Leveraging Unlabeled Electroencephalographic Data to Predict Neurological Recovery for Comatose Patients Following Cardiac Arrest

Isaac Sears[1], Augusto Garcia-Agundez[2], George Zerveas[2], William Rudman[2], Laura Mercurio[3], Corey E. Ventetuolo[4,5], Adeel Abbasi[4], Carsten Eickhoff[6,7]

[1]Warren Alpert Medical School at Brown University, Providence, RI
[2]Department of Computer Science, Brown University, Providence, RI
[3]Department of Emergency Medicine, Section of Pediatric Emergency Medicine, Warren Alpert Medical School at Brown University, Providence, RI
[4]Department of Medicine, Warren Alpert Medical School at Brown University, Providence, RI
[5]Department of Health Services, Policy & Practice, Brown School of Public Health, Providence, Rhode Island, USA
[6]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany
[7]Faculty of Medicine, University of Tübingen, Tübingen, Germany

Abstract

In response to the 2023 George B. Moody PhysioNet Challenge, we propose an automated, unsupervised pre-training approach to boost the performance of models that predict neurologic outcomes after cardiac arrest. Our team, (BrownBAI), developed a model architecture consisting of three parts: a pre-processor to convert raw electroencephalograms (EEGs) into two-dimensional spectrograms, a three-layer convolutional neural network (CNN) encoder for unsupervised pre-training, and a time series transformer (TST) model. We trained the CNN encoder on unlabeled five-minute EEG samples from the Temple University EEG Corpus (TUEG), which included more than 20x the patients available in the PhysioNet competition training dataset. We then incorporated the pre-trained encoder into the TST as a base layer and trained the composite model as a classifier on EEGs from the 2023 PhysioNet Challenge dataset. Our team was not able to submit an official competition entry and was therefore not scored on the test set. However, in a side-by-side comparison on the competition training dataset, our model performed better with a pretrained (competition score 0.351), rather than randomly initialized (competition score 0.211) CNN encoder layer. These results show the potential benefits of leveraging unlabeled data to boost task-specific performance of predictive EEG models.

1. Introduction

The 2023 George B. Moody PhysioNet Challenge (1) invited participants to predict neurological recovery in comatose patients following cardiac arrest through automated analysis of EEG, ECG, and summary clinical data. As noted in the challenge description, EEG data, which reflects spontaneous electrical activity in the brain, has long been used for neurological prognostication, but requires time-intensive and ideally real-time analysis by specialized medical professionals (1). Computational approaches have the potential to both automate and enhance this analysis.

In response to this challenge, our team focused on leveraging additional unlabeled EEG data not available within the official competition dataset to allow us to take advantage of more sophisticated machine learning models. Clinical dataset curation is a resource and time-intensive endeavor, evidenced by the PhysioNet Challenge dataset development process itself, which spanned several years and necessitated the support of large research grants. The result is a dataset that is undoubtedly the largest and highest quality for the specific problem of predicting neurological recovery after cardiac arrest (2, 3). And yet, the 1,020 included patients fall short of the millions of independent samples often required for modern deep learning models in mainstream machine learning fields such as computer vision (4). In these fields, a technique known as transfer learning is commonly used to take advantage of widely available unlabeled data to build base models which can then be further trained (i.e. "fine-tuned") on more specific prediction tasks with smaller amounts of labeled data (5).

While large unlabeled clinical data are not as readily available as the public-domain internet data used by these other models, there is a growing number of large, general use clinical datasets available for research. In this study,

Computing in Cardiology 2023; Vol 50 Page 1 ISSN: 2325-887X DOI: 10.22489/CinC.2023.308

we used the Temple University EEG Corpus (TUEG) (6), with over twenty times the number of patients as the PhysioNet Challenge dataset, to pretrain a base EEG model before fine-tuning on the competition task of post cardiac arrest neurological prognostication. We hypothesized that models of post cardiac arrest prognostication can benefit from unsupervised pretraining on larger unlabeled datasets before being trained on the goal classification task.

