Invariant Spatiotemporal Representation Learning for Cross-patient Seizure Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Automatic seizure type classification from electroencephalogram (EEG) data can help clinicians to better diagnose epilepsy. Although many previous studies have focused on the classification problem of seizure EEG data, most of these methods require that there is no distribution shift between training data and test data, which greatly limits the applicability in real-world scenarios. In this paper, we propose an invariant spatiotemporal representation learning method for cross-patient seizure classification. Specifically, we first split the spatiotemporal EEG data into different environments based on heterogeneous risk minimization to reflect the spurious correlations. We then learn invariant spatiotemporal representations and train the seizure classification model based on the learned representations to achieve accurate seizure type classification across various environments. The experiments are conducted on the largest public EEG dataset, the Temple University Hospital Seizure Corpus (TUSZ) dataset, and the experimental results demonstrate the effectiveness of our method.

1 Introduction

Epilepsy is a pervasive neurological disease that affects individuals all over the world, with considerable cognitive, psychological, and social ramifications [4]. The mainstream approach to epilepsy diagnosis relies on EEG data to classify seizures [8, 9]. However, traditional methods based on human labor are not only costly, but also susceptible to human uncertainty, as these methods require clinicians to meticulously review extensive EEG recordings [17]. As a result, using machine learning techniques to automatically classify seizure type attract increasingly attentions.

Early machine learning methods for accurately classifying EEG data included Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Bayesian methods [19, 32]. With the advancement of deep learning, Convolutional Neural Networks (CNNs) [34] and Recurrent Neural Networks (RNNs) [33] have been introduced. CNN-based methods typically aim at learning spatiotemporal feature representations of EEG signals through convolutional operations [6], exemplified by EEG-DBNet [24] and ACPA-ResNet [41]. RNNs, including CNN-BiRNN and CNN-Bi-LSTM [15, 25], capture temporal dependencies and dynamics. To address non-Euclidean geometric properties overlooked by CNNs and RNNs, Graph Neural Networks (GNNs) have been proposed to model the spatial relationships between EEG electrodes using a graph representation [10, 12, 18]. Methods such as REST [1], DCRNN [35], NeuroGNN [11] integrate GNNs with recurrent structures to enhance classification by capturing spatiotemporal dependencies and dynamic interactions.

However, these aforementioned methods are predominantly patient-specific and rely on a consistent distribution between training and test sets, which limits their ability to address cross-patient problem [40]. This kind of problems can be partially attributed to the spatial-temporal evolution of EEG data, which is common in real-world scenario where data from different patients exhibit significant

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

variability [22, 43]. Thus, for a group of new patients, it is very likely that this shift will impact the performance of models, leading to less precision and reduced generalizability. These challenges underscore the crucial and urgent need to develop robust cross-patient methods.

Previous methods addressing the cross-patient problem can be broadly categorized into three types. The first type involves unsupervised representation learning, particularly domain generalization, to initialize more robust representations for downstream tasks [38, 39, 42]. The second type revolves around supervised models to enhance generalization by employing techniques such as causal learning and invariant risk minimization. These approaches emphasize end-to-end learning strategies, which have been shown to improve robustness to distributional shifts [26–28]. The third type involves optimization-based approaches, including distributionally robust optimization (DRO), which focuses on minimizing the worst-case performance under potential shifts in the data [21, 30]. However, most of these methods ignore the spatiotemporal information, which leads to sub-optimal performance.

In this paper, we proposed a novel spatiotemporal invariant risk minimization loss to solve this problem. Specifically, we first use the invariant mask function to separate the raw EEG feature into the invariant representation and variant representation, and use the self-supervised learning (SSL) to guarantee the preserved invariant information is able to predict the invariant feature at next timestamp. In addition, we use the label information to generate the supervised signal to ensure the preserved invariant information can predict the seizure type. Finally, we use the variance of the gradient toward the mask function to control the time-varying variation of our methods in different patient groups.

We highlight our contributions as follows:

- We use the mask function to capture the invariant spatiotemporal information in the raw EEG data and use such information for self-supervised learning.
- To further control the variation of the loss of the classification model, we use the variance of the gradient as the penalty to achieve invariant learning.
- The experiments on the largest public dataset verify the effectiveness of our method.

