Less is More: Reducing Overfitting in Deep Learning for EEG Classification

Songchi Zhou[1], Shijia Geng[2], Jun Li[3], Deyun Zhang[2], Ziqian Xie[4], Chuandong Cheng[5], Shenda Hong[6,7,*]

[1]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China
[2]HeartVoice Medical Technology, Hefei, China
[3]College of Electronic Science and Engineering, Jilin University, Changchun, China
[4]University of Texas Health Science Center at Houston, TX, US
[5]Department of Neurosurgery, The First Affiliated Hospital of USTC, Hefei, China
[6]National Institute of Health Data Science, Peking University, Beijing, China
[7]Institute of Medical Technology, Health Science Center of Peking University, Beijing, China

Abstract

Although most of the patients' recordings includes large scale long-term physiological time series, the patient-level quantity is relatively small, posing great challenges for machine learning methods. As part of the George B. Moody PhysioNet Challenge 2023, we aim to propose a series of Reducing Overfitting techniques in Deep learning for EEG (coined as RODE) in this scenario, for neurological recovery prognosis. RODE is a simple yet effective machine-learning method, which is mostly powered by generalizable deep-learning features. Specifically, we first pre-train a convolutional neural network on the segment level with a margin-based and mining-based loss, and extract deep features from it. Then we decrease the deep features' complexity using dimensionality reduction methods, which prove to be quite significant for reducing overfitting in deep learning. Finally, we combine the reduced features with static features in patient level, and put them into an ensemble model for classification. Our team, PKU_NIHDS, receives a Challenge score of 0.821 on the hidden validation set and 0.708 on the hidden test set.

1. Introduction

In the 2023 George B. Moody PhysioNet Challenge [1], teams aim to develop automated methods to predict patient outcomes after cardiac arrest from long-term EEG and some other types of physiological time series such as ECG [2]. There is plenty of research focusing on extracting representative features from EEG for training the machine-learning model, which greatly speeds the analytical procedure of long-term EEG and then reduces the labor work for neurologists. Some work first applies knowledge-driven extraction pipelines to collect various kinds of expert features and then combine these features for prediction. In this way, practitioners are required to have certain prior domain knowledge of the data to build an effective method. More recently, deep learning has emerged as a popular data-driven method for end-to-end feature extraction. The deep representations learned by deep learning methods usually boost the performance on many downstream tasks and may not be explored by the traditional feature extraction methods.

In this work, we choose to utilize deep learning for feature extraction from long-term EEG. However, sufficient

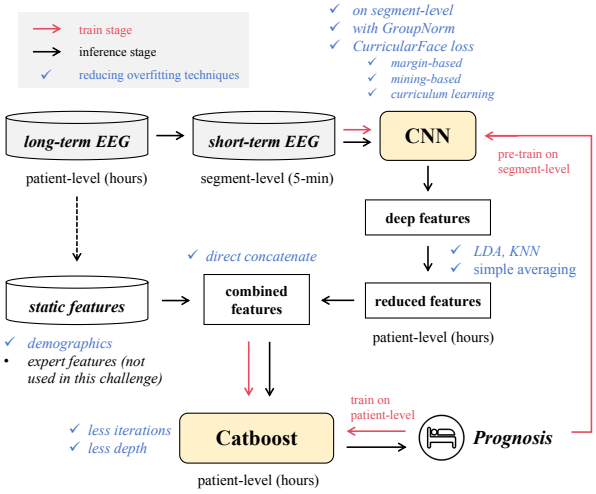

Figure 1. The framework of the proposed method RODE

This work was done when Songchi Zhou and Jun Li were interns at National Institute of Health Data Science, Peking University

This work was supported by the National Natural Science Foundation of China (No.62102008).

*Corresponding author: Shenda Hong

data is needed for a deep learning framework to prevent the common overfitting problem, which is challenging when using the I-CARE dataset [2] on the patient level. Consequently, we Reduce the Overfitting via adopt several useful training settings for the Deep learning process of EEG (RODE), including training on the segment level, use of a margin-based and mining-based loss, and dimensionality reduction with transformations. The obtained features are then combined with static features composed of demographic features and expert features to formulate the final input to the Catboost classifier.

