Cluster Embedding Joint-Probability-Discrepancy Transfer for Cross-Subject Seizure Detection

Xiaonan Cui, Jiuwen Cao, *Senior Member, IEEE*, Xiaoping Lai, *Member, IEEE*, Tiejia Jiang, and Feng Gao

***Abstract*— Transfer learning (TL) has been applied in seizure detection to deal with differences between different subjects or tasks. In this paper, we consider cross-subject seizure detection that does not rely on patient history records, that is, acquiring knowledge from other subjects through TL to improve seizure detection performance. We propose a novel domain adaptation method, named the Cluster Embedding Joint-Probability-Discrepancy Transfer (CEJT), for data distribution structure learning. Specifically, 1) The joint probability distribution discrepancy is minimized to reduce the distribution shift in the source and target domains, and strengthen the discriminative knowledge of classes. 2) A clustering is performed on the target domain, and the class centroids of sources is used as the clustering prototype of the target domain to enhance data structure. It is worth noting that the manifold regularization is used to improve the quality of clustering prototypes. In addition, a correlation-alignment-based source selection metric (SSC) is designed for most favorable subject selection, reducing the computational cost as well as avoiding some negative transfer. Experiments on 15 patients with focal epilepsy from the Children's Hospital, Zhejiang University School of Medicine (CHZU) database shown that CEJT outperforms several state-of-the-art approaches, and can promote the application of seizure detection.**

***Index Terms*— Seizure detection, domain adaptation, transfer learning, correlation-alignment-based source selection.**

I. Introduction

EPILEPSY is a common neurological syndrome caused by abnormal discharge of brain neurons. Seizures are the most important clinical manifestations of epilepsy. Patients have unusual behaviors and sensations during seizures, and sometimes lead to loss of consciousness. Driven by data, researchers have begun to build epileptic seizure detection models through machine learning, correlation analysis, and time-frequency analysis [1] in recent years. By automatically identifying seizure on electroencephalography (EEG), it can provide an objective reference to neurologists for epilepsy diagnosis, treatment and evaluation [2], [3], [4], [5].

Most of the existing EEG-based seizure detection methods focus on patient-dependent scenarios, including training and testing data originating from the same patient, or mixing the collected data together for model training and testing [6], [7], [8]. The patient-dependent forms strongly rely on the patient history records. Patient-dependent algorithms have been extensively studied in the past. The high accuracy of seizure detection methods in this scenario can be attributed to a basic assumption that training and testing data follow the same distribution. However, in real scenarios, it is shown that there are differences in the onset and propagation of abnormal electrical activity in the brain [9]. Moreover, EEG is greatly affected by age and individual differences, especially in children, with the increase of age, the frequency, amplitude and rhythm of EEG background activity are significantly different [10]. Faced with a more diverse data distribution from different subjects, patient-dependent seizure detection methods become insufficient for new patients.

Due to the significant individual differences in EEG signals, the training data and the actual testing data do not obey the assumption of independent identical distribution. *How to establish a cross-subject seizure detection model that can overcome individual differences is a long-standing issue?* Domain adaptation naturally comes to mind, which offers the possibility to generalize a classifier learned from well-labeled source domains to an unlabeled target domain, where observations from source and target domains are often derived from different distributions. In cross-subject seizure detection, the domain consisting of multiple subjects with sufficient labeled data is called the source domain, and the domain consisting of subjects with unlabeled data is named the target domain. In this paper, we propose a domain adaptation-based learning framework to develop a robust cross-subject seizure detection algorithm, which can eliminate the influence of distribution differences between patients. There are two key contributions in the paper: 1) We design a simple but effective evaluation metric for source-domain transferability, the correlation-alignment-based source selection (SSC), to select the most favorable subjects in multi-source transfer learning.

Manuscript received 26 August 2022; revised 21 November 2022 and 9 December 2022; accepted 11 December 2022. Date of publication 14 December 2022; date of current version 1 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U1909209, in part by the National Key Research and Development of China under Grant 2021YFE0100100, in part by the National Key Research and Development Program of China under Grant 2021YFE0205400, in part by the Natural Science Key Foundation of Zhejiang Province under Grant LZ22F030002, and in part by the Research Funding of Education of Zhejiang Province under Grant GK228810299201. *(Xiaonan Cui and Tiejia Jiang contributed equally to this work.) (Corresponding author: Jiuwen Cao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Second Affiliated Hospital of Zhejiang University and registered in Chinese Clinical Trail Registry (ChiCTR1900020726). All patients gave their informed consent prior to their inclusion in the study.

Xiaonan Cui, Jiuwen Cao, and Xiaoping Lai are with the Machine Learning and I-Health International Cooperation Base of Zhejiang Province, Artificial Intelligence Institute, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China (e-mail: xncui@hdu.edu.cn; jwcao@hdu.edu.cn; laixp@hdu.edu.cn).

Tiejia Jiang and Feng Gao are with the National Clinical Research Center for Child Health, Department of Neurology, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China (e-mail: jiangyouze@zju.edu.cn; epilepsy@zju.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3229066

2) We propose a new domain adaptation algorithm, the Cluster Embedding Joint-Probability-Discrepancy Transfer (CEJT), which unifies the cluster learning and joint probability distribution discrepancy. Minimizing the joint probability distribution discrepancy can reduce the difference between domains and strengthen the discriminative knowledge of categories, and the clustering learning can deeply explore the data distribution structure of the target domain. In this study, we validate the proposed cross-subject seizure detection model consisting of the transferability evaluation metric SSC and domain adaptation algorithm CEJT on the dataset collected from the Children's Hospital, Zhejiang University School of Medicine (CHZU).

The remainder of this paper is organized as follows: Section II introduces related work on domain adaptation and transfer learning based seizure detection. Section III describes the details of the seizure detection framework composed of CEJT and SSC. Section IV presents the experimental studies to compare the performance of CEJT with several state-of-the-art (SOTA) domain adaptation methods and to verify the effectiveness of the proposed framework. Finally, Section V draws the conclusions.

II. Related Work

A. Domain Adaptation

The most commonly used domain adaptation approaches include instance-based adaptation and feature representation adaptation [11]. It is generally believed that distribution differences can be compensated by the instance-based adaptation approaches, such as weighting the samples from the source domain to better match the target-domain distribution; or adopting feature transformation-based methods to project the features of the two domains to another subspace with small distribution shift.

Feature-based approaches seek a unified/respective transformation that projects data from two domains into a domain-invariant space to reduce distribution differences between domains while preserving data properties in the original space. Such methods often rely on a distance metric, the maximum mean discrepancy (MMD). MMD measures the distance between two distributions in the reproducing kernel Hilbert space (RKHS). Pan et al. [12] propose the transfer component analysis (TCA) using MMD to learn the transport components across domains in RKHS. TCA assumes that there is a feature map such that the marginal distributions of the two domains are close after the mapping. Joint distribution analysis (JDA) [13] improves the disadvantage in TCA which only considers the marginal distribution shift, and JDA takes the conditional distribution shift into account using the pseudo-label of the target domain. Adaptation regularization based transfer learning (ARTL) [14] builds a domain-invariant classifier by introducing the structural risk loss. Domain-invariant classifiers tend to have better performance than single feature transformations. Manifold embedded distribution alignment (MEDA) [15] is the first to quantitatively evaluate the importance of marginal and conditional distributions when performing distribution alignment. Joint geometrical and statistical alignment (JGSA) [16] breaks the strong assumption that the source and target domains need to be transformed uniformly, learning two coupled projections while reducing the geometric and distributional shifts.

Instance-based adaptation is often not considered separately, but is usually combined with feature matching to achieve domain adaptation. Long et al. [17] state that there are some source-domain samples unrelated to the target domain in feature matching, and propose a transfer joint matching (TJM) by introducing the $l_{2,1}$-norm regularization term to achieve instance weighting. Locality preserving joint transfer (LPJT) [18] explicitly weights samples from the source and target domains, and reduces the influence of outliers through landmark selection.

Deep domain adaptation utilizes deep networks to enhance domain adaptation performance, where discrepancy-based methods have been extensively studied. The deep domain confusion network (DDC) by Tzeng et al. [19] adds an adaptation layer with MMD metric to the convolutional network, and the domain discrepancy loss of the adaptation layer is used to improve the original objective function. Rather than using a single layer and linear MMD, the Deep Adaptation Network (DAN) [20] measures domain discrepancy by considering all task-specific layers and designs an optimal multi-kernel selection strategy to improve the effectiveness of embedding matching. The joint adaptation network (JAN) [21] aligns the joint distribution of features and labels in multiple domain-specific layers based on joint MMD. CORrelation ALignment (CORAL), which learns a linear transformation to align second-order statistics between domains, has been extended to deep networks [22]. Adversarial-based methods encourage domain confusion through adversarial objectives, resulting in domain-invariant representations. The domain-adversarial neural network (DANN) adds an adversarial mechanism to the deep transfer network. Yu et al. [23] prove that the adversarial network also suffers from probability distribution mismatch, and propose a dynamic adversarial adaptation network (DAAN) to dynamically learn domain-invariant representations.

The above findings for shallow methods, most feature matching algorithms consider a linear combination of aligned marginal and conditional distributions, that is not equivalent to a joint distribution. Meanwhile, existing domain-invariant classifiers often use the squared loss and hinge loss, and the learned classifier generally labels the target samples separately, failing to make full use of the target-domain data structure information.

