Effectively Interpreting Electroencephalogram Classification Using the Shapley Sampling Value to Prune a Feature Tree

Kazuki Tachikawa, Yuji Kawai, Jihoon Park and Minoru Asada

Graduate School of Engineering, Osaka University,
2-1 Yamadaoka, Suita, Osaka, 565-0871, Japan.
{kazuki.tachikawa,kawai,jihoon.park,asada}@ams.eng.osaka-u.ac.jp

Abstract. Identifying the features that contribute to classification using machine learning remains a challenging problem in terms of the interpretability and computational complexity of the endeavor. Especially in electroencephalogram (EEG) medical applications, it is important for medical doctors and patients to understand the reason for the classification. In this paper, we thus propose a method to quantify contributions of interpretable EEG features on classification using the Shapley sampling value (SSV). In addition, a pruning method is proposed to reduce the SSV computation cost. The pruning is conducted on an EEG feature tree, specifically at the sensor (electrode) level, frequency-band level, and amplitude-phase level. If the contribution of a feature at a high level (e.g., sensor level) is very small, the contributions of features at a lower level (e.g., frequency-band level) should also be small. The proposed method is verified using two EEG datasets: classification of sleep states, and screening of alcoholics. The results show that the method reduces the SSV computational complexity while maintaining high SSV accuracy. Our method will thus increase the importance of data-driven approaches in EEG analysis.

Keywords: electroencephalogram (EEG), Shapley sampling value (SSV), convolutional neural networks (CNN)

1 Introduction

Deep learning, especially via convolutional neural networks (CNNs), is a promising method of classification of electroencephalogram (EEG) signals. CNNs enable identification of brain states from raw EEG signals and provide higher classification accuracy than conventional machine learning techniques [6, 11]. However, understanding how the models classify the signals is difficult because CNNs have highly complex nonlinear functions. Visualization of features, which contribute to their classification, may engender neurophysiological insights and explanations that can be applied to medical diagnoses.

To identify the interpretable features of EEG, bandpass filters are often applied to EEG signals in standard EEG analysis to separate them into five frequency bands: delta,

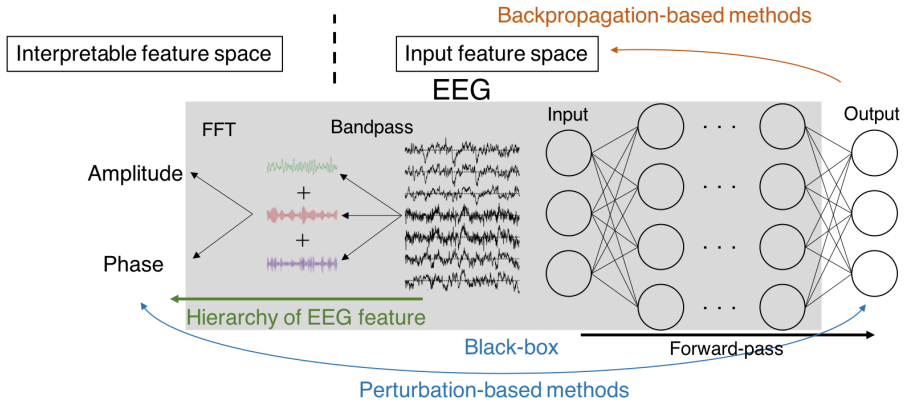

Fig. 1. Overview of methods to interpret EEG classification and a hierarchy of EEG features.

theta, alpha, beta, and gamma. Waves in each band are then analyzed in terms of their amplitude and phase. Therefore, it is useful to quantify the contributions of amplitude and phase in a specific frequency band in the EEG classification.

Various methods to interpret classification of learning models have been proposed. They can be categorized into two groups: *backpropagation-based methods,* and *perturbation-based methods* [1]. Backpropagation-based methods, including layer-wise relevance propagation (LRP) [3], deep learning important features (DeepLIFT) [13], integrated gradients (IG) [16], and deep Shapley additive explanations (SHAP) [8], compute the contributions of all input features in accordance with the backpropagation of class information from an output layer to an input one (as denoted in orange characters in Fig. 1). Their computational cost is relatively low. However, these methods show the contributions only in the input feature space, which is not always interpretable.

