Unsupervised Domain Adaptation for Cross-Subject Few-Shot Neurological Symptom Detection

Bingzhao Zhu[1] and Mahsa Shoaran[2]

***Abstract*— Modern machine learning tools have shown promise in detecting symptoms of neurological disorders. However, current approaches typically train a unique classifier for each subject. This subject-specific training scheme requires long labeled recordings from each patient, thus failing to detect symptoms in new patients with limited recordings. This paper introduces an unsupervised domain adaptation approach based on adversarial networks to enable few-shot, cross-subject epileptic seizure detection. Using adversarial learning, features from multiple patients were encoded into a subject-invariant space and a discriminative model was trained on subject-invariant features to make predictions. We evaluated this approach on the intracranial EEG (iEEG) recordings from 9 patients with epilepsy. Our approach enabled cross-subject seizure detection with a 9.4% improvement in 1-shot classification accuracy compared to the conventional subject-specific scheme.**

I. Introduction

Machine learning (ML) has been an increasingly useful tool in neural engineering in recent years. ML can be used to analyze and classify invasive or noninvasive electrophysiological recordings, enabling timely and accurate prediction of neurological symptoms (or events) in epilepsy [1], [2], [3], Parkinson's disease [4], migraine [5], and other emerging applications. However, despite the recent progress and potential of ML in neurological disease detection, the existing algorithms primarily use a subject-specific scheme, requiring each patient's extended neuronal recordings to train the model. Therefore, employing such algorithms on new patients with limited labeled data has been a challenge. This is particularly the case for invasive recordings, where the duration of recording is typically short due to surgical and ethical concerns (several minutes to days).

To tackle this problem, transfer learning aims to transfer the source domain knowledge to a target domain where labeled data is difficult to acquire [6]. Over the past decade, there has been an extensive literature on domain transfer learning [7], [8], with the goal of eliminating domain shift for better generative or discriminative performance. Among transfer learning approaches, the adversarial domain adaptation introduces an adversarial loss to minimize domain shift and enforce the learned representations to share a common feature space [9], and has obtained a promising performance in image-to-image translation tasks [10], [11]. Although domain adaptation techniques are widely used in computer vision tasks [10], [11], their application in neural engineering and particularly in detecting neurological symptoms is still underexplored [12].

In this paper, we propose a cross-subject seizure detection algorithm based on adversarial networks [9]. We mapped the features from various subjects into a subject-invariant space via the proposed unsupervised adversarial domain adaptation. Following domain adaptation, we trained an ensemble of gradient boosted trees in the subject-invariant feature space to generate cross-subject seizure predictions. The rest of this paper is organized as follows. We describe the classification task and dataset in Section II. The adversarial domain adaptation is introduced in Section III, followed by results in Section IV. Section V concludes the paper.

II. Classification Task and Data Description

In this work, we propose a domain adaptation model for cross-subject seizure detection. This approach was evaluated on continuous iEEG recordings from 9 patients with epilepsy.

A. Seizure detection task and iEEG data

Epileptic seizure detection is a supervised classification problem to differentiate between seizure and non-seizure states of a patient. We studied a total number of 97 seizure events from 9 patients. The iEEG recordings were sampled at 500Hz and annotated as *seizure* or *non-seizure* by domain experts (publicly available at the IEEG Portal [13]). All included subjects gave written informed consent and the study was approved by the Mayo Clinic and University of Pennsylvania Institutional Review Board. We segmented the iEEG recordings of each patient to 1s windows for the subsequent processing.

B. Feature extraction

A set of predictive biomarkers of seizure activity [2] were extracted from the segmented iEEG recordings, followed by domain adaptation and classification. The features and their definitions are as follows: line-length (LLN, $\frac{1}{d}\sum_d |x[n] - x[n-1]|$, $d =$ window size), total power (Pow, $\frac{1}{d}\sum_d x[n]^2$), variance (Var, $\frac{1}{d}\sum_d (x[n]-\mu)^2$, $\mu = \frac{1}{d}\sum_d x[n]$), and band power over delta (δ: 1–4 Hz), theta (θ: 4–8 Hz), alpha (α: 8–13 Hz), beta (β: 13–30 Hz), low-gamma (γ_1: 30–50 Hz), gamma (γ_2: 50–80 Hz), high-gamma (γ_3: 80–150 Hz), and ripple (R: 150–250 Hz) bands.