2 Preliminary

2.1 Problem Setup

The primary objective of the seizure classification task is to predict the seizure type from a given EEG signal clip. These clips were sliced from seizure EEGs using non-overlapping sliding windows with fixed temporal size T. For each sample, we denote $X \in \mathbb{R}^{T \times N \times M}$ as the EEG clip feature after preprocessing, where T is the temporal length of the EEG clip, N is the number of EEG channels/electrodes, and M is the number of features obtained through Fast Fourier Transform (FFT). Meanwhile, we denote y as the seizure class label. For the independent identical distributed scenario, different clips from the same patient may appear in both the training and test sets. However, in real healthcare scenarios, patients in the test sets (a group of new patients) may completely unseen in the training set, leading to the cross-patient problem [44], which can be further formulated as follows: The patient set P is divided into two disjoint subsets, P_T and P_D, such that $P_T \cup P_D = P$ and $P_T \cap P_D = \emptyset$. Here, P_D is used for training, and P_T is used for testing.

2.2 Previous Graph-Based methods for EEGs

Graph Representing. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \boldsymbol{W}\}$ denote the graph structure, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ refers to the set of edge, and $\boldsymbol{W}$ is the adjacency matrix of the graph. In consideration of the distribution of nodes and the physiological properties of the brain, two distinct approaches to graph construction on EEG clips are evident. One undirected distance graph-based approach is to utilize the Euclidean distance between different nodes on standard 10-20 EEG electrode placement as the basis, followed by the threshold Gaussian kernel to determine the weights between v_i and v_j ($v_i, v_j \in \mathcal{V}$):

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{\text{dist}(v_i, v_j)^2}{\sigma^2}\right) & \text{if dist}(v_i, v_j) \leq \zeta \\ 0 & \text{otherwise,} \end{cases}$$

where $\text{dist}(v_i, v_j)$ represents the Euclidean distance between two nodes v_i and v_j, σ is the standard deviation of the distances, while ζ is the threshold for sparsity.

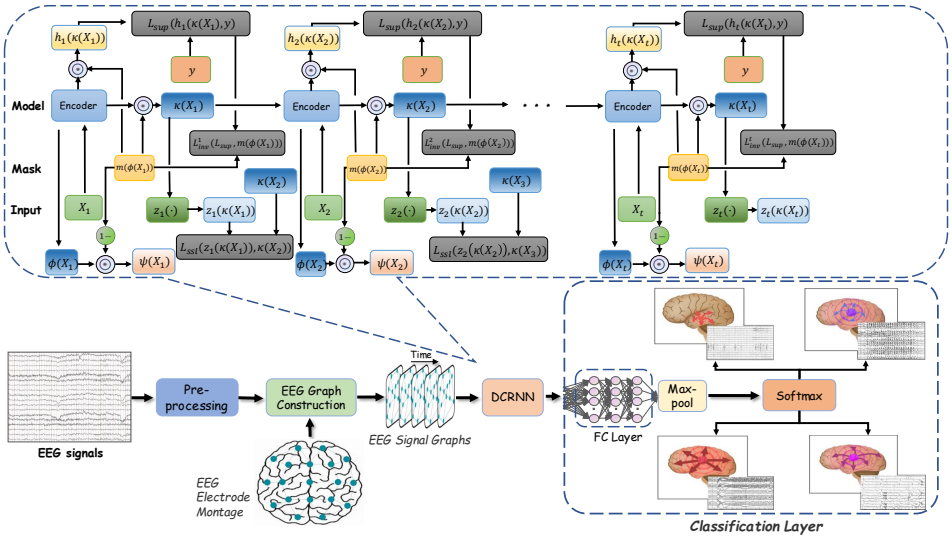

Figure 1: Overview of the proposed spatiotemporal invariant learning method.

An alternative approach, based on a directed correlation graph, particularly focuses on the dynamic connectivity between different nodes. To evaluate the connectivity, only the weights that fall within the most k similar neighbors (including self-edges) are preserved to ensure the sparsity of the graph. The weight can be formulated as follows:

$$W_{ij} = \begin{cases} Corr(\mathbf{X}_{:,i,:}, \mathbf{X}_{:,j,:}) & \text{if } v_j \in \mathcal{C}_k(v_i) \\ 0 & \text{otherwise,} \end{cases}$$

where $X_{:,i,:}$ and $X_{:,j,:}$ denotes the preprocessed signals in v_i and v_j, $Corr(\cdot,\cdot)$ represents the pearson correlation coefficient, and $\mathcal{C}_k(v_i)$ referring to the most k similar neighbors of v_i.

