Prediction of Functional Recovery Post-Cardiac Arrest Using an Ensemble of Extreme Gradient-Boosted Trees

Matthew Kolisnyk [1], Xiaoyu Wang[1], Chao Guo[2], Shigeng Xie[2], Karnig Kazazian[1], Loretta Norton[1,3], Teneille Gofton[1,4], Saptharishi Lalgudi Ganesan[1,4], Adrian M. Owen[1] and Derek Debicki[1,4]

[1] Western University, London, Canada
[2] Dalian University of Technology, Dalian, China
[3] King's University College, London, Canada
[4] Lawson Health Sciences Centre, London, Canada

Abstract

Predicting the outcome of critically ill patients after cardiac arrest is a substantial clinical challenge. As part of the 2023 George B. Moody PhysioNet Challenge, we used an ensemble of extreme gradient boosted (XGB) trees, trained on electroencephalogram (EEG) data across three distinct levels of analysis to predict neurological outcomes. To this end, we preprocessed raw EEG recordings via filtering and motion correction and extracted several features, including connectivity, power spectral density (PSD), and coherence. The first step of model training and optimization occurred at the Recording level, where we trained XGB trees on each feature set (e.g., PSD) derived from every EEG recording. These models predicted the neurologic outcome for each recording. At the Patient level, we computed the median of each feature set for each patient, which we used to make patient-specific predictions. Finally, the Challenge level merged and optimized predictions from the previous levels, thus synthesizing patient and recording specific predictions. Our approach produced successful results on the test set, with Challenge Scores of 0.371 (12 hrs), 0.480 (24 hrs), 0.569 (48 hrs), and 0.678 (72 hrs), resulting in a final ranking of 9/36 (team: WesternUni). The approach showed particular promise at early timepoints, placing 2nd when only using EEG available 12-hours post-ictus.

1. Introduction

Providing an early and accurate prognosis for patients who remain in a coma following cardiac arrest is one of the greatest challenges in critical care [1, 2]. Electroencephalography (EEG) is one diagnostic tool used in the intensive care unit (ICU) to make clinically relevant decisions [3-5]. EEG measures electrical activity produced by the brain via electrodes placed on the scalp. Many different patterns of electrical activity have been identified by clinicians who have characterized how the patterns are related to the course and progression of the injury [5].

However, prognostication using EEG is far from perfect. Indeed, there is debate about the true meaning of specific EEG patterns and whether they are ultimately related to patient outcomes [3-5]. These problems are exacerbated by inter-reliability issues between electroencephalographers when identifying these EEG patterns via visual inspection alone [6]. Hence, there remains a need for objective and quantitative assessments of EEG's relation to neurologic recovery.

Machine learning is emerging as the quantitative method of choice to make predictions from EEG data [7, 8]. One particularly successful machine learning approach is extreme gradient-boosted (XGB) trees [9, 10]. XGB has been applied to several domains and has been the model of choice for several machine learning competitions [9, 11]. XGB learns by aggregating predictions made by a base classifier (e.g., decision trees [12]) in an iterative fashion. By adding successive decision trees to address error from previous steps, the model can learn to predict increasingly difficult training examples and at increasingly faster speeds than other tree-based approaches [13]. This approach relies on the foundation of an ensemble of *weak learners* – the idea that many minimally predictive decision trees, once combined, can produce an effective predictive model (i.e., *strong learner*) [10].

As part of the 2023 George B. Moody PhysioNet Challenge [14], we built a multi-level model that uses XGB trees trained on EEG measures detectable within the first days of ICU admission to inform the likelihood of recovery of critically ill patients.

2. Method

The PhysioNet Challenge consists of ~50,000 EEG recordings from ~1000 cardiac arrest patients [14-16]. Each recording reflected at most an hour of continuous EEG taken before 72 hours post-ictus. Functional outcome was measured by the cerebral performance category (CPC) score, which categorized patients into good (CPC 1-2) and

poor outcome (CPC 3-5) groups. Data from ~600 patients was made available for model training.

2.1. EEG preprocessing

Five EEG electrodes (F3, P4, C3, T4, O1) were band-pass filtered [0.1-30 Hz] and motion-corrected. Motion correction was applied in two steps. First, the moving standard deviation was computed across each recording. Data that exceeded five standard deviations were removed. The moving standard deviation was then recomputed, and segments exceeding four standard deviations were removed from the data. Additionally, data with a moving standard deviation below $10^{-4}\,\mu V$ were also excluded. Contiguous segments of clean data were combined in 200 second epochs.

