Transformer Network with Time Prior for Predicting Clinical Outcome from EEG of Cardiac Arrest Patients

Maurice Rohr[1], Tobias Schilke[1], Laurent Willems[2], Christoph Reich[1], Sebastian Dill[1], Gökhan Güney[1], Christoph Hoog Antink[1]

[1] KIS*MED - AI Systems in Medicine, Technische Universität Darmstadt, Darmstadt, Germany
[2] Neurology and Epilepsy Center Frankfurt Rhine-Main, Goethe-University Frankfurt am Main, Frankfurt am Main, Germany

Abstract

The prognosis of neurologic outcome for patients after cardiopulmonary resuscitation is usually made based on morphologic patterns in the electroencephalogram (EEG). However, the specificity of these patterns is comparably low and the dependency on the investigator is high. Hence, features that are instead learned by neural networks are a promising tool to overcome both problems. As part of the George B. Moody PhysioNet Challenge, we propose the convolutional embedding transformer (CET) that learns input features with a feature encoder and applies a time prior in the cost function to enhance the training process and enable the model to learn more generalized internal feature representations. The input of CET is the raw, downsampled data. For comparison we implement a support vector machine (SVM) model that uses the features based on morphologic patterns. In a 5 fold cross-validation on the training data, CET achieves a mean challenge score of 0.41 and AUPRC of 0.84. The model did not achieve a score on the test set during the official phase (team "IWill-Survive").

1. Introduction

Prognostication in patients with hypoxic encephalopathy (HE) after cardiopulmonary resuscitation (CPR) is a challenging aspect of modern neurocritical care. Apart from clinical and laboratory diagnostics, electroencephalography (EEG) is of particular diagnostic importance due to its continuous monitoring capability. Several morphologic patterns have been shown for poor functional outcome after CPR, *e.g.* burst suppression patterns with identical bursts or isoelectric EEG [1, 2]. In addition to the low specificity of these visually detected EEG patterns, poor inter-rater reliability in the detection of these patterns is often a problem [3]. Computationally learned features based on repetitive or continuous EEG may have the po-

tential to significantly improve the prognosis of functional outcome in HE patients. As part of the George B. Moody PhysioNet Challenge 2023 [4], we present a model based on the transformer architecture for predicting the outcome from EEG at arbitrary time points after cardiac arrest. The model incorporates a time prior, which can be described as enforcing a monotonous increase in the model's belief about the true outcome. The prior is expected to improve model generalization and is initially implemented as a reward term in the cost function. For reference, we implement a Support Vector Machine (SVM) classifier that uses aforementioned features based on morphologic patterns. The classification task is challenging, because the origin of the EEG signal and thus the placement of the electrodes have to be considered in the model. For example, the EEG of the frontal lobe usually contains frequency components associated with eye movement, but if the same patterns show up in the parietal lobe they can be malignant. Furthermore analyzing multiple days of multi-channel EEG recording is difficult for learning due to the high amount of data and sparsity in feature space. Most segments and especially early hours of the EEG recordings will not contain useful information for diagnosis. In the early hours, patients might be still under influence of sedative medication that increases/decreases alpha waves. We expect these effects to be properly tackled by the time prior, while still being able to incorporate less informative data into the training procedure of the model.

2. Methods

The basis of our analysis are versions 1.0 and 2.0 of the training subset of the International Cardiac Arrest REsearch consortium (I-CARE) dataset [4, 5], which consists of up to 72 hours of EEG-recordings of 607 adult patients after cardiac arrest who had return of heart function but stayed comatose. I-CARE 2.0 contains the full 19-channel unipolar recordings whereas 1.0 provides only the best 5

minutes of EEG per hour as 18 channels in bipolar configuration and a signal quality metric. The data is labeled with Cerebral Performance Category (CPC)-scores which indicate the neurological outcome for each patient with 1 (good) to 5 (dead).

