German Medical Data Sciences 2022 - Future Medicine
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI220801

Consistency of Feature Importance Algorithms for Interpretable EEG Abnormality Detection

Felix KNISPEL[a], Alexander BRENNER[b], Rainer RÖHRIG[a], Yvonne WEBER[c], Julian VARGHESE[b], and Ekaterina KUTAFINA[a,d,1]

[a] *Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[b] *Institute of Medical Informatics, University of Münster, Münster, Germany*
[c] *Department of Epileptology, Neurology, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[d] *Faculty of Applied Mathematics, AGH University of Science and Technology, Krakow, Poland*

Abstract. Recent advances in machine learning show great potential for automatic detection of abnormalities in electroencephalography (EEG). While simple and interpretable models combined with expert-comprehensible input features offer full control of the decision making process, these methods commonly lag behind complex deep learning and feature extraction methods in terms of performance. Here we study a feasibility of a bridging solution, where deep learning is combined with interpretable input and an algorithm computing the importance of particular EEG features in the decision process. We built a convolutional neural network with multi-channel EEG frequency bands as input and investigated four different methods for feature importance attribution: Layer-wise Relevance Propagation (LRP), DeepLIFT, Integrated Gradients (IG) and Guided GradCAM. Our analysis showed consistency between the first three methods, and deviating attributions of the fourth method, suggesting the importance of using a package of methods together to ensure the robustness of medical interpretation.

Keywords. Machine Learning Interpretability, Electroencephalography, EEG, Decision Support Techniques

1. Introduction

Modern machine learning (ML) systems are increasingly considered as an important tool in clinical decision support systems. However, some domains, such as electroencephalography (EEG) that measures electrical brain activity, remain challenging. An important issue remains the lack of transparency and interpretability of the best performing models, which typically use raw data or highly complex features as an input to the black box of deep learning neural networks.

[1] Corresponding Author: Dr. Ekaterina Kutafina, PhD. Address: Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Pauwelsstraße 30, 52074 Aachen, Germany. Email: ekutafina@ukaachen.de

In [1] we constructed an ML model to detect abnormalities in EEG recordings on the example of the Temple University Hospital (TUH) dataset [2]. In [3] we worked towards an understanding of the individual computational steps that lead to a classification of an EEG. The importance of individual EEG channels and the generated features in general was analyzed. Another possible direction of improvement is combining expert-interpretable features with deep learning models. In such setup, post hoc interpretability algorithms can be used to determine the importance of a given feature in the decision process. As a result, explanations of model decisions can be given in a user-interpretable format. In this paper, we explored this direction by building up on our previous work. We replaced difficult to comprehend wavelet input features with clinically-relevant frequency bands. These features were mapped to a grid for all electrode positions, generating an approximate head-map per sample. The resulting images were used as input to a deep learning (DL) model. This approach allowed us to consider different post hoc feature attribution methods and discuss their consistency and possible relationship with the results to the medical reasoning.

The choice of the feature attribution methods is based on several works published on EEG interpretability. The Layer-Wise Relevance Propagation (LRP) has been previously applied to BCI-EEG data in [4] to generate heatmaps of feature importance. This way, Sturm and colleagues were able to explain with high-resolution at what point in time and at which electrodes on the head scalp the electrical activity was important in the model's classification. Uyttenhove et al. [5] applied a different feature attribution method, GradCAM, to raw time-series EEG data to obtain clinically plausible attributions. DeepLIFT was used by Jansen et al. [6] to analyze the results of an artificial neural network trained on physiological network data of insomnia patients. Integrated Gradients (IG) is another common method for highlighting feature attributions [7].

2. Methods

MNE (v0.24.1) [8] was used for accessing EEG data. PyTorch (v1.10.1) was used to develop and evaluate the machine learning model. To generate feature attributions using LRP, GradCAM, DeepLIFT and IG, we used Captum (v0.4.1) [9].