In contrast, perturbation-based methods, including the Shapley sampling value (SSV) [14], local interpretable model-agnostic explanations (LIME) [10], and kernel SHAP [8], regard the classifier as a black-box, i.e., they compute the contributions based on pairs of a perturbed (masked or permuted) input and its output (as denoted in blue characters in Fig. 1). These methods can display the contributions in a space representing the perturbation, which differs from the input feature space. By perturbing the classifiers in an interpretable way, we can obtain the contributions in an interpretive feature space. However, the computational cost becomes drastically higher as the number of features increases.

For visualization of signals contributing to EEG classification, several approaches have been proposed. Sturm et al. [15] applied LRP to EEG classifiers. However, LRP cannot directly reflect the contributions in the amplitude-phase form because it is based on backpropagation. Schirrmeister et al. [11] statistically analyzed EEG classifiers using two methods: input-feature unit-output correlation maps (IFUOCM) and input-perturbation network-prediction correlation maps (IPNPCM). IFUOCM computes correlations between the values of output neurons and the input powers of each frequency. IPNPCM calculates correlations of the output values with variations of the perturbed

amplitude or phase [5]. However, the correlations do not always exactly reflect the impact of individual features on the prediction. In addition to the above three methods, two other methods were proposed in [18] and [7], respectively. However, they require modifying input features or specifying the classifier architecture.

Recently, IG and SHAP were shown to be theoretically superior to other methods [8, 16] because they are compatible with the Shapley value (SV) that guarantees the equitable attribution of contributions. The SV assigns the contributions according to the impact on the prediction of each feature. In perturbation-based methods, which can quantify the contributions of any classifier in any feature space, the SSV and kernel SHAP also satisfy the SV axioms [8].

Based on the above review, the SSV is apparently a prominent method for interpreting EEG classification because it can display the contributions in an interpretable space and it satisfies the SV axioms. However, its computational cost is relatively high. In this paper, we apply the SSV to EEG classifiers and propose a pruning method to reduce its computational cost. EEG features form a tree structure comprised of a sensor (electrode) level, frequency-band level, and amplitude-phase level. If the contribution of a feature at a higher level (e.g., sensor level) is very small, the contributions of features at the lower levels of the feature (e.g., frequency-band level of the sensor) should also be small. Therefore, calculation of the contributions of such features can be ignored or pruned. We evaluate the proposed method using two benchmark EEG datasets to confirm the reduction of its computational cost. Furthermore, we demonstrate the higher interpretability of the proposed method compared to IG, IPNPCM, and IFUOCM.

2 Method

The SV was originally proposed to fairly assign the gains to players in cooperative game theory [12]. In its application to classification, the contribution $\phi_i(f,x)$ of the ith feature out of input feature x in a classifier, f, is given as:

$$\phi_i(f,x) = \sum_{S \subseteq x \setminus i} \frac{|S|!\,(M-|S|-1)!}{M!} [f_{S \cup i}(S \cup i) - f_S(S)], \tag{1}$$

where M denotes the number of input features, S denotes all possible subsets of an input feature space except for feature x_i, and $f_S(S)$ indicates the output of classifier f for input S. Basically, the contribution of a feature is defined as how much the output of a classifier is reduced by removal of the feature. The amount of reduction is then averaged over all possible combinations of features. This calculation requires computational cost $O(2^n)$ for input size n and retraining of the classifier for all possible combinations. The SSV approximates the SV using a sampling method to reduce its computational cost [14].

We apply the SSV to EEG classification to identify the contributions of amplitude and phase in each frequency band in each sensor (electrode). These features, i.e., amplitude phase, frequency bands, and sensors, form a tree structure (left side in Fig. 1). We contend that pruning of the tree can reduce the SSV computational cost. First, we

calculate the SSV at the sensor level, specifically to assess the influence of the elimination of a sensor. The number of features (electrodes) at this level is relatively small. The sensor signals with small contributions do not contribute to the classification at frequency-band or amplitude-phase levels. Therefore, it is not necessary to calculate the contributions of such irrelevant sensors at the lower levels. Similarly, calculation of the contributions of amplitude and phase can be ignored if the frequency bands do not contribute to the classification. The pruning can reduce the computational cost, especially if a few feature branches contribute to the classification.

3 Experimental Settings

We conducted experiments using two EEG datasets to verify the validity of the proposed method. One dataset is the PhysioNet polysomnography (PSG) dataset. It easily shows the raw waves and applies the perturbation-based methods because it includes data of only three sensors. Therefore, we applied the proposed method and IG to this dataset and calculated the computational efficiency of our proposed pruning method. The other dataset was the UCI EEG dataset. This dataset contains much more sensor data. Therefore, we empirically compared the proposed method with IPNPCM and IFUOCM by the input flipping method.