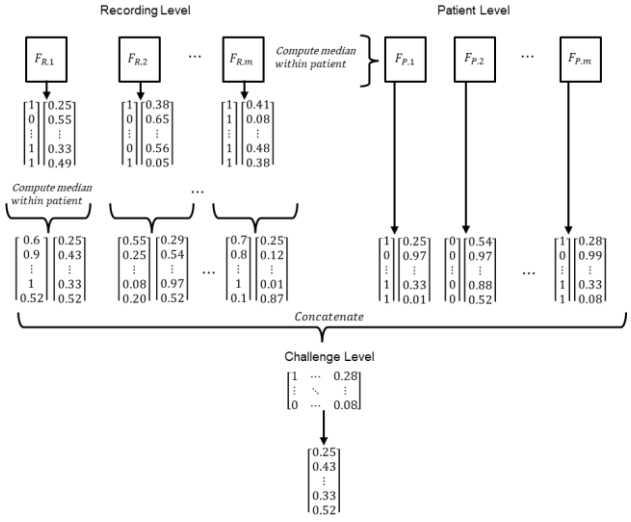

Figure 1. Diagram of the multi-level modelling approach. Our approach consisted of Recording, Patient and Challenge levels. Sets of features are denoted as $F_{L,i}$ where L denotes the level (e.g., R stands for Recording) and i denotes a feature set (e.g., PSD). Arrows denote the training and optimization of an XGB model. Each model outputs an outcome prediction and probability, shown by two vectors following each arrow. The outcome and prediction probability were concatenated for all features in both the Patient and Recording levels, and a final XGB model was trained on predictions from those levels and optimized to maximize the Challenge Score, thus producing a final outcome probability prediction.

2.2. Feature Extraction

Five sets of features were computed on each recording for each epoch of the preprocessed data. Feature extraction was conducted on Python (Version 3.81) using mne, scipy, and FLOOF packages [17, 18]. Power spectral density (PSD) was computed for delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz) and beta (13-30 Hz) frequencies for every 1 Hz band. Spectral coherence and weighted phase lag index (wPLI) were similarly computed for delta, theta and alpha frequencies. Next, the aperiodic component was computed on the average of the 5 EEG electrodes. Finally, basic statistical properties of the signal (e.g., mean, variance, peak, kurtosis) were calculated on the preprocessed data. The 'trajectory' of these features was calculated as the mean difference between successive recordings to obtain information about how these features change over time. Features were averaged across epochs for each patient, and a two-step data cleaning was applied to remove outliers, which involved removing features exceeding five standard deviations and then four standard deviations. While the total number of rejected recordings varied between features, approximately 1/3 of recordings were removed for each EEG feature. Finally, patient clinical information, including age, gender, return of spontaneous circulation, targeted temperature management, and the presence of a shockable rhythm, as well as data quality information (e.g., number of clean epochs, total number of clean recordings), were also used as features.

2.3. Multi-level model training

Model training occurred at three levels: Recording level, Patient level, and Challenge level (see Figure 1).

Recording Level. XGB trees (implemented using the xgboost package [14]) were individually trained on each feature set (e.g., PSD, wPLI). The outcome measure was the functional outcome of a patient assigned to each of their recordings. The tree-structured Parzen Estimators Approach (implemented via the Hyperopt package [19]) was used for hyperparameter optimization. The hyperparameters were optimized to maximize the balanced accuracy of five stratified folds of the training data (implemented via sci-kit learn [20]). Optimization occurred for sixteen hyperparameters, including alpha, lambda, eta, max depth, max delta step, column sampling, and subsampling. Each trained model used the training data to predict outcome and outcome probability for each recording.

Patient Level. The features were aggregated for each patient by taking the median value obtained at the Recording level. Then, XGB models were optimized for each set of features identically to the Recording level. A XGB model using only patient clinical information (e.g., age, gender, ROSC) was also optimized at this stage. Each trained model predicted an outcome and outcome probability for each patient. Notably, when no EEG features were available for a patient (e.g., due to poor data quality), the model trained on only clinical information.

Challenge Level. The median outcome prediction and

probability from the Recording level were computed for each patient and concatenated with the outcome prediction and probability from the Patient level. This resulted in a matrix of outcome predictions and probabilities across features and levels. The 'trajectory' of features, as well as data quality features, were also included in this step. Optimization proceeded identically as the preceding levels except with the objective of maximizing the PhysioNet Challenge Score. This model was used to classify good and poor outcomes as well as CPC scores at 12-, 24-, 48-, and 72-hours post-ictus.

3. Results

Our approach produced successful predicted outcomes on the test set, with Challenge Scores of 0.371 (12 hrs), 0.480 (24 hrs), 0.569 (48 hrs), and 0.678 (72 hrs), resulting in a final ranking of 9/36. Similar results were obtained on the validation set, with Challenge Scores of 0.37 (12 hrs), 0.69 (24 hrs), 0.51 (48 hrs) and 0.52 (72 hrs). For accuracy metrics for train, validation, and test sets, see Figure 2B.

