Generalizable Representation Learning for fMRI-based Neurological Disorder Identification

Anonymous authors
Paper under double-blind review

Abstract

Despite the impressive advances achieved using deep learning for functional brain activity analysis, the heterogeneity of functional patterns and the scarcity of imaging data still pose challenges in tasks such as identifying neurological disorders. For functional Magnetic Resonance Imaging (fMRI), while data may be abundantly available from healthy controls, clinical data is often scarce, especially for rare diseases, limiting the ability of models to identify clinically-relevant features. We overcome this limitation by introducing a novel representation learning strategy integrating meta-learning with self-supervised learning to improve the generalization from normal to clinical features. This approach enables generalization to challenging clinical tasks featuring scarce training data. We achieve this by leveraging self-supervised learning on the control dataset to focus on inherent features that are not limited to a particular supervised task and incorporating meta-learning to improve the generalization across domains. To explore the generalizability of the learned representations to unseen clinical applications, we apply the model to four distinct clinical datasets featuring scarce and heterogeneous data for neurological disorder classification. Results demonstrate the superiority of our representation learning strategy on diverse clinically-relevant tasks.

1 Introduction

Deep learning based approaches have demonstrated success in analyzing brain connectivity based on functional magnetic resonance imaging (fMRI) (Gadgil et al., 2020; Ahmedt-Aristizabal et al., 2021), but the scarcity and heterogeneity of fMRI data still pose challenges in clinical applications such as identifying neurological disorders. fMRI plays a vital role in identifying biomarkers for neurological disorders (Pitsik et al., 2023; Li et al., 2021). However, clinical fMRI datasets are not only high-dimensional and spatiotemporally complex but are also characterized by high variability among subjects and a limited number of data, posing significant challenges for training deep learning models to predict neurological disorders (Akrami et al., 2021). Medical datasets typically contain an abundance of healthy control data but often face a scarcity of clinical data collected for any particular neurological disorder. Simply train a model by aggregating all healthy control and clinical data may cause limited generalization and bias due to the data imbalance and heterogeneity in clinical features. This can in-turn lead to poor performance in the group with clinical pathology (Azizi et al., 2023). Moreover, deep learning models often struggle to achieve satisfactory performance on small-scale clinical datasets, frequently under-performing compared to traditional machine learning methods (Akrami et al., 2024; 2021). Considering the limitations of both data and existing deep learning models, we explore representation learning on fMRIs, aiming to extract meaningful and generalizable features from data by learning inherent functional activity patterns. Driven by the need to learn generalizable representations, we leverage self-supervised learning and meta-learning as the foundation of our representation learning approach.

Self-supervised learning is popular in representation learning and has shown the ability to improve the generalization of features (Reed et al., 2022; You et al., 2020b). In contrast to fully-supervised tasks such as classification or segmentation, self-supervised tasks are typically designed to learn intrinsic features that are not specific to a particular task (Taleb et al., 2020). Contrastive self-supervised learning applied to fMRI classification has demonstrated the ability to prevent over-fitting on small medical datasets and address high intra-class variances (Wang et al., 2022). For our proposed approach, we apply contrastive self-supervised learning, known to be effective in representation learning (Azizi et al., 2023), to the healthy control data to learn more generalizable features.

Now that self-supervised learning is applied to learn generalizable representations from abundant healthy control data, we need an effective approach to transfer the learned knowledge to the scarce clinical data. For this, we adopt a meta-representation learning approach (Liu et al., 2020a), which leverages meta-learning to enhance generalization across domains. Meta-learning has recently gained tremendous attention because of its learning-to-learn mechanism, which strongly increases the generalizability of models across different tasks (Zhang et al., 2019; Liu et al., 2020a; Finn et al., 2017). By employing a bi-level optimization scheme (Finn et al., 2017), the model is trained to generalize effectively to unseen domains. Meta-learning is particularly effective in a low-data regime (Zhang et al., 2019), making it well-suited for clinical applications.

