Model Ensembling for Predicting Neurological Recovery after Cardiac Arrest: Top-down or Bottom-up?

Hongliu Yang[1], Ronald Tetzlaff[1]

[1]Technische Universität Dresden, Dresden, Germany

Abstract

Early electroencephalography (EEG) contains valuable information for predicting neurological recovery in comatose patients after cardiac arrest. As part of the George B. Moody PhysioNet Challenge 2023, our team, TUD_EEG, developed a novel ensembling approach that combines two pipelines with different directions of information transfer between patient-level and segment-level descriptions. Using both EEG and patient clinical information, our model achieved a Challenge score of 0.72 (3rd place out of 34 eligible teams) on the hidden test set.

1. Introduction

The 2023 George B. Moody PhysioNet Challenge [1, 2] invited teams to develop machine learning algorithms for predicting patient outcome after cardiac arrest using longitudinal electroencephalogram (EEG) and other recordings. The algorithm development was based on the ICARE database [3].

EEG recordings monitor electrical activity in the brain, and are commonly used to analyze neurological diseases. In studies to predict the neurological outcome of comatose patients after cardiac arrest, characteristic patterns of EEG have been observed in relation to specific levels of recovery, e.g. suppressed background and burst suppression are highly malignant patterns indicating poor outcome [4, 5]. Note that the differentiation of EEG patterns was done locally at the level of EEG segments. However, the ultimate goal is to predict recovery outcome at the patient level. From a machine learning point of view, the classification task is for weakly labelled data. So far, there has been no careful study of how to optimize the information transfer between the two levels of description in order to obtain accurate and robust prediction results. This is the main issue to be addressed in our model.

Moreover, the longitudinal time evolution of EEG patterns has been found to be important for the post-arrest neurological prognosis [6]. The full potential remains to be explored.

2. Methods

2.1. Dataset

The ICARE database [3] contains 32,712 hours of continuous recordings for 1020 patients from seven hospitals, divided into training, valid and test sets of 607, 107 and 306 patients, respectively. We used only the EEG recordings to predict patient recovery, discarding the electrocardiogram (ECG) and other recordings.

2.2. Preprocessing

The provided EEG recordings with a monopolar montage were first mapped to the ”double banana” bipolar montage. The recordings were then processed with a notch filter to remove power line noise, refined with a bandpass filter with cutoff frequencies $[0.1, 30]$ Hz to obtain the relevant information, resampled to 100 Hz for ease of processing in the following steps, and finally divided into segments of 5 minutes length.

2.3. Feature extraction

The features considered in this study can be roughly divided into four groups.

P features: For each EEG segment, mean power was calculated for 5 frequency bands, $[1-4Hz]$, $[4-8Hz]$, $[8-12Hz]$, $[12-30Hz]$, $[30-100Hz]$, to monitor the change in the frequency domain.

S features: Signal statistics, including mean, standard deviation, skewness and kurtosis, were used to characterize the change in the time domain.

E features: To describe the profile change due to epileptic-form events, e.g. spikes, signal entropy and participation ratio [7] were calculated.

C features: Besides these features characterizing the properties of individual channels, Pearson correlation coefficients and mean phase coherence [7] were used to describe the linear and nonlinear correlation between channels.

ISSN: 2325-887X DOI: 10.22489/CinC.2023.015

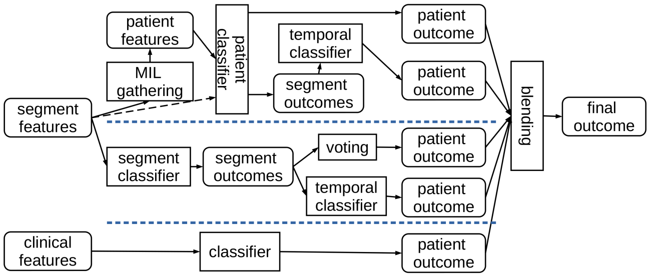

Figure 1. Overview of the model proposed in this study. Three pipelines, separated by thick dashed lines, are the top-down, bottom-up and clinical approaches (from top to bottom).

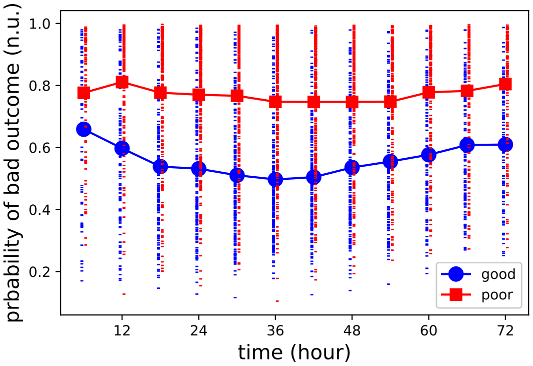

Figure 2. Temporal evolution of segment-level prediction outcomes of the top-down approach. Means for two groups of patients with poor and good outcomes are also shown. Note that, even at the segment level, the predicted probability of a poor outcome is (mostly) higher for patients with a poor outcome than for patients with a good outcome.

