Distilling Invariant Representations With Domain Adversarial Learning for Cross-Subject Children Seizure Prediction

Ziye Zhang, Aiping Liu, *Member, IEEE*, Yikai Gao, Xinrui Cui, Ruobing Qian, and Xun Chen, *Senior Member, IEEE*

***Abstract*—Seizure prediction based on electroencephalogram (EEG) has great potential to improve patients' life quality. Due to the high heterogeneity in distributions of EEG signals among different patients, conventional studies usually show poor generalization ability when transferring the model to new patients, which also leads to difficulties in clinical applications. To alleviate the challenging issue concerning cross-subject domain shift, we propose a transformer-based domain adversarial model. Our model first adopts a pretrained general neural network to extract common features from the EEG signals of available patients. Then, we design a distiller module and a domain discriminator module to perform domain adaptation training based on a small amount of labeled data from the new-coming patient. During the adaptation process, conditional domain adversarial training with the addition of label information is employed to remove patient-related information from the extracted features to learn a common seizure feature space among different patients. Our proposed seizure prediction method is evaluated on the CHB-MIT EEG database. The proposed model achieves a sensitivity of 79.5%, a false alarm rate (FPR) of 0.258/h, and an AUC of 0.814. Experimental results demonstrate that the proposed method can effectively reduce interpatient domain disparity compared to state-of-the-art methods.**

***Index Terms*—Adversarial learning, domain adaptation (DA), electroencephalogram (EEG), seizure prediction, transformer.**

Manuscript received 29 November 2022; revised 24 February 2023; accepted 9 March 2023. Date of publication 14 March 2023; date of current version 12 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 32271431, Grant 61922075, and Grant 82272070; in part by the Research Project of Health Commission of Anhui Province under Grant AHWJ2022b004; in part by the Fundamental Research Funds for the Central Universities under Grant KY2100000123; and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-025. *(Corresponding author: Xun Chen.)*

Ziye Zhang, Aiping Liu, and Yikai Gao are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China.

Xinrui Cui is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

Ruobing Qian is with the Epilepsy Center, Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230001, China.

Xun Chen is with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, and the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230001, China (e-mail: xunchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TCDS.2023.3257055

I. Introduction

EPILEPSY is a common brain disorder that arises from abnormal excessive or synchronous neuronal activity in the brain. It affects approximately 50 million people globally, about half of whom have coexisting physical or psychiatric conditions [1], [2]. Seizures are debilitating, disrupt daily activities, and are associated with an increased risk of premature death [3], [4], [5]. Therefore, the accurate prediction of impending seizures is of great importance to guide physicians and patients to take preventive measures to avoid the harm caused by seizures.

The electroencephalogram (EEG) is a representative signal containing information about the electrical activity of the brain and is also used as an important tool for clinical diagnosis and analysis of epilepsy [6], [7], [8]. Since the predictability of human seizures has been proven in 1975 [9], many outstanding studies have emerged on this issue [10], [11], which also inspired a host of subsequent EEG-based studies on seizure prediction [12], [13]. The majority of current studies classify continuous epileptic EEG signals into four states: 1) ictal (during a seizure); 2) preictal (the period before a seizure); 3) post-ictal (the period after a seizure); and 4) interictal (the period between post-ictal and preictal). We can, therefore, generally convert seizure prediction into a binary classification problem, i.e., distinguishing between preictal and interictal periods. Specifically, if the state is detected as preictal, an alert is issued to warn the patient to take precautionary measures. To distinguish the different EEG states, the previous method consists of three main steps, including preprocessing, feature extraction, and classification [3]. Among them, feature extraction is the most critical step, aiming to extract a discriminative biomarker from the original signal.

Traditional methods design and extract features, including spatiotemporal correlations [14], phase-synchronous patterns [15], spectral power features [16], autoregressive models [17], empirical pattern decomposition [18], and bivariate measures such as intercorrelation [19]. Recently, researchers tend to use deep learning techniques to learn the underlying features of brain signals [20], [21], [22]. While traditional methods often require the construction of complex handcrafted features, deep learning can learn features directly from data, thus, enabling the mining of complex patterns from brain signals to optimize seizure prediction performance.

In [20] and [21], the EEG signal transformed into a time–frequency map by short-time Fourier transform (STFT) was used as an input to a convolutional neural network (CNN). In [23], the wavelet transform was applied to the EEG signal and the resulting wavelet tensor was fed into the CNN to identify preictal patterns. These deep learning models, however, focus only on local information for long sequence processing, while the self-attention mechanism can efficiently extract the global information of the signal without the limitation of sequence length. Based on the self-attention mechanism, Vaswani et al. [24] proposed the transformer model, which has great success in natural language processing and has been increasingly adopted in recent years for computer vision and time-series prediction tasks. Therefore, we considered its application in long-term EEG signal processing.

Currently, state-of-the-art seizure prediction is achieved by machine learning approaches in a patient-specific manner in which one model is trained per subject [22], [25]. Despite the importance of these trials for personalized medicine, their clinical application remains challenging due to the difficulty of replication. Specifically, since seizure generation mechanisms are highly heterogeneous among patients [26], existing models may perform well in one subject, but poorly in another. This high interpatient disparity also leads to the fact that the features obtained by conventional patient-independent methods contain a large amount of patient-related information and make them poor predictors of unseen patients. Moreover, most existing methods require large amounts of data to achieve satisfactory performance, which is difficult to meet in the clinical situation due to invasive surgery risks and privacy regulations.

The main challenge at present is to perform general seizure prediction with few samples of unseen patients. Therefore, we introduce the domain adaptation (DA) algorithm to address this problem, which is a machine learning technique that can alleviate domain shift. Specifically, this work considers existing patients as the source domain and new patients as the target domain, transforming the generalization problem into a DA problem. In some prevalent DA models [27], [28], [29], the minimization of interdomain distances in feature space is adopted as the main objective of optimization to improve the generalization ability of the model to the target domain. To the best of our knowledge, few studies have introduced DA algorithms to epileptic EEG. Existing works on the seizure prediction task using DA methods [30], [31] align the feature distributions extracted from the source and target domains with an artificially set prior distribution, however, the fixed prior distribution may be unreasonable. For seizure prediction tasks, feature extractors attempt to learn domain-invariant features by minimizing the domain disparity between patients. Nevertheless, due to the disparity between patients, there is redundancy in feature representation within the source domain, and the generalization ability over the target domain is affected during the alignment process.

Inspired by these studies, we propose a two-stage domain adaptive transformer-based (AdaTrans) seizure prediction approach. To extract the time–frequency information of the raw EEG signal, we use the STFT [32] for preprocessing, followed by training a feature extractor on the source domain. Then, a distiller is added after the feature extractor to eliminate the nonseizure information while ensuring the discriminability of the feature. A conditional discriminator is utilized to further ensure feature discriminability, and adversarial learning is employed to align the source and target domain distributions to finally obtain a better seizure prediction model for the target patient. Our contributions are as follows.