Table 1. PhysioNet 2023 Challenge Score Results

Time	Challenge Score	Ranking
12 hours	0.371	2/36
24 hours	0.480	8/36
48 hours	0.569	11/36
72 hours	0.678	9/36

The importance of different features across the different levels is reported in Figure 2A. Generally, features gathered at the Recording level have higher average importance (as measured by gain) than at the Patient level. Moreover, some features are important across time points (e.g., PSD), whereas others show greater importance at different time points (e.g., coherence at later time points). Generally, when EEG recordings are fewer in number, the model relied more on clinical information.

4. Discussion

Our method successfully identified the functional outcome of patients following cardiac arrest. Particularly, our approach excelled at predicting neurologic recovery at early time points, including the 2nd highest score in the Challenge at the 12-hour mark, and 1st place on the validation set at 24 hours. Notably, it is typically thought that these early points following resuscitation are too early for prognostication based on clinical assessment alone due to variability in medical stability. However, our results suggest that certain EEG features may contain useful information even at these early timepoints. Of these, PSD and the aperiodic component tended to be the most informative. However, our technique was competitive at later time points, placing 9th in the competition. It may be the case that increasing the number of recordings may have resulted in additional outliers that were not sufficiently cleaned. Given these results, future work should investigate systematic quality and feature differences across time points.

Our approach has several advantages. First, it is flexible and could incorporate numerous features beyond the ones reported here (e.g., weighted-symbolic mutual information [18], motifs [19]), as well as accommodate different classifiers (e.g., deep learning models), which can be different learning objectives. Additionally, the framework is well-suited for making predictions on individual

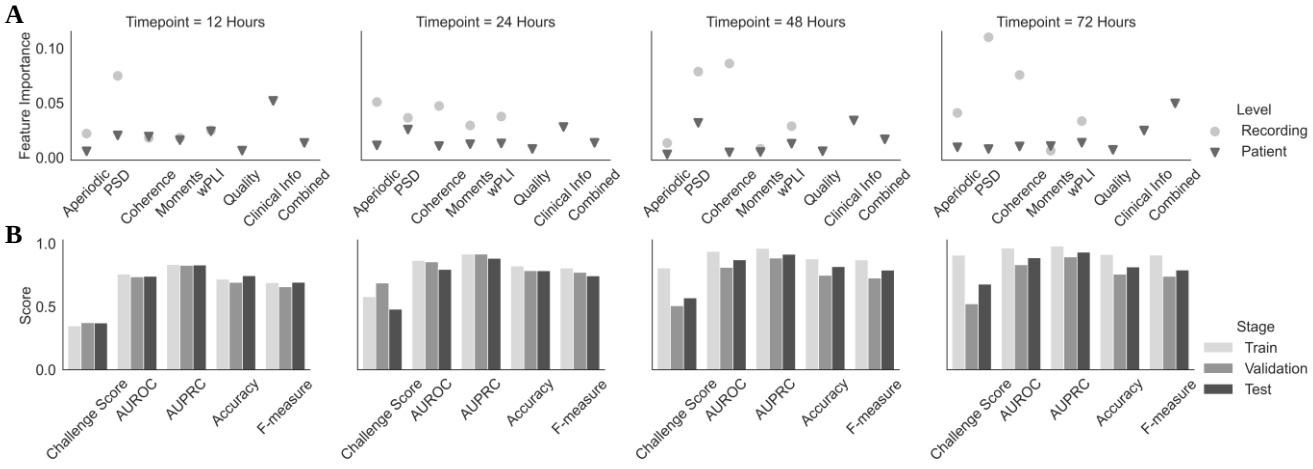

Figure 2. Panel A shows a dot plot with feature importance (gain) for our model calculated on the training set. Features were computed independently at either the Recording level (circles) or Patient level (arrow) across four timepoints (12,24,48,72 hours). For simplicity, the gain is averaged within each feature. For example, the feature importance for PSD is the gain for each of its individual features (e.g., outcome prediction, outcome probability, trajectory information), which is then averaged. Panel B shows bar plots of the scores of various accuracy metrics for predicting outcome -- computed on training, validation, and test sets across each timepoint.

recordings. It could be extended to make predictions on individual epochs of data, signaling its potential for predictions to be made in real-time within ICU settings.

