Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest

Felix Krones[1], Ben Walker[2], Guy Parsons[1], Terry Lyons[2], Adam Mahdi[1]

[1] Oxford Internet Institute, University of Oxford, Oxford, UK, [2] Mathematical Institute, University of Oxford, Oxford, UK

Abstract

This work showcases our team's (The BEEGees) contributions to the 2023 George B. Moody PhysioNet Challenge. The aim was to predict neurological recovery from coma following cardiac arrest using clinical data and time-series such as multi-channel EEG and ECG signals. Our modelling approach is multimodal, based on two-dimensional spectrogram representations derived from numerous EEG channels, alongside the integration of clinical data and features extracted directly from EEG recordings. Our submitted model achieved a Challenge score of 0.53 *on the hidden test set for predictions made* 72 *hours after return of spontaneous circulation and was ranked* 14*th. Our study shows the efficacy and limitations of employing transfer learning in medical classification. With regard to prospective implementation, our analysis reveals that the performance of the model is strongly linked to the selection of a decision threshold and exhibits strong variability across data splits.*

1. Introduction

There are more than six million cardiac arrests annually, with general survival rates varying between 1% and 10% due to geographical disparities [1]. Following successful resuscitation, a significant proportion of survivors are admitted to intensive care units (ICUs) in a comatose state, with severe brain injury emerging as the leading cause of death among this group. In the crucial days following cardiac arrest, medical professionals are often tasked with estimating the likelihood of the patient regaining consciousness. A positive prognosis often leads to ongoing medical care, while a negative prognosis can result in the discontinuation of life support and hence death. There have been reported instances of patients recovering well despite a grim prognosis, raising concerns that negative predictions may inadvertently influence the outcome [2].

This year's Challenge [1, 3] asked to develop an open-source algorithm capable of predicting the extent of recovery from coma after a cardiac arrest. These predictions were to be made using a combination of basic clinical data, EEG, ECG and other signals, with the aim of classifying outcomes into either 'Poor' or 'Good'. In this work we develop a multimodal deep learning approach. Our strategy involves generating two-dimensional spectrogram representations sourced from multi-channel EEG signals and their integration with clinical data, along with features directly extracted from the EEG recordings.

2. Methodology

2.1. Data

The dataset for this study is taken from the International Cardiac Arrest REsearch (I-CARE) consortium and originates from seven academic hospitals across the United States and Europe [4]. It comprises *clinical data*, including age, gender, return time of spontaneous circulation (ROSC), arrest location (out-of-hospital cardiac arrest OHCA), presence of shockable rhythm, use of targeted temperature management (TTM), and *clinical time-series data*, such as continuous electroencephalography (EEG), electrocardiogram (ECG) and partially other recordings (e.g. SpO2). The dataset consists of $1{,}020$ patients (from which 607 were provided for training) who were admitted to an ICU in a comatose state following cardiac arrest. Neurological outcomes were assessed using the Cerebral Performance Category (CPC) scale.

We filtered the EEG signals by applying a band-pass filter over the range 0.5–30 Hz and a notch-filter at 50 and 60 Hz to mitigate artifacts from electrical grids. We employed artifact detection using a sliding window approach, only keeping the cleanest section of each recording.

For part of our model (Sec. 2.2) we converted the EEG signals to spectrograms (Figure 1), using the Python package `librosa`, to use them together with other non-imaging modalities as inputs for the models [5].

2.2. Models and architecture

In our approach, outlined in Figure 2, we evaluated six models for binary prediction of patient outcome at 72 hours after ROSC. Model M1 used clinical features and

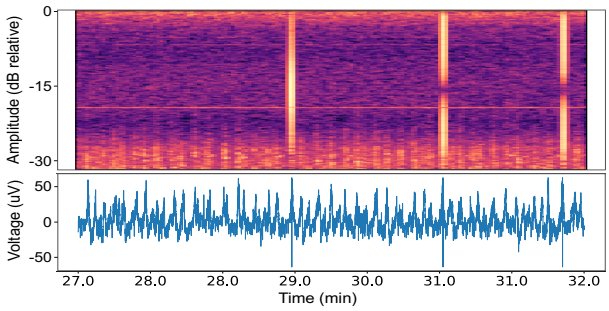

Figure 1. An example of EEG recordings (bottom) for patient 994 at the 8th hour of channel F8 with good outcome. The top displays the corresponding spectrogram (squared amplitude in decibel units rel. to peak power).