C. Train-test split

We split the data into train and test sets using a block-wise approach, in which each block is comprised of one seizure event and the subsequent non-seizure segment. To evaluate the performance, we used the first n blocks for training and the remaining blocks for testing, referred to as 'n-shot learning'

*This work was supported by funding from EPFL and Cornell University.

[1]Bingzhao Zhu is with the School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14850, USA. E-mail: bz323@cornell.edu

[2]Mahsa Shoaran is with the Institute of Electrical Engineering and Center for Neuroprosthetics, EPFL, Geneva 1202, Switzerland. E-mail: mahsa.shoaran@epfl.ch

in the following sections. The block-wise approach is a fair method to evaluate the performance, as we use a number of recorded seizure events to predict a future unseen seizure [2].

III. Adversarial Domain Adaptation

In this section, we consider each subject (i) to be associated with a specific domain ($\mathcal{D}_i = \{\mathcal{X}_i, P_i(\mathbf{X})\}$), from which features are sampled. $P_i(\mathbf{X})$ denotes the distribution of the feature vector $\mathbf{X}$. Our goal is to learn a unique encoder for each subject and map the features from this subject-specific domain to a subject-invariant domain.

A. Model structure

We first consider a simple case with only two patients: one patient from the source domain ($\mathcal{D}_S = \{\mathcal{X}_S, P_S(\mathbf{X})\}$) where exists abundant labeled data for a given task, and the other patient from the target domain ($\mathcal{D}_T = \{\mathcal{X}_T, P_T(\mathbf{X})\}$) where data is expensive to acquire for the same task.

As shown in Fig. 1, the proposed adversarial domain adaptation model consists of three parts: encoder (source encoder E_S, target encoder E_T), decoder (source decoder D_S, target decoder D_T), and subject discriminator (SD). We used the handcrafted features ($\mathbf{X}$) as input to the encoders. The encoders and decoders form an autoencoder, which learns a latent representation (dimension: 2048) of the original input. The subject discriminator is a multilayer perceptron, which takes the latent representations ($E_S(\mathbf{X})$ denoted by green squares, and $E_T(\mathbf{X})$ denoted by red squares) as input. The subject discriminator has two hidden layers with 512 and 128 nodes, respectively. We trained the SD to predict whether the latent representations are from the source or target subject.

1) Adversarial loss: The encoders and subject discriminator form a GAN model [9] for adversarial training. Here, we have encoders for both source and target domains. Let $\mathcal{L}_{\text{adv}}(\mathcal{X}_S, \mathcal{X}_T, E_S, E_T, SD)$ denote the standard supervised loss of SD. We train the subject discriminator by minimizing the loss:

$$\min_{SD} \mathcal{L}_{\text{adv}}\left(\mathcal{X}_S, \mathcal{X}_T, E_S, E_T, SD\right). \tag{1}$$

The goal of the encoders is to minimize the distance between the empirical source and target latent representations $E_S(\mathbf{X}_S)$ and $E_T(\mathbf{X}_T)$. Thus, we trained the encoders to fool the subject discriminator and make the source/target representations indistinguishable from each other:

$$\max_{E_S, E_T} \mathcal{L}_{\text{adv}}\left(\mathcal{X}_S, \mathcal{X}_T, E_S, E_T, SD\right). \tag{2}$$

Overall, the adversarial learning can be formalized as a maximin problem which can be solved using alternating optimization: $\max_{E_S,E_T} \min_{SD} \mathcal{L}_{\text{adv}}\left(\mathcal{X}_S, \mathcal{X}_T, E_S, E_T, SD\right)$.

