Recovery from Coma after Cardiac Arrest: Which Time-Window Counts the Most for Deep Learning Predictions?

Filippo Uslenghi[1], Roberto Sassi[1], Massimo W Rivolta[1]

Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy

Abstract

The George B. Moody PhysioNet Challenge 2023 was dedicated to the development of automated methods for predicting neurological recovery from coma after cardiac arrest. Models were requested to predict a good vs poor neurological outcome using electroencephalograms (EEGs), electrocardiograms (ECGs) and clinical information. Here, we proposed a deep learning model based on a residual network architecture. The model was designed to process only one 5-minute window for each hour up to the 72nd from spontaneous resuscitation, and aggregated the output probabilities of poor outcome using different weighted averages. A 5-fold cross validation technique was used to set the hyperparameters of the model and evaluate the performance on the public dataset. The model's input involved EEG data, heart rate variability (HRV) features extracted from the available ECGs and clinical information. The weighted averages showed improvement over uniform weighting on the hidden validation set (score increased up to 20%), but no improvement in cross-validation. Also, the addition of HRV features and clinical information did not show significant improvement over using only EEG data. The Challenge scores on the public training set, hidden validation set and hidden test set were 0.887, 0.627, and 0.708, respectively (team name: unimi_bisp_squad, ranking: 5).

1. Introduction

Predicting the neurological recovery of comatose patients after cardiac arrest (CA) is a major clinical problem, with many patients never regaining consciousness and others with good outcome predicted as poor. Current guidelines recommend electroencephalogram (EEG) monitoring to identify features predictive of severe brain injury leading to poor outcomes, such as EEG background suppression, burst suppression and seizure. In addition, the combination of EEG monitoring with serum biomarkers and clinical information was found optimal, with the former having the major impact on the prediction [1]. However, continuous monitoring of EEG patterns requires extensive effort by the clinical staff and visual assessment of such patterns may compromise the prediction [2].

The George B. Moody PhysioNet Challenge 2023 [3, 4] was dedicated to the development of automated methods to predict neurological recovery from coma after CA. The International Cardiac Arrest REsearch consortium (I-CARE) assembled a database from seven hospitals in the United States and Europe with a large set of patients who underwent monitoring following CA [5]. The Challenge aimed to design state-of-the-art predictive models to process EEGs, electrocardiograms (ECGs) and clinical information to predict the neurological outcome.

Previous works showed that deep learning (DL) was an effective methodology to predict the neurological recovery using only 5 minute artifact-free EEGs [2]. Also, when the DL model processed the entire monitoring period (up to 3 days, using a recurrent neural network), the performance improved [6]. In these studies, the optimal timing for the prediction after the resuscitation was found different (12 h vs 66 h), keeping this as a matter of investigation.

Another possible approach to predict the neurological outcome is heart rate variability (HRV), which recently provided promising performance [7]. Results suggested HRV features, such as very low frequency power and entropy-based features, being useful for the prediction.

In this study, relying on previous results, we investigated the use of DL and HRV to predict the neurological outcome after CA. The approach was designed to process all available 5 minute recordings of a patient and to weight the predictions according to the time from resuscitation.

2. Methods

2.1. Dataset

The public dataset of the Challenge consisted of a collection of several biosignals from 607 patients who experienced CA and were treated in intensive care units. Among others, signals included 19-channel continuous EEG recordings and one- or two-lead electrocardiograms (ECGs). Additionally, for each patient, several clini-

cal variables were available; namely: age, sex, hospital where they were hospitalized, return of spontaneous circulation (ROSC) in minutes, out-of-hospital cardiac arrest (OHCA), shockable rhythm and targeted temperature management (TTM) in Celsius. Every patient had a different number of recordings depending on the total monitoring time. They could last from hours to days and were provided split into hourly segments.

2.2. Preprocessing

We simulated a bipolar EEG montage by subtracting the leads as follows: Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, Fz-Cz, Cz-Pz. From these signals, we selected one 5-minute window in a given hourly segment least affected by noise. To do this, we segmented each 1-hour EEG signal into consecutive 5-minute windows. We then selected the window with the highest number of good quality channels which occurred first in time. The assessment of EEG quality for each channel was performed using a decision tree based on two features, *i.e.*, the peak-to-peak amplitude (PP) and the average absolute difference (AAD) between consecutive EEG samples. The thresholds for the decision tree were estimated by computing the 2.5 and 97.5 percentiles of the distribution of PP and AAD values of all 18 channels of the EEG recordings of the first version of the dataset, since it contained only high quality EEGs. A good quality channel was then defined as having both PP and AAD within the ranges defined by the percentiles. Thresholds for PP were 0.66 μV and 679.55 μV and for AAD 0.3 μV and 15.96 μV.

