# Combining Complementary Models: Fusing CNNs, RNNs, and XGBoost for Enhanced Outcome Prediction of Comatose Patients after Heart Attack

Shuaixun Wang[1], Siyi Liu[1], Martyn G Boutelle[1]

[1]Imperial College London, London, UK

## Abstract

*Prognostication in comatose patients after cardiac arrest (CA) remains one of the biggest challenges for neurologists in the intensive care unit, as it shapes decisions about continuing or withdrawing life support. Electroencephalogram (EEG) provides valuable and non-invasive insights into patients' neurological status and has been used in many prediction models. However, traditional models often view EEG as stationary data, neglecting the dynamic patterns of EEG signals in response to internal and external perturbations. In addition, the importance of clinical data was underestimated in previous studies. We, team Data Doctors, took part in Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023, and proposed a prediction model based on data fusion, which scored 0.322 and ranked 29th in the official phase. We introduced a specialist system to combine various machine-learning frameworks, including a recurrent neural network (RNNs) for capturing dynamic EEG features, a convolutional neural network (CNNs) for identifying inter-channel EEG interactions, and an eXtreme Gradient Boosting (XGBoost) algorithm to synthesize these features for outcome prediction. The proposed model outperforms each single model, demonstrating the potential to improve outcome prediction accuracy and reliability by fusing complimentary results from different models.*

## 1. Introduction

Our team took part in Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023, which called on groups to create open-source automated systems for predicting patient outcomes after cardiac arrest using electroencephalogram (EEG) recordings and other data (1, 2). Quantitative analysis of EEG background activity provides an alternative prognostic tool (3, 4). EEG reflects real-time brain electrical activity and is sensitive to cerebral ischemia. Specific EEG patterns correlate with neurological outcomes after cardiac arrest. For instance, suppression of delta waves and alpha/theta rhythms predicts poor prognosis.

Machine learning applied to quantitative EEG analysis is a promising approach for early prognostication of neurological outcomes after cardiac arrest. Automated EEG interpretation can provide objective, accurate estimates of the likelihood of good cerebral recovery in individual patients to facilitate clinical decision-making after resuscitation.

In addition to using quantitative EEG features, we also explored prediction based solely on clinical data from the cardiac arrest patients. Clinical variables like age, gender, return of spontaneous circulation (ROSC), out-of-hospital cardiac arrest (OHCA), shockable rhythm, and targeted temperature management (TTM) have established prognostic values after cardiac arrest. We developed a separate machine learning model using only clinical data, then combined the predictions from the EEG models and clinical model. Fusing the multimodal results enabled us to leverage complementary prognostic information from both neurological and circulatory metrics. The ensemble approach of merging predictions from the EEG model and clinical model improved overall performance compared to either individual model. This demonstrates the benefit of synthesizing different data types, including both EEG and clinical circulatory arrest features, for enhanced prognostication accuracy after cardiac arrest.

## 2. Methods

The model was developed using the multi-center cardiac arrest dataset of the International Cardiac Arrest EEG consortium (ICARE) with 1020 adult patients from seven hospitals (5). This dataset includes multimodality monitoring data including EEG, oxygen saturation (SpO2), electrocardiogram (ECG), electromyography (EMG), and so on up to two weeks since 'return of spontaneous circulation'(ROSC). Clinical outcome was determined prospectively in two centers by phone interview (at 6 months from ROSC), and at the remaining hospitals retrospectively through chart review (at 3-6 months from ROSC). Neurologic outcomes were assessed using the Cerebral Performance Category (CPC) scale (1–5). Good outcome was defined as a CPC score of 1 or 2 (minimal to moderate neurologic disability), and poor outcome was defined as a CPC score of 3-5 (severe neurologic disability, persistent coma or vegetative state, or death).

