Neurological Outcome Prediction After Cardiac Arrest: A Multi-Level Deep Learning Approach with Feature and Decision Fusion

Bill Chen[1], Jiamu Yang[1], Ke W Wang[1], Hayoung Jeong[1], Perisa Ashar[1], Leeor Hershkovich[1], Md Mobashir Hasan Shandhi[1], Jessilyn P Dunn[1]

[1]Duke University, Durham, NC, USA

Abstract

Cardiac arrest leads to complex neurological outcomes, demanding accurate predictions to guide post-arrest care. Using the International Cardiac Arrest Research Consortium (I-CARE) dataset, we developed models to discern between "good" and "poor" neurological outcomes post-cardiac arrest. We concatenated clinically relevant, manually extracted EEG features with autoencoder-derived, automatically extracted features to train transformer and Bi-LSTM models. Additionally, we ensembled the predicted probabilities between these deep learning models with a statistical model trained on non-EEG clinical variables. This ensemble approach demonstrated that the transformer excel at capturing long-term temporal dependencies, and the fusion of features and prognosis decisions led to improved model performance in terms of AUROC in predicting neurological outcomes post-cardiac arrest.

1. Introduction

Annually, over 6 million people worldwide experience cardiac arrests. Most survivors incur severe brain injury, and many, once in the ICU, remain comatose [1]. In the days after the arrest, physicians give a prognosis on the patient's chances of regaining consciousness. A positive outlook may extend care, whereas a negative one may lead to discontinuing life support. Yet, the subjectivity in these prognoses can lead to inaccuracies, adversely affecting patient outcomes [2].

Electroencephalography (EEG) is an objective tool to monitor brain activity and estimate neurological recovery after cardiac arrest. However, the analysis of EEG signals requires considerable resources and the expertise of specialized neurologists [3]. Advancements in machine learning (ML) offer automated EEG analysis solutions that can improve prognostic accuracy while making the process more accessible for healthcare systems. Particularly, deep learning (DL) techniques have outperformed traditional ML in handling the complexities of EEG data, with sequential Bidirectional Long Short Term Memory (Bi-LSTM) networks proving adept at capturing dependencies for interval predictions [4]. The same group also demonstrated that the Bi-LSTM model can be further improved by using Convolutional Neural Network (CNN) extracted features and decision fusion with a Random Forest (RF) classifier [5]. Separately, Hessulf et al. showed that an Extreme Gradient Boosting model can achieve high performance using non-EEG clinical data [6].

Utilizing the I-CARE dataset, our study aims to develop a prognostic framework for differentiating "good" versus "poor" neurological outcomes, approached as a binary classification challenge. We improved on existing methods by integrating manually extracted EEG features with those obtained from unsupervised representation learning. We also explored the efficacy of time series transformers in capturing these features and their temporal dependencies. Our approach combines predictions from the deep learning (DL) model with a clinical data-driven model, highlighting the benefits of feature and decision fusion, and evaluating whether transformers outperform traditional sequential models in this domain.

2. Methods

2.1. Dataset and Preprocessing

The dataset comprises continuous EEG readings from over 1,000 comatose patients. Out study uses a subset of 509 subjects with over 24 hours of complete 19-channel EEG recordings. Beyond the EEG data, this dataset incorporates clinical variables such as patient demographics, targeted temperature management (TTM), return of spontaneous circulation (ROSC), and the presence of shockable rhythms [5]. The neurological outcomes are categorized using the Cerebral Performance Category (CPC) scale, which ranges from 1 to 5. A score of 1 indicates optimal neurological function and 5 signifies mortality.

The EEG data was preprocessed with a 6th order Butterworth bandpass filter with cutoff frequencies at [0.5, 30] Hz, also removing the utility frequencies [5]. The filtered

data was then downsampled to a rate of 100 Hz. To generate binary classification labels of patient outcomes, we grouped CPC scores of 1 to 3 and 4 to 5 as "good" and "poor" outcomes, respectively.

2.2. Feature Extraction

We employed both clinically relevant, manual feature extraction and data-driven, automated feature extraction approaches at non-overlapping 5-minute intervals shown in Fig. 1a. The use of both manual and automated approaches enhances the robustness of the features extracted, improving their suitability for model training.

For manual feature extraction, we derived eight distinct features that were previously identified by neurophysiologists as being predictive of poor outcomes [7]. These features include the power bands for δ (0.5-4 Hz), θ (4-7 Hz), α (8-15 Hz), and β (16-30 Hz), as well as the α/δ ratio, Shannon entropy, burst suppression ratio and regularity. Power band features were extracted from the estimated Power Spectral Density (PSD) using the Morlet wavelet. Shannon entropy was calculated to capture the signal's complexity and predictability. The burst suppression ratio, indicative of poor neurological recovery, was extracted using recursive mean and variance estimation. Lastly, regularity was calculated to discern continuous patterns from burst suppression patterns.