The selected window was then filtered with a zero-phase band-pass FIR filter (0.1-15 Hz), designed using the window method with a Hamming window. Then, the window was resampled at 30 Hz (*i.e.*, 9000 samples) and signals were stored into a matrix with dimensions of 18×9000. In case the hourly segment was shorter than 5 minutes, we applied zero-padding to form a 5-minute signal.

In order to extract HRV features, ECG signals went through a preprocessing phase. First, to make ECG signals uniform in terms of sampling rates and to reduce the contribution of high frequency noise, baseline wandering and respiration, ECGs were filtered with a 3^{rd} order Butterworth band-pass filter (0.5 - 40 Hz) and then re-sampled at 128 Hz. Second, the ECG segment which was time-aligned to the EEG window was then retained. Third, ECG quality was assessed by computing the average Pearson's correlation coefficient between the average QRS and each individual complex extracted from the vector magnitude. When such correlation was < 0.8 then the ECG was considered as of a poor quality and the HRV features were then marked as missing values to be imputed. Otherwise, we extracted the inter-beat time interval series (RR) using a custom Pan-Tompkins algorithm and extracted the following HRV features: i) average RR interval (ARR) in s; ii) SDNN in ms; iii) RMSSD in ms; iv) sample entropy (SampEn, $m = 2$, $r = 0.2 \times \text{STD}$, where the estimate of the standard deviation was obtained as $1.4826\times$ median absolute deviation of the RR series); and v) deceleration capacity (DC) in ms ($L = T = s = 5$). For the latter, the approach reported in [8] was used; here, the autocovariance function was directly estimated from the RR series.

2.3. Models

2.4. EEG model

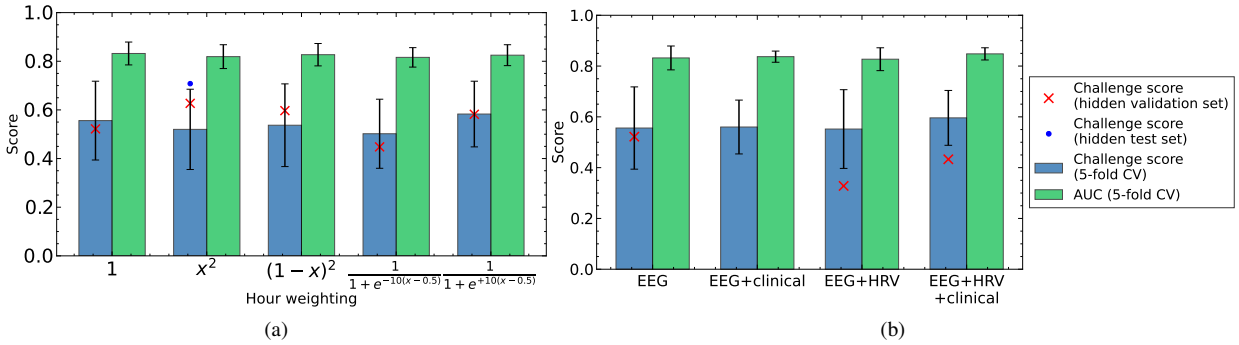
Several DL components were used: i) 1D convolutional layer (ConvL); ii) dropout (Dp); iii) batch normalization (BN); iv) fully connected layer (FCL); v) ReLU activation function (ReLU); vi) max pooling (MP) with size and stride of 4; and flatten layer (FL). The EEG model consisted in an initial ConvL with F filters followed by R residual blocks [9]. Each block was designed as 2 consecutive ConvLs. The skip connection of each block had a ConvL with filter size of 1. The ConvLs of the first block had F filters while the ConvLs of second block had $2F$. Next, 2 FCL and a single neuron as output were added. Each ConvL was followed by BN and Dp. The Dp of first ConvL of each block was followed by ReLU. The input of each block was preceded by ReLU and MP. The first FCL was preceded by ReLU, MP and a FL. Each FCL was followd by BN, Dp and ReLU.