Diffusion Convolutional Recurrent Neural Network. Previous works utilize the diffusion convolutional recurrent neural network (DCRNN) to effectively capture the temporal and spatial dependencies in EEG signals. To capture the temporal dependencies in EEG data, modified gated recurrent units (GRUs) [5] are employed. For spatial dependency, diffusion convolution provides significant insights into the influence exerted by each node on all others, and the extent of this kind of influence can be quantified by applying a bidirectional random walk on the directed graph and calculating a K-step diffusion convolution. The diffusion convolution is defined by:

$$X_{:,m\star\mathcal{G}} f_\theta = \sum_{k=0}^{K-1} \left(\theta_{k,1}(D_O^{-1}W)^k + \theta_{k,2}(D_I^{-1}W^\intercal)^k\right) X_{:,m}, \qquad m \in \{1,\ldots,M\},$$

where X is the preprocessed segment with N nodes and M features at timestamps $t \in \{1,\cdots,T\}$, $\theta \in \mathbb{R}^{K\times 2}$ are the parameters of the filter, and D_O and D_I are the out-degree and in-degree diagonal matrices of the graph. The transition matrices for the diffusion processes are $D_O^{-1}W$ and $D_I^{-1}W^\intercal$.

For undirected graphs, the process converts to ChebNet spectral graph convolution [7], where $X_{:,m}$ is filtered using Chebyshev polynomial bases. The spectral graph convolution can be expressed as

$$X_{:,m\star\mathcal{G}} f_\theta = \mathbf{\Phi}\left(\sum_{k=0}^{K-1} \theta_k \mathbf{\Lambda}^k\right)\mathbf{\Phi}^\intercal X_{:,m} = \sum_{k=0}^{K-1} \theta_k \mathbf{L}^k X_{:,m} = \sum_{k=0}^{K-1} \tilde{\theta}_k T_k(\tilde{L}) X_{:,m}, \qquad m \in \{1,...,M\},$$

where $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ are the eigenvector and eigenvalue matrix of the graph Laplacian $\mathbf{L}$. $T_k(\tilde{L})$ is the k-th Chebyshev polynomial of the scaled Laplacian $\tilde{L}$, allowing for efficient computation without explicit eigenvalue decomposition.

3 Methodology

In a cross-patient scenario, we propose the spatiotemporal invariant risk minimization (ST-IRM) loss, making the prediction model achieves both (a) accurately predicting patient's seizure type in each patient group; (b) The variation of prediction between the different groups is small. Specifically, for a timestamp t, we derive an invariant mask function $m(\cdot)$ to separate the representations of the raw EEG feature into two orthogonal components. We denote the representation of the raw EEG feature as $\phi(X_{:,:,t})$. For simplification of notations, we use X_t instead of $X_{:,:,t}$. The representation in the present paper is obtained by DRCNN. Through the invariant mask function $m(\cdot)$, $\phi(X_t)$ is decomposed into an invariant representation $\kappa(X_t) = m(\phi(X_t))$, and the variant representation $\psi(X_t) = (1 - m(\phi(X_t))) \odot \phi(X_t)$, where $m(X_t) \in [0, 1]^{N \times M}$.

In time-series data, especially EEG data, there should be some correlation of the previous representations X_{t-1} with the current feature X_t [35]. Unlike the previous SSL approach that aims to learn a model $z_t(\cdot)$ to ensure $z_{t-1}(X_{t-1}) \approx X_t$, we claim that preserve the relation between the variant parts, $\psi(X_{t-1})$ and $\psi(X_t)$ may not be helpful due to the spurious correlation. We expect only a good prediction performance between the invariant representations. The proposed SSL loss is as below:

$$\mathcal{L}_{ssl} = \frac{1}{|nT|} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathcal{L}(z_{t-1}(m(\phi(X_{t-1}^i))), m(\phi(X_t^i))),$$

where $\mathcal{L}(\cdot, \cdot)$ is the loss function such as mean-square-error loss and $X_t^i \in \mathbb{R}^{N \times M}$ is the preprocessed signal for sample i at timestamp t. In addition, we want the information preserved by the mask function can not only predict the next invariant representation but also can predict the final seizure type, thus we use the following loss to provide the supervised signal for training the mask function:

$$\mathcal{L}_{sup} = \frac{1}{|n|} \sum_{i=1}^{n} \mathcal{L}(h_T(m(\phi(X_T^i))), y_i),$$

where $h_T(\cdot)$ is the classification model and y_i is the ground truth label.

In addition, an ideal mask function $m(\cdot)$ should be able to capture the invariant representation from the raw EGG data. To address this, environments are created using the K-means clustering method to separate the samples into groups, ensuring that samples within a group share similar characteristics, while those in different groups exhibit distinct features. Thus, a classifier that performs consistently across these environments would truly learn the invariant components and suffer the least from spurious correlations. Assuming there is a total of G groups/environments, and the group indicator of each sample is denoted by g_i. The supervised loss at timestamp t for the group g is given by

$$\mathcal{L}_{sup}^{g,t} = \frac{1}{\sharp\{i : g_i = g\}} \sum_{\{i:g_i=g\}} \mathcal{L}(h_t(m(\phi(X_t^i))), y_i),$$

