An Optimization Approach to EEG Feature Extraction for the Prediction of Neurological Outcome

Allan R Moser, Jackie T Le, Lys K P Kang

Swarthmore College, Swarthmore, PA, USA

Abstract

As part of the George B. Moody PhysioNet Challenge 2023, our team, Swarthbeat, developed a computational approach that uses electroencephalograms (EEGs) to predict the neurological recovery of patients following cardiac arrest. Our method involved selecting a small number of significant features from a much larger set by optimization based on the PhysioNet Challenge score. Significance was determined using a bagged tree ensemble method. The model for our highest ranking entry was trained using this smaller set of features with adaptive boosting. Our model received a Challenge score of 0.52 (17th out of 36 ranked teams) on the hidden test set.

1. Introduction

The goal of the 2023 George B. Moody PhysioNet Challenge is to develop open-source software to predict good and poor neurological outcome for patients after cardiac arrest using longitudinal electroencephalogram (EEG) and other recordings [1, 2]. Data for this challenge is described in Amorim et al. (2023) [3].

Many features can be extracted from EEG data to classify neurological outcome. Additionally, all, some, or a combination of the 19 EEG electrodes can be selected. Team Swarthbeat's approach to this Challenge is to treat the selection of features and electrodes as an optimization problem with the objective being maximization of the true positive rate given a false positive rate of less than 5%.

2. Method

We used the MATLAB example code provided by the Challenge team as a starting point for our processing. Although several hours of patient data were available, including EEG, electrocardiogram, and other signal types, our method used only the last EEG record that passed our quality check.

Pre-processing

1. For each patient, EEG signals were examined starting from the last available record, working backwards to the first record. Records recorded at greater than 72 hours were skipped.
2. Each EEG signal was resampled to 100 Hz using the MATLAB function *resample* which applies an FIR antialiasing lowpass filter and compensates for the delay introduced by the filter.
3. A five-minute segment was extracted from the middle of each EEG signal. We found that using the entire hour-long signal led to a poorer result. After resampling and extracting the five-minute segment, the resulting signals were demeaned.
4. A quality check was performed by examining the number of zero values and standard deviation for each EEG channel. If the number of zeros exceeded half the signal length or the standard deviation was less than 0.001, for any channel, the record was discarded.
5. The hours of the last useful record and first record were saved to be used as features as described below.
6. If no EEG data was found, or none passed the quality check, that patient instance was flagged as not having EEG data.
7. The unipolar EEG signals were converted to bipolar signals by subtracting adjacent channels. The channel number, position, and notation are shown in Figure 1.

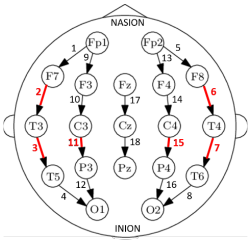

Figure 1. Electrode positions, notation, and numbering for the montage used in this study. Electrodes that optimized the score are shown in red.

Feature Extraction

Three types of features were obtained: (1) patient information; (2) time-domain attributes; and (3) frequency-domain attributes.

Most patient information features were obtained from the header records and are described in Amorim, et al. [3]. Three additional patient information features were

determined from EEG file names: the hours of the first and last EEG recordings and the time difference between these.

Time domain features extracted from the EEG signals were the standard deviations of the 18 bipolar channels.

The majority of the features were determined in the frequency domain and are described below.

- Bandpower – The bandpowers in six frequency regions were calculated using the MATLAB function *bandpower*. We found that the upper limit of the useful frequency range is 26 Hz. The six frequency regions are:

δ: 1 – 3 Hz; θ: 3 – 6 Hz; α: 6 – 10 Hz;
β: 10 – 26 Hz; θ–α–β: 3 – 26 Hz; Total: 0 – 26 Hz.

- Ratio of bandpowers – The ratios of bandpowers were obtained by dividing the bandpowers for the δ/Total, δ/θ, δ/α, δ/β; θ/Total, θ/α, θ/β; α/Total, α/β; and β/Total.
- Slope and R^2 for a linear fit to the log of the power spectral density (PSD) – For each frequency band, a straight line was fit to the log(PSD). The motivation for these features is that the slope of the log of the PSD is indicative of how the power decreases with increasing frequency, *f*. For example, if the bandpower (*P*) decreases as a power of the frequency (i.e., $P(f) \propto 1/f^n$) the slope of the line would be *-n*. R^2 describes of the goodness of fit.
- Standard deviation of the log of the PSD – The standard deviation of the log(PSD) was calculated for each frequency band of the 18 bipolar signals.
- Mean of the magnitude squared coherence estimate – The MATLAB function *mscohere* was used to calculate the magnitude squared coherence estimate defined by