The model took as input a single preprocessed EEG matrix of a patient and outputted the probability of poor outcome for each available signal. A weighted aggregation of the probabilities over 72 hours was computed as the final probability of having a poor outcome for that patient.

In order to build the EEG model, we performed a grid-search on the kernel-size K and the number of filters F of the ConvLs, as well as the number of residual blocks R and neurons of the FCL, by minimizing the 5-fold cross-validation (CV) error (see sec. 2.8 for details about performance evaluation). Once the architecture was defined, we performed a grid/random-search on i) the regularization values of λ_1 and λ_2 of the L_1 and L_2 norms; ii) positive class weighting; iii) the probability of the dropout layers; iv) the value of the learning rate; v) batch size; and vi) number of training epochs. We also tested different gradient descent algorithms such as stochastic gradient descent (SGD), Adam and RMSprop. Here, SGD showed to be more stable and consistent, and it was then retained.

2.5. ECG model

The ECG model was defined as a logistic regression model. The inputs were i) HRV features binarized; and ii)



(a) (b)

Figure 1: Average and standard deviation across the 5 folds of Challenge score and AUC for each aggregation formula (a) and the four models tested (b).

the outcome probability estimated by the EEG. The output was the probability of poor outcome. The binarization was performed using thresholds estimated by means of a Receiver Operating Characteristic (ROC) curve. The optimal threshold was the one associated to the closest top-left corner of the ROC curve. Features were set to 1 if $\mathrm{ARR} \geq 0.74$ s, $\mathrm{SDNN} \geq 41.60$ ms, $\mathrm{RMSSD} \geq 54.16$ ms, $\mathrm{SampEn} \geq 1.06$ and $\mathrm{DC} \geq 9.26$ ms, 0 otherwise.

The logistic regression model was trained using the same learning rate, optimizer, momentum, loss function, positive class weighting of the EEG model and it was trained for 100 epochs (found with a coarse grid-search).

In case the ECG recording was not available at the same time of the EEG, or the ECG was not of sufficient quality, the HRV features were imputed (see sec. 2.7).

2.6. Clinical model

Similarly to the ECG model, we designed a logistic regression model with the same output while inputs were i) clinical features binarized; and ii) outcome probability estimated by the EEG model. From the set of available clinical data we considered age, sex, ROSC, OHCA, schockable rhythm and TTM. Age and ROSC were binarized to 1 if $\mathrm{age} \geq 65$ (arbitrary threshold) and $\mathrm{ROSC} \geq 30.0$ minutes (75^{th} percentile across patients), 0 otherwise. TTM was set to 1 if the patient was treated with temperature management, 0 otherwise. The logistic regression model was trained as the ECG model, but with 50 epochs.

Training	Validation	Test	Ranking
0.887	0.627	0.708	5

Table 1: Challenge score for our final selected entry (team unimi_bisp_squad) for the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

2.7. Data imputation

Different approaches were carried out for imputing missing values. Regarding the ECG recordings, we imputed the missing values using the median of every feature computed over the good quality ECG signals which were time-aligned with the corresponding EEG signals. For the clinical variables, we used the median value for age and ROSC variables. For sex, OHCA and shockable rhythm, since these were categorical variables, the most frequent value across all patients was retained for imputation.

2.8. Training of the model and performance evaluation

A 5-fold CV was chosen for the selection of the hyperparameters and evaluation of the performance. Folds were built only once (same folds for all experiments) and patient's data were not divided into different folds.

The weighted binary cross entropy (BCE) was used to handle class imbalance. We quantified the Challenge score (TPR at a FPR of 0.05) at 72 h as validation metric. The score required only one probability per patient to be computed, hence, probabilities provided by the models for each 5-minute segment were aggregated by a weighted average across hours. The average score across folds was used to select the hyperparameters. The training of the final model was performed using the selected hyperparameters and preprocessing on the public dataset. The training pipeline was then submitted to the Challenge organizers to obtain the official validation score.

