# A Framework for Systematically Evaluating the Representations Learned by A Deep Learning Classifier from Raw Multi-Channel Electroencephalogram Data

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

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

*** Correspondence:**
Corresponding Author
cae67@gatech.edu

**Keywords: explainable AI, deep learning, electroencephalography, major depressive disorder, interaction explainability, spectral explainability**

## ABSTRACT

The application of deep learning methods to raw electroencephalogram (EEG) data is growing increasingly common. While these methods offer the possibility of improved performance relative to other approaches applied to manually engineered features, they also present the problem of reduced explainability. As such, a number of studies have sought to provide explainability methods uniquely adapted to the domain of deep learning-based raw EEG classification. In this study, we present a taxonomy of those methods, identifying existing approaches that provide insight into spatial, spectral, and temporal features. We then present a novel framework consisting of a series of explainability approaches for insight into classifiers trained on raw EEG data. Our framework provides spatial, spectral, and temporal explanations similar to existing approaches. However, it also, to the best of our knowledge, proposes the first explainability approaches for insight into spatial and spatio-spectral interactions in EEG. This is particularly important given the frequent use and well-characterized importance of EEG connectivity measures for neurological and neuropsychiatric disorder analysis. We demonstrate our proposed framework within the context of automated major depressive disorder (MDD) diagnosis, training a high performing one-dimensional convolutional neural network with a robust cross-validation approach on a publicly available dataset. We identify interactions between central electrodes and other electrodes and identify differences in frontal θ, β, and $\gamma_{low}$ between healthy controls and individuals with MDD. Our study represents a significant step forward for the field of deep learning-based raw EEG classification, providing new capabilities in interaction explainability and providing direction for future innovations through our proposed taxonomy.

# INTRODUCTION

In recent years, studies have increasingly applied deep learning approaches to raw electroencephalography (EEG) data. Relative to studies using traditional machine learning and deep learning methods with extracted features, deep learning studies using raw EEG allow for automated feature learning and the discovery of EEG features that might ordinarily be overlooked. This benefit has the potential to enhance model performance. Nevertheless, the use of raw EEG also occasions an important shortcoming. Namely, deep learning models with raw EEG are not as explainable as traditional machine learning [1] or deep learning models [2], [3] applied to extracted features. This has resulted in the development of a subfield of EEG analysis seeking to make deep learning models with raw EEG more explainable. To a large extent, these studies have succeeded. Under certain circumstances, EEG explainability approaches can provide insight into key channels [4]–[11], frequency bands [4], [5], [8], [12]–[18], and waveforms [5], [8], [14], [16], [17]. However, existing methods do not provide insight into interactions between channels. In this study, we present a taxonomy of deep learning-based raw EEG explainability approaches, identifying critical gaps in the capabilities of the field. We then present a series of explainability approaches that form a framework for systematically evaluating what a deep learning model has learned from raw EEG. Specifically, we train a high-performing one-dimensional convolutional neural network (1D-CNN) with a robust cross-validation approach to differentiate between healthy individuals and individuals with clinically diagnosed major depressive disorder (MDD) on multichannel EEG data. We present approaches to (1) identify the relative importance of each channel, (2) identify interactions uncovered by the model between channels, (3) identify key frequency bands in each channel, (4) identify interactions between frequency bands in each channel and other channels, and (5) identify representative samples of each class and the waveforms of importance to their classification. Our identification of spatio-spectral interactions is to our knowledge the first implementation of such a method in raw EEG-based deep learning explainability. Moreover, our study represents a significant step forward for the domain of raw EEG-based deep learning explainability and has the potential to stimulate future advances in the field.

## Modalities Used for Analysis of Neurological and Neuropsychiatric Disorders and Advantage of EEG

Multiple modalities have been used to study neurological and neuropsychiatric disorders. A few of these modalities include EEG [1]–[3], [9], [10], [19]–[27], magnetoencephalography (MEG) [28]–[30], and functional magnetic resonance imaging (fMRI) [30]–[37]. Each modality offers both advantages and disadvantages. For example, fMRI has enhanced spatial resolution relative to EEG and MEG. However, EEG and MEG have significantly improved temporal resolution relative to fMRI, which can afford better insight into the effects of disorders upon brain dynamics. Additionally, relative to MEG, EEG devices can be performed much more cheaply and are more widespread, making it better suited for deployment in a clinical setting. Within the domain of EEG analysis, both task [8], [11], [38] and resting-state [5], [6], [9], [10], [25], [39], [40] analyses are commonly performed. However, most brain activity spontaneously occurs (i.e., reflects brain networks that are unmodulated by any task), so resting-state activity better reflects the activity that is common to an individual with a disorder. Additionally, individuals with a disorder may perform a task less effectively than healthy individuals, which could introduce a confounder in any subsequent analyses [41]. As such, in this study, we focus on explainability for resting-state EEG.

## Features Commonly Included in EEG Analyses

Historically, many features have been extracted from EEG for insight into neurological and neuropsychiatric disorders. These include single-channel features like spectral power [1]–[3], [18], [25], [28], [42]–[45] and multi-channel features like temporal and spectral connectivity [26], [27], [46]–[49]. Spectral power has been associated with disorders like schizophrenia [43], attention deficit hyperactivity disorder (ADHD) [43], obsessive compulsive disorder (OCD) [43], Parkinson's disease, and major depressive disorder [50]. Connectivity features have shown high discriminative power and been associated with a many disorders including schizophrenia [49], MDD [27], and Alzheimer's disease [51]. Importantly, previous studies of MDD have identified effects upon connectivity between all frequency bands [27] and between frontal electrodes, between temporal electrodes, and between temporal and central electrodes [26].

### Transition from Manually Engineered Features to Raw EEG Data

Building upon these features, many studies have trained machine learning [25], [28], [42] and deep learning [2], [3], [18], [19], [39], [45], [52] models on spectral power features. Additionally, a few studies have trained machine learning [47] and deep learning models [48] on extracted connectivity features. These studies have obtained high levels of model performance while simultaneously offering high levels of explainability. They have obtained high levels of explainability largely because many methods have been previously developed to explain traditional machine learning models [53]–[55] and many methods like saliency [56], gradient-weighted class activation mapping (Grad-CAM) [57], and layer-wise relevance propagation (LRP) [58] have been developed within the domain of image classification to explain deep learning models. However, models using extracted features have an inherit limitation. Namely, they restrict the space of features over which models can learn. As such, over time, as the field of deep learning has further developed, an increasing number of studies have begun training deep learning models on raw EEG data [4]–[14], [16]–[18]. Deep learning models employ an automated feature extraction approach that precludes the need for manually engineered features. As such, deep learning models are theoretically able to learn from the entire feature space when applied to raw EEG data. Unfortunately, they also have reduced explainability due to the high dimensionality of the input data.

### Explainability in Models with Manually Engineered Versus Automatically Learned Features

Deep learning models applied to raw EEG are not less explainable because existing explainability methods cannot be applied in the context of EEG. Rather, deep learning models applied to raw EEG are less explainable because the temporal nature of EEG data presents unique problems relative to tabular and image data. Traditional explainability methods cannot be directly translated to provide insight into key frequency bands because the input to models is a time-series. Traditional methods [59] cannot be directly translated to identify key waveforms because to extract useful global insight thousands or hundreds of thousands of samples might need to be analyzed. Traditional methods [60] that account for interactions are also difficult to translate directly to EEG because of the large number of features per sample of EEG data (e.g., a sample may have 19 to 60 channels and be thousands of time points long). As such, over time a growing number of studies have begun seeking to develop explainability methods uniquely adapted to the domain of deep learning-based raw EEG analysis. It should be noted, however, that methods like those developed for multimodal data explainability [6], [7], [61], [62] can be adapted to multichannel EEG data with minimal inconvenience.

### Taxonomy of Explainability Methods for Deep Learning Models Trained on Raw EEG

In this section, we describe a taxonomy of explainability methods for deep learning models trained on raw EEG. As shown in Figure 1, deep learning-based explainability methods for raw EEG can be categorized on a hierarchy with two general levels: (1) based on the traditional features into which they provide insight and (2) based on the mechanisms by which they provide that insight. Existing explainability approaches can generally provide insight into 3 types of features: (1) spatial features (i.e., identifying specific brain regions or electrodes of importance), (2) spectral features (i.e., identifying specific frequency bands of importance), and (3) temporal features (i.e., identifying specific waveforms of interest). Approaches for identifying spatial or multimodal importance typically use some variation of ablation [6], [7], [9], [14], [63], [64], in which information from a particular channel is removed and the effect upon model performance or softmax activations is quantified, or a gradient-based feature attribution (GBFA) approach [65] like LRP [61], [63], in which importance is summed across all time points for each channel.

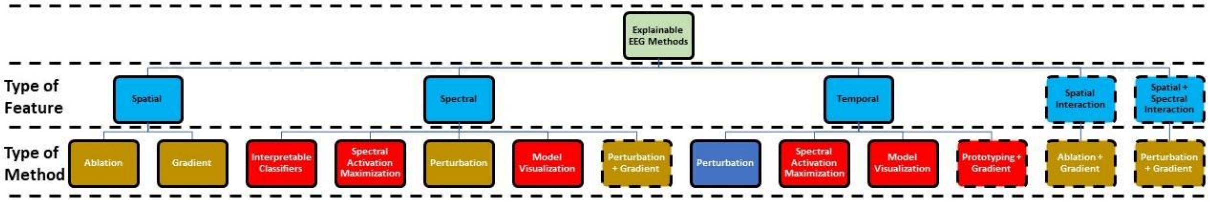

Figure 1. Taxonomy of Explainability Approaches for Deep Learning Models Using Raw EEG Data. The taxonomy has 3 levels that are each separated by black horizontal dashed lines: (1) the overall field of EEG explainability methods, (2) the types of features into which the explainability methods provide insight, and (3) the types of explainability approaches that provides insight into the different feature types. Light blue boxes correspond to specific types of features. Note that the type of method or explainability approach generally corresponds to methods that were first developed outside the domain of EEG analysis and then adapted to the domain. Dark blue, red, and gold boxes show methods that provide local, global, and both local and global explanations. Boxes surrounded by dashed black lines show explainability approaches first proposed in this study.

