frontiers | Frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fninf.2022.872035

A Systematic Approach for Explaining Time and Frequency Features Extracted by Convolutional Neural Networks From Raw Electroencephalography Data

Charles A. Ellis[1,2], Robyn L. Miller[2,3] and Vince D. Calhoun[1,2,3]*

[1] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States, [2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, United States, [3] Department of Computer Science, Georgia State University, Atlanta, GA, United States

In recent years, the use of convolutional neural networks (CNNs) for raw resting-state electroencephalography (EEG) analysis has grown increasingly common. However, relative to earlier machine learning and deep learning methods with manually extracted features, CNNs for raw EEG analysis present unique problems for explainability. As such, a growing group of methods have been developed that provide insight into the spectral features learned by CNNs. However, spectral power is not the only important form of information within EEG, and the capacity to understand the roles of specific multispectral waveforms identified by CNNs could be very helpful. In this study, we present a novel model visualization-based approach that adapts the traditional CNN architecture to increase interpretability and combines that inherent interpretability with a systematic evaluation of the model *via* a series of novel explainability methods. Our approach evaluates the importance of spectrally distinct first-layer clusters of filters before examining the contributions of identified waveforms and spectra to cluster importance. We evaluate our approach within the context of automated sleep stage classification and find that, for the most part, our explainability results are highly consistent with clinical guidelines. Our approach is the first to systematically evaluate both waveform and spectral feature importance in CNNs trained on resting-state EEG data.

Keywords: explainable AI, spectral explainability, EEG, CNNs, sleep stage classification, deep learning

OPEN ACCESS

Edited by:
Pedro Antonio Valdes-Sosa, University of Electronic Science and Technology of China, China

Reviewed by:
Alessio Burrello, University of Bologna, Italy
Anubha Gupta, Indraprastha Institute of Information Technology, Delhi, India

****Correspondence:***
Charles A. Ellis
cae67@gatech.edu

Received: *09 February 2022*
Accepted: *16 May 2022*
Published: *31 May 2022*

Citation:
Ellis CA, Miller RL and Calhoun VD (2022) A Systematic Approach for Explaining Time and Frequency Features Extracted by Convolutional Neural Networks From Raw Electroencephalography Data. Front. Neuroinform. 16:872035. doi: 10.3389/fninf.2022.872035

INTRODUCTION

In recent years, the use of convolutional neural networks (CNNs) in the analysis of raw electroencephalography (EEG) data has grown considerably. These classifiers have the advantage over standard machine learning and deep learning classifiers paired with manual feature extraction in that they don't require any prior assumptions about the important features within the data and that they automate feature extraction. While this is the case, pairing automated feature extraction

with raw time-series data also causes problems with explainability, which is highly important in sensitive domains like healthcare. As such, novel methods for explaining CNNs trained on raw EEG data are needed. In this study, we present a novel approach that pairs a CNN architecture adapted for increased interpretability with a series of systematic model perturbations that provide valuable insight into the features extracted by the CNN and the relative importance of those features. Unlike previous approaches that have mainly provided insight into key frequency feature extracted by CNNs, we provide insight into both the frequencies and waveforms extracted by CNNs.

Prior to the recent trend of applying CNNs to raw electrophysiology data for automated feature extraction, it was common to manually create features and apply machine learning or deep learning approaches with traditional explainability methods when analyzing electrophysiology data. The user-created features typically reflected either time domain or frequency domain (Ince et al., 2008; Kwon et al., 2018; Chen et al., 2019; Ruffini et al., 2019; Ellis et al., 2021g) aspects of the data. A strength of this approach was that it could be applied alongside explainability methods initially developed outside the domain of electrophysiology analysis like layer-wise relevance propagation (LRP) (Bach et al., 2015), Grad-CAM (Selvaraju et al., 2020), and activation maximization (Simonyan et al., 2013). However, the use of user-selected input features also inherently limited the available feature space, thereby limiting the potential performance of classifiers.

As such, more studies have begun to apply CNNs to raw electrophysiology analysis. While this application can improve model performance, applying traditional explainability methods to raw time-series samples makes it very difficult to know what time or frequency features are extracted by classifiers and to draw global conclusions about the importance of extracted features (Sturm et al., 2016). It should be noted that this difficulty is not applicable to identifying spatial importance (Sturm et al., 2016) or modality importance (Ellis et al., 2021a,b,f, 2022), in multichannel or multimodal classification, respectively. However, this difficultyis applicable when trying to understand the temporal and spectral features extracted by classifiers.

In response to the need for improved explainability in CNNs applied to raw electrophysiology data, a new field of explainability for CNN-based raw electrophysiology classification has developed. The vast majority of these methods provide insight into frequency-based features extracted by CNNs. These methods can loosely be divided into four categories: (Ellis et al., 2021g) interpretable architectures (Chen et al., 2019) activation maximization approaches, (Kwon et al., 2018) perturbation approaches, and (Ince et al., 2008) model visualization approaches. Interpretable architectures involve structuring filters in the first convolutional layer such that they only extract spectral features (Borra et al., 2019, 2020). While these methods are very innovative, they still inherently restrict the feature space to frequency features. Several studies have presented methods that use activation maximization to identify spectral features that maximize activation of the CNN (Tsinalis et al., 2016b; Ellis et al., 2021h; Pathak et al., 2021). Two studies examined the effect of sinusoids at different frequencies upon activations of nodes in the early layers of the classifier (Tsinalis et al., 2016b; Ellis et al., 2021h; Pathak et al., 2021), and the remaining study varied the spectral representation of a sample until it maximized the activation of the final output node (Ellis et al., 2021h). Other existing studies involve perturbing canonical frequency bands of samples and examining the effect upon the predictions (Schirrmeister et al., 2017; Ellis et al., 2021e) or performance of a classifier (Nahmias and Kontson, 2020). The last category of methods involves training a CNN with long first-layer filters that can be converted to the frequency domain after training and visualized to examine the spectral features extracted by the model (Tsinalis et al., 2016b). While these methods do provide useful insight into extracted spectra, they do not provide effective insight into extracted time-domain features.