B. Transfer Learning Based Seizure Detection

Domain adaptation has been applied in seizure detection in the past. Yang et al. [24] use the large-margin projected transductive SVM to reduce the distribution difference between training and testing data, and realize the adaptive recognition of EEGs. In [25], the TSK fuzzy system and distribution alignment are jointly optimized, and the proposed TL-SSL-TSK has strong interpretability and adaptability in epilepsy recognition. Jiang et al. [26] introduce a semi-supervised learning

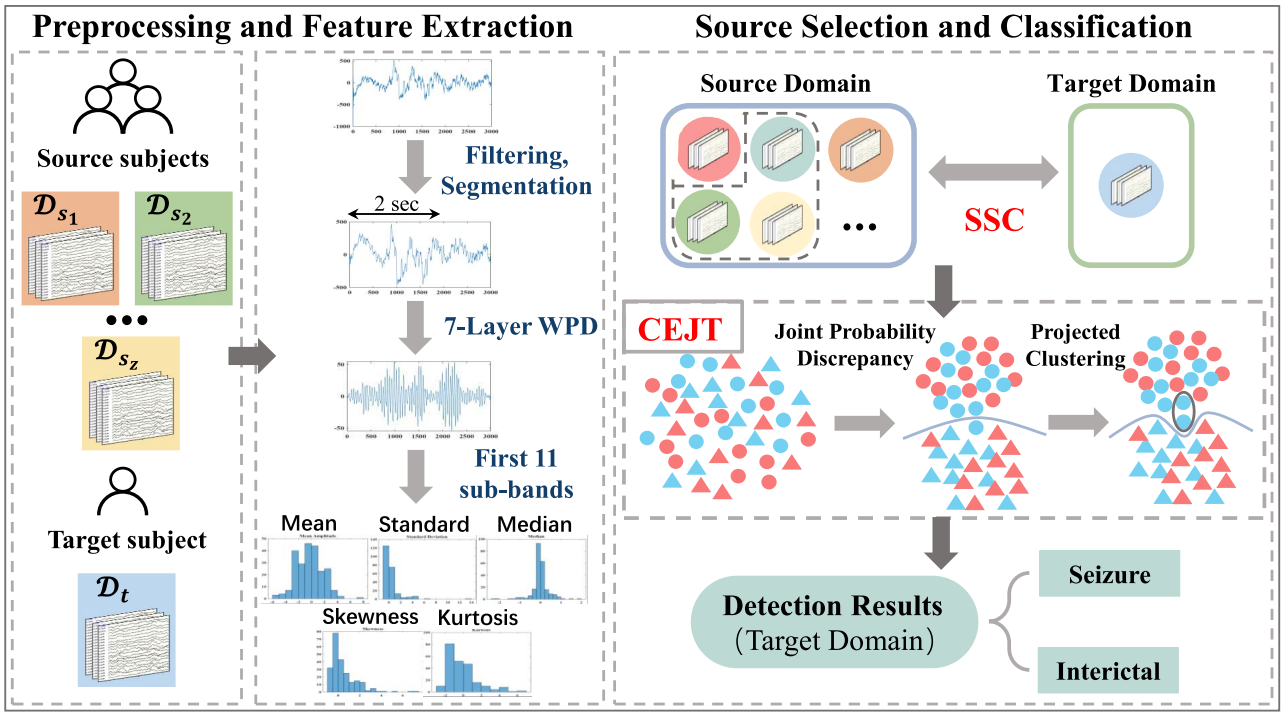


Fig. 1. Flowchart of our proposed cross-subject seizure detection framework, including data preprocessing, feature extraction, source selection and domain adaptation classification. The collected raw EEG signals are first filtered and segmented to extract wavelet packet features. Then, SSC is used to evaluate the discriminability of the source subject and the correlation with the target domain. Finally, CEJT unifies feature matching and projection clustering to build a classification model for seizure detection.

method based on [25] to exploit the unlabeled testing data. In [27], the feedforward neural networks, fuzzy systems, and transductive transfer learning are successfully unified into a generalized hidden-mapping model for seizure recognition. Recently, a cross-domain epilepsy EEG signal classification model with knowledge utilization maximization [28] has been proposed, which makes full use of the data global structure of source and target domain. And a pairwise constraint regularization term is added to utilize the association information between the labeled samples. In [29], from the perspective of error consistency, a regularization used for knowledge transfer is proposed to unify the TSK fuzzy classifier to achieve online calibration. The effectiveness of these algorithms for EEG differences in different states has been proved, but the performance on individual differences is not well studied.

Deep transfer learning has also been used for seizure detection. Zhang et al. [30] convert EEG signals into the time-frequency maps and three fine-tuned deep networks, VGG16, VGG19 and ResNet50, are adopted for classification. In [31], a unified adversarial learning framework is proposed to extract the epilepsy-specific representations while removing inter-patient noises. Cao et al. [32] perform quadratic feature extraction on the mean amplitudes of sub-band spectrum representing brain activity rhythms through a deep pre-trained network and develop a deep network for epileptic state classification. Most of these studies initialize the network parameters or carry out secondary feature extraction through pre-training models, which realize model transfer and cannot effectively solve the problem of individual differences in EEG signals.

III. Dataset and Methods

The proposed cross-subject seizure detection framework, which aims to use data from multiple source subjects to help target subject build domain adaptation classification models, is introduced in this section. Shown in Fig. 1, the multi-channel EEG signal are first filtered and segmented, and then wavelet packet decomposition (WPD) is performed to extract statistical features of EEGs. The correlation-alignment-based source selection (SSC) is then used to evaluate the transferability of subjects in the source domain. Finally, the selected source subjects are used together with the target subjects to build the Cluster Embedding Joint-Probability-Discrepancy Transfer learning (CEJT) classification model.

A. CHZU Dataset and Feature Extraction

The EEG signals used in this study are obtained from the Children's Hospital, Zhejiang University School of Medicine (CHZU). The recording time of EEG signals for each subject is 2 hours or 16 hours, respectively. The EEG signals are collected by the international 10-20 lead system, each record contains 21 scalp EEG channels, and the sampling frequency is 1000 Hz. In the study, 15 children with focal epilepsy are analyzed. We first divide the EEG signal into interictal and ictal states, where the interictal state refers to signal from one hour and more before a seizure onset but one hour after the previous seizure. The EEG signals are further segmented into 2-second frames, and the overlap rate between the two adjacent samples is 50% for ictal state. While for interictal

TABLE I
DATASET SPECIFICATIONS: THE DATA ARE FROM PATIENTS IN CHZU, WHICH INCLUDE 7 FEMALES AND 8 MALES. THE EEG RECORDING TIME OF EACH SUBJECT IS EITHER 2 HOURS OR 16 HOURS. THE SHORTEST SEIZURE DURATION IS 29 SECONDS AND THE LONGEST IS 347 SECONDS

Subject ID	Gender	Age	Number of seizures	Interictal samples / Ictal samples	Diagnostic report
P01	F	8y2m	7	843 / 573	Focal secondary bilateral tonic-clonic seizures with episodic EEG changes originating in the left hemisphere
P02	M	2m8d	8	1572 / 1794	One focal motor seizure and two focal secondary bilateral tonic-clonic seizures
P03	M	2y11m	5	1500 / 233	Focal motor seizures, episodic EEG changes originating in the right central region
P04	F	5m30d	3	900 / 288	Focal seizures, episodic EEG changes in the left hemisphere
P05	M	4m13d	2	600 / 246	Focal secondary bilateral tonic-clonic seizures with episodic EEG changes originating in the left frontal region
P06	M	4m8d	2	300 / 365	Focal secondary bilateral tonic-clonic seizures with episodic EEG changes originating in the left temporal region
P07	F	4m25d	1	300 / 202	Focal secondary bilateral tonic-clonic seizures with episodic EEG changes originating in the left temporal region
P08	F	11y1m	12	1646 / 530	Temporal lobe seizures, irregular δ activity with widespread long-range spikes and sharp waves
P09	F	7y10m	3	900 / 301	Frontal-initiated focal motor seizures with confusion
P10	M	1y	1	222 / 183	Focal seizures with disturbance of consciousness, episodic EEG changes originating in the left anterior temporal region
P11	M	5m16d	3	900 / 157	Migrating focal seizures
P12	M	5y11m	5	300 / 343	Focal seizures, episodic EEG changes originating in the left frontotemporal region
P13	F	27d	2	300 / 212	Episodic EEG changes with right central origin
P14	M	2m13d	2	300 / 689	Episodic EEG changes originating from the right anterior temporal region, synchronized child with head tilted to the right, hands clenched into fists
P15	M	5y5m	2	600 / 162	Focal seizures, episodic EEG changes originating in the right hemisphere

state, there has no overlap between EEG frames. Table I lists the specifications of CHZU dataset.

In order to utilize both the time and frequency domain EEG knowledge, the wavelet packet decomposition (WPD) is adopted for feature extraction. In the experiment, we perform a 7-layer WPD on the pre-processed EEG signal, and the first 11 sub-bands covering 0-40 Hz are selected. Then, 5 statistical features, including the mean amplitude, standard deviation, median, kurtosis and skewness, are extracted on each sub-band. A 55-dimensional feature vector is generated on each EEG channel. Finally, for all 21 channels, each EEG frame is represented by a feature vector of 21×55.1155.

B. Cluster Embedding Joint-Probability-Discrepancy Transfer

1) Problem Settings and Notations: A domain $\mathcal{D}$ contains three parts: feature space $\mathcal{X}$, probability distribution $P(\mathbf{X})$ and label space $\mathcal{Y}$, where $\mathbf{X} \in \mathcal{X}$. For simplicity, We use subscripts s and t to indicate the source domain and the target domain, respectively. The key notations used in this paper and the corresponding descriptions are shown in Table II.