We trained an autoencoder (AE) for automated feature extraction (Fig. 1b). The AE encodes EEG signals into a latent space representation and uses a decoder for signal reconstruction. The architecture for the encoder and decoder was adapted from EEGNet—a compact CNN known for its effectiveness in brain-computer interface tasks [7]. As described in Fig. 1b, EEGNet begins with a temporal convolution, followed by a depthwise convolution for spatial representations. Average pooling was used for dimensionality reduction, and dropout layers were used for regularization. Lastly, a separable convolution is employed, learning the individual feature maps before merging them to provide a detailed representation. After training the AE to maximize the peak signal-to-noise ratio, we used the encoder to automatically extract features from the EEG segments.

2.3. Pipeline and Experiment Setup

We aimed to predict neurological outcomes of each patient using 6-hour epochs of EEG feature data, spanning from 0 to 72 hours post-ROSC. For our analyses, we worked with three distinct EEG feature sets. The first set was based on manually extracted features, the second drew from the autoencoder, and the third combined features from both methods. Each 6-hour epoch was paired with an average of all previous epochs to expand our training samples and offer historical context. Missing data

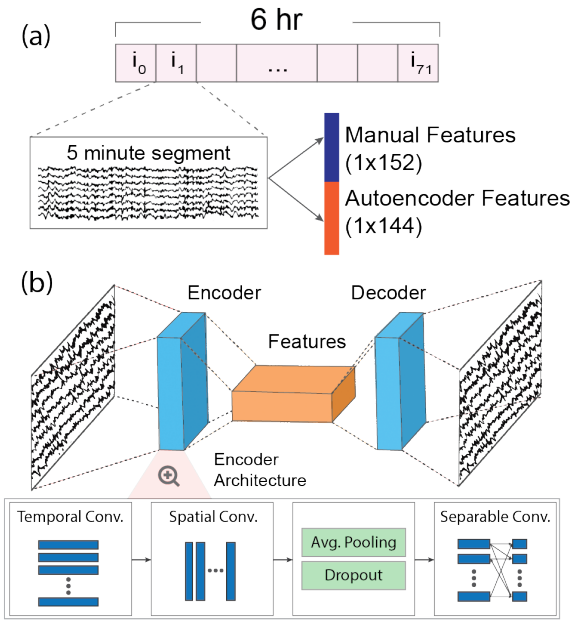

Figure 1. (a) Each non-overlapping 5-minute window is represented with features of size 1x296 after applying both manual and automated feature extractions. (b) The autoencoder used for automated feature extraction consists of an encoder and a decoder adapted from EEGNet.

within epochs was imputed using data from the nearest epoch. We split the dataset into 70% training, 10% validation, and 20% testing, ensuring no overlap of patient data across these sets.

Two deep learning models were employed: a Bi-LSTM and a Transformer with linear and convolutional embeddings, plus randomized positional encoding. Both aimed to minimize the negative log likelihood loss. We applied grid search for hyperparameter tuning. Early stopping was set at 20 unchanged epochs. The Adam optimizer was used with L2 regularization and a learning rate scheduler that adjusted rates if the loss plateaued over 5 epochs. Training was performed on a GeForce RTX 4090 GPU.

For clinical data, we used features including Age, Sex, Hospital, ROSC, out-of-hospital cardiac arrest presence, TTM, and initial rhythm shockability. We trained statistical models such as Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) on this data [6]. The final output combined predictions from these models with our deep learning models to estimate the probability of poor outcomes at each 6-hour epoch.

3. Results

We assessed our models on the test set using Accuracy, F1-Score, and Area Under the Receiver Operating Characteristic curve (AUROC), measured at 6-hour intervals

