Improving Explainability for Single-Channel EEG Deep Learning Classifiers via Interpretable Filters and Activation Analysis*

Charles A. Ellis
Tri-institutional Center for Translational Research in Neuroimaging and Data Science
Georgia State University, Georgia Institute of Technology, Emory University
Atlanta, USA
cellis42@gsu.edu

Robyn L. Miller
Tri-institutional Center for Translational Research in Neuroimaging and Data Science
Georgia State University, Georgia Institute of Technology, Emory University
Atlanta, USA
robyn.l.miler@gmail.com

Vince D. Calhoun
Tri-institutional Center for Translational Research in Neuroimaging and Data Science
Georgia State University, Georgia Institute of Technology, Emory University
Atlanta, USA
vcalhoun@gsu.edu

***Abstract*—Deep learning methods are increasingly being applied to raw electroencephalography (EEG) data. Relative to traditional machine learning methods, deep learning methods can increase model performance through automated feature extraction. Nevertheless, they are also less explainable. Existing raw EEG explainability methods identify relative feature importance but do not identify how features relate to model predictions. In this study, we combine well-characterized first layer filters with a novel post hoc statistical analysis of the filter activations that linearly relates properties of model activations with predictions. We implement our approach within the context of automated sleep stage classification, finding that the model uncovers positive relationships between waveforms resembling sleep spindles and NREM2, between high frequency filters and Awake samples, and between δ activity and NREM3. Our approach represents a significant step forward for raw EEG explainability and has the potential to provide many insights relating learned features and model predictions in future studies.**

Keywords— explainable AI, sleep stage classification deep learning, electroencephalography, activation analysis

I. Introduction

The application of deep learning methods to raw electroencephalography (EEG) data has grown increasingly common in recent years with the eventual goal of developing EEG-based clinical decision support systems (CDSSs) [1]–[5]. While these methods can offer improved performance relative to traditional machine learning methods applied to extracted EEG features [6], [7], they also make explainability more challenging. As such, a growing number of studies have developed EEG-specific explainability methods [8]–[13]. However, while these methods provide important insights, they are mainly feature importance approaches. Namely, they identify the degree to which features are important to the classification being performed, but they do not specify how the important features relate to the classes being identified. This distinction is critical, as simply pointing to important features without knowing why those features are important provides an insufficient level of explainability for clinicians who are ethically obligated to explain their recommendations to patients [14]. In this study, we present a novel explainability approach that combines well-characterized, interpretable first layer filters in a one-dimensional convolutional neural network (1D-CNN) with a post hoc activation analysis that linearly relates activations of learned features to model predictions.

Early applications of traditional machine learning to electrophysiology involved manually extracted features [6], [7]. These methods were often directly interpretable (e.g., elastic net regression) [6] or easily explainable with methods like permutation feature importance (e.g., random forests) [7]. However, as the field of deep learning has grown in popularity, the application of deep learning methods to raw EEG data has become much more common. Deep learning methods with raw EEG data have been used in automated sleep state classification [1], [2], [15] and schizophrenia [3], [16] and major depressive disorder diagnosis [4], [5]. While these methods can improve performance and enable an unbiased approach to the extraction of features via automated feature extraction, they are less explainable relative to traditional machine learning methods.

This lack of explainability has resulted in the development of a variety of domain-specific explainability methods. As recently taxonomized in [4], existing EEG explainability methods generally fall into one of five categories based on the types of features that they identify. They can provide insight into the importance of key (1) frequency bands [8]–[13], (2) waveforms [2], [8], [9], [17], [18], (3) channels [3], [5], [19], (4) spatial interactions [4], and (5) spatio-spectral interactions [4], [5], though the third through fifth categories are restricted to multi-channel EEG data. Many studies use methods that involve data perturbation [1], [3], [12], which can lead to out-of-distribution samples that yield poor explanations [20]. One alternative approach is model visualization (i.e., structuring first layer filters such that characteristics of learned features can be understood) [8], [9], [21], [22].

Funded by NIH R01MH123610, NIH R01MH118695, and NSF 2112455.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

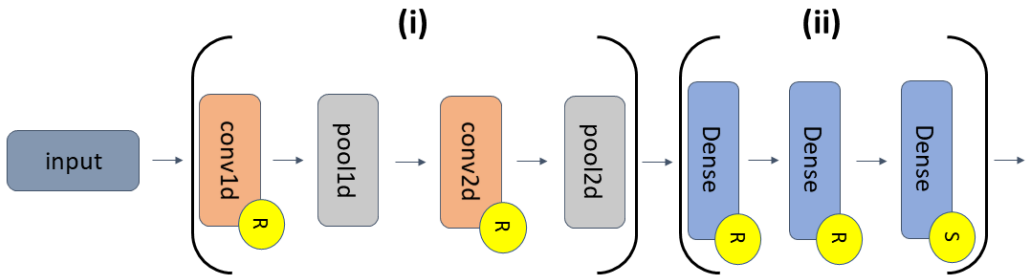

Fig. 1. CNN Architecture. The model has two components: a feature extraction portion (i) and a classifier portion (ii). Component (i) has a 1D convolutional layer (conv1d) with 30 filters of length 200 (i.e., approximately 2 seconds) and a stride of 1, a 1D max pooling layer (pool1d) with a pooling size of 15 and stride of 10, a conv2d layer with 400 30x25 filter and a stride of 1x1, and a pool2d layer with a 10x1 pooling size and 1x2 stride. Component (ii) has 2 dense layers with 500 nodes followed by a dense layer with 5 nodes. Yellow circles indicate activation functions: ReLU (R) and softmax (S).