Let $\mathcal{D}_s = \left\{\left(\mathbf{x}_{s,i}, y_{s,i}\right)\right\}_{i=1}^{n_s} = \{\mathbf{X}_s, \mathbf{Y}_s\}$ denote source-domain EEG samples data, where $\mathbf{x}_{s,i} \in \mathbb{R}^d$ is the feature vector with label $y_{s,i} \in \mathbb{R}^C$. Similarly, we let $\mathcal{D}_t = \left\{\mathbf{x}_{t,j}\right\}_{j=1}^{n_t} = \{\mathbf{X}_t\}$ as unlabeled target-domain data with $\mathbf{x}_{t,j} \in \mathbb{R}^d$. We assume the feature spaces and label spaces between domains are the same: $\mathcal{X}_s = \mathcal{X}_t$ and $\mathcal{Y}_s = \mathcal{Y}_t$. Due to the domain shift $P_s(\mathbf{x}_s, y_s) \neq P_t(\mathbf{x}_t, y_t)$, we devote to seek a

TABLE II
SYMBOL NOTATIONS AND DESCRIPTIONS

Notation	Description
$\mathbf{X}_s/\mathbf{X}_t$	Source/Target original data
$\mathbf{Y}_s/\hat{\mathbf{Y}}_t$	Source/Target label matrix
$\mathbf{P}$	Projection matrix
$\mathbf{F}$	Cluster centroids
$\mathbf{E}/\mathbf{V}$	Source/Target class centroid indicator matrix
$\mathbf{R}_T/\mathbf{R}_D$	Transferability/Discriminability indicator matrix
$\mathbf{L}$	Laplacian matrix
$\mathbf{W}$	Affinity matrix
$\hat{\mathbf{I}}$	Domain indication matrix
$\mathbf{H}$	Centering matrix
$\mathbf{I}_m$	Identity matrix with dimension m
n_s/n_t	Number of source/target samples
d/m	Dimension of original/projected feature
C	Number of categories
n_s^c/n_t^c	Number of source/target samples in class c

latent common space shared across source and target domains through a projection $\mathbf{P} \in \mathbb{R}^{d\times m}$, where the domain shifts are minimized and the discriminative knowledge is transferred from $\mathcal{D}_s$ and $\mathcal{D}_t$. On this basis, we aim to design a adaptive classifier by exploring two learning strategies: distribution adaptation and label propagation. Thus, we adopt the projected clustering to regard the samples within the same cluster in target domain as a whole to emphasize the data distribution structure of target domain. CEJT is formulated by finding a projection to obtain new representations of the respective domains and labels of the target domain, such that 1) in the projected space, the clustering of the target domain is achieved through the class centroids of the source domain, 2) the distribution matching of the same class and distinguishability of different classes in source and target domains are jointly explored, 3) the local manifold is introduced to improve the quality of cluster centroids.

2) Projected Clustering: Projected clustering aims to jointly optimize cluster centroids and labels in the embedding space so that samples within the same cluster can share the same label. In the case that all the source-domain labels are available, the class centroids of the source data can be obtained by calculating the mean of sample features in the identical class after projection. Based on the discriminative structure of the source data and the sample distribution structure information of target data, the pseudo-labels are assigned to the target samples under the guidance of the class centroids. Then, the projected clustering can be expressed as:

$$L_{pc} = \left\|\mathbf{P}^T\mathbf{X}_s\mathbf{E}_s - \mathbf{F}\right\|_F^2 + \alpha\left\|\mathbf{P}^T\mathbf{X}_t - \mathbf{F}\mathbf{\Omega}_t^T\right\|_F^2, \tag{1}$$

where $\alpha > 0$ is a tradeoff parameter, $\mathbf{P} \in \mathbb{R}^{d\times m}$ is the projection matrix, $\mathbf{F} \in \mathbb{R}^{m\times C}$ is the cluster centroids. $\mathbf{E}_s \in \mathbb{R}^{n_s\times C}$ is a constant matrix used to calculate the class centroids of source data in the projected space with each element $\mathbf{E}_{ij} = 1\Big/n_s^j$ if $y_{s,i} = j$, and $\mathbf{E}_{ij} = 0$ otherwise. $\hat{\mathbf{Y}}_t \in \mathbb{R}^{n_t\times C}$ is the one-hot encoded matrix of the predicted labels of the target domain.

3) Joint Probability Distribution Discrepancy: The core goal of domain adaptation is to match the different distributions in the source and target domains. The maximum mean discrepancy (MMD) criterion of marginal distribution and conditional distribution and their linear combination are commonly used for distribution alignment. Here, we use a more natural metric MMD criterion based on the joint probability distribution to measure the distribution difference between the source and target domains. The objective is to increase the discriminability between different classes while align the joint distributions of the source and target domains. Therefore, the joint probability distribution discrepancy is adopted and expressed as:

$$L_{jpd} = \mathcal{M}_T - \mu\mathcal{M}_D \tag{2}$$

with

$$\begin{aligned}\mathcal{M}_T &= \sum_{c=1}^{C} d\left(P_s\left(\mathbf{x}_s, y_s^c\right), P_t\left(\mathbf{x}_t, y_t^c\right)\right)\\ &= \sum_{c=1}^{C} d\left(P_s\left(\mathbf{x}_s\left|y_s^c\right.\right)P_s\left(y_s^c\right), P_t\left(\mathbf{x}_t\left|y_t^c\right.\right)P_t\left(y_t^c\right)\right),\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{M}_D &= \sum_{c=1}^{C}\sum_{\hat{c}\neq c} d\left(P_s\left(\mathbf{x}_s, y_s^c\right), P_t\left(\mathbf{x}_t, y_t^{\hat{c}}\right)\right)\\ &= \sum_{c=1}^{C}\sum_{\hat{c}\neq c} d\left(P_s\left(\mathbf{x}_s\left|y_s^c\right.\right)P_s\left(y_s^c\right), P_t\left(\mathbf{x}_t\left|y_t^{\hat{c}}\right.\right)P_t\left(y_t^{\hat{c}}\right)\right),\end{aligned} \tag{4}$$

where $P_s\left(\mathbf{x}_s\left|y_s^c\right.\right)$ represents the conditional probability, and $P_s\left(y_s^c\right)$ is the prior probability of class c in the source domain. According to the marginal distribution discrepancy $d\left(P_s\left(\mathbf{x}_s\right), P_t\left(\mathbf{x}_t\right)\right) = \left\|\mathbb{E}\left[\mathbf{P}^T\mathbf{x}_s\right] - \mathbb{E}\left[\mathbf{P}^T\mathbf{x}_t\right]\right\|_F^2$ and conditional distribution discrepancy $d\left(P_s\left(\mathbf{x}_s\left|y_s\right.\right), P_t\left(\mathbf{x}_t\left|y_t\right.\right)\right) = \sum_{c=1}^{C}\left\|\mathbb{E}\left[\mathbf{P}^T\mathbf{x}_s\left|y_s^c\right.\right] - \mathbb{E}\left[\mathbf{P}^T\mathbf{x}_t\left|y_t^c\right.\right]\right\|_F^2$ based on MMD, $\mathcal{M}_T$ and $\mathcal{M}_D$ are further expressed as:

$$\begin{aligned}&\mathcal{M}_T\\ &= \sum_{c=1}^{C}\left\|\mathbb{E}\left[\mathbf{P}^T\mathbf{x}_s\left|y_s^c\right.\right]P_s\left(y_s^c\right) - \mathbb{E}\left[\mathbf{P}^T\mathbf{x}_t\left|y_t^c\right.\right]P_t\left(y_t^c\right)\right\|_F^2\\ &= \sum_{c=1}^{C}\left\|\frac{1}{n_s}\sum_{i=1}^{n_s^c}\mathbf{P}^T\mathbf{x}_{s,i}^c - \frac{1}{n_t}\sum_{j=1}^{n_t^c}\mathbf{P}^T\mathbf{x}_{t,j}^c\right\|_F^2\\ &= \left\|\mathbf{P}^T\mathbf{X}_s\mathbf{N}_s - \mathbf{P}^T\mathbf{X}_t\hat{\mathbf{N}}_t\right\|_F^2,\end{aligned} \tag{5}$$

$$\begin{aligned}&\mathcal{M}_D\\ &= \sum_{c=1}^{C}\sum_{\hat{c}\neq c}\left\|\mathbb{E}\left[\mathbf{P}^T\mathbf{x}_s\left|y_s^c\right.\right]P_s\left(y_s^c\right) - \mathbb{E}\left[\mathbf{P}^T\mathbf{x}_t\left|y_t^{\hat{c}}\right.\right]P_t\left(y_t^{\hat{c}}\right)\right\|_F^2\\ &= \sum_{c=1}^{C}\sum_{\hat{c}\neq c}\left\|\frac{1}{n_s}\sum_{i=1}^{n_s^c}\mathbf{P}^T\mathbf{x}_{s,i}^c - \frac{1}{n_t}\sum_{j=1}^{n_t^{\hat{c}}}\mathbf{P}^T\mathbf{x}_{t,j}^{\hat{c}}\right\|_F^2\\ &= \left\|\mathbf{P}^T\mathbf{X}_s\mathbf{M}_s - \mathbf{P}^T\mathbf{X}_t\hat{\mathbf{M}}_t\right\|_F^2,\end{aligned} \tag{6}$$

where C is the number of categories, $\mu > 0$ is a trade-off parameter, and $\mathbb{E}[\cdot]$ denotes the mathematical expectation operation. Besides, $\mathbf{N}_s = \mathbf{Y}_s/n_s$ and $\hat{\mathbf{N}}_t = \hat{\mathbf{Y}}_t/n_t$, in which $\mathbf{Y}_s = \left[\mathbf{y}_{s,1}; \ldots; \mathbf{y}_{s,n_s}\right] \in \mathbb{R}^{n_s\times C}$ and $\hat{\mathbf{Y}}_t = \left[\hat{\mathbf{y}}_{t,1}; \ldots; \hat{\mathbf{y}}_{t,n_t}\right] \in \mathbb{R}^{n_t\times C}$ are the one-hot coding matrices of the true labels of the source samples and the predicted labels of the target samples, respectively. $\mathbf{M}_s = \mathbf{F}_s/n_s$ with $\mathbf{F}_s = \left[\mathbf{Y}_s(:,1), \ldots, \mathbf{Y}_s(:,C)\right] \otimes \mathbf{1}_{C-1}$ (the symbol $\otimes$ denotes the Kronecker product operation, and $\mathbf{1}_{C-1}$ is the all-one vector of dimension $C-1$), and $\hat{\mathbf{M}}_t = \hat{\mathbf{F}}_t/n_t$ with $\hat{\mathbf{F}}_t = \left[\hat{\mathbf{Y}}_t(:,2:C), \ldots, \hat{\mathbf{Y}}_t(:,[1:C]\setminus\{c\}), \ldots, \hat{\mathbf{Y}}_t(:,1:C-1)\right]$ ($\hat{\mathbf{Y}}_t(:,[1:C]\setminus\{c\})$ represents all but the c-th column of $\hat{\mathbf{Y}}_t$). $\mathcal{M}_T$ measures the distribution difference between the same classes of the source and target domains, and $\mathcal{M}_D$ measures the distribution difference between different classes of the two domains. Converted to the trace form, the joint probability distribution discrepancy can be rewritten as:

$$L_{jpd} = \mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{R}_T - \mu\mathbf{R}_D\right)\mathbf{X}^T\mathbf{P}\right), \tag{7}$$

where

$$\mathbf{R}_T = \begin{bmatrix}\mathbf{N}_s\mathbf{N}_s{}^T & -\mathbf{N}_s\hat{\mathbf{N}}_t^T\\ -\hat{\mathbf{N}}_t\mathbf{N}_s{}^T & \hat{\mathbf{N}}_t\hat{\mathbf{N}}_t^T\end{bmatrix}, \tag{8}$$