EEG summary statistics before employing a Random Forest classifier. In model M2, we added features extracted from EEG spectrograms using a DenseNet121-CNN [6] into the classifier. In model M3, these DenseNet features were further aggregated over time and channels. Model M4 introduced an intermediate fusion step for clinical features into the last layer of the DenseNet architecture. In models M5 and M6, we introduce additional output from a ridge regression classifier. This classifier is trained on features extracted from the EEG signals using the Random Convolutional Kernel Transform (ROCKET) [7]. We obtain the regularisation strength of the classifier through leave-one-out cross-validation over 10 log-evenly spaced values ranging from 10^{-3} and 10^{3}. ROCKET employs 10^4 kernels, with lengths randomly selected from the set $\{7, 9, 11\}$. Each kernel comprises 4 features and a maximum of 32 dilations. These settings align with the default parameters from sktime's `RocketClassifier`. Model M5 omitted the intermediate fusion present in model M4, whereas model M6 included it.

2.3. Scoring

The Challenge's official score emphasises the False Positive Rate (FPR) of incorrectly predicting a poor patient outcome. The scoring method selects the highest decision threshold that maintains the FPR below 5%, and subsequently evaluates the True Positive Rate (TPR) for predicting poor outcomes. Mathematically the score is defined, given a threshold θ drawn from the predictions, as

$$\max_{\theta:\mathrm{FPR}_\theta \leq 0.05} \mathrm{TPR}_\theta \tag{1}$$

Two distinct types of scores were reported. The 'Validation Score' was the assessment of our models provided by the Challenge organisers on the hidden set. The 'CV Score' and 'CV AUC' were computed as the mean and standard deviation, derived from the five-fold cross-validation performed on the local held-out set.

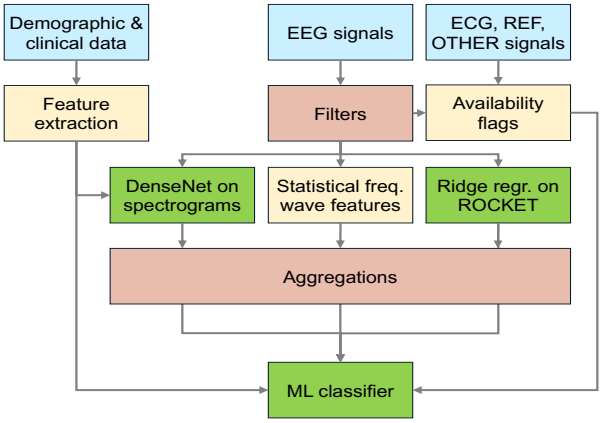

Figure 2. A schematic diagram of our model architecture. The colour codes are the following: blue: input data; red: filters and aggregation; yellow: pre-defined features; green: trainable models.

3. Results

Table 1 shows the feature availability and their overall key characteristics for the training data provided by the organisers of the Challenge collected from 607 patients.

Feature	Value	Missing
Age [years], mean (SD)	61 (16)	1
Sex ['Men'], N(%)	417 (69)	0
ROSC [minutes], mean (SD)	23 (19)	304
OHCA ['True'], N(%)	442 (78)	41
Shockable rhythm ['True'], N(%)	297 (52)	32
TTM, N at 33/36/na °C	448/61/98	0
Outcome ['Poor'], N(%)	382 (52)	0

Table 1. Summary of clinical data and patient outcome for the available 607 patients. ROSC is the time from cardiac arrest to return of spontaneous circulation; OHCA is 'True' for out of hospital arrests; TTM is the temperature management where 'na' stands for no TTM was applied.

Table 2 provides the official scores of the Challenge on the hidden validation set, along with the score and AUC obtained from our local cross-validation. Reported are the mean and standard deviation for the six models proposed in this work. Our final submission (model M5) achieved a Challenge score of 0.53 on the hidden test set. Figure 3 shows the relative feature importance of the classifier of our locally best performing model M6. The EEG signal information were the most important features while demographics and other clinical features exhibit lower importance. Figure 4 shows the dependency of error rates on the selection of a particular threshold. Figure 5 shows the ROC curve of our locally best performing approach (model M6) for which AUC=0.854.