Unfortunately, both model visualization and perturbation-based approaches typically only provide a measure of feature importance. While feature importance can aid model validation, they do not explain the role of features in model predictions. This could be problematic for clinical settings in which clinicians need to not only be able to explain to patients what features a model relied upon but also explain what role those features relate to model predictions [14].

In this study, we present a novel post hoc statistical analysis approach that builds upon existing model visualization approaches [8], [22] by identifying linear relationships between properties of activations for well-characterized filters and model predictions. More plainly, our approach identifies key waveforms and how those waveforms relate to the probability of being assigned to a given class. We present our approach within the context of automated sleep stage classification, comparing the insights obtained from our novel analysis with those that can be obtained through an existing filter perturbation-based feature importance analysis. Our approach offers enhanced insights into how our model uses extracted features and represents a key step forward for the domain of EEG explainability.

II. Methods

In this study, we trained a 1D-CNN with long first layer filters, visualized the first layer filters in the time and frequency domains, assigned the filters to clusters based on their spectral power, perturbed each cluster of filters to identify its importance, and performed linear regression to relate the mean and variance of the activations for each cluster to model predictions. Sleep stages included Awake, rapid eye movement (REM), Non-REM 1 (NREM1), NREM2, and NREM3.

A. Dataset and Preprocessing

In this study, we used the Sleep Cassette data from the Sleep-EDF Expanded dataset [23] that is publicly available on PhysioNet [24]. The dataset was originally collected in [25]. Because the dataset was publicly available, Institutional Review Board approval at our institution was unnecessary. The dataset consisted of 153 day-night 20-hour recordings from 78 participants. All but 3 participants had 2 recordings from consecutive days. Similar to previous studies, we used data from the FPz-Cz electrode that was recorded at 100 Hertz [1], [2]. Using the Rechtschaffen and Kales manual [26], experts assigned 30-second segments of data to different sleep stages: Awake, REM, NREM1 NREM2, NREM3, and NREM4.

NREM3 and NREM4 were combined into a single NREM3 class based on clinical guidelines [27]. Due to the use of full-day recordings, the dataset was heavily imbalanced towards the Awake class. To balance the data, we removed Awake samples at the start and end of each recording. Prior to separating the data into 30-second segments, we z-scored each recording separately. Our final dataset consisted of 85,034 Awake, 21,522 NREM1, 69,132 NREM2, 13,039 NREM3, and 25,835 REM samples.

B. Model Development

We adapted the 1D-CNN architecture initially developed in [21] and later used in [8], [22]. We implemented it in Keras 2.2.4 [28] and Tensorflow 1.15.0 [29]. The architecture was first developed to work with much longer samples that might contain multiple sleep stages. As our goal was to develop and evaluate a new explainability approach using existing sleep stage literature as a ground truth, we thought it better to train the architecture on 30-second samples that were less likely to contain multiple sleep stages, and optimizing performance for that sample length did require some modification. Fig. 1 shows our architecture. We used a 10-fold cross-validation approach that involved assigning 63, 7, and 8 participants to training, validation, and test sets, respectively. Note that assigning all samples from a given participant to either the training, validation, or test set in a particular fold and not mixing across sets prevented an inflation of model performance that is relatively common in the field and has been identified as a key problem in previous studies [30]. We used a class-weighted categorical cross entropy loss function to address class imbalances, and we used stochastic gradient descent with an adaptive learning rate that began at 0.015 and decreased by 10% if validation accuracy did not increase within a period of 5 training epochs. We also applied early stopping if 20 epochs passed without an improvement in validation accuracy and trained for a maximum of 30 epochs. We used a batch size of 100 during training and assessed model test performance by calculating the F1 score, recall, and precision for each class in each fold. All explainability analyses were performed on the model from the fold with the highest weighted F1 score across classes. In the fold used for explainability, we also computed a confusion matrix.

C. Filter Clustering and Visualization

To characterize the features extracted by the model, we converted all first layer filters to the frequency domain using a fast Fourier transform (FFT) and extracted their spectral power. We then clustered the spectral power values using k-means

Table 1. Model Performance Results.

