Functional Outcome Prediction After Cardiac Arrest Using Machine Learning and Network Dynamics of Resting-State Electroencephalography

Charlotte Maschke[1,2], Kira Dolhan[1,2], Beatrice P. De Koninck[1,2,4,5], Miriam Han[1,2], Stefanie Blain-Moraes [1,2,3]

[1]Montreal General Hospital, McGill University Health Centre, Montreal, Canada
[2]Integrated Program in Neuroscience, McGill University, Montreal, Canada
[3]School of Physical and Occupational Therapy, McGill University, Montreal, Canada
[4]Department of Psychology, University of Montreal, Montreal, Canada
[5]CIUSSS du Nord-de-l'Île-de-Montréal Research Center, Montreal, Canada

Abstract

Accurately predicting outcome after cardiac arrest (CA) is necessary to address the ethical, economical, and societal issues following this highly prevalent condition in the Western population. This research was conducted as part of the Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023 (team name: EEGnition). We used machine learning on spectral features and network dynamics of resting-state electroencephalography (EEG) to predict the long-term functional outcome of comatose patients following CA. We extracted features from six categories of data: (i) demographic and clinical; (ii) electrocardiogram recording; (iii) general EEG recording; (iv) EEG spectral analysis; (v) EEG criticality and complexity analysis; and (vi) EEG functional connectivity. A binary classification into good and poor outcomes was then performed using AutoML - an automatized machine learning pipeline. Our model received a challenge score of 0.525 (ranked 16th out of 36 teams in the final challenge evaluation) on the hidden test set. The best performing features included the number and frequency of spectral peaks, Shannon entropy and functional connectivity. Our approach provides evidence for the prognostic value of network dynamics from EEG recorded acutely following CA.

1. Introduction

The widespread occurrence of cardiac arrest (CA) generates a myriad of ethical, economical, and societal consequences. Approximately 80% of CA patients who are successfully resuscitated are comatose as a result of post-cardiac arrest brain injury [1]. Among these, 50-60% either do not survive or live with severe disability [2]. Critical decisions about how aggressively to pursue care are made in the days following the return of spontaneous circulation (ROSC), including the decision of whether or not to withdraw life-sustaining treatment. Thus, early and accurate identification of patients who are likely to make a meaningful recovery is critical in shaping further treatment and care trajectories [1].

Resting-state electroencephalograms (EEG) are often recorded for up to 72 hours in the intensive care unit post-CA; this data has been shown to have strong prognostic potential [3]. Automated machine learning (ML) classification models may be applied to this EEG data to identify the most promising features for predicting patient prognosis and optimizing prognostic accuracy [4].

As part of the *2023 George B. Moody PhysioNet Challenge,* we used a multi-site meta-database [5]–[7], to predict the long-term functional outcome of comatose patients following CA. We developed an open-source ML algorithm to predict the 6-month functional outcome of CA patients using clinical data and physiological signals.

2. Methods

2.1 Dataset

We analyzed the open-source I-CARE database (International Cardiac Arrest Research Consortium Database) [5]. 607 patients were included in the training dataset, 107 patients were used as a hold-out validation set, and 306 patients were included in the hold-out test set. The dataset included EEG, electrocardiogram (ECG), demographic information, and treatment information. The dataset also provided functional outcome measures using the Cerebral Performance Category (CPC) scale at 6 months post-ROSC. For this study, CPC scores were binarized into two groups: a CPC score of 2 and lower was considered a "Good Outcome" and scores of 3 to 5 were considered a "Poor Outcome".

2.2. Preprocessing

EEG data was preprocessed using the MNE Python toolbox [8]. Data was bandpass-filtered between 0.1 and 40 Hz, notch-filtered at the corresponding utility frequency and resampled to 125 Hz. The data was then segmented

into non-overlapping 10 second epochs. For further analyses, we used a subject-specific threshold (i.e., lowest standard deviation to keep 5 minutes of data) to select 5 minutes (i.e., 30 epochs) which were least contaminated by non-physiological artifacts from each hour of recording. To minimize computational resources, only ten percent of the available EEG recordings per participant were analyzed (e.g., for 30 hours of EEG data, we analyzed every third hour of available recording). We then extracted features from six categories: (i) demographic and clinical; (ii) ECG; (iii) general EEG recording; (iv) EEG spectral analysis; (v) EEG criticality and complexity; and (vi) EEG functional connectivity.