where $\sharp$ denotes the cardinal number of the set. It represents the supervised loss within the g-th group. Combining the group-based supervised loss, the overall invariant risk loss at timestamp t is composed of two major terms:

$$\mathcal{L}_{inv}^t = \mathbb{E}_{g \in \mathcal{G}} \mathcal{L}_{sup}^{g,t} + \lambda \left\| \mathrm{Var}_{g \in \mathcal{G}} \left(\nabla_{\Theta^{\mathrm{m}}} \mathcal{L}_{sup}^{g,t} \odot m(\phi(X_t)) \right) \right\|^2,$$

where Θ^{m} is the parameter of the mask function, and λ is the hyper parameter for tuning. The previous term can be naively computed by $\frac{1}{n} \sum_{g \in \mathcal{G}} \mathcal{L}_{sup}^{g,t}$, suggesting the overall supervised loss at timestamp t; while the second term penalizes the classifier to perform consistently across groups. The variance depicts the variation across the environments: the lower the variance is, the more consistent performance the classifier obtains, thus, the better invariant presentation the classifier has learned with. In the second term, we multiply the gradient with the mask function for scaling. For further incorporating the spatiotemporal information, because the more information being observed, the more accurate classification should be, we propose the weight decay loss below:

$$\mathcal{L}_{inv} = \sum_{t=1}^{T} w^{T-t} \mathcal{L}_{inv}^t,$$

where $w \in (0, 1)$ is the weight decay rate, which is a hyper-parameter for tuning. The final proposed ST-IRM loss is:

$$\mathcal{L}_{ST-IRM} = \mathcal{L}_{ssl} + \alpha \mathcal{L}_{sup} + \beta \mathcal{L}_{inv},$$

where α and β are the hyper-parameters. An overview of the proposed method is given in Figure 1.

Table 1: Performance comparison of different methods under 12-second and 60-second scenario.

	12-s			60-s		
Method	F1	Recall	Precision	F1	Recall	Precision
CNN-LSTM	0.596 ± 0.035	0.654 ± 0.030	0.647 ± 0.036	0.623 ± 0.028	0.661 ± 0.030	0.647 ± 0.036
LSTM	0.690 ± 0.043	0.724 ± 0.033	0.725 ± 0.041	0.692 ± 0.011	0.718 ± 0.007	0.717 ± 0.017
Dense-CNN	0.657 ± 0.069	0.690 ± 0.053	0.694 ± 0.049	0.653 ± 0.085	0.704 ± 0.057	0.659 ± 0.118
MSTGCN	0.670 ± 0.031	0.719 ± 0.023	0.734 ± 0.029	0.647 ± 0.046	0.696 ± 0.027	0.694 ± 0.030
NeuroGNN	0.647 ± 0.040	0.710 ± 0.024	0.744 ± 0.030	0.698 ± 0.044	0.733 ± 0.042	0.714 ± 0.056
Corr-DCRNN	0.729 ± 0.058	0.756 ± 0.041	0.752 ± 0.047	0.672 ± 0.038	0.712 ± 0.021	0.705 ± 0.029
Dist-DCRNN	0.713 ± 0.044	0.735 ± 0.043	0.734 ± 0.045	0.695 ± 0.028	0.735 ± 0.013	0.738 ± 0.021
PANN-DCRNN	0.728 ± 0.052	0.753 ± 0.042	0.755 ± 0.041	0.684 ± 0.023	0.717 ± 0.016	0.720 ± 0.024
ST-InvDCRNN(ours)	**0.748** ± **0.038**	**0.772** ± **0.028**	**0.764** ± **0.043**	**0.713** ± **0.043**	**0.741** ± **0.024**	**0.742** ± **0.037**

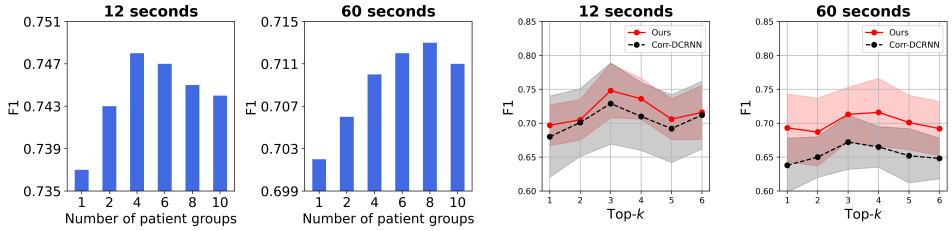

Figure 2: F1 under different numbers of patient groups (the two subfigures on the left) and different values of hyper-parameter top-k to control the graph sparsity (the two subfigures on the right).