In this work, we introduce a novel representation learning strategy, Meta Transfer of Self-supervised Knowledge (MeTSK), which leverages meta-learning to generalize self-supervised features from large-scale control datasets (source domain) to scarce clinical datasets (target domain). MeTSK not only enables effective knowledge transfer from source to target domains but also enhances the model's ability to generalize to new and unseen clinical data in challenging applications by leveraging the learned generalization from control features to scarce clinical features. In summary, our contribution is three-fold:

- We are the first to propose a novel representation learning approach for fMRI data to achieve generalization across various challenging neurological disorder classification tasks with limited data;
- The proposed approach can serve as a reliable feature extractor for future challenging tasks with limited data, where deep learning methods typically fail.
- We address the heterogeneity and scarcity of clinical fMRI data through the integration of meta-learning and self-supervised learning.

Our experiments are designed to demonstrate, i). the improved knowledge transfer from source to target datasets, we use a neurological disorder classification task to evaluate the performance on the target domain when applying MeTSK in a knowledge transfer task setting. ii) The generalization of representations to unseen clinical datasets. We evaluate the MeTSK model pre-trained with a source and a target dataset on unseen clinical datasets. Direct training of deep learning models on these challenging clinical datasets performed poorly due to limited training data and the heterogeneity of features, also resulting in worse performance compared to simpler machine learning classifiers. So acquiring generalizable representations is crucial for these clinical datasets. According to (Kumar et al., 2022), fine-tuning can distort good pre-trained features and degrade downstream performance under large distribution shifts. Here we explored linear probing for evaluating the generalization of representations on distinct neurological disorder classification tasks. As we show below, we are able to consistently achieve superior classification performance for diverse neurological disorder identification tasks compared to linear classification directly using traditional functional connectivity features as input.

2 Related Work

fMRI data are widely used for identifying neurological disorders such as Alzheimer's Disease (AD) (LaMontagne et al., 2019), epilepsy (Gullapalli, 2011), Parkinson's Disease (PD) (Badea et al., 2017), and Attention-Deficit/Hyperactivity Disorder (ADHD) (Bellec et al., 2017). These disorders cause atypical brain activity that can be characterized by analyzing fMRI data. In a traditional setting, features such as functional connectivity between brain regions (Van Den Heuvel & Pol, 2010) and Amplitude of Low-Frequency Fluctuation (ALFF) features (Zou et al., 2008), are extracted from raw fMRIs for statistical analysis or as inputs to machine learning classifiers to identify neurological biomarkers (Akrami et al., 2024). Given the inherent graph structure of fMRI data, where brain regions can be considered as nodes and functional connectivity measures as the edges, Graph Neural Networks (GNNs) (Li et al., 2021; Gadgil et al., 2020; Wang et al., 2022) have emerged as the predominant approaches in the literature for deep learning. GNNs usually take graph-structured data as input and perform graph convolution based on the neighboring relations (defined by edges) between graph nodes (Kipf & Welling, 2016). Li et al. (2021) proposed a GNN model with a novel pooling strategy for Autism Spectrum Disorder (ASD) classification; Zhang et al. (2023) applied a local-to-global GNN to ASD and AD classification, which uses a population graph to make use of inter-subject correlations in the dataset. In this work, we adopted a popular spatio-temporal GNN as our backbone model.

In fMRI analysis, heterogeneity and disparities across datasets pose significant challenges for the generalization of models. Most existing methods developed for fMRIs focus on adapting models between closely related domains, such as generalizing across imaging sites within the same dataset. These methods are often designed to address site-specific variations, such as differences in imaging protocols or scanner types (Li et al., 2020; Shi et al., 2021; Liu et al., 2023). Other approaches, such as fine-tuning (Raghu et al., 2019) and multi-task learning (Huang et al., 2020), attempt to adapt to target domains but often fail to learn transferable features due to domain discrepancies (Liu et al., 2020b; Raghu et al., 2019) and may distort the learning of target-domain specific features Chen et al. (2019).

Despite these efforts, there is no existing work exploring the generalization across domains with fundamentally different characteristics, such as healthy control data and clinical data from patients with neurological disorders. To mitigate this gap, we explore a representation learning strategy that seeks to enhance generalization to unseen domains not available during training. This challenge, typical in real-world clinical applications (Akrami et al., 2024; Badea et al., 2017), where new datasets may vary significantly and are often scarce, makes traditional deep learning methods less applicable. Our approach also serves as a reliable feature extractor for small-scale datasets, thereby enabling robust neurological disorder identification.

