HyperEnsemble Learning from Multimodal Biosignals to Robustly Predict Functional Outcome after Cardiac Arrest

Morteza Zabihi¹, Alireza Chaman Zar¹,², Pulkit Grover², Eric S. Rosenthal¹

¹Massachusetts General Hospital, Boston, MA, USA
²Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Aligned with the George B. Moody PhysioNet Challenge 2023, we aim to create an open-source algorithm to predict coma recovery 3-6 months following return of spontaneous circulation after cardiac arrest. Our algorithm employs multimodal biosignals from a dataset spanning five U.S. and European hospitals. The preprocessing pipeline enhances signal quality through bandpass and notch filtering, resampling, and z-score normalization. EEG channels are organized into 21 bipolar montages and standardized ECG lead selection. We extract 362 EEG expert-based features, encompassing temporal, spectral, time-frequency, and connectivity metrics. ECG features include 16 heart rate variability indices, features indicating atrial tachyarrhythmia, and shockable rhythms. Our classification approach involves initial positional encoding, followed by an ensemble of CatBoost combined with stacked ensembles with diverse base learners in the first layer. To improve robustness and reduce bias, our ensemble strategy employs various cross-validation committees, including confounder-isolating cross-validation. Our proposed predictive framework demonstrates an average challenge score mechanism of $0.651 \pm 0.077$ through 5-fold cross-validation. Our "AftRhythm" scored 0.792 on the hidden test set and officially ranked 1st in the PhysioNet Challenge 2023.

1. Introduction

Predicting functional outcomes of comatose cardiac arrest patients demands timely, precise, and robust approaches that inform clinical decision-making [1]. Traditional prognostic approaches, like visual inspection of electroencephalography (EEG) by neurophysiologists, are limited by subjectivity, risking bias and lack of reproducibility [2]. Machine learning (ML) for EEG analysis [3], [4] is increasingly utilized, and the opportunity exists for simultaneously using ECG cardiac telemetry and EEG neurotelemetry to predict functional outcomes [5].

We sought to utilize longitudinal EEG and ECG recordings available in cardiac arrest patients during the initial 72 hours following return of spontaneous circulation (ROSC) to predict a Cerebral Performance Category (CPC) scale dichotomous outcome: "good" (CPC 1-2) or "poor" (CPC 3-5). We introduce an ML framework (Figure 1) for predicting functional outcomes that integrates multimodal features, introducing some novel handcrafted EEG features, positional encoding via gradient boosting trees, confound-isolating cross-validation (CV) strategies, and a combination of stacking and boosting ensembles (HyperEnsemble).

2. Materials and methods

The training dataset included 607 adult cardiac arrest patients in coma, monitored continuously for hours to days with EEG and ECG, as previously described [6]–[8].

2.1. Preprocessing

EEG and ECG signals were bandpassed (0.1-45 Hz), resampled at 125 or 128 Hz, z-score normalized, and segmented into non-overlapping 3-minute segments. Given ECG lead inconsistencies, we integrated the first lead into our framework. We created a range of effective EEG channel montages through empirical validation: F3-C3, C3-P3, F4-C4, C4-P4, Fz-Cz, P3-O1, T5-O1, P4-O2, T6-O2, Fp1-Fp2, T3-T4, T5-T6, Fp1-T3, Fp2-T4, T4-O1, T3-O2, Fp1-Cz, Fp2-Cz, O1-O2, P3-P4, and Cz-Pz.

2.2. EEG quantitative measures

Single-lead measures: We extracted 43 features from 6 EEG channels, C3-P3, C4-P4, Fp1-Fp2, Fz-Cz, T3-T4, and Cz-Pz, including various temporal (Hjorth Activity, Mobility, and Complexity) and spectral (power below 1Hz [9], power ratios of 4-12/12-30Hz, 4-12/8-35Hz, and the alpha-to-delta power ratio features. Mean and variance were also computed for signal components across varying scales through wavelet decomposition, employing Daubechies 4 wavelets with level 5. We also extracted the interquartile range of the first difference of zero-crossing...
Figure 1. In the proposed predictive framework, preprocessing (1) and feature extraction (2) from 3-minute windows were followed by aggregation and sinusoidal positional encoding (3), wherein quantiles of features were computed at every h-hour intervals and aggregated windows were assigned a sinusoidal pattern. Training set construction (4) involved conventional and confound-isolating (CI) cross-validation. Finally, HyperEnsemble learning (5) leveraged diverse classifiers from data grouped using varying quantiles and h values to generate an averaged final prediction probability.

2.3. ECG quantitative measures

Heart Rate Variability (HRV) measures: The first 10
features are HRV measures derived from R-R intervals. Two features gauge the occurrence of abrupt R-R interval changes exceeding a 50-millisecond threshold, and the percentage of these among all R-R intervals. Subsequently, we computed SD1, SD2, the ratio SD1/SD2 and SD2/SD1, capturing both short- and long-term HRV [13]. Further, we extracted the cardiac sympathetic index by relating the dimensions of an ellipse defined by SD1 and SD2 and the Cardiac Vagal Index by logarithmically transforming the product of SD1 and SD2. A modified variation was utilized by computing the square of SD1 divided by SD2. Two additional features included the squared differences between R-R intervals and the ratio of the standard deviation to the mean of R-R differences [14].