2.1. Data

The openly available Temple University Hospital (TUH) Abnormal EEG Corpus v2.0.0 [2] is used as our example. 2993 EEG sessions from 2383 subjects are included. Of those sessions, 93% were recorded with a sampling frequency of 250 Hz, the remaining sessions were sampled with 256 or 512 Hz. 1472 recordings are labeled as abnormal, the remaining 1521 as normal. The labeling of EEGs by the dataset authors is based on visual analysis of frequency, voltage, waveform, regulation, locus, reactivity and interhemispheric coherence [2]. The recordings are split into a training set of 2717 recordings and an evaluation set consisting of 276 recordings. We trained models exclusively on the training set, and report model accuracy and feature attributions on the evaluation set.

2.2. Preprocessing

Similar to Brenner et al. [1] and Mortaga et al. [3], we extracted the first minute of every EEG recording. All 21 electrodes were used. The 60 seconds of raw data were re-sampled to 250 Hz, band-filtered (1-50 Hz), and then divided into 11 sliding windows, each of 10 second length with a 5 second overlap. Each segment was checked against a threshold for maximal amplitude. If more than 50% of all electrodes exceeded this threshold, the segment was removed. Lastly, on each remaining segment we calculated power spectral density (PSD) for the five clinically relevant frequency ranges (Delta: 1-3 Hz, Theta: 4-7 Hz, Alpha: 8-13 Hz, Beta: 14-30 Hz, Gamma: >30 Hz) using Welch's method with Hann windowing and 50% overlap. Our final dataset consisted of 27761 segments in the training set and 2813 segments in the testing set, with 5 PSD values for every single one of the 21 electrodes. Each segment was transformed into a 5x5x7 matrix that can be interpreted as a 5x7 image with five channels corresponding to the five frequency bands. All 21 electrodes were mapped as illustrated on Figure 1 – positions in the matrix without annotated electrodes were assigned zero-values to fill up the matrix format. These "images" consisting out of 175 individual features were used as input to the DL model.

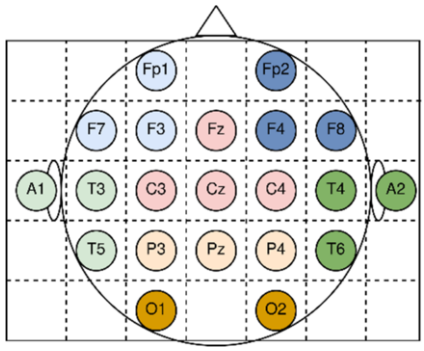

Figure 1. Grid used for mapping electrode positions into a 2D square. The colors illustrate the grouping per brain region used in algorithm comparison.

2.3. Deep Learning model

We constructed a Convolutional Neural Network (CNN) model to classify our segments as normal or abnormal. Our model consists of one convolutional layer (with 3x3 filters), a pooling layer, one fully connected layer with 128 nodes and a final output layer with two nodes representing the class probabilities. Due to the splitting of each EEG recording into 11 segments, we hypothesize that not all segments of an abnormal EEG exhibit abnormal morphologies. This may result in a decreased classification accuracy, as possibly normal-looking segments are labeled as abnormal. To mitigate this problem, we implemented a voting mechanism as in [1].

2.4. Post hoc feature importance attribution

To understand and interpret the importance of our input features, we generated feature importance values using LRP [4], Guided GradCAM [5], DeepLIFT [6], and IG [7] for all 2813 testing segments. Note, that for GradCAM, we obtain relevance scores for the first convolutional layer. These algorithms make use of the given sample, network

structure, learned network weights, as well as outcome class of interest. For each of the 175 individual input features, the algorithms return a (positive or negative) real number, indicating an importance of this feature to the outcome of interest. Positive attributions for a feature indicate that high feature values contribute towards the outcome of interest, while negative attributions indicate that low feature values contribute to the outcome of interest. Large absolute feature attributions are therefore considered important features. As our primary interest lies in the reasoning for why an EEG is considered abnormal by the model, we attributed feature importance for the abnormal class.