For the most part, existing approaches that provide insight into insight into time-domain features of EEG are of limited utility. One study perturbs windows of a time-domain sample (Pathak et al., 2021). However, when datasets consist of thousands of samples and there is no way to combine insights from the perturbation of each sample, that approach does not provide useful global conclusions on the nature of the time-domain features extracted. Another study used activation maximization to optimize the spectral content of a sample (Ellis et al., 2021h). While the method does yield a sample in the time domain that maximizes activation for a particular class, it does not provide insight into the relative importance of different time domain features. A couple of other studies have used activation maximization for insight into the time domain (Yoshimura et al., 2019, 2021). However, they were only applied to networks trained on samples that were around 30 time points long, and EEG samples can be hundreds to thousands of time points long. Additionally, a previous study showed that these methods do not generalize well to sample lengths relevant for EEG analysis (Ellis et al., 2021h). It should be noted that existing explainability methods can be applied to some forms of electrophysiology like electrocardiograms (ECG) that have regularly repeated waveforms (Porumb et al., 2020; Frick et al., 2021). However, these applications rely upon the regular repetition of waveforms, and that repetition is not present in forms of electrophysiology like resting-state magnetoencephalography and EEG. Methods which provide useful insight into time domain features (Ellis et al., 2021d) extracted by CNNs for EEG involve training a CNN with filters in the first layer that are long enough to extract distinct waveforms (Tsinalis et al., 2016b; Lawhern et al., 2018). The filters can then be visualized and perturbed to examine the relative importance of each waveform within the filters to the classifier performance. We developed this method in a previous study (Ellis et al., 2021d), and we expand upon it here to provide a systematic approach for explaining CNNs trained on raw electrophysiology time-series.

In this study, we use sleep stage classification as a testbed to demonstrate the utility of our approach. Sleep stage classification has several noteworthy characteristics that make it ideal for our application (Ellis et al., 2021g). The domain of sleep stage classification has well-characterized spectral and temporal features, so we can evaluate whether our explainability results are consistent with established scientific knowledge (Iber et al., 2007;

Chen et al., 2019). There are multiple large publicly available datasets within the domain of sleep stage classification that help with reproducibility of analyses (Quan et al., 1997; PhysioNet, 2002; Khalighi et al., 2016; Kwon et al., 2018). Multiple studies have already presented explainability methods for the domain of sleep stage classification, which will make it easier to compare our findings with results from previous studies (Tsinalis et al., 2016b; Ellis et al., 2021c,e,h; Pathak et al., 2021).

In summary, in this study, we present a novel systematic approach for gaining insight into the features extracted by CNNs on raw, resting-state EEG data and into the relative importance of those features. We train a CNN for automated sleep stage classification with a publicly available dataset and structure the first layer of the architecture such that we can visualize the waveforms extracted by the model. We convert the first layer filters of the model to the frequency domain and identify clusters of filters with distinct spectral characteristics. We then examine the relative importance of each cluster of filters. After identifying the importance of each cluster, we examine how much of that importance is attributable to spectra and to multispectral waveforms within each cluster. Unlike previous methods that only provided insight into key spectral features, we provide insight into both key spectral features and waveforms. **Figure 1** shows an overview of our methods.

MATERIALS AND METHODS

In this study, we use EEG sleep stage data to train a CNN. We then visualize the first layer filters of the CNN, convert them to the frequency domain, and cluster them. After clustering the filters, we ablate each cluster to determine their importance and perturb their frequency and time domain representations for insight into the importance of the spectra and waveforms within each cluster.

Dataset and Data Preprocessing

In this study, we used the Sleep Cassette subset of the PhysioNet (Goldberger et al., 2000). Sleep-EDF Database Expanded (PhysioNet, 2002). The Sleep Cassette subset contains 153 20-h recordings from 78 healthy individuals. Each individual had two subsequent recordings of day-night periods while at home. The dataset includes electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), oro-nasal airflow, and rectal body temperature. However, in our study, we just used EEG from the FPz-Cz electrode recorded at 100 Hertz (Hz). The data was assigned by experts to Awake, REM, NREM1, NREM2, NREM3, and NREM4 stages in 30-s intervals.

We segmented the data into 30-s samples based on the expert assigned intervals. To alleviate data imbalances, we removed Awake data from the start of the recordings and part of the end of the recordings. Using clinical guidelines, we made NREM3 and NREM4 a single class. After removing samples, we separately z-scored the EEG data from each recording. **Table 1** shows the resulting distribution of samples in each class.

Model Development

Here, we discuss our model development and evaluation approach. We implemented our model architecture and training in Keras 2.2.4 (Chollet, 2015) and Tensorflow 1.15.0 (Abadi et al., 2016).

Architecture

We utilized a 1D-CNN architecture that had long filters in the first convolutional layer to make it easier to apply explainability methods and gain insight into extracted waveforms and frequency bands. The architecture that we used was originally developed in Tsinalis et al. (2016b) for 150-s segments that frequently included multiple sleep stages. We adapted it to make it compatible with our shorter 30-s segments that only included a single sleep stage. This adaptation made it easier to compare our explainability results to domain knowledge on the key characteristics of each sleep stage. **Figure 2** shows our classifier architecture.

Cross-Validation and Training Approach

When developing our architecture and training our classifier, we used a 10-fold cross-validation approach. In each fold, we randomly assigned 63, 7, and 8 subjects to training, validation, and test groups, respectively. To address class imbalances, we weighted our categorical cross entropy loss function. We also used a stochastic gradient descent optimizer with a batch size

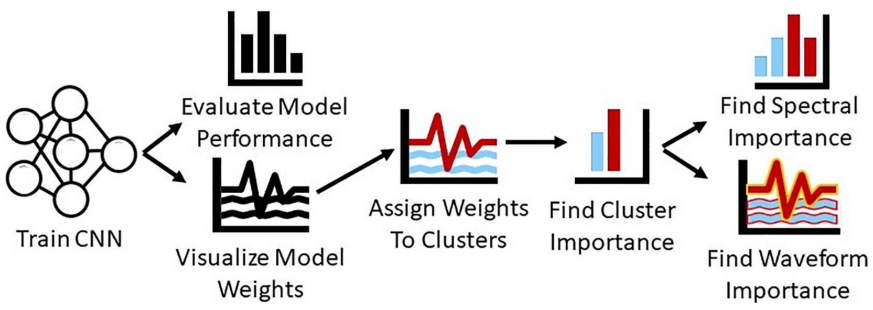

FIGURE 1 | Overview of methods. We train a CNN for sleep stage classification and evaluate its performance. We then visualize the first layer filters of the CNN, convert them to the frequency domain, and cluster them. After clustering the filters, we ablate each cluster to determine its importance and perturb their frequency and time domain representations to determine how much of cluster importance is attributable to the spectra and waveforms within each cluster.

of 100 and an adaptive learning rate with an initial value of 0.015 that decreased by 10% after each set of five epochs that did not have a corresponding increase in validation accuracy. We applied early stopping if 20 epochs occurred without an increase in validation accuracy. The maximum number of training epochs was 30. Additionally, we shuffled the training data between each epoch.

Performance Evaluation

When evaluating the test performance of our model, we computed the precision, recall, and F1-score for each class in each fold. After computing the metrics for each fold, we calculated the mean and standard deviation of each metric across folds. The equations below show the formulas for precision, recall, and F1-score.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Explainability – Identifying Clusters of Filters

To better understand the global extraction of features by the model, we selected the model from the fold with the highest weighted F1-score on the test data. We then used a Fast Fourier Transform (FFT) to convert the 30 filters in the first convolutional layer into the frequency domain. Next, we calculated the spectral power between 0 and 50 Hz. We then performed two rounds of k-means clustering using scikit-learn (Pedregosa et al., 2011). In the first round of clustering, we used 50 initialization and applied the silhouette method to determine the optimal number of clusters. After determining the optimal number of clusters, we redid the clustering with 100 initializations with the optimal number of clusters and examined the spectra of the filters within each cluster.