3 Methods

In this section, we introduce our proposed strategy, MeTSK, which improves the generalization of self-supervised fMRI features from a control dataset to clinical datasets. The proposed network architecture consists of a feature extractor that learns generalizable features from both source (control) and target (clinical) domains, and source and target heads to learn domain-specific features for the source and target domain, respectively. The bi-level optimization strategy is applied to learn generalizable features using a Spatio-temporal Graph Convolutional Network (ST-GCN) (Gadgil et al., 2020) as the backbone model. After MeTSK is trained on the control and clinical datasets, its generalization capability will be evaluated on unseen clinical datasets using linear probing. The methodology of MeTSK as well as the representation learning pipeline are shown in Fig. 1.

3.1 Feature Extractor: ST-GCN

We adopt a popular model for fMRI classification, ST-GCN (Gadgil et al., 2020), as the backbone architecture to extract graph representations from both spatial and temporal information. A graph convolution and a temporal convolution are performed in one ST-GCN module shown in Fig. 2, following the details in (Gadgil et al., 2020). The feature extractor includes three ST-GCN modules. The target head and the source head share the same architecture, which consists of one ST-GCN module and one fully-connected layer.

To construct the graph, we treat brain regions parcellated by a brain atlas (Glasser et al., 2016) as the nodes and define edges using the functional connectivity between pairs of nodes measured by Pearson's correlation coefficient (Bellec et al., 2017). We randomly sample fixed-length sub-sequences from the whole fMRI time series to increase the size of training data by constructing multiple input graphs containing dynamic temporal information. For each time point in each node, a feature vector of dimension C_i is learned. So for the r-th sub-sequence sample from the n-th subject, the input graph $X_i^{(n,r)}$ to the i-th layer has a dimension of $P \times L \times C_i$, where P is the number of brain regions or parcels (nodes), L is the length of the sampled sub-sequence, and $C_0 = 1$ for the initial input. In ST-GCN, a graph convolution (Kipf & Welling, 2016), applied to the spatial graph at time point l in the i-th layer, can be expressed as follows.

$$X_{i+1}^{(n,r,l)} = D^{-1/2}(A+I)D^{-1/2}X_i^{(n,r,l)}W_{C_i \times C_{i+1}} \tag{1}$$

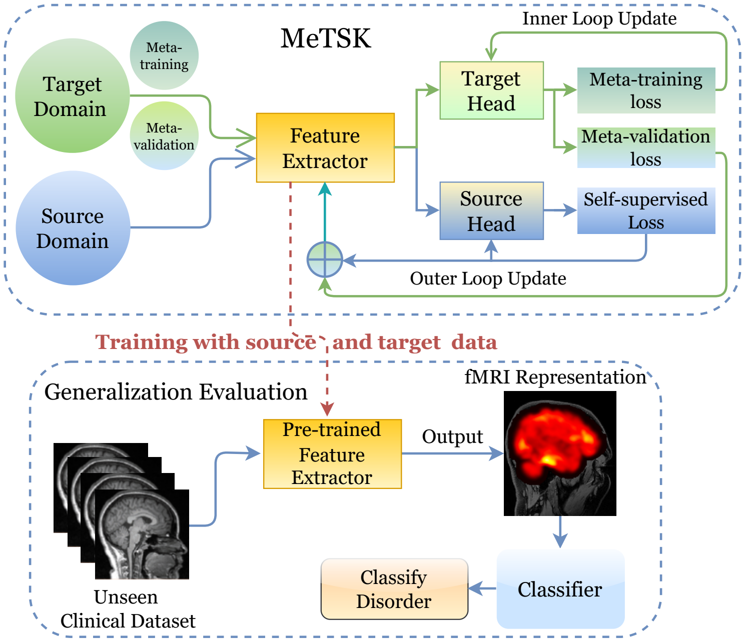

Figure 1: An illustration of MeTSK for generalizable representation learning. In MeTSK, two optimization loops are involved in training. The inner loop only updates the target head, while the outer loop updates the source head and feature extractor. The representation learning pipeline involves first training MeTSK on the source and target data, and then evaluate the learned representations on unseen clinical datasets using neurological disorder classification tasks.