4 Experiments

4.1 Experimental Settings

Datasets. Following previous works [20, 31, 36], we utilized the Temple University Hospital EEG Seizure Corpus (TUSZ) dataset, which is the largest public dataset for our experiments. Specifically, we use the version v1.5.2 of the TUSZ dataset. The TUSZ dataset contains 5,612 EEG signals, and 3,050 annotated seizure events from over 300 patients, covering eight seizure types. The EEG signal was recorded using 19 electrodes from the standard 10-20 system [14].

Data preprocessing and Experiment Details. Following the preprocessing approach of Tang et al. [35], we transform the raw EEG signals into the frequency domain, as seizures are associated with brain electrical activity in specific frequency bands [37]. Following prior methodologies [2, 3], EEG recordings were resampled to 200Hz and segmented into non-overlapped 60-second windows (clips), and only clips that contain a single type of seizure are considered. If a seizure event ends and another begins within the same clip, it is truncated and zero-padded to preserve a 60-second duration. Each 60-second clip is then segmented into 1-second intervals. The Fast Fourier Transform (FFT) algorithm is applied to each segment to obtain the logarithmic amplitudes of non-negative frequency components, as is outlined in Tang et al. [35]. Consequently, each 60-second clip is transformed into a sequence of 60 log-amplitude vectors. In addition, following recent studies on seizure type classification [2, 3, 35], we use weighted F1-score as the main evaluation metric with precision and recall as well to measure the classification performance. See Appendix B for more experiment protocols and details.

Baselines. We compare our proposed method with CNN-based method: **DenseCNN** [29], RNN-based method: **LSTM** [13], and hybrid approach that combine CNN and RNN: **CNN-LSTM** [2]. We also compared our method with GNN-based methods: **MSTGCN** [16], **Dist-DCRNN** [35], **Corr-DCRNN** [35], **NeuroGNN** [11], and **PANN** [44].

4.2 Performance Analysis

Table 1 shows the performance of our method compared with various baseline methods, evaluating with three metrics, i.e., weighted F1, Recall, and Precision scores. First, DCRNN-based models achieve competitive performance among all baselines. Second, our method significantly outperforms the baselines under both scenarios with 12-second and 60-second clip windows. Note that we adopt DCRNN as a backbone in the experiment, which is shown in Figure 1, and the superior against DCRNN-based methods demonstrates the effectiveness of our invariant learning framework.

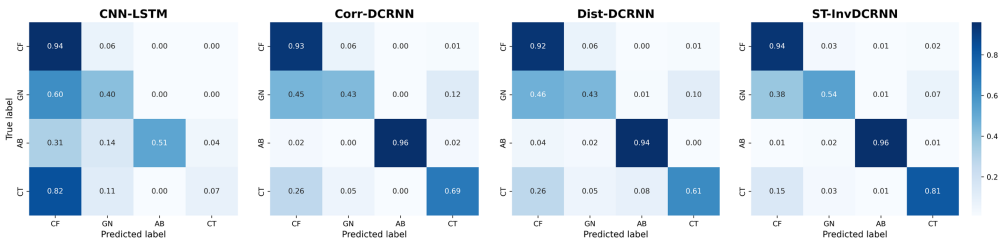

Figure 3: Confusion matrices for four classes of seizures.

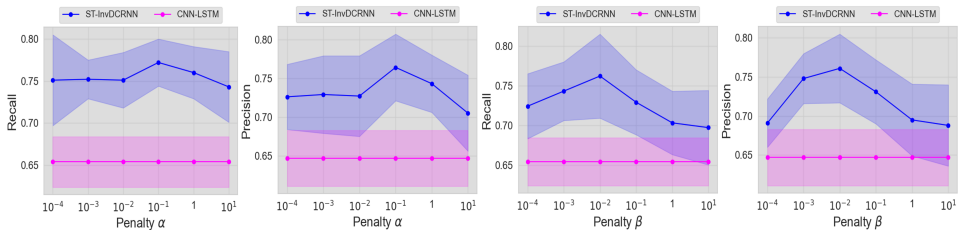

Figure 4: 12-second Performance under different penalty weights.