Explainability – Examining Importance of Each Cluster of Filters

After identifying clusters of filters, we sought to understand the relative importance of each of the clusters. We applied two methods to this end: ablation and layer-wise relevance propagation (LRP). In our ablation approach, we replaced each cluster of filters with zeros and measured the percent change in the weighted F1-score and class-specific F1-scores following ablation. A large negative percent change in performance after ablation corresponds to increase cluster importance.

We implemented LRP using the Innvestigate toolbox (Alber et al., 2019). LRP (Bach et al., 2015) is a popular gradient-based feature attribution method (Ancona et al., 2018). Rather than examining the effect of perturbing the model, it utilizes the gradients and activations of the network to estimate relevance (i.e., importance). LRP can output both positive and negative relevance. Positive relevance indicates that particular features provide evidence for a sample being assigned to the class it is ultimately assigned to by the classifier. In contrast, negative relevance indicates features that provide evidence for a sample being assigned to classes other than what it is ultimately assigned to by the classifier. In this study, we used the αβ relevance rule with an α of 1 and a β of 0 to filter out all negative relevance and only propagate positive relevance. The equation below shows the αβ-rule.

$$R_j = \sum_k \left(\alpha \frac{\left(a_j w_{jk}\right)^+}{\sum_{0,j} \left(a_j w_{jk}\right)^+} - \beta \frac{\left(a_j w_{jk}\right)^-}{\sum_{0,j} \left(a_j w_{jk}\right)^-} \right) R_k$$

Where the subscript k corresponds to a value for one of K nodes in a deeper layer and j corresponds to a value for one of J nodes in a shallower layer. The activation output by the shallower layer is referred to as a_j, and the model weights are referred to by w. The relevance is split into positive and negative portions when propagated backwards. The variables α and β control how much positive and negative relevance are propagated backwards, respectively.

Layer-wise relevance propagation typically gives relevance values for individual features, but we wanted to gain insight into filter importance. As such, we computed the relevance for the first layer convolutional layer activations of all test samples, computed the percentage of relevance assigned to each cluster for each subject, summed the relevance values within each predicted class and cluster, and then scaled the relevance values of each respective cluster by the total number of filters minus the number of filters in each cluster divided by the total number of filters. After understanding the relative importance of each cluster of filters, we sought to understand the important components of each cluster of filters. The equation below shows how we computed the importance for each class and cluster.

$$NR_C = \sum_s \frac{\sum_{f,c,s} R_{fcs}}{\sum_{f,s} R_{fs}} * \frac{F - F_c}{F}$$

Where NR_C indicates the normalized relevance within a particular cluster (C) and class, R indicates relevance, f indicates a filter, c indicates a cluster, s indicates a subject within cluster C, F indicates the total number of filters, and F_C indicates the total number of filters in a cluster.

Explainability – Examining Why Filter Clusters Are Important Spectrally

While visualizing the frequency domain of each cluster of filters provided insight into the frequencies learned by the classifier,

TABLE 1 | Distribution of samples.

	Awake	NREM1	NREM2	NREM3	REM	Total
Number	85,034	21,522	69,132	13,039	25,835	214,562
Percent	39.63	10.03	32.22	06.08	12.04	100

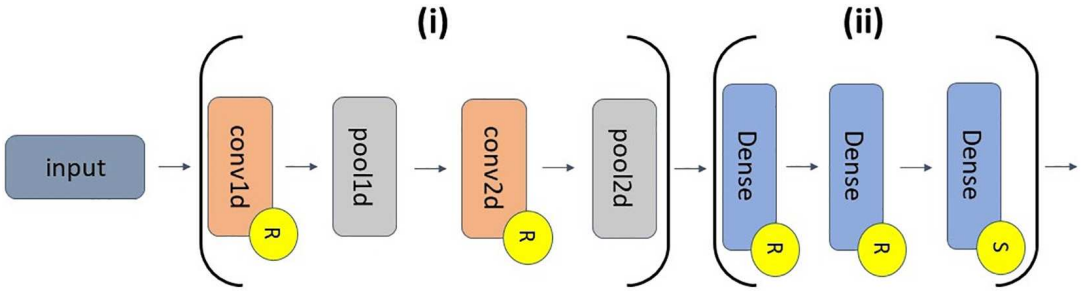

FIGURE 2 | CNN architecture. The feature extraction and classifier segments of the classifier are indicated by Sections (i) and (ii), respectively. (i) has of a 1D convolutional (conv1d) layer (30 filters, length of 200, stride of 1), a 1D max pooling (pool1d) layer (pooling size of 15, stride of 10), a conv2d layer (400 filters, 30 × 25 filter size, stride of 1 × 1), and a pool2d layer (pooling size of 10 × 1, stride of 1 × 2). (ii) has 2 dense layers (500 nodes) and an output dense layer (5 nodes). Layers with an "R" or an "S" have ReLU or softmax activation functions, respectively.

they did not necessarily indicate the relative importance of each of the frequency bands within each cluster of filters. As such, we performed a separate analysis to understand the relative importance of each frequency band within each cluster of filters. To this end, we (Ellis et al., 2021g) iteratively converted each cluster of filters to the frequency domain using an FFT, (Chen et al., 2019) replaced coefficients of a particular frequency band with zeros, (Kwon et al., 2018) converted the perturbed coefficients back to the time domain, (Ince et al., 2008) used the perturbed model to generate predictions for the test data, and (Ruffini et al., 2019) examined the change in F1-scores following perturbation. When perturbing the frequency domain of the filters, we perturbed five frequency bands that were in units of Hertz (Hz) – δ (0 – 4 Hz), θ (4 – 8 Hz), α (8 – 12 Hz), β (12 – 25 Hz), and γ (25 – 50 Hz).

Explainability – Examining Why Clusters Are Important Temporally

To understand the relative importance of the waveforms within each of the filters, we developed a novel weight perturbation approach including the following steps: (Ellis et al., 2021g). We calculated the class-specific F1-scores and weighted F1-score of the model on the test data (Chen et al., 2019). We used a sliding window approach to ablate (i.e., replace with values of zero) some of the first-layer filters weights of the model. Note that each filter in the first layer had a length of 200 weights (Kwon et al., 2018). We used the model with ablated weights to predict labels for the test data and calculated the percent change in F1-scores (Ince et al., 2008). We designated the percent change in class-specific F1-score as the importance of the point at the center of the window (Ruffini et al., 2019). We restored the model weights to their original pre-ablation values (Bach et al., 2015). We moved the sliding window along the filter with a specific step size and repeated steps 2 through 5 until all weights in the filter had corresponding importance values (Selvaraju et al., 2020). We repeated steps 2 through 6 for the next filter until importance values were obtained for weights in all filters. The sliding window had a window length of 25 points and a step size of one point. To ensure that each individual weight had an assigned importance value, we zero-padded the filters prior to the sliding window ablation process.