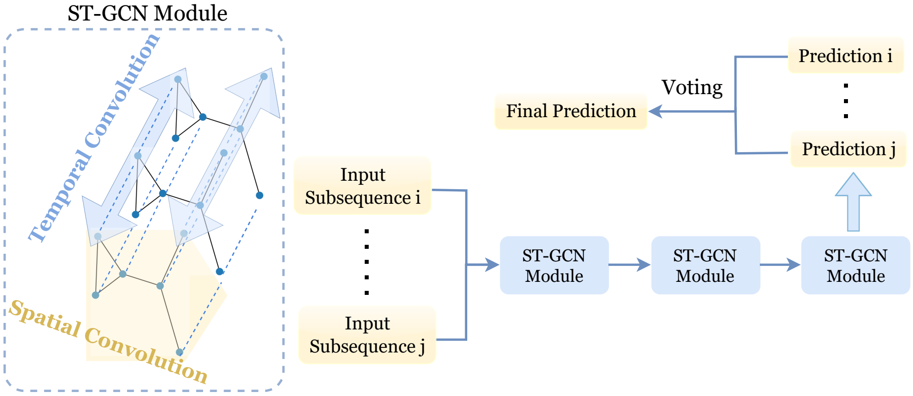

Figure 2: An illustration of the ST-GCN model architecture. Spatial graph convolution is first applied to the spatial graph at each time point. Then 1D temporal convolutions are performed along the resulting features on each node. Multiple sub-sequences are randomly sampled from the whole time series as input graphs for training.

where A is the adjacency matrix consisting of edge weights defined as Pearson's correlation coefficients, I is the identity matrix, D is a diagonal matrix such that $D_{ii} = \sum_j A_{ij} + 1$, and W is a trainable weight matrix. We then apply 1D temporal convolution to the resulting sub-sequence of features on each node. A voting strategy is applied to combine predictions generated from different sub-sequences.

3.2 Bi-level Optimization

Assume there exists a source domain (healthy controls) $\mathcal{S}$ with abundant training data $X_{\mathcal{S}}$ and a target domain (clinical) $\mathcal{T}$, where the training data $X_{\mathcal{T}}$ is limited. A feature extractor $f(\phi)$, a target head $h_{\mathcal{T}}(\theta_t)$, and a source head $h_{\mathcal{S}}(\theta_s)$ are constructed to learn source features $h_{\mathcal{S}}(f(X_{\mathcal{S}};\phi);\theta_s)$ as well as target features $h_{\mathcal{T}}(f(X_{\mathcal{T}};\phi);\theta_t)$, where ϕ, θ_t, and θ_s are model parameters.

We introduce a bi-level optimization strategy to perform gradient-based update of model parameters (Finn et al., 2017; Liu et al., 2020a). The model first back-propagates the gradients through only the target head in several fast adaptation steps, and then back-propagates through the source head and feature extractor. Each step in a nested loop is summarized as follows:

Outer loop (M iterations): `Step 1`. Initialize the target head and randomly sample target meta-training set $X_{\mathcal{T}_{tr}}$ and meta-validation set $X_{\mathcal{T}_{val}}$ from $X_{\mathcal{T}}$, where $X_{\mathcal{T}_{tr}} \bigcap X_{\mathcal{T}_{val}} = \emptyset$, $X_{\mathcal{T}_{tr}} \bigcup X_{\mathcal{T}_{val}} = X_{\mathcal{T}}$.