1) We propose a transformer-based DA seizure prediction framework, which achieves promising generalization ability with few samples and makes it feasible in clinical practice.
2) We utilize the distiller and conditional inputs in the DA process to ensure feature discriminability while removing patient-related information from the features. The disparity between the target and source domains is minimized by adversarial training to improve the generalization ability and prediction performance of the model.
3) We compare the performance with several state-of-the-art DA algorithms on a popular data set in the cross-subject seizure prediction task and achieve superior performance.

II. Methods

A. Data Set

This work uses the CHB-MIT [33], [34], a scalp EEG data set collected at Children's Hospital Boston for model evaluation. This data set records EEG signals from 23 patients with intractable epilepsy, with a total duration of nearly 983 h and containing 198 seizures. According to the international 10–20 system, most patients contained 23 channel recordings, and all these multichannel EEG signals were acquired at a sampling rate of 256 Hz. Following the setting in [21], we define the preictal period as 30 min before the seizure and the interictal period as at least 4 h before onset and at least 4 h after the seizure. In cases of two seizures within a 30-min period, only the preictal period of the preceding seizure is retained without evaluating the latter seizure. In addition, this work only considers patients with more than 2 and fewer than 10 seizures per day. On the one side, training requires sufficient data, on the other side, we believe that seizure prediction for patients with too frequent seizures is not very critical. Based on these definitions, we select 13 patients for the experiment, and a detailed description of the considered cases is listed in Table I.

B. Preprocessing

The EEG records in the CHB-MIT database were contaminated by power line noise at 60 Hz. Therefore, we exclude components in the 57–63-Hz and 117–123-Hz frequency ranges and the DC component (0 Hz). In addition, we rearrange the EEG records of each patient by electrode order and select the common 18 channels for the experiment. Due to the severe data imbalance between preictal and interictal periods, we divide the consecutive recordings into 50% overlapping 30-s EEG segments, which are then converted to time–frequency maps by STFT to generate more data.

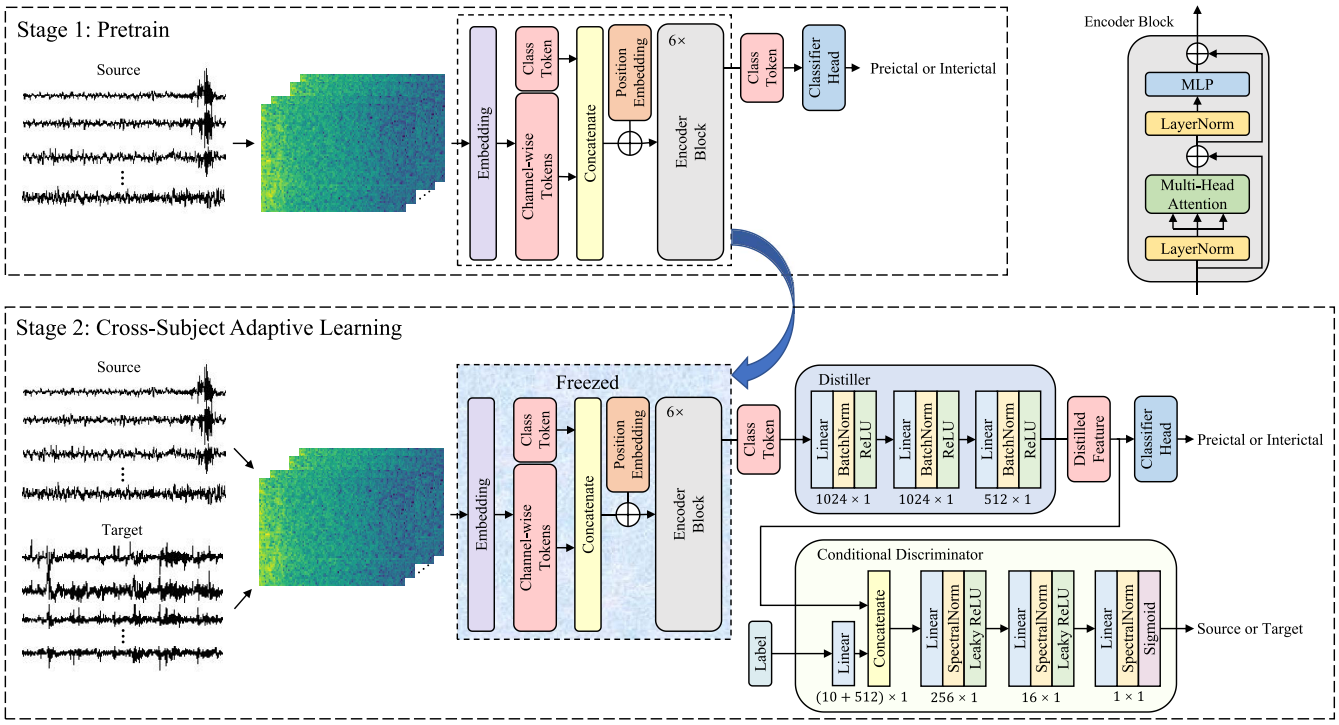

Fig. 1. Overall flow chart of our proposed training scheme and the architecture of the employed modules. In stage 1, the EEG signals of the source-domain patients are converted into time–frequency maps and fed to the transformer-based feature extractor for pretraining. In stage 2, we freeze the pretrained feature extractor and insert the distiller and conditional discriminator. Data from patients in the source and target domains are used as input for domain-adversarial training to learn the seizure-related common feature space. Note that the discriminator additionally inputs labels to constrain the domain-invariant features to contain seizure information. The architectures of the distiller, conditional discriminator, and encoder block are given in the figure.

TABLE I
DETAILED DESCRIPTION OF THE SELECTED PATIENTS IN THE CHB-MIT SCALP EEG DATABASE

Subject	Gender	Age	No. of seizures	Interictal hours
Pat1	F	11	7	17
Pat2	M	11	3	22.9
Pat3	F	14	6	21.9
Pat5	F	7	5	13
Pat9	F	10	4	12.3
Pat10	M	3	6	11.1
Pat13	F	3	5	14
Pat14	F	9	5	5
Pat18	F	18	6	23
Pat19	F	19	3	24.9
Pat20	F	6	5	20
Pat21	F	13	4	20.9
Pat23	F	6	5	3

C. Transformer-Based Domain Adaptation Seizure Prediction

In this section, we elaborate on the specific training process of the proposed two-stage approach, and the overall structure of AdaTrans is illustrated in Fig. 1. In stage one, we pretrain a transformer model using the source-domain data. Then, in stage two, we freeze the backbone of the model trained in stage one, connect the distiller after the encoder blocks to remove the patient-related information from the original features through conditional adversarial training in the source and target domains, and project the features into the domain-invariant subspace while maintaining feature discriminability.

