Computationally Efficient Early Prognosis of the Outcome of Comatose Cardiac Arrest Survivors Using Slow-Wave Activity Features in EEG

Miikka Salminen[1], Juha Partala[1,2], Eero Väyrynen[1], Jukka Kortelainen[1,3]

[1]Cerenion Oy, Oulu, Finland
[2]Oulu University Secure Programming Group, Biomimetics and Intelligent Systems Group, University of Oulu, Oulu, Finland
[3]Physiological Signal Analysis Team, Center for Machine Vision and Signal Analysis, MRC Oulu, University of Oulu, Oulu, Finland

Abstract

This study, part of 'Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023', evaluated a computationally efficient method in predicting cardiac arrest (CA) survivors' prognoses using electroencephalography (EEG) recordings of a dataset provided for participants. Authors' team Cerenion developed a random forest based machine learning algorithm. A feature set of channel-by-channel root mean square power of a well-described neurophysiological EEG phenomenon called slow-wave activity (SWA), with time elapsed since CA, was used.

Five-fold cross-validation, using 80 % of the provided training set of EEG recordings from 607 out of 1020 patients, was used for evaluation. The held-out 20 % of data were used for testing and evaluating a final model trained on the full 80 % of the training data.

Cross-validated results, evaluated at 72 hours after CA, for predicting the outcome were: AUROC 70 %, AUPRC 78 %, accuracy 68 %, F-measure 64 %. Evaluating the challenge metric on the training data at times of 12, 24, 48, and 72 hours after CA provided scores of: 0.32, 0.40, 0.64, and 0.58, respectively. The hidden validation and test sets were not used, earning no rank. The results show promise in using SWA power features in predicting the outcomes of comatose CA patients.

1. Introduction

George B. Moody PhysioNet Challenge 2023[1, 2] provided a dataset of EEG recordings and associated metadata of comatose cardiac arrest (CA) survivors[3] and a framework for teams tasked to compete in developing an open-source solution for predicting such patients' outcomes. The possible outcomes are defined as either good or poor, but can further be broken down into finer Cerebral Performance Category (CPC) scores that codify the patient's state on a five point range from good neurological function (1) to deceased (5), with conditions such as moderate (2) or severe neurological disability (3) and unresponsive wakefulness syndrome (UWS) (4) in between. CPC values 1 and 2 are considered good outcome, and the remaining 3, 4, and 5 poor outcome.

As a business operating in this exact field (see the disclosure in Acknowledgments) the authors' team Cerenion set out to contribute a simplified version of its commercial method for the benefit of the field of research: namely, the use of electroencephalogram's (EEG) slow-wave activity (SWA) features as a part of a machine learning (ML) algorithm for assessing the patients' outcomes in a computationally efficient way and, to reduce the costs of healthcare, as early as possible during an inpatient's stay at the intensive care unit (ICU).

SWA, i.e. frequencies under 1 Hz, refers to a well-described neurophysiological phenomenon in EEG. Kortelainen et al. have shown SWA's predictive performance on forming prognoses for comatose CA survivors in the ICU[4] – especially during the first 12 hours since CA[5].

A subset of electrodes on a hypoxic ischemic encephalopathy patient's forehead provide similar results on the slow-wave EEG as a full cap[6]. The solution in this study builds upon that finding, emphasizing computational efficiency and robustness in situations where obtaining a full set of data and metadata would be impractical.

2. Method

The software implemented in this study was built on the example Python code provided by the Challenge organizers[2]. This was necessitated by having to submit the challenge entry for scoring with the hidden validation

ISSN: 2325-887X DOI: 10.22489/CinC.2023.096

and test datasets. In addition to Python's runtime and standard library, NumPy, SciPy, and scikit-learn were used for implementing the algorithm.

2.1. Data

The I-CARE v2.0 dataset[3] by International Cardiac Arrest REsearch consortium (I-CARE) comprising EEG recordings of 607 out of a total of 1020 patients available for training was used. Specifically, data from all patients in the available training set were included, i.e. no patients were left out for any reason such as too noisy data. However, some individual recordings from some patients were left out due to being too short for calculating the features. The hidden validation and test sets comprising rest of the full 1020 patient dataset were not used for evaluation.