Approaches for identifying spectral importance generally fall within one of four categories. (1) They use interpretable classifiers with filters designed to extract specific frequencies [4], [66]. While highly innovative, these classifiers still inherently limit the space of possible features. (2) They use methods like activation maximization [56]. Two studies have sought to identify the frequencies at which sinusoids maximize the activation of early convolutional layers [14], [67], and one study sought to optimize the multi-spectral content of a sample to maximize the activation of the final softmax output layer in a class-specific manner [16]. These approaches do not actually indicate the importance of frequency bands to a classifier. Rather, they show a set of frequencies that are extracted at a particular layer or representative of a particular class. As such, they can produce a highly useful representation of what a model has learned. It should be noted that in methods like that in [16] there are theoretically multiple possible representations that could have high class-specific activations, so resulting samples may not contain all of the features important to the capacity of the model to identify each class. (3) A number of studies have used the fast Fourier transform to convert to and from the frequency domain wherein the perturbed specific frequency bands and examined the resulting effect upon model performance [9], [13] or predictions [12], [68]. These methods are highly effective. However, in some models, if performed on training data or sometimes even test data, perturbation may have a negligible effect upon model performance or activations, and an explanation may not be obtained. As an example, note how perturbation of frequency bands in the Awake class sample, which had an extremely high activation, had minimal effect upon the model activation in [16]. (4) Several studies have created specialized CNN architectures with extended first layer filters capable of extracting distinct waveforms [5], [40], [67]. The filters can then be converted to the frequency domain and visualized. These methods provide a very effective way for understanding what frequencies were extracted by a model. Additionally, when combined with perturbation, they can also provide an effective approach for estimating spectral importance. Nevertheless, they require the development of a highly specialized architecture and are thus incompatible with many architectures developed within the field.

Approaches for temporal waveform importance generally fall into one of three categories. Additionally, while these approaches can be effective to a degree, they all have key shortcomings, and there is generally significant room for continued innovation within this type of EEG explainability. (1) Windows of individual samples can be perturbed. Those windows that cause the largest change in softmax layer activation can then be considered important [14]. While this approach can give insight into the importance of individual waveforms for the classification of an individual sample, the perturbation of individual time windows may not always have a significant effect upon the softmax layer activation. A key shortcoming is that the identified waveforms also cannot be assumed to be of global significance, and it is impractical to perturb windows across a dataset with thousands or hundreds of thousands of samples. (2) Similar to approaches in spectral importance, activation maximization [56] can be applied to identify key waveforms [16]. Activation maximization approaches have been applied to other types of time-series classification [69], [70]. The methods in these two studies optimize the content of a sample in the time domain and are effective for short time-series (i.e., around 30 time points long). However, when applied to longer time-series like those found in resting-state EEG, they tend to do a very poor job creating recognizable waveforms [16]. This led to one study optimizing the spectral content of a sample to create waveforms [16]. This approach obtains more realistic waveforms than the approaches shown in other domains for shorter time-series [69], [70] but still leaves room for improvement. (3) Lastly, model visualization approaches that provide insights into important spectral features can also identify the importance of waveforms when paired with perturbation of model filters [5], [8], [40], [67]. However, while these approaches give the clearest insight into identified waveforms, they also require the design of special architectures that may not be able to obtain high levels of classification performance for all applications. It should also be noted that there is a tentative fourth category for identifying temporal waveform significance. GBFA methods like LRP can be applied to identify the relative importance of waveforms. This is a tentative category because while the approach has been used to identify important time points [71], it has not yet been applied to identify key waveforms. As demonstrated in several fMRI classification studies, this approach can also provide global insight into patterns of importance distribution [72], [73].

While significant advancement has been made in the field of deep learning-based explainability for raw EEG, there is still significant room for continued innovation and development. As previously described, models may sometimes not be sensitive enough to existing perturbation approaches to produce a significant change in model softmax activations or performance, which can prevent the approaches from providing usable explanations [16]. Additionally, the most effective approaches for insights into temporal waveform importance require the use of specially designed classifiers [5], [8], [40], [67], and there is a need for approaches that can be applied to any deep learning architecture. Lastly, while we have not yet mentioned this shortcoming and some explainability approaches have been applied to multichannel EEG data, many existing EEG explainability approaches have been developed within the context of single channel sleep stage classification. This is likely due to (1) the well-characterized features of sleep stages [74], (2) the comparative ease of developing models for sleep stage classification, and (3) the public availability of large EEG sleep stage datasets [75]–[77]. As such, there is a need to extend these approaches to multichannel EEG data, which is often used for more complex classification tasks [9], [21], [24], [26], [27], [78]–[81]. Related to this problem of multichannel explainability, is that, to the best of our knowledge, no existing approaches have sought to provide insight into interactions between different frequency bands and channels, which is a key limitation given the relative importance of connectivity-based features in models using traditional feature extraction [46]–[48].

In this study, we expand on the taxonomy that we previously presented. We present a systematic framework for evaluating the features of a deep learning classifier that includes inter-channel and spatio-spectral interaction - a novel feature type for deep learning-based raw EEG explainability. As such, our framework encompasses spatial, spectral, temporal, and interaction-based explanations. We (1) identify the relative importance of each channel, (2) identify interactions uncovered by the model between channels, (3) identify key frequency bands in each channel, (4) identify interactions between frequency bands in each channel and other channels, and (5) identify representative samples of each class and the waveforms of importance to their classification. We present our approach within the context of explaining a 1D-CNN trained on data from individuals with MDD (MDDs) and healthy controls (HCs). Our framework represents a significant step forward for the field of deep learning-based raw EEG explainability, and we hope that it will inspire future methods also capable of solving the problems that we presented in our taxonomy of explainability approaches.

## METHODS

In this section, we describe our proposed framework. As detailed in Figure 2, we (1) used multi-channel resting-state EEG data from 28 healthy controls (HCs) and 30 individuals with MDD (MDDs). (2) We trained a one-dimensional convolutional neural network (1D-CNN) for classification and evaluated model performance. (3) We applied layer-wise relevance propagation (LRP) to identify the relative importance of each channel, and (4) we applied a combination of ablation and LRP to identify interactions in the representations learned by the model between channels. We applied a combination of LRP and spectral perturbations to identify (5) the relative importance of each canonical frequency band in each channel and (6) interactions between the representations learned by the model for the canonical frequency bands in each channel and every other channel. (7) We applied a combination of a novel prototyping approach and LRP to identify representative samples of each class and identify important waveforms that the model used to differentiate them. Our code is publicly available on GitHub and can be found at: https://github.com/cae67/MultichannelExplainabilityFramework.git.

### Description of Data Acquisition and Preprocessing

We used a publicly available scalp EEG dataset [23] consisting of 30 MDDs and 28 age-matched HCs between the ages of 12 to 77 that has been used in multiple studies [24], [26], [78]. The data can be found at https://figshare.com/articles/dataset/EEG_Data_New/4244171. While we were not involved with data collection, we detail the collection procedures below. MDD participants met the diagnostic criteria for MDD defined in the Diagnostic and Statistical Manual-IV (DSM-IV) [82]. Common symptoms of MDD include a depressed mood, a loss of interest or pleasure, changes in appetite or weight, psychomotor agitation, feelings of worthlessness or excessive guilt, diminished ability to concentrate, and frequent thoughts of death [83]. HCs were determined to be healthy following examination for psychiatric conditions. To avoid potential confounding effects of medication, all MDDs underwent a two-week washout period prior to the first EEG recordings. All participants gave informed consent prior to data collection that was approved by the human ethics committee of the Hospital Universiti Sains Malaysia (HUSM) in Kelantan, Malaysia.

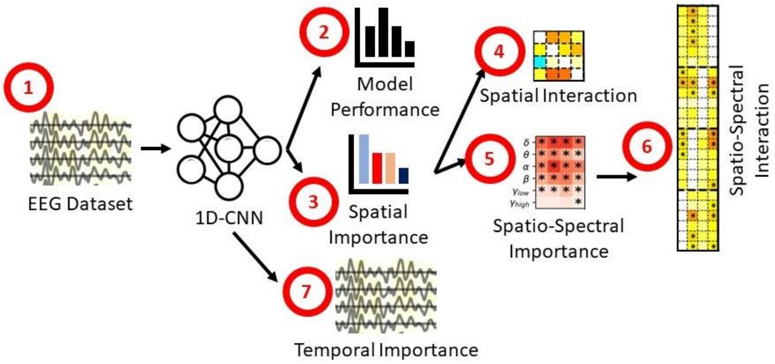

Figure 2. Overview of Methods. (1) We used a publicly available resting-state EEG dataset containing data from healthy individuals and individuals with major depressive disorder (MDD). (2) We first trained a one-dimensional convolutional neural network for automated MDD diagnosis and evaluated overall model performance. (3) We applied layer-wise relevance propagation (LRP) to identify the relative importance of each electrode (i.e., spatial importance). (4) We combined LRP with ablation to quantify how much the amount of LRP relevance assigned to each channel changed following the perturbation of other channels (i.e, spatial interaction). (5) We combined LRP with spectral perturbation to quantify how much the amount of LRP relevance assigned to a channel changed following the perturbation of frequency bands within that channel (i.e., spatio-spectral importance). (6) We combined LRP with spectral perturbation to quantify how much the amount of LRP relevance assigned to a channel changed following the perturbation of frequency bands within other channels (i.e., spatio-spectral importance). (7) Lastly, we combined a prototyping approach with LRP to identify representative samples of each class and to identify the relative importance of waveforms in each of those samples (i.e., temporal importance).

While separate recordings were performed at resting state with both eyes open and eyes closed for each participant, we only used data from recordings with eyes closed in this study. Participants were instructed to sit in a semi-recumbent position and minimize head movements and eye blinks. The Brain Master Discovery amplifier (Make: Brain Master, Model: Discovery 24e, Manufacturer: Brainmaster Technologies Inc.) was used to amplify EEG signals from the sensors. Recordings were performed for 5 to 10 minutes with a sampling rate of 256 Hertz (Hz) using a standard 10-20 format with 64 electrodes. The data were band pass filtered from 0.1 to 70 Hz and were notch filtered at 50 Hz to remove line-related noise. EEG data were recorded with the linked ear reference and were re-referenced to the infinity reference [84].