**Atrial tachyarrhythmia measures**: We also extracted three features for identifying atrial tachyarrhythmia: AF Evidence, ATEvidence, and OrgIndex [15].

**Shockable measures**: We extracted three features inspired to identifying shockable rhythms [16]: 1) the difference between the maximum and minimum values of the band passed (6.5-30 Hz) ECG segment; 2) the proportion of time for which the first difference squared ECG segment remains below a predefined threshold; and 3) a measure in which the cumulative sum of the normalized power spectrum was first calculated, after which bandwidth was computed by discerning frequencies at which a given proportion of power is enclosed.

### 2.4. Static covariates

Four static variables were used as predictors and carried forward across rolling windows, including age, ROSC, out-of-hospital cardiac arrest (OHCA), and shockable rhythm. Sex and hospitals were considered confounders.

### 2.5. Classification framework

**Aggregating and positional encoding**: For every h-hour interval \((h=5, 6, 12)\), the 3-minute sequential segments were combined using quantiles (0.88 or 0.89) derived from individual feature values. Following the aggregation, sinusoidal positional encoding [17] was employed and added to feature values to imbue the feature vectors with the relative temporal order.

**Constructing training subsets**: To build diverse training sets for our proposed ensemble learning approach, we utilized three distinct CV strategies: 1) Five-fold CV generating 5 replicated datasets from the initial training data formed by withholding a random 20% subset to produce 5 intersecting training subsets; 2) Confounder-isolating [18] CV (cluster analysis) utilizing k-means clustering to group training data based on potential confounding variables, namely hospital type and sex, which led to three overlapping training subsets from the original dataset, where each iteration removed a distinct cluster; and 3) Confounder-isolating CV (propensity score), a variant employing logistic regression with predictors like hospital types and sex to predict outcomes then stratifying the population into four non-overlapping quantiles of the predicted probabilities and subsequently generating four intersecting training subsets by excluding a distinct group each time. By constructing these training subsets, we aimed to enhance robustness and reduce bias.

**HyperEnsemble learning**: Random undersampling was used to mitigate class imbalance within each subset, with each resulting subset inputted into three ensemble classifiers, whose probabilities were averaged to calculate final prediction probabilities. The first ensemble classifier employed was CatBoost, an ML method based on gradient boosting over decision trees [19]. The second ensemble was a stacking ensemble, wherein the base learners were a multilayer perceptron (MLP) comprising three hidden layers of 20 neurons, a support vector machine with a linear kernel, a CatBoost, extremely randomized trees, and a linear discriminant analysis classifier. The meta-learner was an MLP with two hidden layers, each with five neurons. The third classifier was another stacking ensemble with three CatBoost classifiers as base learners and logistic regression as the meta-learner. Apart from the base learners themselves, a pivotal distinction between this stacking ensemble and the previous ensemble resides in the nature of input data being generated by employing distinct aggregating time resolutions, i.e., \(h=5, 6, 12\) hours.

### 3. Results

Performance of the proposed framework for two variations of CV is shown in Table 1. CV1 tested patientwise 5-fold CV. CV_{tr}, tested CV on data isolated from one hospital and 20% of the remaining patients.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>CV1</th>
<th>CV_{tr}</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Challenge Score</td>
<td>0.691±0.083</td>
<td>0.651±0.077</td>
<td>0.792</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.80±0.07</td>
<td>0.78±0.06</td>
<td>NR</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.80±0.08</td>
<td>0.79±0.05</td>
<td>NR</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.89±0.02</td>
<td>0.87±0.02</td>
<td>0.916</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.93±0.01</td>
<td>0.91±0.02</td>
<td>0.957</td>
</tr>
<tr>
<td>F1</td>
<td>0.80±0.04</td>
<td>0.79±0.03</td>
<td>0.791</td>
</tr>
</tbody>
</table>

**AUPRC**, area under the precision-recall curve; **AUROC**, area under the receiver operating characteristic curve. **NR**, not reported by the PhysioNet Challenge 2023.

### 4. Discussion and conclusion

We developed an open-source framework for predicting the outcomes of comatose cardiac arrest patients. Our approach included established and novel EEG and ECG
features with positional encoding using non-deep learning models. By adapting ensemble learning and implementing diverse CV committees, we increased the robustness of our framework. The proposed solution focuses on a restricted set of EEG channels and features due to the computational constraints set by the PhysioNet challenge; a broader range of EEG channels and additional features may yield improved performance. Future directions include evaluating the response of clinicians and surrogates to simulated predictions and silent real-time evaluation.

References


Address for correspondence:
Morteza Zabihi.
Massachusetts General Hospital, Boston, MA
mzabihi@mgh.harvard.edu