	Awake	NREM1	NREM2	NREM3	REM
F1	91.60±02.09 (94.16)	41.04±03.51 (40.52)	79.91±02.80 (83.91)	68.82±12.79 (76.96)	68.35±05.37 (74.91)
Recall	88.18±03.54 (90.97)	45.39±05.45 (56.15)	77.83±04.28 (81.69)	63.98±13.37 (82.57)	71.93±06.01 (67.23)
Precision	95.40±01.84 (97.59)	38.22±05.89 (31.69)	82.22±02.14 (86.25)	75.85±14.59 (72.06)	65.80±08.67 (84.56)

clustering with 50 initializations, sweeping from 2 to 15 clusters. To select the optimal number of clusters, we computed three cluster quality metrics: the silhouette score [31], Calinski-Harabasz score [32], and Davies-Bouldin score [33]. We then ranked the number of clusters based on the 3 scores, computed the average rank, and selected the number of clusters with the best overall rank. We lastly visualized the filters in each cluster in both the time and frequency domains. Clustering simplified subsequent analyses by allowing us to analyze a smaller number of filter groups with similar features rather than having to characterize and analyze all filters separately.

D. Filter Perturbation Analysis

To identify the relative importance of each cluster of filters using a perturbation approach, we calculated the weighted and class-specific recall for the test data from the fold with the highest test performance. We then replaced all filters within a given cluster with zeros and recalculated the test performance. We lastly computed the percent change in test performance following the perturbation of each cluster. This analysis is adapted from [8] and provides an important point of comparison for our novel filter activation analysis with existing approaches.

E. Filter Activation Analysis

While our earlier analyses characterized the features extracted by the model and the importance of those features, they did not identify the nature of that importance. We next examined how the model may have used those extracted features and sought to determine how that use might relate to its predictions. We extracted two activation values for each test sample: the (1) mean and (2) variance of the activations for each filter following the first max pooling layer. We then calculated the mean of those values across filters within each cluster and z-scored the values for each class separately. We next performed 10 linear ordinary least squares regression analyses with (1) the mean activations and (2) the variance of the activations as the independent variables and the predicted probability of the clinician-assigned sleep stage as the dependent variable. We performed separate analyses for each sleep stage and activation value (i.e., 5 sleep stages x 2 activation values = 10 analyses) and applied false discovery rate correction ($p < 0.05$) to the p-values generated in each analysis. To identify relationships between the activation values and predicted probabilities, we visualized the regression coefficients and computed the coefficient of determination (R^2).

III. Results and Discussion

In this section, we describe and discuss our model performance results, extracted features, filter cluster importance, and the relationships between filter cluster activations and model predictions. We further discuss the unique aspects of our approach and its broader implications for the field.

A. Examining Model Performance

Table 1 shows the mean and standard deviation of model test performance for each class across all folds and for the performance in the top fold used in the explainability analyses (shown in parentheses). Fig. 2 shows a confusion matrix of the test performance for the top fold. Performance was well above chance-level (i.e., 20%) for all classes, with the model having highest F1 scores for the two classes with the largest number of samples (i.e., the Awake class followed by the NREM2 class). Although REM had roughly twice the number of samples as NREM3, model performance for both classes was highly comparable, which could indicate that the model found NREM3 features to be more clearly defined. Model performance for the smallest class (NREM1) was lowest. NREM1 performance has been lowest in a number of studies [13], [34]–[36], and some studies have specifically developed methods with the goal of increasing NREM1 performance [34]. This could be attributable to the relatively low number of NREM1 samples or the similarity of NREM1 to Awake and REM [27], [35]. Model performance for the fold used in explanations was mostly above the mean performance across folds. Additionally, based on the top fold confusion matrix, many Awake samples were assigned to the NREM1 class, and many NREM1 samples were assigned to NREM2, followed by REM and Awake. Some NREM3 samples (i.e., approximately 20%) were assigned to NREM2, and some REM samples were assigned to NREM1 and NREM2. These misclassifications likely affect our explainability analysis results.

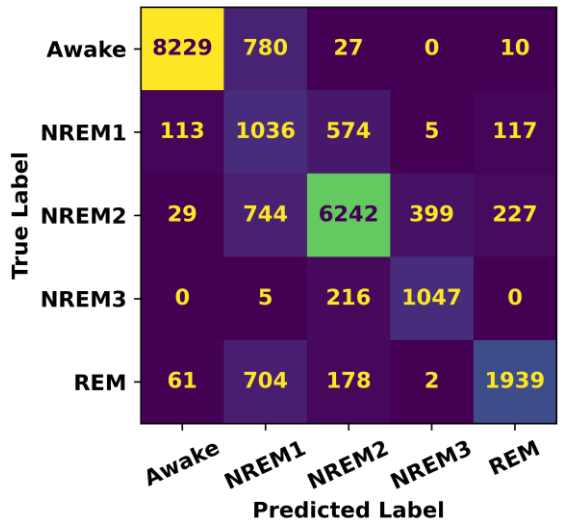

Fig. 2. Top Fold Confusion Matrix. True and predicted labels are shown on the x- and y-axes, respectively. Brighter colors indicate that a larger portion of samples reside within a classification group.