Due to the high levels of correlation present between scalp EEG channels, we only used a subset of electrodes. This approach is similar to those of other studies of neuropsychiatric disorders [9], [20], [25], [85]. Specifically, we used the Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2 electrodes. We downsampled the data from 256 Hz to 200 Hz. To increase the number of samples available for classification, we used a 25-second sliding window with a 2.5-second sliding step size to separate the recordings into epochs. After dividing the data into epochs, we channel-wise z-scored the epochs for each participant separately. Our final dataset consisted of 2,950 SZ epochs and 2,942 HC epochs. Importantly, we did not remove any samples with extreme amplitude values.

### Description of Model Development

We adapted an architecture (Figure 3) that was originally developed in [85] for schizophrenia classification and was later used in [78] for MDD classification. We implemented the model in Keras 2.2.4 [86] to maintain compatibility with explainability libraries. Input samples had dimensions of 5,000 time points x 19 channels. Relative to [85], we added multiple batch normalization layers and converted ReLU activation functions to ELU activation functions. We also modified the training approach from [78] in an effort to enhance model performance. As mentioned in [27], many previous studies involving classification of MDD EEG data have used poor cross-validation approaches, which can lead to an inflation of model performance. In an effort to enhance the generalizability of our models, we used a 10-fold stratified group shuffle split cross-validation approach to ensure that samples from the same participants were not simultaneously distributed across training, validation, and test sets in the same fold. Approximately, 75%, 15%, and 10% of samples were assigned to training, validation, and test sets, respectively.

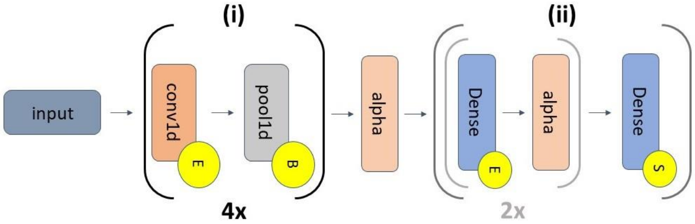

Figure 3. Model Architecture. The model can be subdivided into two segments that are separated by an alpha dropout layer (alpha) – feature extraction (i) and classification (ii). The feature extraction segment repeats 4 times, and the light grey inset within the classification segment repeats twice. Segment (i) has 4 one-dimensional convolutional layers (conv1d) that are each followed by max pooling layers (pool1d). The conv1d layers have 5, 10, 10, and 15 filters and have kernel sizes of 10, 10, 10, and 5. The pool1d layers have pool sizes and strides of 2. Segment (ii) has 3 dense layers with 64, 32, and 2 nodes, respectively. All alpha layers have dropout rates of 0.5. Yellow circles containing an “E”, “B”, or “S” indicate ELU activations, batch normalization, and softmax activations, respectively. Note that conv1d and dense layers have max norm kernel constraints with max values of 1.

During model training, we used a data augmentation approach that has previously been used in [15] to double our training set size. After separating the data into training, validation, and test sets in each fold, we duplicated the training data and augmented the duplicate data via the addition of Gaussian noise (mean = 0, standard deviation = 0.7). We then trained the model on the combined original and augmented data. To account for class imbalances that might randomly occur in the allocation of the training set, we used a class-weighted categorical cross-entropy loss function. We used an Adam optimizer [87] with a learning rate of 0.0075 and a batch size of 128 samples. We trained for a maximum of 35 epochs, using early stopping to end training if validation accuracy (ACC) did not improve after 10 consecutive epochs. To help ensure the generalizability of the model, we also selected the model from the epoch with the peak balanced validation accuracy (BACC). When assessing model test performance, we calculated the mean and standard deviation of the sensitivity (SENS), specificity (SPEC), ACC, and BACC across folds. All convolutional and dense layers, except for the final dense layer which had a softmax activation function with Glorot normal initialization [88], were initialized with He normal initialization [89]. Explainability analyses were performed on the test data from the model with the highest overall BACC.

## Description of Spatial Importance Approach

We applied the αβ-rule [90] of LRP [58], [91] for insight into the relative importance of each channel. LRP is a popular approach in the domain of explainability for image classification and has also been used extensively in the domain of neuroimaging and neurological time-series classification [5], [39], [40], [62], [63], [72], [73], [78], [92]–[97]. We implemented LRP using the Innvestigate library [98]. LRP involves multiple steps. (1) A sample is forward passed through a network. (2) A total relevance value of 1 is assigned to the output node corresponding to the class of interest. (3) The total relevance is iteratively propagated from layer to layer back through the network to the input space using a relevance rule. LRP can propagate both positive (i.e., identifying features that provide evidence for the class of interest) and negative (i.e., identifying features that provide evidence for a class other than the class of interest) relevance. To simplify our analysis by only examining relevance for samples corresponding to their true class, we used the αβ-rule. The αβ-rule has α and β terms, where α and β control positive and negative relevance propagation, respectively. We used α = 1 and β = 0. The equation below shows the αβ-rule.

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

Where the subscripts $k$ and $j$ correspond to values for one of $K$ nodes in a deeper layer and one of $J$ nodes in a shallower layer, respectively. The model weights are referenced by $w$, and $a_j$ is the shallower layer activation output.

We output relevance corresponding to the true classes of all test samples in the model with the highest test BACC. While LRP theoretically propagates relevance in a manner that sums to 1, practically, the total relevance can sometimes diverge. As such, after extracting relevance for each test sample, we normalized the absolute relevance of each sample to sum to 100 percent. Specifically, we summed the total absolute relevance assigned to each sample, divided the absolute relevance for each time point and channel by the total absolute relevance, and multiplied by 100. We next summed the total percent of absolute relevance assigned to each channel to estimate spatial importance for each sample. Lastly, to obtain class-specific spatial importance estimates, we averaged separately across HC, MDD, and HC + MDD samples.

The last spatial analysis that we performed sought to determine whether the average spatial importance for each channel was significantly above a uniform spatial distribution of relevance (i.e., where relevance for each channel equals total percent of absolute relevance divided by 19 channels). To this end, we performed a 1-sample t-test comparing the mean relevance of each channel to 100 percent / 19 channels. We then applied false discovery rate (FDR) correction [99] with α = 0.001 to reduce the likelihood of false positive test results. We performed this analysis for HC, MDD, and combined HC and MDD groups.

### Description of Spatial Interaction Approach

After identifying the relative importance of each channel, we sought to understand whether the model uncovered interactions between channels. To this end, we combined spatial LRP as detailed in the previous section with ablation. (1) We output the percent of absolute relevance for each sample and channel $C$ (see previous section). (2) We ablated channel $c$ of the test samples by replacing it with zeros. While we could have used line-related noise-based ablation approach similar to [63], we elected to use zeros, as line noise was notch filtered during data acquisition. (3) We re-output the percent of absolute relevance for each sample and channel $C$. (4) We calculated the absolute change in relevance belonging to each channel $C$.

The thought process behind our approach was that if a model has uncovered interactions between channels $c1$ and $c2$, then the model should rely upon $c1$ to interpret information in $c2$ and vice versa. Thus, if information in channel $c1$ is removed via ablation and the model uncovered a relationship between channels $c1$ and $c2$, the relevance of channel $c2$ should decrease because the model is no longer able to use the information in channel $c2$ as effectively. Additionally, if the relevance of channel $c2$

increases, then that indicates that the model compensated for the loss of channel $c1$ by relying more upon $c1$. As such, in our approach, we repeated steps 1 through 4 for each of $C$ channels and measured the effect of the ablation of channel $c$ upon spatial relevance for all of $C-1$ channels. Our approach relies upon the idea that the relevance of each channel is a combination of its relevance independent of other channels and of its interaction with all other channels.

$$Rel_c = Independent\_Rel_c + \sum_{i=1, i \neq c}^{C} Interaction\_Rel_{c,i}$$

Where $Rel_c$ is the total relevance assigned to channel $c$, $Independent\_Rel_c$ is the relevance independently assigned to channel $c$, and $Interaction\_Rel_{c,i}$ is the relevance of channel $c$ that results from interactions between channel $c$ and channel $i$, which is not channel $c$.

After outputting the change in relevance of all channels following the ablation of all channels, we sought to determine whether the interactions were statistically significant. To this end, we performed paired two-sample t-tests comparing the relevance assigned to channels before and after the ablation of each channel $c$. We next applied FDR correction [99] with α = 0.001 to reduce the likelihood of false positive test results. We performed this analysis separately for HC, MDD, and combined HC and MDD groups.

### Description of Spatio-Spectral Importance Approach

We next sought to uncover the relative importance of each canonical frequency band in each channel. We analyzed the canonical frequency bands: $\delta$ (0 – 4 Hz), $\theta$ (4-8 Hz), α (8 – 12 Hz), β (12 – 25 Hz), $\gamma_{low}$ (25 – 45 Hz), and $\gamma_{high}$ (55 – 100 Hz). Note that most of $\gamma_{high}$ was removed during the band pass filtering of the preprocessing, so analyzing $\gamma_{high}$ enabled us to sanity check our findings, as $\gamma_{high}$ importance should theoretically be very low. Our spatio-spectral importance analysis consisted of multiple steps. (1) We output the percent of absolute relevance for each test sample and channel $C$ (see previous sections). (2) We converted each sample to the frequency domain using a fast Fourier transform (FFT). (3) We assigned coefficients corresponding to frequency band $f$ in channel $c$ to values of zero. We could have randomly permuted coefficient values or reassigned them from a Gaussian distribution [10], [13], [15], [68]. However, doing so would have required repeatedly perturbing each channel and frequency band dozens of times, which would have been computationally prohibitive given subsequent steps. (4) We re-output the percent of absolute relevance for each sample and channel $C$. (5) We calculated the absolute change in relevance assigned to each channel $c$ following the perturbation of frequency band band $f$ in channel $c$.

After outputting the change in relevance of all channels following the perturbation of frequency bands, we sought to determine whether the frequency bands in each channel had statistically significant importance. To this end, we performed paired, two-sample, two-tailed t-tests comparing the relevance assigned to channels before and after the ablation of each channel $c$. We next applied FDR correction [99] with α = 0.001 to reduce the likelihood of false positive test results. We performed this analysis separately for HC, MDD, and combined HC and MDD groups.

### Description of Spatio-Spectral Interaction Approach

After identifying the relative importance of each frequency band in each channel, we sought to determine whether the model uncovered interactions between frequency bands $f$ in each channel $c$ with all other channels $C$ not including $c$. This analysis was highly similar to that described in the section, “Description of Spatio-Spectral Importance Approach”. The only difference between the two analyses was that in step 5, we calculated the absolute change in relevance assigned to all channels $C$ not including channel $c$ following the perturbation of frequency band $f$ in channel $c$. After obtaining the percent change in relevance assigned to channels $C$ following the the perturbation of frequency band $f$ in channel $c$, we employed the same paired, two-tailed, two-sample t-test approach followed by FDR correction described in the previous section.