4.3 In-Depth Analysis

To comprehensively evaluate the proposed invariant learning method, we conduct four in-depth analyses on the number of patient groups, the value of hyper-parameter top-k, the classification confusion matrix, and the 12-second performance under different penalty weights respectively. Note that the patients are clustered into groups according to their EEG recordings, and the two subfigures on the left of Figure 2 show that different numbers of groups result in varying performance. In the scenario of 12-second clip windows, the best choice for group number is 4, while in the 60-second case, the best value is 8. Our method outperforms Corr-DCRNN with top-k ranging from 1 to 6, and the highest F1 is achieved when top-k is around 3 for both scenarios. In addition, we provide the results of the recall metric and the confusion metrics in Appendix B. Figure 3 shows the confusion matrices for four seizure classification models, highlighting the superior performance of our method. The ST-InvDCRNN reduces misclassifications and confusion between seizure types, notably 0.81 for the CT class and 0.54 for GN seizures, outperforming baseline models. Figure 4 compares ST-InvDCRNN and CNN-LSTM performance across different penalty parameters (α and β) for recall and precision. ST-InvDCRNN consistently outperforms CNN-LSTM, especially at intermediate penalty values. For Penalty α, ST-InvDCRNN peaks at $\alpha = 10^{-1}$, achieving 0.772 recall score and 0.764 precision score, while CNN-LSTM shows lower scores. Similarly, for Penalty β, ST-InvDCRNN reaches its best performance at $\beta = 10^{-1}$, with 0.762 recall score and 0.761 prescision score. Overall, ST-InvDCRNN delivers better classification results.

5 Conclusion

Epilepsy remains a significant global health challenge, with traditional EEG-based diagnostic methods posing limitations due to their reliance on clinician review. With the recent advancement of deep learning, techniques such as CNNs, RNNs, and GNNs are proposed to automatically classify the seizure type. However, existing methods often lack cross-patient robustness and guarantee, which is very common in practice. In addition, most of the methods addressing the cross-patient problem ignore the spatiotemporal information. To bridge this gap, we propose a spatiotemporal invariant risk minimization approach that addresses these challenges by adopting self-supervised learning and capturing time-varying invariant features. Experimental results on the largest public dataset verify the effectiveness of our approach, demonstrating its potential to advance epilepsy diagnosis in the cross-patient scenario. One of the possible limitations is to investigate a more efficient way to learn the model parameters and reduce the complexity while maintaining the classification performance.

References

[1] Arshia Afzal, Grigorios Chrysos, Volkan Cevher, and Mahsa Shoaran. Rest: Efficient and accelerated eeg seizure analysis through residual state updates. *arXiv preprint arXiv:2406.16906*, 2024.

[2] David Ahmedt-Aristizabal, Tharindu Fernando, Simon Denman, Lars Petersson, Matthew J Aburn, and Clinton Fookes. Neural memory networks for seizure type classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 569–575. IEEE, 2020.

[3] Umar Asif, Subhrajit Roy, Jianbin Tang, and Stefan Harrer. Seizurenet: Multi-spectral deep feature learning for seizure type classification. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 77–87. Springer, 2020.

[4] Ettore Beghi. The epidemiology of epilepsy. *Neuroepidemiology*, 54(2):185–191, 2020.

[5] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, page 103. Association for Computational Linguistics, 2014.

[6] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[8] Jessica Falco-Walter. Epilepsy—definition, classification, pathophysiology, and epidemiology. In *Seminars in neurology*, volume 40, pages 617–623. Thieme Medical Publishers, Inc., 2020.

[9] Robert S Fisher, J Helen Cross, Jacqueline A French, Norimichi Higurashi, Edouard Hirsch, Floor E Jansen, Lieven Lagae, Solomon L Moshé, Jukka Peltola, Eliane Roulet Perez, et al. Operational classification of seizure types by the international league against epilepsy: Position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):522–530, 2017.

[10] Arash Hajisafi, Haowen Lin, Sina Shaham, Haoji Hu, Maria Despoina Siampou, Yao-Yi Chiang, and Cyrus Shahabi. Learning dynamic graphs from all contextual information for accurate point-of-interest visit forecasting. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–12, 2023.

[11] Arash Hajisafi, Haowen Lin, Yao-Yi Chiang, and Cyrus Shahabi. Dynamic gnns for precise seizure detection and classification from eeg data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 207–220. Springer, 2024.

[12] Jiatong He, Jia Cui, Gaobo Zhang, Mingrui Xue, Dengyu Chu, and Yanna Zhao. Spatial–temporal seizure detection with graph attention network and bi-directional lstm architecture. *Biomedical Signal Processing and Control*, 78:103908, 2022.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Richard W Homan, John Herman, and Phillip Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and Clinical Neurophysiology*, 66(4):376–382, 1987. ISSN 0013-4694.

[15] Chengbin Huang, Weiting Chen, and Guitao Cao. Automatic epileptic seizure detection via attention-based cnn-birnn. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 660–663, 2019. doi: 10.1109/BIBM47256.2019.8983420.

[16] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and H Lehman Li-wei. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021.