The high dimensionality of the data stemming from the number of channels and more importantly from the length of the recordings is a problem for all existing machine learning methods. In traditional machine learning this is usually approached by using extensive preprocessing and hand-crafted features that try to cover the most important aspects of the signal. This approach is implemented as a baseline with a SVM. In Deep Learning multiple approaches exist. Our novel solution is a combination of incorporating prior knowledge and down sampling.

2.1. SVM

We employ an SVM classifier with a radial basis function kernel and and the regularization parameter C=0.3 as a baseline model. The input of the SVM are 8 *features* extracted from 5 minute epochs of the EEG recording. in addition to the five quantitative features proposed in [6] , namely power, shannon entropy, alpha-to-delta ratio, regularity and coherence in the delta band, we use the quality measure specified in I-Care 1.0 and the corresponding hour of recording after return of spontaneous circulation (ROSC), as well as a burst-suppression feature. A burst-suppression background with or without discharges is regarded as highly malignant [7]. This is best described by a suppression below 10 μV more than 50% of the time. Hence we test if 1 s segments stay below this threshold and then compute the ratio of suppressed segments to non-suppressed segments over all channels as a measure. Before feature computation, the EEG signals are filtered with a zero-phase bandpass filter (0.5 to 30 Hz). Then, regularity, power, Shannon Entropy and alpha-delta ratio are computed on segments of 10 seconds per channel and then averaged over the segments and the channels. Delta coherence is computed over pairwise channels and then averaged. The final outcome prediction is computed by averaging the SVM predictions for all epochs of a patient.

2.2. Transformer with time prior

In order to incorporate the time prior, a model with temporal encoding and/or recursive processing is necessary. Therefore, the Convolutional Embedding Transformer (CET) in Figure 1 consists of a residual network (ResNet) to create a feature embedding that is concatenated to a sinusoidal positional encoding. The output serves as the input to a causal-masked transformer encoder layer with M=4 sequential layers. The input of the model is batches (B) of the bandpass filtered (0.5 to 16 Hz)
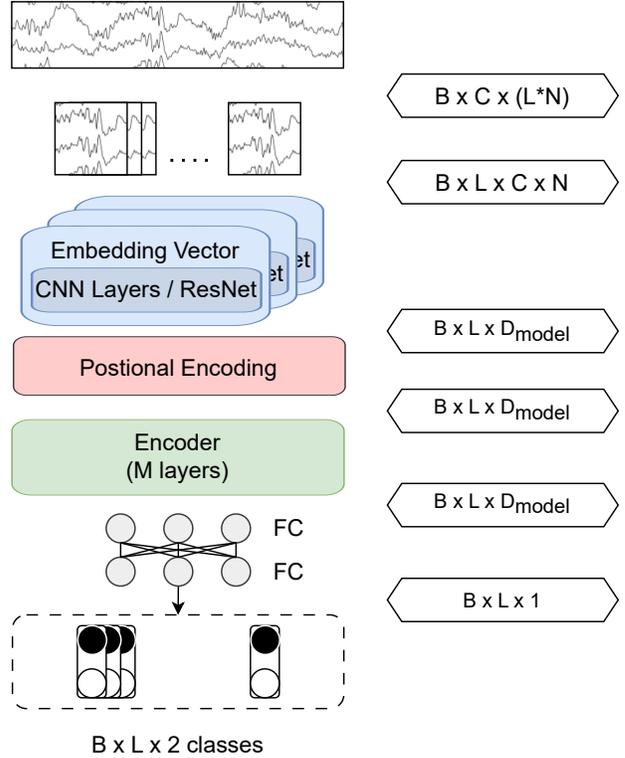

Figure 1. Convolutional Embedding Transformer (CET) architecture which learns local ResNet Embeddings. The final decision is generated in a fully connected block with shared weights for all L encoder sequence embeddings

and resampled (f_s =32 Hz) EEG signals, that are split in L=2160 segments with duration N=320 samples (C=18 channels). Further parameters are D_{model}=256, D_{deep}=64 and D_{binary}=1. The 6-layer ResNet consists of two 1D-convolutions of kernels sized 5 and padding 2 interleaved with group-normalization and LeakyReLU activations.