2.5. Normalization and comparison of feature importance attribution

Due to varying distribution and scale of different feature attribution methods, we applied the following normalization. Outlier removal was performed by first calculating the sum of all attributions for each segment and method. Based on the top and bottom third percentile of attributions for each method, we removed outlier segments. Following this, we normalized every attribution left across electrodes and channels to unit norm using L2-normalization before further processing the attributions.

To view feature importance not only per sample but also across the evaluation set, we averaged attributions for abnormality across all samples. Further, we grouped electrodes into lobes (Frontal left: Fp1, F3, F7; Frontal right: Fp2, F4, F8; Temporal left: A1, T3, T5; Temporal right: A2, T4, T6; Parietal: P3, Pz, P4; Occipital: O1, O2; Central: Fz, Cz, C3, C4). To compare similarity between different methods, we calculated Mean Squared Error (MSE). For every combination of two feature attribution methods, we calculated MSE in attribution differences of a fixed feature (pairing of electrode and frequency band) for every single sample. We then averaged resulting MSE across all features, resulting in one MSE-based score for every combination of methods.

3. Results

During preprocessing, around 7% of all EEG segments were removed due to amplitude thresholding. After training the model, we achieved an accuracy of around 78% on the 2813 segments of the testing set. When using the voting mechanism, accuracy increases to 81.4%. Model performance remains within acceptable range for the study purpose on the test set. This is on par with previous results of 80.15% accuracy in [3], outperforms the results of de Diego [2] (78.8%), but lags behind more complex models such as BD-Deep4 (85.4%) [10] or ChronoNet (86.6%) [11]. Feature attributions were calculated for all test segments using all four methods. After removing 445 of these segments due to outlier values, we normalized and plotted results in Figure 2 and Figure 3. They visualize the attributions generated by the four methods for different electrodes/lobes and frequency bands.

Results of the MSE-based similarity between methods are presented in Figure 4. A single MSE value is difficult to interpret, but Figure 4 shows high agreement between the results of LRP, IG, and DeepLIFT compared to agreement between GradCAM and any other method. Similar conclusion can be drawn from Figures 2 and 3.

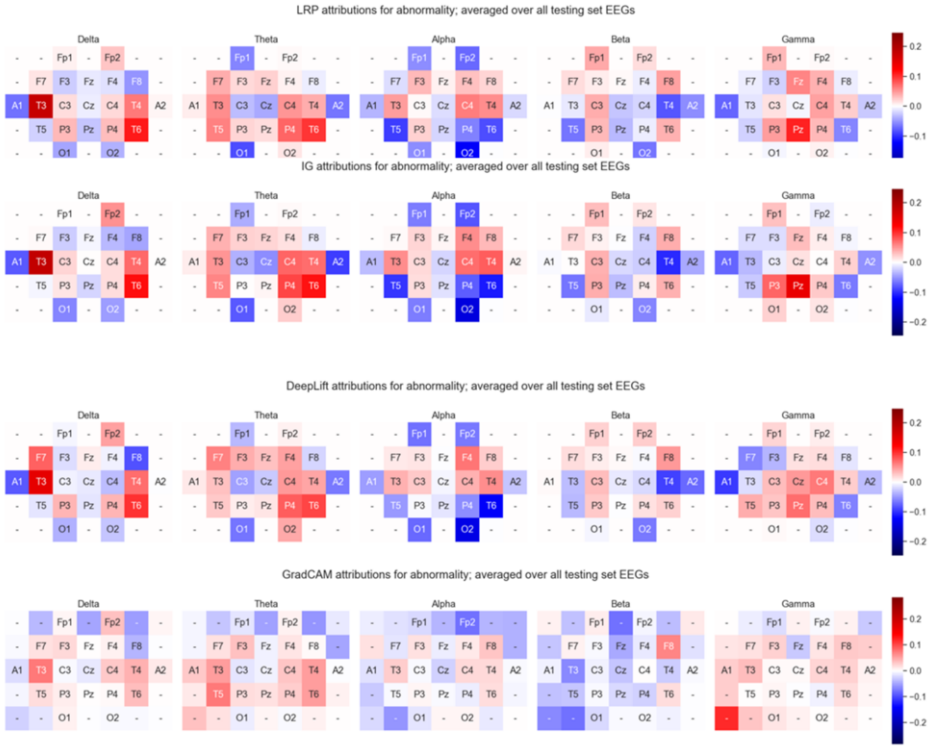

Figure 2. Attributions for abnormality generated by the feature attribution methods.

