IMPROVING EXPLAINABILITY FOR SINGLE-CHANNEL EEG DEEP LEARNING CLASSIFIERS VIA INTERPRETABLE FILTERS AND ACTIVATION ANALYSIS

Charles A. Ellis, Robyn L. Miller, Vince D. Calhoun

Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS): Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, Georgia, United States

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 and 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.

***Index Terms*—** explainable AI, deep learning, electroencephalography

1. 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 do provide important insights, they are mainly feature importance approaches. Namely, they identify 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 will be insufficient 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 an activation analysis that linearly relates learned features to model predictions.

Early applications of traditional machine learning to electrophysiology generally involved the use of 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. To name a few applications, deep learning methods with raw EEG data have been used for automated sleep state classification [1], [2] and schizophrenia [3] and major depressive disorder diagnosis [4], [5]. While these methods can improve performance and allow for an unbiased approach to the extraction of features via their automated feature extraction capabilities, relative to traditional machine learning methods, they are less explainable.

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], [15], [16], (3) channels [3], [5], [17], (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 provide poor explanations [18]. One alternative approach that has provided useful insights is model visualization (i.e.,

structuring first layer filters such that characteristics of features can be understood) [8], [9], [19], [20].

Unfortunately, both model visualization and perturbation-based approaches typically only provide a measure of feature importance. While this insight can be useful when validating a model, it still leaves much to be desired for model validation. Furthermore, it is insufficient for a clinical setting in which clinicians need to not only be able to explain to patients what features a model relied upon but also explain how 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 by identifying linear relationships between properties of activations for well-characterized filters and model predictions. 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 a pre-existing model visualization 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.

2. METHODS

In this study, we trained a one-dimensional convolutional neural network (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 their relative importance to the model, and performed linear regression to identify relationships between the mean and variance of the activations for each cluster of filters and the model predictions.

2.1. Dataset and Preprocessing

In this study, we used the Sleep Cassette data from the Sleep-EDF Expanded dataset [21] that is publicly available on PhysioNet [22]. Because the dataset was publicly available, Institutional Review Board approval at our institution was unnecessary. The dataset consisted of 20-hour recordings from 78 participants. All but 3 participants, which had 1 recording, had 2 recordings. Similar to previous studies, we used data from the FPz-Cz electrode that was recorded at 100 Hertz [1], [2]. A trained expert assigned 30-second segments of data to different sleep stages: Awake, rapid eye movement (REM), Non-REM 1 (NREM1), NREM2, NREM3, NREM4.

Data from NREM3 and NREM4 were combined into a single NREM3 class based on clinical guidelines [23]. Due to the length of the recordings, the dataset was heavily imbalanced towards the Awake class. To rectify this issue, we removed Awake samples at the start of each recording and many Awake samples at the end of each recording. Prior to separating the data into 30-second segments, we z-scored each recording separately. After preprocessing, our dataset consisted of 85,034 Awake, 21,522 NREM1, 69,132 NREM2, 13,039 NREM3, and 25,835 REM samples.

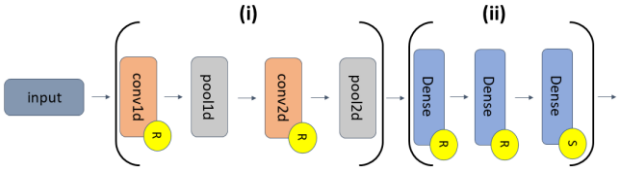

Figure 1. CNN Architecture. The model has two components: a feature extraction portion (i) and a classifier portion (ii). Component (i) has a 1D convolutional layer (conv1d) with 30 filters of length 200 (i.e., approximately 2 seconds) and a stride of 1, a 1D max pooling layer (pool1d) with a pooling 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 correspond to activation functions: ReLU (R) and softmax (S).

2.2. Model Development

We adapted the 1D-CNN architecture initially developed in [19] and later used in [8], [20]. 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, 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. Figure 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 [24]. 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 mean F1 score across classes. In the fold used for explainability, we also computed a confusion matrix.

2.3. Filter Clustering and Visualization

We sought to characterize the features extracted by the model. To this end, we converted all first layer filters to the frequency domain using a fast Fourier transform (FFT) and

Table 1. Model Performance Results.