In summary, our contributions include: 1) We propose a novel automated method powered by deep learning for neurologic outcome prognosis. Specifically, we adopt effective procedures to reduce overfitting in deep learning and thus enable the model to learn more generalizable deep features; 2) Our method achieves impressive performance on the clinical metrics, which is promising to help clinical workers make better prognosis decisions for comatose patients after cardiac arrest.

2. Methods

The overall framework of our proposed method is shown in Figure 1. Generally, a deep neural network is trained on the longitudinal EEG to extract deep features and an ensemble classifier combines the deep features with static features for patient outcome prediction. We elaborate the designs for reducing overfitting in Table 1.

2.1. Data preprocessing

We choose 4 EEG channels of F3, F4, P3, and P4 and recompute them into two bipolar channel data, i.e. F3-P3, F4-P4. The longitudinal EEG is filtered to keep the bandpass frequencies between 0.1 and 30 Hz and then resampled to 100 Hz to reduce the data size. We only keep the EEG that was recorded ahead of 72 hours after ROSC (return of spontaneous circulation) for model training. Note that we do not apply further normalization to the EEG.

We split the long-term EEG into the 5-minute segment level and ensure the resulting segment EEG pertains to a single recording hour, which is essential for data consistency. Normally, a 5-minute time window of EEG is sufficiently long to recognize irregular brain rhythm, so we label each time window with the outcome label of the patient it is derived from. This process could generate many more samples for deep learning than directly training on the patient level, and more importantly, is more feasible considering the large quantity of long-term EEG.

2.2. Deep learning method

Convolutional neural networks (CNN) have been widely used in the field of signal processing, especially for the analysis of physiological time series [3]. In this work, we utilize an improved version of CNN [4] as the backbone for automatic feature extraction of the longitudinal EEG.

After consideration, it occurs to us that, even though every 5-minute EEG segment is different in morphology, frequency domain, and so on, they all belong to the same patient, which interestingly shows some correspondence of different facial views of the same person. We then formulate the EEG segment classification as a typical face recognition problem. We use a popular deep metric learning method in the face recognition field to guide the CNN to learn discriminative features of the EEG. Though self-supervised learning has been widely implemented in the EEG analysis [5], we would mainly use labels to help the deep learning model learn an efficient metric to distinguish data samples of different outcomes. Therefore, our method is in the context of supervised deep metric learning.

Usually, the softmax-based loss is used in the time series classification problems, which may fail to learn representative features with only limited data samples. Consequently, margin-based loss functions are proposed to prompt the model to learn more discriminative deep features which could be more generalizable to mitigate overfitting. There are also some mining-based methods for allocating higher weights to hard samples, which proved to boost the model performance. In this work, we apply the CurricularFace loss [6], in which the margin-based setting and the mining-based setting are jointly learned and enhanced by curriculum learning, to train the deep neural network.

The vanilla softmax loss is usually defined as $L = -\log \frac{e^{W_{y_i} x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j x_i + b_j}}$, where x_i is the output feature vector of CNN with respect to the i-th sample, y_i is the label of sample i, n is the number of classes. W_j denotes the transformation part related to class j, and b_j denotes the bias which is usually ingored. By applying the l_2 normalization, W_j is normalized to 1 and the deep feature x_i is both normalized and then rescaled to s. Now $W_j x_i = ||W_j||||x_i||cos\theta_j = s(cos\theta_j)$. Thus, the softmax is rewritten as $L = -\log \frac{e^{s(cos\theta_{y_i})}}{\sum_{j=1}^{n} e^{s(cos\theta_j)}}$.

In CurricularFace, the setting of positive and negative similarity functions are parameterized as $P(cos\theta_{y_i}) = cos(\theta_{y_i} + m)$, $N(t, cos\theta_j) = cos\theta_j$ when $P(cos\theta_{y_i}) - cos\theta_j \geq 0$, and $N(t, cos\theta_j) = cos\theta_j(t + cos\theta_j)$ when $P(cos\theta_{y_i}) - cos\theta_j < 0$. m is the margin coefficient and t is a parameter adaptively learned by the curricular learning method. With Exponential Moving Average (EMA), t is updated via $t^{(k)} = \alpha r^{(k)} + (1 - \alpha)t^{(k-1)}$, where $t^{(0)} = 0$, $r^{(k)} = \sum_i cos\theta_{y_i}$ is the average of the simi-