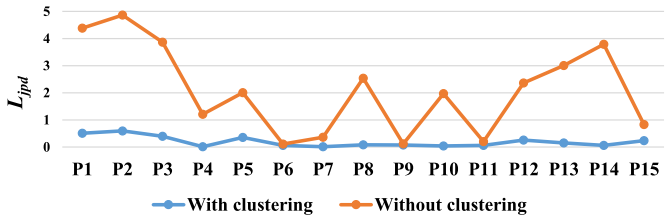


Fig. 2. L_{jpd} values in clustering and without clustering.

$$\mathbf{R}_D = \begin{bmatrix} \mathbf{M}_s \mathbf{M}_s{}^T & -\mathbf{M}_s \hat{\mathbf{M}}_t^T \\ -\hat{\mathbf{M}}_t \mathbf{M}_s{}^T & \hat{\mathbf{M}}_t \hat{\mathbf{M}}_t^T \end{bmatrix}. \tag{9}$$

To verify the effectiveness of the proposed method, the clustering based joint probability distribution discrepancy L_{jpd} obtained from the EEGs of 15 subjects in CHZU dataset (Table I) is derived. Meanwhile, comparisons to L_{jpd} obtained on without using the clustering method are also presented. As shown in Fig. 2, a smaller L_{jpd} value on almost all subjects can be derived in our proposed method than not adopting clustering. The comparison indicates that applying clustering can bring a positive effect to the joint probability distribution difference, thus enhancing the seizure detection performance.

4) Structure Consistency: The quality of cluster centroids plays an important role in whether the algorithm can accurately classify samples in the target domain. In real applications, many high-dimensional data are generally considered to reside in low-dimensional manifolds space with nonlinear geometric structures. Relevant studies [33] have shown that introducing the local manifold structure can improve the clustering performance of non-linear characteristic data. As one trivial but effective trick, we add a Laplacian regularization term to exploit the similar geometrical property of nearest points as:

$$L_{sc} = \frac{1}{2} \sum_{i,j=1}^{n_s+n_t} \mathbf{W}_{ij} \left\| \mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j \right\|^2 = \mathrm{tr}\left(\mathbf{P}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P} \right), \tag{10}$$

where $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t]$, $\mathbf{W}$ is the affinity matrix, defined as:

$$\mathbf{W}_{ij} = \begin{cases} \dfrac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}, & \mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}_j) \ \text{ or } \ \mathbf{x}_j \in \mathcal{N}_p(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ represents the inner product of $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathcal{N}_p(\mathbf{x}_i)$ denotes the set of p-nearest neighbors of point $\mathbf{x}_i$. The Laplacian matrix is $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with diagonal entries $\mathbf{D}_{ii} = \sum_{j=1}^{n_s+n_t} \mathbf{W}_{ij}$.

5) Regularization: In the knowledge transfer and manifold regularization, the structure of the data is constrained, but we do not want to lose the data attributes of the target domain. To avoid information loss, we introduce a regularization term to preserve the energy of the original signal:

$$\left\| \mathbf{X}_t - \mathbf{P}\mathbf{P}^T \mathbf{X}_t \right\|_F^2. \tag{12}$$

After performing several algebraic steps and constant term removal, the minimization problem of (12) can be written as:

$$L_r = -\mathrm{tr}\left(\mathbf{P}^T \mathbf{X} \hat{\mathbf{I}} \mathbf{X}^T \mathbf{P} \right) \tag{13}$$

where $\hat{\mathbf{I}}$ is a diagonal matrix defined as $\hat{\mathbf{I}}_{ii} = 1$ if $\mathbf{x}_i \in \mathbf{X}_t$, otherwise $\hat{\mathbf{I}}_{ii} = 0$.

6) Overall Formulation and Optimization Procedure: Then, by combining (1), (7), (10) and (13), we arrive at the final CEJT formulation:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{F}} \ & \left\| \mathbf{P}^T \mathbf{X} \mathbf{E} - \mathbf{F} \right\|_F^2 + \alpha \left\| \mathbf{P}^T \mathbf{X} \mathbf{V} - \mathbf{F} \mathbf{Y}^T \right\|_F^2 \\ & + \beta \|\mathbf{P}\|_F^2 + \lambda \mathrm{tr}\left(\mathbf{P}^T \mathbf{X} (\mathbf{R}_T - \mu \mathbf{R}_D) \mathbf{X}^T \mathbf{P} \right) \\ & + \rho \mathrm{tr}\left(\mathbf{P}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P} \right) - \mathrm{tr}\left(\mathbf{P}^T \mathbf{X} \hat{\mathbf{I}} \mathbf{X}^T \mathbf{P} \right) \\ \text{s.t.} \ & \mathbf{P}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{P} = \mathbf{I}_m \end{aligned} \tag{14}$$

where $\beta > 0$, $\lambda > 0$ and $\rho > 0$ are penalty parameters, $\mathbf{E} = \left[\mathbf{E}_s; \mathbf{0}_{n_t \times C} \right]$, $\mathbf{V} = \mathrm{diag}\left(\mathbf{0}_{n_s \times n_s}, \mathbf{I}_{n_t} \right)$ and $\mathbf{Y} = \left[\mathbf{0}_{n_s \times C}; \hat{\mathbf{Y}}_t \right]$. $\mathbf{H}$ is a centering matrix defined as $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{1}_n$, $n = n_s + n_t$. The constraint $\mathbf{P}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{P} = \mathbf{I}_m$ is introduced to avoid trivial solutions.

In (14), the labels of the target domain are needed for the projection clustering and the calculation of joint probability discrepancy. It is very difficult to obtain the best $\hat{\mathbf{Y}}_t$ by optimizing (14), so we solve it by assigning the label of each target sample to the nearest class centroid in optimization. Then:

$$\left(\hat{\mathbf{Y}}_t \right)_{ik} = \begin{cases} 1, & \text{if } \ k = \arg\min_j \left\| \mathbf{P}^T \mathbf{x}_{t,i} - \mathbf{F}(:, j) \right\|_2^2 \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

In addition, there are two variables $\mathbf{P}$ and $\mathbf{F}$ to optimize. We update each of them alternately while keeping the other variables fixed. When other variables are fixed, the optimization problem of $\mathbf{F}$ becomes:

$$\min_{\mathbf{F}} \left\| \mathbf{P}^T \mathbf{X} \mathbf{E} - \mathbf{F} \right\|_F^2 + \alpha \left\| \mathbf{P}^T \mathbf{X} \mathbf{V} - \mathbf{F} \mathbf{Y}^T \right\|_F^2 \tag{16}$$

Then, by taking the derivative of (16) with respect to $\mathbf{F}$, and setting the derivative to zero, we get:

$$\mathbf{F} = \left(\mathbf{P}^T \mathbf{X} \mathbf{E} + \alpha \mathbf{P}^T \mathbf{X} \mathbf{V} \mathbf{Y} \right) \left(\alpha \mathbf{Y}^T \mathbf{Y} + \mathbf{I} \right)^{-1} \tag{17}$$

Next, substituting (17) into (14) to replace $\mathbf{F}$, the optimization of $\mathbf{P}$ can be written as:

$$\begin{aligned} \min_{\mathbf{P}} \ & \mathrm{tr}\left(\mathbf{P}^T \begin{pmatrix} \mathbf{X} \mathbf{M} \mathbf{X}^T + \lambda \mathbf{X} (\mathbf{R}_T - \mu \mathbf{R}_D) \mathbf{X}^T \\ + \rho \mathbf{X} \mathbf{L} \mathbf{X}^T - \mathbf{X} \hat{\mathbf{I}} \mathbf{X}^T + \beta \mathbf{I}_d \end{pmatrix} \mathbf{P} \right) \\ \text{s.t.} \ & \mathbf{P}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{P} = \mathbf{I}_m \end{aligned} \tag{18}$$

where $\mathbf{Z} = (\mathbf{E} + \alpha \mathbf{V} \mathbf{Y}) \left(\alpha \mathbf{Y}^T \mathbf{Y} + \mathbf{I} \right)^{-1}$ and $\mathbf{M} = (\mathbf{E} - \mathbf{Z})(\mathbf{E} - \mathbf{Z})^T + \alpha \left(\mathbf{V} - \mathbf{Z} \mathbf{Y}^T \right) \left(\mathbf{V} - \mathbf{Z} \mathbf{Y}^T \right)^T$. According to the constrained optimization theory, the Lagrange multiplier Λ is introduced for optimization and the Lagrange function of (18) is:

$$\begin{pmatrix} \mathbf{X} \mathbf{M} \mathbf{X}^T + \lambda \mathbf{X} (\mathbf{R}_T - \mu \mathbf{R}_D) \mathbf{X}^T \\ + \rho \mathbf{X} \mathbf{L} \mathbf{X}^T - \mathbf{X} \hat{\mathbf{I}} \mathbf{X}^T + \beta \mathbf{I}_d \end{pmatrix} \mathbf{P} = \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{P} \Lambda \tag{19}$$

where $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$. Then the optimal solution is obtained by calculating the eigenvectors of (19) corresponding to the m-smallest eigenvalues. The proposed CEJT is summarised in Algorithm 1.

