A Gradient-based Approach for Explaining Multimodal Deep Learning Classifiers

Charles A. Ellis
Tri-institutional Center for Translational Research in Neuroimaging and Data Science
Georgia State University, Georgia Institute of Technology, and Emory University
Atlanta, USA
cae67@gatech.edu

Rongen Zhang
Department of Computer Information Systems
Georgia State University
Atlanta, USA
rzhang6@gsu.edu

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

Darwin A. Carbajal
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, USA
darwin.carbajal@gatech.edu

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

May D. Wang
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, USA
maywang@gatech.edu

2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE) | 978-1-6654-4261-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBE52308.2021.9635460

***Abstract*—In recent years, more biomedical studies have begun to use multimodal data to improve model performance. Many studies have used ablation for explainability, which requires the modification of input data. This can create out-of-distribution samples and lead to incorrect explanations. To avoid this problem, we propose using a gradient-based feature attribution approach, called layer-wise relevance propagation (LRP), to explain the importance of modalities both locally and globally for the first time. We demonstrate the feasibility of the approach with sleep stage classification as our use-case and train a 1-D convolutional neural network with electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) data. We also analyze the relationship of our local explainability results with clinical and demographic variables to determine whether they affect our classifier. Across all samples, EEG is the most important modality, followed by EOG and EMG. For individual sleep stages, EEG and EOG have higher relevance for awake and non-rapid eye movement 1 (NREM1). EOG is most important for REM, and EEG is most relevant for NREM2-NREM3. Also, LRP gives consistent levels of importance to each modality for the correctly classified samples across folds but inconsistent levels of importance for incorrectly classified samples. Our statistical analyses suggest that medication has a significant effect upon patterns learned for EEG and EOG NREM2 and that subject sex and age significantly affects the EEG and EOG patterns learned, respectively. Our results demonstrate the viability of gradient-based approaches for explaining multimodal electrophysiology classifiers and suggest their generalizability for other multimodal classification domains.**

Keywords*—*Explainability, Multimodal Fusion, Automated Sleep Staging, Electrophysiology

I. Introduction

A growing number of biomedical studies have begun incorporating data from multiple modalities [1]–[4]. This growth has occurred because complementary modalities enhance data richness and can improve classifier performance [2]. However, multimodal data also increases the difficulty of model interpretation, and most multimodal studies have not incorporated explainability methods that could provide insight into the relative contributions of each modality [3]. In this study, we present a novel application of Layer-wise Relevance Propagation (LRP) [5], a member of the gradient-based feature attribution (GBFA) family [6] of explainability methods, for explaining multimodal deep learning classifiers.

In recent years, a few studies used explainability methods like forward feature selection (FFS) [1], impurity [4], and ablation [4][7] to identify the relative importance of each modality to classifiers trained on multimodal data. However, some of these methods are incompatible with high-performing deep learning frameworks. For example, FFS requires retraining classifiers repeatedly, so it is impractical for computationally intensive deep learning classifiers. Furthermore, impurity methods are only applicable to decision tree-based models.

Unlike FFS and impurity, ablation is applicable to most classifier types, is relatively easy to implement, and is not computationally intensive. Different variations on ablation also yield both global [8] and local [9] explanations. Global methods explain modality importance to the overall classifier, and local methods explain modality importance to the classification of individual samples. The relationships of local explanations with demographic and clinical variables also provide insight into the effects of those variables upon the classifier [9]. However, despite ablation's abovementioned advantages, it also has key weaknesses. Like perturbation, ablation requires data modification. This modification can create samples that are out of the data distribution upon which the classifier was trained [10]. Moreover, in deep learning classifiers with automated feature extraction, ablation can cause extracted features to be outside the distribution of features extracted from the dataset. This potential creation of out-of-distribution samples or features can lead to incorrect explanations. Furthermore, ablation identifies how model performance changes when the information in a modality becomes unavailable. As such, it is prudent to be cautious while adapting such methods to new domains [8]. In domains like electrophysiology (EP), a value of zero for a modality is abnormal and would likely be too dissimilar to real-life samples, which could adversely affect explainability results.