The dataset includes multiple EEG recordings for each patient in the dataset, recorded at different times after the onset of a CA; the recordings include the information of when exactly after the CA they were recorded. Importantly, the metadata also includes the outcomes (good or poor) and the CPC values (1, 2, 3, 4, or 5) 3–6 months after the Return of Spontaneous Circulation (ROSC), which are used as the ground truth labels, or classes, in the method of this study. Table 1 summarizes the statistics on outcomes in the provided dataset. The metadata also has hospital-specific identifiers to distinguish where the recordings were made.

Outcome	Patients #	Patients %
Good	225	37 %
Poor	382	63 %
Total	607	100 %

Table 1. Summary of patients in good and poor outcome classes in the given dataset.

The sampling frequencies of the EEG recordings vary record-by-record from 200 Hz to 2048 Hz. The EEG channels available in all recordings are: C3, C4, Cz, F3, F4, F7, F8, Fz, Fp1, Fp2, O1, O2, P3, P4, Pz, T3, T4, T5, and T6, i.e. a subset of the channels in the International 10–20 system. ECG data and further metadata with descriptions are also available in the dataset, although not utilized in this study.

No measures were taken to account for the slightly imbalanced classes within the data, nor to control other aspects such as patient's age or sex. The data were split into a training and a held-out test set using a random 80/20 split with a constant random seed for reproducibility. The training data was further split into five different 80/20 training and validation sets in a 5-fold cross-validation scheme.

2.2. Feature Set

The only metadata feature selected for the feature set is Time Since Cardiac Arrest (TSCA), which was available for all patients. TSCA is the number of minutes that has elapsed since the patient suffered a CA. Other metadata were intentionally left out to increase the robustness of the solution in cases when including such metadata is not possible.

Signal power features of Slow-Wave Activity (SWA) were chosen for training.

Before preprocessing, the signal data was first organized for consistency across patients and recordings and to accommodate the study's preferences of ease-of-use and computational performance:

1. ECG and other non-EEG channels were removed.
2. A subset of EEG channels was selected on electrodes labeled F7, F8, Fz, Fp1, Fp2, T5 and T6, allowing for easy access single-use electrodes on the patients' heads' hairless areas, and further reducing the computational requirements.
3. The used EEG channels were organized into an exact, same order, for consistency across feature vectors.

The selected data was preprocessed as follows:

1. Constant detrending, i.e. subtraction of arithmetic mean of individual EEG signal channels.
2. Low-pass filtering of the detrended signal with a 12th order digital infinite impulse response (IIR) Butterworth filter with a cutoff frequency of 1.0 Hz, implemented with a cascade of second-order sections for numerical stability. The generated filter was evaluated as being sufficiently stable by confirming that the absolute values of its poles are smaller or equal to one, and the poles themselves are nonzero.
3. A length of one second long segment was discarded from the beginning of each signal channel to account for any filter border effect artifacts.

The features from the EEG signal were computed as follows:

1. The whole length of the preprocessed signal was split into non-overlapping segments of 35 seconds each.
2. Channel-wise root mean squared (RMS) powers were calculated in accordance with equation (1) from the segments' samples (x_i) for the whole length (n) of the segment.

$$x_{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {x_i}^2} \tag{1}$$

The final feature vectors were built for all of the 35-second segments with the first feature being TSCA and the rest being channel-wise RMS powers.

A patient's full set of features comprises all feature vectors generated from all the 35-second segments. E.g. if a patient had a total of 1 hour of EEG recorded, 102 feature vectors were generated.

2.3. Labeling

Ground truth labels were extracted for each patient.

For the outcome model, the label (class) for each feature vector was set to be the corresponding patient's outcome, i.e. either good or poor.

For the CPC model, each feature vector's label was set to be the CPC number of the patient.