3.1 PhysioNet polysomnography dataset

The PhysioNet PSG dataset is a publicly available sleep PSG dataset from PhysioNet [4]. It includes data of 20 healthy subjects (ten males; ten females) of ages ranging from 25 to 34 years. We employed EEG (Fpz-Cz and Pz-Oz electrodes) and electrooculography (EOG) signals in this dataset. Their sampling rates were 100 Hz, and the duration of epochs was 30 s. During the first night of the experiment, PSG was used to train the classifier; during the last night, it was used to test it. We constructed a six-layered CNN, as shown in Fig. 2, to classify the data into five sleep stages: Rem, Wake, N1, N2 and N3. Its classification accuracy for test data is 81%. The sleep stages are officially labeled based on the EEG and EOG signals. For example, the class N3 is defined as the large low-frequency power (delta band) in EEG. Therefore, the power of the delta band is expected to contribute to CNN classification for N3.

We compared the results of our proposed method with those of IG [16]. We applied them to CNN and visualized their results on randomly chosen N3 data. In addition, we evaluated the proposed pruning method in terms of its accuracy and computational cost which is the number of calculations of model outputs. We randomly chose 400 data and computed their SSVs with and without pruning. A branch was pruned when the contribution was smaller than one-fifth that of the most contributed feature. We regarded the SSV of 1,000 samples per feature as the true value, i.e., the SV, and evaluated the difference between the true value and the value estimated by the SSV with pruning.

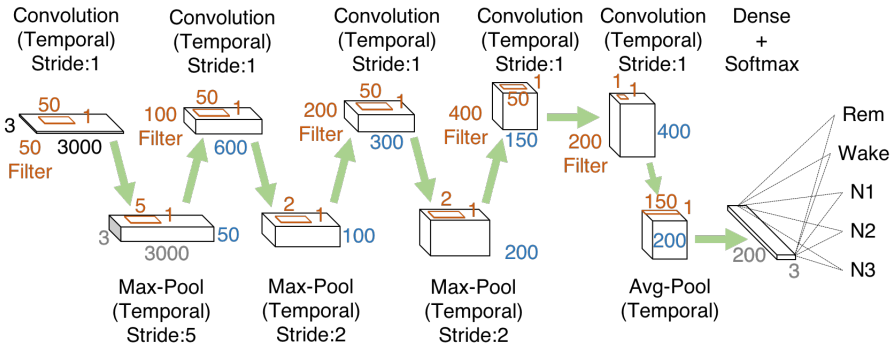

Fig. 2. CNN architecture in the PhysioNet experiment.

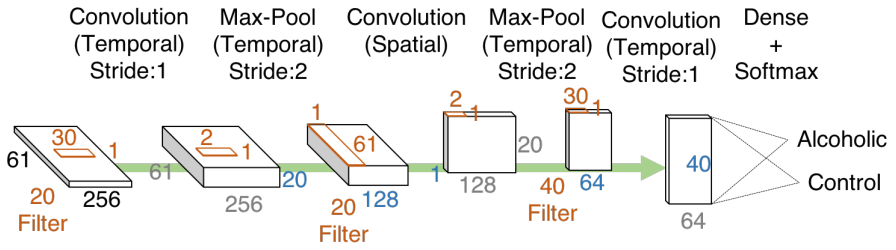

Fig. 3. CNN architecture in the UCI EEG experiment.

3.2 UCI EEG dataset

The UCI EEG dataset is a publicly available event-related EEG dataset in the UCI Machine Learning Repository [2]. The dataset is comprised of EEG data collected using 64 electrodes at 256 Hz. It contains 120 data items for one subject, each obtained within 1 s and labeled as "alcoholic" or "control." The number of subjects is 122 (77 alcoholics and 45 controls). We evaluated the accuracy of CNN using ten-fold cross-validation, resulting in 75% accuracy, as shown in Fig. 3. We compared the results of the SSV and IPNPCM. We applied them on 30 randomly chosen data items. In addition, we empirically quantified the power of explanation using "frequency-band-level flipping" based on "pixel flipping" [3]. Then, the output values of the classifier were evaluated while removing the most highly contributed features. The output decrease means the power of the explanation.