`Step 2.` **Inner loop** (k update steps): Only target head parameters θ_t are updated using optimization objective $\mathcal{L}_{\mathcal{T}}$ (see below) for the target task. The parameter α is the inner loop learning rate, and θ_t^j is the target head parameter at the j-th update step.

$$\theta_t^{j+1} = \theta_t^j - \alpha \nabla_{\theta_t^j} \mathcal{L}_{\mathcal{T}}(h_{\mathcal{T}}(f(X_{\mathcal{T}_{tr}};\phi^i);\theta_t^j)) \tag{2}$$

`Step 3`: After the inner loop is finished, freeze the target head and update feature extractor parameters ϕ and source head parameters θ_s. The target loss $\mathcal{L}_{\mathcal{T}}$ and source loss $\mathcal{L}_{\mathcal{S}}$ are defined in the following section. The parameter β is the outer loop learning rate, and λ is a scaling coefficient.

$$\begin{aligned}\{\theta_s^{i+1}, \phi^{i+1}\} = \{\theta_s^i, \phi^i\} - \beta(&\nabla_{\theta_s^i,\phi^i} \mathcal{L}_{\mathcal{S}}(h_{\mathcal{S}}(f(X_{\mathcal{S}};\phi^i);\theta_s^i)) \\ &+ \nabla_{\phi^i} \lambda \mathcal{L}_{\mathcal{T}}(h_{\mathcal{T}}(f(X_{\mathcal{T}_{val}};\phi^i);\theta_t^k)))\end{aligned} \tag{3}$$

The target head, source head and feature extractor are updated in an alternating fashion. The target head is first trained on $X_{\mathcal{T}_{tr}}$ in the inner loop. In the outer loop, the feature extractor and source head are trained to minimize the generalization error of the target head on an unseen set $X_{\mathcal{T}_{val}}$ as well as to minimize the source loss. In this way, the feature extractor encodes features beneficial for both domains and the source head extracts features from the source domain that enable generalization to the target domain.

3.3 Contrastive Self-supervised Learning

To further boost the generalizability of features, we apply a graph contrastive loss (You et al., 2020a) to perform a self-supervised task on the source domain. We randomly sample sub-sequences $X^{(n,r_1)}$, $X^{(n,r_2)}$ ($r1 \neq r2$) from the whole fMRI time series for subject n as the input graph features (Gadgil et al., 2020), which can be viewed as an augmentation of input graphs for ST-GCN. $X^{(n,r_1)}$ and $X^{(n,r_2)}$ should produce similar output graph features even though they contain different temporal information. The graph contrastive loss enforces similarity between graph features extracted from the same subject and dissimilarity between graph features extracted from different subjects (Chen et al., 2020b), so that the model learns invariant functional activity patterns across different time points for the same subject and recognizes inter-subject variances. A cosine similarity is applied to measure the similarity in the latent graph feature space (You et al., 2020a).

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{N}\sum_{n=1}^{N} -\log \frac{\exp\left(sim(\tilde{X}_{\mathcal{S}}, n, n)/\tau\right)}{\sum_{m=1, m\neq n}^{N} \exp\left(sim(\tilde{X}_{\mathcal{S}}, n, m)/\tau\right)} \tag{4}$$

$$sim(X, n, m) = \frac{(X^{(n,r_1)})^\top X^{(m,r_2)}}{\|X^{(n,r_1)}\| \cdot \|X^{(m,r_2)}\|} \tag{5}$$

where $\tilde{X}_{\mathcal{S}} = h_{\mathcal{S}}(f(X_{\mathcal{S}};\phi);\theta_s)$ is the generated graph representation, τ is a temperature hyper-parameter, and N is the total number of subjects in one training batch. By minimizing the graph contrastive loss on the source domain, the model produces consistent graph features for the same subject and divergent graph features across different subjects, which may be related to latent functional activities that reveal individual differences. Such intrinsic features are likely to be invariant across domains, promoting better generalization to unseen data.

The optimization objective $\mathcal{L}_{\mathcal{T}}$ of the target domain depends on the target task. In a classification task with class labels $Y_{\mathcal{T}}$, we adopt the Cross-Entropy loss. The total loss for the proposed MeTSK strategy is then:

$$\begin{aligned} \mathcal{L}_{meta} &= \mathcal{L}_{\mathcal{S}} + \lambda \mathcal{L}_{\mathcal{T}} \\ \mathcal{L}_{\mathcal{T}} &= -\sum_{\text{classes}} Y_{\mathcal{T}} \log(h_{\mathcal{T}}(f(X_{\mathcal{T}};\phi);\theta_t)) \end{aligned} \tag{6}$$