Clinical data collected at admission includes age, sex, hospital ID, arrest location (in-hospital or out), cardiac rhythm at resuscitation (categorized as shockable or non-shockable), and the time interval from cardiac arrest to ROSC. To ensure privacy, ages above 89 are coded as "90". Post-arrest temperature is typically managed via a closed-loop feedback

ISSN: 2325-887X DOI: 10.22489/CinC.2023.297

device (TTM), barring contraindications like severe hypotension or delayed admission. Temperature settings include 36C, 33C, or unregulated (see Table 1).

| CPC group | CPC 1 | CPC 2 | CPC 3 | CPC 4 | CPC 5 |
|---|---|---|---|---|---|
| Number of patients | 181 | 44 | 20 | 9 | 353 |
| Age (years) | 58 | 58 | 65 | 56 | 63 |
| Female (%) | 50 (8.2%) | 12(2.0%) | 6(1.0%) | 3(0.5%) | 116(19.1%) |
| ROSC(average mins) | 20 | 18 | 19 | 24 | 25 |
| Shockable Rhythm | 133(21.9%) | 31(5.1%) | 6(1.0%) | 5(0.8%) | 122(20.1%) |
| TTM (33) | 136(22.4%) | 34(5.6%) | 11(1.8%) | 6(1.0%) | 261(43.0%) |
| TTM (36) | 20(3.3%) | 4(0.7%) | 5(0.8%) | 0(0.0%) | 32(5.3%) |

VFib = Ventricular Fibrillation; TTM = Targeted Temperature Management. EEG start time in hours (h) is relative to the time of cardiac arrest. Age and EEG data are shown as mean.

Table 1. Patient characteristics grouped by CPC scores.

Although multiple types of data are monitored, not all the data are collected continuously. Fig. 1 illustrates the availability of different data types over a period of 72 hours for three out of 1020 patients. EEG recordings are initiated soon after ROSC and continued up to 14 days post-cardiac arrest. Other physiological signals like ECG and EMG have more intermittent monitoring. Clinical variables such as age, sex, medical history are documented only at admission. This highlights the heterogeneous nature of the multimodal dataset, with EEG providing continuous quantitative brain monitoring while other data types are more sporadically sampled. There are some approaches to impute the time series (6, 7). However, we assume the discontinuity of data is also an important feature, thus no imputation is performed.

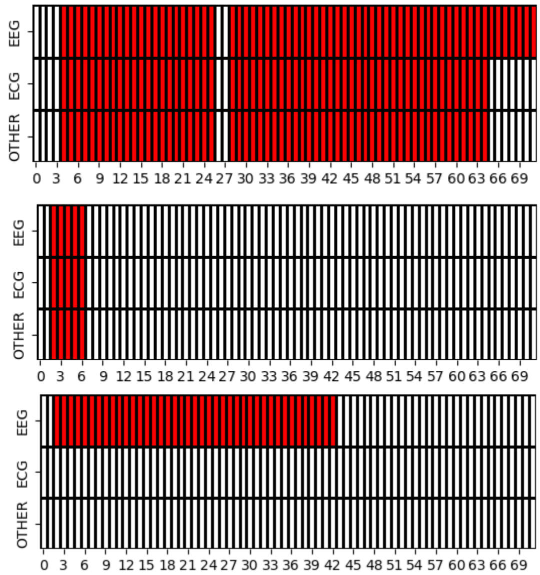

Figure 1. Existence of different data for patient 284 (top), 296 (middle) and 326 (bottom). Red block indicates the corresponding data exist at that hour.

Machine learning frameworks such as XGBoost, CNNs, and RNNs have consistently exhibited superior performance in a broad range of predictive modeling scenarios. However, their efficacy tends to wane when applied to datasets that are both smaller in size and riddled with noise. To counteract these limitations, we propose a specialist system that incorporates fusion logic to enhance the predictive capabilities of these standard machine learning approaches.