$$C_{xy} = \left|P_{xy}^2\right| \Big/ \left(P_{xx} P_{yy}\right)$$

where P_{xx} and P_{yy} are the PSD's of signals *x* and *y*, and P_{xy} is the cross-PSD of signals *x* and *y*. Signals *x* and *y* correspond to EEG signals on opposite sides of the brain. These features were calculated for all pairs of signals. The central electrode signals, Fz-Cz and Cz-Pz, were not used.
- Cross ratio of bandpowers – The ratio of the difference in bandpowers to average bandpower was calculated for signals on opposite sides of the brain for each frequency band. As was the case for the magnitude squared coherence estimate, these features were calculated for all pairs of signals excluding the central electrodes.

In total, there are 622 features. The feature classes are summarized below:

- 12 patient information features: hospital, age, sex (3 indicators for male, female, other), ROSC, OHCA, VFIB, TTM, first hour, last hour, and hour difference;
- 18 time-domain features: 18 standard deviations;
- 108 bandpower features:18 channels x 6 bands;
- 216 features for PSD: slope, R^2, and standard deviations for 3 x 18 channels x 4 bands (the Total and θ–α–β frequency bands were not used);
- 180 bandpower ratios: 18 channels for the 10 ratios described above;
- 40 coherence features: 8 pairs of channels for 5 bands (the Total frequency band was not used);
- 48 cross PSD ratios: 8 pairs of channels for 6 bands.

These features were calculated once for 597 of the 607 training instances and saved to a spreadsheet. (Ten instances did not have useful EEG data.) This facilitated the investigation of different classification models, and the optimization and cross-validation strategies.

Selection of Classification Models

We used the MATLAB *classificationLearner* app to explore a variety of supervised machine learning classifiers. *classificationLearner* provides 32 models, including several flavors of decision trees, linear and quadratic discriminants, logistic regression, naïve Bayes, support vector machines, k-nearest neighbors, neural networks, and ensemble methods such as boosted and bagged trees. This tool uses k-fold validation (with 5-fold as the default) to compute an overall classification accuracy and provides a confusion matrix and ROC curve for each model.

The best overall accuracy using all 622 features was obtained using ensemble tree methods; in particular bootstrap aggregation (MATLAB *TreeBagger*), adaptive boosting (MATLAB *AdaBoost*), and random under-sampling boosting (MATLAB *RUSBoost*). *AdaBoost* gave the highest overall classification accuracy of 78.1%, however, the false positive rate was 37.1%. *RUSBoost* yielded a lower overall accuracy (76.9%) but had a better false positive rate (23.5%). *TreeBagger* had performance similar to that of *AdaBoost* (overall accuracy: 77.4%, false positive rate 38.9%). All other models available in *classificationLearner* yielded substantially poorer performance with overall classification accuracies on the order of 60% and higher false positive rates. The preliminary studies using *classificationLearner* provided only the overall classification accuracy, and thus were not useful for our objective of optimizing based on the Challenge score. However, they did provide insight that boosted and bagged tree ensemble methods were the best choice for this task, so we limited our consideration of classification methods to these two. Additionally, the lower false positive rate of *RUSBoost* motivated us to utilize balanced datasets (with equal numbers of poor and good outcome instances) in the code developed for our Challenge entries. A final note regarding the choice of classification methods: *TreeBagger* appeared to give the best results in our cross-validation studies and enabled the determination of feature significance. However, we found that *AdaBoost* produced better scores on the validation data used for the official entries.

Optimization Methodology

A 5-fold cross-validation strategy with balanced datasets was utilized to estimate classification scores and feature significance.