3.4 Linear Probing

The bi-level optimization strategy and contrastive self-supervised learning work together to learn robust, domain-invariant features, enhancing the ability to generalize effectively to unseen clinical datasets. Here we evaluate the generalization of learned representations through linear probing, a crucial method for evaluating the quality of representations learned by the model (Chen et al., 2020a; Kumar et al., 2022). Linear probing involves freezing the parameters of a pre-trained model and training a linear classifier on the output. The intuition behind linear probing is that good representations should be linearly separable between classes (Chen et al., 2020a). We apply the MeTSK model pre-trained on the source and target domains to directly generate features for the unseen fMRI data without any fine-tuning. We then input these features into a linear classifier to perform neurological disorder classification.

4 Datasets

We use HCP (Van Essen et al., 2013) as our source dataset due to its large size. For target datasets, we use the ADHD (Bellec et al., 2017) datasets, and the ABIDE dataset (Craddock et al., 2013) during the training of the MeTSK model. We then introduce four independent clinical datasets as held-out domains for evaluating generalization performance.

4.1 HCP Dataset

The healthy control data (source domain) is drawn from the Human Connectome Project (HCP) S1200 dataset (Van Essen et al., 2013). The HCP database includes 1,096 young adult (ages 22-35) subjects with resting-state-fMRI data collected at a total of 1200 time-points for each of four sessions. The preprocessing of fMRI follows the minimal preprocessing procedure in (Gadgil et al., 2020; Glasser et al., 2013). Finally, the brain was parcellated into 116 Regions of Interest (ROIs) using the Automated Anatomical Labeling (AAL) atlas in (Tzourio-Mazoyer et al., 2002). The AAL atlas was defined based on brain anatomy. It divides the brain into 116 regions, including 90 cerebrum regions and 26 cerebellum regions. These 116 regions form the nodes of our graph. The fMRI data were reduced to a single time-series per node by averaging across each ROI.

4.2 ADHD-Peking Dataset

The Attention-Deficit/Hyperactivity Disorder (ADHD-200) consortium data from the Peking site (Bellec et al., 2017) includes 245 subjects in total, with 102 ADHD subjects and 143 Typically Developed Controls (TDC). To model the situation where the clinical target data set is small, we use only the subset of the larger ADHD database that was collected from the Peking site. In the ablation studies, we explore the impact of using the complete set. We use the preprocessed data released on (`http://preprocessed-connectomes-project.org/adhd200/`). During preprocessing, the initial steps involve discarding the first four time points, followed by slice time and motion correction. The data is then registered to the Montreal Neurological Institute (MNI) space, processed with a band-pass filter (0.009Hz - 0.08Hz), and smoothed using a 6 mm Full Width at Half Maximum (FWHM) Gaussian filter. The fMRI data consisted of 231 time points after preprocessing. As a final step, the ADHD-Peking data were re-registered from MNI space to the same AAL atlas as for the HCP subjects, and the average time-series computed for each ROI.

4.3 ABIDE-UM Dataset

The Autism Brain Imaging Data Exchange I (ABIDE I) (Craddock et al., 2013) collects resting-state fMRI from 17 international sites. Similar to the ADHD dataset, we use only the subset of data from the UM site, which includes 66 subjects with Autism Spectrum Disorder (ASD) and 74 TDCs (113 males and 27 females aged between 8-29). We downloaded the data from `http://preprocessed-connectomes-project.org/abide/`, where data was pre-processed using the C-PAC pre-processing pipeline (Craddock et al., 2013). The fMRI data underwent several preprocessing steps: slice time correction, motion correction, and voxel intensity normalization. The data was then band-pass filtered (0.01–0.1 Hz) and spatially registered to the MNI152 template space using a nonlinear method. All fMRIs have 296 time points. As a final step, the ABIDE-UM data were re-registered from MNI space to the same AAL atlas as for the HCP subjects, and the average time-series computed for each ROI.