### Description of Temporal Prototyping and Importance Approach

We lastly sought to uncover any key waveforms differentiating MDDs from HCs. To this end, we combined a prototyping-based approach to identify samples that ideally represented each class and applied LRP to identify the relative importance of each time point and channel for those samples. Our approach consisted of several stages. (1) We input all test samples for both classes into the model and output the activations from the final convolutional layer. (2) We applied principal component analysis (PCA) with 3 components to reduce the dimensionality of the extracted activations for samples in both classes. (3) We applied k-means clustering to the 3 principal components of the activations with 100 initializations sweeping from 2 to 10 clusters. We selected the optimal number of clusters using the maximum silhouette score [100]. We performed clustering separately for each class. (4) We selected the samples closest to the cluster centroids for each class. (5) We output normalized absolute LRP relevance for each sample using the αβ-rule. (6) We applied a moving average with a window size of 20 time points to the relevance assigned to each channel to make visualizing relevance distributions easier. This analysis was two-fold. Firstly, it enabled us to identify samples representative of each class that could be visually inspected for differences. Secondly, it gave insight into how the model analyzed the samples temporally (e.g., Was the relevance temporally distributed or focused on specific highly localized time points? Was the model focused exclusively on unique waveforms or only a few of many similar waveforms?).

## RESULTS

In this section, we describe our model performance, spatial importance, spatial interaction, spatio-spectral importance, spatio-spectral interaction, and temporal prototyping and importance results.

### Model Performance Results

Table 1 shows the classification test performance of our model across all folds. Mean performance for all metrics was above 80%. The model more effectively identified MDDs than HCs, as SENS was near 90% while SPEC was closer to 80%. Additionally, SPEC had a slightly higher standard deviation than SENS. BACC and ACC differed some for specific folds but were on average highly similar. The model test performance for the fold used for explainability was 100% across all metrics.

Table 1. Model Performance Results

| | ACC | BACC | SENS | SPEC |
|---|---|---|---|---|
| **Mean** | 84.90 | 85.57 | 89.03 | 82.11 |
| **Standard Deviation** | 09.39 | 09.93 | 13.92 | 17.43 |

### Spatial Importance

Figure 4 shows the average total absolute relevance for HCs, MDDs, and both classes combined as well as the t-test results comparing the relevance per channel to a uniform distribution of relevance (i.e., 100% relevance / 19 channels = 5.26% relevance per channel). Across classes, O1 and O2 were generally far below uniform. Additionally, F3 and C4 were unimportant across classes. Frontal (Fp2, F7, and F8) and parietal (Pz and P4) were consistently highly important across classes. Additionally, a few electrodes were of great importance for one class but not the other. C3 and T4 were highly important for MDD, and T6 was highly important for HC. T5 and P3 were also of moderate importance for HCs.

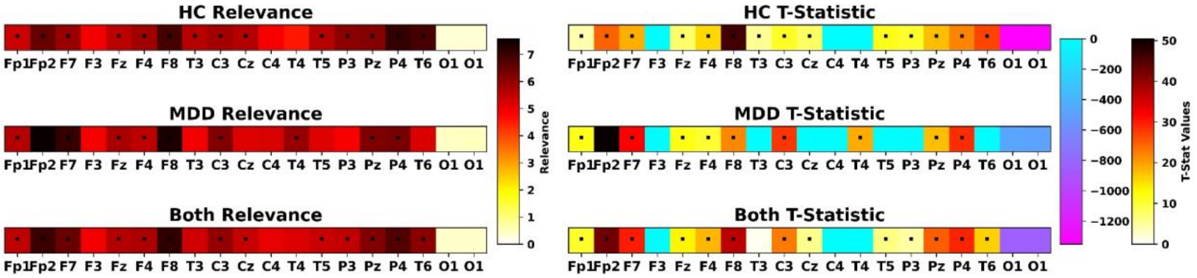

Figure 4. Spatial Importance Results. The leftmost panels show heatmaps of the average relevance for HCs, MDDs, and both classes in descending order. The heatmap to the right of the leftmost panels indicates the amount of relevance corresponding to the heatmap values. The rightmost panels show heatmaps of the t-statistics that resulted from our one-sample, two-tailed t-tests comparing the relevance of each channel to a uniform distribution of relevance (i.e., 100% of relevance / 19 channels). Panels values for HCs, MDDs, and both classes in descending order. Black dots indicate channels with statistically significant p-values following FDR correction (α = 0.001). The two color bars to the right of the leftmost panels indicate the t-statistic values corresponding to the channels with relevance below the uniform distribution and above the uniform distribution. Note that the names corresponding to each channel are displayed along the x-axis.

### Spatial Interactions

Figure 5 shows the results for our spatial interaction analysis. Channels that had a negative change in relevance following the perturbation of another channel can be considered to have an interaction with that channel. The P3 electrode which was of moderate importance to identifying HCs had a reduction in relevance when some frontal, temporal, and parietal electrodes were perturbed, indicating that the model likely relied upon information present in other electrodes to effectively use the information in the P3 electrode. Several other parietal and temporal electrodes (P4 and T6) also had reductions in relevance following the perturbation of some frontal, central, and parietal electrodes in HCs. While HCs tended to have reductions in more posterior electrodes (P3, P4, and T6) following the perturbation of other electrodes, MDDs tended to have reductions in central electrode relevance following the perturbation of other central electrodes. This was particularly the case for C3, which had MDD relevance above the uniform distribution, and Cz, which had MDD relevance slightly below the uniform distribution. While the HCs and MDDs did seem to have interactions between electrodes, there were comparatively fewer interactions similar across both classes. It should be noted that HCs seemed to have slightly more negative interactions than MDDs. Additionally, the model had reductions in relevance for a number of electrodes in both classes following the perturbation of occipital electrodes O1 and O2.

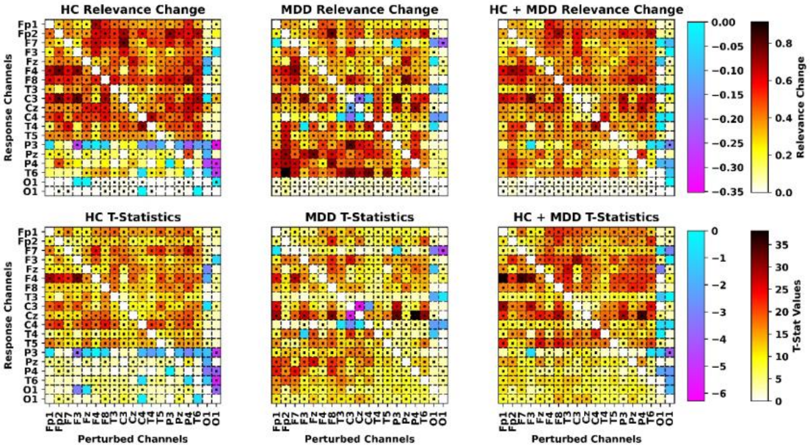

Figure 5. Spatial Interaction Results. From left to right, panels show average relevance interactions for HCs, MDDs, and both classes. The top panels show the change in relevance (relevance 2 – relevance 1), and the bottom panels show t-statistics for the paired, two-tailed, two-sample t-tests comparing the importance of channels before another channel was ablated versus after another channel was ablated. The x-axis indicates channels that were ablated, and the y-axis indicates channels in which a change in relevance is measured (i.e., response channels). Black dots indicate channel combinations in which there was a significant change in the relevance of a response channel after ablation following FDR correction (α = 0.001). Heatmaps to the right of the top and bottom rows of panels indicate the relative magnitude of the change in relevance and the t-statistic values for the t-tests, respectively. Note that values along the left-to-right diagonal were replaced with zeros, as the percent of relevance of a channel decreased much strongly when that channel was itself ablated and such extreme values prevented a visualization of the change in relevance of other channels.

While there were a few channels with negative changes in relevance following the perturbation of other channels, a larger number of channels had significant increases in relevance following the perturbation of other channels. In HCs, the model tended to increase relevance to frontal, central, and temporal electrodes following the ablation of any electrode. The model had particularly large increases in relevance for F4 and C3 following the perturbation of Fp2 and F3. In MDDs, the model tended to increase relevance assigned to central, temporal, and parietal electrodes following the ablation of all but parietal and thoracic electrodes.

## Spatio-Spectral Importance

Figure 6 shows the importance of each channel and the change in channel relevance following the perturbation of each canonical frequency band. The perturbation of most frequency bands did cause a significant reduction in relevance assigned to their corresponding channels. Nevertheless, there were differences in the magnitude of those effects between HCs and MDDs. The model relied upon θ more strongly and across more channels for identifying HCs than MDDs, and the model relied upon β and $\gamma_{low}$ across a wider range of channels for identifying MDDs than HCs. Importantly, the model relied upon α for identifying both classes. The model did not rely extensively upon $\gamma_{high}$. The overall most important frequency and channel combinations for HCs were more posterior parietal (P3, PZ, P4) and temporal (T6) θ and α and frontal (Fz, F4, F8) θ. Overall most important frequency and channel combinations for MDDs were Fp2 α and β, and F7, F8, T4 β and $\gamma_{low}$.