1) Transformer: In recent years, the transformer has gained great success in natural language processing and has been increasingly adopted for computer vision and time-series prediction tasks. Some work has also demonstrated its effectiveness in the field of seizure prediction [35], [36]. In this work, we employ transformer as our backbone to analyze EEG signals and extract high-dimensional features. After the time–frequency transform, the original EEG signal becomes a 3-D matrix (T, C, F), where T, C, and F denote time, channel, and frequency, respectively. The standard transformer takes 3-D data (B, N, D) as input, where B, N, and D represent batch, the number of tokens, and the token dimension. We regard the time dimension T as the batch dimension B, and after two fully connected layers of embedding, each channel is regarded as a token while the frequency dimension F is projected to D dimension ($D = 512$ in this article). Then, following [37], we prepend a trainable class token to the sequence of embedded channelwise tokens for the subsequent classification task. To retain the position information, we add trainable 1-D position embeddings to the tokens and obtain embedded features, which are used as the input to the encoder block. The structure of the encoder block is similar to [37], where each block includes multihead self-attention consisting of MLP blocks and multiple self-attentions. Layernorm (LN) is applied before each block and residual connections after each block [38], [39]. The inputs of the self-attention layer are $\mathbf{q}$ (query), $\mathbf{k}$ (key), and $\mathbf{v}$ (value), and its output matrix is formulated as shown in the following:

$$\text{attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_j}}\right)\mathbf{v}. \tag{1}$$

Multihead self-attention goes a step further, allowing the model to focus on different positions and jointly attend to information from different representation subspaces, which is expressed as follows:

$$\text{multihead}(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad (2)$$

$$\text{head}_i = \text{attention}\left(\mathbf{q}W_i^q, \mathbf{k}W_i^k, \mathbf{v}W_i^v\right) \quad (3)$$

where h is the number of attention heads and W_i^q, W_i^k, and W_i^v are the learned projection matrices.

After the embedded input goes through the feature extractor f_e consisting of six encoder blocks, class token z_{cls} is extracted and fed into a classification head f_{cls} consisting of a single linear layer to predict whether it is preictal or interictal.

2) Adversarial Learning: In purpose of better transferring the pretrain model of stage one to the target domain, we introduce a module based on adversarial learning. Adversarial learning is an emerging deep learning approach that is used in the area of DA to learn domain invariant features. It usually contains a generator and a discriminator. The generator G will generate some fake inputs and try to confuse them with the real ones. The discriminator D will determine if the input samples are artificially generated. During training, the adversarial module finds the Nash equilibrium between G and D, and the distribution of the fake data progressively approximates the real data. The optimization objective of adversarial learning can be described as

$$\min_G \max_D \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}\left[\log D(\boldsymbol{x})\right] + \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_f}\left[\log\left(1-D(\tilde{\boldsymbol{x}})\right)\right] \quad (4)$$

where $\mathbb{P}_r$ is the real data distribution and $\mathbb{P}_f$ denotes the fake data distribution. After the above optimization process, the adversarial module can align $\mathbb{P}_f$ with $\mathbb{P}_r$.

To alleviate the training instability problem of conventional GAN, we use WGAN [40], [41] as our adversarial module in our experiments. One drawback of the conventional adversarial learning framework is the training instability problem, i.e., gradient disappearance in the generator if the discriminator is trained to the optimum in each training iteration. WGAN changes the optimization objective to the Wasserstein distance rather than the Jensen–Shannon divergence adopted by the original GAN. Wasserstein distance is used to measure the mass transportation cost from one distribution to another. In WGAN, the improved objective function can be written as

$$\min_G \max_{D\in\mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}\left[D(\boldsymbol{x})\right] + \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_f}\left[D(\tilde{\boldsymbol{x}})\right] \quad (5)$$

where $\mathcal{D}$ is a set of 1-Lipschitz functions designed to introduce Lipschitz continuity to improve the discriminator [41].

In the practical application of this work, we add spectral norm regularization to the discriminator so as to satisfy the 1-Lipschitz continuity [42]. The Lipschitz norm of the discriminator D is constrained to be less than one by doing the spectral normalization as shown in (6) to the weight matrix W for each layer of the discriminator D, so that the discriminator satisfies the 1-Lipschitz continuity

$$\hat{W_{\text{SN}}}(W) = W/\sigma(W) \quad (6)$$

where for each layer, the spectral norm $\sigma(W)$ is the maximum singular value of W.

In order to learn domain-invariant information, we hope to utilize the above principles in the feature space. We assume the existence of a universal seizure feature distribution $p(z)$ among patients. We define a feature distiller f_{dist} as the generator G and a conditional discriminator f_D. The class token z_{cls} produced by the feature extractor is further refined by f_{dist} and mapped to a patient-independent subspace to deceive the discriminator. The discriminator f_D determines whether the input distilled feature z_{dist} is from the source domain. In addition, we implicitly constrain the extracted domain invariant features to be seizure-related by providing labels to the discriminator. The specific network parameters are shown in Fig. 1.

3) Training and Testing Strategy: Let the entire data set be D, $D = \{X_1, X_2, \ldots, X_N\}$, where X_i denotes the EEG time–frequency maps of the ith patient after STFT transform. In order to better evaluate our method, we adopt the leave-one-patient-out (LOPO) principle by considering $N-1$ patients as the source domain and the remaining patient as the target domain. For a patient with p seizures, there are p matching preictal periods. We split the interictal records into p segments and concatenate them with the preictal records to obtain p seizure records. In stage one, we train a model on the source-domain data based on the cross-entropy loss l_{CE}

$$l_{\text{CE}} = -\left[y_{\text{src}}\log(f_{ada}(x_{\text{src}})) + (1-y_{\text{src}})\log(1-f_{ada}(x_{\text{src}}))\right] \quad (7)$$

where x_{src} denotes the source-domain input and y_{src} denotes the corresponding label.

During training, we select one seizure record from each patient in the source domain as the test set and the remaining data as the training set. Due to the limited available data set and to prevent the model from overfitting, we select 25% of the late samples from the preictal and interictal records in the training set as the validation set following [21], and the remaining samples are used for training.

In stage two, we freeze f_e and add the adversarial module consisting of f_{dist}, f_D. In purpose of learning the domain-invariant feature representation and extracting seizure-related information, we optimize our model by a zero-sum game between f_{dist} and f_D according to the following equation:

$$l_{\text{adv}} = \mathbb{E}(D(z_{\text{true}}, y_{\text{src}})) - \mathbb{E}\left(D\left(z_{\text{fake}}, y_{trg}\right)\right) \quad (8)$$

where z_{true} denotes the features obtained by distilling the source-domain data, z_{fake} denotes the features obtained by distilling the target domain data, and y denotes the corresponding label.