2) Reconstruction loss and mode collapse: With the adversarial training, we expect the latent space to be a subject-invariant representation of the inputs. However, the source and target encoders may simply learn to produce the same output (e.g., all zeros for latent representation), making it impossible for the subject discriminator to distinguish. In this scenario, the subject-invariant space cannot represent the inputs. This failure is referred to as mode collapse, which is a common issue with GAN training [14].

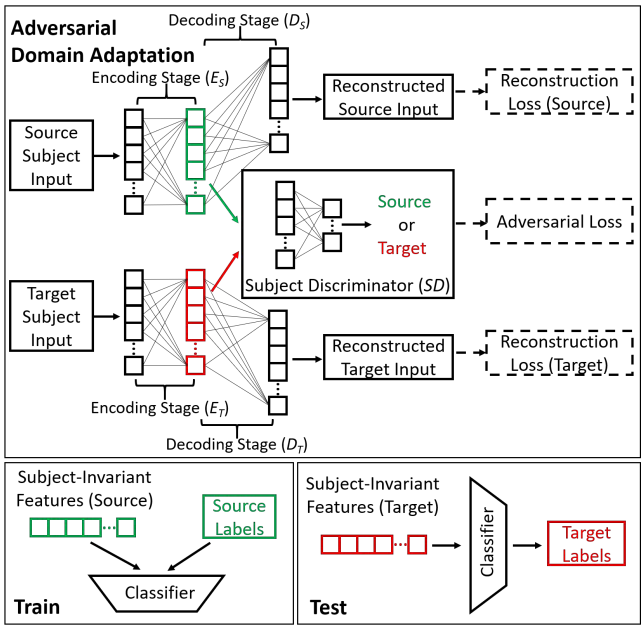

Fig. 1. Model structure of the proposed unsupervised adversarial domain adaptation (top), train and test approaches (bottom). The encoding and decoding stages form an autoencoder for each subject, which learns a latent representation of the input feature vectors. The subject discriminator takes the latent representation as input and is trained to distinguish the data from different subjects. The encoding stages are trained to fool the subject discriminator. Our goal is to learn encoders that map the input features to a subject-invariant space (denoted by green and red blocks). Following domain adaptation, a classifier is trained with the subject-invariant features to predict seizures on a target subject.

To avoid mode collapse, we reconstructed the inputs from the latent representations using a decoding stage. The decoders enforce the latent representations to preserve similar information as the inputs. We calculated the reconstruction loss $\mathcal{L}_{\text{rec}}$ using the mean squared error (MSE) and L_1 norm was used to regularize the autoencoder.

$$\mathcal{L}_{\text{rec}}(E, D) = \frac{1}{N}\sum_{n=1}^{N} \|X^{(n)} - D(E(X^{(n)}))\|^2 + \lambda(\|E\|_1 + \|D\|_1), \tag{3}$$

where N represents the mini-batch size, X_n denotes the n-th sample, and λ is the regularization coefficient, which was empirically set to 3e-5 in this work. In addition to maximizing $\mathcal{L}_{\text{adv}}$, we also train the encoders and decoders to minimize the reconstruction loss. Overall, the encoders and decoders are trained with the following formula:

$$\min_{E_S, E_T, D_S, DT} -\mathcal{L}_{\text{adv}}\left(\mathcal{X}_S, \mathcal{X}_T, E_S, E_T, SD\right) + \alpha(\mathcal{L}_{\text{rec}}(E_S, D_S) + \mathcal{L}_{\text{rec}}(E_T, D_T)), \tag{4}$$

where α controls the trade-off between the adversarial loss and reconstruction loss.

B. Multi-subject domain adaptation

Previous literature on domain adaptation has only focused on transferring from one source domain to the target [7]. However, for cross-subject seizure detection, we need to consider each patient as a unique domain and transfer

Algorithm 1 Multi-Subject Domain Adaptation.

