. (2018). Cnn fixations: An unraveling approach to visualize the discriminative image regions. *IEEE Trans. Image Process.* 28, 2116–2125. doi: 10.1109/TIP.2018.2881920
- Özdenizci, O., Wang, Y. E., Koike-Akino, T., and Erdogmus, D. (2020). Learning invariant representations from EEG via adversarial inference. *IEEE Access* 8, 27074–27085. doi: 10.1109/ACCESS.2020.2971600
- Petsiuk, V., Das, A., and Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv [Preprint]*. arXiv:1806.07421.
- Pfurtscheller, G., and Da Silva, F. L. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clin. Neurophysiol.* 110, 1842–1857. doi: 10.1016/s1388-2457(99)00141-8
- Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv [Preprint]*. arXiv:1605.01713.
- Simon, M., Schmidt, E. A., Kincses, W. E., Fritzsche, M., Bruns, A., Aufmuth, C., et al. (2011). EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. *Clin. Neurophysiol.* 122, 1168–1178. doi: 10.1016/j.clinph.2010.10.044
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). "Deep inside convolutional networks: Visualising image classification models and saliency maps" in *Proceedings of the international conference on learning representations*, (Princeton, NJ: Citeseer).
- Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A. (2014). Striving for simplicity: The all convolutional net. *arXiv [Preprint]*. arXiv:1412.6806.
- Sturm, I., Lapuschkin, S., Samek, W., and Müller, K. (2016). Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Methods* 274, 141–145. doi: 10.1016/j.jneumeth.2016.10.008

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic attribution for deep networks," in *Proceedings of the international 34th conference on machine learning*, (Sydney, NSW).
- Tjoa, E., and Guan, C. (2020). Quantifying explainability of saliency methods in deep neural networks. *arXiv [Preprint]*. arXiv:2009.02899.
- Wei, C., Wang, Y., Lin, C., and Jung, T. (2018). Toward drowsiness detection using non-hair-bearing EEG-based brain-computer interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 400–406. doi: 10.1109/TNSRE.2018.2790359
- Yuan, H., Cai, L., Hu, X., Wang, J., and Ji, S. (2020). "Interpreting image classifiers by generating discrete masks," in *Proceedings of the IEEE transactions on pattern analysis and machine intelligence*, (Piscataway, NJ). doi: 10.1109/TPAMI.2020.3028783
- Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *Proceedings of the European conference on computer vision*, (Cham). doi: 10.1007/978-3-319-10590-1\_53
- Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (Piscataway, NJ). doi: 10.1109/CVPR.2016.319
- Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *arXiv [Preprint]*. arXiv:1702.04595.
- Zou, X., Kolouri, S., Pilly, P. K., and Krichmar, J. L. (2018). Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.* 126, 1084–1102. doi: 10.1016/j.neunet.2020.01.031