2.4. Models

Our model consists of three pipelines, which are described in detail below. An overview is given in Figure 1.

2.4.1. Top-down approach

As discussed, the recordings are only labelled at the patient level. Therefore, it is natural to first construct a patient-level global feature set by gathering local features of all segments belonging to each individual patient. The standard classifier can then be used to predict the recovery outcome by binary classification of the patient-level global features. The proposal is in the realm of multiple instance learning [8], and the gathering method is essential to its success. There are many options for the MIL gathering, e.g. extracting the mean, standard deviation, or maximum of instances in a bag. We adopted the so-called earth-mover's-distance (EMD) [9] to measure the difference between bags of segment-level features of a pair of patients. Here, each bag of feature instances is viewed as a distribution, and the EMD metric quantifies the minimum cost of transforming one distribution into the other. The use of the EMD metric has been shown to achieve leading performance on common MIL tasks (see Figure 15 in [8]). As noted, patients in the ICARE database have recordings of varying lengths. The metric can handle such inhomogeneous bags without data imputation, which can introduce artifacts.

For patient-level classification, we used a support vector machine (SVM). A Gaussian kernel was applied to convert the EMD distance into a similarity measure:

$$S = exp(-\gamma EMD). \tag{1}$$

where the parameter γ is to be estimated via cross-validation.

To account for the temporal evolution of EEG patterns [6], we applied the trained patient-level classifier to features of individual segments to obtain local predictions at the time of each EEG segment. Using these segment-level predictions as input features, a temporal classifier, e.g. logistic regression, was trained to provide a further patient-level prediction of recovery outcome.

In this top-down approach, the basic classification is at the patient level, and all subsequent operations are based on it.

2.4.2. Bottom-up approach

The bottom-up approach is motivated by an observation on the temporal evolution of the segment-level predictions of the top-down approach, see Figure 2. Note that the segment-level predictions are mostly consistent with the patient-level labels, i.e. the predicted probability of a poor outcome is higher for patients with a poor outcome than for patients with a good outcome. This is also supported by the mean predictions for the two groups. It is therefore plausible to start directly from the segment-level classification.

Given the large number of samples, we used the light gradient boosting machine (LightGBM) [10] as our segment-level classifier. The patient-level labels were simply distributed to the corresponding segments. The obtained segment-level predictions were aggregated by soft voting to obtain a patient-level prediction of recovery outcome.

Similar to the top-down approach, a temporal classifier was trained with the collections of segment-level predictions of individual patients as input features. The output is another patient-level prediction incorporating information about the temporal evolution of the EEG.

Models	Parameters
SVM	C: regularization parameter γ: kernel parameter in Equation (1)
LGBM	num_leaves min_child_samples
model blending	weights of the average

Table 1. Hyperparametrs optimized with cross validation.

In contrast to the top-down approach, the basic classification for the bottom-up approach is at the segment level.

2.4.3. Clinical information approach

Clinical features were classified using a generalized linear model, Poisson regression. The set of patient clinical information provided was selected via a grid search to eliminate the less relevant items.

2.4.4. Model ensembling

All patient-level predictions from the above approaches were combined to produce a final prediction using their weighted average. For patients with missing data for a particular approach, the prediction was set to the mean value of the approach for the train set.

2.4.5. Model training

The hyper-parameters of the SVM and LightGBM classifiers and the linear meta-model of model blending were optimized with 3-fold cross validation (CV) using the Hyperopt package [11]. The parameters involved are listed in Table 1.

3. Results

The official challenge score for characterizing prediction performance is the true positive rate at a false positive rate of 0.05 for predicting poor outcome. Other metrics include the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPRC). A 3-fold cross-validation was used to obtain the scores on the public training set.

The performance comparison of different feature combinations was presented in Table 2. The best combination marked in bold was used for the rest of the study.

Continuous longitudinal EEG recordings, and/or full coverage of the 10-20 system, may not always be available/feasible. We also tested how the model performed with a reduced number of samples or channels. The scores are presented in Table 3.

The ICARE dataset consists of data from seven different hospitals in the USA and Europe. To show the inhomogeneity intrinsic to the data, we presented in Table 4 leave-one-out cross-validation scores.

Features	challenge score	AUROC	AUPRC
PS	0.60 ± 0.11	0.86 ± 0.03	0.91 ± 0.03
C	0.62 ± 0.08	0.85 ± 0.02	0.91 ± 0.02
PSC	$\mathbf{0.63 \pm 0.11}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.92 \pm 0.03}$
PSE	0.60 ± 0.11	0.86 ± 0.02	0.91 ± 0.02
PSEC	0.62 ± 0.10	0.87 ± 0.03	0.92 ± 0.03

Table 2. Prediction scores on the public training set for different feature combinations. The best feature combination used in the rest of the study is marked in bold. AUROC: area under the receiver operating characteristic curve. AUPRC: area under the precision-recall curve. For feature combinations, the letters P, S, E, C stand for the corresponding feature groups as described in Section 2.3.