	Awake	NREM1	NREM2	NREM3	REM
F1	91.60±02.09 (92.65)	41.04±03.51 (41.75)	79.91±02.80 (82.74)	68.82±12.79 (87.11)	68.35±05.37 (77.58)
Recall	88.18±03.54 (89.16)	45.39±05.45 (45.10)	77.83±04.28 (81.37)	63.98±13.37 (91.58)	71.93±06.01 (81.37)
Precision	95.40±01.84 (96.43)	38.22±05.89 (38.87)	82.22±02.14 (84.16)	75.85±14.59 (83.05)	65.80±08.67 (74.13)

extracted their spectral power. We then applied k-means clustering with 50 initializations on the spectral power values, sweeping from 2 to 15 clusters and selecting the number of clusters with the maximal silhouette value [25]. Upon identifying an optimal number of clusters, we visualized the filters in each cluster in both the time and frequency domains. Our use of clustering greatly simplified subsequent analyses by allowing us to analyze a smaller number of groups of filters with similar features rather than having to analyze all filters individually.

2.4. Filter Perturbation Analysis

We next sought to identify the relative importance of each cluster of filters using a perturbation approach. To this end, we calculated the weighted recall 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 of filters. This analysis is adapted from [8] and provides an important point of comparison for our novel filter activation analysis by relating it to existing approaches.

2.5. Filter Activation Analysis

While our earlier analyses characterized the features extracted by the model and the relative 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. To that end, we extracted two values for each test sample. We calculated 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 corresponding to each class separately. We next performed 10 sets of 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 through 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).

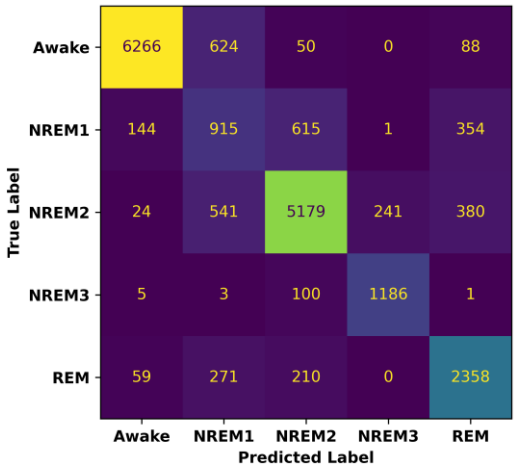

Figure 2. Top Fold Confusion Matrix. True and predicted labels are shown on the x- and y-axes, respectively. Lighter colors indicate that a larger portion of samples reside within a classification group.

3. 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.

3.1. 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). Figure 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, and some studies have specifically developed methods with the goal of increasing NREM1 performance [26]. Model performance for the fold used in explanations was generally above the mean performance across folds.