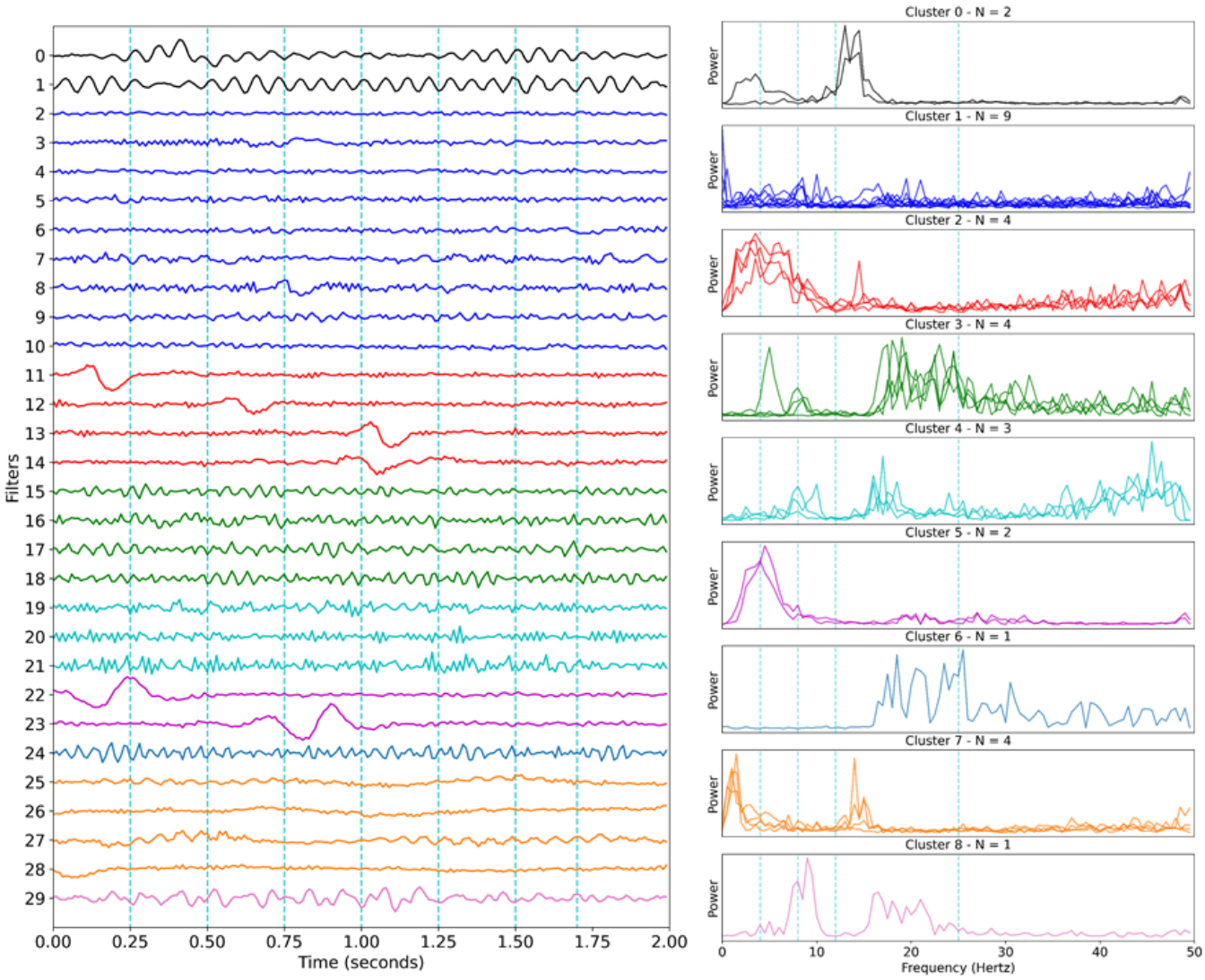

Fig. 3. Filter Waveforms and Spectral Clusters. The left panel shows the filters generated by the CNN. Clusters of filters are arranged based on color. Those colors correspond to the colors of the rightmost panel, which shows the spectral power of filters in each cluster. The title for each subpanel indicates the corresponding cluster number and number of filters assigned to each cluster. Gray dashed lines at 4, 8, 12, and 25 Hz reflect the boundaries of each canonical frequency band in order of δ (0-4Hz), θ (4-8Hz), α (8-12Hz), β (12-25Hz), and γ (25-50Hz). The y-axis of the rightmost panel indicates spectral power. Filters in the leftmost panel are min-max normalized to enable easier visual inspection of the waveforms.