2. Methods

2.1 EEG Signal Pre-Processing Pipeline

In attempting to pretrain our model on one dataset (TUEG) and fine-tune on another (the PhysioNet Challenge dataset), it was essential to build a data processing pipeline that reduced inter-dataset variability as much as possible. Our data processing pipeline also served to compress the raw EEG signal, which consisted of 30,000 data points per five-minute recording, into a smaller, more information-dense representation. These data processing steps were applied to both the unlabeled TUEG pretraining dataset and the labeled PhysioNet dataset.

First, we standardized each channel of the raw EEG data. To account for outliers in the signal due to artifacts, the scaling process centered the data for each channel on the channel's median value and scaled it according to the interquartile range as implemented in scikit-learn's *robust_scale* function. Next, a zero-phase low-pass finite impulse response filter was applied using the Scipy *resample_poly* function to remove high-frequency noise from the EEG signal while introducing minimal signal distortion. EEGs in the TUEG dataset included studies sampled at alternate frequencies than the standard 100 Hz sampling found in the challenge dataset. The *resample_poly* function was also used to resample the signal up or down to 100 Hz. Recordings in the TUEG dataset were also truncated to five minutes to match the challenge dataset. To avoid padding short recordings only TUEG patients with recordings greater than five minutes were selected for inclusion. To avoid including potentially noisy segments at the beginning and end of the recordings, only the middle five minutes of these arbitrarily long recordings were included in the final pretraining dataset. As a final step, the 18 unipolar TUEG channels were re-referenced to match the PhysioNet bipolar channels using the MNE library's *set_bipolar_reference* function.

2.2 Spectrogram Generation

Following pre-processing, the recordings were converted to spectrograms: two-dimensional matrix representations of the EEG signal with time plotted against frequency (**Figure 1**). Spectrograms were generated by taking consecutive Fourier transformations of overlapping windows of the EEG signal such that the value of the spectrogram at a given frequency *f* and time *t* represents the intensity of that frequency in a window of the EEG signal centered around that time. The default values of the SciPy *signal.spectrogram* method were used for the size of windows used to generate the spectrograms as well as for the window overlap. Spectrograms were generated for each channel in each EEG recording. Spectrogram frequencies were clipped to between 0.5 and 30 Hz, a band of frequencies that includes all four EEG wave classes (delta, theta, alpha, and beta). The result was uniform-shaped multi-channel 18 x 75 x 133 spectrograms (EEG channels x frequency x time).

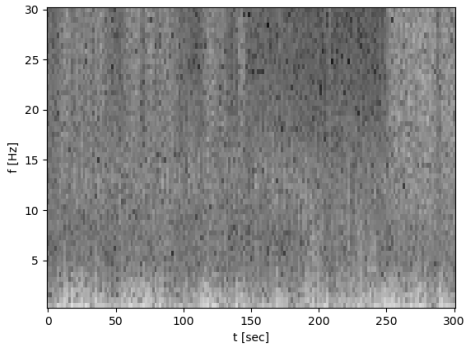

Figure 1: An example spectrogram representing a single channel of EEG data.

2.3 Model Details

Our model architecture consisted of two parts: (1) a three-layer convolutional neural network (CNN) encoder that took 18 x 75 x 133 spectrograms as input and returned 128 x 364 representations of the original data and (2) a time series transformer (TST) that operated on the spectrogram representations generated by the encoder to estimate the probability of neurological recovery. The CNN was chosen for the first part of this model architecture because it is designed to handle three-dimensional (number of channels x height x width) image data like the EEG spectrograms generated by our data processing pipeline (7). As an encoder, the CNN was also necessary to further reduce the size of the input EEG data before it was passed to the TST. For a given input vector, the TST's memory requirements are more substantial than the CNN. Encoding the input spectrograms into a more information-dense representation allowed us to test versions of the TST with more memory requirements than we otherwise would have been able to if the data was passed to the TST unencoded.

The second layer of our model architecture, the TST, is a modern transformer-based deep learning model designed specifically for multivariate time series classification tasks, such as the 2023 PhysioNet challenge (8). The TST was used because it has been shown to achieve state-of-the art performance on a variety of non-

medical timeseries classification tasks and has even been applied successfully to EEG classification tasks (9). The TST is configurable with nine hyperparameters, as shown in **Table 1**. We used the suggested hyperparameters presented in the original paper that were shown to be generally high performing on the benchmark tasks on which the TST was originally tested, with one exception: the batch size was reduced from 128 to 32.

