Variational Autoencoders for Electroencephalogram Feature Extraction in Patients with Coma after Cardiac Arrest

Adel Hassan[1], Liam Ferreira[1]

[1]Baylor College of Medicine, Houston, United States of America

Abstract

Many survivors of cardiac arrest subsequently end up in a coma state, and these patients will go onto achieve varying levels of neurological recovery, ranging from brain death to full recovery. Electroencephalogram (EEG) analysis can be used to predict the neurological outcome of a cardiac arrest patient, but the patterns are complex and human analysis is time-consuming.

We trained a variational autoencoder to extract features from EEGs and used those features in a random forest classifier to predict neurological outcome after cardiac arrest. The resulting model was able to differentiate between good neurological outcome, defined as Cerebral Performance Category (CPC) of 1 or 2, versus poor neurological outcome, defined as CPC of 3-5. The final model had a true positive rate of 0.257 and a false positive rate of 0.05. These results demonstrate that it is possible to use variational autoencoders to extract EEG features that are useful for downstream tasks. This opens the door to more interpretable models for EEG analysis in the future.

This article was part of 'Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023'.

1. Introduction

Cardiac arrest is a condition with high morbidity and mortality, with the majority of survivors experiencing some degree of permanent brain damage [1]. Accurate prognoses can help families make informed decisions, but predicting the neurological outcome of any given patient can be challenging. Electroencephalograms (EEGs) record the activity at the cortex of the brain using scalp electrodes [2], and they have been used to make more informed predictions about the neurological prognosis of patients status-post cardiac arrest [3].

However, human analysis of EEGs is labor-intensive, time-consuming, and susceptible to errors. Therefore, machine learning can improve both the time to get a prognosis and the accuracy of that prognosis. For example, Zheng et al used a convolutional neural network with a long-short term memory to predict CPC after cardiac arrest with are under the receiver operating characteristics curve (AUC) of 0.91 [4]. However, existing machine learning methods lack interpretability, which hinders adoption in the clinical environment.

One novel architecture that has been proposed for interpretable feature extraction is the variational autoencoder (VAE). Autoencoders compress multi-dimensional data into a low-dimensional latent space. Unlike traditional autoencoders, VAEs enforce an additional constraint, namely that similar values of the latent variables should result in similar reconstructions. This tends to force models toward encoding disentangled latent variables that are often more easily interpretable [5].

This work was part of 'Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023' for the team “HeartsAndMinds” [6, 7].

2. Methods

2.1. Database

The I-CARE database includes data from 607 patients who were treated for cardiac arrest at 7 hospitals across the U.S. and Europe, as well as clinical variables including age, sex, location of arrest (in-hospital vs out-of-hospital), type of cardiac rhythm at time of resuscitation (shockable vs non-shockable), targeted temperature management, and the CPC outcome [8].

2.2. Algorithm Training

We devised a variational autoencoder (VAE) to encode segments of the EEG strip into a compressed latent space. The input data for the VAE consisted of 1024-frame segments of the EEG, containing 1024 voltage measurements from each of 6 scalp electrodes. Although several recordings had 18 or more electrodes, for our algorithm, we selected the 6 electrodes that were present in all recordings in order to reduce imputation. Namely, these were F3, F4, P3, P4, T3, and T4. The I-CARE (International Cardiac Arrest REsearch consortium) database contains over 32,000 hours of EEG recordings [8], so to reduce the number of variables, we truncated the recordings, considering only the first 5 minutes of each

hour-long recording. These 5 minutes of EEG data were split into 1024-frame-long segments and provided as input for the VAE.

The code for the VAE was based on a GitHub repository by Subramanian [9], and the final code is publicly available at the following URL (https://github.com/firejake308/cinc-2023). The structure of the encoder part of the VAE consisted of three one-dimensional convolution layers, with the first having 128 output channels, 256 for the second, and 512 for the third. Batch normalization was applied to each layer, and leaky rectified linear units (ReLU) were used as the activation function. The output of the convolutional layers was provided in parallel to two linear layers to produce two vectors of 400 variables each per batch, one representing the mean and one representing the standard deviation of each latent variable. A random number is then sampled from the normal distribution described by that mean and standard deviation, producing the value of the final latent variable. These values were then fed through a decoder to try to restore the original signal, where the structure of the decoder was a single linear layer, followed by three one-dimensional transposed convolutional layers, each with batch normalization and leaky ReLU activation. The final layer of the decoder was a standard one-dimensional convolution layer (no transpose) with tanh activation to allow for output of negative voltages.

Due to initially unstable training, a random shortcut term was added to simplify the prediction task during the initial batches of training, inspired by diffusion models [10]. Namely, at the time of decoding latent variables back to the original voltage signals, a copy of the initial input was fed through the encoder, skipping over the linear layers and the random sampling, and provided to the decoder to assist with reconstructing the initial voltages. However, each batch was assigned a probability of shuffling, and based on that probability, the values of the shortcut path would be randomly shuffled, have a noise added to them, or set to zero. The probabilities for randomizing samples for each batch were sampled from a uniform distribution ranging from zero to one, with no modifications throughout the training process.