	Items	Example Options	Notes	Implementations
Deep learning (Section 2.2)	Training objects	- Segment-level EEG - Patient-level EEG	The former produces more training samples for model training and is more feasible in the context of long-term EEG classification	Segment-level
	The loss function	- Margin-based - Mining-based	The margin-based loss poses challenges for DNN to learn more discriminative features and the mining-based loss allows DNN to assign a different weight for easy and hard samples	CurricularFace
	Normalizations	- BatchNorm - GroupNorm	BatchNorm reduces the individual impact of samples on the training process and GroupNorm is more appropriate in general settings with varied batch size	GroupNorm
Feature combinations (Section 2.3)	Segment → Patient	- Simple averaging - Learnable averaging	Learnable averaging introduces extra parameters to optimize which is more prone to cause overfitting but may lead to better results when suitably trained	Simple averaging
	Dimensionality reduction	- Supervised methods - Unsupervised methods	The ensemble model may degenerate due to redundant deep features and supervised methods could utilize label information to better learn the reduction process	LDA+KNN
	+Static features	- Direct concatenate - Importance weight	The feature importance of static features and deep features need to be compared and adjusted thoughtfully	Direct concatenation
Ensemble classifier (Section 2.4)	Model type	- Xgboost - Catboost	Generally, GBDT (Gradient Boosting Decision Tree) methods are popular candidates for ensemble learning with respect to heterogeneous features (static & deep features)	Catboost
	Parameter tuning	- Maximum number of trees - Tree depth	As for parameters such as the maximum number of trees (*iterations* in Caboost) and the tree depth (*depth* in Caboost), a small value may help reduce overfitting	Iterations: 1000→100 Depth: 6→4

Table 1. A recipe for reducing overfitting in EEG classification

larity value in the k-th batch, and α is the the momentum coefficient. Therefore, the final CurricularFace loss can be derived from the equations above and written as:

$$L = -\log \frac{e^{s(cos\theta_{y_i}+m)}}{e^{s(cos\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{n} e^{sN(t^k,cos\theta_j)}}$$

GroupNorm is used to replace BatchNorm for better performance. In CurricularFace, m is set to 0.5, s is set to 64, and α is set to 0.99.

2.3. Combine static features and deep features with dimensionality reduction

Static features are composed of basic demographic features and hand-crafted expert features of EEG. Demographic features are extracted from the patient metadata, including age, sex, ROSC, OHCA (out-of-hospital cardiac arrest), shockable rhythm, and TTM (targeted temperature management). After applying one-hot encoding on the sex variable, all these features are put together to formulate the final patient demographic features. Expert features of EEG can be derived from the careful design of extraction methods such as frequency-domain analysis, time-domain analysis, and so on. In this work, we assume that the deep features of EEG contain sufficient information to discriminate each class, so we apply no extra feature engineering.

After training, the CNN generates a high-dimensional deep feature vector from EEG segments, posing challenges for the ensemble classifier in concatenating and correlating it with static features. To address this, we employ two dimensionality reduction methods to reduce the dimension of deep features. These reduced features are then concatenated with static features and fed into the ensemble classifier, mitigating overfitting issues.

The first method is Linear Discriminant Analysis (LDA), which is a classic supervised dimensionality method to project the input into the most discriminative directions. The second method is based on the nearest neighbor search method, which would be used to compute the number of close-by embeddings for each class given a fixed size of K. The number of dimensions reduced by the second method (KNN) corresponds to the number of classes in the dataset. Then the LDA output is concatenated with the KNN output to formulate the reduced features. In implementation, the LDA-reduced dimension is 1 and the k nearest neighbors in KNN is set to 5.

The deep features of all EEG segments of the same patient are simply averaged on each dimension to form one single vector representing the information learned from the long-term EEG. Herein we do not apply learnable pooling methods in case introducing more unnecessary parameters could intensify overfitting.

2.4. Ensemble classifier

Catboost [7] is used as the ensemble classifier to make the eventual prognosis. The reduced feature is combined with static features as the final input to the Catboost model. Note that the dimensionality reduction methods mentioned above are already integrated into the pipeline of Catboost and we thus take advantage of them.