RESULTS

In this section, we describe the performance of our model and the results of each of the explainability analyses that we performed.

Model Performance

Table 2 shows our model performance results. The performance of our model was reasonable overall. The classifier generally had highest performance for the Awake class, followed by performance for NREM2. Additionally, performance for Awake and NREM2 had a low standard deviation and was generally consistent across folds. Performance for NREM3 and was comparable to REM. NREM3 had a noticeably higher mean recall than REM and a slightly higher mean F1-score than REM. REM had a markedly higher mean precision than NREM3. However, REM had much lower levels in variation of performance across folds than NREM3. Interestingly, NREM1 performance was above chance level but still much lower than the performance for all other classes across metrics. NREM1 had lower levels of variation in precision and the F1-score across folds but higher levels of variation in recall. Our classification performance was somewhat lower than the performance of the classifier in Tsinalis et al. (2016b). This is likely due to our choice to use 30-s samples from one sleep stage rather than 150-s segments with multiple sleep stages that included state transitions. This choice aided our goal in evaluating the utility of our proposed explainability approach, as it made it easier to directly compare our explainability results with the well-characterized features of individual sleep stages.

Results for Clustering Spectra

We identified three clusters with our clustering approach. **Figure 3** shows the filters for each cluster in the time domain and in the frequency domain. Clusters 0, 1, and 2 were assigned 3, 16, and 11 of the 30 filters, respectively. Cluster 0 was the smallest cluster. It contained sinusoids of varying amplitude that were predominantly associated with the lower β band, although

one of the filters had some α activity. Cluster 1 was the largest cluster. Although it contained a mixture of frequencies, it was predominantly composed of upper β band activity and some upper θ and lower α. In contrast to the other clusters, Cluster 2 was primarily composed of filters extracting δ and lower θ-band activity. Interestingly, several filters did extract lower β-band activity. There were a number of dominant low frequency waveforms in Cluster 2 filters that were not purely sinusoidal (e.g., in filters 20, 21, 26, and 28).

Results for Cluster Importance

Figure 4 shows the overall importance of each cluster to the classifier using both LRP and ablation. With a few exceptions both methods yield similar results. The results are the same for NREM2, NREM3, and REM. They show that Cluster 2 is most important across the three classes. Additionally, across both methods Cluster 1 is second most important for REM and NREM3, while Cluster 0 is second most important for NREM2. For NREM1 and Awake, there are some key differences. Namely, LRP finds Cluster 2 followed by Cluster 1 to be most important. In contrast, ablation finds Cluster 2 followed by Cluster 1 to be most important. For both methods, Cluster 0 is of low to moderate importance for all classes except NREM2. Additionally, the weighted F1-score of the ablation indicates that Clusters 1 and 2, followed by Cluster 0, are most important.

Results for Cluster-Specific Spectral Perturbation

Our previous analyses characterized and identified the relative importance of each cluster of filters. To examine the importance of different spectra within each cluster, we perturbed the frequency bands in each cluster and examined their effect on classifier performance. **Figure 5** shows the results for our spectral perturbation analysis. β was the most important band within Cluster 0. While there were small levels of δ and α activity (as shown in **Figure 3**) in the cluster, the bands had little to no importance. Similar to in Cluster 0, β was the most important band in Cluster 1, having strong effects on NREM1 and minor effects on NREM3 and REM. Importantly, δ and θ were most important in Cluster 2. Perturbation of δ had strong effects upon all classes except Awake, and perturbation of θ had a strong effect on REM, with low to moderate effects on NREM1 and NREM2. Interestingly, the Awake class was not strongly affected by the perturbation of frequency bands in any of the classes. The strongest effects of spectral perturbation across clusters and bands were those of Cluster 2 δ and θ upon REM.

Results for Temporal Filter Ablation

After examining the key frequencies of each cluster of filters, we examined the waveforms extracted by the filters. **Figure 6** shows the results of our temporal ablation analysis. The change in weighted F1-score identified the overall impact of the temporal ablation upon the classifier performance. Ablation of the Cluster 1 windows had a slight impact upon classifier performance across most filters and windows, while ablation of Cluster 0 windows had no noticeable effect. Ablation of Cluster 2 windows resulted in large levels of localized importance in parts of filters 21, 26, and 29.

Similar waveforms showed impacts upon class-specific F1-scores. However, there were also some variations. Note that references to timepoints in this paragraph refer to the first-layer filters which are 2 s long (i.e., 200 weights per filter, with a data sampling rate of 100 Hz) and correspond to the x-axis of **Figure 6**. Interestingly, timepoints 0.8 to 1.1 s of filter 20 were of modest importance for NREM1 and REM but not for other classes. Timepoints from 0 to 0.5 s of filter 21 in Cluster 2 had high importance across all classes, with highest importance for NREM1, NREM2, and REM and noticeably less importance for Awake and NREM3. Timepoints 1.00 to 1.25 s of filter 26 had little importance for Awake but moderate (i.e., NREM1, NREM2, NREM3) to high (i.e., REM) levels of importance for other classes. Timepoints 1.0 to 1.25 s of filter 28 were important for NREM1, NREM2, and REM. Timepoints 0.25 to 0.5 s of filter 29 had moderate levels of importance for NREM2, NREM3, and REM. Interestingly, timepoints 0.5 to 1.0 s of filter 17 were important for NREM2 and NREM3, with other parts of the filter also being important for NREM3. Part of filter 12 were also noticeably important for NREM1.

DISCUSSION

In this section, we discuss how the results from our analyses relate to one another to explain the model more effectively and how those results compare to existing knowledge from the sleep domain. Unless otherwise specified, when we discuss the results within the context of domain knowledge, we are comparing them to the AASM Manual (Iber et al., 2007).

Developing a High Performing Classifier

Classifier performance was highest for the Awake class, which makes sense given that the Awake class had the largest number of samples and that it has clear differences from the other classes. Although NREM1 performance was low, that is acceptable. Most sleep stage classification studies have difficulty classifying NREM1 effectively (Tsinalis et al., 2016a; Supratak et al., 2017;

TABLE 2 | Model test performance.