[17] Yizhang Jiang, Dongrui Wu, Zhaohong Deng, Pengjiang Qian, Jun Wang, Guanjin Wang, Fu-Lai Chung, Kup-Sze Choi, and Shitong Wang. Seizure classification from eeg signals using transfer learning, semi-supervised learning and tsk fuzzy system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12):2270–2284, 2017.

[18] Dominik Klepl, Min Wu, and Fei He. Graph neural network-based eeg classification: A survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[19] Alicia Guadalupe Lazcano-Herrera, Rita Q Fuentes-Aguilar, and Mariel Alfaro-Ponce. Eeg motor/imagery signal classification comparative using machine learning algorithms. In *2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–6. IEEE, 2021.

[20] Yang Li, Yu Liu, Wei-Gang Cui, Yu-Zhu Guo, Hui Huang, and Zhong-Yi Hu. Epileptic seizure detection in eeg signals using a unified temporal-spectral squeeze-and-excitation network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(4):782–794, 2020. doi: 10.1109/TNSRE.2020.2973434.

[21] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8662–8670, 2021.

[22] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[24] X Lou, X Li, H Meng, J Hu, M Xu, Y Zhao, J Yang, and Z Li. Eeg-dbnet: A dual-branch network for temporal-spectral decoding in motor-imagery brain-computer interfaces. 2024.

[25] Yahong Ma, Zhentao Huang, Jianyun Su, Hangyu Shi, Dong Wang, Shanshan Jia, and Weisu Li. A multi-channel feature fusion cnn-bi-lstm epilepsy eeg classification and prediction model based on attention mechanism. *IEEE Access*, 11:62855–62864, 2023.

[26] Bijan Mazaheri, Atalanti Mastakouri, Dominik Janzing, and Michaela Hardt. Causal information splitting: Engineering proxy features for robustness to distribution shifts. In *Uncertainty in Artificial Intelligence*, pages 1401–1411. PMLR, 2023.

[27] Michael Oberst, Nikolaj Thams, Jonas Peters, and David Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

[28] Advait U Parulekar, Karthikeyan Shanmugam, and Sanjay Shakkottai. Pac generalization via invariant representations. In *International Conference on Machine Learning*, pages 27378–27400. PMLR, 2023.

[29] Khaled Saab, Jared Dunnmon, Christopher Ré, Daniel Rubin, and Christopher Lee-Messer. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *NPJ digital medicine*, 3(1):59, 2020.

[30] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[31] Rijad Sarić, Dejan Jokić, Nejra Beganović, Lejla Gurbeta Pokvić, and Almir Badnjević. Fpga-based real-time epileptic seizure classification using artificial neural network. *Biomedical Signal Processing and Control*, 62:102106, 2020. ISSN 1746-8094.

[32] MNAH Sha'Abani, N Fuad, Norezmi Jamal, and MF Ismail. knn and svm classification for eeg: a review. In *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019*, pages 555–565. Springer, 2020.

[33] Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. Epileptic seizures detection using deep learning techniques: a review. *International journal of environmental research and public health*, 18(11):5780, 2021.

[34] Supriya Supriya, Siuly Siuly, Hua Wang, and Yanchun Zhang. Epilepsy detection from eeg using complex network techniques: A review. *IEEE Reviews in Biomedical Engineering*, 16:292–306, 2021.

[35] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*, 2022.

[36] Punnawish Thuwajit, Phurin Rangpong, Phattarapong Sawangjai, Phairot Autthasan, Rattanaphon Chaisaen, Nannapas Banluesombatkul, Puttaranun Boonchit, Nattasate Tatsaringkansakul, Thapanun Sudhawiyangkul, and Theerawit Wilaiprasitporn. Eegwavenet: Multiscale cnn-based spatiotemporal feature extraction for eeg seizure detection. *IEEE Transactions on Industrial Informatics*, 18(8):5547–5557, 2022. doi: 10.1109/TII.2021.3133307.

[37] Alexandros T Tzallas, Markos G Tsipouras, and Dimitrios I Fotiadis. Epileptic seizure detection in eegs using time–frequency analysis. *IEEE transactions on information technology in biomedicine*, 13(5): 703–710, 2009.

[38] Haiyang Yang, Shixiang Tang, Meilin Chen, Yizhou Wang, Feng Zhu, Lei Bai, Rui Zhao, and Wanli Ouyang. Domain invariant masked autoencoders for self-supervised learning from multi-domains. In *European Conference on Computer Vision*, pages 151–168. Springer, 2022.

[39] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

[40] Zhizhang Yuan, Daoze Zhang, YANG YANG, Junru Chen, and Yafeng Li. Ppi: Pretraining brain signal model for patient-independent seizure detection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 69586–69604. Curran Associates, Inc., 2023.