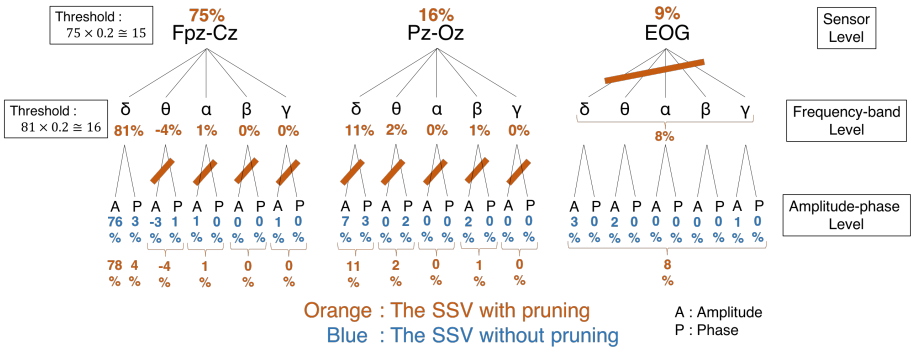

Fig. 4. Example of the results of the Shapley sampling value (SSV). Orange and blue percentages indicate the SSVs with and without pruning, respectively. Orange diagonal lines represent pruning.

4 Results

4.1 Results for the PhysioNet PSG dataset

An example of the SSV result on the randomly chosen N3 data is shown in Fig. 4. The contributions of features are described as percentages of the SSVs in trees. Orange and blue percentages denote the contributions with and without pruning, respectively. The figure shows that the power of the delta band in the Fpz-Cz electrode is the most important for this classification, which corresponds to the definition of N3. The percentages of the features are 78% for the SSV with pruning and 76% for the SSV without pruning, suggesting that the pruning effect on the accuracy is minimal. Fig. 5 shows an example of the IG result, where the colors on raw EEG signals indicate their contributions. IG shows the contributions in the input space, i.e., raw EEG signals. Therefore, this means of visualization is difficult to interpret and requires additional analysis to identify the important frequency bands.

Fig. 6 shows the effects of pruning on the accuracy (left panel) and computational cost (right panel). The horizontal axes indicate the number of samplings per feature in both panels. The solid and broken curves indicate the values for the SSV with and without pruning, respectively. The green, red, and blue curves represent the results of classification for the N2 class, all classes, and the N3 class, respectively. The left panel shows the approximation errors of the SSVs with respect to the true SVs. The results show that the errors of the SSVs with pruning are the same level as those without pruning, especially in the range of samples per feature from 10 to 50. The right panel shows that pruning reduces the computational cost to approximately two-thirds. These results suggest that the proposed method realizes effective the SV estimation.

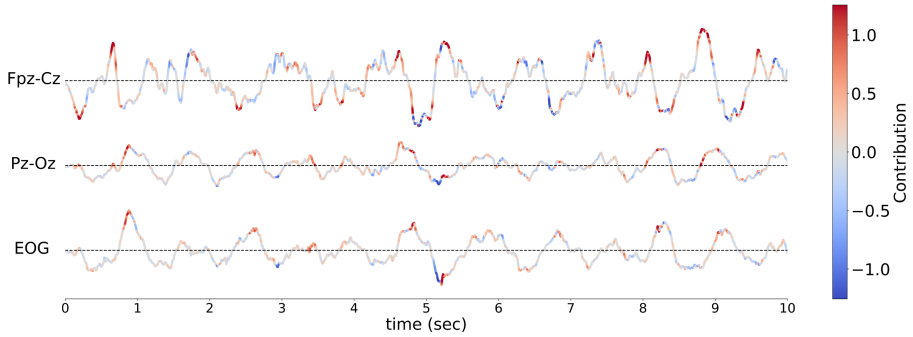

Fig. 5. Example of the results of ingredient gradients [16]. Curves indicate raw EEG signals and their colors represent their contributions at the given time.