	Awake	NREM1	NREM2	NREM3	REM
Precision	94.90 ± 02.67	38.47 ± 04.85	79.83 ± 04.56	63.02 ± 12.25	67.27 ± 08.53
Recall	89.12 ± 02.44	41.10 ± 10.46	79.12 ± 05.77	75.55 ± 15.04	68.65 ± 06.44
F1	91.88 ± 01.82	38.56 ± 04.37	79.20 ± 01.67	67.89 ± 11.09	67.41 ± 04.80

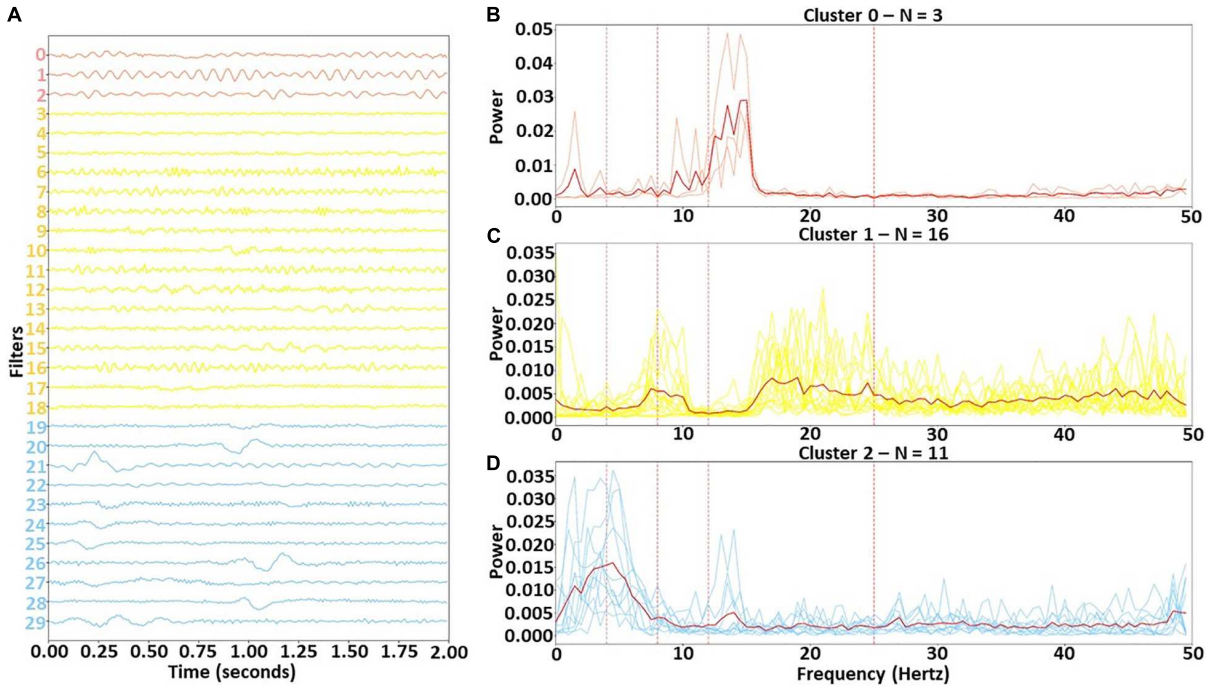

FIGURE 3 | Identified clusters in time and frequency domain. Panel **(A)** shows the time domain of the identified clusters, while panels **(B–D)** show the frequency domain of the filters in Clusters 0, 1, and 2, respectively. In panel **(A)**, the filters are arranged vertically based on cluster. The solid red lines in panels **(B)** through **(D)** are the cluster centers, and the vertical dashed lines show boundaries of each frequency band. Light red, yellow, and light blue lines represent Clusters 0, 1, and 2, respectively.

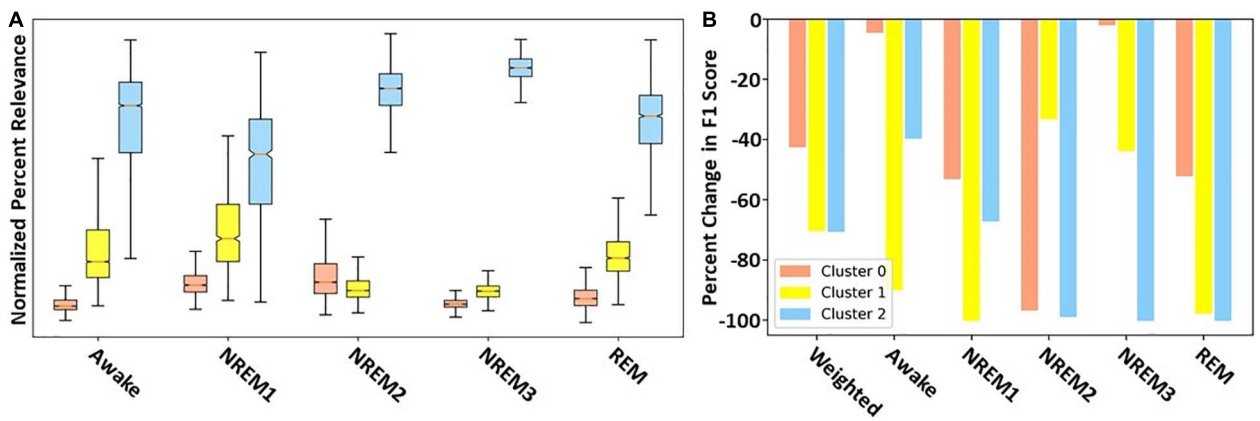

FIGURE 4 | Filter cluster importance with LRP and ablation. Panels **(A,B)** show the global importance of each cluster of filters using LRP and ablation, respectively. Importantly, red, yellow, and blue values show importance of Clusters 0, 1, and 2, respectively.

Chambon et al., 2018; Michielli et al., 2019), possibly because it is one of the smaller sample groups and can appear similar to Awake and REM (Quan et al., 1997; Goldberger et al., 2000; PhysioNet, 2002; Iber et al., 2007; Pedregosa et al., 2011; Chollet, 2015; Abadi et al., 2016; Khalighi et al., 2016; Tsinalis et al., 2016a; Ancona et al., 2018; Alber et al., 2019; Michielli et al., 2019; Ellis et al., 2021c). Some studies have gone so far as to develop hierarchical models specifically designed to improve performance for NREM1 classification (Michielli et al., 2019). After Awake, the performance of the classifier for NREM2, NREM3, and REM seem to be loosely related to the number of samples in each class.