In the adversarial process, the classification loss is also added to ensure the prediction accuracy of the obtained features, so the optimization objective of the whole model can be written as

$$\min_{G,C}\max_D l_{\text{adv}} + l_{\text{CE}}. \quad (9)$$

In the clinical situation, despite the numerous available epileptic EEG data sets, the sample size of the target patient is usually not sufficient to train a satisfactory deep learning

network. Hence, this study seeks to perform seizure prediction based on a pretrain model of existing patients and a few target patient data. To simulate the extreme case of small sample size, only one seizure record of the target patient is used for training. To ameliorate the excessive data imbalance between the source and target domains, we downsample the source-domain data in stage two. Specifically, for each patient in the source domain, we randomly select their single seizure record for DA training in this stage. To improve the robustness of the evaluation, we use the leave-one-out cross-validation (LOOCV) strategy. For the target patient with p seizure records, $p-1$ seizure records are used as the test set and the remaining seizure records are used for training. The procedure is repeated p times until each seizure record has been used for training and the average result is calculated.

D. Postprocessing

In this work, we use a k-of-n approach [21] to eliminate the effect of isolated false-positive predictions occurring during interictal periods. In particular, the alarm is set for the presence of at least k positive predictions in n consecutive windows. In our experiments, we set $k = 8$ and $n = 10$, which means that an alarm is set if at least 240 s are positive predictions during the last 300 s. In addition, to prevent alarms from being triggered continuously within a short period, we have set the refractory period to 30 min.

III. Experiments and Results

To evaluate the effectiveness of AdaTrans, we perform extensive experiments on the EEG recordings of 13 patients with 64 seizures in the CHB-MIT scalp EEG data set and compare them with conventional methods and several state-of-the-art DA methods. In this section, we will present the evaluation metrics, experimental setup, and results.

A. Evaluation Metrics

Before introducing the evaluation metrics, the seizure occurrence period (SOP) and the seizure prediction horizon (SPH) need to be defined. SOP is the period during which a seizure is predicted, while SPH is the duration between a seizure alarm and the start of SOP [43]. If an alarm is received during the SPH period but no seizure is observed in the SOP, it is considered a false alarm. Otherwise, the prediction is correct. In this study, SPH is set to 5 min, and SOP is set to 30 min.

This work considers two major metrics: 1) sensitivity (Sen, the proportion of seizures correctly predicted) and 2) false alarm rate (FPR, the average number of false alarms per hour) to evaluate the model performance [21]. Existing seizure prediction tasks can often be divided into two categories: 1) segment based and 2) event based, and we believe that event-based prediction is closer to clinical situations and more instructive in practical applications. Specifically, the formulas for sensitivity and FPR in the event-based situation can be written in the following form:

$$\text{sensitivity} = \frac{\text{TP}}{\text{FN}+\text{TP}} \times 100\% \tag{10}$$

TABLE II
Prediction Performance Comparison of Our Proposed Approach With Conventional Methods on the CHB-MIT Data Set

Target	Patient-Independent		Finetune		AdaTrans	
	Sen(%)	FPR(/h)	Sen(%)	FPR(/h)	Sen(%)	FPR(/h)
Pat1	71.4%	1.591	73.8%	0.383	95.2%	0.246
Pat2	33.3%	0.348	33.3%	0.000	66.7%	0.000
Pat3	33.3%	0.864	47.2%	0.318	60.0%	0.155
Pat5	0.0%	0.000	35.0%	0.429	70.0%	0.375
Pat9	25.0%	0.324	33.3%	0.497	91.7%	0.793
Pat10	50.0%	0.500	50.0%	0.400	80.0%	0.254
Pat13	60.0%	0.072	60.0%	0.054	85.0%	0.072
Pat14	20.0%	0.200	40.0%	0.701	70.0%	0.501
Pat18	33.3%	0.292	36.7%	0.008	53.3%	0.050
Pat19	66.7%	0.960	66.7%	0.620	66.7%	0.160
Pat20	100.0%	0.200	100.0%	0.088	100.0%	0.013
Pat21	75.0%	0.227	100.0%	0.242	100.0%	0.197
Pat23	40.0%	0.233	90.0%	0.739	95.0%	0.544
Avg.	46.8%	0.447	58.9%	0.345	79.5%	0.258

[1] Sen, sensitivity; FPR, false alarm rate; Avg., average result.

$$\text{FPR} = \frac{\text{FP}}{\text{time(interictal)}} \tag{11}$$

where TP indicates the number of correctly predicted seizures, FN indicates the number of seizures that failed to be predicted, and FP indicates the number of false alarms.

In addition, this work uses the area under the receiver operating characteristic curve (AUC) for model evaluation.

B. Experimental Setup

In this section, we present the experimental setup. In stage one, adopted with the SGD optimizer, the learning rate λ_{pre} is set to 0.001, the weight decay is set to $5e-5$, and the maximum training epoch is set to 100. In stage two, we use the RMSprop optimizer to train the adversarial module and use the Adam optimizer to train the classification head f_{cls} with momentum parameters setting to $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The maximum learning rates of the two optimizers are set to $\lambda_{gan} = 1e-3$ and $\lambda_{\text{cls}} = 1e-3$, respectively. The maximum training epoch is set to 300. We use the Python library and the Pytorch framework to perform our experiments on a server equipped with NVIDIA 3080 GPUs.

C. Results

1) Compared With Conventional Methods: To demonstrate the superiority over conventional methods, we compare with two cross-subject seizure prediction approaches: 1) patient-independent and 2) finetune. These approaches are often used when considering the generalization ability of the model to new patients, but often have poor performance limited by the high degree of heterogeneity between patients. Here, based on our proposed transformer feature extractor, we use EEG data from defined source-domain patients for training and use EEG data from target domain patients for testing. For the patient-independent approach, we directly test the model f_{pre} obtained by training on the source domain with target domain data. For the fine-tuning approach, we fix the embedding layer and encoder blocks of f_{pre}, and then follow LOOCV strategy, fine-tuning the classification head f_{cls} with one seizure record of the target patient and testing it with the remaining seizure records. The sensitivity and FPR of these approaches are shown in Table II. The AUC values for each patient are shown in Fig. 2.

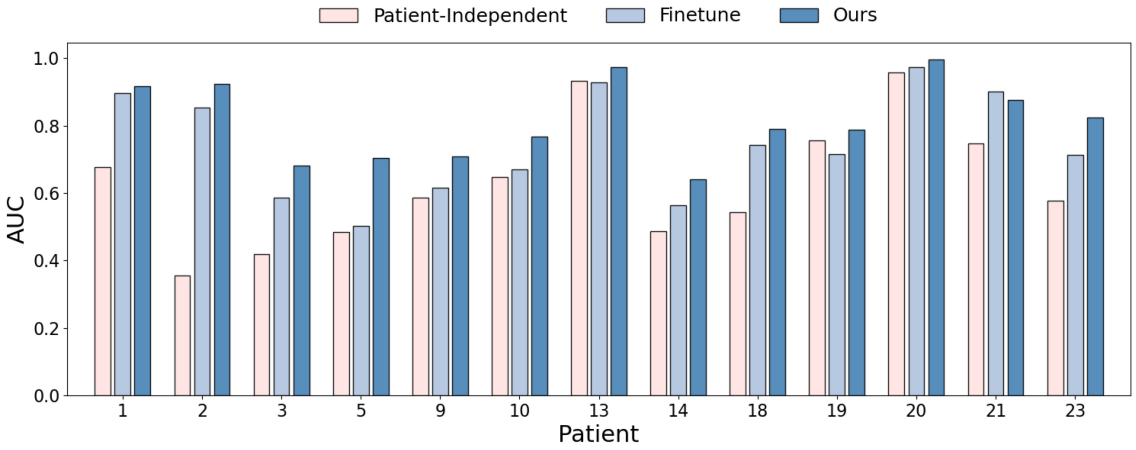

Fig. 2. Comparison of different conventional methods in the CHB-MIT sEEG data set. AUC, the area under the receiver operating characteristic curve.