Three fusion methods have been implemented and compared, including weighted sum, Dempster-Shafer theory, and fuzzy logic. Weighted sum simply sums predicting probability of different models by assigning them different weights. Dempster-Shafer theory, also known as evidence theory or belief function theory, is a framework for modeling uncertainty and reasoning with partial information. It is an alternative to traditional probability theory and allows the explicit representation of uncertainty and ignorance. Fuzzy logic is based on fuzzy set theory, which allows for partial membership in sets defined by vague concepts like "tall" or "high skill." By using membership functions that map inputs to degrees of membership, fuzzy logic systems are capable of modeling complex relationships and patterns that elude capture by conventional binary logic and statistical methods. Consequently, fuzzy logic is particularly well-suited for addressing the inherent uncertainties and imprecisions that pervade numerous real-world datasets.

We developed three models to extract complimentary information from the dataset. XGBoost is utilized to handle patient demographic information, CNNs are utilized to scrutinize the frequency spectra, and RNNs are deployed to capture the temporal variations inherent in EEG data.

Specifically, the XGBoost model ingests categorical and numerical features like age, gender, and health history to predict disease risk. The XGBoost model stands out as a highly adaptable, precise, and interpretable tool, offering valuable insights into the importance of various features. This makes it particularly well-suited for predicting disease risk based on a mix of categorical and numerical variables such as age, gender, and health history. Moreover, XGBoost has the capability to automatically manage missing values, a common challenge in healthcare datasets. Several methods have been tried to explore the optimal parameters for XGBoost, including grid search, random search, and Bayesian optimization. The optimal parameters are found by random search, which gives {'reg_lambda': 100, 'reg_alpha': 0.1, 'max_depth': 3, 'learning_rate': 0.001, 'gamma': 0.1, 'colsample_bytree': 0.3}.

The CNNs are employed to scrutinize 2D representations of EEG frequency data for brain state classification. To construct these 2D matrices, we initially calculate the frequency spectra of EEG signals on an hourly basis. These are then aggregated into a 2D matrix of dimensions 72x512, as illustrated in Fig. 1. Here, '72' represents a span of 72 hours, and '512' denotes the 512 spectral data points computed for each hourly EEG segment. Within this matrix, each row signifies the frequency spectrum of EEG data for a given hour, while each column tracks changes in a specific frequency component over the course of 72 hours. CNNs are particularly effective for handling 2D spatial data due to their convolutional layers, which contain filters designed to identify spatial patterns and features. As these filters are applied across the entire matrix, CNNs can discern significant patterns irrespective of their spatial location. By applying these filters across the entire image, CNNs are able to recognize patterns regardless of where they appear in the image. We hypothesize that variations within specific frequency bands may serve as crucial features linked to patient outcomes, thus warranting the use of CNNs for data processing. The hyperparameters of CNN is determined empirically to achieve the highest accuracy.

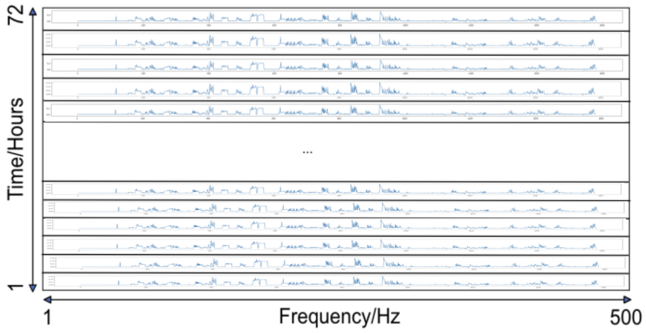

Figure 2. An illustration of structure of data for CNNs

Lastly, the RNN models the sequential changes in entropy and complexity of EEG. Several nonlinear parameters related to entropy and complexity are calculated for EEG data hourly, including Hjorth parameters, relative roughness, and decorrelation. RNNs are selected for this task precisely because of the capability to process time-dependent data. They excel in handling dynamic data types such as text, speech, and time series, owing to their recurrent connections. These connections enable the retention of information across sequential time steps. At each given time step, RNNs ingest new input while concurrently updating internal states based on both the current inputs and preceding states. This iterative process furnishes the RNNs with a form of temporal memory, thereby allowing them to capture and learn time-based dependencies and relationships. The hyperparameters of RNNs is determined empirically to achieve the highest accuracy.

