Causal Representation Learning for Cross-Patient Seizure Classification

Chunyuan Zheng[1] **Yan Lyu**[1] **Taojun Hu**[1] **Xiaoxin Liu**[1] **Xiao-Hua Zhou**[1*]
[1]Peking University, China

Abstract

Automated classification of epileptic seizure types from electroencephalogram (EEG) signals is important for accurate diagnosis, yet current state-of-the-art deep learning models often fail to generalize to unseen patients due to covariate distribution shifts. To overcome this limitation, this paper introduces Patient-Invariant Causal Representation Learning (PInvCRL), a novel framework that disentangles the patient covariate representation into the invariant and variant parts, and only uses the invariant part for final model training. Specifically, our method first employs a Diffusion Convolutional Recurrent Neural Network (DCRNN) to extract informative spatiotemporal representation from original multi-channel EEG signals. The representations are then clustered using k-means to establish an initial set of patient "environments". Subsequently, we initiate an iterative learning process that dynamically refines these environments while concurrently learning a covariate mask, which decomposes the covariate vector into a patient-invariant component, which captures the core seizure patterns stable across environments, and a residual patient-variant component. Finally, a Multi-Layer Perceptron (MLP) is trained based on these robust, invariant representations for the final seizure classification. We validate PInvCRL on the largest public Temple University Hospital Seizure Corpus (TUSZ) dataset, demonstrating that by explicitly removing patient-variant information, our model achieves a state-of-the-art performance in the cross-patient seizure classification scenario.

1 Introduction

Epilepsy is a globally prevalent neurological disorder, affecting millions of people [5]. Accurate seizure classification is fundamental for effective diagnosis to benefit society [6]. One of the significant challenges is to process the Electroencephalography (EEG) signal effectively during the diagnosis procedure [1]. However, the manual interpretation of lengthy EEG recordings by neurologists is labor-intensive, time-consuming, and subject to inter-rater variability [12, 25]. With the rapid development of the combination of electronic health and artificial intelligence [14, 13], machine learning-based methods for seizure classification have therefore attracted increasing attention [7, 21]. For example, Graph Neural Networks (GNNs) have emerged as a powerful tool to model the brain's network topology relations [26, 17], and architectures like the Diffusion Convolutional Recurrent Neural Network (DCRNN) have shown particular promise by integrating GNNs with recurrent units to capture complex spatiotemporal dynamics simultaneously [15].

Despite these advances, a critical limitation persists: most models are trained and tested under the assumption that data distributions are identical between the training set and test set, which is a condition rarely met in clinical practice. When a model is deployed on new patients not seen during training, its performance often degrades [20, 24]. This is because EEG signals contain a mixture of covariates: some are "invariant" and truly indicative of a specific seizure type, while others are

*Corresponding author. Email: azhou@math.pku.edu.cn

Accepted to the NeurIPS 2024 Workshop on Causal Representation Learning.

"variant" or "spurious" [29], which is beneficial for classification in the training set but not in the test set with another distribution. Traditional machine learning based models often conflate these covariate types, learning spurious correlations that fail to generalize to unseen patients with different covariate distributions [3].

Several recent works have explored approaches to improve cross-patient seizure classification. Zhang et al. [31] propose to use adversarial learning based on the patient ID to train a more generalizable classifier. Wu et al. [30] presented a spatiotemporal invariant representation learning framework with self-supervised consistency learning loss and prespecified environments based on a clustering algorithm. Despite these efforts, existing methods still face limitations in explicitly and dynamically finding invariant covariate representation that causally governs seizure types, which motivates our proposed causal representation learning approach [19].

In this work, we introduce a new framework, **Patient-Invariant Causal Representation Learning (PInvCRL)**, which treats individual patients (or clusters of similar patients) as different "environments" and learn representations that are robust across them using the invariant risk minimization techniques [3, 18]. Our contributions are summarized below:

- We propose a novel, multi-stage method that can tackle the cross-patient seizure classification problem by finding the invariant representation dynamically.
- Specifically, we first use a DCRNN to extract powerful spatiotemporal covariates from raw EEG signals. Then we adapt the concept of environment-based invariant learning, using k-means clustering for initialization and an iterative process to learn a covariate mask that separates patient-invariant representations from variant ones. Finally, we train an MLP classifier on these patient-invariant representations.
- Experiments on the large-scale TUSZ dataset that PInvCRL improves cross-patient seizure classification by mitigating patient-specific confounding factors.