Parameter Name	Value
Activation	Gaussian Error Linear Unit
Dropout	0.1
Learning Rate	0.001
Positional Encoding	Learnable
Model Dimension	128
Num Attention Heads	16
Num Encoder Blocks	3
Batch Size	32

Table 1. The hyperparameters used for our implementation of the TST.

Two different versions of this two-part model architecture were tested: (1) a pretrained version, where the CNN encoder was first trained using the larger TUEG dataset to generate spectrogram representations with minimal information loss before being combined with the TST classification module and (2) an un-pretrained version where the weights of the CNN encoder were randomly initialized. In both versions, the weights of the CNN layer were fine-tuned, rather than frozen.

To pretrain the CNN encoder, its architecture was mirrored to create an autoencoder: a type of deep learning model capable of learning to efficiently represent input data at lower dimensions. The first part of the autoencoder, the encoder, maps the input down to a lower-dimensional vector. The second part of the autoencoder, the decoder, attempts to map this compressed representation back to the original inputs, reconstructing them with minimal error. This task is non-trivial because of the information bottleneck induced by the lower-dimensional intermediate representation between the encoder and the decoder. Throughout the training process, weights in the encoder and decoder are jointly adjusted such that minimal information loss occurs as the encoder maps its inputs to lower-dimension representations. Thus, the autoencoder learns information-efficient transformations that reduce spectrogram data to approximately a quarter of its original size with minimal information loss.

The CNN autoencoder was trained on batches of 32 examples from the TUEG dataset, using with an Adam optimizer set to a learning rate of 1e-4 and the mean squared error (over all pixels of a sample's spectrograms) as the loss function. Training was set to end after 100 epochs or until the average mean squared error failed to decrease for 10 consecutive epochs. The learnable model parameters from the epoch with the lowest mean squared error loss were saved and used in the final encoder that would be incorporated with the rest of the model for the competition classification task.

Both when employing a pretrained encoder and when not, classification models were trained and evaluated using the same k-fold cross-validation pipeline to enable side-by-side comparison: the PhysioNet Challenge dataset was split into five equal subsets. For each split, the one-fifth subset was held out as a test set and the other four-fifths were used as the training set. The result were five scores for each model. The averages and standard deviations of these scores are reported in the results section. The specific metrics used were (1) the PhysioNet competition score, which was the maximum true positive rate at a threshold false positive rate of five percent, and (2) the area under the receiver operating characteristic curve (AUROC).

3. Results

After applying inclusion criteria, 14,927 of the 14,983 subjects (99.6 %) who received EEGs at Temple University hospitals between 2002 and 2017 were used for pretraining. This dataset represented a 24.6-fold increase in the number of independent training examples over the PhysioNet Challenge dataset, which included 607 cardiac arrest patients. Although our team did not submit an official-phase entry for the 2023 PhysioNet Challenge, and therefore has no scores to report for the competition's hidden test dataset, we present cross-validation results for both our pre-trained and un-pretrained models on the training dataset. The pretrained model outperformed the un-pretrained model, with an average cross-validation competition score of 0.351 (standard deviation 0.058) and AUROC 0.77 (standard deviation 0.042) vs. un-pretrained competition score 0.211 (standard deviation 0.076) and AUROC 0.653 (standard deviation 0.033).

4. Discussion and Conclusion

In this study we present a novel modeling approach capable of learning from unlabeled data to improve performance on the task of predicting neurological recovery after cardiac arrest from EEG data. Our findings suggest that pre-training should be attempted whenever possible to help boost performance of classification tasks. By directly comparing a pretrained and un-pretrained version of the model, we show that there is benefit to pretraining for this task and model architecture. In theory, the benefit of pretraining on the unlabeled TUEG dataset is due to the fact that the CNN autoencoder was able to learn compressed representations of five-minute EEG segments that were generalizable across datasets. By leveraging these learned compressed representations in the

fine-tuning phase, the dimensionality of the classification task was effectively reduced for the pretrained model, and fewer examples were therefore required to achieve a certain level of performance.