B. Identifying Extracted Features and Their Importance

We identified 9 clusters of first layer filters with unique spectral properties. Fig. 3 shows the filter clusters in the time and frequency domains, and Fig. 4 shows the perturbation-based importance of each cluster of filters to model performance. Cluster 0 was characterized primarily by lower β activity, with some δ activity. It was extremely important to the identification of NREM2 and, to a lesser degree, NREM 3. It seemed to have waveforms similar to sleep spindles, which are found in NREM2 [27]. Additionally, NREM3 is primarily characterized by δ activity, which could explain why cluster 0 was important for identifying NREM3. Cluster 1 was very noisy with little average power in one frequency band, which could explain why it was largely unimportant to the model. Given that it contained nearly a third of the model filters, the model may have been overly complex. Cluster 1 was of minor importance for Awake and REM, which could be explained by how REM is defined as being low amplitude and mixed frequency [27].

Cluster 2 mainly extracted δ and θ activity. Its extraction of δ could explain its importance for NREM3. Additionally, eye blink artifacts appear between 0.5 – 2 Hz and were not filtered from the dataset, which could explain why cluster 2 was important for Awake. Cluster 3 was characterized by upper β activity and was important for identifying NREM1 and REM. Its waveforms are reminiscent of vertex sharp waves which appear in the upper β band in NREM1. It is unclear why Cluster 3 was important for identifying REM, but that could be related to the misclassification of many REM samples as NREM1.

Cluster 4 extracted upper γ activity with some β activity. Its perturbation had a moderate effect on Awake performance, which is reasonable given that γ most often occurs during wakefulness [37]. Cluster 5 mainly extracted δ and θ activity, and its perturbation had strong effects on NREM2 and REM

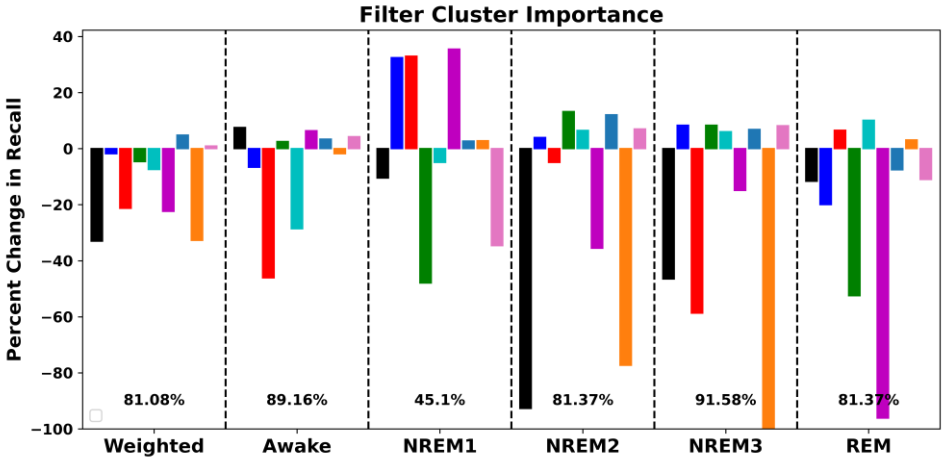

Fig. 4. Cluster Perturbation Importance. The figure shows the percent change in weighted and class-specific recall after filters in each cluster were replaced with zeros. The y-axis shows the change in recall. The x-axis shows the corresponding class. Within each class, each cluster is shown from left to right as a bar plot in colors corresponding to those of Fig. 3.

performance. The waveforms in Cluster 5 resemble inverted k-complexes, which are an important aspect of NREM2 [27]. Cluster 5 also shares frequencies and waveforms similar to sawtooth waves that are present during REM [27].

Cluster 6 extracted upper β activity and some γ activity. Its perturbation had negligible effects on model performance, with the exception of REM for which it only had a small effect. That could be because Cluster 6 had only one filter. It should be noted, however, that γ has been associated with REM [38]. Cluster 7 was characterized by δ activity with some lower β activity. It was very important for the identification of NREM2 and NREM3. Its importance for NREM3 is attributable to its slow waveform δ activity, and its β activity resembles the sleep spindles found in NREM2 [27]. Cluster 8 was the only cluster to extract α activity. It also extracted some θ and β and was mainly important for identifying NREM1. Its importance for NREM1 could be attributable to its extraction of θ [27].

C. *Identifying Linear Relationships Between Activations and Model Predictions*

Fig. 5 shows the linear regression coefficients, FDR-corrected p-values, and R^2 values from our analysis relating the mean and variance of the test activations to model predictions. The NREM3 regression analysis had the highest R^2, followed by the Awake class, for both the activation mean and variance. The activation mean R^2 values tended to be higher than for the variance, and across both measures, the NREM1, NREM2 and REM R^2 were more modest than those of NREM3 and Awake. When analyzing explanations, we compare them to the American Academy for Sleep Medicine (AASM) Manual [27].

The filter perturbation analysis identified cluster 0 to be related to NREM2 and NREM3 classification, and our activation analysis provided further insight. Namely, increased activation mean and variance corresponded to increased likelihood of being identified as NREM2, and increased activation mean but negligible variance corresponded to increased likelihood of NREM3. This indicates that the model sought near continuous δ activity when identifying NREM3. In contrast, while it expected a strong presence of waveforms resembling sleep spindles in NREM2, it also expected these waveforms to only be periodically present.