978-1-7281-6338-3/21/$31.00 ©20XX IEEE

Like ablation, GBFA methods produce both local and global explanations. In contrast to FFS and ablation, GBFA methods offer multimodal time-series explainability without modifying data and are applicable to many deep learning frameworks. LRP is a popular GBFA method [5] for insight into classifiers trained on multimodal data. To demonstrate the viability of this approach, we train a 1-dimensional (1D) convolutional neural network (CNN) for automated sleep stage classification with electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) data from a popular online dataset [11]. We use automated sleep stage classification because it is a representative multimodal classification task with clinical needs for model explainability [7]. We apply LRP locally for insight into the classification of samples over time and globally to show the relative importance of each modality to each classification group. We further examine the relationship between the local explainability results and demographic and clinical variables to determine whether the classifier learned patterns associated with those variables.

II. Methods

Here we provide a description of our methods. Using multimodal data, we train a CNN to discriminate between each sleep stage and apply LRP to explain the decisions of the classifier. The dataset, preprocessing, and classifier that we use are the same as those which we presented in [8]. The key innovation of this work is its explainability approach.

A. Description of Data

We use the Sleep Telemetry dataset from the Sleep-EDF Database [11] on Physionet [12]. The dataset includes 44, approximately 9-hour recordings from 22 subjects (7 male and 15 female). Each subject has a recording following the administration of a placebo and temazepam. Subject age has a mean of 40.18 years and a standard deviation of 18.09 years. The dataset consists of EEG, EOG, and EMG with a sampling rate of 100 Hertz (Hz), as well as a polysomnogram of the sleep stage at each time point. For EEG, we use the FPz-Cz electrode. Sleep stages include: awake, movement, rapid eye movement (REM), non-REM 1 (NREM1), NREM2, NREM3, and NREM4. A marker at 1 Hz intervals indicates whether an error occurred in the sleep telemetry device.

B. Description of Data Preprocessing

We separate each recording into 30-second segments and extract the corresponding labels from the polysomnogram. We discard movement samples and samples that correspond with recording errors, and we consolidate the NREM3 and NREM4 stages into NREM3 [13]. We then z-score each modality within each recording. The dataset has 42,218 samples and is very imbalanced. Awake, NREM1, NREM2, NREM3, and REM stages compose 9.97%, 8.53%, 46.8%, 14.92%, and 19.78% of the dataset, respectively. Data preprocessing was performed in MATLAB R2020b [14].

C. Description of CNN

We adapt a 1D-CNN architecture originally developed for EEG-based sleep stage classification to our multimodal dataset [15], and implement it in Keras 2.2.4 [16]. The architecture, model hyperparameters, and training approach are described in [8]. We use 10-fold cross validation with training, validation, and test sets composed of 17, 2, and 3 randomly assigned subjects, respectively. To measure classifier performance, we generate a confusion matrix showing the distribution of sample classification across all folds. Further details on the precision, recall, and F1 score of the classifier are included in [8].

D. Description of Explainability Approach

We use LRP to explain the relative importance of each modality [5]. LRP provides local explanations for the classification of each individual sample. In LRP, a sample is fed into the neural network and classified. A total relevance of 1 is assigned to the output node for its respective class, and that total relevance is propagated back through the network via relevance rules until a portion of that total relevance is assigned to each of the points in the input sample. Both positive and negative relevance propagate through the network. Positive relevance shows the features that support the assignment of a sample to the class to which it is assigned. Negative relevance identifies the features that support the sample being assigned to other classes. We used the ε and $\alpha\beta$ relevance rules [17]. The parameter ε in the ε-rule enables relevance to be filtered when propagated through the network. Increasing ε filters out smaller relevance values, reducing the noise in the explanation. The parameters α and β in the $\alpha\beta$-rule control the degree to which positive and negative relevance are propagated through the network, respectively. While the ε-rule propagates both negative and positive relevance, the $\alpha\beta$-rule can propagate only positive relevance (i.e., when $\alpha = 1$ and $\beta = 0$). We use the ε-rule ($\varepsilon = 0.01$ and 100) and the $\alpha\beta$-rule ($\alpha = 1$, $\beta = 0$). The ε-rule ($\varepsilon = 0.01$) enables an approximately baseline examination of modality importance where both positive and negative relevance are propagated and little relevance is filtered. The ε-rule ($\varepsilon = 100$) explains the model with less noise, and the $\alpha\beta$-rule enables an understanding of modality importance that only indicates which modality is important to the classified sleep stage. To obtain a "global" explanation, we combine the local explanations for all samples in the test set of each fold. For each fold, we calculate the percent of absolute relevance assigned to each modality across samples to identify their relative importance. We do this for all test samples and for each classification group (e.g. awake classified as awake or NREM1 classified as NREM2).