3.2. Identifying Extracted Features and Their Relative Importance

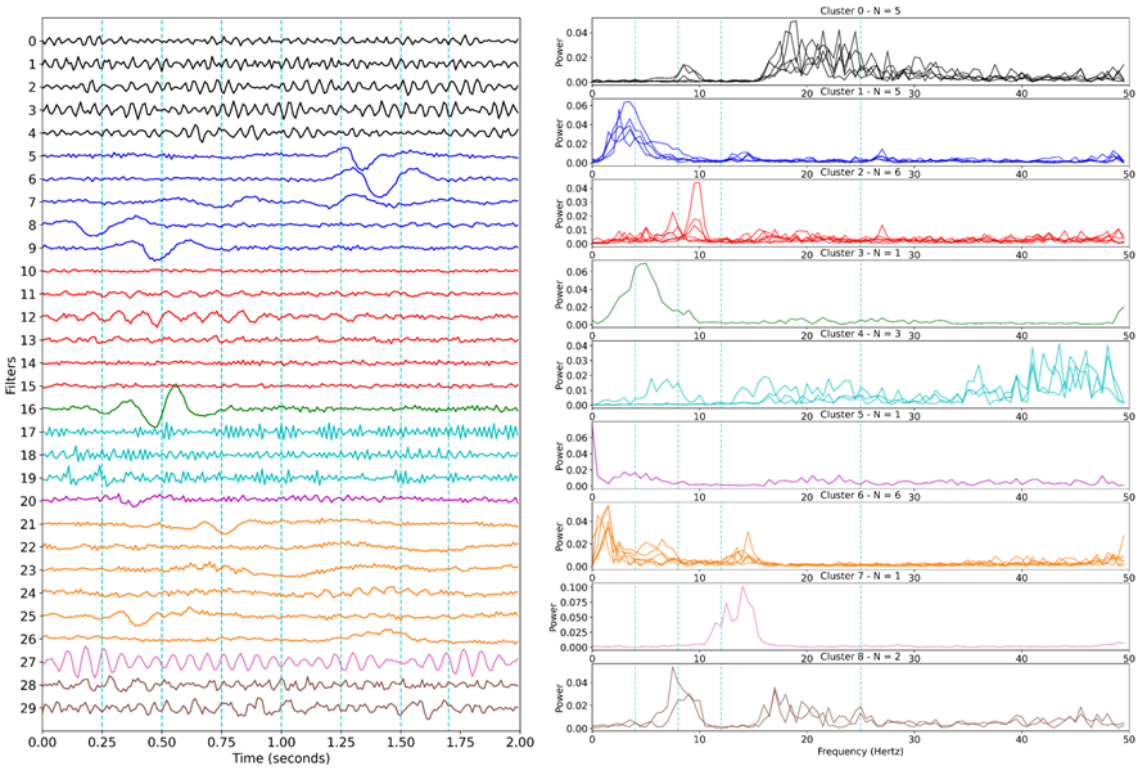

Figure 3. Filter Waveforms and Spectral Clusters. The left panel shows the filters generated by the CNN. Clusters of filters are arranged based on color. Those colors correspond to the colors of the rightmost panel, which shows the spectral power of filters in each cluster. The title for each subpanel 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 leftmost panel indicates spectral power. Filters in the leftmost panel are min-max normalized to enable easier visual inspection of the waveforms.

We identified 9 clusters of first layer filters with unique spectral properties. Figure 3 shows the filter clusters in both the time and frequency domains, and Figure 4 shows the perturbation-based relative importance of each cluster of filters to the model performance. Cluster 0 was characterized primarily by upper β activity, with some lower α activity and was extremely important to the identification of NREM1. It seemed to have waveforms similar to vertex sharp waves, which are found in NREM1 [23]. Cluster 1 primarily extracted δ activity with some lower θ activity, and its waveforms seem to resemble k-complexes. It played a larger role in the identification of NREM2 and REM than many other clusters. Clusters 2 and 3 primarily extracted α and a mix of δ and θ activity, respectively. Perturbation of cluster 2 had a relatively small effect upon classifier performance relative to other clusters, and cluster 3 was extremely important to the identification of REM. Cluster 4 extracted γ activity and some θ and β activity, but its perturbation seemed to have a relatively small effect on model performance. Cluster 5 seemed to extract relatively little information, though it did seem to extract some δ and θ activity, and its perturbation had a relatively small effect on model performance. Cluster 6 extracted δ activity, and similar to cluster 7, it also extracted some lower β activity. Cluster 6 was instrumental to the identification of NREM3, which is characterized by δ activity [23]. Cluster 7 was highly important for the identification of NREM2, possibly because it extracted waveforms in a frequency range similar to sleep spindles [23]. Cluster 8 extracted upper θ and lower α activity, along with upper β activity. Its perturbation had a relatively low impact upon model performance.

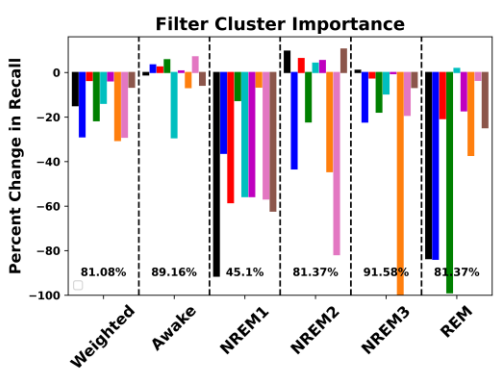

Figure 4. Cluster Perturbation Importance. The figure shows the percent change in weighted and class-specific recall after filters in each cluster were replaced with zeros. The y-axis shows the change in recall. The x-axis shows the corresponding class. Within each class, each cluster is shown from left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