2 Methodology

Our proposed PInvCRL framework is designed to learn invariant covariate representations that are the same across all patients.

Problem Setup. An input EEG clip is represented as $X \in \mathbb{R}^{T \times K \times M}$, where T is the number of time steps, K is the number of EEG channels (brain regions), and M is the covariate dimension after preprocessing. Our goal is to predict the seizure class label y. In addition, in our paper, we have four seizure types, i.e., $y \in \{0, 1, 2, 3\}$. In the cross-patient seizure classification task, the dataset is partitioned such that patients in the training set do not appear in the validation or test sets, simulating a real-world clinical scenario. Formally, the cross-patient scenario means that the distribution of X in the training set is not equal to the test set, but $P(Y \mid X)$ is the same.

Next, we introduce the proposed method, whose overall workflow consists of two main stages.

2.1 Stage 1: Spatiotemporal Covariate Extraction with DCRNN

DCRNN Encoder. We model the relationship between EEG channels as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$, where nodes $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ represent the K channels and the adjacency matrix W encodes their spatial relationships or functional connectivity. We employ a DCRNN as our primary covariate encoder. The DCRNN replaces standard matrix multiplications in a Gated Recurrent Unit (GRU) with diffusion convolution operations, allowing it to capture both spatial dependencies via graph convolutions and temporal dependencies via the recurrent structure. For an input sequence $X = (X^{(1)}, \ldots, X^{(T)})$, the DCRNN computes a sequence of hidden states $H = (H^{(1)}, \ldots, H^{(T)})$, where each $H^{(t)} \in \mathbb{R}^{K \times d}$ is a rich spatiotemporal representation of the EEG data in time t, and $d \ll M$ is the hidden dimension of the GRU. The details of the diffusion convolution step can be found at [15]. For training speed consideration, we first train a DCRNN encoder with a readout MLP layer, following Tang et al. [27], with cross entropy as the classification loss, then we fix the learned representation in the following stage.

2.2 Stage 2: Patient-Invariant Causal Representation Learning

This stage is the core of our PInvCRL framework. We take the final hidden state matrix $H^{(T)} \in \mathbb{R}^{K\times d}$ as the covariate matrix, which contains a d-dimensional covariate vector for each of the K brain regions, to avoid the high dimension problem caused by the original data while maintaining the enough information of predicting y. We then flatten this matrix into a vector $h_{flat} \in \mathbb{R}^{Kd}$. This vector serves as the informative input for the invariant learning algorithm. It separates the flattened representation h_{flat} into an "invariant" component that is stable across different patient environments and a "variant" component that captures patient-specific noise.

Invariant and Variant Representations. We introduce a learnable float vector, the invariant mask $m \in \mathbb{R}^{Kd}$, with values constrained in $[0, 1]$. This mask is used to decompose h_{flat}. The invariant representation Φ and the variant representation Ψ are defined as:

$$\Phi = m \odot h_{flat},$$
$$\Psi = (1 - m) \odot h_{flat},$$

where $\odot$ denotes the element-wise product. The goal is to learn a mask m such that Φ contains only generalizable, seizure-specific invariant information.

Environment Initialization via Clustering. To learn the invariant covariate representation under the invariant risk minimization framework, we require an initial partition of the training data into distinct "environments". Thus, for faster convergence, we perform k-means clustering on $h_{flat,1}, \ldots, h_{flat,n}$ to partition the training data into E clusters. Each cluster constitutes an initial environment, based on the hypothesis that samples with similar high-level spatiotemporal covariates may share common spurious characteristics.

Iterative Environment Partitioning and Mask Generation. We learn the optimal mask m through an iterative process that alternates between refining the environments and updating the mask, as inspired by Liu et al. [18] and Du et al. [4].

1. **Environment Partition:** Holding the mask m fixed, we refine the environment assignments. We first train E simple, environment-specific classifiers, $\Gamma^{(e)}$, using only the **variant** representation Ψ as input. The goal is for each classifier to specialize in the spurious correlations of its environment using the following loss

$$\arg\min_{\Theta_e} \mathcal{L}\big(\Gamma^{(e)}(h_{flat,i} \cdot \Psi_i \mid \Theta_e)\big|R_e\big),$$

where R_e means the training samples in the corresponding environment e and $\Psi_i = (1 - m) \cdot h_{flat,i}$.