Require: recordings from N_s subjects

$E_{1,\ldots,N_s}, D_{1,\ldots,N_s}, SD \leftarrow$ random initialization

for number of iterations **do**

Sample mini-batches of N samples from all subjects $\left\{\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_1^{(N)}, \ldots, \mathbf{X}_{N_s}^{(N)}\right\}$

Update the subject discriminator SD by gradient decent:

$$\nabla_{SD} \frac{1}{N} \sum_{i=1}^{N_s} \sum_{n=1}^{N} -log(SD^i(E_i(\mathbf{X_i^{(n)}})))$$

for $i \in 1, \ldots, N_s$ **do**

Update encoder and decoder E_i, D_i by gradient decent:

$$\nabla_{E_i, D_i} \frac{1}{N} \sum_{n=1}^{N} [log(SD^i(E_i(\mathbf{X_i^{(n)}})))$$
$$+\alpha(\|\mathbf{X}_i^{(n)} - D_i(E_i(\mathbf{X}_i^{(n)}))\|^2 + \lambda(\|E\|_i + \|D\|_i))]$$

end for

end for

the feature vectors from multiple subjects to a subject-invariant domain. Here, we extended the domain adaptation framework to enable multi-subject seizure detection. The subject discriminator predicts the patient index (rather than only a single source or target), and patients are alternately considered as target while others are considered as source. We used the cross-entropy loss for $\mathcal{L}_{\text{adv}}$:

$$\mathcal{L}_{\text{adv}} = -\sum_{i=1}^{N_s} \mathbb{E}_{\mathbf{X} \sim P_i(\mathbf{X})}[log(SD^i(E_i(\mathbf{X})))], \qquad (5)$$

where $N_s = 9$ is the total number of patients, SD outputs a vector of size N_s, and the i-th entry of the subject discriminator output (SD^i) indicates the probability that $\mathbf{X}$ belongs to subject i.

1) Learning procedure: Eq. 1-4 show the learning objectives for the domain adaptation with two subjects: a source patient and a target patient. Here, we introduce the learning procedure for multiple patients. We alternatingly considered one patient as the target and all other patients as source. The algorithmic pseudocode is shown in Algorithm. 1. Our goal is to leverage the labeled data from source patients to make predictions for a target patient. The domain adaptation process is essentially unsupervised, mapping different subjects' data into a common feature space. A discriminative model was trained on the subject-invariant features to generate predictions for the target patients. We tested several settings. For example, 0-shot learning does not require any labeled recordings from the target patient. So we trained the discriminative model only on the labeled data from source patients. For n-shot learning, n labeled seizure blocks from the target patient were used for training, in addition to the extensive data from source patients.

2) Convergence Analysis: The encoders map the input subject feature distribution ($P_i(\mathbf{X})$) to a latent space distribution ($Q_i(\mathbf{z})$). In the cross-subject learning scheme, we would like $Q_i(\mathbf{z})$ to be invariant across patients (i.e., $Q_i(\mathbf{z}) = Q_1(\mathbf{z})$ for all $i \in 1, \ldots, N_s$). Previous work has proven that adversarial training can reduce the shift between target and source domains [9]. In this work, the subject discriminator performs a multi-class classification task and patients are alternately considered as the target (Algorithm. 1). Following the framework in [9] , we recognize that Algorithm. 1 minimizes the Jensen-Shannon Divergence ($JSD(Q_1, \ldots, Q_{N_s})$) of latent space distributions [15]. Given that $JSD(Q_1, \ldots, Q_{N_s})$ is always non-negative and becomes zero if and only if all distributions are the same (i.e., $Q_i(\mathbf{z}) = Q_1(\mathbf{z})$ for $i \in 1, \ldots, N_s$), Algorithm. 1 will converge to a subject-invariant space given sufficient capacity.