TABLE III
PREDICTION PERFORMANCE COMPARISON WITH EXISTING DA METHODS ON THE CHB-MIT DATABASE

Target	DAN [27]		DAAN [28]		DSAN [29]		AdaTrans	
	Sen(%)	FPR(/h)	Sen(%)	FPR(/h)	Sen(%)	FPR(/h)	Sen(%)	FPR(/h)
Pat1	90.5%	0.304	73.8%	0.167	85.7%	0.265	95.2%	0.246
Pat2	50.0%	0.000	33.3%	0.000	33.3%	0.000	66.7%	0.000
Pat3	6.7%	0.000	33.3%	0.227	26.7%	0.182	60.0%	0.155
Pat5	95.0%	1.662	55.0%	1.108	60.0%	1.215	70.0%	0.375
Pat9	50.0%	0.879	25.0%	0.303	58.3%	0.836	91.7%	0.793
Pat10	43.3%	0.054	23.3%	0.069	36.7%	0.108	80.0%	0.254
Pat13	70.0%	0.036	70.0%	0.072	90.0%	0.107	85.0%	0.072
Pat14	60.0%	0.450	60.0%	0.450	60.0%	0.450	70.0%	0.501
Pat18	26.7%	0.000	43.3%	0.033	53.3%	0.050	53.3%	0.050
Pat19	0.0%	0.000	83.3%	0.840	33.3%	0.100	66.7%	0.160
Pat20	90.0%	0.050	90.0%	0.050	85.0%	0.050	100.0%	0.013
Pat21	100.0%	0.182	100.0%	0.258	100.0%	0.227	100.0%	0.197
Pat23	100.0%	0.408	95.0%	0.816	100.0%	0.719	95.0%	0.544
Avg.	60.2%	0.310	60.4%	0.338	63.3%	0.332	79.5%	0.258

Table II shows that the fine-tuning approach is better than the patient-independent approach, and our model outperforms every other conventional approach to a great extent. It is reasonable that the existing patient-independent methods do not have a high generalization ability due to the high heterogeneity among patients, while fine-tuning uses the target domain data to be able to adjust the classification boundaries to some extent and achieves better performance. Based on DA methods, AdaTrans shows a significant advantage in generalization ability with a sensitivity of 79.5% and an average FPR of 0.258/h.

2) Compared With DA Methods: Further, we compare the proposed approach with existing DA methods. However, few studies of DA methods have been proposed in the seizure prediction task. Therefore, to evaluate the effectiveness of our proposed model, we have to adopt DA methods from other fields. Specifically, we implement three prevalent DA methods: deep adaptation network (DAN) [27], dynamic adversarial adaptation network (DAAN) [28], and deep SubDA network (DSAN) [29]. The sensitivity and FPR are shown in Table III. The AUC values for each patient are shown in Fig. 3. Among these methods, AdaTrans achieves the best performance. A detailed discussion of these DA methods is presented below.

DAN: DAN is a representative work in the field of deep transfer learning that introduces deep networks for DA. It uses a multikernel MMD proposed by Arthur Gretton to replace the original single-kernel MMD [27]. The original MMD is to map the source and target domain in a regenerative kernel Hilbert space (RKHS) with an identical mapping and then calculate the mean discrepancy between two domains as the optimization target. As for MK-MMD, it proposes to use multiple kernels to construct the overall kernel, which solves the problem of the restricted kernel function of MMD. As can be shown in Table III, the results of DAN are the worst compared to DAAN and DSAN because it only focuses on the alignment of domain marginal distributions, which impairs the discriminative ability of the features in the process of adaptation. It is apparent from Pt3 and Pt19 that the alignment destroys the discriminative hyperplane of the features due to the excessive gap between different patient domains, resulting in poor prediction performance.

DAAN: DAAN is a deep DA method that considers aligning both the marginal probability distribution and the conditional probability distribution of source and target domains [28]. It can dynamically and quantitatively analyze the relative importance of both for transfer learning. Specifically, DAAN aligns the marginal distributions based on a global discriminator, expects to align the conditional distributions based on multiple subdiscriminators, and adaptively adjusts the weights of the marginal and conditional distributions in the objective function based on the loss of the classifier. From the results, DAAN predicts better than DAN because it takes into account the

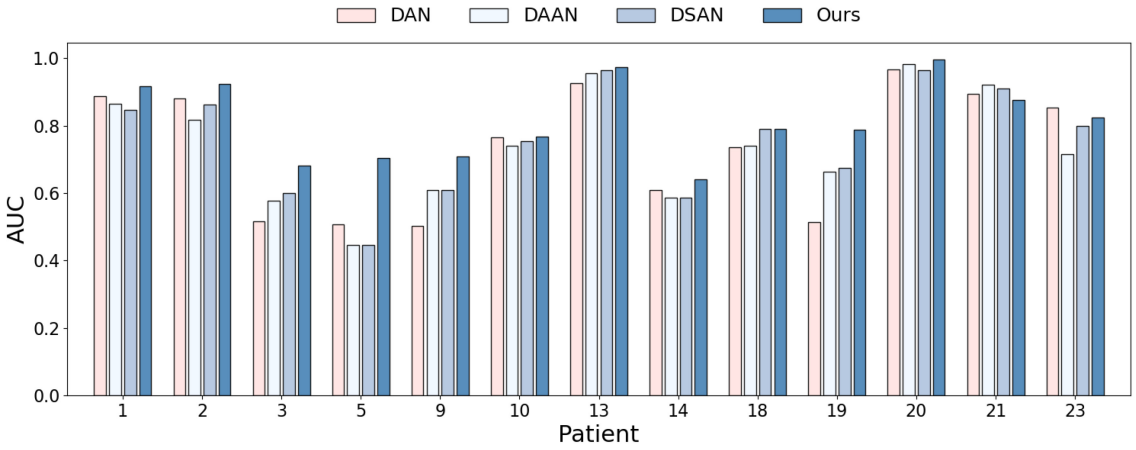

Fig. 3. Comparison of different DA approaches in the CHB-MIT sEEG data set. The AUC is selected as the evaluation metric.

effect of conditional distributions of different classes, but the prediction performance is also not satisfactory because the adversarial process may destroy the feature space obtained from the pretraining process.