Model	Features	Validation Score	CV Score	CV AUC
M1	Clinical data, EEG summary stats, Signal flags	0.520	0.327 (0.137)	0.688 (0.046)
M2	+ DenseNet on spectrograms	0.540	0.484 (0.157)	0.824 (0.023)
M3	+ aggregation of features over time and channels	0.567	0.527 (0.090)	0.811 (0.049)
M4	+ intermediate fusion	0.328	0.537 (0.104)	0.818 (0.038)
M5	+ ROCKET features without intermediate fusion	0.627	0.447 (0.085)	0.836 (0.033)
M6	+ ROCKET features with intermediate fusion	not provided	0.567 (0.085)	0.854 (0.015)

Table 2. The six models proposed and tested in this work (Sec. 2.2). The Validation Score is the performance on the Challenge's hidden validation set. CV Score is the mean (standard deviation) Challenge score from the five-fold cross validation on the training data (Sec. 2.3). CV AUC is the AUC on the same data. The '+' should be read as 'in addition'.

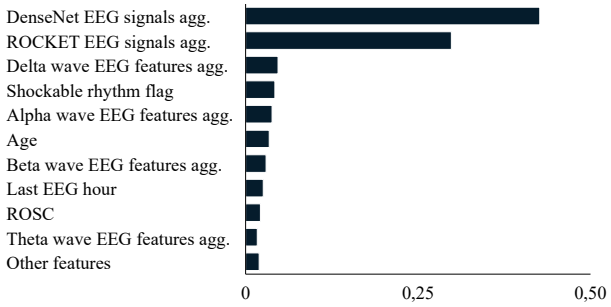

Figure 3. Relative feature importance plot for the locally best performing model M6. Here 'agg.' means aggregated over channels and time using the mean prediction and a majority voting for the DenseNet features.

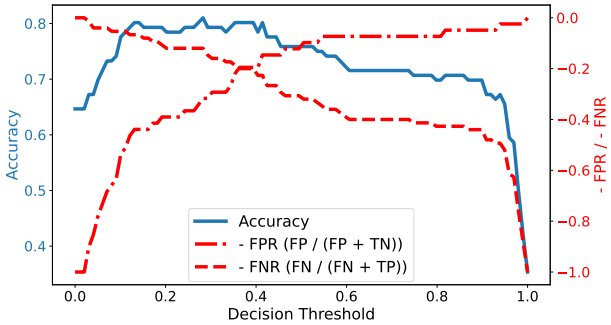

Figure 4. Metrics for different decision thresholds for the positive finding class. Accuracy (blue), False Positive Rate (red dashed) and False Negative Rate (red dashed-dotted) of the 'Poor' outcome label for different decision thresholds of the best performing model.

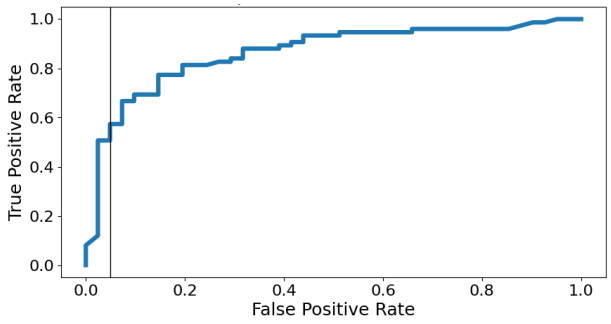

Figure 5. ROC curve (blue solid) of our final model (AUC = 0.854) with the indicated (black line) 5% FPR threshold.

4. Discussion

4.1. Main findings and limitations

In this study, we developed and tested six models for forecasting post-cardiac arrest coma recovery using a multimodal approach [5, 8]. Model M1 served as our benchmark. Incorporating CNN-extracted features from spectrograms in M2 led to the most significant AUC improvement in local cross-validation. To improve robustness, we aggregated features from different hours and channels in model M3. This aggregation effectively reduced the number of features used in the final classifier and improved the score, which was highly sensitive to minor variations on the left side of the AUC curve (Figure 5). Although model M4's intermediate fusion improved local score, its validation performance declined, indicating overfitting. Adding features extracted by ROCKET to the final classifier substantially improved performance on the validation set (0.627), but had limited impact on the local cross-validation set. Furthermore, performance deteriorated on the hidden test set (0.53), leading us to conclude that enhancing the robustness of our methodology remains an open research question. Additionally, given that ROCKET features did not markedly improve local results, it is still unclear how much additional information these features provide compared to those already extracted. Our results show that although multimodal machine learning approaches (models M2-M6) achieved strong performance on the local held-out set (CV AUC > 0.81), generalisability concerns emerged, which need further considerations.