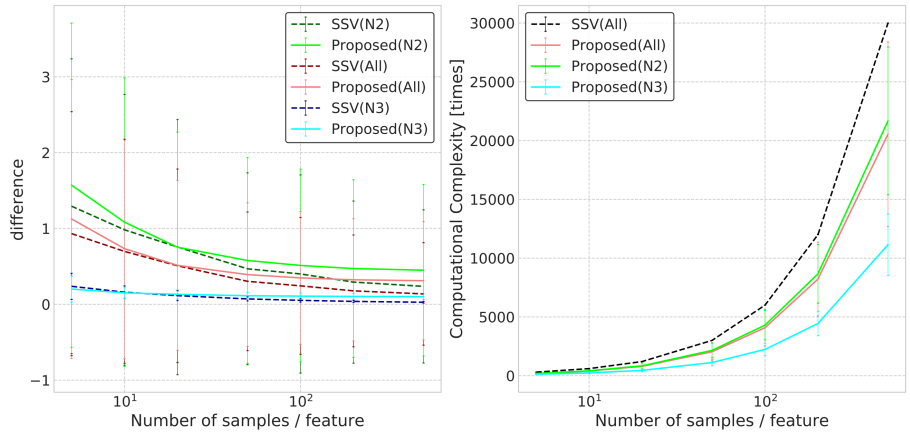

Fig. 6. Comparison of the results of the SSV with and without pruning. Left: difference between the true Shapley value and the SSV. Right: computational cost.

4.2 Results for the UCI EEG dataset

The results of the proposed SSV and IPNPCM are shown in Fig. 7 and Fig. 8. The SSV demonstrates significant contributions of amplitude in the delta and gamma bands and of phase in the delta band. IPNPCM contributes amplitude in the beta and gamma bands and phase in the delta band. The beta band was not addressed by the SSV because of the already mentioned problem of the correlation. Fig. 9 shows the results of the "frequency-band-level flipping" for the SSV methods with pruning (blue curve), IPNPCM (orange curve), and IFUOCM (green curve). The classification scores of the SSV (with pruning) significantly decreases compared to those of IPNPCM and IFUOCM, suggesting that the proposed SSV more appropriately elucidates the classification.

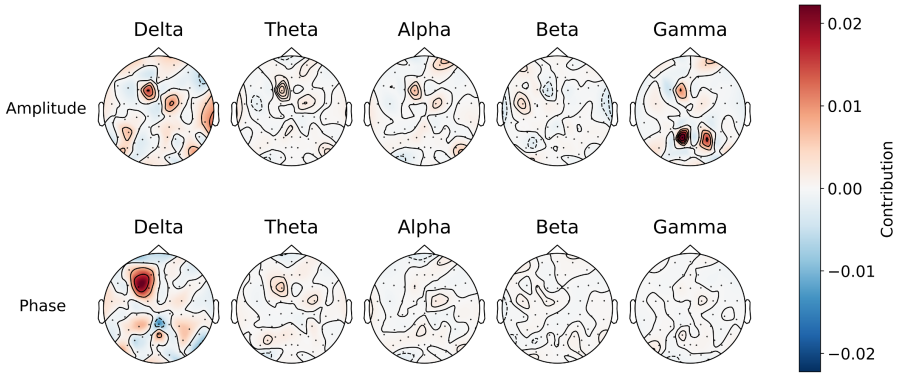

Fig. 7. Averaged contributions for 100 data items, visualized by the SSV with pruning.

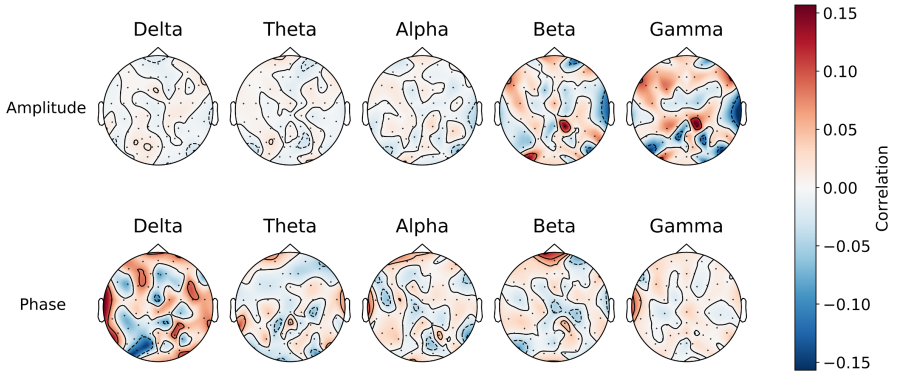

Fig. 8. Averaged contributions for 100 data items, visualized by IPNPCM [11].

5 Discussion and Conclusion

In this paper, we proposed a pruning method in the SV sampling and demonstrated that the method can effectively quantify the contributions of features in CNN classifiers. We verified the proposed method when applied to two tasks: classification of sleep stages and alcoholic screening. In the first experiment, the SSV assigned the largest contributions to amplitude in the delta band (Fig. 4), which was consistent with the definition of the N3 sleep stage. In the second experiment, the SSV displayed the contributions of amplitude in the delta and gamma bands and phase in the delta band (Fig. 7). A recent review of EEGs of alcoholics demonstrated that many studies focus on the gamma band for screening the event-related potentials of alcoholics [9]. In addition, Tcheslavski and Gonen [17] found significant differences of the power and coherence in the lower frequency bands between alcoholics and controls. These results correspond to our visualization, suggesting that the proposed method can explain the contributing features.