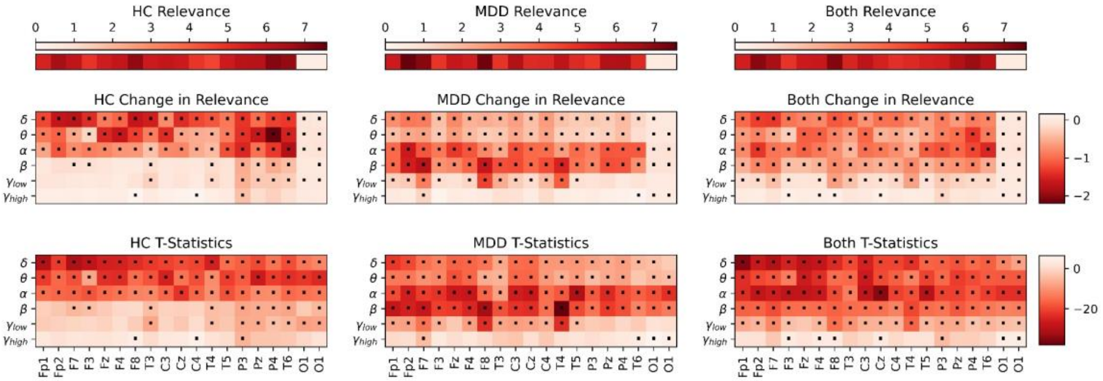

Figure 6. Spatio-Spectral Importance Results. The left, middle, and right columns of panels show results for HCs, MDDs, and both HCs and MDDs, respectively. The top row of panels indicates spatial relevance of each channel. The middle row of panels indicates average change in relevance of specific channels across samples following the perturbation of frequency bands within those channels. The bottom row of panels indicates the t-statistics for the two-sample, two-tailed, paired t-tests comparing the relevance of a channel before versus after perturbation of a frequency band within that channel. Black dots indicate channel and frequency band combinations in which there was a significant reduction in the relevance of a channel after perturbation following FDR correction (α = 0.001). The x-axis shows channels, and the y-axis shows frequency bands. Their corresponding color bars are to the right of the middle and bottom rows, and the corresponding color bars for the top row are located above the panels.

## Spatio-Spectral Interactions

Figure 7 shows spatio-spectral interactions for each frequency band and channel and other channels. For HCs, the model identified a reduction in relevance for parietal and occipital (P3, Pz, P4, T6, O1, O2) following the perturbation of most other channels and frequency bands. However, P3 was the only channel with consistently statistically significant reductions in relevance following the perturbation of other channels. For MDDs, several channels along the frontal and central planes (F7, C4) tended to have strong reductions in relevance following the perturbation of most channels, and a couple other channels in the frontal and central planes had reductions in relevance following the perturbation of a few channel and frequency combinations (Fp1, T3, C3). The channel F7 was the only channel with consistently statistically significant reductions in relevance following the perturbation of frequency bands in other channels. Those channels that did not have negative changes in relevance generally had strongly positive changes. Few channels had near-zero positive changes, though some had near-zero negative changes.

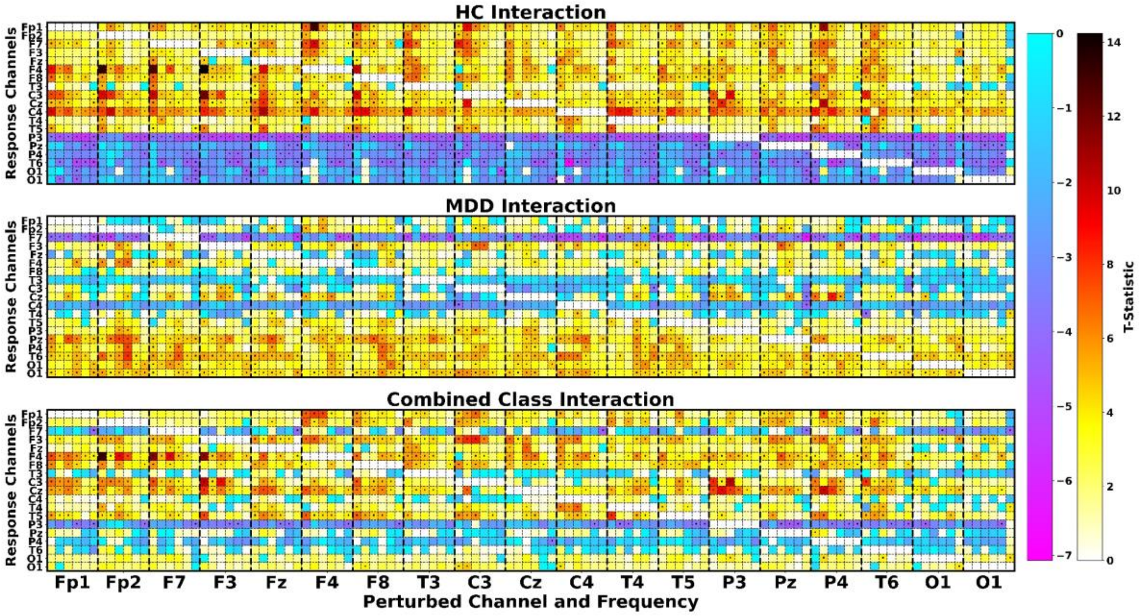

Figure 7. Spatio-Spectral Interaction Results. The top, middle, and bottom panels show heatmaps of the t-statistics from the two-sample, two-tailed, paired t-tests comparing the amount of relevance in a channel before versus after the perturbation of frequency bands in other channels for HCs, MDDs, and both classes, respectively. Perturbed channels and frequency bands are arrayed along the x-axis. Perturbed channels are separated by thick vertical dashed lines, and from left to right within each set of vertical dashed lines are shown results for the perturbation of $\delta$ (0 – 4 Hz), $\theta$ (4-8 Hz), α (8 – 12 Hz), β (12 – 25 Hz), $\gamma_{low}$ (25 – 45 Hz), and $\gamma_{high}$ (55 – 100 Hz) frequency bands. Channels in which a change in relevance was measured are arrayed along the y-axis (i.e., response channels). The color bars to the right of the figure are shared by all panels and indicate the value of the t-statistics in the heatmaps. Black dots indicate channel and frequency band combinations in which there was a significant change in the relevance of a channel after perturbation following FDR correction (α = 0.001).

## Temporal Prototyping and Importance

Figure 8 shows the extracted CNN features with dimensionality reduced via PCA. It also shows the clusters and samples closest to each cluster centroid. We identified 3 HC clusters and 2 MDD clusters based on the maximal silhouette scores for clustering each class. Figure 9 shows the 3 HC and 2 MDD samples closest to each cluster centroid along with an overlayed LRP relevance heatmap that highlights the most important regions of each time-series. MDD and HC clusters were highly separable. HC cluster 2 seemed to have a representation more comparable to that of the MDD clusters. HC clusters 1 and 3 were relatively similar, though they seemed to be somewhat separable along the vertical axis. Interestingly, the samples for the HC 1, HC 3 and MDD 2 clusters were higher along vertical axis and tended to have more high frequency activity, with HC 1 and HC 3 seeming to have high amounts of $\gamma_{low}$ and MDD 2 having activity at the boundary of β and $\gamma_{low}$. HCs had consistent levels of θ oscillations, which were not present in MDDs. Additionally, both MDDs seemed to contain β oscillations that were not found in HCs. Both HCs and MDDs seemed to have α oscillations. LRP relevance tended to be more highly concentrated in HCs than MDDs, and MDD 2 relevance was much more concentrated than MDD 1 relevance, which had a more diffuse representation as shown through PCA. In HCs, rather than selecting unique waveforms, the model seemed to focus on a few of many reoccurring waveforms (e.g., θ waveforms). This was also often the case for MDD 2, though some unique bursts of high frequency activity (e.g., between 19 and 20 seconds in Figure 9) were also highlighted. In MDD 1, many high amplitude bursts of parietal α were relevant, and these bursts sometimes co-occurred with β activity in central electrodes highlighted (e.g, around 1 second in Figure 9).

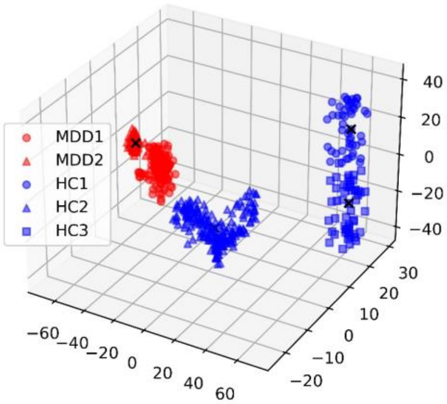

Figure 8. PCA Extracted Activation Clustering Results for Temporal Explainability. Three principal components were used to reduce the dimensionality of activations for the test samples from the final convolutional layer. HC and MDD reduced activations are shown in blue and red, respectively. Two MDD and 3 HC clusters were optimal. Different clusters for each class are each indicated by markers with a different shape, as shown in the legend to the left of the plot. A black “x” is used to mark the samples closest to the cluster centroids that were used in subsequent temporal explainability analyses.

## DISCUSSION

In this study, we make two major contributions. (1) We define a taxonomy of explainability approaches for deep learning models trained on raw EEG data. (2) We present a framework for systematically evaluating a deep learning model trained on raw multichannel EEG data that provides insight into each type of feature (i.e., spatial, spectral, and temporal) found in our taxonomy while also expanding upon that taxonomy to provide insight into new types of features (i.e., spatial and spatio-spectral interactions). Importantly, to the best of our knowledge, our methods for examining spatial and spatio-spectral interactions are the first of their kind for raw EEG deep learning classification. Additionally, our approach for spectral importance enables a more sensitive identification of key frequency bands than previous approaches by examining the change in relevance of individual channels (rather than the change in a softmax activation or model accuracy) following perturbation. Both our novel spectral and interaction explainability approaches can provide both local and global insights. Lastly, in contrast to previous approaches that required a specially designed architecture or required that a model be sensitive to the perturbation of input samples, our approach for temporal importance offers an approach for global importance estimation that is applicable to a variety of deep learning classifiers. As a whole in recent years, the field of explainable deep learning for EEG has made great progress, but existing approaches still leave much to be desired. The collection of novel methods presented in our study represents a significant step forward for the field, and the taxonomy that we propose represents a key advancement that will provide guidance for future studies and developments.