E. Description of Statistical Analyses

We perform statistical analyses to examine the effects of clinical (i.e., medication) and demographic (i.e., age and sex) variables upon the classifier. For each correct classification group (e.g., awake classified as awake, NREM1 classified as

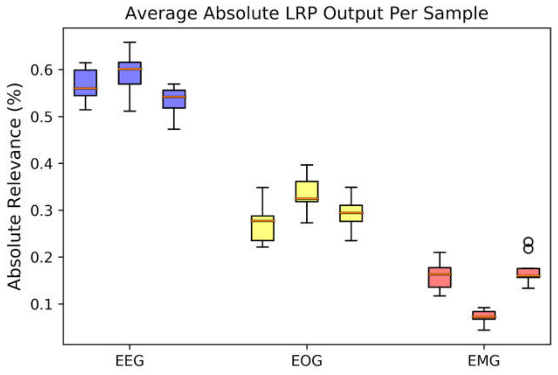

Fig. 1. LRP-based global explainability. Plot shows explainability results for all folds. Blue, yellow, and red boxes are for EEG, EOG, and EMG, respectively. Within each trio, from left to right are relevance results for the LRP ε-rule (0.01), ε-rule (100), and α-β-rule.

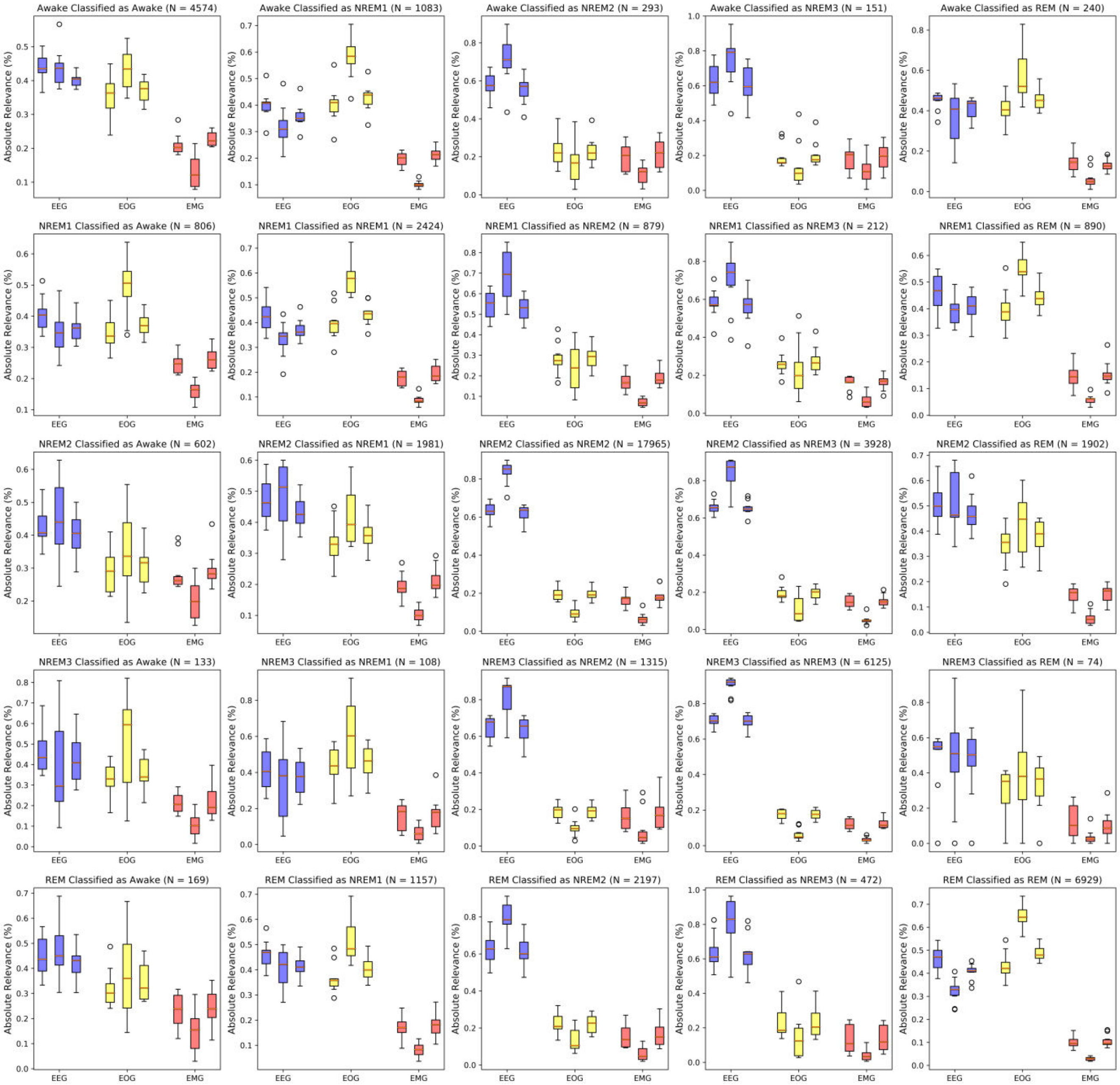

Fig. 2. LRP Results for Each Classification Group. Plot shows explainability results for all folds. Blue, yellow, and red boxes are for EEG, EOG, and EMG, respectively. Each trio, from left to right, shows relevance results for the LRP ε-rule (0.01), ε-rule (100), and αβ-rule. The title of each panel indicates the number of samples in each classification group. The panels on the left-to-right diagonal show correctly classified groups. The panels off the left-to-right diagonal show incorrectly classified groups.

NREM1) and modality, we train an ordinary least squares regression model with sample age, sex, and medication as independent variables and the percent of absolute relevance belonging to a modality as the dependent variable. We then examine the resulting coefficient and p-values. The use of regression enables us to examine the effects of each clinical or demographic variable while controlling for the effects of the other variables. We use false discovery rate correction (α = 0.05) to correct for multiple comparisons.

III. Results and Discussion

Here we describe and discuss the LRP results. We also discuss the study limitations and potential future work.

A. *Global Explainability Results*

Fig. 1 and 2 show LRP results for all samples and for each classification group, respectively. When all classes are considered, all LRP rules indicate that EEG is the most important modality, followed by EOG, and EMG. For an ε of 100, when low relevance values are filtered out, the EEG and EOG show an increase in importance while EMG importance decreased. Interestingly, both the ε-rule (ε=100) and the αβ-rule give more importance to EMG than the ε-rule (ε=0.01). These results are comparable to the ablation results in [8], though the importance of EOG and EMG appears to be greater for LRP than for ablation.