Identifying Filter Clusters With Distinct Spectral Features and Quantifying Their Relative Importance

In our first analysis, we visualized the first layer filters of the model from the fold with the highest F1-score. We then

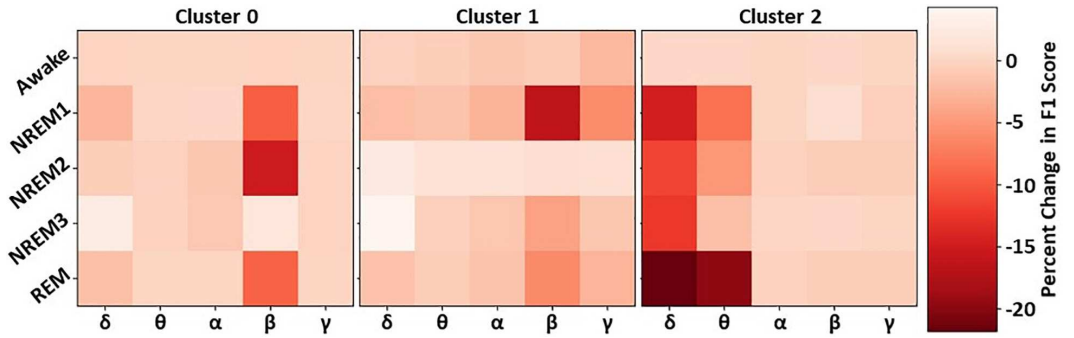

FIGURE 5 | Effect of cluster-level spectral perturbation on each class. The leftmost, middle, and rightmost panels show the percent change in class-specific F1-score that resulted from perturbing each frequency band in Clusters 0, 1, and 2, respectively.

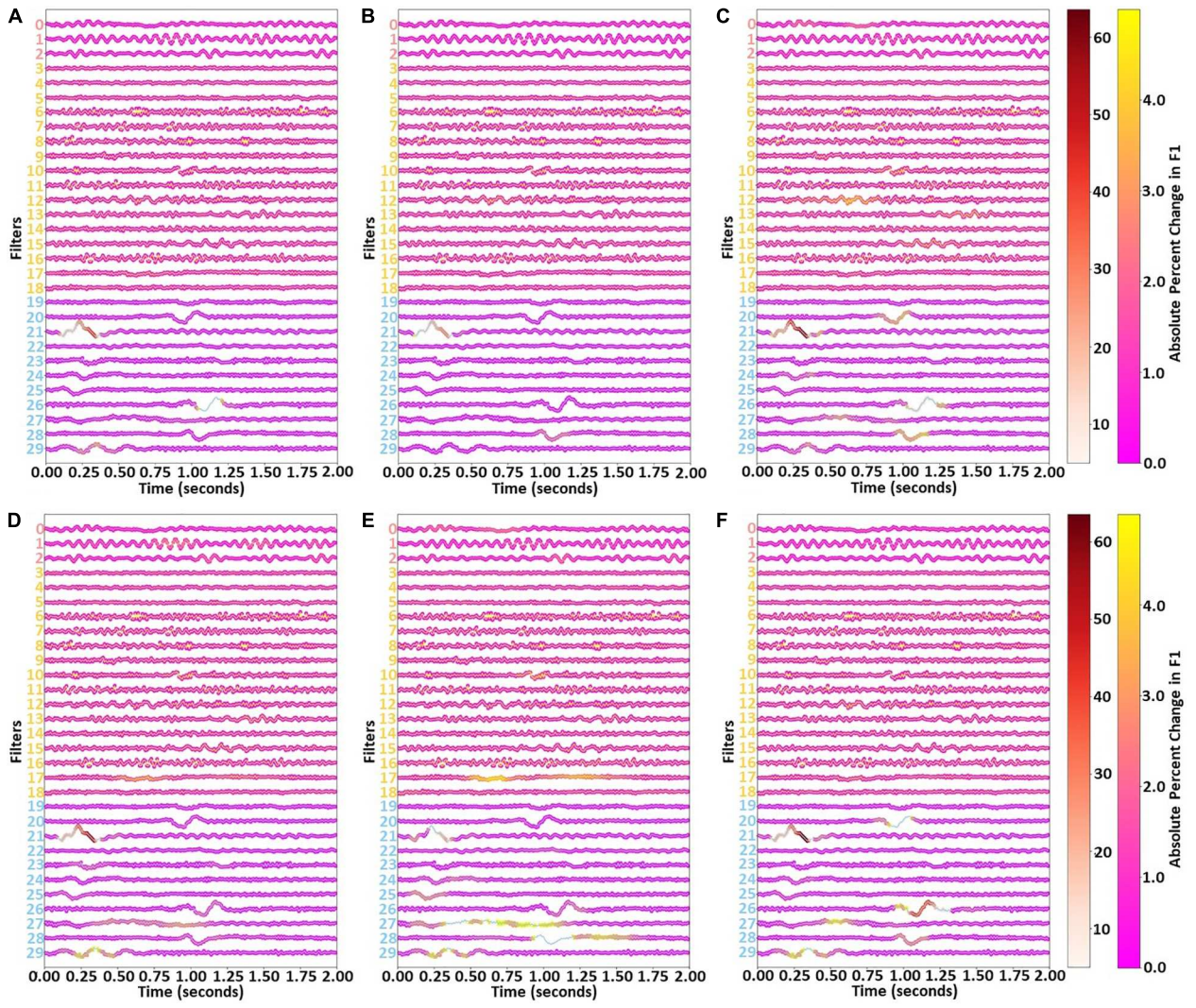

FIGURE 6 | Effects of filter temporal ablation. Panels **(A–F)** show the effect of ablating each set of 25 filter weights upon the weighted F1-score and Awake, NREM1, NREM2, NREM3, and REM class-specific F1-scores, respectively. Red, yellow, and blue lines show weights for Clusters 0, 1, and 2, respectively. Markers at each time point and the two colormaps indicate the effect of ablating the surrounding 24 time points upon the F1-score. There are two colormaps to enable better visualization of importance for both low and high levels of importance.

clustered the filters in the frequency domain to identify sets of spectrally distinct filters. Interestingly, all canonical frequency bands were present in at least one of the filter clusters. Cluster 0 contained large amounts of lower β activity and is highly important for identifying NREM2. This makes sense given that NREM2 often contains sleep spindles that appear in the lower β band (i.e., 12 – 14 Hz). Cluster 0 was also important for identifying NREM1 and REM and was overall least important for identifying classes other than NREM2. Interestingly, previous studies have shown increased levels of β activity during REM (Vijayan et al., 2017). Cluster 1 was characterized as having some lower frequency activity (upper θ and lower α) but predominantly higher frequency activity (upper β and γ). It was very important for identifying Awake, NREM1, and REM samples. The often erratic and high frequency activity of Awake could explain why the cluster was important for identifying the Awake class. Incidentally, NREM1 is often characterized by low amplitude, mixed frequency activity within the θ range, which could indicate why the cluster was important for NREM1. Moreover, as REM is often characterized as multispectral, it makes sense that a cluster having multiple distinct frequency bands would be important for identifying REM. Cluster 2 was characterized by low frequency δ and θ activity and was very important for identifying NREM2, NREM3, and REM, with moderate importance for NREM1. Its importance for NREM1 could be attributed to its extraction of low amplitude θ activity. NREM2 often includes K-complexes that appear within the δ band. Given that Cluster 2 extracts δ activity and extracts waveforms in a number of filters (i.e., 20, 23, 24, 25, 26, and 28) that resemble k-complexes, it is very reasonable that Cluster 2 would be important for identifying NREM2. Importantly, the main feature of NREM3 is δ activity, which could explain why Cluster 2 is so important for NREM3. Lastly, REM has previously been associated with high levels of frontal θ activity (Vijayan et al., 2017).