2.2. Feature Extraction

Demographic and clinical features

We extracted patient sex (male/female/unknown) and age (years) from the available demographic data. From the patient clinical information, we included the treating hospital, time from the CA to ROSC, the location of the CA (i.e., in or out of hospital), the type of cardiac rhythm recorded upon resuscitation (i.e., shockable rhythm versus non-shockable) and the use of targeted temperature management (TTM) during the recording (yes/no).

ECG features

ECG features were calculated using the first available hour of data. Time- and frequency-domain features of heart rate variability were calculated using Neurokit2 toolbox [9].

General EEG recording features

General recording information included the number of existing recordings, patient-specific preprocessing thresholds, the time post-ROSC of the first available recording and each channel's signal mean and standard deviation, which was calculated on epochs of each available hour of recording.

Spectral features

We calculated spectral power in the delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz) and gamma (30-40 Hz) bandwidths. Power spectral density was calculated on every channel individually, using Welch's method, and averaged over the corresponding frequency bandwidth. We extracted the spectral slope, number of existing peaks and peak frequency, and power of the first identified peak using the fitting oscillations and 1/f algorithm [10]. Spectral features were calculated on epochs of each available hour of EEG data.

Criticality and Complexity features

Criticality-related EEG features include the deviation from the criticality coefficient, avalanche repertoire size, branching ratio and the Fano factor, which were calculated using a custom Python toolbox [11]. Avalanche criticality measures were estimated on epochs of each available hour of recording. The complexity of the EEG signal was estimated using Lempel-Ziv complexity, Petrosian fractality, Katz fractal dimension, permutation entropy, Shannon entropy and the number of optimal embedding dimension. Due to limited computational resources, only Lempel-Ziv complexity was estimated on epochs of each available hour of recording. Other complexity measures were estimated solely on epochs from the first available recording. All complexity measures were calculated on each epoch and channel individually using the Neurokit2 toolbox [9] and subsequently averaged over time.

Functional Connectivity features

Functional connectivity (FC) was estimated using the weighted (wPLI) and directed phase lag index (dPLI) in the theta, alpha and beta bandwidths [12], [13]. Both matrices were calculated on each epoch individually using MNE-Python [8] and were averaged over time. FC matrices were averaged to yield the mean connectivity from one electrode to all other existing electrodes. Due to limited computational resources, FC features were calculated solely for epochs from each patient's first available hour of EEG.

Longitudinal feature development

We recorded the longitudinal development of features that were calculated on epochs from all available hours of recording by calculating the slope of the linear regression for the individual whole-brain averaged feature over time.

Spatial averaging of features

All features (excluding avalanche features) were calculated on every channel individually. Due to variability in available channels between patients, individual values were averaged over brain regions (i.e., anterior, posterior, left, right, frontal, central, parietal, temporal, occipital). Region-averaged features of the first available recording were combined with time-resolved and demographic features, yielding a final total of 302 features.

2.3. Machine learning

ML models were trained using the six feature categories, resulting in a total of 302 training features. Model training and evaluation was conducted using AutoML mljar-supervised [14] – an automated Python ML package using scikit-learn [15]. A binary classification model was initialized using the algorithms compete mode, 100 golden features, and 4 top models to improve (mode = 'Compete', golden_features = 100, top_models_to_improve = 4) and trained to predict the binary functional outcome measure (i.e., good or poor).

AutoML was employed to test a variety of ML

algorithms on the training set, to quantify algorithm performance, and to perform automatized hyperparameter tuning. For classification models, AutoML evaluated model performance using the logloss metric. Within one run, AutoML performed and compared models from eight categories: linear models, decision tree, random forest, extra trees, light gradient-boosting machine (LightGBM), eXtreme Gradient boosting (XGBoost), CatBoost, neural network and nearest neighbors. The final ensemble algorithm was then constructed by combining various ML algorithms into one conglomerate algorithm, where individual algorithms are weighted according to their performance using a fivefold cross validation. The final model predicted the binary outcome and the probability of an unfavorable outcome.

