Predicting Neurological Outcomes of Comatose Cardiac Arrest Patients Using Transformer Neural Networks with EEG Data

Jefferson Dionisio[1], Che Lin[1,2,3,4,5,6], Lian-Yu Lin[1,7], Wen-Chau Wu[1,8,9,10]

Abstract

As part of the 2023 PhysioNet Challenge, our team FINDING_MEMO utilized Transformer to predict outcomes using patient EEG data since it excels at dealing with sequential data like EEG. We mainly used the Transformer encoder block's multi-head self-attention to generate representations from the input and leverage several hidden layers to form the final prediction. Using the latest EEG from every patient, our team achieved the challenge score of 0.42 with the hidden validation set (ranked 36th out of 73 invited teams) and obtained a result of 0.37 (ranked 29th out of 36 qualified teams). Our results show a consistent performance across varying EEG recording durations in both the validation and test set. Our team also had the second-best score when evaluated, with only 12 hours of available recordings in the test set. Such promising results showcase the models' generalizability and clinical potential in predicting outcomes for comatose patients, especially for limited available EEG recordings.

1. Introduction

This paper comes from our team FINDING_MEMO's participation in the "Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023" [1-2]. This year's challenge objective was predicting the neurological outcomes of comatose patients following cardiac arrest. Teams were tasked with classifying their outcomes as either "good" or "bad" using the cerebral performance category (CPC) scale based on data such as electroencephalogram (EEG), electrocardiogram (ECG), and other clinical information. EEG, which captures electrical brain activity, serves as a valuable indicator of a patient's neurological recovery. This challenge's primary evaluation metric is the true positive rate at a false positive rate of 0.05.

The dataset [3] consists of 1020 patients, which were further divided into 60% training, 10% validation, and 30% test set. All subsequent experiments were trained and cross-validated using the 607 patients training dataset and evaluated with the remaining datasets for the leaderboard.

2. Methods

2.1 Inclusion and Exclusion Criteria

In recent years, numerous studies [4-6] have emphasized the potential of EEG signals in unraveling the intricate workings of the brain, particularly in predicting outcomes for post-anoxic comatose patients. These studies consistently demonstrated the highest sensitivity and specificity when leveraging earlier EEG signals. They have effectively showcased EEG's potential in addressing brain-related disorders. Consequently, we focused exclusively on EEG channels for all our experiments, excluding others such as those from ECG.

This paper covers various trials, including early and recent experiments, to assess our model's ability to predict neurological outcomes in comatose patients, highlighting successes and areas for improvement.

In our primary experiment (TRS1), which serves as our team's main submission, we exclusively utilized the most recent hour of EEG signals alongside each patient's clinical data. The objective of this trial was to evaluate the model's predictive performance while working with a limited dataset of EEG information, specifically focusing on just one representative hour. It is worth noting that prior research studies [4-6] have hinted at the superior predictive potential of utilizing earlier EEG time points. However, our decision to focus on the last hour of data was influenced by a significant issue: the presence of extensive missing data in the initial hours of the dataset.

Furthermore, it is essential to consider that the challenge at hand emphasizes the development of predictive models using a 72-hour data window rather than shorter time frames, such as 12 or 24 hours. Given this context, it becomes more logical to incorporate later data from the early EEG records provided, aligning our approach with the challenge's requirements.

In our subsequent experiments, denoted as TRS2, TRS3, and TRS4, we employ varying subsets of available EEG

[1] Smart Medicine and Health Informatics Program, National Taiwan University (NTU), Taipei, Taiwan
[2] Department of Electrical Engineering, NTU
[3] Graduate Institute of Communication Engineering, NTU
[4] School of Medicine, NTU
[5] Center for Advanced Computing and Imaging in Biomedicine, NTU
[6] Center for Biotechnology, NTU
[7] Department of Internal Medicine, NTU Hospital, Taipei, Taiwan
[8] Institute of Medical Device and Imaging, NTU
[9] Department of Radiology, College of Medicine, NTU
[10] Graduate Institute of Clinical Medicine, NTU

data from the initial hours for training and evaluating the model: the first 24 hours, the first 48 hours, and the first 72 hours, respectively. The primary objective of these experiments was to assess whether the predictive performance improves when using shorter time intervals (24 and 48 hours) compared to the 72-hour dataset. Additionally, we aimed to investigate whether the significant amount of missing data in the earlier hours negatively impacts predictive performance. Furthermore, we intended to explore whether using multiple hours of data surpasses the predictive capability achieved when relying on just one hourly data per patient.