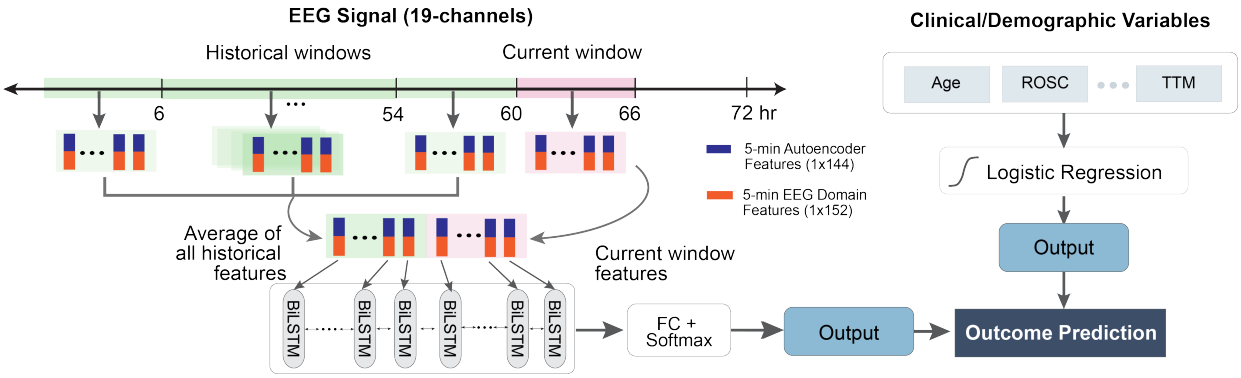

Figure 2. For every 6-hour epoch, the pipeline concatenates the average features from all historical windows with the features from the current window. BiSTM and Transformer models are applied to the combined features. The output of these models is then combined with the output from a logistical regression model that was trained with clinical variables. *FC*: Fully connected layers.

from 12 hours post-ROSC. Fig. 3 illustrates the averaged AUROC over time, contrasting the transformer (3a) and Bi-LSTM (3b) models trained with different feature sets (manual vs. autoencoder), and their combination. Beyond 30 hours, the combined features yielded superior results compared to using either alone. Notably, integrating probabilistic estimates from the LR clinical data model enhanced AUROC across all time points.

During the initial 12 to 30 hours post-ROSC, both models exhibited low AUROC scores, with a noticeable performance improvement between 48 and 60 hours, before a post-60-hour decrease. Isolating the feature sets revealed distinct model behaviors: the transformer consistently outperformed on autoencoder-derived features, while the Bi-LSTM showed strength with manually extracted features, matching or surpassing the combined features at 18 and 42 hours. The transformer's average AUROC followed a stable, predictable path, unlike the Bi-LSTM's, which displayed more erratic and variable progression over time.

For a more holistic view on model performance, Table 1 shows a comparative display of Accuracy and F1-Score for the deep learning and clinical data models at the 48 hour post-ROSC mark. Contrary to the AUROC performance, both the transformer and Bi-LSTM achieved higher accuracy and F1-Score using the autoencoder-derived feature as oppose to the combined features. Among the three clinical data models we examined, LR resulted in the best performance across all metrics, substantiating its integration into our ensemble for final decision fusion.

The incorporation of a logistic regression clinical model has notably improved early predictions when EEG data is limited. As time progresses, the benefit of this addition diminishes as the deep learning model's performance strengthens. Nonetheless, the clinical model's contribution to early prognostication remains a valuable asset for initial assessments post-ROSC.

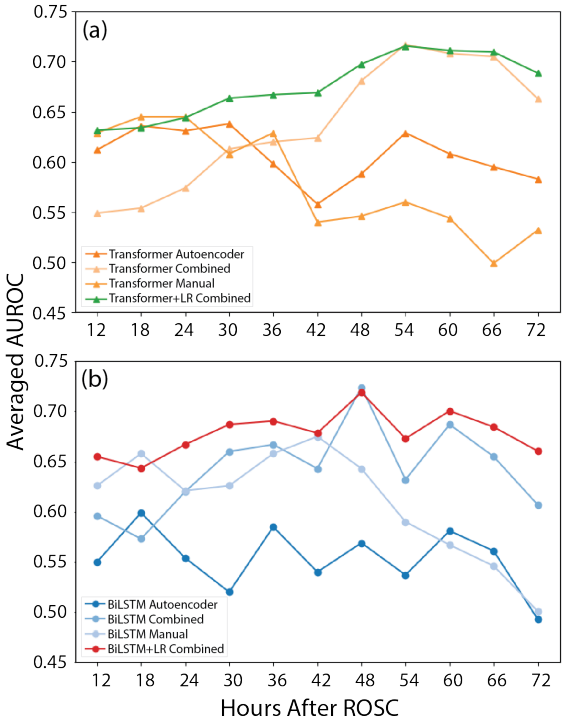

Figure 3. AUROC at every 6 hour interval for (a) transformer models (b) Bi-LSTM models with each feature set.s

4. Discussion

Our study highlights that by combining autoencoder-derived and manually extracted features, we were able to improve the AUROC for transformer and Bi-LSTM models. The transformer model, in particular, display a marked increase in performance over time. Furthermore, we demonstrated that by ensembling the deep learning model predictions with those from the clinical data model

Table 1. Additional Metrics for Models at 48 Hours

Model	Feature	Accuracy	F1-Score
Bi-LSTM	Manual	0.55	0.62
	Autoencoder	0.60	0.68
	Combined	0.59	0.62
Transformer	Manual	0.60	0.68
	Autoencoder	0.67	0.75
	Combined	0.63	0.70
LR		**0.81**	**0.88**
SVM	Clinical	0.71	0.80
RF		0.74	0.81

strengthens the prognostic performance.