Feature selection

Missing feature values were inputted using the median value of the feature. Features were normalized before ML training occurred. AutoML calculated "golden features" by combining pairs of features (i.e., multiplication, division, addition, or subtraction of the two features) to obtain transformed potential 'golden' feature. This process was repeated for all possible pairwise feature combinations (up to a maximum of 250,000 pairs). New features were ranked in terms of logloss values. The 100 best performing features were then selected as "golden features" and used as inputs for certain ML algorithms. We also identified 'selected features' -- original features that contribute most to the ML algorithm with the smallest logloss value – by using pseudo features with random values as additional input for the best performing ML algorithm. We then conducted a permutation-based feature importance analysis. If features were ranked as more important than the random feature for more than 2 of the five folds during validation, they were considered selected features. Features which were not included in golden or selected features were dropped from the given ML algorithm.

2.4. Model evaluation

The best performing models were selected using logloss values of binary predictions on a fivefold cross validation of the training data. Additionally, we calculated the model's accuracy (Acc), precision, and area under the curve (AUC). The model's performance on the hidden cross-validation and test data was evaluated using the challenge score: the true positive rate (TPR) for predicting a poor outcome given a false positive rate (FPR) of less than or equal to 5% at 72 hours post ROSC. The challenge score was evaluated four times individually, using data from the first 12-, 24-, 48- and 72-hours post-ROSC.

3. Results

Performance on training data

The models with the best performance were Xgboost, CatBoost, LightGBM, Neural Network and RandomForest. Individual model performance on the training set is reported in Table 1. The Ensemble method combined these models to perform a final decision (see Table 1), achieving a classification accuracy of 74%.

Table 1. Individual performance of selected model and performance of Ensemble method.

Model	Logloss	Acc	Precision	AUC
Xgboost	0.54	0.74	0.94	0.77
CatBoost	0.56	0.73	1	0.75
LightGBM	0.56	0.72	1	0.75
RandomForest	0.55	0.71	1	0.75
Neural Net	0.69	0.71	1	0.71
Ensemble	**0.53**	**0.74**	**1**	**0.78**

The top ten percent of features which best distinguished patients according to their functional outcome were: the presence of shockable rhythms, the number and frequency of spectral peaks, Shannon entropy, and the functional connectivity (wPLI and dPLI) in the theta, alpha and beta frequency band.

Final challenge performance

On the holdout validation data, our model achieved a challenge score of 0.358 at 12 hours, 0.463 at 24 hours, 0.522 at 48 hours and 0.627 at 72 hours post-ROSC. On the holdout test data, our model achieved a challenge score of 0.243 at 12 hours, 0.411 at 24 hours, 0.475 at 48 hours and 0.525 at 72 hours post-ROSC (see Table 2).

Table 2. Challenge Score (True positive rate at a FPR of 0.05) for our final selected entry (team EEGnition), including the ranking of our team on the hidden test set.

Training	Validation	Test	Ranking
0.843	0.627	0.525	16/36

4. Discussion and Conclusion

The aim of this study was to develop a ML model that could accurately predict long-term functional recovery in patients post-CA using clinical data and longitudinal EEG and ECG data. Using the resting-state EEG recordings within 72 hours post-ROSC, we were able to predict the functional outcome with a challenge score of 0.525 using an Ensemble method.

Our ML model strongly relied on features from the first available hour of EEG recording, rather than the longitudinal features. These results corroborate with previous studies showing that early EEG recordings have a greater prognostic information with higher specificity

compared to post-48-hour recordings, even in the presence of TTM and light to moderate sedatives [16]. The inclusion of EEG spectral features and dynamic network properties – such as entropy and functional connectivity – in the ML algorithm yielded greater discriminative power, highlighting their potential importance in predicting outcomes in this patient population.

The results of this study needs to be interpreted in light of several limitations. First, automized ML tools such as AutoML are a valid framework for the exploration of different types of algorithms but provide relatively low interpretability and are not well suited for the clinical use. Second, our model's predictability may be influenced by confounding factors such as the inclusion of patients who were withdrawn from life-sustaining treatment in the natural death category (CPC of 5) and the potential inclusion of deaths resulting from non-brain related complications, such as metabolic conditions, which fall beyond our model's scope. Third, TTM treatment requires sedative administration during the initial 24 hours post-CA, which may confound the accuracy of our prediction model due to the lack of information regarding administration parameters. Fourth, while automated EEG preprocessing pipelines are widely available for recordings from healthy adults, the automatic preprocessing of EEG from brain-injured patients without the validation of a trained experimenter can introduce severe confounds, as pathological signal characteristics can be automatically removed as noise. The validation of data quality by a human experimenter would be crucial for the development of a clinical tool from this work.