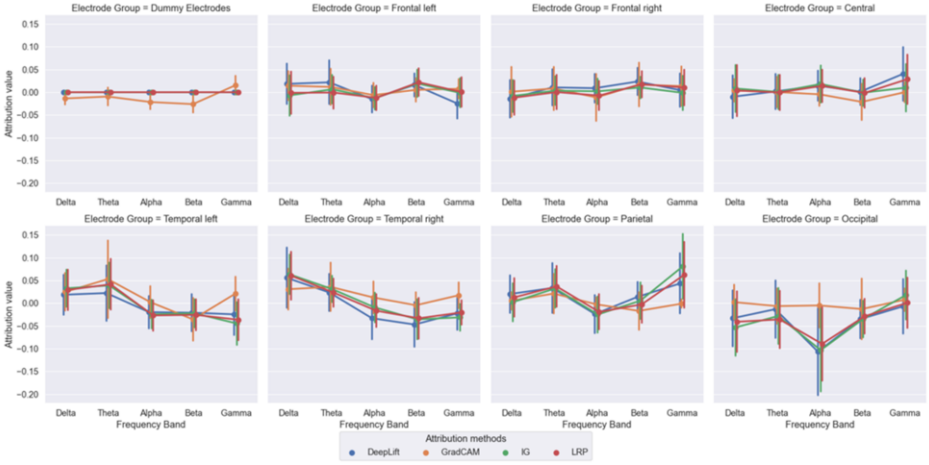

Figure 3. Average attribution values for abnormality, broken down into lobes, methods and frequency bands.

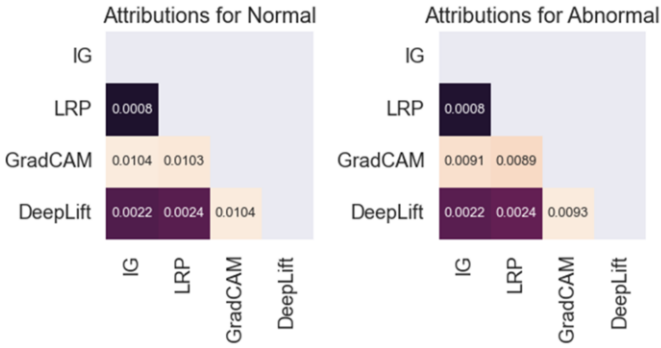

Figure 4. Pairwise mean squared error (MSE) between different attribution methods. Lower MSE indicates higher level of agreement between two methods.

4. Discussion

For the purpose of this study, we have constructed a DL model that uses interpretable frequency-based features extracted from the raw EEG data. The usage of a CNN preserves information about spatial position of the electrodes. As a main paper goal, four different methods for post hoc feature importance assessment were compared by a) using the medically relevant view of the electrode mesh on the head, b) grouping electrodes into larger brain regions and presenting averaged importance attributions separated by frequency bands and brain region and c) computing easy-to-compare MSE.

While our model's performance is notably similar to that of some other models from literature, we note that inferior performance compared to more complex models such as ChronoNet could possibly be attributed to our comparatively simple feature presentation as well as smaller network size.

The head view shows clear visual similarities between three methods with GradCAM being an exception. This is further confirmed by the MSE table (Figure 4) and Figure 3 (line plot). In Figure 2, we can also observe that, in agreement with the expectations, "dummy" electrodes on the 2D grid show little importance. We hypothesize that GradCAM's non-zero feature attribution for these electrodes can be traced back to the method's unique usage of convolutional filters and its handling of Rectified Linear Units in the neural network. Frontal lobes and the central part of the head in general have slightly higher importance. Signals from two temporal lobes (similar directions), parietal (directions opposite to the temporal lobes) and occipital have the strongest effects on the model prediction. High and low frequency bands are showing consistent effects. Interestingly, in the occipital lobe the alpha band clearly stands out, which is in agreement with the fact that large alpha band fluctuation related to e.g. closed/open eyes conditions manifest in the occipital lobe. The medical interpretation of the results should be taken with cautiousness, due to the fact that the labeling of the discussed data set is not controlling for drug usage and it has been reported that medications used to treat epilepsy can affect certain frequency bands [12].