Confirming Spectral Importance of Clusters

Based on our initial identification of clusters of filters and the relative importance of the clusters for each class, we suggested a number of reasons why the filters might be important to particular classes. However, our previous cluster ablation analysis did not provide definitive evidence regarding the importance of the spectral features present in each cluster. Here, we examine the effects of perturbing the canonical frequency bands within each cluster and examine their relative impact upon classifier performance. Overall, the effect of perturbing the individual bands within each cluster does not sum up to the effect of ablating each cluster, which could indicate that all of the useful information in the filters was not found in the spectral features extracted. For example, the perturbation of frequency bands across all clusters did not affect performance for Awake, which could indicate waveforms, rather than frequency bands, were important for identifying Awake.

Our previous cluster importance analysis indicated that Cluster 0 was highly important for NREM2, with moderate importance for NREM1 and REM, and little to no importance for Awake and NREM3. Our spectral perturbation analysis confirmed that β was the only important frequency band in Cluster 0. Additionally, perturbing β in Cluster 0 had a larger impact upon NREM2 than the perturbation of any other band in Clusters 1 and 2. REM is often characterized as having high β activity (Vijayan et al., 2017), which could explain the importance of Cluster 0 β upon REM, but NREM1 is not typically associated with β, which could indicate that the classifier incorrectly associated β with NREM1 and could explain the poor performance of the classifier for the sleep stage. The effect of perturbing β in Cluster 1 upon NREM1 was the largest effect of any pair of classes and bands within Cluster 1. Interestingly, perturbation of individual frequency bands in Cluster 1 seemed to have very little impact upon Awake and REM. Given the importance of the cluster for identifying the stages, that could indicate that the classifier did not rely solely upon extracted spectral features or upon any single frequency band within the cluster when identifying Awake and REM. Instead, the classifier might rely more upon extracted waveforms. The perturbation of δ and θ in Cluster 2 was particularly impactful across multiple classes, particularly the REM class that is characterized as having high levels of θ activity (Vijayan et al., 2017). Similar to our previous hypotheses and consistent with clinical guidelines, perturbing θ in Cluster 2 did impact NREM1 performance, and perturbing δ impacted NREM3 performance.

Examining Importance of Extracted Waveforms

Our analysis of the importance of the canonical frequency bands within each cluster did not fully explain the importance of each cluster to the individual sleep stages. As such, to more fully understand the importance of each cluster, we sought to examine the importance of individual waveforms within the filters to performance for each class. The importance of Cluster 0 to NREM2 was primarily explained by the extraction of β activity. However, by perturbing filters 1 (i.e., 0.75 to 1.25 s) and 2 (1.0 to 1.25 s) in Cluster 0, we can see some time windows where waveforms resembling sleep spindles seem to be of some importance. While perturbing individual frequency bands in Cluster 1 seemed to have little impact, the perturbation of individual waveforms in Cluster 1 seemed to generally have more of an impact than the perturbation of most time windows in other clusters, which could indicate that the temporal characteristics of the filters were more important than their spectral characteristics. This could also explain the importance of Cluster 1 for identifying Awake and NREM1. Lastly, as Cluster 2 had more low frequency activity, there were more clearly discernable waveforms of strong importance to the classifier performance. Perturbation of filter 21 (i.e., 0.00 to 0.50 s) was of high but varying importance across all classes. Low frequency oscillations in a several filters (i.e., 27 and 29) were important for identifying NREM2. Additionally, waveforms resembling k-complexes in multiple filters (i.e., 23, 25, 26, 28) were also of some importance for identifying NREM2. Low frequency activity was present across multiple filters. Like for NREM2, the classifier relied heavily upon δ waveforms

when identifying NREM3. Additionally, Filter 21, which has a waveform resembling a vertex sharp wave from (0.00 to 0.50 s), was particularly importance across nearly all classes, and was most important in NREM1, NREM2, and REM. The importance of vertex sharp waves in NREM1 could explain its importance for the class.

Limitations and Next Steps

One of the key contributions of this paper is the combination of a model architecture that is structured to enable increased interpretability with a systematic approach for examining the key spectral and temporal features learned by the classifier. While the filter size of our first layer enabled the visualization of the extracted filters, it also made the training and evaluation of the architecture very computationally intensive. Future iterations of this analysis approach could likely find sufficient levels of explanatory insight with filter lengths equivalent to 0.5 or 1.0 s of signal, while also having a model that could be trained and evaluated with more computational efficiency. Additionally, this study only used data from one electrode. The use of one electrode is common in sleep stage classification, but less so in other domains of EEG analysis. Future iterations of this approach could be generalized to multichannel data by perturbing filters when they are applied to individual channels but not to other channels. Lastly, we applied our approach to sleep stage classification so that we could evaluate its efficacy within a well-characterized domain. In the future, our approach might be applied for biomarker identification in domains that are poorly characterized. It would also be possible to examine the effect of perturbation upon the probability of individual samples belonging to a class or upon subject-specific performance metrics for the purpose of personalized biomarker identification.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.physionet.org/content/sleep-edfx/1.0.0/.

AUTHOR CONTRIBUTIONS

CE helped with conception of the manuscript, performed the analyses, and wrote and edited the manuscript. RM helped with the conception and editing of the manuscript. VC helped with the editing of the manuscript and provided funding for the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the NIH grant R01MH123610 and NSF grant 2112455.

ACKNOWLEDGMENTS

We thank those who collected the Sleep-EDF Database Expanded on PhysioNet.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, (Savannah, GA: USENIX Association).

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., et al. (2019). INNvestigate neural networks! *J. Mach. Learn. Res.* 20, 1–8.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). "Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks," in *International Conference on Learning Representations*, (Vancouver: ICLR).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., 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.0130140doi

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

Borra, D., Fantozzi, S., and Magosso, E. E. E. G. (2019). "Motor Execution Decoding via Interpretable Sinc-Convolutional Neural Networks," in *Mediterranean Conference on Medical and Biological Engineering and Computing [Internet]*, (New York: International Publishing), 1515–1525. doi: 10.1007/978-3-030-31635-8_188

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., and Gramfort, A. A. (2018). deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 758–769. doi: 10.1109/TNSRE.2018.2813138

Chen, H., Song, Y., and Li, X. (2019). Use of deep learning to detect personalized spatial-frequency abnormalities in EEGs of children with ADHD. *J. Neural Eng.* 19:16. doi: 10.1088/1741-2552/ab3a0a

Chollet, F. (2015). *Keras*. San Francisco: GitHub.

