doi:10.3233/SHTI210538

Towards Interpretable Machine Learning in EEG Analysis

Maged MORTAGA[a], Alexander BRENNER[b] and Ekaterina KUTAFINA[a,c,1]
[a] *Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany*
[b] *Institute of Medical Informatics, University of Münster, Münster, Germany*
[c] *Faculty of Applied Mathematics, AGH University of Science and Technology, Krakow, Poland*

Abstract. In this paper a machine learning model for automatic detection of abnormalities in electroencephalography (EEG) is dissected into parts, so that the influence of each part on the classification accuracy score can be examined. The most successful setup of several shallow artificial neural networks aggregated via voting results in accuracy of 81%. Stepwise simplification of the model shows the expected decrease in accuracy, but a naive model with thresholding of a single extracted feature (relative wavelet energy) is still able to achieve 75%, which remains strongly above the random guess baseline of 54%. These results suggest the feasibility of building a simple classification model ensuring accuracy scores close to the state-of-the-art research but remaining fully interpretable.

Keywords. EEG, supervised machine learning, epilepsy, decision support techniques

1. Introduction

Electroencephalography (EEG) is a method of assessing the brain electrical activity [1]. EEG plays a key role for example in neurology as a diagnostic and monitoring tool for epilepsy or sleep disorders. The important medical role of EEG technology convolved with the rapid methodological development of data analysis in the recent years led to multiple solutions aiming in automatic EEG data analysis. Seizure detection and prediction together with epileptiform abnormalities detection remain on the leading target positions in this research. With the increasing complexity of the used models we observe improvements in the scores achieved on specific datasets, but the clinical usability remains rather limited.

This problem recently became well recognized in computational medicine [2], and in EEG research in particular [3, 4]. One important approach to improving the situation is working on interpretability of deep artificial neural networks (DNN) which are often used to perform classification and prediction tasks on medical data [5]. In this paper we are dealing rather with shallow networks embedded into a multi-step pipeline which includes preprocessing, feature selection and a stabilizing voting system. Our goal is to

[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.

take a step backwards and examine the potential of simplifying the methodology to the level, when it is possible to return to a direct discussion of the results with the medical experts experienced in manual EEG analysis. This step is done by dissecting the shallow artificial network-based EEG classification pipeline from Brenner et al. [6]. It was designed to classify normal and abnormal EEG recordings from TUH Abnormal data corpus (v.1.1.2) [7]. The resulting accuracy of 81% was later outperformed by several other approaches, which due to their complexity would be much more difficult to simplify and interpret. Roy et al. [8] reported the accuracy score of 86% achieved through employing ChronoNet, a cross-over of convolutional and recurrent neural networks. Recent preprint of Fernando et al. [9] introduces Neural Memory Networks with the result of 93%.

The paper of Brenner et al. [6] presents a relatively simple machine learning approach which is based on wavelet features and consists of several optimization steps. This simplicity enables us to examine the influence of different features and optimization steps on the final accuracy score. The presented work can be seen as research towards interpretability of the machine learning solutions for EEG analysis, with the practical goal of facilitating the feedback between medical professionals and data scientists which is necessary for development of clinically usable methods.

2. Methods

2.1. Data

The data used for the classification is the TUH Abnormal EEG Corpus (v.2.0). Lopez and colleagues [6] used a decision tree to label an EEG session as either normal or abnormal. The original data set contains 2993 records and is readily split into a training and evaluation data set. In Brenner et al. [6] following previous works [10] the single EEG channel T5 - O1 was used and only 60 sec of the recording was considered. The raw data was resampled to 250 Hz, band filtered (1-50 Hz) and then a moving window of length 10 seconds and overlap 5 seconds was applied to split records into 11 segments each. While keeping track of the correspondences of the segments to the original records, pieces with abnormally high amplitude (above 100 μV) were discarded from further processing. On the remaining segments 6 decomposition levels of Symlets wavelet transform (order 7) were extracted. Statistics were computed over the set of wavelet coefficients to form representative feature vectors (maximum, minimum, mean, standard deviation, wavelet entropy, relative wavelet energy). Given 6 statistical values for each set of wavelet coefficients, the resulting feature vectors hold 42 values per segment.

2.2. Starting point: the original classification pipeline (Figure 1)

Using MATLAB's PatternNet tool, Brenner et al. [6] achieved an accuracy of 80.51% with a network structure with 2 layers, 15 nodes each. 10 instances of such a network were trained on the dedicated train set and applied to the EEG segments. The classification per record was implemented via a two-phase voting system. Firstly, the classification results from 10 networks on the single segments are averaged to decide on the segment label. Secondly, a majority voting is applied to the 11 overlapping segments from one record to decide about the final classification. Since certain segments of the

EEG set were discarded due to amplitude filtering, in some EEG sessions less than 11 segments were used in the voting process.