It is crucial to note that during the training phase, patients who lacked data for 24, 48, or 72-hour hours were excluded from contributing to the model's training in the respective experiment. This ensures that our analysis is based on patients with actual available data for each experiment, allowing us to draw meaningful conclusions regarding predictive performance from EEG.

2.2. Feature Extraction and Preprocessing

TRS1 was an initial trial designed primarily to evaluate the model's capabilities. Consequently, most of the feature extraction and signal processing steps utilized default settings provided by the challenge to assess how effectively the model could enhance performance under controlled conditions.

For TRS1, the signal processing involved bandpass filtering within the range of [0.1, 30] Hz, followed by resampling to reduce noise. The resulting signals were then normalized to a range of [-1, +1]. Only the F3, F4, P3, and P4 channels were employed, with bipolar referencing applied to F3-P3 and F4-P4. EEG features from each reference were extracted from the power spectral density (PSD) of alpha, beta, delta, and theta waves. In conjunction with the 8 available clinical features, each patient was characterized by a total of 16 features.

In contrast, TRS2, TRS3, and TRS4 represented subsequent trials to refine the model in various aspects. These experiments followed the EEG feature extraction processes outlined in [7] and [8]. In these trials, all 19 EEG channels were used, employing 18 bipolar referencing according to the longitudinal bipolar montage. Each of the 18 bipolar references contributed the same 4 features, resulting in a total of 72 EEG features. Combined with clinical data, each patient profile consisted of 80 features per hourly EEG data.

Signal preprocessing for TRS2, TRS3, and TRS4 involved bandpass filtering within the range of [0.5, 30] Hz, followed by resampling to 100 Hz for noise reduction, followed by normalization to values from -1 to +1.

These used sequentially organized data as input to assess the model's capacity for handling time series data. For TRS2, all hourly EEG data from the first 24 hours were utilized, with missing hours filled in with NaN values. Subsequently, each hourly data point was concatenated with the corresponding clinical data, as they all originated from the same patient. TRS3 and TRS4 employed 48 and 72 hours of data, respectively. To address missing values, imputation was performed using the mean value for each feature. Both EEG and patient features were further normalized to values from 0 to 1.

2.3. Vanilla Transformer

Inspired by [7] and [8], who used CNN-LSTM and bidirectional LSTM models for time series, we adopted Transformers for our implementation. Transformers excel at handling sequential data, which is crucial for our task given EEG data's time-dependent nature and the challenge's objective of predicting future outcomes from earlier data.

The Transformer model, introduced in [9], includes encoder and decoder blocks that employ multi-head self-attention for creating data representations. The multi-head mechanism calculates distinct attention weights for different input parts, enabling selective attention to more important parts. The encoder generates a feature embedding from the input, while the decoder constructs an output based on the input embedding.

Transformers are adept at capturing sequential patterns and have diverse applications, especially in handling sequential time series data, where their sequence-processing capabilities shine. Some variants, like Gated Transformer Networks (GTN) [10], tackle specific time series challenges, exhibiting exceptional classification performance across 13 multivariate time series datasets. GTN outperformed other benchmarks significantly, demonstrating Transformers' potential in handling various sequential time series data types.

2.4. Model Implementation

Two parallel Transformer models were trained for each experiment: one for outcome and another for CPC prediction. Figure 1 illustrates the model architecture. Both models make use of the Transformer encoder to generate the final prediction, omitting the decoder component. This is because the decoder's primary function is to reconstruct the embedding to match the input's top level, which is not necessary for these tasks. Both models take the pre-processed patient features as input and produce an embedding through the embedding layer with the same dimensions as the input.

For TRS1, since we used only the most recent hour of EEG recording, this resulted in one row of data per patient. The feature rows from the different patients were then combined into a single training dataset. We then randomized the order and trained the models in batches of 10, which aided in enhancing their generalizability.