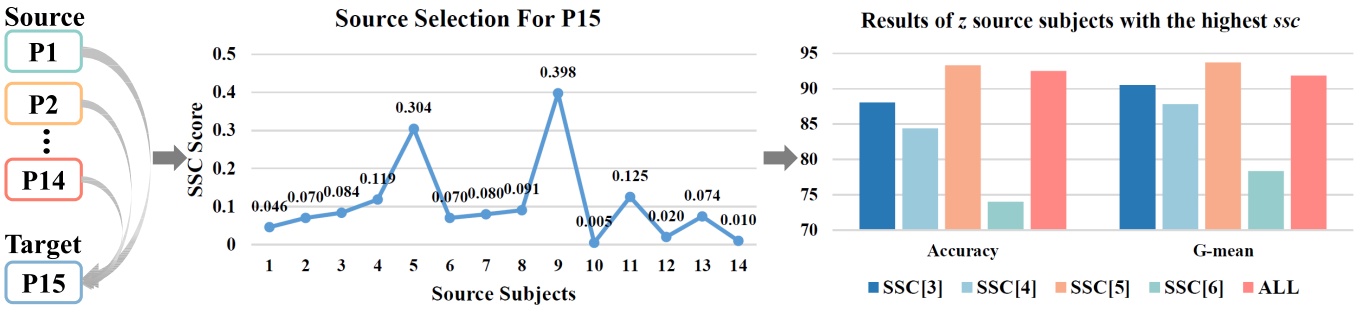


Fig. 3. The process of source selection: For P15, we first calculate the ssc scores of each source subject and P15 according to equation (20). Based on the results of different number of source subjects, the top 5 source subjects are finally selected.

Algorithm 1 CEJT

Input: Source data $\{\mathbf{X}_s, \mathbf{Y}_s\}$, target $\mathbf{X}_t$, penalty parameters α, β, λ and ρ, subspace dimensionality $m = 50$, maximum iteration $T = 10$

Output: Target label matrix $\hat{\mathbf{Y}}_t$.

1) Train the weak classifier to initialize the target label $\hat{\mathbf{Y}}_t$.
2) Compute the graph Laplacian matrix $\mathbf{L}$.
3) **repeat**
   a) Construct $\mathbf{R}_D$ and $\mathbf{R}_T$ by (8) and (9).
   b) Update $\mathbf{P}$ by solving the generalized eigenvalue problem in (19).
   c) Update $\mathbf{F}$ by Equation (17).
   d) Update $\hat{\mathbf{Y}}_t$ by Equation (15).
4) **until** Convergence or max iteration

C. Correlation-Alignment-Based Source Selection

Correlation alignment (CORAL) [34] minimizes the domain shift by the second-order statistics of source and target distributions. Inspired by the correlation alignment, we design an evaluation metric for source selection to find subjects that have a high correlation with the target domain. It thus can reduce the computational cost while avoide some negative transfer.

Assume there is a target domain $\mathbb{T}$ with unlabeled feature matrix $\mathbf{X}_t$, there have z labeled source domains $\mathbb{S}_i = \{\mathbf{X}_{s,i}, \mathbf{Y}_{s,i}\}_{i=1}^{z}$, where $\mathbf{X}_{s,i}$ is the feature matrix of the i-th source domain, the SSC between the i-th source domain and the target domain is defined as:

$$ssc\left(\mathbb{S}_i, \mathbb{T}\right) = \frac{dis\left(\mathbb{S}_i\right)}{dif\left(\mathbb{S}_i, \mathbb{T}\right)} = \frac{\sum_{c=1}^{C-1} \sum_{\hat{c}=c+1}^{C} \left\| \mathbf{C}_{\mathbb{S}_i^c} - \mathbf{C}_{\mathbb{S}_i^{\hat{c}}} \right\|_F^2}{\left\| \mathbf{C}_{\mathbb{S}_i} - \mathbf{C}_{\mathbb{T}} \right\|_F^2}, \quad (20)$$

where $\mathbf{C}_{\mathbb{S}_i}$ is the covariance of $\mathbf{X}_{s,i}$, $\mathbf{C}_{\mathbb{T}}$ is the covariance of $\mathbf{X}_t$, and $\mathbf{C}_{\mathbb{S}_i^c}$ represents the covariance of the c-th category in the source domain. The $dif\left(\mathbb{S}_i, \mathbb{T}\right)$ measures the distribution difference between the i-th source domain and the target domain, and $dis\left(\mathbb{S}_i\right)$ measures the inter-class discriminability of the i-th source domain. For the target domain $\mathbb{T}$, a larger $ssc\left(\mathbb{S}_i, \mathbb{T}\right)$ indicates a higher transferability of the i-th source domain. Therefore, we select $\hat{z} \in (1, z)$ source subjects with the highest $ssc\left(\mathbb{S}_i, \mathbb{T}\right)$.

We take the subject P15 as an target domain example to show the process of source selection in Fig. 3, where in the testing, all the rest subjects are taken as the source domain data. First, the ssc scores of P15 and each source subject are calculated, and the top $\hat{z}$ are selected. Then we test the effect of different number of source subjects on the classification results. Obviously, $\hat{z} = 5$ performs the best, also slightly better than using all subjects as the source (ALL). Similar results can be obtained for other patients when tested independently as the target domain.

IV. Results and Discussions

A. Experimental Settings

To show the effectiveness of the proposed transfer learning algorithm, experimental studies on the CHZU focal epilepsy dataset are carried out in this section. The accuracy, G-mean, sensitivity and F_1 score are used as the performance measure:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (21)$$

$$\text{G} - \text{mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \cdot \frac{\text{TN}}{\text{TN} + \text{FP}}} \quad (22)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (23)$$

$$\text{F}_1 = \frac{2 \cdot \text{P} \cdot \text{R}}{\text{P} + \text{R}} \quad (24)$$

where TP, TN, FP and FN denote the true positive, true negative, false positive and false negative detection, respectively. P and R are precision and recall rate, calculated by

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (25)$$

On the one hand, the proposed domain adaptation algorithm is compared with 2 classical intelligent methods without transfer learning abilities, i.e., SVM and KNN. On the other hand, the proposed CEJT algorithm is also compared with 7 classical domain adaptation approaches, i.e., TCA [12], JDA [13], TJM [17], ARTL [14], MEDA [15], JGSA [16] and our previous work Joint-Probability-Discrepancy-Based Domain Adaptation (JPDDA). By the way, JPDDA learns a domain-invariant classifier with structural risk minimization, while performing joint probability distribution discrepancy minimization, and manifold consistency maximization. Domain adaptation algorithms TCA, JDA, TJM and JGSA

learn through a transformation on all data in $\mathbf{X}_s$ and $\mathbf{X}_t$ for a common feature space across the source and target domains. Then the classification model is trained on the mapped source data using SVM. In our experiments, the parameters of the learning algorithms are optimized on the given search grids. For KNN, the optimal number of nearest neighbors is selected from $\{1, 2, \ldots, 10\}$. The best value of the trade-off parameter in SVM is searched on $\{2^{-6}, 2^{-4}, 2^{-2}, 2^0, 2^2, 2^4\}$. Besides TCA, other domain adaptation algorithms need to iteratively update the target-domain labels during the feature matching, where the iteration number is set to be $T = 10$. For TCA, JDA, TJM, JGSA and CEJT, the dimension of the common feature space is set to 50, and the manifold feature dimension of MEDA is also set to be 50. The optimal distribution adaptation parameters in all transfer learning algorithms (e.g., λ in our CEJT) are searched in the range of $\{2^{-6}, 2^{-4}, 2^{-2}, 2^0, 2^2, 2^4\}$. For the domain invariant classifiers ARTL, MEDA, JPDDA and CEJT, we obtain the optimal manifold regularization parameters by searching within $\{2^{-6}, 2^{-4}, 2^{-2}, 2^0, 2^2, 2^4\}$. Finally, the tradeoff parameter μ in CEJT is set to be 1, and the tradeoff parameter α is set to be 0.25. Specifically, in the proposed SSC method, we selected 5 source subjects with the highest $ssc(\mathbb{S}_i, \mathbb{T})$ for each subject to compose the source domain.

B. Comparisons Among Different Learning Methods

The detailed results of different algorithms for each subject are listed in the Table III. To be more clarity, the highest accuracy, G-mean, sensitivity and F_1 score are highlighted in bold font in the table. The results show that CEJT has the best classification performance. Obviously, domain adaptation algorithms are generally better than non-transfer learning algorithms, and the reason is that domain adaptation methods take into account the distribution differences between the source and target domains. It is worth noting that compared with the combination of feature transformation and classifier (e.g. TCA, JDA, TJM and JGSA), domain-invariant classifier (e.g. ARTL, MEDA, JPDDA and CEJT) perform better in jointing feature matching and classification. JPDDA and CEJT are superior to the state-of-the-art domain adaptation algorithms, thanks to the fact that the joint probability distribution difference strengthens the discriminative knowledge of classes while aligning the source and target domains. This is also confirmed by the tSNE visualization shown in Fig. 4, the EEG features extracted by JPDDA and CEJT are more distinguishable between different categories after the feature transformation. The clustering learning in CEJT further improves the performance by utilizing the data distribution structure of the target domain.

Further, we perform the statistical tests on the performance of the proposed algorithm and existing methods. The nonparametric Friedman test is used to evaluate whether the difference in performance among different methods is statistically significant. The rank of each algorithm is determined. The post-hoc test is then performed to verify that the difference between the top-ranked algorithm and the others is significant. Table IV shows the results of the Friedman test. In the table, if the p-value is less than the significance level $\alpha = 0.05$, it indicates the null hypothesis that all methods have the same classification performance is rejected. The proposed CEJT ranks first, outperforming other algorithms. Based on the results of the Friedman test, the post-hoc test is further performed to compare CEJT with other algorithms. The results in Table V show that the proposed algorithm significantly outperforms ARTL as well as algorithms ranked lower than ARTL. Meanwhile, it can be seen from Tables III and IV that the proposed algorithm outperforms MEDA and JPDDA to some extent although the improvement is not statistically significant.