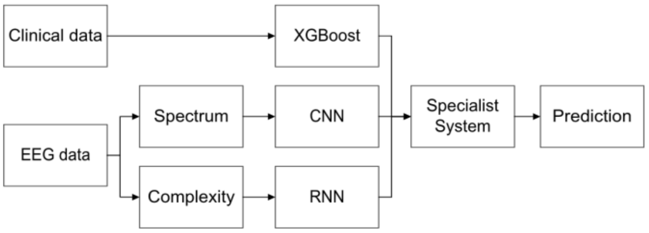

Figure 3. Structure of proposed machine learning model

The weights for weighted sum method are determined by going through all the possible weights for these three models within the range of 0 to 1. The fusing probabilities for Dempster-Shafer theory are determined by multiplying the probabilities for 0 of all three models as the predicting probability for 0 and multiplying the probabilities for 1 of all three models as the predicting probability for 1, and then normalize the results. The fusing process for fuzzy logic is much more complex. The first step is to determine the fuzzy regions. Here we defined three regions, namely low probability to be poor, medium probability to be poor and high probability to be poor. Then each output from the three models will be assigned a fuzzy value. For example, a prediction probability of 0.8 from XGBoost gives a fuzzy value of high probability to be poor. After transforming the outputs of three models into fuzzy region, we can summarize the fuzzy rules for prediction.

By fuzzifying the inputs to these models, our specialist system is able to improve predictive performance across all three modalities. The fuzzified demographic data helps XGBoost better assess risk gradients, the spectral inputs help the CNNs discern nuanced frequency patterns, and the temporal inputs allow the RNNs to better detect motif changes over time. Our approach demonstrates how data fusion can enhance disparate machine learning techniques applied to diverse biomedical data types.

## 3. Results

For analyzing risk factors, various statistical tests were employed to compare all the clinical features among different outcomes. Age and ROSC were found to be significantly different between different outcomes. For the age, elder individuals tended to have poorer outcomes. For the ROSC, a longer duration of ROSC is significantly associated with poorer outcomes.

We then explored the importance of the top two features by comparing the classification accuracy using the top two features and all features. The XGBoost model used here is using the default parameters rather than optimized parameters. From table 2 we can find that using two features have the similar performance as using all the features, suggesting that it is enough to conclude ages and ROSC as features when design prediction models based on clinical data.

| | Accuracy | F1 score |
|---|---|---|
| Two features | 0.5934 | 0.6992 |
| All features | 0.6264 | 0.7018 |

Table 2. Performance difference between using two features and all features (Results obtained on the training set).

We first compared the fusing results of weighted sum, dempster-Shafer theory, and fuzzy logic. From table 3, weighted sum outperforms the other two and thus was chosen to construct the specialist system.

| Fusing methods | Weighted Sum | Dempster-Shafer | Fuzzy Logic |
|---|---|---|---|
| Accuracy | 0.7857 | 0.3242 | 0.3956 |
| Specificity | 0.5254 | 1.0000 | 1.0000 |
| F1 score | 0.8517 | 0 | 0.1912 |

Table3. Comparison of performance of different fusing methods (Results obtained on the training set).

We then compare the performance of XGBoost, CNN, RNN and fusing results. From table 4 and fig. 4 we can see that there is a significant improvement in performance by fusing the prediction results from the three models. The challenge scores for each single model and fusing model have been calculated. However, only score for XGBoost has been obtained due to excessive computations required to get features for CNN and RNN. XGBoost scores 0.359 on the training set, 0.209 on the validation set, and 0.322 on the test set.