In the case of TRS2, TRS3, and TRS4, we took a different approach. Instead of mixing the data, we extracted the EEG features per hour and arranged them in sequential order per patient prior to training. For instance, as illustrated in Figure 2, the training process with TRS4 involved stacking all 72 hours of extracted EEG features with the clinical data for each patient in sequential order, with each H representing the corresponding hour for each feature row. For each patient, we trained the model using their data as a single batch. In cases where data was missing, such as the third hour of data for patient 1, we used mean imputation to fill in the gaps for the respective features. We employed this strategy to ensure that the data of each patient remained distinct and were not mixed with other patients' data. Masking was also applied to these three implementations to mask out the imputed hours, aiding the models to focus more on the available data.

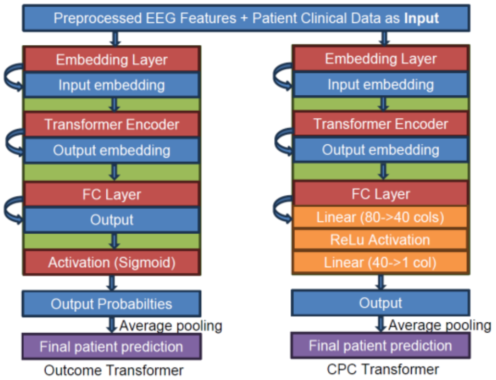

Figure 1. Transformer model architectures, Outcome Transformer (left) and CPC Transformer (right), using the preprocessed EEG features and clinical data as the input.

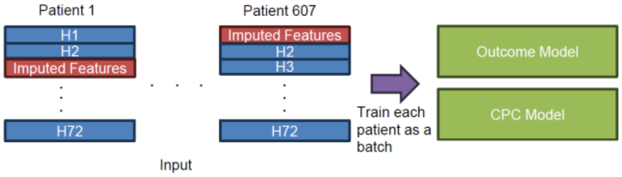

Figure 2. Model training pipeline for TRS4 with 'H' representing each hour of features and the red boxes representing the imputed missing hours.

The input embedding is processed by two layers of stacked Transformer encoders, utilizing 11 heads for TRS1 and 8 heads each for TRS2, TRS3, and TRS4. Multi-head attention is employed to form the output embedding. In the latter three implementations, a masking vector is applied as an additional input to the encoder.

The transformer for outcome prediction employs BCEWithLogitsLoss as its loss function and uses RAdam as the optimizer in all implementations. In contrast, the transformer for CPC uses L1Loss as its loss function and employs RMSProp as the optimizer. We selected these hyperparameters through several rounds of tuning and trained them for 5 epochs each.

We then use a fully connected layer to finalize the outputs, applying the Sigmoid activation function for the outcome model to calculate the probabilities. In the last step, during both training and testing, an average pooling operation is performed on each column of data to derive the final prediction for the current patient, with the specific pooling determined by the batch size.

For cross-validation, TRS1 employs k-fold (k=5), while TRS2, TRS3, and TRS4 utilize GroupKFold (k=5) to ensure that data groups per patient remain intact and unmixed with data from other patients.

3. Results

Table 1. Cross-validation results across experiments.

Model	CS	ROC	PRC	F1	MSE	MAE
TRS1	**0.26**	**0.77**	**0.84**	**0.79**	3.53	1.36
TRS2	0.24	0.68	0.72	0.64	3.46	1.11
TRS3	**0.26**	0.75	0.81	0.74	3.09	**1.02**
TRS4	0.15	0.73	0.80	0.72	**3.07**	1.03

Table 1 provides an overview of the cross-validation results obtained across experiments, using the following metrics: CS (Challenge Score), ROC (Area under the Receiver Operating Characteristic curve), PRC (Area under the Precision-Recall curve), F1 (F1-Score), MSE (Mean Squared Error), and MAE (Mean Absolute Error).

Table 2. Results from the public leaderboard for TRS1.

Model	12H	24H	48H	72H
training	0.41	0.44	0.43	0.43
validation	0.39	0.40	0.40	0.42
test	0.37	0.38	0.40	0.37

Table 2 provides a comparison of our team's challenge scores using TRS1 when evaluated with the training set, validation set, and test set across various available EEG recording durations for evaluation.

4. Discussion and Conclusions

In Table 1, TRS1 outperforms the other experiments on CS, ROC, PRC, and F1. However, the difference among them is relatively small, suggesting potential for further improvements with the other three implementations.