Further, we visualize the decision boundary obtained by the JPDDA and the proposed CEJT for comparisons in Fig. 5, where in the figure, the data of the subject P06 is used as the target domain. The squared loss as a structural risk function of JPDDA, making it possible to classify the target domain by labeling the samples individually. While CEJT introduces the clustering to take advantage of the data structure of the target domain, which can adjust the labels of the target domain by clusters. Obviously, Fig. 5 confirms our assumption.

C. Ablation Study

We conduct the ablation experiments and analyze the significance of each loss in CEJT. The joint probability distribution discrepancy, structure consistency, and regularization are removed sequentially, and the average accuracy and G-mean are shown in Table VI. When the weight of the joint probability distribution discrepancy λ is set to 0, our method degenerates to a traditional clustering algorithm. With Table III, it is found that the overall performance is better than some transfer learning algorithms. A possible explanation is that in projection clustering, the class centroid of the source domain guides the clustering of the target domain, playing the role of aligning the distribution. When the weight of structural consistency ρ is set to be 0, the average accuracy and G-mean drops severely. It confirms that the structural consistency affects the quality of cluster centroids. In addition, it can be observed that the overall performance decreases slightly after removing the regularization term, indicating that focusing on the preservation of the original information can appropriately improve the performance.

D. Comparison Among Different Source Selection Strategies

This subsection validates the effectiveness of the proposed source selection strategy in finding the most beneficial source subjects. Fig. 6 shows the classification results when using different source selection methods: Euclidean distance (L_2), Earth Mover's distance (EMD), $\mathcal{A}$-distance and CORAL distance of source and target domains, Domain Transferability Estimation (DTE) [35]. ALL sources without selection is also included for comparison. As observed, the proposed SSC algorithm is superior to other selection strategies in both classification accuracy and G-mean score, even slightly higher than the unused selection strategy, which greatly reduces the computational cost and avoids the negative transfer caused by unrelated subjects to some extent. Specifically, compared with

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS ON CHZU DATASET

Subject	Evaluation Index	SVM	KNN	TCA	JDA	TJM	JGSA	ARTL	MEDA	JPDDA	CEJT
P01	Accuracy	67.16	57.49	76.62	75.49	65.96	87.71	84.18	84.18	87.29	**90.47**
	G-mean	59.02	27.04	68.15	66.33	45.29	85.73	82.15	81.70	83.78	**88.29**
	Sensitivity	62.97	49.57	72.12	70.78	58.81	86.13	82.58	82.30	84.80	**88.75**
	F_1	62.88	42.55	72.83	71.31	55.36	86.91	83.20	83.70	86.10	**89.78**
P02	Accuracy	85.27	78.20	91.75	92.42	90.74	95.62	94.53	95.03	95.54	**95.96**
	G-mean	75.56	59.19	83.97	88.32	81.34	**91.63**	91.50	91.54	91.45	91.56
	Sensitivity	77.17	64.72	84.99	88.63	82.79	**91.92**	91.67	91.76	91.74	91.15
	F_1	78.72	66.24	87.80	89.46	86.03	**93.77**	92.42	93.03	93.64	**93.77**
P03	Accuracy	48.54	49.52	54.72	52.55	50.24	59.63	58.88	58.08	63.93	**70.80**
	G-mean	41.87	36.57	49.60	45.54	37.59	59.28	57.71	56.39	63.57	**70.63**
	Sensitivity	50.39	51.96	56.51	54.53	52.67	60.38	59.95	59.29	64.74	**70.66**
	F_1	45.20	43.15	52.11	49.00	43.98	59.40	58.25	57.19	63.70	**70.66**
P04	Accuracy	78.25	70.74	76.40	80.78	79.40	75.36	75.07	76.86	81.71	**83.96**
	G-mean	82.66	48.12	84.41	85.38	77.36	81.77	82.19	84.57	88.19	**89.94**
	Sensitivity	82.90	53.55	85.28	85.64	77.41	82.32	82.88	85.36	88.71	**90.37**
	F_1	69.18	51.83	68.45	72.02	68.22	66.85	66.80	68.82	73.66	**76.12**
P05	Accuracy	88.77	73.05	92.79	91.96	76.95	**95.51**	93.03	93.03	95.39	95.39
	G-mean	82.14	49.01	88.29	85.40	45.53	**94.40**	89.91	89.76	94.19	94.19
	Sensitivity	83.21	59.17	88.8	86.42	60.37	**94.43**	90.17	90.05	94.23	94.23
	F_1	85.39	59.44	90.78	89.37	60.18	**94.54**	91.30	91.27	94.39	94.39
P06	Accuracy	77.44	50.83	80.00	80.30	65.26	74.89	86.62	86.92	88.27	**95.49**
	G-mean	76.80	33.69	79.73	80.07	60.59	73.71	86.96	87.27	88.67	**95.73**
	Sensitivity	79.42	55.09	81.78	82.05	68.36	77.09	87.81	88.08	89.32	**95.77**
	F_1	77.09	42.43	79.79	80.10	62.95	74.31	86.60	86.90	88.27	**95.47**
P07	Accuracy	77.69	66.53	82.87	80.88	77.09	83.47	86.45	**87.65**	84.86	86.25
	G-mean	70.89	48.00	75.78	73.30	66.67	79.69	84.71	**85.75**	84.39	85.62
	Sensitivity	73.73	59.63	78.71	76.56	71.86	80.83	85.03	**86.11**	84.42	85.67
	F_1	74.64	57.03	80.21	77.83	72.51	81.92	85.63	**86.85**	84.31	85.70
P08	Accuracy	59.70	69.12	67.69	61.12	**79.14**	66.77	58.55	58.23	57.31	59.15
	G-mean	48.38	47.67	58.82	48.48	53.22	**66.65**	53.76	54.55	62.15	65.22
	Sensitivity	51.16	55.03	60.48	51.79	62.41	**66.65**	54.37	54.93	63.21	66.61
	F_1	50.62	55.23	59.45	51.38	**64.18**	62.31	52.36	52.54	55.35	57.25
P09	Accuracy	73.11	81.35	68.19	78.93	78.68	78.10	83.68	84.01	89.34	**91.26**
	G-mean	59.53	55.09	45.66	61.62	41.77	77.30	72.89	73.89	83.51	**87.08**
	Sensitivity	62.93	64.56	53.90	66.49	58.36	77.32	74.96	75.73	84.15	**87.42**
	F_1	63.25	67.06	54.01	68.19	58.37	73.9	76.65	77.30	85.31	**88.14**
P10	Accuracy	62.47	65.68	61.48	60.74	67.65	72.35	65.68	67.16	73.33	**81.23**
	G-mean	53.99	59.38	52.22	51.50	64.02	71.39	62.27	61.71	72.21	**81.61**
	Sensitivity	59.96	63.51	58.87	58.14	66.08	71.70	64.18	65.15	72.60	**81.97**
	F_1	58.23	62.61	56.83	56.07	65.89	71.83	63.98	64.55	72.76	**81.23**
P11	Accuracy	59.41	69.25	76.35	70.86	**79.75**	77.58	76.16	79.28	79.56	75.88
	G-mean	67.69	51.33	82.64	76.15	79.43	70.45	78.31	80.96	**83.34**	82.98
	Sensitivity	69.07	55.13	83.22	76.58	79.43	71.06	78.38	81.00	83.53	**83.73**
	F_1	53.82	52.97	68.99	63.17	70.36	65.46	67.35	70.53	**71.55**	68.78
P12	Accuracy	69.83	45.10	70.14	67.81	61.74	69.35	71.70	74.49	88.49	**95.65**
	G-mean	69.05	25.83	68.85	65.60	57.61	68.15	71.84	74.60	88.64	**95.72**
	Sensitivity	71.24	47.79	71.41	69.30	63.53	70.59	72.11	74.63	89.11	**95.73**
	F_1	69.00	36.39	69.53	66.76	59.69	68.78	71.69	74.47	88.48	**95.63**
P13	Accuracy	72.07	67.49	75.39	63.48	75.39	72.27	83.59	80.27	92.19	**93.36**
	G-mean	64.62	63.77	70.68	59.31	67.06	72.88	81.79	76.96	92.23	**94.07**
	Sensitivity	68.28	65.01	72.57	61.02	71.25	73.01	82.19	78.05	92.23	**94.19**
	F_1	68.47	65.21	73.15	61.08	71.65	72.08	82.74	78.80	91.99	**93.27**
P14	Accuracy	54.40	31.45	66.53	63.70	53.69	64.11	67.95	70.07	78.26	**86.55**
	G-mean	58.87	24.71	71.80	69.10	58.11	68.51	70.69	72.60	82.68	**89.41**
	Sensitivity	61.81	47.32	74.85	73.19	66.01	70.29	71.16	72.97	83.93	**89.78**
	F_1	54.26	28.05	66.30	63.62	53.56	63.59	66.64	68.64	77.48	**85.49**
P15	Accuracy	80.58	79.00	76.64	87.27	86.09	83.86	86.75	83.07	86.09	**93.31**
	G-mean	69.97	50.78	68.24	78.08	76.85	71.95	88.59	84.27	87.52	**93.72**
	Sensitivity	71.67	60.31	69.40	79.30	78.10	73.98	88.65	81.44	87.56	**89.81**
	F_1	71.38	61.72	67.76	80.31	78.74	74.91	82.85	75.66	81.98	**83.50**
Average	Accuracy	70.31	63.65	74.50	73.89	72.52	77.11	78.19	78.56	82.77	**86.31**
	G-mean	65.40	45.35	69.92	68.95	60.83	75.57	77.02	77.10	83.10	**87.05**
	Sensitivity	68.39	56.82	72.85	72.02	67.82	76.51	77.73	77.79	83.61	**87.05**
	F_1	65.47	52.79	69.86	69.31	64.77	74.03	75.23	75.35	80.59	**83.94**

the CORAL distance, SSC not only considers the distribution difference between the source domain and the target domain, but also measures the discriminability of the source domain, making it more robust.

E. Comparison Among Different Seizure Detection Methods

We compare the proposed approach with a set of competitive state-of-the-art (SOTA) seizure detection algorithms.