We also examine the importance of each modality for the correct or incorrect classification of each class. The results in Fig. 2 fit with sleep scoring guidelines, as EEG differentiates all stages while EOG and EMG differentiate awake, REM, and NREM samples [13]. The diagonal of Fig. 2 shows the LRP results for correctly classified samples. For the awake stage, the CNN relies mostly on EEG and EOG data. For correctly classifying NREM1, the CNN model places importance either more on EOG than EEG (ε=100) or about equally on EEG and EOG (ε=0.01 and αβ-rule). However, EMG is the least relevant in correctly classifying NREM1 for all relevance rules. For correctly classifying NREM2 and NREM3, EEG has more than 3 times as much relevance as EOG and EMG for all rules. This coincides with sleep scoring guidelines, as EEG in NREM samples is often very distinct [13]. For correctly classified REM samples, EOG is the most relevant (ε rule with ε=100 and α-β rule). However, for ε-rule (ε=0.1), EEG and EOG are equally relevant for REM classification. EMG is the least relevant for REM classification across all rules. Our results corroborate the well-documented importance of EOG in classifying awake and REM. EOG is important because it tracks eye movements which are more common during awake and REM stages. We also note that the relevance across folds of the test samples that the CNN correctly classified had lower variance than the that of the misclassified samples. The lower variance of the correctly classified samples indicates that the features learned by the CNN for correct classification are likely similar across all 10 folds, which could indicate that the architecture is learning generalizable features or that the subjects randomly assigned to each test group are comparable. However, the greater variance in relevance across folds for the incorrectly classified samples could indicate that the classifiers make different mistakes in each fold or identify different ungeneralizable patterns in the training data.