In Table 2, when evaluated with the validation set, the model displayed an encouraging upward trend as the evaluation included more recording hours. However, when evaluated with the test set, the scores exhibited an upward trend up to the 48-hour mark, followed by a slight decline at 72 hours. This suggests the need for ongoing improvements to achieve better results at the crucial 72-hour point, which was the primary focus of the challenge.

Notably, despite the marginal decrease in scores at different available time points, it's significant to observe that the scores remained consistent when assessed with 48 hours of available recordings per patient for both validation and test sets. Another noteworthy point is that our team achieved the second-best score when evaluated with 12 hours of available recordings, with only one other team surpassing us with a score of 0.38. This underscores the exceptional performance of our model when dealing with fewer available recordings, a vital attribute in the clinical context where missing data is a common challenge.

Another noteworthy aspect is the consistent performance of our model across different timeframes, resulting in challenge scores that closely resemble each other. This consistency underscores our model's ability to generalize effectively, performing at a high level even when confronted with earlier time points, despite being primarily trained on the most recent hour of data. Interestingly, our experiments indicate that using a single representative hour of EEG recording from each patient outperforms models that use more extensive hours of recordings and this may be due to significant missing data, especially in the initial hours, as shown in Figure 3.

It is important to note that deep learning models, including Transformers, often excel with larger datasets. Given that this study used data from only 607 patients for training, there is a high likelihood of significantly improved model performance with larger training datasets. Nevertheless, the results obtained from this study already show promising clinical potential from using Transformers for outcome prediction of comatose cardiac arrest patients using EEG and clinical data.

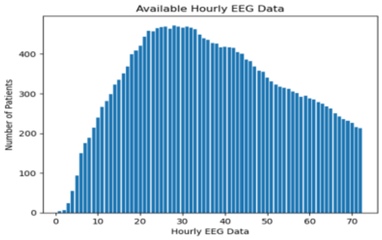

Figure 3. Hourly available EEG data

Acknowledgments

We would like to acknowledge Taiwan's NSTC, MOHW, and MOE's funding for our projects under MOST 110-2221-E-002-112-MY3, MOHW112-TDU-B-221-124003, and NTU-112L900701, respectively.

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, 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.

[4] 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 Jul 14;85(2):137-43. doi: 10.1212/WNL.0000000000001742. Epub 2015 Jun 12. PMID: 26070341; PMCID: PMC4515041.

[5] Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, van den Bergh WM, Foudraine NA, Kornips FHM, Drost G, Scholten E, Bosch FH, Beishuizen A, van Putten MJAM, Hofmeijer J. Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study. Ann Neurol. 2019 Aug;86(2):203-214. doi: 10.1002/ana.25518. Epub 2019 Jun 24. PMID: 31155751; PMCID: PMC6771891.

[6] Sondag L, Ruijter BJ, Tjepkema-Cloostermans MC, Beishuizen A, Bosch FH, van Til JA, van Putten MJAM, Hofmeijer J. Early EEG for outcome prediction of postanoxic coma: prospective cohort study with cost-minimization analysis. Crit Care. 2017 May 15;21(1):111.

[7] 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 MJAM, Westover MB. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation. 2021. Dec;169:86-94.

[8] Zheng WL, Amorim E, Jing J, Wu O, Ghassemi M, Lee JW, Sivaraju A, Pang T, Herman ST, Gaspard N, Ruijter BJ, Tjepkema-Cloostermans MC, Hofmeijer J, van Putten MJAM, Westover MB. Predicting Neurological Outcome From Electroencephalogram Dynamics in Comatose Patients After Cardiac Arrest With Deep Learning. IEEE Trans Biomed Eng. 2022 May;69(5):1813-1825.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is All You Need." In Advances in Neural Information Processing Systems (NeurIPS) (pp. 30-38).

[10] Minghao Liu, Shengqi Ren, Siyuan Ma, Jiahui Jiao, Yizhou Chen, Zhiguang Wang, and Wei Song. Gated transformer networks for multivariate time series classification. arXiv preprint arXiv:2103.14438, 2021.

Address for correspondence:

Che Lin
No. 1, Sec. 4, Roosevelt Rd, Da'an Dist., Taipei, Taiwan - 10617
chelin@ntu.edu.tw