Ellis, C. A., Carbajal, D. A., Zhang, R., Miller, R. L., Calhoun, V. D., and Wang, M. D. (2021a). An Explainable Deep Learning Approach for Multimodal Electrophysiology Classification. *bioRxiv* [Preprint]. doi: 10.1101/2021.05.12.443594

Ellis, C. A., Carbajal, D. A., Zhang, R., Sendi, M. S. E., Miller, R. L., Calhoun, V. D., et al. (2021b). "A Novel Local Ablation Approach For Explaining Multimodal Classifiers," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering*, (Kragujevac: IEEE).

Ellis, C. A., Miller, R. L., and Calhoun, V. D. A. (2021e). "Novel Local Explainability Approach for Spectral Insight into Raw EEG-Based Deep Learning Classifiers," in *21st IEEE International Conference on BioInformatics and BioEngineering*, (Serbia: IEEE).

Ellis, C. A., Miller, R. L., and Calhoun, V. D. A. (2021c). Gradient-based Spectral Explainability Method for EEG Deep Learning Classifiers. *bioRxiv*. [Preprint]. doi: 10.1101/2021.07.14.452360

Ellis, C. A., Miller, R. L., and Calhoun, V. D. A. (2021d). Model Visualization-based Approach for Insight into Waveforms and Spectra Learned by CNNs. *bioRxiv* [Preprint]. doi: 10.1101/2021.12.16.473028

Ellis, C. A., Miller, R. L., Calhoun, V. D., and Wang, M. D. A. (2021f). "Gradient-based Approach for Explaining Multimodal Deep Learning Classifiers," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering*, (kragujevac: IEEE).

Ellis, C. A., Sendi, M. S. E., Miller, R., and Calhoun, V. A. (2021g). "Novel Activation Maximization-based Approach for Insight into Electrophysiology

Classifiers," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Housto: IEEE).

Ellis, C. A., Sendi, M. S., Willie, J. T., and Mahmoudi, B. (2021h). "Hierarchical Neural Network with Layer-wise Relevance Propagation for Interpretable Multiclass Neural State Classification," in *10th International IEEE/EMBS Conference on Neural Engineering*, (Italy: IEEE), 18–21.

Ellis, C. A., Sendi, M. S. E., Zhang, R., Carbajal, D. A., Wang, M. D., Miller, L., et al. (2022). Novel Methods for Elucidating Modality Importance in Multimodal Electrophysiology Classifiers. *bioRxiv* [preprint]. doi: 10.1101/2022.01.01.474276

Frick, T., Glüge, S., Rahimi, A., Benini, L., and Brunschwiler, T. (2021). "Explainable Deep Learning for Medical Time Series Data," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, (Germany: Springer), 244–256.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank. PhysioToolkit, and PhysioNet: components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101, e215–e220. doi: 10.1161/01.cir.101.23.e215

Iber, C., Ancoli-Israel, S., Chesson, A. L., and Quan, S. F. (2007). *The AASM Manual for Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications*. Westchester, IL: American Academy of Sleep Medicine.

Ince, N., Goksu, F., Pellizzer, G., Tewfik, A., and Stephane, M. (2008). "Selection of spectro-temporal patterns in multichannel MEG with support vector machines for schizophrenia classification," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (Vancouver: IEEE), 3554–3557. doi: 10.1109/IEMBS.2008.4649973

Khalighi, S., Sousa, T., Santos, J. M., and Nunes, U. (2016). ISRUC-Sleep: a comprehensive public dataset for sleep researchers. *Comput. Methods Progr. Biomed.* 124, 180–192. doi: 10.1016/j.cmpb.2015.10.013

Kwon, Y. H., Shin, S. B., and Kim, S. D. (2018). Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system. *Sensors* 18:1383. doi: 10.3390/s18051383

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

Michielli, N., Acharya, U. R., and Molinari, F. (2019). Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Comput. Biol. Med.* 106, 71–81. doi: 10.1016/j.compbiomed.2019.01.013

Nahmias, D. O., and Kontson, K. L. (2020). "Easy Perturbation EEG Algorithm for Spectral Importance (easyPEASI): A Simple Method to Identify Important Spectral Features of EEG in Deep Learning Models," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (New York: ACM), 2398–2406. doi: 10.1145/3394486.3403289

Pathak, S., Lu, C., Nagaraj, S. B., van Putten, M., and Seifert, C. S. T. Q. S. (2021). Interpretable multi-modal Spatial-Temporal-seQuential model for automatic Sleep scoring. *Artif. Intell. Med.* 114:102038. doi: 10.1016/j.artmed.2021.102038

Pedregosa, F., Weiss, R., and Brucher, M. (2011). Scikit-learn: machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.1080/13696998.2019.1666854

PhysioNet (2002). *The Sleep-EDF database [Expanded]*. New York: IEEE.

Porumb, M., Stranges, S., Pescapè, A., and Pecchia, L. (2020). Precision Medicine and Artificial Intelligence: a Pilot Study on Deep Learning for Hypoglycemic Events Detection based on ECG. *Sci. Rep.* 10, 1–16. doi: 10.1038/s41598-019-56927-5

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., et al. (1997). The Sleep Heart Health Study: design, rationale, and methods. *Sleep* 20, 1077–1085.

Ruffini, G., Ibañez, D., Castellano, M., Dubreuil-Vall, L., Soria-Frisch, A., Postuma, R., et al. (2019). Deep Learning With EEG Spectrograms in Rapid Eye Movement Behavior Disorder. *Front. Neurol.* 10:806. doi: 10.3389/fneur.2019.00806

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* 128, 336–359.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep Inside Convolutional Networks: visualising Image Classification Models and Saliency Maps. *arXiv* [preprint]. doi: 10.48550/arXiv.1312.6034

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K. R. (2016). Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Methods* 274, 141–145. doi: 10.1016/j.jneumeth.2016.10.008

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1998–2008. doi: 10.1109/TNSRE.2017.2721116

Tsinalis, O., Matthews, P. M., and Guo, Y. (2016a). Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. *Ann. Biomed. Eng.* 44, 1587–1597. doi: 10.1007/s10439-015-1444-y

Tsinalis, O., Matthews, P. M., Guo, Y., and Zafeiriou, S. (2016b). Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. *arXiv* [preprint]. doi: 10.48550/arXiv.1610.01683

Vijayan, S., Lepage, K. Q., Kopell, N. J., and Cash, S. S. (2017). Frontal beta-theta network during REM sleep. *Elife* 6:e18894. doi: 10.7554/eLife.18894

Yoshimura, N., Maekawa, T., and Hara, T. (2019). Preliminary Investigation of Visualizing Human Activity Recognition Neural Network. 2019 12th Int Conf Mob Comput Ubiquitous Network. *ICMU* 2019, 4–5.

Yoshimura, N., Maekawa, T., and Hara, T. (2021). "Toward Understanding Acceleration-based Activity Recognition Neural Networks with Activation Maximization," in *2021 International Joint Conference on Neural Networks*, (New York: IEEE).

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.