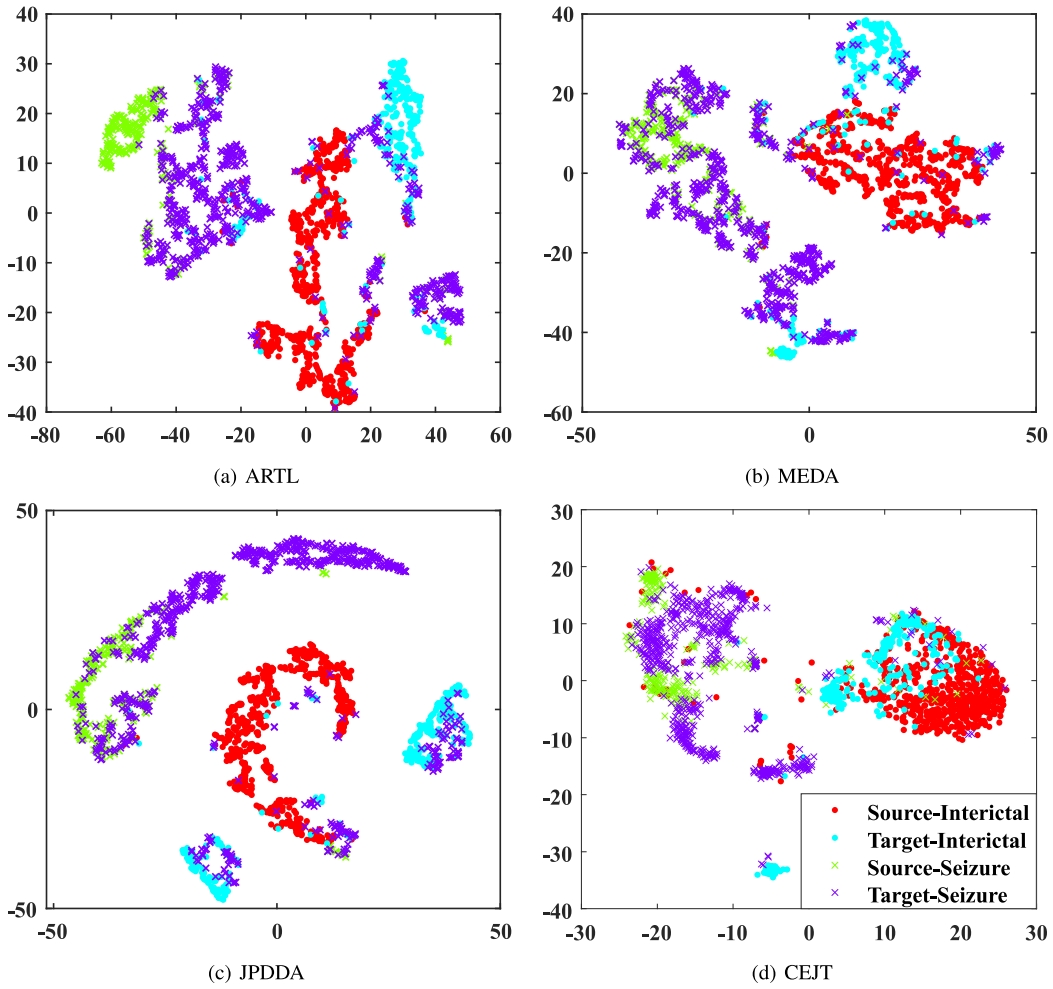


Fig. 4. t-SNE visualization of the data distributions with different domain adaptation approaches, when transferring P7 (source) to P16 (target). Compared with ARTL and MEDA, JPDDA and CEJT consider the MMD distance information between classes, and the inter-class discriminability is more obvious. Thanks to projection clustering, the source and target domains are better aligned and the samples are more compact after CEJT domain adaptation.

TABLE IV
FRIEDMAN TEST ON CLASSIFICATION PERFORMANCE OF COMPETITIVE ALGORITHMS

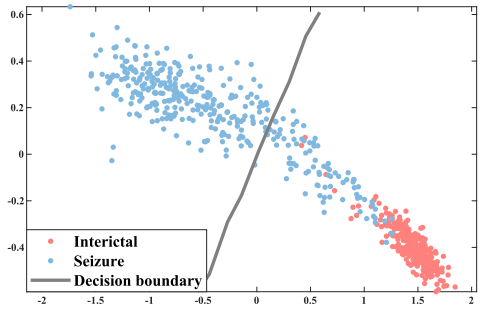
Algorithms	Rank	p-value	Null hypothesis
CEJT	**8.97**	2.68E-10	Rejected
JPDDA	8.00		
MEDA	6.67		
ARTL	6.23		
JGSA	6.06		
TCA	4.57		
TJM	4.53		
JDA	4.47		
SVM	3.13		
KNN	2.37		

TABLE V
HOLM POST-HOC TEST BETWEEN CEJT AND OTHER METHODS

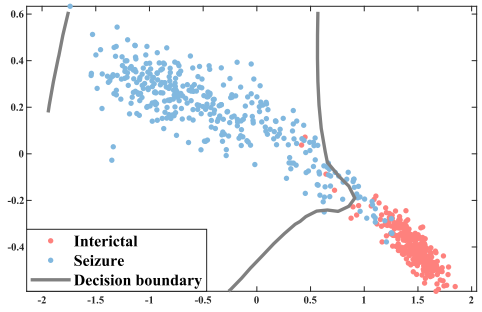
Rank(i)	Algorithms	p-value	Holm $= \alpha/i, \alpha = 0.05$	Null hypothesis
1	JPDDA	0.3813	0.05	Not Rejected
2	MEDA	0.0373	0.025	Not Rejected
3	ARTL	0.0133	0.016667	Rejected
4	JGSA	0.0086	0.0125	Rejected
5	TCA	6.75E-05	0.01	Rejected
6	TJM	5.95E-05	0.008333	Rejected
7	JDA	4.59E-05	0.007143	Rejected
8	SVM	1.27E-07	0.00625	Rejected
9	KNN	2.27E-09	0.005556	Rejected

Moreover, we try to compare with the simplest deep domain adaptation methods. The SOTA algorithms included for comparisons are:

- Cao et al. [32] adopt a stacked generalization model built on multiple CNNs with diverse activation functions and learning strategies for probability feature learning, and propose a novel adaptive weighting for fusion.
- Jiang et al. [36] extract regional multi-channel cross correlation EEG features combined with the convolutional autoencoder model based EEG feature dimensionality reduction method and ensemble classification model to achieve early seizure detection.
- Zhang et al. [31] improve the seizure-specific representations by eliminating inter-subject noise through



(a) JPDDA



(b) CEJT

Fig. 5. Decision boundaries visualization of JPDDA and CEJT on P06. Compared to JPDDA which uses squared loss, CEJT introduces clustering learning that, as hypothetically, utilizes the data structure of the target domain to adjust its labels.

TABLE VI
ABLATION STUDY ON THE PROPOSED ALGORITHM

Projected Clustering	√	√	√	√
Joint Probability Distribution Discrepancy		√	√	√
Structure Consistency	√		√	√
Regularization	√	√		√
Average Accuracy	79.27	75.89	84.07	**86.31**
Average G-mean	81.47	76.46	84.42	**87.05**

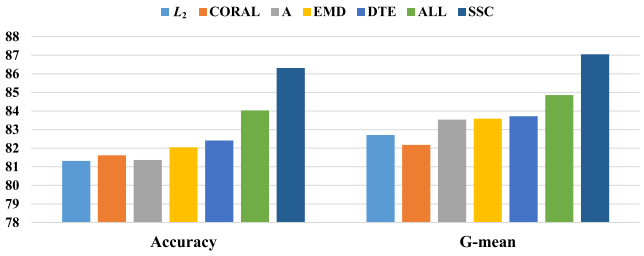


Fig. 6. Comparison of the selection methods of different source subjects.

adversarial training, resulting in a better cross-subject seizure detection model.

The compared deep domain adaptation methods with the mean amplitudes of sub-band spectrum (MAS) [32] as input include:

- Deep Domain Confusion (DDC) [19], which is a single-layer deep adaptation method with the MMD loss.
- Deep CORAL (DCORAL) [22], which is a deep neural network with the CORAL loss.

TABLE VII
CLASSIFICATION ACCURACY AND G-MEAN (%) OF DIFFERENT SEIZURE DETECTION MODEL

Subject	Evaluation Index	Cao	Jiang	Zhang	DDC	DCORAL	DANN	Ours
P01	Accuracy	62.92	72.81	72.74	79.80	78.81	80.86	**90.47**
	G-mean	61.79	67.38	59.74	79.44	77.17	79.48	**88.29**
P02	Accuracy	92.34	90.24	60.86	94.28	94.95	95.20	**95.96**
	G-mean	89.74	90.13	68.60	90.13	90.40	90.13	**91.56**
P03	Accuracy	64.47	68.03	52.61	62.89	68.45	68.77	**70.80**
	G-mean	63.82	67.60	48.54	62.96	67.34	67.71	**70.63**
P04	Accuracy	82.17	81.19	62.32	79.28	81.25	**85.92**	83.96
	G-mean	55.79	45.31	55.43	82.13	64.97	64.48	**89.94**
P05	Accuracy	91.02	91.49	64.42	**97.04**	96.57	96.57	95.39
	G-mean	83.65	84.10	70.07	**95.71**	94.73	94.73	94.19
P06	Accuracy	93.68	83.46	73.38	86.77	93.53	92.93	**95.49**
	G-mean	93.99	83.69	74.51	87.18	93.87	93.24	**95.73**
P07	Accuracy	85.66	80.28	64.34	**86.45**	84.86	84.46	86.25
	G-mean	85.13	81.67	54.17	**86.40**	86.02	85.90	85.62
P08	Accuracy	70.40	73.16	61.35	**82.99**	71.51	82.49	59.15
	G-mean	70.14	71.86	63.17	**81.28**	73.04	77.51	65.22
P09	Accuracy	76.52	81.52	64.53	83.18	85.43	85.35	**91.26**
	G-mean	56.86	73.33	63.90	74.40	75.94	70.18	**87.08**
P10	Accuracy	55.80	67.90	57.78	52.35	65.43	66.67	**81.23**
	G-mean	55.50	67.94	28.37	46.89	60.98	63.49	**81.61**
P11	Accuracy	**87.80**	85.71	55.91	73.89	86.38	85.62	75.88
	G-mean	70.83	68.39	56.63	53.60	67.10	69.49	**82.98**
P12	Accuracy	62.21	60.81	51.79	73.87	70.14	69.98	**95.65**
	G-mean	62.25	60.82	43.07	71.43	69.42	68.46	**95.72**
P13	Accuracy	73.83	77.15	66.21	89.84	89.84	90.23	**93.36**
	G-mean	65.47	71.15	63.77	89.59	89.36	89.92	**94.07**
P14	Accuracy	36.80	30.74	51.87	73.71	40.85	43.48	**86.55**
	G-mean	32.06	34.07	51.68	75.59	41.25	45.21	**89.41**
P15	Accuracy	87.01	91.08	82.15	87.14	89.89	88.84	**93.31**
	G-mean	74.18	88.11	78.79	81.68	84.60	81.54	**93.72**
Average	Accuracy	74.84	75.70	62.82	80.23	79.86	81.16	**86.31**
	G-mean	68.08	70.37	58.70	77.23	75.75	76.10	**87.05**

- Domain adversarial neural network (DANN) [37], which is a deep network that achieves the efficient domain transfer being indistinguishable between source and target data.