In summary, we demonstrated the feasibility of an automated ML model for the prediction of long-term functional outcome post CA using network dynamics and spectral properties of resting-state EEG. Future endeavors dedicated to optimizing predictive tools that minimize the false positive rate can have a tremendous impact on the treatment decisions immediately post-resuscitation for this clinical population.

References

[1] C. Sandroni, T. Cronberg, and M. Sekhon, "Brain injury after cardiac arrest: pathophysiology, treatment, and prognosis," *Intensive Care Med*, vol. 47, no. 12, pp. 1393–1414, Dec. 2021, doi: 10.1007/s00134-021-06548-2.

[2] S. A. Amacher *et al.*, "Long-term Survival After Out-of-Hospital Cardiac Arrest: A Systematic Review and Meta-analysis," *JAMA Cardiology*, vol. 7, no. 6, pp. 633–643, Jun. 2022, doi: 10.1001/jamacardio.2022.0795.

[3] J. H. Lee, D. H. Lee, B. K. Lee, D. K. Kim, and S. J. Ryu, "Association Between Procalcitonin Level at 72 Hours After Cardiac Arrest and Neurological Outcomes in Cardiac Arrest Survivors," *Ther Hypothermia Temp Manag*, vol. 13, no. 1, pp. 23–28, Mar. 2023, doi: 10.1089/ther.2022.0019.

[4] W.-L. Zheng *et al.*, "Predicting Neurological Outcome in Comatose Patients after Cardiac Arrest with Multiscale Deep Neural Networks," *Resuscitation*, vol. 169, pp. 86–94, Dec. 2021, doi: 10.1016/j.resuscitation.2021.10.034.

[5] Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, Karukonda V, Lee JW, Herman ST, Adithya S, Gaspard N, Hofmeijer J, van Putten MJAM, Sameni R, Reyna MA, Clifford GD, Westover MB. "The International Cardiac Arrest Research (I-CARE) Consortium Electroencephalography Database." *Critical Care Medicine* 2023 (in press); doi:10.1097/CCM.0000000000006074

[6] 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. Accessed: Sep. 08, 2023. [Online]. Available: https://moody-challenge.physionet.org/2023/

[7] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.

[8] MNE Developers, "MNE — MNE 0.21.2 documentation." Accessed: Dec. 01, 2020. [Online]. Available: https://mne.tools/stable/index.html

[9] D. Makowski *et al.*, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav Res*, vol. 53, no. 4, pp. 1689–1696, Aug. 2021, doi: 10.3758/s13428-020-01516-y.

[10] T. Donoghue *et al.*, "Parameterizing neural power spectra into periodic and aperiodic components," *Nat Neurosci*, vol. 23, no. 12, pp. 1655–1665, Dec. 2020, doi: 10.1038/s41593-020-00744-x.

[11] J. O'Byrne, "EdgeofPy." Sep. 20, 2023. Accessed: Oct. 13, 2023. [Online]. Available: https://github.com/jnobyrne/edgeofpy

[12] C. J. Stam and E. C. W. van Straaten, "Go with the flow: Use of a directed phase lag index (dPLI) to characterize patterns of phase relations in a large-scale model of brain dynamics," *NeuroImage*, vol. 62, no. 3, pp. 1415–1428, Sep. 2012, doi: 10.1016/j.neuroimage.2012.05.050.

[13] M. Vinck, R. Oostenveld, M. van Wingerden, F. Battaglia, and C. M. A. Pennartz, "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias," *Neuroimage*, vol. 55, no. 4, pp. 1548–1565, Apr. 2011, doi: 10.1016/j.neuroimage.2011.01.055.

[14] "AutoML mljar-supervised." Accessed: Sep. 01, 2023. [Online]. Available: https://supervised.mljar.com/

[15] "scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation." Accessed: Sep. 01, 2023. [Online]. Available: https://scikit-learn.org/stable/

[16] B. J. Ruijter *et al.*, "Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study," *Annals of Neurology*, vol. 86, no. 2, pp. 203–214, Aug. 2019, doi: 10.1002/ana.25518.

Address for correspondence:
Stefanie Blain-Moraes.
1650 Cedar Ave, Montreal, Quebec H3G 1A4.
stefanie.blain-moraes@mcgill.ca