We start with the Binary Cross Entropy loss that computes the loss after the last segment of a sequence with maximum length L (e.g. 72 *hours*, that is zero-padded to achieve the same length for all samples in the dataset).

$$\mathcal{L}_{\text{BCE}}(y_{n,L}, \hat{y}_n) = -\hat{y}_n \log(y_{n,L}) \\ + (1 - \hat{y}_n) \log(1 - y_{n,L}), \quad (1)$$

where $y_{n,i}$ is the predicted probability for a negative outcome (CPC-Score $\in \{3, 4, 5\}$) for the i-th segment of patient n and $\hat{y}_n$ the respective binary ground truth label. The loss can be easily generalized to more classes. One problem is that we cannot be sure which time segments of the signal contain good features to decide the outcome, but the more consecutive time segments the model inspects, the higher the certainty of the model about the outcome

should get. By introducing a time prior that rewards an increase in the models belief in the (correct) classifications as it receives a continuous input of hourly data, we aim to incorporate knowledge about the outcome while allowing the model to be less certain for smaller examples. Phrased differently, that means that penalization will be decreased for earlier predictions if they employ less data. This should work similar to data-augmentation as we split recordings with one label into multiple segments with their own label in a sense and improves gradient flow. Similar approaches with slightly different reasoning have been applied to vision transformers before [8]. We suggest four slightly different approaches to a time prior where we always assume $y_{n,i}$ and $y_{n,i+1}$ to be actual consecutive samples:

1. **Weighted reward**

$$\mathcal{L}_{\text{total},1} = \sum_i \mathcal{L}_{\text{BCE}}(y_{n,i}, \hat{y}_n) + \mathcal{L}_{\text{time}} \qquad (2)$$

$$\mathcal{L}_{\text{time}} = -\sum_i \alpha_i \cdot |y_{n,i} - 0.5|, \qquad (3)$$

where $\alpha_i = i/L$ with i being the absolute position of the latest sample.

2. **Monotonicity** (based on [9])

$$\mathcal{L}_{\text{total},2} = \sum_i \mathcal{L}_{\text{BCE}}(y_{n,i}, \hat{y}_n) + \mathcal{L}_{\text{time}}, \qquad (4)$$

$$\mathcal{L}_{\text{time}} = -(\max(\sum_{i=1}^{L-1} \max(0, y_{i+1} - y_i),$$

$$\sum_{i=1}^{L-1} \max(0, -(y_{i+1} - y_i))), \qquad (5)$$

3. **Exponential weighting**:

$$\mathcal{L}_{\text{total},3} = \sum_i^L \gamma^{L-i} \mathcal{L}_{\text{BCE}}(y_{n,i}, \hat{y}_n), \qquad (6)$$

4. **Sigmoidal weighting**

$$\mathcal{L}_{\text{total, 4}} = \frac{1}{L} \sum_i^L \phi_\gamma(i) \mathcal{L}_{\text{BCE}}(y_{n,i}, \hat{y}_n), \qquad (7)$$

$$\phi_\gamma(i) = \frac{1}{1 + e^{-\gamma i}}, \qquad (8)$$

where in both cases γ is a tune-able parameter. In the evaluation $\mathcal{L}_{\text{total, 4}}$ is used and γ is chosen such that 95% certainty is assigned to the prediction 36 hours after ROSC. The loss is averaged over a batch of samples.