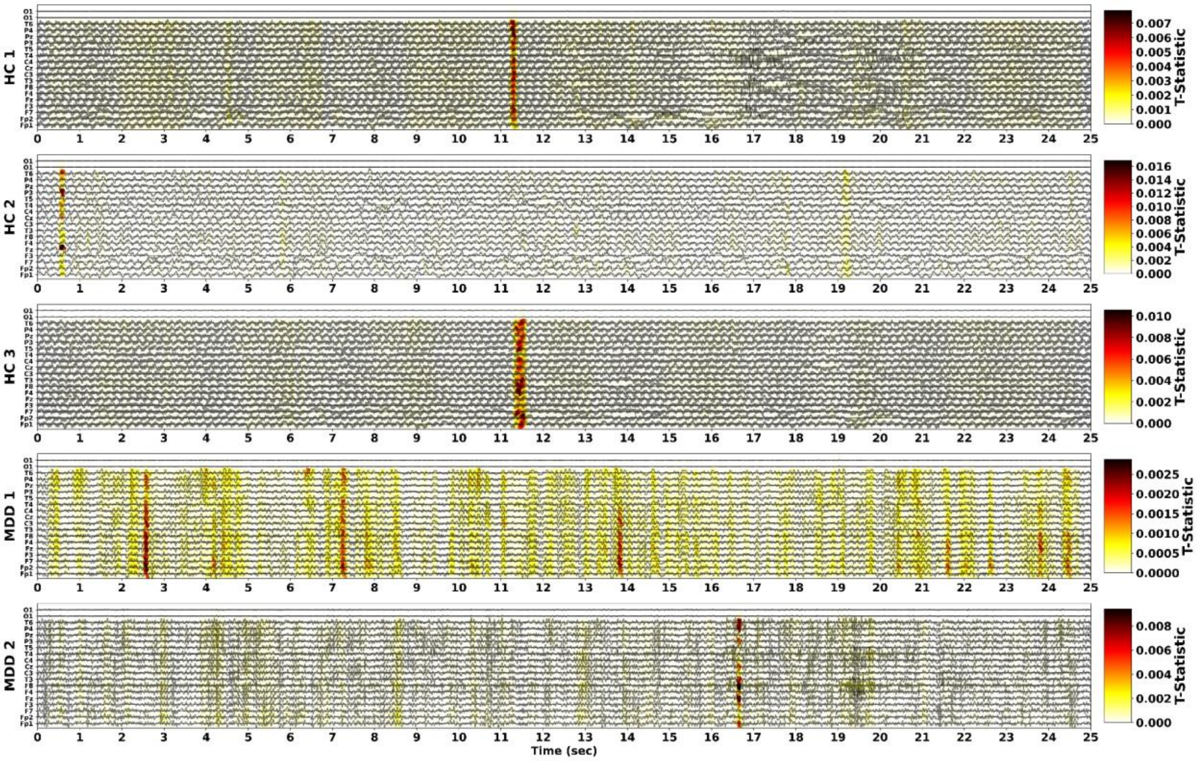

Figure 9. Prototyping and Temporal Importance Results. The time-series of samples closest to the cluster centers of clusters HC 1, HC 2, HC 2, MDD 1, and MDD 2 (as shown in Figure 8) are displayed in panels from top to bottom. The x-axis shows time in seconds, and the y-axis shows EEG channels. The mean of the data for each channel was subtracted for easier display. A heatmap of LRP importance is overlayed on the time-series. Color bars to the right of the panels show the values of the LRP relevance out of 100% for their respective panels.

## Training a High Performing Model with a Robust Cross-Validation Approach

We developed a model for the classification of individuals with MDD and healthy controls. Our overall model performance was very high (i.e., greater than 80% across all metrics), and the model performance from the fold used in explanations was at 100% across all metrics, increasing the likelihood of the potential generalizability of our explainability findings. Relative to previous studies performing automated diagnosis of MDD using raw EEG data with robust cross-validation techniques, our model obtained higher performance [78], and relative to studies using extracted features with traditional machine learning approaches and robust cross-validation techniques, our model obtained comparable or higher performance [27]. There were some studies that obtained higher test performance than our model using either raw EEG data [80], [81] or extracted features [3], [21], [24], [26]. However, it appears based on the descriptions of their cross-validation approaches that those studies allowed data from the same study participants to leak across training, validation, and test sets within the same folds. That leakage would inflate model test performance and prevent the test performance from actually giving an indication of the generalizability of the patterns learned by their models. This problem in the field is unfortunately relatively common and has been described more extensively in previous studies [27]. Our stratified group shuffle split cross-validation approach protected against this leakage and helped ensure the reliability of our performance findings.

### Identifying Electrodes and Electrode Interactions Important to the Identification of Healthy Individuals and Individuals with MDD

The model relied upon frontal and parietal information for identifying both classes indicating that it was able to uncover discriminatory activity for both classes in those areas. However, the model also relied upon more central electrodes for identifying MDDs and upon more posterior electrodes for identifying HCs. This finding is interesting when combined with our channel interaction results. Namely, the model tended to rely upon more central interactions when identifying MDDs and more posterior interactions when identifying HCs. This finding of spatially widespread effects of MDD is consistent with previous studies that found it necessary to rely upon information from spatially distributed electrodes to obtain high levels of performance [27]. Additionally, it is interesting that while the model did not rely heavily upon occipital (O1 and O2) electrodes, the model did seem to uncover widespread interactions between those electrodes and other electrodes across the scalp. As such, while information in the O1 and O2 electrodes may have not been highly discriminative, that information may have helped the model interpret information in other electrodes.

### Examining Why Electrodes are Important by Identifying Their Important Composite Frequency Bands and Frequency Band Interactions

Our findings of spatial importance are further illuminated within the context of our findings on spatio-spectral importance and interactions. While some frequency bands like $\alpha$ that have well-characterized importance in MDD [50], [79] were of widespread importance to the model for identifying both classes, some combinations of frequency bands and channels were important to specific classes. The importance of the posterior electrodes that were important for identifying HCs can be attributed to the presence of $\theta$ and $\alpha$ in those electrodes, and the presence of interactions between more posterior electrodes with other electrodes is also found in spatio-spectral interactions where there are widespread interactions with all frequency bands across most channels. The importance of the central electrodes to MDDs is attributable to $\alpha$ in those electrodes. While spatial importance indicates that frontal electrodes are important to both classes, spatio-spectral importance indicates that frontal electrodes are important to each class for different reasons. In HCs, frontal $\theta$ is more important, and in MDDs, frontal $\beta$ and $\gamma_{low}$ are more important. Importantly, frontal $\theta$ has been identified as discriminatory between HCs and MDDs [26], and frontal and central $\beta$ and $\gamma_{low}$ have previously been associated with inattention in MDD [101]. Additionally, our finding of low $\gamma_{high}$ importance supports the reliability of the methods, as much of $\gamma_{high}$ was filtered during initial signal amplification. It is curious that spatio-spectral interactions tend to be more widespread than spatial interactions. This is potentially attributable to how our use of zero-out ablation for identifying spatial interactions and use of perturbation for identifying spatio-spectral interactions interacted differently with the model. Previous studies have shown the importance of choosing ablation and perturbation methods specific to the target domain [63]. While it would have been ideal to be able to use a line noise-related ablation approach in our spatial interactions, the data that we used was publicly available, and line noise was notch filtered during the data collection process. As such, the model would not have learned to consider line noise as neutral information, and the line noise-related ablation approach would thus not be viable.

### Identification of Characteristic Samples for Each Class and Key Waveforms Within Those Samples

Our dimensionality reduction and clustering approach seemed to uncover some underlying structure in the representations learned by the model. For example, clusters higher along the vertical dimension seemed to have more high frequency activity, and there seemed to be high levels of separation between MDD and HC clusters. Additionally, the model did not seem to uncover unique waveforms of particularly high importance in HCs and MDD 1. Rather it seemed to primarily focus in a highly temporally localized manner on specific waveforms that were identical to many other waveforms found across the 25-second samples. This indicates that while we used 25-second sample sizes, it may have been possible to train an effective classifier with much shorter samples. For MDD 2, the model seemed to have a much more diffuse PCA representation and much more temporally distributed relevance. Additionally, the identified waveforms also illuminate our spatio-spectral importance findings. Specifically, we previously identified that θ was highly important to identifying HCs, and in the samples identified via our prototype approach, HCs had consistently high levels of θ oscillations that were not found in MDDs. Both HCs and MDDs had high levels of α oscillations, which explains why α was important for identifying both classes, and MDDs had high levels of β oscillations, which explains the importance that the model placed upon β for identifying MDDs.

### Limitations and Next Steps

There are several new opportunities for future research directions that are spawned by this study. Our prototyping approach could potentially be expanded upon in future studies. The waveforms identified with our prototyping approach seemed to align well with our previously identified spatio-spectral importance estimates. However, future studies might apply more local explainability approaches to the identified samples to determine how well the findings for the identified prototypes fit with findings for the entire dataset. If there is a high degree of alignment between the findings for each of the prototypes and the global dataset findings, future studies could potentially adapt more robust methods like SHAP [60] for insight into spectral, spatial, and interaction importance that would otherwise not be viable for application with whole EEG datasets given their computational complexity. Additionally, while we applied LRP to provide a measure of channel importance that could change following perturbation and that approach should be broadly applicable to both CNNs and models with recurrent units, future studies might apply approaches similar to ours within the context of other architectures by measuring changes in model attention. Lastly, all of our analyses were performed in the sensor space. Future efforts might use inverse modeling to obtain source space signals and train models on those signals. Resulting explanations could provide enhanced insights into specific brain regions associated with classification performance.

While our proposed explainability approaches were highly effective and present new opportunities for future research, our study methods and findings do have some limitations. Some of the limitations are not unique to our study but rather a problem for the overall field of deep learning-based studies using explainability methods. Specifically, model explanations are not meant to provide an exhaustive investigation of which features could possibly be discriminatory between individuals with MDD and healthy controls. EEG data can be very rich, and as has been shown in previous studies [9], [25], there are often multiple sets of features upon which a model can rely when performing a classification. As such, if our presented methods were to be used in an attempt to obtain exhaustive insight into all of the features that might possibly be useful for diagnosing MDD, it would be necessary to use a more robust training procedure (e.g., many more folds than is the popular practice) and use multiple independently collected datasets. Another limitation of our study findings is related to our dataset size. If we were to try to make generalizable claims about which features are most important for diagnosing MDD, we would need a much larger dataset. Lastly, there are several limitations to the methods proposed in this specific study. Namely, we only perturbed features once, and it would be ideal if we could examine the effects of multiple perturbations upon the same features. Perturbing features only once is relatively common within studies ablating whole channels or modalities, so it is not overly problematic for our spatial interaction analysis. Within spectral importance analyses, it is more common to perturb individual frequency bands more than once; however, due to the computational complexity of repeatedly outputting LRP explanations, we elected to just replace the coefficients of each frequency band with zeros. That said, our use of statistical testing to identify the most important features does help ensure the reliability of our findings. Our use of spectral perturbation may also have caused some edge effects, though this is also a potential problem for all spectral perturbation explainability methods. Alternative approaches might consider applying windows to samples to attenuate any edge effects or performing notch filtering.