Our work clearly demonstrates both benefits and risks of using feature attribution methods to explain model decisions. On the one hand, calculated feature attributions confirm that the model puts relevance on properties of the EEG signal that are in line with medical expectations. The consistency of some methods allows us to place a certain

trust in the functioning of the respective feature attribution methods. On the other hand, inconsistency of others highlights the need for extensive testing and comparison of multiple feature attribution algorithms in the context of EEG data.

In the further work, we will work towards preparing a small data set where both labeled normal and abnormal fragments will be available from the patients who are in an ongoing diagnostic process and therefore are not yet receiving pharmacological intervention. The control for alertness level and age is equally important, as those factors are known to affect the EEG spectrum strongly. Meanwhile, analysis of text-files provided in the TUH dataset for every EEG, describing patient and recording, will be performed. This could enable meaningful validation of feature attributions on the level of individual EEG recordings. Moreover, other datasets in the TUH Corpus contain labels for specific types of abnormalities in EEG recordings. These may be used to further validate obtained results.

5. Conclusion

The paper presents one of several possibilities of constructing human-interpretable ML models for EEG data. The constructed ML model establishes reasonable accuracy, is light weighted and allows to easily test different algorithms for post hoc feature importance attribution. Three of the four tested algorithms showed very consistent results. The inconsistency of the fourth one suggests that if the proposed approach is used for clinical purposes, several different algorithms should be tested to increase the robustness of the importance interpretation.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.

Author contributions: FK and AB executed the computations; EK, JV and RR supervised computational side of the work; YW supervised the medical side; FK, EK and AB drafted the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

References

[1] Brenner A, Kutafina E, Jonas SM. Automatic recognition of epileptiform EEG abnormalities. In: Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. IOS Press; 2018. p.171-5.

[2] de Diego, S L. Automated interpretation of abnormal adult electroencephalograms. Temple University. 2017.

[3] Mortaga M, Brenner A, Kutafina E. Towards interpretable machine learning in EEG analysis. In: German Medical Data Sciences 2021: Digital Medicine: Recognize–Understand–Heal. IOS Press; 2021. p. 32-8.

[4] Sturm I, Lapuschkin S, Samek W, Müller KR. Interpretable deep neural networks for single-trial EEG classification. Journal of neuroscience methods. 2016;274:141-5.

[5] Uyttenhove T, Maes A, Van Steenkiste T, Deschrijver D, Dhaene T. Interpretable epilepsy detection in routine, interictal eeg data using deep learning. In: Machine Learning for Health. PMLR; 2020. p. 355-66.

[6] Jansen C, Penzel T, Hodel S, Breuer S, Spott M, Krefting D. Network physiology in insomnia patients: Assessment of relevant changes in network topology with interpretable machine learning models. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2019;29(12):123129.
[7] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: International conference on machine learning. PMLR; 2017. p. 3319-28.
[8] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. Frontiers in neuroscience. 2013:267.
[9] Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for PyTorch; 2020.
[10] Tibor Schirrmeister R, Gemein L, Eggensperger K, Hutter F, Ball T. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. arXiv e-prints. 2017:arXiv-1708.
[11] Roy S, Kiral-Kornek I, Harrer S. ChronoNet: a deep recurrent neural network for abnormal EEG identification. In: Conference on artificial intelligence in medicine in Europe. Springer; 2019. p. 47-56.
[12] Ouyang CS, Chiang CT, Yang RC, Wu RC, Wu HC, Lin LC. Quantitative EEG findings and response to treatment with antiepileptic medications in children with epilepsy. Brain and Development.