DSAN: Based on the idea of subdomain adaptation, DSAN uses a local MMD (LMMD) to align the distributions, achieving excellent results in recent years in the DA area [29]. DSAN puts the same class of data into a subdomain by using class labels as a division basis. Then, instead of focusing on global alignment, the relevant subdomains are aligned separately. This approach maximizes the discriminative ability of the features and relatively well alleviates the problem of excessive differences between patients in seizure prediction tasks. Experimental results show that DSAN outperforms DAN and DAAN in prediction, achieving a performance second only to our method. The poor performance of DSAN on some patients may be due to aligning only the subdomains and ignoring the overall fine-grained information, which is prone to overfitting for few samples in the target domain.

Based on the above results, we presume that the adversarial learning-based framework is relatively superior in alleviating patient heterogeneity, as all the methods based on adversarial learning have achieved good performance and generalization ability in seizure prediction. The results of DAAN and DSAN show that the idea of subdomain alignment can further improve the generalization ability of the models. AdaTrans achieves the best performance compared to the existing popular DA methods by fine-tuning the feature distribution through the distiller while the conditional information input comes to implicitly constrain the alignment of conditional distributions, ensuring that the features still have good discriminative ability during the adaptation process. In the experiments, although the data vary greatly between different patients, we do not distinguish them within the source domain, which also leads to great redundancy in the source domain for subsequent DA training. In the field of DA, on the other hand, there exists some work focused on multisource DA, such as [44], which may be a better match for seizure prediction tasks. In our future work, we will further explore these methods to understand the correlation weights between different domains and work on achieving better model generalization ability.

TABLE IV
ABLATION STUDY ON THE DIFFERENT COMPONENTS

Method	Sen(%)	FPR(/h)	AUC
w.o. distiller	61.6%	0.269	0.772
w.o. cGAN	72.6%	0.345	0.777
Ours	**79.5%**	**0.258**	**0.814**

3) Ablation Study and Visualization: To explore the effect of different components in AdaTrans on the prediction results, we perform ablation experiments under the constraints of the fixed experimental setup. This work discusses two components in AdaTrans: 1) the feature distiller f_{dist} in the adversarial process and 2) the conditional input of the discriminator. The experimental results are listed in Table IV, where w.o. distiller means removing f_{dist} during the DA process, and w.o. cGAN means to romove the conditional input of the discriminator f_D. From the experimental results, it can be seen that removing either the refining operation in the adversarial process or the conditional input of the discriminator leads to performance degradation in the CHBMIT data set. In particular, we find that removing f_{dist} brings a drastic performance degradation, which indicates that this module effectively removes patient-related information from the features by adjusting the feature space during DA. It also demonstrates that f_{dist} has a significant impact on the prediction performance and the generalization of the model. The conditional input of the model constrains the domain alignment process in stage two to retain seizure-related information, which also contributes to ensuring the discriminative ability of the features and improving the model performance to some extent from the results.

To better visualize the effect of f_{dist}, we present the t-SNE plots [45] of Pat2 before and after f_{dist}. As shown in Fig. 4, different scatter colors represent the training and test sets, and different scatter shapes represent different feature categories. It can be seen that the original feature distributions shown in the left panel confuse the preictal and interictal data, while the distilled feature distributions shown in the right panel distinguish the features of the two categories well. This further demonstrates the effectiveness of f_{dist} for improving model performance.

TABLE V
RESULTS OF RECENT DA METHODS IN THE SEIZURE PREDICTION TASK ON THE CHB-MIT SEEG DATA SET

Method	Dataset	Features	Processing	No. of seizures	Sen(%)	FPR(/h)	AUC	Interictal Distance[1]	SOP	SPH
MMD-AAE [30]	MIT, 16 patients	STFT spectral images	segment-based	74	73%	0.24	0.75	30min	30min	0
MAAE [31]	MIT, 10 patients	Riemannian manifold feature	segment-based	54	82%	0.13	0.84	30min	30min	0
This work	MIT, 13 patients	STFT spectral images	event-based	64	79.50%	0.258	0.814	240min	30min	5min

[1] Interictal Distance is the period between interictal and an onset.

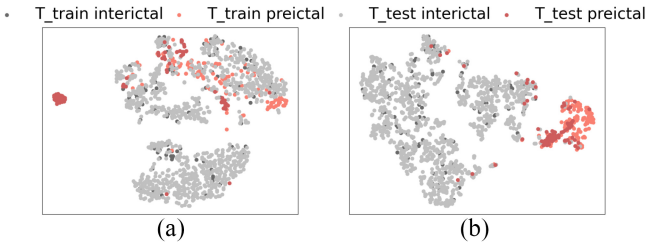

Fig. 4. Visualizations of features before and after f_{dist} with the t-SNE technique. ("T" indicates the target domain, "train" is the training set, and "test" is the test set.) (a) Original target. (b) Distilled target.

To show the prediction process of AdaTrans more intuitively, Fig. 5 compares the prediction outcomes of our method for Pat1, Pat13, Pat18, and Pat20 with their labels. Overall, AdaTrans can predict the seizures of the target patients relatively well, but there are still problems of omission and false alarms. For the majority of seizures, the model successfully predicts and accurately locates the onset of preictal period. However, AdaTrans have omission and false alarm issues for some patients, such as Pat13 and Pat18, which indicates that despite the acceptable sensitivity achieved by our method, there is still space to improve the feature discriminability within the target domain.

IV. DISCUSSION

This work proposes a seizure prediction method based on DA, which aims to alleviate the problem of high variability in patient data during model transferring. In general, patient-specific models perform better than patient-independent models, but the small amount of data in clinical settings makes it difficult to meet their training requirements. In addition, the conventional fine-tuning or patient-independent methods are limited by the domain gap between patients and perform poorly. Our approach is based on patient-independent models pretrained on existing patients, using a small amount of data from unknown patients for DA to obtain a common seizure feature space. Conditional adversarial training is employed to extract domain-invariant features while maintaining feature discriminative ability, and patient-related information is removed by the distiller f_{dist} to achieve excellent generalization ability and prediction performance on the target patients. According to Table II and Fig. 4, it can be assumed that our method improves the performance on the target domain by eliminating the patient-related information in the patient-independent model through conditional domain adversarial while retaining the common seizure information.