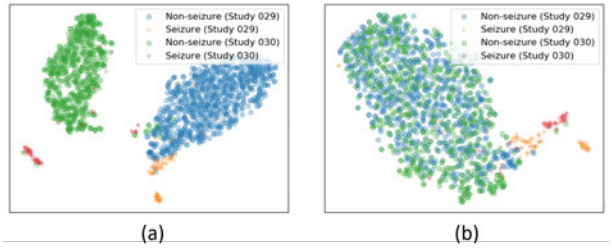

Fig. 2. t-SNE visualization of the data distribution from two patients; (a) Visualization of the data distribution before domain adaptation. (b) Visualization of the subject-invariant feature space. After domain adaptation, the data from different patients become indistinguishable.

IV. Results

We tested the proposed algorithm for seizure detection from iEEG recordings of 9 epilepsy patients. We first mapped the input feature vectors into a subject-invariant space, using the proposed unsupervised adversarial training. A discriminative model (gradient boosted trees [16]) was trained in the subject-invariant space to make predictions for each patient.

A. t-SNE visualization of data distribution

We used t-SNE [17] to visualize the high-dimensional data distribution by mapping each data sample to a location in a 2-dimensional space. As shown in Fig. 2, we plot the data distribution of two patients before and after domain adaptation. We used different colors and markers to show which class the points are belonging to (seizure or non-seizure). In Fig. 2(a), Study 029 has a different distribution from Study 030. The domain adaptation process successfully removed the between-subject variation and brought their distributions closer to each other (Fig. 2(b)), enables cross-subject classification. Visualization was obtained with $\alpha = 0.01$ (see α in Eq. 4 or Algorithm. 1), which we kept for the following experiments.

B. Cross-subject seizure detection

We first trained the encoders to map the features into a subject-invariant space, using the domain adaptation process depicted in Algorithm. 1. The size of mini-batches (N) was set to 32. We used the Adam optimizer [18] (learning rate of 1e-5) to update both encoders and subject discriminator for 100 epochs. Next, 100 gradient boosted trees with a maximum depth of 4 were trained on the subject-invariant features to predict the probability of epileptic seizures [16], using the 0-shot and n-shot learning schemes described above. In the n-shot learning scheme, we assigned different weights to the data from source and target patients. The samples from source patients received a weight of 0.01 while the data from target patients had a sample weight of 1. We applied

TABLE I

PERFORMANCE OF CONVENTIONAL SUBJECT-SPECIFIC (SS) AND CROSS-SUBJECT (CS) SEIZURE DETECTION METHODS.

Subject #	0-shot	1-shot		2-shot		3-shot	
	CS	SS	CS	SS	CS	SS	CS
Study 004-2	0.791 ± 0.091	0.889± 0.047	**0.935± 0.027**	**0.947± 0.033**	0.915± 0.041	N/A	N/A
Study 022	0.809 ± 0.058	0.787± 0.056	**0.962± 0.012**	0.910± 0.052	**0.960± 0.007**	0.875± 0.050	**0.915± 0.023**
Study 024	0.695 ± 0.124	0.874± 0.030	**0.949± 0.012**	0.837± 0.061	**0.944± 0.011**	0.914± 0.015	**0.938± 0.012**
Study 026	0.715± 0.099	0.658± 0.158	**0.931± 0.009**	0.931± 0.011	**0.953± 0.009**	0.928± 0.024	**0.957± 0.008**
Study 029	0.773± 0.034	**0.813± 0.059**	0.785± 0.070	0.942± 0.024	**0.944± 0.022**	N/A	N/A
Study 030	0.793± 0.044	**0.977± 0.005**	0.974± 0.010	**0.983± 0.009**	0.979± 0.003	0.976± 0.025	**0.990± 0.003**
Study 033	0.659± 0.049	0.888± 0.009	**0.901± 0.002**	0.885± 0.005	**0.887± 0.003**	0.920± 0.004	**0.923± 0.005**
Study 037	0.514± 0.145	0.631± 0.043	**0.801± 0.012**	0.597± 0.071	**0.751± 0.037**	**0.994± 0.004**	0.989± 0.004
Study 038	0.596± 0.029	**0.882± 0.019**	0.858± 0.027	**0.896± 0.005**	0.882± 0.027	0.915± 0.011	**0.929± 0.011**
Average	0.705± 0.030	0.822± 0.031	**0.899± 0.012**	0.881± 0.012	**0.913± 0.007**	0.932± 0.008	**0.949± 0.005**

the reweighting scheme to address the following concerns: (1) In few-shot learning, the training samples from source patients were more than the samples from target patients by an order of magnitude. (2) Compared to the data from source patients, the target patient data is more informative in predicting seizures on that patient.