[41] Zhang Yutian, Huang Shan, Zhang Jianing, and Fan Ci'en. Design and implementation of an emotion analysis system based on eeg signals. *arXiv preprint arXiv:2405.16121*, 2024.

[42] Xiang Zhang, Lina Yao, Manqing Dong, Zhe Liu, Yu Zhang, and Yong Li. Adversarial representation learning for robust patient-independent epileptic seizure detection. *IEEE journal of biomedical and health informatics*, 24(10):2852–2859, 2020.

[43] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. Dynamic graph neural networks under spatio-temporal distribution shift. *Advances in neural information processing systems*, 35: 6074–6089, 2022.

[44] Zongpeng Zhang, Taoyun Ji, Mingqing Xiao, Wen Wang, Guojing Yu, Tong Lin, Yuwu Jiang, Xiaohua Zhou, and Zhouchen Lin. Cross-patient automatic epileptic seizure detection using patient-adversarial neural networks with spatio-temporal eeg augmentation. *Biomedical Signal Processing and Control*, 89: 105664, 2024.

Appendix

A Experimental Details

Following previous works, we divide the clips and patients of the TUSZ dataset into training, validation, and test sets. The number of EEG clips is 1,925, 450, and 521 for the three sets respectively, while the number of patients is 179, 22, and 34. Note that the patient sets are disjoint for training, validation, and test sets to study the cross-patient seizure classification robustness.

We tune the following hyper-parameters on the validation set.

- $lr\_init \in [1e-5, 5e-3]$, the initial learning rate;
- top-$k \in \{1, 2, 3, 4, 5, 6\}$, the number of neighbors included in the correlation graphs for each node;
- $K \in \{2, 3, 4\}$, the maximum diffusion step;
- $d \in [0, 0.7]$, the dropout probability in the prediction networks.
- $e \in [20, 40, 60, 80, 100]$, the number of training epochs.

During the training, each batch has 40 EEG clips and the cosine annealing learning rate scheduler [23] is adopted. Our experiments are conducted on a computing platform of NVIDIA GeForce RTX 3090 and Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz.

B Additional Evaluation Results

Figure 5 shows the weighted F1 and the Recall scores to evaluate the performance of our method under different number of patient groups, for both 12-second and 60-second clip windows. We can observe that as the number of patient groups increases, the Recall-score has a similar pattern as the weighted F1-score, achieving the highest value at 4 for the 12-second case and 8 for the 60-second case.

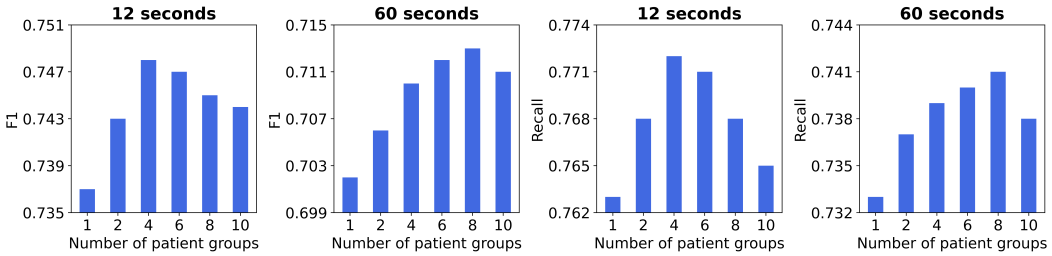

Figure 5: Performance under different numbers of patient groups.

Figure 6 shows the weighted F1 and the Recall scores to compare the performance of our method with Corr-DCRNN under different top-k values, for both 12-second and 60-second clip windows. As the value of top-k ranges from 1 to 6, the trend for both weighted F1 and Recall scores is increasing until a peak at around 3, followed by a slight decrease.

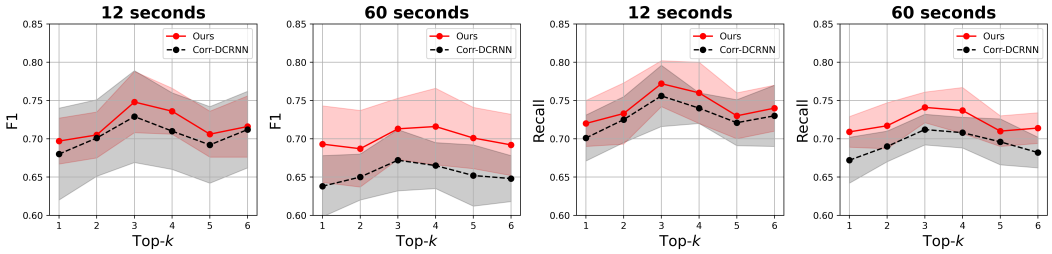

Figure 6: Performance under different values of top-k.