There remains potential for improvement at each step and level of our machine learning approach. For the purposes of the Challenge, a 72-hour restriction on training the data was imposed, which meant that expediency had to be prioritized, sometimes at the cost of robust data quality checking. For example, a lack of proper data cleaning led to many outliers within the feature set, which we addressed only with a quick two-step outlier detection method. Another limitation was using only five channels for feature selection and excluding electrocardiography, which likely reduced model performance. Given that our approach used three model training levels, including a computationally heavy approach that used every recording (rather than averaging recordings across patients), it would have benefited from increased training time.

In conclusion, we used a novel multi-level machine learning approach to predict functional outcomes of patients who remained unresponsive following cardiac arrest using EEG. This study adds to the growing body of research supporting the combination of EEG and machine learning to assist clinical prognostication [7, 8]. These results highlight this machine learning technique's potential to provide early and accurate prognostication, which is desperately needed to improve goal-directed patient care in the ICU.

Acknowledgments

Thank you to the patients, caregivers, and hospital staff for their essential contributions to this study.

References

[1] Kamps MJA, *et al.* "Prognostication of neurologic outcome in cardiac arrest patients after mild therapeutic hypothermia: a meta-analysis of the current literature," *Intensive Care Med*, vol. 39, no. 10, pp. 1671-1682, 2013

[2] Weijer C, Bruni T, Gofton T, et al. "Ethical considerations in functional magnetic resonance imaging research in acutely comatose patients." *Brain*, vol. 139, no. 1, pp. 292-299, 2016

[3] Kondziella, D., *et al.*, 'European Academy of Neurology Guideline on the Diagnosis of Coma and Other Disorders of Consciousness'. *European Journal of Neurology*, vol. 27 no. 5, pp. 741–56, 2020.

[4 D. Friedman, J. Claassen, and LJ. Hirsch, "Continuous Electroencephalogram Monitoring in the Intensive Care Unit'. *Anesthesia and Analgesia*, vol. 109, no. 2, pp. 506–23, 2009.

[5] LJ. Hirsch, *et al.*, "American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2012 Version". *Journal of Clinical Neurophysiology* vol. 30, no. 1, pp. 1-27, 2013

[6] E. Westhall *et al.*, 'Interrater variability of EEG interpretation in comatose cardiac arrest patients', *Clinical Neurophysiology*, vol. 126, no. 12, pp. 2397–2404, Dec. 2015.

[7] M. Amiri *et al.*, 'Multimodal prediction of residual consciousness in the intensive care unit: the CONNECT-ME study', *Brain*, Sep. 2022.

[8] J. Claassen *et al.*, 'Detection of Brain Activation in Unresponsive Patients with Acute Brain Injury', *N Engl J Med*, vol. 380, no. 26, pp. 2497–2505, Jun. 2019.

[9] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 785–794, Aug. 2016.

[10] J. H. Friedman, 'Greedy function approximation: A gradient boosting machine.', *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[11] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin, 'Parallel boosted regression trees for web search ranking', in *Proceedings of the 20th international conference on World wide web*, in WWW '11. New York, NY, USA: Association for Computing Machinery, pp. 387–396, Mar. 2011.

[12] J. R. Quinlan, 'Induction of decision trees', *Mach Learn*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[13] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. The MIT Press, 2012.

[14] Reyna MA, Amorim E *et al.*, 'Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023'. *Computing in Cardiology 2023*, vol 50, pp. 1-4, 2023

[15] Amorim E, *et al.*, 'The International Cardiac Arrest Research (I-CARE) Consortium Electroencephalography Database', *Critical Care Medicine*, October 2023.

[16] Goldberger AL, *et al.*, 'PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals', *Circulation*, vol. 101, no.23, June 2000.

[17] A. Gramfort *et al.*, 'MEG and EEG data analysis with MNE-Python', *Frontiers in Neuroscience*, vol. 7, 2013.

[18] T. Donoghue *et al.*, 'Parameterizing neural power spectra into periodic and aperiodic components', *Nat Neurosci*, vol. 23, no. 12, Dec. 2020,

[19] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[20] T. Donoghue *et al.*, 'Parameterizing neural power spectra into periodic and aperiodic components', *Nat Neurosci*, vol. 23, no. 12, Art. no. 12, Dec. 2020.

[21] J.-R. King *et al.*, 'Information Sharing in the Brain Indexes Consciousness in Noncommunicative Patients', *Current biology*, vol. 23, Sep. 2013.

[22] C. Duclos *et al.*, 'Brain network motifs are markers of loss and recovery of consciousness', *Sci Rep*, vol. 11, no. 1, p. 3892, Feb. 2021.

Address for correspondence:
Matthew Kolisnyk
Western Institute of Neuroscience, Western Interdisciplinary Research Building, Western University, 1151 Richmond Street, London, Ontario, N6A 3K7. mkolisny@uwo.ca