## CONCLUSION

The application of deep learning methods to raw EEG data is becoming increasingly common. However, relative to other methods that use traditional machine learning or deep learning with extracted features, deep learning models applied to raw EEG data are less easily explainable. As a result, a field of research has developed seeking to explain these models. In this study, we propose a taxonomy of the explainability methods that have been developed for deep learning models trained on raw EEG. We then introduce an explanatory framework consisting of a series of methods that build upon our proposed taxonomy. In addition to providing insights into key spatial, spectral, and temporal features like existing approaches, the methods in our framework also provide insight into spatial and spatio-spectral interactions uncovered by models. We present our framework within the context of a 1D-CNN trained for automated major depressive disorder diagnosis, identifying interactions between central electrodes and other electrodes and identifying differences in frontal θ, β, and $\gamma_{low}$ between healthy individuals and individuals with major depressive disorder. Our study represents a significant step forward for the field of deep learning-based raw EEG classification, providing new capabilities in interaction explainability and providing directions for future research innovations through our proposed taxonomy.

## REFERENCES

[1] M. Manjusha and R. Harikumar, "Performance analysis of KNN classifier and K-means clustering for robust classification of epilepsy from EEG signals," *Proc. 2016 IEEE Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2016*, pp. 2412–2416, 2016, doi: 10.1109/WiSPNET.2016.7566575.

[2] G. Ruffini *et al.*, "Deep Learning With EEG Spectrograms in Rapid Eye Movement Behavior Disorder," *Front. Neurol.*, vol. 10, Jul. 2019, doi: 10.3389/fneur.2019.00806.

[3] C. Uyulan *et al.*, "Major Depressive Disorder Classification Based on Different Convolutional Neural Network Models: Deep Learning Approach," *Clin. EEG Neurosci.*, vol. 52, no. 1, pp. 38–51, 2021, doi: 10.1177/1550059420916634.

[4] D. Borra, S. Fantozzi, and E. Magosso, "Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination," *Neural Networks*, vol. 129, pp. 55–74, Sep. 2020, doi: 10.1016/j.neunet.2020.05.032.

[5] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "A Systematic Approach for Explaining Time and Frequency Features Extracted by Convolutional Neural Networks From Raw Electroencephalography Data," *Front. Neuroinform.*, vol. 16, no. May, pp. 1–11, 2022, doi:

10.3389/fninf.2022.872035.

[6] C. A. Ellis *et al.*, “A Novel Local Ablation Approach For Explaining Multimodal Classifiers,” in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, 2021, pp. 1–6.

[7] C. A. Ellis, R. Zhang, D. A. Carbajal, R. L. Miller, V. D. Calhoun, and M. D. Wang, “Explainable Sleep Stage Classification with Multimodal Electrophysiology Time-series,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 0–3.

[8] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces,” *J. Neural Eng.*, vol. 15, no. 5, 2018, doi: 10.1088/1741-2552/aace8c.

[9] C. A. Ellis, A. Sattiraju, R. Miller, and V. Calhoun, “Examining Effects of Schizophrenia on EEG with Explainable Deep Learning Models,” in *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*, 2022, pp. 301–304, doi: 10.1109/BIBE55377.2022.00068.

[10] C. A. Ellis, R. L. Miller, and V. D. Calhoun, “A Convolutional Autoencoder-based Explainable Clustering Approach for Resting-State EEG Analysis,” in *bioRxiv*, 2023, pp. 3–6.

[11] M. Shim, H. J. Hwang, D. W. Kim, S. H. Lee, and C. H. Im, “Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features,” *Schizophr. Res.*, vol. 176, no. 2–3, pp. 314–319, 2016, doi: 10.1016/j.schres.2016.05.007.

[12] R. T. Schirrmeister *et al.*, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.

[13] D. O. Nahmias and K. L. Kontson, “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*, Aug. 2020, pp. 2398–2406, doi: 10.1145/3394486.3403289.

[14] S. Pathak, C. Lu, S. B. Nagaraj, M. van Putten, and C. Seifert, “STQS: Interpretable multi-modal Spatial-Temporal-seQuential model for automatic Sleep scoring,” *Artif. Intell. Med.*, vol. 114, no. January, p. 102038, 2021, doi: 10.1016/j.artmed.2021.102038.

[15] C. A. Ellis, A. Sattiraju, R. Miller, and V. Calhoun, “Examining Effects of Schizophrenia on EEG with Explainable Deep Learning Models,” 2022.

[16] C. A. Ellis, M. S. E. Sendi, R. Miller, and V. Calhoun, “A Novel Activation Maximization-based Approach for Insight into Electrophysiology Classifiers,” 2021.

[17] C. A. Ellis, R. L. Miller, and V. D. Calhoun, “A Model Visualization-based Approach for Insight into Waveforms and Spectra Learned by CNNs,” in *IEEE*, 2021, pp. 1–4.

[18] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, “A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, 2018.

[19] K. Singh, S. Singh, and J. Malhotra, “Spectral features based convolutional neural network for

accurate and prompt identification of schizophrenic patients," *Proc. Inst. Mech. Eng. Part H J. Eng. Med.*, vol. 235, no. 2, pp. 167–184, 2021, doi: 10.1177/0954411920966937.

[20] A. Shoeibi *et al.*, "Automatic Diagnosis of Schizophrenia in EEG Signals Using CNN-LSTM Models," *Front. Neuroinform.*, vol. 15, no. November, pp. 1–16, 2021, doi: 10.3389/fninf.2021.777977.

[21] A. Saeedi, M. Saeedi, A. Maghsoudi, and A. Shalbaf, "Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach," *Cogn. Neurodyn.*, vol. 15, no. 2, pp. 239–252, 2021, doi: 10.1007/s11571-020-09619-0.

[22] S. Yasin, S. A. Hussain, S. Aslan, I. Raza, M. Muzammel, and A. Othmani, "EEG based Major Depressive disorder and Bipolar disorder detection using Neural Networks: A review," *Comput. Methods Programs Biomed.*, vol. 202, 2021, doi: 10.1016/j.cmpb.2021.106007.

[23] W. Mumtaz, L. Xia, M. A. M. Yasin, S. S. A. Ali, and A. S. Malik, "A wavelet-based technique to predict treatment outcome for Major Depressive Disorder," *PLoS One*, vol. 12, no. 2, pp. 1–30, 2017, doi: 10.1371/journal.pone.0171409.

[24] H. W. Loh, C. P. Ooi, E. Aydemir, T. Tuncer, S. Dogan, and U. R. Acharya, "Decision support system for major depression detection using spectrogram and convolution neural network with EEG signals," *Expert Syst.*, vol. 39, no. 3, pp. 1–15, 2022, doi: 10.1111/exsy.12773.

[25] C. A. Ellis, A. Sattiraju, R. Miller, and V. Calhoun, "Examining Reproducibility of EEG Schizophrenia Biomarkers Across Explainable Machine Learning Models," in *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*, 2022, pp. 305–308, doi: 10.1109/BIBE55377.2022.00069.

[26] R. A. Movahed, G. P. Jahromi, S. Shahyad, and G. H. Meftahi, "A major depressive disorder classification framework based on EEG signals using statistical, spectral, wavelet, functional connectivity, and nonlinear analysis," *J. Neurosci. Methods*, vol. 358, no. November 2020, p. 109209, 2021, doi: 10.1016/j.jneumeth.2021.109209.

[27] C. Te Wu *et al.*, "Resting-state EEG signal for major depressive disorder detection: A systematic validation on a large and diverse dataset," *Biosensors*, vol. 11, no. 12, 2021, doi: 10.3390/bios11120499.

[28] N. Ince, F. Goksu, G. Pellizzer, A. Tewfik, and M. Stephane, "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*, 2008, pp. 3554–7.

[29] S. M. Kia, F. Pedregosa, A. Blumenthal, and A. Passerini, "Group-level spatio-temporal pattern recovery in MEG decoding using multi-task joint feature learning," *J. Neurosci. Methods*, 2017, doi: 10.1016/j.jneumeth.2017.05.004.

[30] T. J. Gawne *et al.*, "A multimodal magnetoencephalography 7 T fMRI and 7 T proton MR spectroscopy study in first episode psychosis," *npj Schizophr.*, vol. 6, no. 1, pp. 1–9, 2020, doi: 10.1038/s41537-020-00113-4.

[31] M. S. E. Sendi *et al.*, "Aberrant Dynamic Functional Connectivity of Default Mode Network in Schizophrenia and Links to Symptom Severity," *Front. Neural Circuits*, vol. 15, no. March, pp. 1–14, 2021, doi: 10.3389/fncir.2021.649417.

[32] C. A. Ellis, M. S. E. Sendi, R. L. Miller, and V. D. Calhoun, "An Unsupervised Feature Learning Approach for Elucidating Hidden Dynamics in rs-fMRI Functional Network Connectivity," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 4449–4452.

[33] E. Zendehrouh *et al.*, "Aberrant Functional Network Connectivity Transition Probability in Major Depressive Disorder," in *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020, pp. 1493–1496.

[34] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "Explainable Fuzzy Clustering Framework Reveals Divergent Default Mode Network Connectivity Dynamics in Schizophrenia," *bioRxiv*, 2023.

[35] M. S. E. Sendi, C. A. Ellis, R. L. Milller, D. H. Salat, and V. D. Calhoun, "The relationship between dynamic functional network connectivity and spatial orientation in healthy young adults," *bioRxiv*, 2021.

[36] M. S. E. Sendi *et al.*, "The link between brain functional network connectivity and genetic risk of Alzheimer's disease," *bioRxiv*, 2021, doi: 10.1002/alz.050101.

[37] M. S. Salman, E. Verner, H. J. Bockholt, Z. Fu, and V. D. Calhoun, "Machine Learning Predicts Treatment Response in Bipolar Major Depression Disorders," *BIBE 2021 - 21st IEEE Int. Conf. Bioinforma. Bioeng. Proc.*, 2021, doi: 10.1109/BIBE52308.2021.9635339.

[38] A. Vahid, M. Mückschel, S. Stober, A. K. Stock, and C. Beste, "Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control," *Commun. Biol.*, vol. 3, no. 1, 2020, doi: 10.1038/s42003-020-0846-z.

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

[40] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "A Model Visualization-based Approach for Insight into Waveforms and Spectra Learned by CNNs," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2022, vol. 2022-July, pp. 1643–1646, doi: 10.1109/EMBC48229.2022.9871414.