4.4 Post-traumatic Epilepsy Dataset

We use the Maryland Traumatic Brain Injury (TBI) MagNeTs dataset (Gullapalli, 2011) for generalization performance evaluation. All subjects suffered a traumatic brain injury. Of these we used acute-phase (within 10 days of injury) resting-state fMRI from 36 subjects who went on to develop PTE and 36 who did not (Gullapalli, 2011; Zhou et al., 2012). The dataset was collected as a part of a prospective study that includes longitudinal imaging and behavioral data from TBI patients with Glasgow Coma Scores (GCS) in the range of 3-15 (mild to severe TBI). The individual or group-wise GCS, injury mechanisms, and clinical information is not shared. The fMRI data are available to download from FITBIR (`https://fitbir.nih.gov`). In this study, we used fMRI data acquired within 10 days after injury, and seizure information was recorded using follow-up appointment questionnaires. Exclusion criteria included a history of white matter disease or neurodegenerative disorders, including multiple sclerosis, Huntington's disease, Alzheimer's disease, Pick's disease, and a history of stroke or brain tumors. The imaging was performed on a 3T Siemens TIM Trio scanner (Siemens Medical Solutions, Erlangen, Germany) using a 12-channel receiver-only head coil. The age range for the epilepsy group was 19-65 years (yrs) and 18-70 yrs for the non-epilepsy group.

Pre-processing of the MagNeTs rs-fMRI data was performed using the BrainSuite fMRI Pipeline (BFP) (`https://brainsuite.org`). BFP is a software workflow that processes fMRI and T1-weighted MR data using a combination of software that includes BrainSuite, AFNI, FSL, and MATLAB scripts to produce processed fMRI data represented in a common grayordinate system that contains both cortical surface vertices and subcortical volume voxels (Glasser et al., 2013). As described above, the pre-processed data were then mapped to the same AAL atlas as used with the other datasets. Regional time-series were then generated for each of the 116 parcels by averaging over the corresponding region of interest.

4.5 Alzheimer's Disease Dataset

Open Access Series of Imaging Studies (OASIS-3) (LaMontagne et al., 2019) is a longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer Disease (AD), including 1379 participants: 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42-95 years old. Resting-state fMRI data was used for the classification of Alzheimer's Disease and is publicly available at `https://www.oasis-brains.org`. We randomly selected a subset of the OASIS-3 dataset, consisting of 42 subjects diagnosed as AD and 42 cognitively normal subjects for the downstream clinical task. We are using a small subset to simulate the case where only limited disease data is available. There are 164 time points in each fMRI scan. The preprocessing steps used for OASIS-3 dataset are the same as used for the PTE dataset. We perform binary classification between PD and control subjects.

4.6 Parkinson's Disease Datasets

TaoWu Dataset: The Parkinson's Disease (PD) dataset collected by the group of Tao Wu (Badea et al., 2017) includes both T1-weighted and resting-state fMRI scans from 20 patients diagnosed with PD and 20 age-matched normal controls. All fMRI scans have 239 time points and were collected from a Siemens Magnetom Trio 3T scanner.

Table 1: A comparison of mean AUCs of 5-fold cross-validation on ADHD dataset and ABIDE dataset using different methods: fine-tuning, multi-task learning, the proposed strategy MeTSK, and other baseline methods.

Method	Source & Target	Target-only	ADHD-Peking	ABIDE-UM
SVM	✗	✓	0.6182 ± 0.0351	0.6286 ± 0.0635
RF	✗	✓	0.6117 ± 0.0503	0.6266 ± 0.0612
MLP	✗	✓	0.6203 ± 0.0468	0.6312 ± 0.0724
LSTM (Gadgil et al., 2020)	✗	✓	0.5913 ± 0.0510	0.5936 ± 0.0622
STAGIN (Kim et al., 2021)	✗	✓	0.5638 ± 0.0468	0.5812 ± 0.0684
ST-GCN (Baseline)	✗	✓	0.6215 ± 0.0435	0.6051 ± 0.0615
FT	✓	✗	0.6213 ± 0.0483	0.6368 ± 0.0454
MTL	✓	✗	0.6518 ± 0.0428	0.6345 ± 0.0663
MeTSK (ours)	✓	✗	$\mathbf{0.6981 \pm 0.0409}$	$\mathbf{0.6967 \pm 0.0568}$