Compared with existing conventional and DA methods, AdaTrans obtains a benefit of about 21% and 16%, respectively. The experimental results in Table II show that the introduction of DA techniques allows the model to have a better generalization ability than conventional methods. Moreover, the comparison with several state-of-the-art DA algorithms shows the effectiveness of conditional adversarial learning and the elimination of patient-related information for seizure prediction. As shown in Table III, AdaTrans achieves better sensitivity compared to DAN and DSAN, which suggests that adversarial learning may be more suitable for epileptic EEG compared to kernel methods. Also, DAAN and DSAN can better maintain the discriminative ability of features due to the consideration of conditional distributions, which is further ensured by our approach and better performance is obtained. We also note that these DA methods perform significantly worse than AdaTrans, which may be due, on the one hand, to the fact that these methods are less suitable for processing EEG and, on the other hand, to the fact that these methods directly fine-tune the feature extractor and excessively corrupt the feature space over the source domain. Based on the results, our two-stage approach may be more applicable to this clinical limited data situation and may be able to alleviate the problem of catastrophic forgetting in the source domain.

According to the experimental results, we find that the existing classical DA methods are not well adapted to the field of epilepsy. During our research, we also find a few works that applied DA to seizure prediction tasks. Peng et al. [30] used the MMD-AAE method to align each patient's feature distribution with an artificial prior distribution in an attempt to obtain a common feature space. Based on [30], Peng et al. [31] introduced Riemannian manifolds to compute a prior distribution for alignment. These works have achieved satisfactory results, but the prior distribution in which the seizure features are artificially specified may not seem reasonable. These works used the reconstruction loss of EEG signals as part of the objective function, which we believe is detrimental to the extraction of common seizure features. The reconstruction objective function aims to preserve the integrity of the signal, including patient-related information in it, which is detrimental to the generalization ability of the model. However, the vague representation of data selection strategies and experimental procedures makes it difficult to reproduce their work. Moreover, the previous works are segment-based, and, thus, cannot be directly compared. We list these methods and results in Table V.

Finally, we discuss the limitation of this study and future works. In this work, we did not consider the recognition differences between channels, which is an important issue in this field. Several studies in recent years have demonstrated that the addition of channel selection technique can lead to better

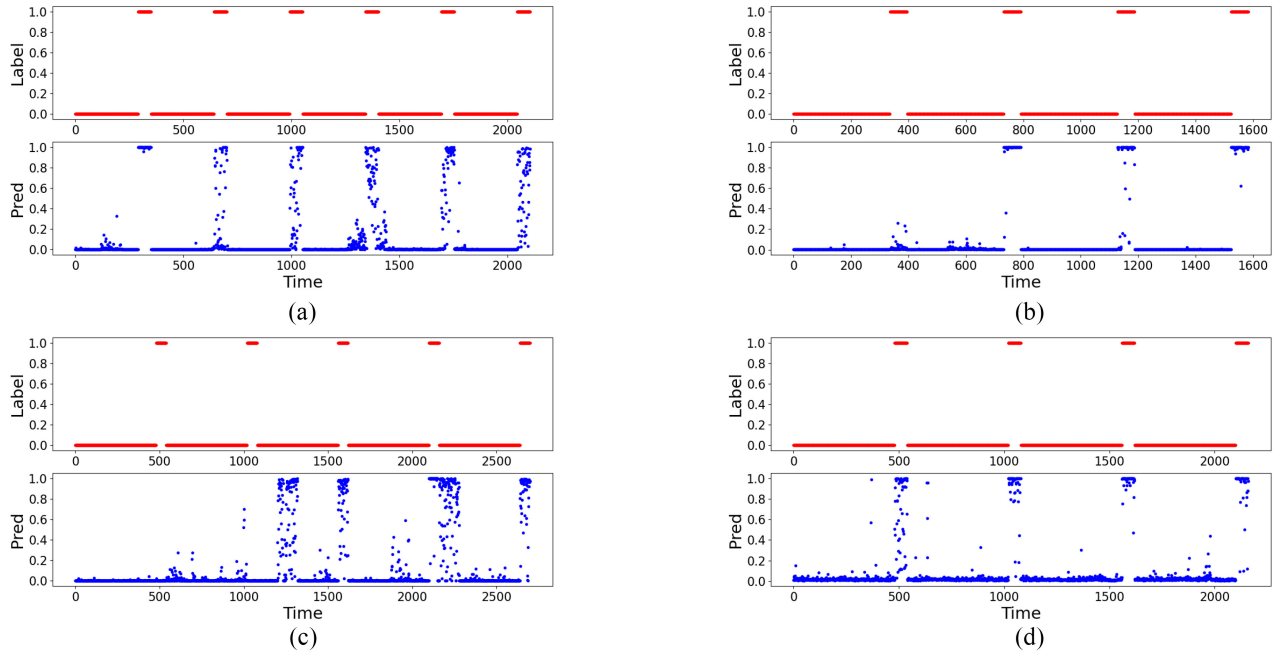

Fig. 5. Comparison of labels and prediction outcomes in testing. ("Pred" indicates the prediction results, and the subtitle shows the target patient corresponding to the subfigure.) (a) Result tested on Pat 1. (b) Result tested on Pat 13. (c) Result tested on Pat 18. (d) Result tested on Pat 20.

prediction [46], [47]. In the future, we will consider the combination of this technique with DA for seizure prediction. The present method is based on adversarial learning, which inherits the associated drawbacks, especially, the instability of training. This study is evaluated on the CHB-MIT data set, which is relatively small for deep learning. In future work, we will consider evaluating our method using a larger data set and further optimizing the network structure to effectively use limited data for learning. Although we artificially fix preictal as in the mainstream paradigm, in the future, we will consider adaptive preictal selection, which might lead to more precisely data classification. In addition, the division within the source domain in this study is relatively coarse and does not consider the relevance of DA by different patients for unknown patients. Considering the high heterogeneity among patients, multisource DA may be able to alleviate the existing problem of high disparity in distributions within the source domain. In future work, we will consider further refinement of the model on this basis.

V. Conclusion

In this study, we propose a domain-adaptive transformer-based (AdaTrans) seizure prediction approach to alleviate the problem of high interpatient EEG heterogeneity. Specifically, we add a distiller to the model of available patients and then perform conditional domain adversarial training using a small amount of labeled data from a new patient. The adversarial process with the addition of label information is employed to learn a common seizure feature space among different patients. The experimental results on the CHB-MIT scalp EEG data set show that AdaTrans effectively reduces the impact of interpatient domain disparity and obtains better generalization ability compared to the state-of-the-art methods. This study provides a promising solution to the problem of cross-subject DA in seizure prediction.

References

[1] *Epilepsy: A Public Health Imperative*, World Health Org., Geneva, Switzerland, 2019.

[2] R. E. Stirling, M. J. Cook, D. B. Grayden, and P. J. Karoly, "Seizure forecasting and cyclic control of seizures," *Epilepsia*, vol. 62, pp. S2–S14, Feb. 2021.

[3] K. Rasheed et al., "Machine learning for predicting epileptic seizures using EEG signals: A review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 139–155, Jul. 2020. [Online]. Available: https://doi.org/10.1109/RBME.2020.3008792

[4] M. F. Pinto et al., "On the clinical acceptance of black-box systems for EEG seizure prediction," *Epilepsia Open*, vol. 7, no. 2, pp. 247–259, 2022.