Cluster 1 activation mean and variance were positively associated with REM, and Cluster 1 activation mean was negatively associated with NREM1. This makes sense for REM given its multi-spectral nature [27]. For NREM1, the correspondence of the negative Cluster 1 coefficient to a strong increase in model performance after perturbation is interesting.

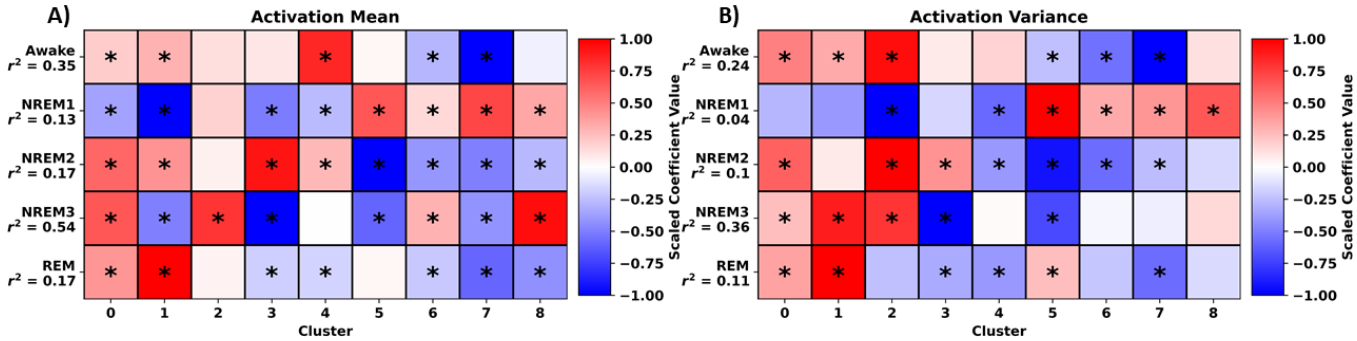

Fig. 5. Linear Regression Results Relating Mean and Variance of Activations to Model Predicted Probabilities. Panels A and B show results for the activation mean and variance, respectively. The x-axis indicates the class and the corresponding R^2 for the analysis. The y-axis indicates the corresponding cluster. The heatmap shows scaled coefficient values, where each coefficient value was scaled by the max magnitude of other coefficients in its respective regression analysis. Asterisks indicate a significant p-value following FDR correction ($p < 0.05$).

Interestingly, the Cluster 2 mean activation was only significantly related to NREM3, while the activation variance was related to Awake, NREM1, NREM2, and NREM3. This provides complementary insight to our filter perturbation analysis. We hypothesized that the importance of Cluster 2 for Awake could be attributed to eye blinks, and our finding of Cluster 2 activation variance being highly related to Awake classification corroborates this finding, as the presence of eye blinks would be intermittent. The likelihood of NREM1 negatively corresponded to the mean and variance of Cluster 2 activations, which could relate to the increase in NREM1 model performance following the perturbation of Cluster 2. It is odd that the filter perturbation analysis did not find a large change in NREM2 performance after the perturbation of Cluster 2, while Cluster 2 activation variance was highly related to NREM2 predicted probabilities. Similar to our findings for the filter perturbation analysis, increased Cluster 2 activation mean and variance corresponded to increased NREM3 likelihood.

Interestingly, while our filter perturbation analysis only identified Cluster 3 to be important to identifying NREM1 and REM, our activation analysis identified Cluster 3 to be related to predictions for NREM1, NREM2, NREM3, and REM. It is counterintuitive that decreased activation mean corresponded to an increased likelihood of being classified as NREM1. This could be related to the amount of variance accounted for in the NREM1 regression. For NREM2 and REM, higher activation mean and variance corresponded to a higher predicted probability. This could explain why the many misclassified NREM2 and REM samples were assigned to the NREM1 class.

Similar to the filter perturbation analysis findings, Cluster 4 activation mean most strongly related to the Awake class. For Cluster 5, increased activation mean and variance corresponded to higher NREM1 predicted probabilities and lower NREM2 and NREM3 predicted probabilities. For NREM1, this coincides with the filter perturbation analysis findings. However, this finding is a bit more perplexing for NREM2 and NREM3. It is also odd for REM, because our filter perturbation analysis indicated that Cluster 5 perturbation singlehandedly dropped REM performance to nearly 0%. This likely indicates that the relationship between Cluster 5 activations and NREM2, NREM3, and REM predictions were highly nonlinear. Like our filter perturbation analysis, our activation analysis found Cluster 6 to be unimportant to model performance.