2.4. Models

Supervised learning was performed using random forests. Random forest was chosen due to its good performance – both in terms of classification and computational cost. No hyperparameters or external objective functions were optimized for, and the default of Gini impurity was used as the random forest's criterion. A default number of 100 estimators was used, and a constant random seed was set to enforce reproducibility.

In cross-validation folds, the training data was the 80 % data selected for that fold out of the initial 80 % split of the available 607 patients. In the final evaluation phase, the initial full 80 % split of the 607 patients, i.e. 485 patients, were used for training. Full length, i.e. all recordings of patient, were always used during training.

In each cross-validation fold, and separately for the final evaluation, two models were trained using the available training data. A binary classifier was trained to predict either of the two outcomes based on the given feature vectors. A regression model was trained to predict CPC based on the given feature vectors.

The trained models were used to give individual predictions for all features of a patient. This was done separately for each patient in the current fold's validation set (during cross-validation) or held-out test set (using the final model). The threshold for making either prediction for the outcome was set at 50 % probability.

Finally, using the full set of predictions on all feature vectors of a patient, the patient's predicted outcome and CPC were selected by popular vote. Proportions of the most popular outcomes amongst the total predictions for a patient were selected as the probabilities for that outcome.

2.5. Evaluation

The predicted outcomes and CPCs for the validation or test set's patients were compared against the corresponding ground truth labels of the patients. These statistical metrics were calculated to evaluate the correctness of the outcome predictions: Area Under Receiver Operating Characteristics (AUROC), Area Under Precision Recall Curve (AUPRC), accuracy, and F-measure. Challenge score – the true positive rate at a false positive rate of 0.05 at each hospital – was included. Mean squared error (MSE) and mean absolute error (MAE) were correspondingly calculated for evaluating the CPC model.

For the cross-validated results, the metrics were averaged over the cross-validation folds, and a standard deviation was calculated to measure the variance. Furthermore, the challenge's evaluation times of 12, 24, 48, and 72 hours after CA were used in the cross-validated outcome results' evaluation: Given any of the four listed timestamps, only the predictions made on feature vectors up to that timestamp were selected. The selected predictions were then tested as described in chapter 2.4. and evaluated with the above metrics. If a patient didn't yet have predictions until the given timestamp, the more prevalent poor outcome, with CPC of 5, at the probability of the proportion of poor outcome out of the total number of patients, was used instead.

3. Results

Table 2 summarizes the cross-validated results on outcome prediction. 3 out of 5 folds produced a Not a Number (NaN) result for challenge score for all timestamps, implying that it wasn't possible to calculate. Thus, NumPy's nanmean and nanstd functions were used.

	12 h	24 h	48 h	72 h
Challenge score	.32±.15	.40±.26	.64±.22	.58±.23
AUROC	.58±.06	.70±.03	.70±.03	.70±.02
AUPRC	.69±.05	.76±.05	.77±.03	.78±.02
Accuracy	.64±.03	.70±.05	.68±.03	.68±.04
F-measure	.52±.04	.65±.06	.62±.03	.64±.04

Table 2. The cross-validated results, using only the training data, evaluated at 12 h, 24 h, 48 h, and 72 h after CA. Values are averages over 5-fold cross-validation folds with standard deviations. Leading zeros before points omitted for legibility.

The evaluation metrics for the final outcome model trained on the specified subset of training set and tested against the full lengths of EEG of the held-out test set were challenge score 0.56, AUROC 83 %, AUPRC 86 %, accuracy 75 %, F-measure 71 %, and for the final CPC model MSE 2.65 and MAE 1.37.

The models were not tested nor evaluated using the hidden validation and test sets, earning no rank in the PhysioNet Challenge 2023.

4. Discussion and Conclusions

The approach in this study shows encouraging results in assessing a comatose CA patient's probable outcome in the I-CARE v2.0 dataset. The metrics calculated for the evaluation of both the cross-validated and final model's results show that SWA power features provide insight on a CA patient's recovery early on after the onset of the CA, that the authors believe can be used by healthcare personnel in the ICU. Overall, the study succeeded in further demonstrating the usefulness of using feature vectors that are based on SWA.