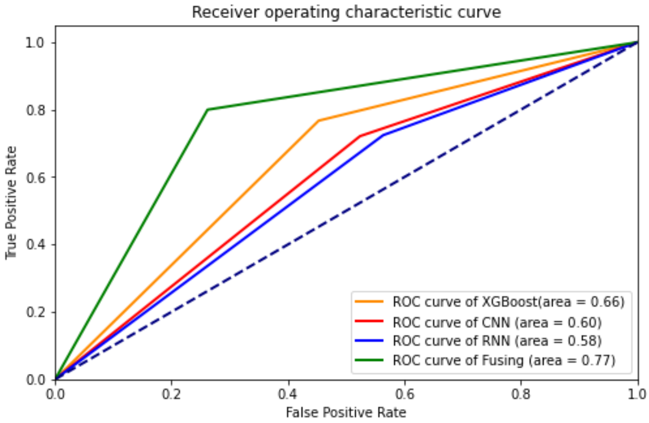

Figure 4. ROC curves for XGBoost, CNN, RNN and fusing results (Results obtained on the training set).

| Models | XGBoost | CNN | RNN | Fusing Results |
|---|---|---|---|---|
| Accuracy | 0.7033 | 0.6648 | 0.6374 | 0.7857 |
| Precision | 0.7674 | 0.7214 | 0.7244 | 0.8000 |
| Recall | 0.8049 | 0.8211 | 0.7480 | 0.9106 |
| Specificity | 0.4915 | 0.3390 | 0.4068 | 0.5254 |
| F1 score | 0.7857 | 0.7681 | 0.7360 | 0.8517 |
| Challenge Score | 0.3220 | N/A | N/A | N/A |

Table 4. Comparison of performance among XGBoost, CNN, RNN and fusing results (Challenge score obtained on the test set, the others obtained on the training set).

## 5. Conclusion

Early testing shows that adding the specialist system improves predictive accuracy across multiple datasets and modeling tasks. The specialist system appears to act as a regularizer that makes the XGBoost, CNN, and RNN models more robust to noise and variability. We hypothesize the specialist system allows the models to better learn subtle nuances and patterns that are obscured when using only crisp, deterministic values.

In this work, we detail the architecture and training process of our proposed specialist system. We present experimental results demonstrating improved accuracy on several benchmark datasets and models. The ability to enhance state-of-the-art techniques like XGBoost, CNNs, and RNNs by fusion shows the potential of hybrid intelligent systems. Our specialist system provides a simple but powerful approach to improving predictive modeling performance.

There are some limitations though. First, the overall performance is not high. This can be a result of some complexity parameters that require a long time to calculate, including Lyapunov exponent, multiscale entropy and Lempel-Ziv complexity, are abandoned due to limited running time. The performance of RNN can be further improved by incorporating more nonlinear features. Second, both RNN and CNN are affected by the missing data. Therefore, both learn from the pattern of missing data. This can offset the complimentary effects. Third, we failed to get challenge scores for RNN, CNN and fusing results due to limited computing resources and time limitations. Finally, if we consider a correct prediction as either of the models give a correct prediction, the accuracy is 0.9286. However, the fusing accuracy we’ve got is 0.7857. There should exist fusing architecture that can further increase the prediction performance.

## Acknowledgment

This work is supported by China Scholarship Council.

## 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):E215-20.

[2] Reyna MA AE, 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] Lee S, Zhao X, Davis KA, Topjian AA, Litt B, Abend NS. Quantitative EEG predicts outcomes in children after cardiac arrest. Neurology. 2019;92(20):e2329-e38.

[4] Ghassemi MM, Amorim E, Alhanai T, Lee JW, Herman ST, Sivaraju A, et al. Quantitative Electroencephalogram Trends Predict Recovery in Hypoxic-Ischemic Encephalopathy*. Critical Care Medicine. 2019;47(10):1416-23.

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

[6] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports. 2018;8(1):6085.

[7] Luo Y, Cai X, Zhang Y, Xu J. Multivariate time series imputation with generative adversarial networks. Advances in neural information processing systems. 2018;31.

Address for correspondence:

Martyn Boutelle
South Kensington Campus, Imperial College London, London, SW7 2AZ, UK
m.boutelle@imperial.ac.uk