2.3. *Pipeline interpretability: modifying single steps*

In this paper we start from the readily extracted features and reconstruct the pipeline illustrated in Figure 1, to confirm the starting point for our accuracy. In the next steps we modify several parts of the pipeline to understand their influence on the overall accuracy score. We have tested a) the choice of the EEG channel used as the input, b) the effect of the voting and averaging, c) different network architectures (number of layers and nodes) and d) the choice of the statistical wavelet features used as the input. Moreover, we investigated the results of replacing the neural network as a main classification block with a simple single-feature thresholding (Figure 2).

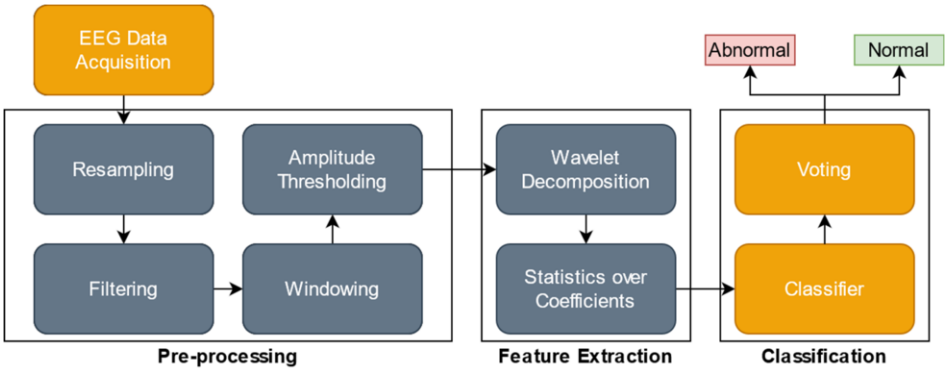

Figure 1. The original classification pipeline from [6], in orange the modified parts are marked.

3. Results

Step 1. Verifying the baseline results. We re-implemented the original pipeline [6] and obtained the accuracy of 80.15%, which is close to the reported 80.51% with the difference explained by the update of the dataset and the random initialization of the network weights.

Step 2. Other single channels. T5-O1 channel was already earlier reported as a promising choice [11]. The default setup was tested on the other 20 EEG channels of the TCP montage to confirm the superiority T5 - O1 channel (see Figure 3). The worst performing channel appears to be FP1-F7 with 69.57%. Both highest and lowest scores are achieved in the left hemisphere.

Step 3. Removing the voting. Each 60 seconds long data fragment is originally divided into 11 overlapping segments which inherit the original label. The predicted labels are used to vote for the whole record. Therefore, after removing the voting, the results are not directly comparable. Nevertheless, we can treat each 10 sec segments as a representation of the original data and assess the segment-based accuracy of 77.09%.

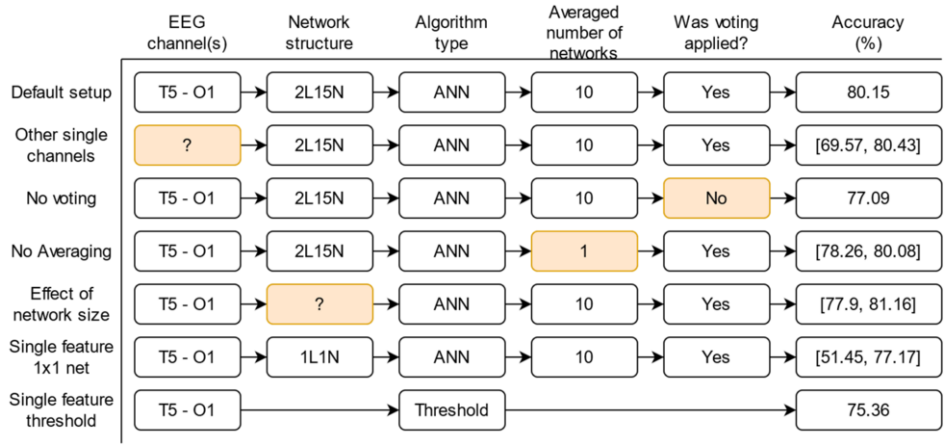

Figure 2. Modifications of the original pipelines. The orange color marks the parts where the modification was introduced.