The approach in this study takes the deliberate tradeoff of preferring simplicity and computational efficiency over maximum predictive performance. The RMS power features of EEG's SWA are computationally very simple and can be used for real-time analysis even on low-power devices. Nevertheless, several steps could be taken to improve the predictive performance, including taking possible biases in the dataset better into account, even within the constraints of the selected simple features.

The model could be better optimized, especially considering the objective of the challenge score. Hyperparameters such as the cutoff frequency of the IIR filter or the length of each segment used for feature vector could be tuned. There's a trade-off between the length of a segment and how many independent segments are available for forming the conclusion, and such trade-off is bound to have an optimal point somewhere. Conversely, nonindependent segments could be selected by allowing them to overlap, which would also increase the computational costs, possibly turning the approach not suitable for real-time monitoring on a low-performance device.

Lots of predictions – some of which disagree with each other – are made for each patient, one for each 35 second feature vector for hours of EEG data. Different algorithms could be experimented on for determining the final result, i.e. on how to make the final prediction from thousands of predictions. Moreover, the decision based on each prediction's probability could be done with a different threshold to favor one outcome over the other.

Additional features or mechanisms for handling noise, such as artifacts caused by sweating, could be introduced for better prognoses at the cost of higher computational requirements. As more metadata is entered over time, separately trained models with better performance could be taken into use for further calculations on the patient's prognosis and the historical EEG retrospectively reanalyzed.

Only the very simple RMS power features were calculated, and thus downsampling would have increased the computational cost. For more complex SWA features downsampling may make sense from computational cost point-of-view, since the EEG signals of interest are below 1 Hz in frequency, and may have a lot of redundancy on high sampling frequencies.

Conversely, to reduce the computational costs further, removal of the TSCA feature could be experimented on as well. This would allow using the method solely on the basis of EEG – with zero requirements for healthcare personnel on inputting any patient metadata.

This study used TSCA as a substitute for ROSC, which is typically favored in related studies for similar purposes. As the patient may be comatose for days, the EEG segments can have TSCA or ROSC values ranging from 0 to thousands of minutes. Further study would be needed to tell if there is a difference in predictive performance between TSCA and ROSC.

Acknowledgments

Cerenion Oy[1] assumed all costs of this study. Disclosure of a possible conflict of interest: Cerenion Oy's C-Trend® software is a CE-marked medical device intended to be used for the EEG-based evaluation of brain function of e.g. comatose CA survivors (at the time of this publication, not FDA 510(k)-cleared and not available for sale in the United States). Cerenion Oy holds several patents related to the technology and reserves all patent rights.

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 2023 (in press); doi:10.1097/CCM.0000000000006074.

[4] Kortelainen J, Väyrynen E, Huuskonen U, Laurila J, Koskenkari J, Backman JT, et al. Pilot Study of Propofol-induced Slow Waves as a Pharmacologic Test for Brain Dysfunction after Brain Injury. Anesthesiology 2017;126(1):94–103.

[5] Kortelainen J, Ala-Kokko T, Tiainen M, Strbian D, Rantanen K, Jouko L, et al. Early recovery of frontal EEG slow wave activity during propofol sedation predicts outcome after cardiac arrest. Resuscitation 2021;165:170–6.

[6] Kortelainen J, Väyrynen E, Juuso I, Laurila J, Koskenkari J, Ala-Kokko T. Forehead electrodes sufficiently detect propofol-induced slow waves for the assessment of brain function after cardiac arrest. Journal of Clinical Monitoring and Computing. 2019;34(1):105–10.

Address for correspondence:

Miikka Salminen c/o Cerenion Oy
Kyllikinportti 2, 00240 Helsinki, Finland
miikka.salminen@cerenion.com

[1] Cerenion Ltd., www.cerenion.com, Finnish business ID 2832753-1.