To our knowledge, this study is the first to apply the concept of EEG transfer learning to boost the performance of a model that could potentially aid in a specific clinically relevant task, such as prognostication of neurological recovery after cardiac arrest. Most other studies that have applied the concept of transfer learning to EEG modeling have done so in the context of improving engineering tasks related to Brain Computer Interfaces (BCI). Examples include (10), a study that applied the concept of transfer learning to better characterize mental workload from EEGs, and (11), a study that pretrained on unlabeled EEG data to improve performance on a series of standard BCI benchmarking tasks. Although these studies do not focus on clinically relevant tasks, their methods parallel our study and we arrive at similar results: for the majority of tasks, regardless of domain, pre-training on unlabeled data improves the performance of task-specific EEG models.

Our study also has certain limitations. First, it should be noted that we used only EEG data, and not other demographic (e.g., gender, age, etc.) or clinical (e.g., targeted temperature management, time since cardiac arrest) data that had been provided. However, our best-performing EEG-only models were outperformed by other teams' un-pretrained models that used non-EEG data. Based on this comparison, it is likely that pretraining with more EEG data, although beneficial, is not as beneficial as training with more variables that are relevant to the prediction task. For this study, we chose to use only EEG data to keep the study focused on relative performance improvements from pretraining, rather than obtaining a maximum competition score. Second, our method of pretraining, which leveraged a CNN autoencoder, is only one relatively simple method among many other possible pretraining methods. Although implementation will be more difficult, the more sophisticated methods presented in (10) and (11) may potentially lead to further performance improvements. Lastly, the scope of this study did not include any analysis of how the CNN autoencoder was able to create useful features from EEG data. Do the compressed representations correspond to summary information of the alpha, beta, delta, and theta waves that are used in conventional EEG analysis, or something else? Analysis of these compressed representations with modern interpretability algorithms is another potentially interesting direction for future projects.

In summary, leveraging unlabeled EEG data has been shown to improve task-specific modeling performance in a variety of non-clinical domains. Our study applies this technique to a specific clinical problem: predicting neurological outcomes in comatose patients after cardiac arrest. Many clinical problems lack specifically labeled datasets that are large enough to fully employ modern deep learning modeling techniques. Transfer learning methods that leverage large unlabeled datasets, such as the model we present in this study, may serve as a bridge that allows these more specific clinical problems to benefit from analysis by modern deep learning algorithms.

References

1. Reyna M, Amorim E, Sameni R, Weigle J, Elola A, Bahrami Rad A, al. e. Predicting neurological recovery from coma after cardiac arrest: The George B. Moody PhysioNet Challenge 2023. Computing in Cardiology. 2023;50:1-4.
2. AL G, LA A, L G, JM H, PC I, RG M, al. e. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2000;101(23):e215-e20.
3. Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, Karukonda V, et al. The International Cardiac Arrest Research Consortium Electroencephalography Database. Crit Care Med. 2023.
4. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. 2015;115(3):211-52.
5. Iman M, Arabnia HR, Rasheed K. A Review of Deep Transfer Learning and Recent Advancements. Technologies. 2023;11(2):40.
6. Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. Front Neurosci. 2016;10:196.
7. Dubreuil-Vall L, Ruffini G, Camprodon JA. Deep Learning Convolutional Neural Networks Discriminate Adult ADHD From Healthy Individuals on the Basis of Event-Related Spectral EEG. Front Neurosci. 2020;14:251.
8. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A Transformer-based Framework for Multivariate Time Series Representation Learning. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining; Virtual Event, Singapore: Association for Computing Machinery; 2021. p. 2114–24.
9. Potter İY, Zerveas G, Eickhoff C, Duncan D, editors. Unsupervised Multivariate Time-Series Transformers for Seizure Identification on EEG. 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA); 2022 12-14 Dec. 2022.
10. Yin Z, Zhang J. Cross-session classification of mental workload levels using EEG and an adaptive deep learning model. Biomedical Signal Processing and Control. 2017;33:30-47.
11. Kostas D, Aroca-Ouellette S, Rudzicz F. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. Front Hum Neurosci. 2021;15:653659.

Address for correspondence:
Isaac Sears
222 Richmond St., Box G-9558, Providence RI, 02903
isaac_sears@brown.edu