The baseline CET model $\text{CET}_{\text{class}}$ is aggregated using a global average pooling followed by a two-layered fully connected classification head and the loss in equation 1. In the time prior approach CET_{time}, the encoder output is processed column-wise by the same classification head and evaluated using $\mathcal{L}_{\text{total, 4}}$.
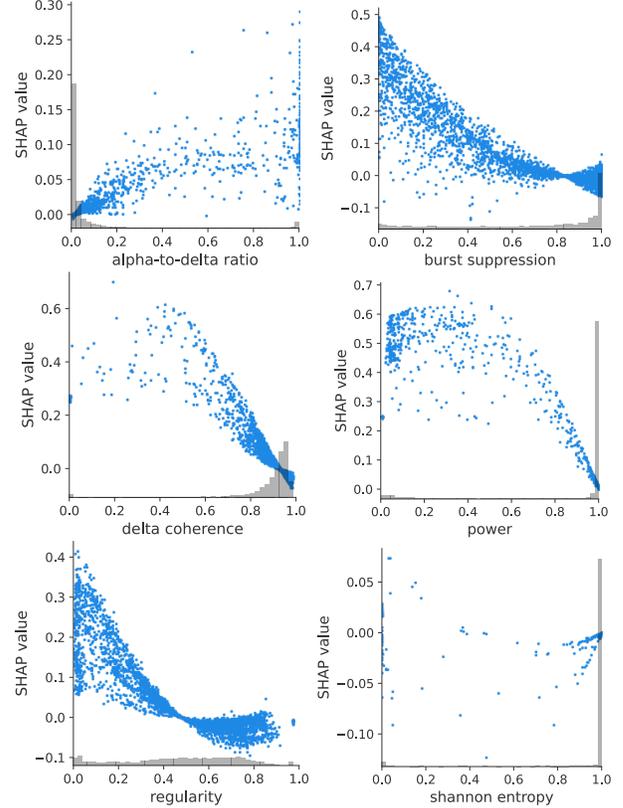

Figure 2. SHAP values (blue) and their distribution (grey) for all features considered in the SVM based on I-Care 1.0. Besides the signal-based features we also included meta-information such as the hour of recording after ROSC.

3. Evaluation and Results

We inspect the SHAP feature importance values for the morphologic features in Figure 2. Then, in two experiments we evaluate the baseline SVM using 10 fold cross-validation on I-Care 1.0 and 2.0 first with the full feature set and second with a reduced feature set (exclusion of shannon energy, A-D ratio and burst suppression based on the SHAP values).

As can be seen in Figure 3,the evaluation on I-Care 1.0, consistently outperforms the evaluation on I-Care 2.0. The reduced feature set did not change the metrics significantly.

Figure 4 shows in a stratified group 5-fold cross-validation on I-Care 1.0, that CET_{time} achieves similar F1 and AUPRC metrics compared to $\text{CET}_{\text{class}}$, but scores lower on the challenge score with high variability between runs.

4. Discussion

Given the high quality data prepared in I-Care 1.0, the feature-based SVM performed well enough for a baseline,
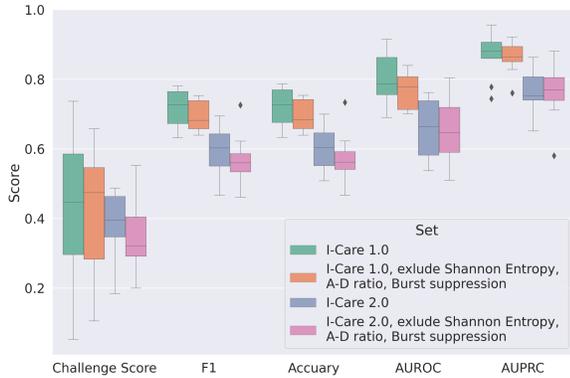
Figure 3. Evaluation Scores for Cross-Validation on I-Care 1.0 and I-Care 2.0 datasets. The model performs significantly better on the (cleaner) I-Care 1.0
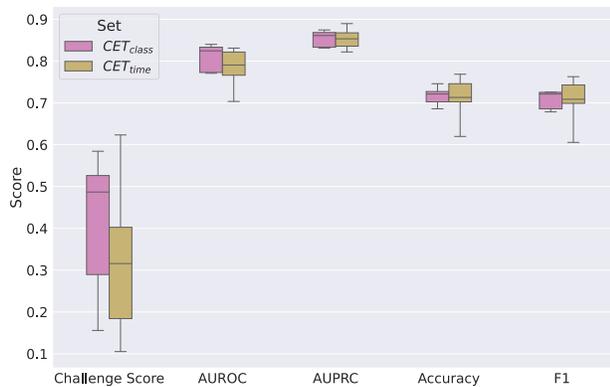