Cross Validation Strategy:

For the training data, 63% of the patients (382/607) had

poor outcomes while 37% (225/607) had good outcomes. To obtain balanced datasets for training, the poor-outcome class was randomly sampled to obtain the same number of instances as the good-outcome class. This random sampling was performed some number of times (typically 10 to 20) to ensure that training was exposed to all poor-outcome instances. For each iteration of balanced subset selection, 5-fold cross validation was performed to estimate the Challenge score. Additionally, a measure of feature significance was obtained. As described in the Feature Selection subsection, 597 training instances had good EEG data. Of these, 221 were good-outcome instances. After randomly selecting 221 poor-outcome instances, these two groups were randomized and divided into 5 subsets with 44 instances in four of them and 45 in the remaining one. The good and poor outcome instances in each subset were combined to yield 5 balanced sets containing 88 (or 90 in one case) instances. Training was performed using 4 of these subsets (~352 instances) to obtain a model. This model was then used to predict the outcome probabilities of the instances in the remaining set. The sets were permuted so that each subset was treated as a test set. A score was obtained for each of the subsets, and those scores averaged to obtain a predicted score for each balanced subset iteration. Final predicted scores were obtained by averaging over the balanced subset iterations.

Feature Significance Estimation:

To select an optimum subset of 622 features described above, we utilized the *OOBPredictorImportance* option available in *TreeBagger,* which predicts feature importance using out-of-bag instances. This method involves training an ensemble of trees and using each tree to predict the outcome of instances not seen during training. The out-of-bag accuracy is stored for each tree. The values of features are then randomly perturbed and the out-of-bag accuracy is again calculated. A large increase in error indicates a high importance for that feature while a small change suggests low significance.

Estimates of feature significance were aggregated in an initial phase of the cross-validation strategy. For this phase, 20 randomly balanced subsets were used. With 5-fold cross validation, this yielded 100 significance estimates for each of the 622 features. These individual significance estimates were then averaged to obtain a final significance value. Figure 2 shows a bar graph of the significance for these features.

Selection of features to maximize the Challenge score:

The selection of features to maximize the Challenge score was accomplished by varying the threshold for feature significance. For each threshold, the subset of features exceeding that value was used with the cross-validation strategy to make a prediction for the Challenge score. The threshold yielding the maximum score resulted in 24 features selected from attributes scattered across several electrodes and attribute groups. The score predicted by cross validation was 0.581, however, our official Challenge entry using these features for the validation data produced a score of only 0.478. This strategy turned out to be an ineffective one since all of the training data was exposed during the feature significance determination, thus resulting in overtraining.

Accordingly, we adopted an alternative strategy yielding better scores that involved eliminating groups of attributes. If a class of attributes had low significance for all EEG channels, we eliminated that group from consideration. This resulted in a much smaller set of features: Age, VFIB; δ, θ, α, and β bandpowers; ratio of bandpowers for δ/θ, and δ/α; slope and R^2 for only the δ band; and coherence for δ, θ, and α bands. Using all EEG channels, this reduced the number of features to 170. The Challenge entry using these features resulted in a score of 0.687.

A final phase of optimization was used to determine if a subset of EEG channels would improve our score. Rather than investigating individual channels, we chose to consider channels in pairs for opposite sides of the brain (e.g., Fp1-F7 / Fp2-F8), since the coherence estimate utilized these pairs. Additionally, we combined the central channels (Fz-Cz / Cz-Pz) into one pair. With 9 pairs, there are 511 possibilities so we were able to exhaustively search for the optimum combination. Several combinations yielded similar scores, however, all of the combinations with high scores in our cross-validation study included either F7-T3 / F8-T4, or T3-T5 / T4-T6 or both. For our Challenge entries, we choose two of these combinations; one with a larger number of channels (10): F7-T3 / F8-T4; T3-T5 / T4-T6; Fp1-F3 / FP2-F4; C3-P3 / C4-P4; P3-O1 / P4-O2 and one with a smaller number (6): F7-T3 / F8-T4; T3-T5 / T4-T6; C3-P3 / C4-P4. The entry using 6 channels produced our best score on the validation data. The 6 channels are show in red on Fig. 1.

In order to implement balanced sets for the code used in our Challenge entries, we randomly sampled the poor-outcome cases to obtain the same number as good-outcome cases ten times and constructed ten different models. The code used to predict the outcome for validation or test data determined the class probability by averaging the probabilities for the ten models. If no EEG data was available, the value of VFIB was used for the prediction.