[5] B. Maimaiti et al., "An overview of EEG-based machine learning methods in seizure prediction and opportunities for neurologists in this field," *Neuroscience*, vol. 481, pp. 197–218, Jan. 2022.

[6] D. Zeng, K. Huang, C. Xu, H. Shen, and Z. Chen, "Hierarchy graph convolution network and tree classification for epileptic detection on electroencephalography signals," *IEEE Trans. Cogn. Devel. Syst.*, vol. 13, no. 4, pp. 955–968, Dec. 2021.

[7] J. Cao, J. Zhu, W. Hu, and A. Kummert, "Epileptic signal classification with deep EEG features by stacked CNNs," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 4, pp. 709–722, Dec. 2020.

[8] J. Cao, D. Hu, Y. Wang, J. Wang, and B. Lei, "Epileptic classification with deep-transfer-learning-based feature fusion algorithm," *IEEE Trans. Cogn. Devel. Syst.*, vol. 14, no. 2, pp. 684–695, Jun. 2022.

[9] S. S. Viglione and G. O. Walsh, "Proceedings: Epileptic seizure prediction," *Electroencephalogr. Clin. Neurophysiol.*, vol. 39, no. 4, pp. 435–436, 1975.

[10] J. Martinerie et al., "Epileptic seizures can be anticipated by non-linear analysis," *Nat. Med.*, vol. 4, no. 10, pp. 1173–1176, 1998.

[11] K. Lehnertz and C. E. Elger, "Can epileptic seizures be predicted? evidence from nonlinear time series analysis of brain electrical activity," *Phys. Rev. Lett.*, vol. 80, no. 22, p. 5019, 1998.

[12] C. Li, Y. Zhao, R. Song, X. Liu, R. Qian, and X. Chen, "Patient-specific seizure prediction from electroencephalogram signal via multi-channel feedback capsule network," *IEEE Trans. Cogn. Devel. Syst.*, early access, Oct. 5, 2022, doi: 10.1109/TCDS.2022.3212019.

[13] Y. Qi, L. Ding, Y. Wang, and G. Pan, "Learning robust features from nonstationary brain signals by multiscale domain adaptation networks for seizure prediction," *IEEE Trans. Cogn. Devel. Syst.*, vol. 14, no. 3, pp. 1208–1216, Sep. 2022.
[14] Y. Wang et al., "Estimating brain connectivity with varying-length time lags using a recurrent neural network," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1953–1963, Sep. 2018.
[15] M. Z. Parvez and M. Paul, "Epileptic seizure prediction by exploiting spatiotemporal relationship of EEG signals using phase correlation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, pp. 158–168, 2016.
[16] Y. Park, L. Luo, K. K. Parhi, and T. Netoff, "Seizure prediction with spectral power of EEG using cost-sensitive support vector machines," *Epilepsia*, vol. 52, no. 10, pp. 1761–1770, 2011.
[17] L. Chisci et al., "Real-time epileptic seizure prediction using AR models and support vector machines," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 5, pp. 1124–1132, May 2010.
[18] D. Cho, B. Min, J. Kim, and B. Lee, "EEG-based prediction of epileptic seizures using phase synchronization elicited from noise-assisted multivariate empirical mode decomposition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, pp. 1309–1318, 2017.
[19] P. Mirowski, D. Madhavan, Y. LeCun, and R. Kuzniecky, "Classification of patterns of EEG synchronization for seizure prediction," *Clin. Neurophysiol.*, vol. 120, no. 11, pp. 1927–1940, 2009.
[20] D. Liang et al., "Semisupervised seizure prediction in scalp EEG using consistency regularization," *J. Healthc. Eng.*, vol. 2022, Jan. 2022, Art. no. 1573076. [Online]. Available: https://doi.org/10.1155/2022/1573076
[21] N. D. Truong et al., "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Netw.*, vol. 105, pp. 104–111, Sep. 2018.
[22] Y. Gao et al., "Pediatric seizure prediction in scalp EEG using a multiscale neural network with dilated convolutions," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–9, 2022.
[23] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, "Focal onset seizure prediction using convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 2109–2118, Sep. 2018.
[24] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[25] S. Zhang, D. Chen, R. Ranjan, H. Ke, Y. Tang, and A. Y. Zomaya, "A lightweight solution to epileptic seizure prediction based on EEG synchronization measurement," *J. Supercomput.*, vol. 77, no. 4, pp. 3914–3932, 2021.
[26] S. D. Shorvon, F. Andermann, and R. Guerrini, *The Causes of Epilepsy: Common and Uncommon Causes in Adults and Children*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[27] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[28] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, 2019, pp. 778–786.
[29] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, "Transfer learning with dynamic distribution adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 1, pp. 1–25, 2020.
[30] P. Peng, Y. Song, L. Yang, and H. Wei, "Seizure prediction in EEG signals using STFT and domain adaptation," *Front. Neurosci.*, vol. 15, Jan. 2022, Art. no. 825434.
[31] P. Peng, L. Xie, K. Zhang, J. Zhang, L. Yang, and H. Wei, "Domain adaptation for epileptic EEG classification using adversarial learning and Riemannian manifold," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103555.
[32] D. Cordes et al., "Energy-period profiles of brain networks in group fMRI resting-state data: A comparison of empirical mode decomposition with the short-time Fourier transform and the discrete wavelet transform," *Front. Neurosci.*, vol. 15, May 2021, Art. no. 663403.
[33] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, MIT Division Health Sci. Technol., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.
[34] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[35] A. Bhattacharya, T. Baweja, and S. P. K. Karri, "Epileptic seizure prediction using deep transformer model," *Int. J. Neural Syst.*, vol. 32, no. 2, 2022, Art. no. 2150058.
[36] J. Yan, J. Li, H. Xu, Y. Yu, and T. Xu, "Seizure prediction based on transformer using scalp electroencephalogram," *Appl. Sci.*, vol. 12, no. 9, p. 4158, 2022.
[37] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.
[38] Q. Wang et al., "Learning deep transformer models for machine translation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist.*, 2019, pp. 1810–1822.
[39] A. Baevski and M. Auli, "Adaptive input representations for neural language modeling," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.
[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5769–5779.
[42] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.
[43] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, "Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic," *Physica D, Nonlinear Phenom.*, vol. 194, nos. 3–4, pp. 357–368, 2004.
[44] S. Zhao et al., "Multi-source distilling domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12975–12983.
[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[46] A. Affes, A. Mdhaffar, C. Triki, M. Jmaiel, and B. Freisleben, "Personalized attention-based EEG channel selection for epileptic seizure prediction," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117733.
[47] J. S. Ra, T. Li, and Y. Li, "A novel permutation entropy-based EEG channel selection for improving epileptic seizure prediction," *Sensors*, vol. 21, no. 23, p. 7972, 2021.