The VAE was then trained to compress the input data into a latent space of 400 variables, then recreate the initial EEG recording of 1024 frames and 6 electrodes, corresponding to compression down to 6.5% of the initial size. Due to practical considerations near the deadline, the VAE was only trained for 24 hours rather than training for the full 72 hours permitted by the Moody Challenge.

The latent variables for all 1024-frame segments in the 5-minute period were then averaged together to produce a single, average latent representation of the signal for each recording. These latent variables were then trended over several hours to produce the final EEG features for the classification model. Namely, the mean latents across all hours and the slope and y-intercept of the linear regression of the latent variables over time were all provided to the final classifier.

The ultimate prediction of the CPC was made by a random forest model. The random forest model had access to the EEG features described in the previous paragraph, in addition to the patient's age, sex, time between cardiac arrest and return of spontaneous circulation (ROSC), out-of-hospital cardiac arrest (OHCA) status, rhythm type, and target temperature, if targeted temperature management was used.

2.3. Testing and Validation

The model was trained on data from 607 patients and tested on a separate, hidden test set of patients. The metric reported by the official challenge was the true positive rate (TPR) when holding the false positive rate at 0.05. This TPR was reported for models trained on 12 hours, 24 hours, 48 hours, and 72 hours of EEG data per patient, but our model did not use the additional EEG recordings due to the self-imposed time limit.

3. Results

Table 1. Binary classifier performance metrics.

	Train	Validation	Test
TPR	0.921	0.403	0.257
AUROC	0.989	0.707	0.698
Accuracy	0.942	0.673	0.683

TPR = true positive rate
AUROC = area under the receiver operating characteristics curve

The final challenge score achieved by the model was 0.257 on the test set (Table 1), meaning that the random forest model using VAE latent variables was able to differentiate between CPC $\geq$ 3 versus CPC < 3 with a true positive rate of 0.257 when the decision threshold was set to have a false positive rate of 0.05. On the hidden test set, the model also had an area under the receiver operating characteristics curve (AUROC) of 0.257. Our team was not officially ranked, but our 72-hour.test-set score fell between the teams ranked 32 and 33.

The model had equal performance regardless of whether it was given 12, 24, 48, or 72 hours of EEG data.

4. Discussion

The final challenge score of 0.257 represented performance that was significantly better than chance alone, which would be 0.05. There was likely significant overfitting involved, since the AUROC decreased from 0.989 on the training set to 0.707 on the validation set and 0.698 on the test set. However, it should be noted that the

VAE-based algorithm was unable to improve its performance when given 24, 48, or 72 hours of EEG recordings due to the limit on training time. Given that even this limited VAE learns generalizable features, it is possible that when exposed to more training data, the VAE may learn even more generalizable features to better encapsulate the data, or it may also be necessary to expand the latent space.

In terms of interpretability, the addition of the random noise term made it difficult to interpret the significance of the latent variables. It is worth noting that in previous iterations without the random term, each latent variable represented a prototype recording, and the overall reconstruction was built by a weighted sum of these prototypes. However, after a random term was added to facilitate the training process, the random noise introduced additional confounding variables, and this interpretability was lost. An ideal scenario would likely find a way to facilitate training without the random term in order to preserve interpretability.

Acknowledgments

No conflicts of interest to disclose.

References

[1] Young GB. Neurologic Prognosis after Cardiac Arrest. N Engl J Med 2009; 361(6):605–11.

[2] Constant I, Sabourdin N. The EEG signal: a window on the cortical brain activity: EEG in pediatric anesthesia. Pediatr Anesth 2012; 22(6):539–52.

[3] Admiraal MM, Ramos LA, Delgado Olabarriaga S, et al. Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest. Clin Neurophysiol 2021; 132(9):2240–47.

[4] Zheng W-L, Amorim E, Jing J, et al. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation 2021; 169:86–94.

[5] Higgins I, Matthey L, Pal A, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. International Conference on Learning Representations 2017. 2017.

[6] Reyna MA*, Amorim E*, Sameni S, Weigle J, Elola A, Bahrami Rad A, Seyedi S, Kwon H, Zheng, WL and Ghassemi M, van Putten MJAM, Hofmeijer J, Gaspard N, Sivaraju A, Herman S, Lee JW, Westover MB**, Clifford GD**. Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023. Computing in Cardiology 2023; 50: 1-4.

[7] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation; 101(23): e215-e220.

[8] 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.

[9] Subramanian AK. PyTorch-VAE. GitHub repository 2020. GitHub 2020.

[10] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. 2020. arXiv 2020.

Address for correspondence:

Adel Hassan
1 Baylor Plaza, Houston, TX 77030
Adel.Hassan@bcm.edu