We also observed differences in different time series model architectures favoring different kinds of features sets. Transformers excel with autoencoder-derived features, possibly due to their capacity to process non-linear, long-range dependencies [8]. In contrast, the Bi-LSTM model demonstrates a propensity for manually extracted features, reflecting its capacity to process immediate temporal sequences. Moreover, the transformer model's performance not only remains consistently higher over time but also aligns more closely with the clinical timeline, where prognostication is customarily deferred several days post-ROSC, thereby underscoring its suitability for capturing the prognostic patterns essential for delayed decision-making in clinical settings.

Lastly, we observed a discernible change in the AUROC commencing at 24 hours post-ROSC, as well as a pronounced decline at the 72-hour mark. This is most likely due to the density difference of training segments: there are 3000 segments at 30 hours and 1600 at 72 hours. The temporal alignment of peak model performance with periods of heightened data availability suggests that the models benefit from the increased volume and richness of the contextual data provided during these intervals.

We recognize that the study has several limitations. The concatenation approach used for the manual feature extraction from individual channels could benefit from dimensionality reduction to mitigate noise. The lack of artifact removal in our methodology introduces additional noise that future studies should address. An exploration into the specific contributions of attention mechanisms, perhaps with an attention-based LSTM model, can lead to insights on why the transformer is better at modeling nuanced dependencies [9].

5. Conclusion

Our multi-level deep learning strategy shows potential for predicting neurological outcomes following cardiac arrest. Fusing manually extracted and autoencoder-derived EEG features, alongside ensembling with a clinical model, our framework improves predictive accuracy. The transformer models outperform Bi-LSTM for autoencoder and combined features. Including LR predictions from clinical data also enhances early post-ROSC predictive performance in the absence of EEG. Future efforts should optimize feature extraction and increase the model's tolerance to noise. This work establishes a baseline for automatic neurological prognostication in ICU settings.

References

[1] Shinozaki K, Nonogi H, Nagao K, Becker LB. Strategies to improve cardiac arrest survival: a time to act. Acute Medicine Surgery 2016;3.

[2] Rundgren M, Westhall E, Cronberg T, Rosén I, Friberg H. Continuous amplitude-integrated electroencephalogram predicts outcome in hypothermia-treated cardiac arrest patients. Critical Care Medicine 2010;38:1838–1844.

[3] Hofmeijer J, Beernink TM, Bosch FH, Beishuizen A, Tjepkema-Cloostermans MC, van Putten MJ. Early eeg contributes to multimodal outcome prediction of postanoxic coma. Neurology 2015;85:137 – 143.

[4] Zheng WL, Amorim E, Jing J, Wu O, Ghassemi MM, Lee JW, Sivaraju A, Pang TD, Herman ST, Gaspard N, Ruijter BJ, Tjepkema-Cloostermans MC, Hofmeijer J, van Putten MJ, Westover B. Predicting neurological outcome from electroencephalogram dynamics in comatose patients after cardiac arrest with deep learning. IEEE Transactions on Biomedical Engineering 2021;69:1813–1825.

[5] Zheng WL, Amorim E, Jing J, Ge W, linda Qiao, Wu O, Ghassemi MM, Lee JW, Sivaraju A, Pang TD, Herman ST, Gaspard N, Ruijter BJ, Sun J, Tjepkema-Cloostermans MC, Hofmeijer J, van Putten MJAM, Westover MB. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation 2021;.

[6] Hessulf F, Bhatt DL, Engdahl J, Lundgren P, Omerovic E, Rawshani A, Helleryd E, Dworeck C, Friberg H, Redfors B, Nielsen N, Myredal A, Frigyesi A, Herlitz J, Rawshani A. Predicting survival and neurological outcome in out-of-hospital cardiac arrest using machine learning: the scars model. eBioMedicine 2023;89.

[7] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance B. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of Neural Engineering 2016;15.

[8] Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L. Transformers in time series: A survey. In International Joint Conference on Artificial Intelligence. 2022; .

[9] Karim F, Majumdar S, Darabi H, Chen S. Lstm fully convolutional networks for time series classification. IEEE Access 2017;6:1662–1669.

Address for correspondence:

Dr. Jessilyn Dunn

534 Research Dr, Room 448, Durham, NC 27708

jessilyn.dunn@duke.edu