While Cluster 7 perturbation had negligible effects on Awake and REM performance, lower Cluster 7 activation mean and variance were strongly related to increased Awake probabilities. This suggests that the model relied on low levels of Cluster 7 activation when identifying Awake, NREM2, and REM. Our filter perturbation analysis likely did not identify Cluster 7 Awake, NREM2, and REM importance because the ablation simply lowered activation values further. Given that our filter perturbation analysis indicated that Cluster 7 was very important for NREM3, it is odd that our activation analysis did not identify a similar relationship.

Cluster 8 mean activation was highly and positively related to NREM3 and moderately related to NREM1 (positive), NREM2 (negative), and REM (negative) predicted probabilities. Cluster 8 activation variance was highly related to NREM1. This finding coincides with the filter perturbation analysis findings for NREM1 and suggests that while the model sought higher levels of Cluster 8 activation in NREM1, it also expected that the presence of Cluster 8 waveforms would vary over time. Our finding for NREM3 could explain why the model assigned many NREM3 samples to NREM2 (i.e., it looked for activity in the β-band related to sleep spindles).

D. Further Discussion on Approach Development

Our approach consisted of multiple key components, some of which build upon existing studies and others that were unique contributions of this study. The first key component of our approach was the use of long first layer filters. These filters enabled the extraction of waveforms that we could visualize both in the time and frequency domains to characterize the patterns being extracted by the model. Multiple previous studies have used long first layer filters for similar purposes [8], [9], [21], [22]. These explainability methods are collectively regarded as model visualization approaches [4]. Most studies directly visualize these filters without identifying commonalities across filters. To make it easier to understand what patterns are collectively being extracted by the model filters, we clustered the frequency domain representations of the filters (i.e., our second key component). A couple previous studies have used a similar filter clustering approach [8], [22]. At this point, our approach diverges from existing approaches. A couple previous studies have used perturbation or gradient-based [39] explainability methods to gain insight into the relative importance of clusters of filters for each class that a model is seeking to identify [8], [22]. Nevertheless, the explainability analyses in those studies provide limited insight into how each filter relates to model predictions. To obtain this level of insight, we added an additional, novel component to our analysis. Namely, we analyzed the relationships between the activations for each cluster of filters and model predictions. By examining linear regression coefficients relating these values, we were able to determine whether the increased presence (i.e., activation mean) or increased variability of presence (i.e., activation variance) of a given cluster of filters corresponded to an increased probability of belonging to a particular sleep stage. It should be noted that this novel analysis of model activations was principally useful because we were previously able to characterize each cluster of model filters.

E. Broader Implications of Approach

Enhanced explainability remains an important next step if deep learning methods are to be applied alongside EEG in a clinical context. While existing feature importance methods identify what features are important, they do not provide insight into how important features relate to model predictions. As such, this is a key next step for the field. Methods that give similar explanations into how feature values relate to predictions like partial dependence plots [40], individual conditional expectation plots [41], and SHAP dependence plots [42] have been developed for tabular data. Future approaches might seek to adapt these methods for insight into key EEG frequency bands by converting time-series into the frequency domain and examining how different Fourier coefficients correspond to predictions. These approaches could foreseeably be applicable

to any deep learning model, regardless of architecture. It should be noted, however, that adaptations of these approaches would be computationally intensive and would likely be unable to provide the same level of insight into time domain features (i.e., waveforms) that our approach provides. This is because one cannot easily scale a waveform to examine the effect upon model predictions. Our activation analysis could foreseeably be applied alongside model visualization approaches in other time-series analysis domains beyond EEG data [43], [44].

F. Limitations and Next Steps

It should be noted that while the extracted mean and variance of the activations accounted for part of the variance of the model predictions for each class, they did not account for all of the variance. While our activation analysis can provide useful insights and represents a helpful approach moving forward, given the non-linearities present in CNNs, they may not present a full picture of model behavior. As such, future studies may find it helpful to use more complex (but still interpretable) regression methods. It might also prove helpful to analyze other properties of the activations that are univariate (e.g., max activation) or multivariate (e.g., correlation). It could also be helpful to examine how the logits of model predictions relate to model activations and compare that with our findings for model predicted probabilities. Also, many applications of EEG require the use of multichannel data (e.g., schizophrenia [3], [16] or major depressive disorder [4], [5] diagnosis). It would be helpful, though technically challenging to extend model visualization explainability methods to multichannel EEG.

IV. Conclusion

Deep learning is increasingly being applied to raw EEG data for automated feature extraction. It presents the possibility of increased model performance but makes explainability much more challenging. Existing EEG explainability methods mainly provide measures of feature importance, meaning that they identify how important features are to model performance but do not identify how those features relate to model predictions. Developing this capability is an important next step for EEG explainability and will be vital if explainability methods are to be included in EEG clinical decision support systems in the future. In this study, we present a novel analysis that seeks to relate activations for well-characterized first layer filters to model predictions in automated sleep stage classification. Our findings frequently conform to existing literature, supporting the viability of our approach. We identified increased activation in filters with properties resembling sleep spindles in NREM2 samples, increased activation in filters with high frequency activity in Awake samples, and increased activation in filters with δ activity in NREM3 samples. Our proposed approach represents a significant step forward for the field of explainability in deep learning for raw EEG data and has the potential to provide many new insights in future studies.