Figure 4. Comparison of evaluation scores between CET_{class} and CET_{time} using a 5-fold cross-validation on I-Care 1.0.

but far from the best ranked validation scores (0.82). Interestingly as can be seen from Figure 3, the reduced feature set reduces the variance of metrics in the cross-validation while slightly improving AUPRC on I-Care 2.0. The feature importance values (Figure 2) provide a deeper understanding. For example, a high burst suppression value does not influence the output, although most recordings show a high value. Similarly, there is a negative correlation between signal power and positive outcome. Interestingly, a low as well as a high signal regularity influence the outcome positively while values in between are mostly associated with a negative outcome. As expected ROSC and signal quality show no influence on the outcome. Since most features show insufficient performance based on the challenge score especially on the longer data, a deep learning based approach seems more promising.

The CET model is far from optimized in multiple regards, such as architecture and tunable hyper-parameters, yet performs already comparable to the SVM. Based on Figure 4, the effect of the time prior remains inconclu-

sive and average pooling appears more robust to changes in the dataset compared to the architectural changes implemented to facilitate use of the time prior. And thus leading to lower fluctuations in prediction performance. However, in future work we plan to compare the proposed time priors more thoroughly and adapt the transformer architecture to stabilize the training.

References

[1] Hirsch LJ, Fong MWK, Leitinger M, LaRoche SM, Beniczky S, Abend NS, et al. American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2021 Version. Clin Neurophysiol 2021;38(1):1–29.

[2] Willems LM, Rosenow F, Knake S, Beuchat I, Siebenbrodt K, Strüber M, Schieffer B, Karatolios K, Strzelczyk A. Repetitive Electroencephalography as Biomarker for the Prediction of Survival in Patients with Post-Hypoxic Encephalopathy. J Clin Med 2022;11(21).

[3] Muhlhofer W, Szaflarski JP. Prognostic Value of EEG in Patients after Cardiac Arrest-An Updated Review. Curr Neurol Neurosci Rep 2018;18(4):16.

[4] Reyna MA, Amorim E, Sameni R, Weigle J, Elola A, Rad AB, Seyedi S, Kwon H, Zheng WL, Ghassemi MM, et al. Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023. In Computing in Cardiology, volume 50. 2023; 1–4.

[5] Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, Karukonda V, et al. The International Cardiac Arrest Research Consortium Electroencephalography Database. Critical Care Medicine 2023;.

[6] Tjepkema-Cloostermans MC, van Meulen FB, Meinsma G, van Putten MJAM. A Cerebral Recovery Index (CRI) for Early Prognosis in Patients after Cardiac Arrest. Crit Care 2013;17(5):R252.

[7] Westhall E, Rossetti AO, van Rootselaar AF, Wesenberg Kjaer T, Horn J, Ullén S, et al. Standardized EEG Interpretation Accurately Predicts Prognosis After Cardiac Arrest. Neurology 2016;86(16):1482–1490.

[8] Bayramli B, Hur J, Lu H. RAFT-MSF: Self-Supervised Monocular Scene Flow using Recurrent Optimizer. IJCV 2023;1–13.

[9] Gupta A, Shukla N, Marla L, Kolbeinsson A, Yellepeddi K. How to Incorporate Monotonicity in Deep Networks While Preserving Flexibility? In NeurIPS 2019 Workshop on Machine Learning with Guarantees. 2019; .

Address for correspondence:

Maurice Rohr
Merckstraße 25, 64283 Darmstadt, Germany
rohr@kismed.tu-darmstadt.de