B. Local Explainability Results

Fig. 3 shows LRP local explainability results (ε=100) for 120 minutes from Subject 12. The figure shows a full sleep cycle from awake through REM. The top panel shows the class assigned to each sample by the CNN model (green line) and the class to which the sample actually belonged (black line). The middle panel shows the electrophysiological recordings of the different modalities: EEG in blue, EOG in yellow, and EMG in red. The bottom panel shows the portion of absolute relevance (%) assigned to each modality (same color scheme as the middle panel) when the CNN made its classification decision. This indicates the relative importance of each modality when classifying each sleep stage. In the first approximately 17 minutes, the model properly predicted and classified the awake stage by heavily relying upon EOG. This makes sense given that EOG tracks eye movements, which both Awake and REM states produce, and also explains why the model places high relevance on EOG for REM classification between 100-120 minutes. The local LRP results also show that the model heavily relied on EEG when classifying NREM1, NREM2, and NREM3 samples. However, the model misclassified NREM1-NREM3 (e.g., at around 70 - 80 minutes) when it placed greater importance on other modalities like EMG and EOG (red EMG and yellow EOG spikes on middle and bottom panel)

C. Statistical Analysis Results

Fig. 4 shows the statistical analysis results for the ordinary least squares regression. Interestingly, in the majority of instances, the local explainability results have statistically significant relationships with medication, sex, and age. As such, the variables likely affect the EEG, EMG, and EOG data across sleep stages, and the classifier likely learns patterns related to these variables. Interestingly, for Awake samples, temazepam administration is correlated with a decrease in absolute relevance assigned to EEG and a corresponding increase in absolute relevance assigned to EMG. This change