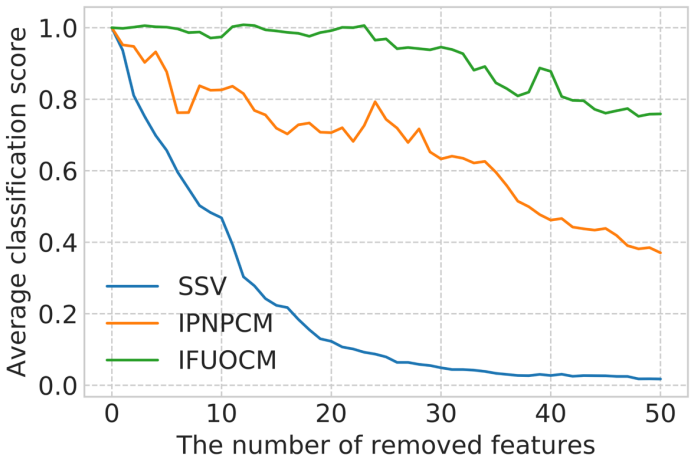

Fig. 9. Result of frequency-band-level flipping for the SSV with pruning, IPNPCM, and IFUOCM [11]. The classification score is normalized so that the scores of the original input are 1.0.

Moreover, the conducted experiments produced the following four results. 1) The proposed method effectively interpreted the EEG classification in the amplitude-phase feature space, while gradient-based methods, including IG, could not explain them in such an interpretable feature space (Fig. 5). 2) Our pruning method reduced the computational cost while maintaining the estimation accuracy (Fig. 6). 3) The SSV explanation is superior to the IPNPCM explanation and IFUOCM explanation in terms of the pixel-flipping evaluation (Fig. 9). 4) The contributions visualized using the proposed method are consistent with previous findings on EEG biomarkers.

Although pruning reduces the SSV computational cost, the cost is still much greater than those of backpropagation-based methods. We must address this problem to apply the SSV to medical data that have a very large number of features. We surmise that combining our method with a backpropagation-based method, such as IG, can enable a more feasible visualization technique.

Acknowledgements

This work was supported by the Center of Innovation Program from Japan Science and Technology Agency and JST CREST Grant Number JPMJCR17A4, Japan.

References

1. Ancona, M., Ceolini, E., Oztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: Proceedings of the 6th International Conference on Learning Representations (2018)
2. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 10(7), e0130140 (2015)
4. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet. Circulation 101(23), e215-e220 (2000)
5. Hartmann, K.G., Schirrmeister, R.T., Ball, T.: Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding. In: Proceedings of the 6th International Conference on Brain-Computer Interface. pp. 1-6 (2018)
6. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: A compact convolutional network for EEG-based brain-computer interfaces. arXiv:1611.08024 (2016)
7. Li, Y., Murias, M., Major, S., Dawson, G., Dzirasa, K., Carin, L. and Carlson, D.E: Targeting EEG/LFP synchrony with neural nets. In: Advances in Neural Information Processing Systems. pp. 4623-4633 (2017)
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4768-4777 (2017)
9. Mumtaz, W., Vuong, P.L., Malik, A.S., Rashid, R.B.A.: A review on EEG-based methods for screening and diagnosing alcohol use disorder. Cognitive Neurodynamics pp. 1-16 (2018)
10. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135-1144. ACM (2016)
11. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping 38(11), 5391-5420 (2017)
12. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games 2(28), 307-317 (1953)
13. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. arXiv:1704.02685 (2017)
14. Shtrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41(3), 647-665 (2014)
15. Sturm, I., Lapuschkin, S., Samek, W., Muller, K.R.: Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods 274, 141-145 (2016)
16. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. arXiv:1703.01365 (2017)
17. Tcheslavski, G.V., Gonen, F.F.: Alcoholism-related alterations in spectrum, coherence, and phase synchrony of topical electroencephalogram. Computers in Biology and Medicine 42(4), 394-401 (2012)
18. Vilamala, A., Madsen, K.H., Hansen, L.K.: Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. arXiv:1710.00633 (2017)