3. Experiments

We conducted three experiments. In the first one, since the dataset was unbalanced with 225 patients with good outcome and 382 with poor outcome (positive class), we evaluated different weights for the samples of the poor out-

come class in the BCE loss. We tested the values: 1, 1/4, 1/10 and 225/382. In the second one, the final probability of poor outcome for each patient was obtained by aggregating the probabilities computed from all the available 5-minute windows over the 72 hours. Here, we tested the following weights: 1, x^2, $(1-x)^2$, $\frac{1}{1+e^{-10(x-0.5)}}$, $\frac{1}{1+e^{10(x-0.5)}}$ where $x = \mathrm{hours}/72$ (weights were then normalized to sum to 1). In the third one, we trained and compared all models separately and one combining EEG, ECG and clinical data. Models were trained with uniform weights across hours (*i.e.*, 1) for averaging the output probabilities.

4. Results

The optimal hyperparameters were: i) $K = 91$ samples, $F = 32$, $R = 2$, 256 and 128 neurons in the first and second FCLs; ii) batch size of 70, 350 epochs with learning rate of 1×10^{-3} and momentum of 0.9; iii) $\lambda_1 = 1 \times 10^{-5}$ and $\lambda_2 = 1 \times 10^{-4}$; iv) Dp probability of 0.3. In the first experiment, we identified the optimal positive class weighting as 1/4, that we then used for other experiments. In the second experiment, the Challenge score did not vary substantially across the weighted average formulas in CV (≈ 0.53; Fig. 1a). On the hidden validation set, weighting differently the probabilities seemed beneficial for prediction at 72 h, with a score of 0.63 for the weight x^2, which was approximately a 20% improvement with respect to uniform weights (0.63 vs 0.52). Yet, the other formula weighting more late recordings, *i.e.*, $\frac{1}{1+e^{-10(x-0.5)}}$, obtained a reduction of 13% (0.45 vs 0.52). When weighting more recordings near ROSC, using $(1-x)^2$ and $\frac{1}{1+e^{10(x-0.5)}}$, the improvement was approximately 15% (0.60 vs 0.52) and 12% (0.58 vs 0.52). In the third experiment, similar results were obtained across the four models, showing no improvements when ECG and clinical data were added to the EEG model's outputs (Fig. 1b). In CV, the combined model achieved slightly higher performance with respect to that of the EEG model. Also, variability across folds seemed lower for models with clinical variables in input with respect to EEG only (0.11 vs 0.16). For an additional comparison, Figure 1 also reports the area-under-the-ROC-curve (AUC) for each experiment. The Challenge scores achieved with x^2 weights and EEG only are reported in Table 1.

5. Discussion and conclusions

Here, we presented a DL model processing only one 5-minute window in each hour after ROSC and weighted all predictions to obtain the final probability of poor outcome. While tests performed on the public set and hidden validation set suggested that first hours after ROSC (3^{rd} and 5^{th} formulas in Fig. 1a) had more importance as compared to uniform weights, the final score on the hidden test set, which weighted more later hours, supported an opposite conclusion. The optimal timing still requires investigation.

Acknowledgments

Part of this research was supported by the MUSA - Multilayered Urban Sustainability Action - project, funded by the European Union - NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D".

References

[1] Oddo M, Rossetti AO. Early multimodal outcome prediction after cardiac arrest in patients treated with hypothermia. Crit Care Med 2014;42:1340–1347.

[2] Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, Tromp SC, Drost Gea. Outcome prediction in postanoxic coma with deep learning. Crit Care Med 2019;47:1424–1432.

[3] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[4] Reyna MA, Amorim E, Sameni R, Weigle J, Elola 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.

[5] Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, et al. The International Cardiac Arrest Research (I-CARE) Consortium Electroencephalography Database. Crit Care Med 2023;in press.

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

[7] Endoh H, Kamimura N, Honda H, Nitta M. Early prognostication of neurological outcome by heart rate variability in adult patients with out-of-hospital sudden cardiac arrest. Crit Care 2019;23:323.

[8] Rivolta MW, Stampalija T, Frasch MG, Sassi R. Theoretical value of deceleration capacity points to deceleration reserve of fetal heart rate. IEEE Trans Biomed Eng 2020;67:1176–1185.

[9] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016; 770–778.

Address for correspondence:

Massimo W Rivolta

Dipartimento di Informatica, Università degli Studi di Milano

Via Celoria 18, 20133, Milan, Italy

massimo.rivolta@unimi.it