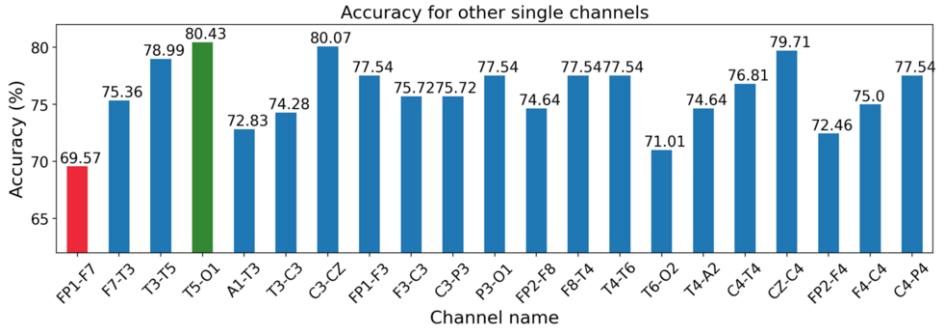

Figure 3. Accuracies of the pipeline based on different single EEG channels.

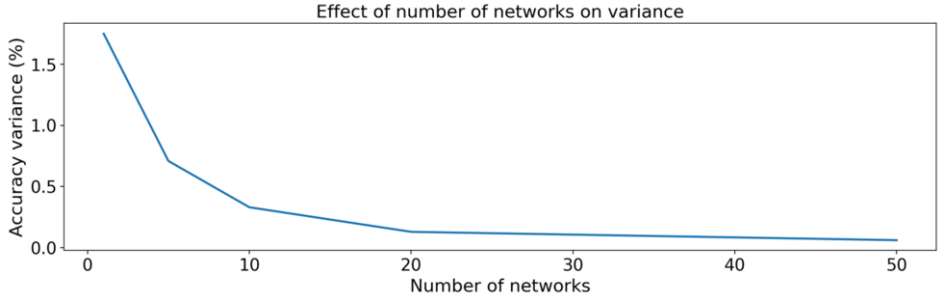

Figure 4. The dependence between the number of averaged networks and the resulting accuracy variance.

Step 4. Effect of averaging. 10 independent runs of the pipeline with removed averaging resulted in accuracy between 78.26% and 80.08%. To investigate the relationship between the accuracy variation and the number of averaged by the pipeline networks, we tried a different number of averaged networks and computed the corresponding variance

in accuracy percentage. Figure 4 supports the originally chosen 10 networks as an adequate number to optimize the ratio between the computation time and the variance.

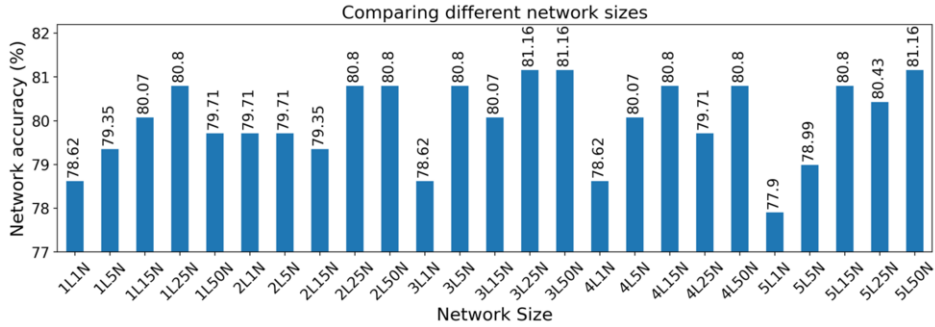

Figure 5. Different number of layers and nodes and the resulting pipeline accuracy.

Step 5. Different network structures. We tested from 1 to 5 numbers of layers. The number of nodes in each layer was kept identical and took values N=1, 5, 15, 25, 50. Figure 5 illustrates the best result of 81.16% and worst of 77.9%. The trivial setup with 1 layer and one node achieved 78.62%, which is, considering the model complexity differences, remarkably close to the best one.

Step 6. Single feature in 1x1 net. The surprisingly high performance of a simplistic 1x1 net suggested the next simplification step, namely reducing the number of the input features from the original 42 to 1. On Figure 6 we presented the performance of the different features fed into 1x1 net, grouped by the wavelet decomposition level. Relative wavelet energy on D3 (feature nr. 30) achieved the best accuracy of 77.17 %.

Step 7. Threshold-based approach. Finally, the relatively high score obtained with the simplistic network of one layer and one node suggested one more step towards simplification, namely putting a threshold on the feature chosen in step 6. The threshold was optimized to the value 23 based on the best accuracy on the train set and resulted in 75.36% accuracy on the test set.