References

[1] 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.

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

[3] 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.

[4] C. A. Ellis, A. Sattiraju, R. L. Miller, and V. D. Calhoun, "A Framework for Systematically Evaluating the Representations Learned by A Deep Learning Classifier from Raw Multi-Channel Electroencephalogram Data," *bioRxiv*, 2023.

[5] C. A. Ellis, A. Sattiraju, R. L. Miller, and V. D. Calhoun, "Novel Approach Explains Spatio-Spectral Interactions in Raw Electroencephalogram Deep Learning Classifiers," 2023.

[6] M. L. Bringas Vega *et al.*, "An Age-Adjusted EEG Source Classifier Accurately Detects School-Aged Barbadian Children That Had Protein Energy Malnutrition in the First Year of Life," *Front. Neurosci.*, vol. 13, Nov. 2019, doi: 10.3389/fnins.2019.01222.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–9, 2020, doi: 10.1186/s12911-020-01332-6.

[15] I. Rojas, G. Joya, and A. Catala, "Deep Learning Using EEG Data in Time and Frequency Domains for Sleep Stage Classification," in *International Conference on Neural Information Processing*, 2017, vol. 10306 LNCS, pp. V–VII. doi: 10.1007/978-3-319-59153-7.

[16] A. Shoeibi *et al.*, "Automatic Diagnosis of Schizophrenia in EEG Signals Using CNN-LSTM Models," *Front. Neuroinform.*, vol. 15, pp. 1–11, 2021, doi: 10.3389/fninf.2021.777977.

[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 *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2021, pp. 1–4.

[18] 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.

[19] C. A. Ellis *et al.*, "Novel methods for elucidating modality importance in multimodal electrophysiology classifiers," *Front. Neuroinform.*, vol. 17, no. March, pp. 1–14, 2023, doi: 10.3389/fninf.2023.1123376.

[20] C. Molnar, *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*, 2018th-08–14th ed. Lean Pub, 2018. [Online]. Available: http://leanpub.com/interpretable-machine-learning

[21] 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

[22] 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.

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

[24] G. AL *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, [Online]. Available: http://circ.ahajournals.org/content/101/23/e215.full

[25] M. S. Mourtazaev, B. Kemp, A. H. Zwinderman, and H. A. C. Kamphuisen, "Age and gender affect different characteristics of slow waves in the sleep EEG," *Sleep*, vol. 18, no. 7, pp. 557–564, 1995, doi: 10.1093/sleep/18.7.557.

[26] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* Washington DC: US Government Printing Office, 1968.

[27] 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.

[28] F. Chollet, "Keras." GitHub, 2015. [Online]. Available: https://github.com/fchollet/keras

[29] M. Abadi *et al.*, "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, 2016, pp. 265–283. doi: 10.1016/0076-6879(83)01039-3.

[30] 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.

[31] 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.

[32] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.

[33] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, 1979.

[34] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Comput. Biol. Med.*, vol. 106, no. December 2018, pp. 71–81, 2019, doi: 10.1016/j.compbiomed.2019.01.013.

[35] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders," *Ann. Biomed. Eng.*, vol. 44, no. 5, pp. 1587–1597, 2016, doi: 10.1007/s10439-015-1444-y.

[36] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, 2017, doi: 10.1109/TNSRE.2017.2721116.

[37] O. J. Ahmed and S. S. Cash, "Finding synchrony in the desynchronized EEG: The history and interpretation of gamma rhythms," *Front. Integr. Neurosci.*, vol. 7, no. JUL, pp. 1–7, 2013, doi: 10.3389/fnint.2013.00058.

[38] S. Takeuchi, T. Mima, R. Murai, H. Shimazu, Y. Isomura, and T. Tsujimoto, "Gamma oscillations and their cross-frequency coupling in the primate hippocampus during sleep," *Sleep*, vol. 38, no. 7, pp. 1085-1091G, 2015, doi: 10.5665/sleep.4818.

[39] 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.

[40] J. H. Friedman, "Greedy Function Approximation : A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[41] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *J. Comput. Graph. Stat.*, vol. 24, no. 1, p. 44=65, 2015, doi: 10.1080/10618600.2014.907095.

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

[43] S. Ha and S. Choi, "Convolutional Neural Networks for Human Activity Recognition using Multiple Accelerometer and Gyroscope Sensors," *2016 Int. Jt. Conf. Neural Networks*, pp. 381–388, 2016, doi: 10.1109/IJCNN.2016.7727224.

[44] M. Kravchik and A. Shabtai, "Detecting Cyber Attacks in Industrial Control Systems Using Convolutional Neural Networks," no. 1, pp. 72–83, 2018, doi: 10.1145/3264888.3264896.