4.2. Previous work

The majority of the predictive models developed to date using EEG signals predominantly employ CNNs, using

colour channels for diverse signals. Notably, these models are mostly constructed from datasets consisting of fewer than 300 patients [9]. Only a handful of studies have focused on predicting coma outcomes [9–11]. For example, [10] achieved an AUC of 91% at 66 hours after return of spontaneous circulation with a sensitivity of 66% for poor outcome prediction at a specificity of 95%.

4.3. Future research

Several potential avenues for future research are available. For instance, one can consider refining the modelling of time dependencies [10] and exploring self-supervised pre-training to boost network performance. The incorporation of other time-series data, such as ECG (not fully exploited in this study due to substantial missing data) could also lead to valuable insights, particularly when integrated with EEG signals [12]. This year's Challenge, however, demonstrated that the best-performing and most robust models were those that had undergone substantial feature engineering and data pre-processing [12]. The winning team, for example, extracted 362 expert-based EEG features plus additional ECG features [12].

Code availability

Our complete code is available on GitHub at https://github.com/felixkrones/physionet_challenge_2023.

Acknowledgement

FK was partially supported by the Friedrich Naumann Foundation. BW and TL were partially funded by the Hong Kong Innovation and Technology Commission. TL was funded in part by the EPSRC [EP/S026347/1], The Alan Turing Institute [EP/N510129/1], the Data Centric Engineering Programme (Lloyd's Register Foundation G0095), the Defence and Security Programme (UK Government funded), and the Office for National Statistics. Funding bodies had no influence on the research. No conflicts of interest existed.

References

[1] Reyna MA, Amorim E, Sameni R, Weigle J, Elola A, Bahrami A, Seyedi S, Kwon H, Zheng WL, Ghassemi M, van Putten MJAM, Hofmeijer J, Gaspard N, Sivaraju A, Herman S, Lee JW, Westover BM, Clifford GD. Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023. Computing in Cardiology 2023;50:1–4.

[2] Forgacs PB, Devinsky O, Schiff ND. Independent Functional Outcomes after Prolonged Coma following Cardiac Arrest: A Mechanistic Hypothesis. Annals of Neurology 2020;87(4):618–632.

[3] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[4] Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, Karukonda V, Lee JW, Herman ST, Sivaraju A, 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. In Critical Care Medicine 2023 (in press); doi:10.1097/CCM.0000000000006074.

[5] Walker B, Krones F, Kiskin I, Parsons G, Lyons T, Mahdi A. Dual Bayesian ResNet: A Deep Learning Approach to Heart Murmur Detection. In 2022 Computing in Cardiology (CinC), volume 498. 2022; 1–4.

[6] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 4700–4708.

[7] Dempster A, Petitjean F, Webb GI. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery 2020;34:1454–1495.

[8] Duvieusart B, Krones F, Parsons G, Tarassenko L, Papież B, Mahdi A. Multimodal Cardiomegaly Classification with Image-Derived Digital Biomarkers. In Medical Image Understanding and Analysis. 2022; 13–27.

[9] Jonas S, Rossetti AO, Oddo M, Jenni S, Favaro P, Zubler F. EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. Human brain mapping 2019; 40(16):4606–4617.

[10] Zheng WL, Amorim E, Jing J, Ge W, Hong S, Wu O, Ghassemi M, Lee JW, Sivaraju A, Pang T, Herman ST, Gaspard N, Ruijter BJ, Sun J, Tjepkema-Cloostermans MC, Hofmeijer J, van Putten MJ, Westover MB. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation 2021;169:86–94. ISSN 0300-9572.

[11] Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, Tromp SC, Drost G, Kornips FH, Beishuizen A, Bosch FH, Hofmeijer J, van Putten MJ. Outcome Prediction in Postanoxic Coma With Deep Learning. Critical Care Medicine 2019;47(10):1424–1432.

[12] Zabihi M, Zar AC, Grover P, Rosenthal ES. HyperEnsemble Learning from Multimodal Biosignals to Robustly Predict Functional Outcome after Cardiac Arrest. Computing in Cardiology 2023;50:Preprint.

Address for correspondence:

Felix Krones
OII, University of Oxford
1 St Giles, Oxford, OX23JS, UK
felix.krones@oii.ox.ac.uk