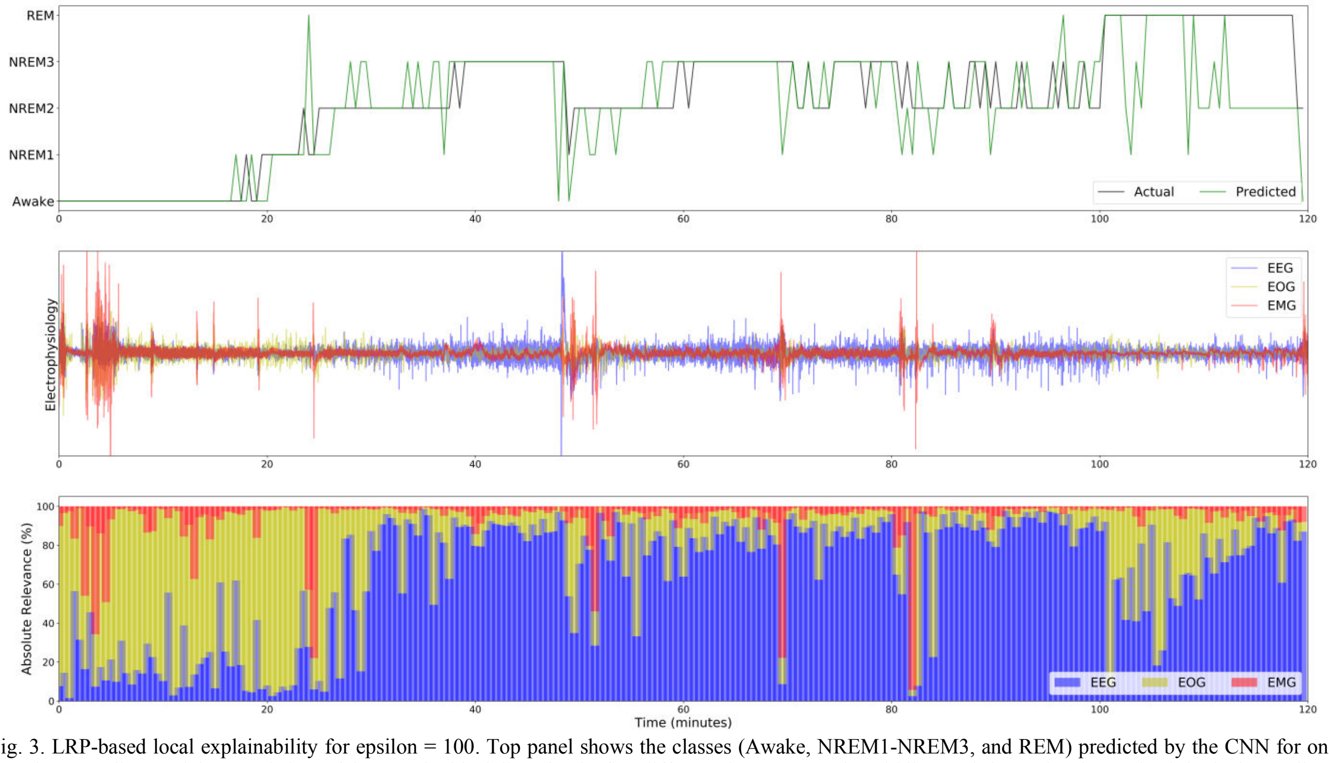

Fig. 3. LRP-based local explainability for epsilon = 100. Top panel shows the classes (Awake, NREM1-NREM3, and REM) predicted by the CNN for one sample (green line) and the actual class of the sample (black line) for the five different sleep stages. The middle panel shows the electrophysiological recordings of the different modalities: EEG in blue, EOG in yellow, and EMG in red. The bottom panel shows the importance that each modality has to the identification of each sleep stage for the CNN model.

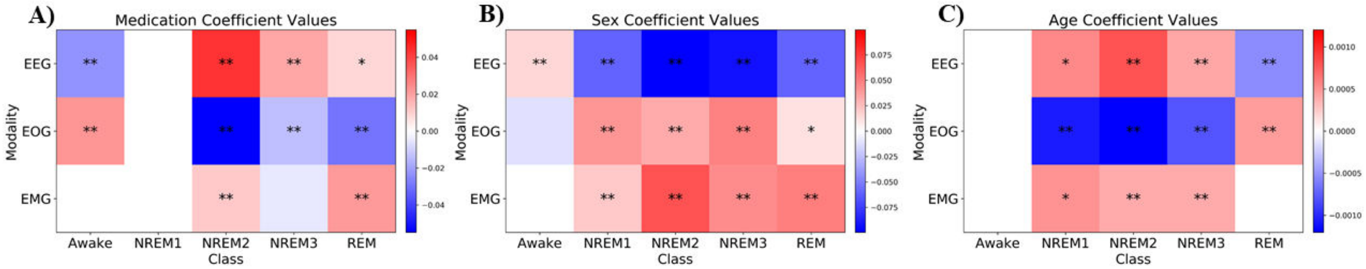

Fig. 4. Statistical Analysis Results for ε = 100. Panels A, B, and C show the results of the statistical analyses for medication, sex, and age. On the x-axis of each panel is the class associated with each column, and on the y-axis is the modality associated with each row. The heatmap values show the resulting model coefficients. The color bar to the right of each panel indicates the scale of the coefficient values. Blocks colored white did not have significant corrected p-values (i.e., $p > 0.05$). Colored blocks without an asterisk are significant with a $p < 0.05$ after FDR correction. Colored blocks with one or two asterisks have p-values of $p < 0.01$ and $p < 0.001$ after FDR correction, respectively. A positive coefficient for medication indicates an increase in absolute relevance from subjects with placebo to subjects with temazepam. A positive coefficient for sex indicates an increase in absolute relevance from female subjects to male subjects. A positive coefficient for age indicates an increase in absolute relevance with an increase in age.

in importance following temazepam administration is particularly pronounced for NREM2 where absolute EEG relevance increases and absolute EOG relevance decreases. Also, male subjects tend to have less absolute EEG relevance and modest increases in EMG and EOG relevance relative to female subjects across all stages (except Awake). This relationship is very pronounced in NREM2. The Awake stage had slightly more EEG relevance in males than females but no differences in EMG and EOG relevance. Age does not have significant effects on modality importance for the Awake stage. However, age affects most other modalities and sleep stages. For age, the amount of absolute relevance assigned to EEG increases most with age in NREM2 and corresponds with a decrease in EOG absolute relevance. More EEG and EMG relevance for NREM1 and NREM3 also correspond with less EOG relevance. In summary, medication most strongly affects the patterns learned by the classifier for EEG and EOG data in NREM2. Subject sex strongly affects the patterns learned for EEG data, and subject age most strongly affects patterns learned for NREM EOG data.