Data	challenge scores	AUROC	AUPRC
F0.1	0.61 ± 0.10	0.87 ± 0.02	0.92 ± 0.02
C4	0.57 ± 0.09	0.86 ± 0.03	0.91 ± 0.03
C2	0.56 ± 0.08	0.84 ± 0.03	0.90 ± 0.03
C1	0.53 ± 0.09	0.83 ± 0.03	0.89 ± 0.03

Table 3. Prediction scores on the public training set for reduced data amount or fewer electrode contacts. $F0.1$: 10% of randomly selected 5m segments of each patient. C4: using only 4 channels, 'F7-T3', 'F8-T4', 'F3-C3', 'F4-C4'. C2: using 2 channels, 'F3-C3', 'F4-C4'. C1: using one channel, 'F3-C3'.

Finally, the results of the official ranking of our team are summarized in Table 5).

4. Discussion and Conclusions

In conclusion, to predict the neurological outcome of comatose patients after cardiac arrest, we proposed a model ensembling method that combines two pipelines with reversed direction of information flow between patient-level and segment-level descriptions. Experiments showed that our method can achieve state-of-the-art results for this purpose (see Table 5).

The success made possible by the bottom-up approach confirms that the information indicative of the patient's recovery outcome is distributed across the majority of EEG

Hospital	A	B	D	E	F
Challenge score	0.70	0.41	0.68	0.90	0.52
AUROC	0.92	0.70	0.86	0.92	0.80
AUPRC	0.94	0.86	0.93	0.98	0.88

Table 4. The leave-one-out cross-validation scores show the imhomogeneity of the data between hospitals.

Training	Validation	Test	Ranking
0.96	0.69	0.72	3/34

Table 5. The official Challenge scores of our team, **TUD_EEG**, and the ranking on the hidden test set. The 3-fold cross validation score was 0.63 ± 0.11 on the public training set.

segments. This is in contrast to typical multi-instance learning, where a bag's label is often determined by a few key instances. The weakly labelled problem can therefore be transformed into a standard classification where the patient labels are distributed over the corresponding EEG segments. This greatly increases the sample size, which reduces the chance of overfitting and may be responsible for the increase in performance.

Ablation experiments of different feature combinations showed that channel-wise (P, S, E) and cross-channel (C) features contain highly overlapping information for the prediction task. So far, only hand-crafted features have been used. The proposed ensembling method can be easily extended to include other features, e.g. those learned with a deep neural network, which could lead to further performance improvements.

Our experiments showed that the prediction performance is almost unchanged when the amount of data is reduced by 90% (see Table 3). Together with our discussion above that useful information for predicting patient outcome is distributed across the majority of EEG segments, this is a good indication that *continuous longitudinal* EEG recording may not be essential, i.e. periodically sampled short-term recordings may do the job.

Performance becomes worse as the number of channels used decreases. However, even with a single bipolar channel, the drop in performance is not severe (see Table 3). The results call for further studies on the plausibility of using fewer EEG channels, which could be of great benefit for ambulatory monitoring or in developing countries.

The large variance of the cross-validation scores in Tables 2, 3, and especially 4 and the result of other studies [12] show the existence of significant data inhomogeneity in the ICARE database [3] considered.

Acknowledgments

We thank the Center for Information Services and High Performance Computing at TU Dresden for the generous allocation of computing time.

The work was supported by the Innovation Projects MedTech ALERT of Else Kröner Fresenius Center for Digital Health of the TU Dresden and the University Hospital Carl Gustav Carus.

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, Kahndare 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] Westhall E, Rossetti AO, van Rootselaar AF, Kjaer TW, Horn J, Ullén S, et al. Standardized EEG interpretation accurately predicts prognosis after cardiac arrest. Neurology 2016;86(16):1482–1490.

[5] Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, van den Bergh WM, Foudraine NA, Kornips FHM, et al. Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study. Annals of Neurology 2019;86(2):203–214.

[6] Khazanova D, Douglas VC, Amorim E. A matter of timing: EEG monitoring for neurological prognostication after cardiac arrest in the era of targeted temperature management. Minerva Anestesiologica 2021;87(6):704–713.

[7] Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. Brain 2007; 130(2):314–333.

[8] Amores J. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence 2013; 201:81–105.

[9] Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 2000;40:99–121.

[10] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 2017;30:3146–3154.

[11] Bergstra J, Yamins D, Cox D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In International Conference on Machine Learning. PMLR, 2013; 115–123.

[12] Zheng WL, Amorim E, Jing J, Ge W, Hong S, Wu O, et al. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation 2021;169:86–94.

Address for correspondence:

Hongliu Yang
Technische Universität Dresden, Faculty of Electrical and Computer Engineering, 01062 Dresden, Germany
hongliu.yang@tu-dresden.de