After training, we re-assign each training sample to the environment e whose classifier $\Gamma^{(e)}$ predicts its seizure type most accurately

$$e(i) = \arg\max_{e\in\mathcal{E}} \Gamma^{(e)}(h_{flat,i} \cdot \Psi_i \mid \Theta_e).$$

This step clusters samples based on their shared spurious covariates. This process is repeated until the environment assignments converge.

2. **Mask Generation:** With stable environment partitions $\{\mathcal{R}_e|e \in E\}$, we update the invariant mask m. The objective is to find a mask that minimizes the prediction variance of a model trained on the **invariant** representation Φ across all environments. This is framed as an Invariant Risk Minimization problem. The loss function to optimize m is:

$$\mathcal{L}_{mask} = \mathbb{E}_{e\in E}\mathcal{L}^e + \alpha \left\|\mathrm{Var}_{e\in E}(\nabla_{\Theta^{mask}}\mathcal{L}^e)\right\|^2 + \lambda\|m\|^2.$$

Here, $\mathcal{L}^e$ is the loss of a shared model on environment e. The first term optimizes for average performance, while the second term (with hyperparameter α) penalizes gradient variance, pushing the model to learn from covariates that are equally predictive in all environments. The third term is an L_2 regularization on the mask.

These two procedures are alternated, progressively refining both the environment groupings and the invariant mask m.

Final Prediction with MLP. After the iterative process converges, the optimized invariant mask m is fixed. We compute the final invariant representation $\Phi_{final} = m \odot h_{flat}$ for all data. This representation, theoretically stripped of patient-specific noise, is then used to train a final, lightweight Multi-Layer Perceptron (MLP) as the seizure classifier with the cross-entropy loss.

	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
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
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
PInvCRL (ours)	$\mathbf{0.740 \pm 0.040}$	$\mathbf{0.765 \pm 0.015}$	$\mathbf{0.767 \pm 0.050}$	$\mathbf{0.709 \pm 0.030}$	$\mathbf{0.739 \pm 0.019}$	$\mathbf{0.744 \pm 0.030}$

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

3 Experiments

3.1 Experimental Settings

Datasets. We conduct experiments on the Temple University Hospital EEG Seizure Corpus (TUSZ) v1.5.2, following the experimental settings in previous works [16, 23, 28]. As the largest public EEG dataset for seizure research, TUSZ contains 5,612 EEG recordings collected from over 300 patients. EEG signals are recorded using 19 electrodes placed according to the international 10-20 system [10].

Data Preprocessing and Experimental Protocols. Our preprocessing procedure is consistent with that used in Tang et al. [27] and other recent works [2]. Initially, all EEG recordings are resampled to 200 Hz and segmented into fixed-length clips of 12 seconds and 60 seconds without overlap. Subsequently, each clip is subdivided into 1-second non-overlapping segments. Covariate extraction is performed by applying the Fast Fourier Transform (FFT) to every 1-second window for frequency components, following the method in Tang et al. [27]. For the evaluation, the model performance is primarily evaluated using the weighted F1-score, precision, and recall. We adopt the same range in [30] for the hyperparameter tuning.

Baselines. We consider the following existing baselines. For convolutional models, we include **DenseCNN** [22]; for recurrent models, we include **LSTM** [9]; and for hybrid convolution-recurrent approaches, we include **CNN-LSTM** [2]. Furthermore, we evaluate our method in comparison with graph-based methods, including **MSTGCN** [11], **NeuroGNN** [8], **DCRNN** [27], and **PANN** [31].

3.2 Performance Analysis

Table 1 presents the primary results of our experiments. Our proposed PInvCRL framework, with its final MLP classifier, consistently outperforms all baseline models for both 12-second and 60-second clip lengths. The DCRNN baseline already demonstrates a strong performance compared to non-graph methods, affirming the utility of graph-based spatiotemporal modeling. Most importantly, our PInvCRL-MLP model achieves a performance gain over the DCRNN baseline. This performance leap directly validates our central hypothesis: by explicitly modeling and removing patient-variant information, the final classifier learns from a more robust and generalizable covariate, leading to superior cross-patient performance. The improvement is consistent across different clip lengths, demonstrating the framework's effectiveness in disentangling invariant and variant representations from complex spatiotemporal covariates.