3.3. Identifying Linear Relationships Between Activations and Model Predictions

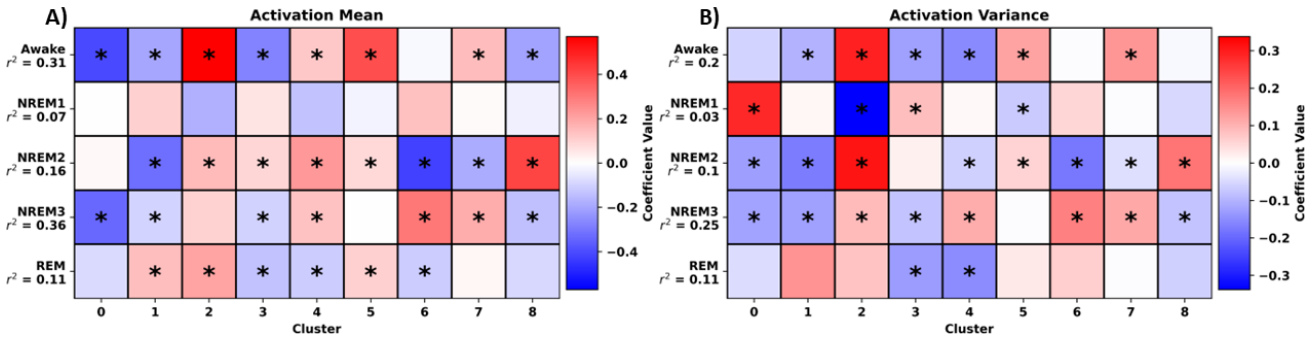

Figure 5. Linear Regression Results Relating Mean and Variance of Activations to Model Predicted Probabilities. Panels A and B show results for the activation mean and variance, respectively. The x-axis indicates the class and the corresponding R^2 for the analysis. The y-axis indicates the corresponding cluster. The heatmap shows coefficient values, and asterisks indicate a significant p-value following FDR correction ($p < 0.05$).

Figure 5 shows the linear regression coefficients, FDR-corrected p-values, and R^2 values resulting from our analysis relating the mean and variance of the model test activations to model predictions. The regression analyses had the highest R^2 for the NREM3 class, followed by the Awake class, for both the mean and variance of the activations. The R^2 for the mean of the activations tended to be higher than for the variance, and across both measures, the R^2 for NREM2 and REM were significant but more modest than those of NREM3 and Awake. The NREM1 R^2 was extremely small.

When analyzing explanations, we compare them to the American Academy for Sleep Medicine (AASM) manual [23]. Similar to the filter perturbation analysis, cluster 0 activation variance was positively related to a sample being classified as NREM1. It should be noted, however, that while the coefficient value was significant with correction ($p < 0.05$), though the R^2 was quite low. Cluster 1 mean activation was negatively associated with NREM2 and positively associated with REM predicted probabilities, though it was not significantly related to NREM2 ($p > 0.05$). Additionally, cluster 1 activation variance had a significant negative association with NREM2 predicted probabilities. Cluster 2, which had activity in frequencies similar to sleep spindles, had a significant relationship with NREM2. Interestingly, cluster 3 activation was negatively correlated with the REM predicted probability. It would be expected that the relatively large low frequency waveform present in cluster 3 would have low activations for REM. Cluster 4 had high a positive relationship with Awake predicted probability, which typically has high frequency activity. Interestingly, cluster 5 mean activation was highly related to Awake predicted probability. Cluster 5 does have some low amplitude, higher frequency waveforms, so it is possible that the model relied upon those for identifying that class. Cluster 6 activation mean and variance was positively associated with NREM3 predicted probabilities and negatively associated with NREM2 predicted probabilities, which makes sense given the importance of δ activity to NREM3 [23]. Oddly, cluster 7 activation, which based on our perturbation analysis was highly important to NREM2 and seemed to have sleep spindles, was actually inversely related to NREM2 predicted probabilities. This is possibly because the model relied upon complex non-linear relationships with other features when using cluster 7 for identifying NREM2. Cluster 8 was also potentially positively related to NREM2. This could be explained by the fact that cluster 8 did have some activity at around 11 Hz, which is the lower boundary of sleep spindles.

3.4. 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 the linear regression activation analysis can provide useful insights and represents a helpful analysis moving forward, given the non-linearities present in CNN, they may not present a full picture of model behavior. In the future, it might prove helpful to analyze other properties of the activations that are univariate (e.g., max activation) or multivariate (e.g., correlation). Additionally, many applications of EEG require the use of multichannel data (e.g, schizophrenia or major depressive disorder diagnosis). It would be helpful, though technically challenging to extend model visualization explainability methods to multichannel EEG.

4. 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 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.

5. ACKNOWLEDGEMENTS

This work was funded NIH R01MH123610, NIH R01MH118695, and NSF 2112455.

11. 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," in *bioRxiv*, 2023, pp. 2–6.

[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] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.