D. Limitations and Future Work

LRP is a member of a broad class of GBFA methods. As such, other GBFA methods could potentially provide better explanations. Some metrics quantify the quality of explanations produced by different explainability methods [18], and those metrics could potentially identify the methods that provide the highest quality explanations for multimodal EP time-series. Furthermore, when we output LRP results, we use each rule for propagating relevance through the whole network. Another study showed that using different rules in different parts of a CNN can improve explanations, especially in deeper networks [17]. For our statistical analyses, we only analyze the statistical relationships of demographic and clinical variables with local explanation results for correctly classified samples. Future studies could analyze the relationship of variables with local explanations for incorrectly classified samples to better understand how the variables may adversely affect classification performance. Also, because we adapt a CNN architecture originally developed for EEG-based sleep stage classification, the architecture may not optimally extract features from EOG and EMG. This could cause EEG to be overutilized by the classifier and could explain the relatively high level of importance that our explainability approach found for the modality. We emphasize that this would not in any way invalidate the explainability results, as they would still reveal what the classifier prioritizes. The broader scientific implications of the explainability analysis might just be limited. As such, examining model architectures that might better extract EOG or EMG features could be helpful. Also, although our classification performance is marginally below the state of the art, our novel explainability approach, rather than our classifier, is the focus of our study. Using LRP with a better classifier could provide more generalizable explanations and contribute to novel biomarker identification.

IV. Conclusion

In this study, we implement a gradient-based model-introspection technique in a novel application for insight into the importance of each modality in multimodal EP data. Our approach offers an advantage over the popular ablation approaches that have previously been used to find the relative importance of each modality to a classifier. Specifically, our approach does not require modifying data, which sometimes causes ablation to yield inaccurate explanations. Also, similar to ablation, our approach provides both local and global explainability. Because of its well-characterized clinical guidelines and need for explainability, we use sleep stage classification as a test bed and train a classifier to discriminate between sleep stages using multimodal data. We implement LRP, a popular GBFA method, to identify the relative importance of each modality to the CNN. Our results corroborate documented findings on the importance of EEG and EOG in classifying Awake and NREM1, EOG for REM, and EEG for NREM2-NREM3. Interestingly, the CNN gives consistent levels of importance to each modality for correctly classified samples across folds and inconsistent importance for incorrectly classified samples. Our statistical analysis of clinical and demographic variables indicates that the variables likely affect the patterns learned by the classifier for all modalities and most sleep stages. Our study provides a novel approach for explaining multimodal EP classifiers. Further, our approach has the potential to be applied to multimodal classification problems in other domains.

Acknowledgment

We thank Felipe Giuste and Wenqi Shi for their advice and assistance with computational resources for this project. We thank Mohammad Sendi for helping edit the paper.

References

[1] M. S. Mellem, Y. Liu, H. Gonzalez, M. Kollada, W. J. Martin, and P. Ahammad, “Machine Learning Models Identify Multimodal Measurements Highly Predictive of Transdiagnostic Symptom Severity for

Mood, Anhedonia, and Anxiety," *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, vol. 5, no. 1, pp. 56–67, Jan. 2020, doi: 10.1016/j.bpsc.2019.07.007.

[2] B. Zhai, I. Perez-Pozuelo, E. A. D. Clifton, J. Palotti, and Y. Guan, "Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, 2020, doi: 10.1145/3397325.

[3] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2019, doi: 10.1109/TBME.2018.2872652.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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," *bioRxiv*, pp. 0–3, 2021.

[9] C. A. Ellis *et al.*, "A Novel Local Ablation Approach For Explaining Multimodal Classifiers," *bioRxiv*, pp. 1–6, 2021.

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

[11] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, 2000, doi: 10.1109/10.867928.

[12] 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.

[13] 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.

[14] "MATLAB version 2020b." The Mathworks Inc., Natick, Massachusetts, 2020.

[15] M. Youness, "CVxTz/EEG_classification: v1.0," 2020. https://github.com/CVxTz/EEG_classification (accessed Jan. 05, 2021).

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

[17] 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.

[18] 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.