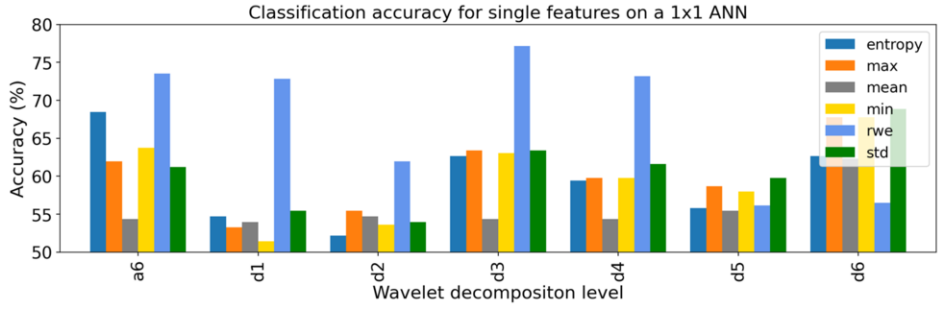

Figure 6. Accuracy achieved by single statistical features combined with a simplistic network of 1 layer and 1 node.

4. Discussion

Interpretability of machine learning [12] is being recognized as an important part of medical AI development. In this paper we took a simple approach which nevertheless allowed us to learn many things about the discussed computational pipeline. The model was segmented into pieces which could be easily modified or removed (Figure 1). The complete pipeline started from 80.15% accuracy achieved by multiple averaged shallow networks with an added voting system. After several steps of degrading we found out that even simple thresholding on a single feature (relative wavelet energy on D3 coefficients) can get accuracy as good as 75.36% (with the random guess baseline of 54.35%).

Experiments with different computational steps supported the importance of channel and feature choice. Specific ANN structure (number of layers and nodes) did not seem to have that large influence, likely due to MATLAB internal optimization algorithms. In further work we plan to reconstruct the pipeline in Python to gain better control of the computational steps.

Averaging multiple randomly initiated networks results in essentially improved robustness and makes the results more reproducible.

Interestingly, we discovered that different setups and initializations keep consistently correctly classifying a certain subset of the data. The differences in the accuracy can be attributed to much smaller subsets which are classified less consistently. In the further work we plan to collaborate with the medical professionals to better understand this issue. This collaboration can also help with a labelling problem which occurs in our computational setup. The labels attributed to the full EEG records are automatically inherited by smaller analyzed fragments. Inheritance of "abnormal" labels may potentially result in an error which can be recognized only by a trained professional.

The approach presented in this paper can be successfully used in typical for medically relevant biosignals setup, when machine learning solution is embedded into multi-step analysis, which includes e.g. feature extraction. It can be potentially combined with testing simple and easily interpretable machine learning algorithms, such as decision trees and regression models [4] as well as with deep nets interpretability study [5].

5. Conclusion

The development of methods for automatic EEG analysis is very promising but requires careful tracking of the underlying mechanisms. In this paper we attempted to study one of the machine learning models for classification of normal and abnormal EEG data to investigate the influence of different computational steps. The obtained results provide an important insight about most relevant computational parts (e.g. averaging, channel choice) and suggesting further analytical steps. In particular, the reported difference in feature importance can be studied via employing better interpretable features, such as a combination of frequencies and entropies instead of statistics of wavelet coefficients.

Declarations

Conflict of Interest: The authors declare no conflict of interest.

Contributions of the authors: AB and MM executed the computations, EK supervised the work, MM and EK drafted the manuscript, all authors substantially revised it and approved the final version. All authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

[1] Niedermeyer, E., da Silva, F.L. Electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins (2005).
[2] Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput & Applic. 2020 Dec;32(24):18069–83.
[3] Sturm I, Lapuschkin S, Samek W, Müller K-R. Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods. 2016 Dec;274:141–5.
[4] Vázquez MA, Maghsoudi A, Mariño IP. An Interpretable Machine Learning Method for the Detection of Schizophrenia Using EEG Signals. Front Syst Neurosci. 2021 May 28;15:652662.
[5] Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018 Feb 1;73:1–15.
[6] Brenner A, Kutafina E, Jonas SM. Automatic Recognition of Epileptiform EEG Abnormalities. Studies in Health Technology and Informatics. 2018;171–5.
[7] Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. Front Neurosci 2016:10.
[8] Roy S, Kiral-Kornek I, Harrer S. ChronoNet: A Deep Recurrent Neural Network for Abnormal EEG Identification. arXiv:180200308. 2018 May 17
[9] Fernando T, Denman S, Ahmedt-Aristizabal D, Sridharan S, Laurens K, Johnston P, et al. Neural Memory Plasticity for Anomaly Detection. arXiv:191005448 [cs, stat] 2019 Oct 11
[10] López S, Suarez G, Jungreis D, Obeid I, Picone J. Automated Identification of Abnormal Adult EEGs. IEEE Signal Process Med Biol Symp. 2015 Dec;2015.
[11] Shah, V., Golmohammadi, M., Ziyabari, S., Von Weltin, E., Obeid, I., Picone, J.: Optimizing channel selection for seizure detection. In: 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), IEEE (2017) 1–5
[12] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. PNAS. 2019 Oct 29;116(44):22071–80.