Given the imbalanced nature of the seizure detection task, we evaluated the classification performance using the area under the ROC curve (AUC scores). Table. I compares the classification performance with/without cross-subject knowledge. In the conventional subject-specific (SS) setting, we trained the classifiers by only using the data from a target patient (e.g., n seizure blocks in n-shot). For the cross-subject (CS) setting, we further incorporated the knowledge from source patients. We ran the proposed domain adaptation approach for 5 independent trials and reported the average performance (AUC scores) $\pm$ standard deviation. 3-shot learning on two patients (Study 004-2 and Study 029) is not applicable (N/A), since only 3 seizure events are available in both patients. As shown in this Table, 0-shot learning achieved an average AUC score of 0.705, which is much better than the chance level (0.5). For 1-, 2-, 3-shot learning, CS outperforms the SS in terms of average classification performance. However, as we used more labeled samples from the target patient (i.e., moved from 1-shot to 3-shot learning), the difference become less significant. Overall, cross-subject learning achieved a superior performance compared to the subject-specific setting, which indicates the importance of leveraging cross-subject knowledge. In addition to seizure detection, the proposed approach has the potential to help various neurological disorders and symptom detection tasks where training data is generally limited, which remains as future work.

V. CONCLUSION

In this paper, we proposed a novel cross-subject seizure detection framework based on adversarial domain adaptation, by mapping the features from different subjects into a subject-invariant space and applying cross-subject learning. With unsupervised domain adaptation, we achieved a better performance compared to the conventional subject-specific approach, particularly when the training data is limited (few-shot learning). The proposed model efficiently incorporates the knowledge from previous patients to enable high-accuracy seizure detection in new patients.

REFERENCES

[1] Muhammad Awais Bin Altaf, Judyta Tillak, Yonatan Kifle, and Jerald Yoo, "A 1.83 μj/classification nonlinear support-vector-machine-based patient-specific seizure classification soc," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*. IEEE, 2013, pp. 100–101.

[2] Mahsa Shoaran, Benyamin Allahgholizadeh Haghi, Milad Taghavi, Masoud Farivar, and Azita Emami-Neyestanak, "Energy-efficient classification for resource-constrained biomedical applications," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 4, pp. 693–707, 2018.

[3] Bingzhao Zhu, Masoud Farivar, and Mahsa Shoaran, "Resot: Resource-efficient oblique trees for neural signal classification," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 4, pp. 692–704, 2020.

[4] Lin Yao, Peter Brown, and Mahsa Shoaran, "Improved detection of Parkinsonian resting tremor with feature engineering and Kalman filtering," *Clinical Neurophysiology*, vol. 131, no. 1, pp. 274–284, 2020.

[5] Bingzhao Zhu, Gianluca Coppola, and Mahsa Shoaran, "Migraine classification using somatosensory evoked potentials," *Cephalalgia*, vol. 39, no. 9, pp. 1143–1155, 2019.

[6] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, pp. 031005, 2018.

[7] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[8] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] 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*, 2020.

[13] Joost B Wagenaar, Gregory A Worrell, Zachary Ives, MATTHIAS Dümpelmann, Brian Litt, and Andreas Schulze-Bonhage, "Collaborating and sharing data in epilepsy research," *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, vol. 32, no. 3, pp. 235, 2015.

[14] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[15] Jianhua Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, 2017, pp. 3146–3154.

[17] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[18] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.