The overall performance of all the compared methods is reported in the Table VII. It is clearly observed that our method outperforms some state-of-the-art seizure detection models and simple deep domain adaptation methods, validating the effectiveness of our method in cross-subject seizure detection. The superiority of our method can be found on 11 subjects. It is worth noting that the model in [31] performs poorly, and one possible explanation is that the validation of the algorithm is in an ideal situation where the number of interictal and ictal samples is the same. In this study, most subjects has more interictal samples than ictal samples, which is more consistent with the actual situation. In addition, for the subject P08, the performance of the compared algorithms is much better than the proposed algorithm. The reason may be due to that the extracted wavelet packet features are not strong enough for EEG representation for this case, which also reminds us not only to pay attention to the establishment of classification models in the future, more attention needs to be paid to the attributes of the features themselves.

TABLE VIII
ACCURACY (%) OF DIFFERENT AMOUNTS OF TARGET DATA FOR TRANSFER

Percentage used for transfered		P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	Average
1/3	Transfer	87.5	97.2	69.6	87.2	95.0	95.9	91.6	56.1	86.3	75.6	74.2	93.5	94.1	83.6	93.7	**85.4**
	Test	86.2	95.2	66.6	81.6	94.3	91.4	88.7	53.0	81.2	77.4	85.2	94.4	90.4	84.1	91.7	84.0
	All	87.5	95.9	67.6	83.4	94.6	92.9	89.6	54.0	82.9	76.8	74.8	94.1	91.6	83.9	92.4	84.1
1/2	Transfer	90.1	96.3	70.1	81.5	96.2	95.2	91.2	57.7	90.5	84.2	75.0	97.2	92.6	88.9	95.5	**86.8**
	Test	89.7	95.8	67.4	81.3	94.8	91.3	92.8	57.9	90.7	77.8	73.0	94.7	91.4	85.5	91.1	85.0
	All	89.9	96.0	68.8	81.4	95.5	93.2	92.0	57.8	90.6	81.0	74.0	96.0	92.0	87.2	93.3	85.9
2/3	Transfer	91.3	95.3	71.0	81.1	95.6	93.7	86.9	56.9	89.9	82.2	72.1	95.1	92.4	88.8	91.7	**85.6**
	Test	89.8	96.7	66.1	90.2	93.3	92.3	89.8	55.3	88.0	76.3	73.6	95.8	92.4	83.3	89.8	84.8
	All	90.8	95.8	69.4	80.8	94.8	93.2	87.9	56.3	89.9	80.3	72.6	95.3	92.4	87.0	91.1	85.2

F. Analysis of Varying Number of Transferred Target Data

In above experiments, all unlabeled target data are used for transfer learning. In this section, we show the performance with varying amount of target data in transfer learning. First, the unlabeled target data (denoted as All) is divided into two parts, one part is used to measure the distribution difference with the source domain in the training model (denoted as Transfer), and the other is used for testing which is not visible during training (denoted as Test). One-third, one-half, and two-thirds of the target data are used to measure distribution differences, respectively, as shown in Table VIII. There is no doubt that in any case the average accuracy of Transfer learning is better than that for Test. When one-third of the target data is used for transfer learning, the average accuracy of the Test is the lowest (84.0%). When more data is used in transfer learning, the average accuracy of Test is improved, reaching a maximum of 85.0%. The average accuracy of All reaches the highest 85.9% when half of the data is used for transfer learning, only 0.4% lower than that recorded in the Table III. Therefore, it is sufficiently feasible to use partial data to measure the distribution of the target domain for transfer learning. This also shows that the proposed framework has low data constraints for practical applications.

V. CONCLUSION

The effectiveness of domain adaptation has been demonstrated in seizure detection to cope with variations among different subjects or tasks. In the proposed cross-subject seizure detection framework, when the number of source subjects is large, the source selection evaluation metric SSC can reduce the computational cost and reduce the impact of irrelevant subjects on the subsequent classification modeling. CEJT organically unifies clustering learning, feature matching and discriminative structure, and performs well in solving individual differences in EEG signals. However, the number of source subjects selected is obtained through simple experiments, which lacks the individual adaptability, and we will make new explorations on this issue in the future.

ETHICAL STANDARDS

This study has been approved by the Second Affiliated Hospital of Zhejiang University and registered in Chinese Clinical Trail Registry (ChiCTR1900020726). All patients gave their informed consent prior to their inclusion in the study.

REFERENCES

[1] M. K. Siddiqui, R. Morales-Menendez, X. Huang, and N. Hussain, "A review of epileptic seizure detection using machine learning classifiers," *Brain Informat.*, vol. 7, no. 1, pp. 1–18, Dec. 2020.
[2] M. Wang et al., "Multidimensional feature optimization based eye blink detection under epileptiform discharges," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 905–914, 2022.
[3] Z. Xu, T. Wang, J. Cao, Z. Bao, T. Jiang, and F. Gao, "BECT spike detection based on novel EEG sequence features and LSTM algorithms," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1734–1743, 2021.
[4] R. Zheng, J. Cao, Y. Feng, X. Zhao, T. Jiang, and F. Gao, "Seizure prediction analysis of infantile spasms," *IEEE Trans. Neural Syst. Rehabil. Eng.*, early access, Nov. 17, 2022, doi: 10.1109/TNSRE.2022.3223056.
[5] J. Cao et al., "Two-stream attention 3D deep network based childhood epilepsy syndrome classification," *IEEE Trans. Instrum. Meas.*, early access, Nov. 7, 2022, doi: 10.1109/TIM.2022.3220287.
[6] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal–spectral squeeze-and-excitation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 782–794, Apr. 2020.
[7] D. Hu, J. Cao, X. Lai, J. Liu, S. Wang, and Y. Ding, "Epileptic signal classification based on synthetic minority oversampling and blending algorithm," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 2, pp. 368–382, Jun. 2021.
[8] F.-G. Tang, Y. Liu, Y. Li, and Z.-W. Peng, "A unified multi-level spectral–temporal feature learning framework for patient-specific seizure onset detection in EEG signals," *Knowl.-Based Syst.*, vol. 205, Oct. 2020, Art. no. 106152.
[9] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021.
[10] G. Dal Canto, S. Pellacani, G. Valvo, G. Masi, A. R. Ferrari, and F. Sicca, "Internalizing and externalizing symptoms in preschool and school-aged children with epilepsy: Focus on clinical and EEG features," *Epilepsy Behav.*, vol. 79, pp. 68–74, Feb. 2018.
[11] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2014.
[12] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[13] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2013, pp. 2200–2207.
[14] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.
[15] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 402–410.

[16] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1859–1867.

[17] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1410–1417.

[18] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6103–6115, Dec. 2019.

[19] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[20] 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.

[21] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[22] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–450.

[23] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 778–786.

[24] C. Yang, Z. Deng, K.-S. Choi, Y. Jiang, and S. Wang, "Transductive domain adaptive learning for epileptic electroencephalogram recognition," *Artif. Intell. Med.*, vol. 62, no. 3, pp. 165–177, 2014.

[25] Y. Jiang et al., "Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2270–2284, Dec. 2017.

[26] C. Yang, Z. Deng, K.-S. Choi, and S. Wang, "Takagi–Sugeno–Kang transfer learning fuzzy logic system for the adaptive recognition of epileptic electroencephalogram signals," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1079–1094, Oct. 2016.

[27] L. Xie, Z. Deng, P. Xu, K.-S. Choi, and S. Wang, "Generalized hidden-mapping transductive transfer learning for recognition of epileptic electroencephalogram signals," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2200–2214, Jun. 2019.

[28] K. Xia, T. Ni, H. Yin, and B. Chen, "Cross-domain classification model with knowledge utilization maximization for recognition of epileptic EEG signals," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 53–61, Jan. 2021.

[29] Y. Zhang et al., "Epilepsy signal recognition using online transfer TSK fuzzy classifier underlying classification error and joint distribution consensus regularization," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 5, pp. 1667–1678, Sep. 2021.

[30] B. Zhang et al., "Cross-subject seizure detection in EEGs using deep transfer learning," *Comput. Math. Methods Med.*, vol. 2020, May 2020, Art. no. 7902072.

[31] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, "Adversarial representation learning for robust patient-independent epileptic seizure detection," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2852–2859, Oct. 2020.

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

[33] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–15.

[34] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1–8.

[35] W. Zhang and D. Wu, "Manifold embedded knowledge transfer for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1117–1127, May 2020.

[36] T. Jiang, J. Zhu, D. Hu, W. Gao, F. Gao, and J. Cao, "Early seizure detection in childhood focal epilepsy with electroencephalogram feature fusion on deep autoencoder learning and channel correlations," *Multidimensional Syst. Signal Process.*, vol. 33, no. 4, pp. 1–21, 2022.

[37] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.