3. Results

We reduced the initial 622 features from patient-information, 1 time-domain class, and 7 frequency-domain classes to 170 by thresholding on a significance measure determined using bootstrap aggregation (*TreeBagger*) out-of-bag instances. If all features of a given class had significance falling below a threshold, determined by maximizing the Challenge score, that class was eliminated. We further reduced the number of features to 59 by exhaustively searching combinations of EEG channels that

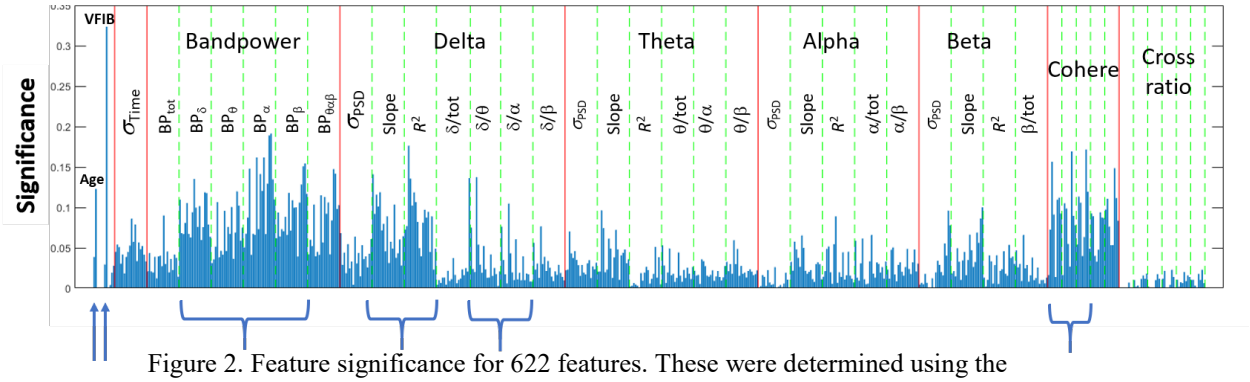

Figure 2. Feature significance for 622 features. These were determined using the *OOBPredictorImportance* option in *TreeBagger*. The blue brackets and arrows at the bottom of the figure indicate which features were selected by optimizing the Challenge score.

maximized the Challenge score. Using multiple balanced datasets and an adaptive boosting classifier (*AdaBoost*) we achieved the scores shown in Table 1.

Training	Validation	Test	Ranking
0.53±0.17	0.72	0.52	17/36

Table 1. True positive rate at a false positive rate of 0.05 (the official Challenge score) for our final selected entry (team Swarthbeat), including the ranking of our team on the hidden test set. We used 10 randomly selected balanced datasets and 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

4. Discussion and Conclusions

Our approach involved starting with a large number of features which we expected to have a great deal of redundancy. The strategy of winnowing this set by optimizing on the Challenge score revealed that only about 1/10 of these features were needed to achieve a relatively high score. Patient features that proved significant were age and shockable rhythm (VFIB), with VFIB being the most significant feature. This is not surprising since it is known from other studies that a non-shockable rhythm has a high correlation with brain death after resuscitation [4]. One also expects that younger people, on average, have a higher probability of recovery. All other significant features were obtained from the frequency domain. These include the bandpowers in all frequency bands. For most of the other features, the delta band proved most important. Fitting a line to the log of the PSD for the delta band proved particularly informative, with a steeper slope and better fit being correlated with better neurological outcome. An opportunity for future work would be to better understand the relationship of these features to the pathophysiology of hypoxic-ischemic brain injury.

A drawback of our method for selecting features, based solely on significance, is overtraining. Additionally, the score was only weakly correlated with the selection of EEG channels. We believe maintaining localization information from electrodes is important, however, it may prove useful to investigate signals constructed from other combinations of electrodes. Another potential area for improvement would be to incorporate information based on pathophysiology domain knowledge in addition to our frequency-domain derived features.

Acknowledgments

This work used the Strelka Computing Cluster, which is supported by the Swarthmore College Office of the Provost. Conference attendance for students was supported by the Sigma Xi Conference Fund and the Provost's Office Professional Opportunities Fund.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–e220.

[2] Reyna MA, Amorim E, Sameni R, Weigle J, Elola A, Bahrami Rad A, et al. Predicting neurological recovery from coma after cardiac arrest: The George B. Moody PhysioNet Challenge 2023. Computing in Cardiology 2023;50:1–4.

[3] Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, Karukonda V, et al. The International Cardiac Arrest Research (I-CARE) Consortium Electroencephalography Database. Critical Care Medicine, October 19, 2023.

[4] Sandroni T, et al. Brain injury after cardiac arrest: pathophysiology, treatment, and prognosis. Intensive Care Medicine 2021; 47(12):1393-1414.

Address for correspondence:

Allan R Moser
500 College Ave., Swarthmore College, Engineering Dept., Swarthmore, PA, USA 19081
amoser2@swarthmore.edu