[41] Q. Yu and V. D. Calhoun, "Resting-State Functional Network Disturbances in Schizophrenia," *Brain Netw. Dysfunct. Neuropsychiatr. Illn.*, pp. 187–215, 2021, doi: 10.1007/978-3-030-59797-9_10.

[42] R. Boostani, K. Sadatnezhad, and M. Sabeti, "An efficient classifier to diagnose of schizophrenia based on the EEG signals," *Expert Syst. Appl.*, vol. 36, no. 3 PART 2, pp. 6492–6499, 2009, doi: 10.1016/j.eswa.2008.07.037.

[43] J. J. Newson and T. C. Thiagarajan, "EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies," *Front. Hum. Neurosci.*, vol. 12, no. January, pp. 1–24, 2019, doi: 10.3389/fnhum.2018.00521.

[44] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "A Systematic Approach for Explaining Time and Frequency Features Extracted by CNNs from Raw EEG Data," *bioRxiv*, 2022.

[45] M. Rashed-Al-Mahfuz, M. A. Moni, S. Uddin, S. A. Alyami, M. A. Summers, and V. Eapen, "A Deep Convolutional Neural Network Method to Detect Seizures and Characteristic Frequencies Using Epileptic Electroencephalogram (EEG) Data," *IEEE J. Transl. Eng. Heal. Med.*, vol. 9, no. January,

pp. 1–12, 2021, doi: 10.1109/JTEHM.2021.3050925.

[46] M. A. Vázquez, A. Maghsoudi, and I. P. Mariño, “An Interpretable Machine Learning Method for the Detection of Schizophrenia Using EEG Signals,” *Front. Syst. Neurosci.*, vol. 15, no. May, pp. 1–11, 2021, doi: 10.3389/fnsys.2021.652662.

[47] C. Phang, C. Ting, F. Noman, and H. Ombao, “Classification of EEG-Based Brain Connectivity Networks in Schizophrenia Using a Multi-Domain Connectome Convolutional Neural Network,” pp. 1–15.

[48] K. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, “A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals,” *Comput. Biol. Med.*, vol. 99, no. May, pp. 24–37, 2018, doi: 10.1016/j.compbiomed.2018.05.019.

[49] L. Zhang, “EEG Signals Classification Using Machine Learning for the Identification and Diagnosis of Schizophrenia,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 4521–4524, 2019, doi: 10.1109/EMBC.2019.8857946.

[50] A. A. Fingelkurts, A. A. Fingelkurts, H. Rytsälä, K. Suominen, E. Isometsä, and S. Kähkönen, “Composition of brain oscillations in ongoing EEG during major depression disorder,” *Neurosci. Res.*, vol. 56, no. 2, pp. 133–144, 2006, doi: 10.1016/j.neures.2006.06.006.

[51] M. Hata *et al.*, “Functional connectivity assessed by resting state EEG correlates with cognitive decline of Alzheimer’s disease - An eLORETA study,” *Clin. Neurophysiol.*, vol. 127, no. 2, pp. 1269–1278, 2016, doi: 10.1016/j.clinph.2015.10.030.

[52] D. Maheshwari, S. K. Ghosh, R. K. Tripathy, M. Sharma, and U. R. Acharya, “Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals,” *Comput. Biol. Med.*, vol. 134, no. May, p. 104428, 2021, doi: 10.1016/j.compbiomed.2021.104428.

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[54] C. Molnar, *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*, 2018th-08–14th ed. Lean Pub, 2018.

[55] L. E. O. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[56] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” Dec. 2013, [Online]. Available: http://arxiv.org/abs/1312.6034.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

[58] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS One*, vol. 10, no. 7, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[59] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[60] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," 2017.

[61] C. A. Ellis, R. L. Miller, V. D. Calhoun, and M. D. Wang, "A Gradient-based Approach for Explaining Multimodal Deep Learning Classifiers," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, 2021, pp. 0–5.

[62] C. A. Ellis, D. A. Carbajal, R. Zhang, R. L. Miller, V. D. Calhoun, and M. D. Wang, "An Explainable Deep Learning Approach for Multimodal Electrophysiology Classification," *bioRxiv*, pp. 12–15, 2021.

[63] C. A. Ellis *et al.*, "Novel Methods for Elucidating Modality Importance in Multimodal Electrophysiology Classifiers," *bioRxiv*, 2022.

[64] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2069–2072, doi: https://doi.org/10.1145/3357384.3358160.

[65] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks," in *International Conference on Learning Representations*, 2018, pp. 1–16.

[66] D. Borra, S. Fantozzi, and E. Magosso, "EEG Motor Execution Decoding via Interpretable Sinc-Convolutional Neural Networks," in *Mediterranean Conference on Medical and Biological Engineering and Computing*, 2019, vol. 1, pp. 1515–1525, doi: 10.1007/978-3-030-31635-8.

[67] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks," *arXiv*, 2016, [Online]. Available: http://arxiv.org/abs/1610.01683.

[68] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "A Novel Local Explainability Approach for Spectral Insight into Raw EEG-Based Deep Learning Classifiers," in *21st IEEE International Conference on BioInformatics and BioEngineering*, 2021, pp. 0–5.

[69] N. Yoshimura, T. Maekawa, and T. Hara, "Preliminary Investigation of Visualizing Human Activity Recognition Neural Network," *2019 12th Int. Conf. Mob. Comput. Ubiquitous Network, ICMU 2019*, pp. 4–5, 2019, doi: 10.23919/ICMU48249.2019.9006643.

[70] N. Yoshimura, T. Maekawa, and T. Hara, "Toward Understanding Acceleration-based Activity Recognition Neural Networks with Activation Maximization," 2021.

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

[72] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "Towards Greater Neuroimaging Classification Transparency via the Integration of Explainability Methods and Confidence Estimation Approaches," *Informatics Med. Unlocked*, vol. 37, 2023, doi: https://doi.org/10.1016/j.imu.2023.101176.

[73] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "Pairing Explainable Deep Learning Classification with Clustering to Uncover Effects of Schizophrenia Upon Whole Brain Functional Network Connectivity Dynamics," *bioRxiv*, 2023.

[74] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, "The AASM Manual for Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications." 2007.

[75] "PhysioNet: The Sleep-EDF database [Expanded]." .

[76] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, no. October 2017, pp. 180–192, 2016, doi: 10.1016/j.cmpb.2015.10.013.

[77] S. F. Quan *et al.*, "The Sleep Heart Health Study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997, doi: 10.1093/sleep/20.12.1077.

[78] C. A. . Ellis, A. Sattiraju, R. L. . Miller, and V. D. . Calhoun, "Novel Approach Explains Spatio-Spectral Interactions in Raw Electroencephalogram Deep Learning Classifiers," in *bioRxiv*, 2023, pp. 2–6.

[79] S. Mahato and S. Paul, "Detection of major depressive disorder using linear and non-linear features from EEG signals," *Microsyst. Technol.*, vol. 25, no. 3, pp. 1065–1076, 2019, doi: 10.1007/s00542-018-4075-z.

[80] A. Thakare, M. Bhende, N. Deb, S. Degadwala, B. Pant, and Y. P. Kumar, "Classification of Bioinformatics EEG Data Signals to Identify Depressed Brain State Using CNN Model," *Biomed Res. Int.*, vol. 2022, no. 1, 2022, doi: 10.1155/2022/5214195.

[81] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, 2018, doi: 10.1016/j.cmpb.2018.04.012.

[82] *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)*, 4th ed. Washington DC: American Psychiatric Association Inc., 2000.

[83] S. H. Kennedy, "Core symptoms of major depressive disorder: Relevance to diagnosis and treatment," *Dialogues Clin. Neurosci.*, vol. 10, no. 3, pp. 271–277, 2008, doi: 10.31887/dcns.2008.10.3/shkennedy.

[84] L. Marzetti, G. Nolte, M. G. Perrucci, G. L. Romani, and C. Del Gratta, "The use of standardized infinity reference in EEG coherency studies," *Neuroimage*, vol. 36, no. 1, pp. 48–63, 2007, doi: 10.1016/j.neuroimage.2007.02.034.

[85] S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, and U. R. Acharya, "Deep convolutional neural network model for automated diagnosis of Schizophrenia using EEG signals," *Appl. Sci.*, vol. 9, no. 14, 2019, doi: 10.3390/app9142870.

[86] F. Chollet, "Keras," *GitHub*, 2015. https://github.com/fchollet/keras.

[87] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.

[88] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, 2010.

[89] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1026–1034, 2015, doi: 10.1109/ICCV.2015.123.

[90] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Cham: Springer International Publishing, 2019.

[91] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017, doi: 10.1109/TNNLS.2016.2599820.

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

[93] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "An Approach for Estimating Explanation Uncertainty in fMRI dFNC Classification," *2022 IEEE 22nd Int. Conf. Bioinforma. Bioeng.*, 2022.

[94] C. A. Ellis, R. L. Miller, and V. D. Calhoun, "Neuropsychiatric Disorder Subtyping Via Clustered Deep Learning Classifier Explanations," in *bioRxiv*, 2022, pp. 12–15.

[95] A. W. Thomas, H. R. Heekeren, K.-R. Müller, and W. Samek, "Analyzing Neuroimaging Data Through Recurrent Deep Learning Models," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.09945.

[96] J. M. Mayor-Torres, S. Medina-DeVilliers, T. Clarkson, M. D. Lerner, and G. Riccardi, "Evaluation of Interpretability for Deep Learning algorithms in EEG Emotion Recognition: A case study in Autism," pp. 1–12, 2021, [Online]. Available: http://arxiv.org/abs/2111.13208.

[97] W. Yan *et al.*, "Discriminating Schizophrenia From Normal Controls Using Resting State Functional Network Connectivity: A Deep Neural Network and Layer-wise Relevance Propagation Method," 2017.

[98] M. Alber *et al.*, "INNvestigate neural networks!," *J. Mach. Learn. Res.*, vol. 20, no. 93, pp. 1–8, 2019.

[99] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[100] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[101] S. C. Roh, E. J. Park, M. Shim, and S. H. Lee, "EEG beta and low gamma power correlates with inattention in patients with major depressive disorder," *J. Affect. Disord.*, vol. 204, pp. 124–130, 2016, doi: 10.1016/j.jad.2016.06.033.