3. Results

On the public training set, we apply a 5-fold cross-validation to fully analyze the model performance. The first evaluation metric is the official Challenge score, with area under the receiver operating characteristic (AUROC)

	Score	AUROC	AUPRC
vanilla softmax	0.497±0.165	0.852±0.017	0.881±0.021
w/o reduction	0.678±0.041	0.894±0.018	0.930±0.005
default catboost	0.691±0.075	0.901±0.022	0.933±0.018
RODE	**0.723**±0.050	**0.906**±0.026	**0.937**±0.018

Table 2. Results of the ablation study of 5-fold cross validation on the public training set. Score refers to the true positive rate at a false positive rate of 0.05.

Training	Validation	Test	Ranking
0.723 ± 0.05	0.821	0.708	5/36

Table 3. The official Challenge score for our final selected entry with the ranking of our team on the hidden test set out of 36 official entries and 75 unofficial entries.

and area under Precision-Recall (PR) curve (AUPRC) as typical metrics for evaluating binary classification performance. For the purpose of ablation studies, we show the results for the metrics value of different settings. Specifically, we utilize the vanilla softmax loss for model training and further explore the effect of dimensionality reduction methods and the influence of some Catboost hyperparameters regarding the overfitting problem.

As is shown in Table 2, the proposed framework outperforms the baselines across all metrics, obtaining 0.723 on average for the CinC Challenge score, 0.906 on average for AUROC, and 0.937 on average for AUPRC. The ablation results indicate that each modification in the framework will help to boost prediction performance. Table 3 presents the summarized Challenge results.

4. Discussion and Conclusions

In most clinical scenarios, such as the I-CARE dataset [2], the patient-level quantity is relatively small but the amount of overall physiological recordings in each patient could be rather large, which is liable to result in overfitting in automated machine learning methods. To tackle this, we adopt several essential procedures to reduce overfitting in the proposed framework, especially for the deep learning part, which excels at extracting deep representations but is usually constrained by the issue of overfitting. Specifically, we formulate the EEG segmentation as a face recognition task and utilize advanced deep metric learning methods to train the CNN. In this process, EEG representations from various views are learned for each patient, which greatly helps the network extract discriminative and generalizable deep features. We then use dimensionality reduction methods on the deep features to produce reduced features with more compact EEG representations, which further help to mitigate the overfitting problem. In addition, we explore the effects of different parameters of the ensemble classifier. As is shown in the experimental results, the proposed framework achieves satisfactory performance on the cross-validation setting and also the Challenge hidden validation set, with prominent flexibility and scalability. To aid future research, we summarize our efforts of reduce overfitting, presented in Table 1.

Nevertheless, there is still some future work worthy of further exploration. First, more dimensionality methods could be used to reduce the dimension of the DNN output, such as Neighborhood Component Analysis (NCA). Second, we did not spend much time on the hyperparameter tuning process of the ensemble classifier considering the time cost of training this vast amount of time series data. At last, transfer learning methods and self/un-supervised learning methods that pre-train the deep learning model on a large and multiple dataset could be tested.

In this work, we propose RODE using long-term EEG for the outcome prognosis of patients after cardiac arrest. Our objective is to reduce overfitting in deep learning, a crucial step with the potential to significantly enhance the modeling of physiological time series, including EEG, ECG, PPG, PCG, PSG, and more.

References

[1] 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. volume 50. IEEE, 2023; 1–4.

[2] 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 10 2023;.

[3] Perslev M, Jensen M, Darkner S, Jennum PJ, Igel C. U-time: A fully convolutional network for time series segmentation applied to sleep staging. In NeurIPS, volume 32. 2019; .

[4] Hong S, Xu Y, Khare A, Priambada S, Maher K, Aljiffry A, et al. Holmes: health online model ensemble serving for deep learning models in intensive care units. In KDD. 2020; 1614–1624.

[5] Rafiei MH, Gauthier LV, Adeli H, Takabi D. Self-supervised learning for electroencephalography. IEEE Trans Neural Netw Learn Syst 2022;.

[6] Huang Y, Wang Y, Tai Y, Liu X, Shen P, Li S, et al. Curricularface: adaptive curriculum learning loss for deep face recognition. In CVPR. 2020; 5901–5910.

[7] Dorogush AV, Ershov V, Gulin A. Catboost: gradient boosting with categorical features support. arXiv preprint arXiv181011363 2018;.

Address for correspondence:

Shenda Hong

National Institute of Health Data Science, Peking University, Beijing, 100191, China

hongshenda@pku.edu.cn