4 Conclusion

In this paper, we introduced PInvCRL, a novel framework for cross-patient seizure classification. By adapting and concretizing principles of invariant representation learning, we have developed a method that can explicitly decompose learned EEG covariates into patient-invariant and patient-variant components. Our two-stage approach—combining a powerful DCRNN encoder, an iterative environment-based learning procedure initialized by k-means clustering, and a final MLP classifier—effectively isolates the core, generalizable biomarkers of seizure types from confounding patient-specific patterns. Experimental results on the large-scale TUSZ dataset confirm that training a classifier on these invariant representations leads to improvements in performance over baselines. One of the potential limitations is that this paper flattens the representation learned by DCRNN, rather than considering the spatiotemporal information explicitly.

References

[1] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using EGG signals. *Computers in biology and medicine*, 100:270–278, 2018.

[2] David Ahmedt-Aristizabal, Tharindu Fernando, Simon Denman, Lars Petersson, Matthew J Aburn, and Clinton Fookes. Neural memory networks for seizure type classification. In *The IEEE Engineering in Medicine and Biology Society*, 2020.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. Invariant representation learning for multimedia recommendation. In *the ACM International Conference on Multimedia*, 2022.

[5] Robert S Fisher, Walter Van Emde Boas, Warren Blume, Christian Elger, Pierre Genton, Phillip Lee, and Jerome Engel Jr. Epileptic seizures and epilepsy: Definitions proposed by the international league against epilepsy and the international bureau for epilepsy. *Epilepsia*, 46(4): 470–472, 2005.

[6] Robert S Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J Helen Cross, Christian E Elger, Jerome Engel Jr, Lars Forsgren, Jacqueline A French, Mike Glynn, et al. Ilae official report: A practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, 2014.

[7] Meysam Golmohammadi, Saeedeh Ziyabari, Vinit Shah, Iyad Obeid, and Joseph Picone. Deep architectures for spatio-temporal modeling: Automated seizure detection in scalp EEGs. In *IEEE International Conference on Machine Learning and Applications*, 2018.

[8] Arash Hajisafi, Haowen Lin, Yao-Yi Chiang, and Cyrus Shahabi. Dynamic gnns for precise seizure detection and classification from EEG data. In *Pacific Asia Knowledge Discovery and Data Mining*, 2024.

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

[10] Richard W Homan, John Herman, and Phillip Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and Clinical Neurophysiology*, pages 376–382, 1987.

[11] 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*, pages 1977–1986, 2021.

[12] Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O'Brien, et al. Epileptic seizure prediction using big data and deep learning: Toward a mobile system. *EBioMedicine*, 27:103–111, 2018.

[13] Xiang Li, Shunpan Liang, Yulei Hou, and Tengfei Ma. Stratmed: Relevance stratification between biomedical entities for sparsity on medication recommendation. *Knowledge-Based Systems*, 284:111239, 2024.

[14] Xiang Li, Shunpan Liang, Yu Lei, Chen Li, Yulei Hou, Dashun Zheng, and Tengfei Ma. Causalmed: Causality-based personalized medication recommendation centered on patient health state. In *ACM International Conference on Information and Knowledge Management*, 2024.

[15] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

[16] 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*, 2020.

[17] Yang Li, Yu Liu, Yu-Zhu Guo, Xiao-Feng Liao, Bin Hu, and Tao Yu. Spatio-temporal-spectral hierarchical graph convolutional network with semisupervised active learning for patient-specific seizure prediction. *IEEE Transactions on Cybernetics*, 52(11):12189–12204, 2021.

[18] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, 2021.

[19] Yanyi Peng and Xiang Li. Enhancing the model robustness and generalization in sentence sentiment classification based on causal representation learning techniques. In *AAAI Workshop on Artificial Intelligence with Causal Techniques*, 2025.

[20] Shivarudhrappa Raghu, Natarajan Sriraam, Yasin Temel, Shyam Vasudeva Rao, and Pieter L Kubben. EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Networks*, 124:202–212, 2020.

[21] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Chrononet: A deep recurrent neural network for abnormal EEG identification. In *Artificial Intelligence in Medicine*, pages 47–56. Springer, 2019.

[22] 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.

[23] 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*, page 102106, 2020.

[24] Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018.

[25] Ali Hossam Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[26] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11 (3):532–541, 2018.

[27] 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.

[28] 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*, pages 5547–5557, 2022.

[29] Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2020.

[30] Yuntian Wu, Yuntian Yang, Jiabao Sean Xiao, Chuan Zhou, Haochen Sui, and Haoxuan Li. Invariant spatiotemporal representation learning for cross-patient seizure classification. In *NeurIPS 2024 Workshop NeuroAI*, 2024.

[31] 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*, page 105664, 2024.