Neurocon Dataset: The Neurocon dataset is provided by the Neurology Department of the University Emergency Hospital Bucharest (Romania) Badea et al. (2017), which includes 27 PD patients and 16 normal controls (with 2 replicate scans per subject). Both the rs-fMRI and T1-weighted scans were collected from a 1.5-Tesla Siemens Avanto MRI scanner. All fMRI scans have 137 time points. Both the Neurocon and TaoWU datasets were downloaded from `https://fcon_1000.projects.nitrc.org/indi/retro/parkinsons.html`. We pre-processed TaoWu and Neurocon fMRI data using the same preprocessing pipeline (BFP) that was applied to the PTE dataset.

5 Experiments and Results

5.1 Evaluation of Representation Transferability

To validate the effectiveness of MeTSK when training with a source and a target domain, we first evaluate the knowledge transfer from the HCP data (healthy controls) to ADHD-Peking data and ABIDE-UM data (clinical data). We designed experiments for tasks that perform ADHD v.s. TDC classification and ASD v.s TDC classification, respectively. We evaluate different strategies and compare their effectiveness in enhancing the knowledge transfer from a healthy dataset (source) to a clinical dataset (target).

For comparison, we designed (i) a baseline model using a ST-GCN with a supervised task directly trained on the target dataset (Baseline), (ii) a ST-GCN model fine-tuned on the target dataset after pre-training on HCP data (FT), (iii) a model performing multi-task learning on both source and target datasets simultaneously (MTL), and (iv) the proposed strategy, MeTSK. We incorporated MTL and FT methods for comparison in order to investigate whether MeTSK is superior to traditional approaches in terms of knowledge transfer. We compared several baseline methods: a Linear Support Vector Machine (SVM), a Random Forest Classifier (RF), a Multi-Layer Perceptron (MLP) consisting of three linear layers, an LSTM model for fMRI analysis (Gadgil et al., 2020), and a model combining a transformer and graph neural network (STAGIN) (Kim et al., 2021). For the SVM, RF, and MLP, the inputs are flattened functional connectivity features, calculated using the Pearson's correlation coefficient between fMRI time-series across pairs of brain regions defined in the AAL atlas. LSTM and STAGIN, on the other hand, utilize raw fMRI time-series as their input.

We use 5-fold cross-validation to split training/testing sets on ADHD-Peking/ABIDE-UM data and use all HCP data for training. Model performance is evaluated using the average area-under-the-ROC-curve (AUC) as shown in Table 1. MeTSK achieved the best mean AUC of 0.6981 and 0.6967 for both target datasets, which is a significant improvement compared to the baseline model trained only on target data. MeTSK also surpassed the performance of fine-tuning and multi-task learning. The results demonstrate that the MeTSK strategy possesses the capability to enhance the knowledge transfer from healthy data to clinical data.

5.2 Evaluation of Representation Generalizability

Now we have a MeTSK model trained on HCP and ADHD-Peking data, we evaluate its generalization to four clinical datasets characterized by their small sample sizes and the inherent challenges they present in the identification of neurological disorders. The generalization performance was evaluated on challenging neurological disorder classification tasks. Specifically, the PTE dataset was used for classifying PTE subjects and non-PTE subjects; the OASIS-3 dataset was used for binary classification distinguishing Alzheimer's Disease (AD) from cognitively normal individuals; for the TaoWu and Neurocon datasets, we conducted binary classification to differentiate between Parkinson's Disease (PD) patients and healthy control subjects.

5.2.1 fMRI-based Foundation Models for Generalization Comparison

We compare our representation learning method with foundation models to assess how well our approach generalizes to unseen clinical data relative to these powerful, large-scale models. Foundation models are designed to capture broad, generalizable representations across a wide variety of downstream tasks. By contrasting MeTSK's learned representations with those from foundation models, we can evaluate the effectiveness of our method in generating robust and generalizable features that perform well even with limited clinical data.

We compared our MeTSK model to a large pre-trained fMRI model, as detailed in (Thomas et al., 2022). This model involves pre-training a Generative Pretrained Transformer (GPT) (Radford et al., 2019) on extensive datasets comprising 11,980 fMRI runs from 1,726 individuals across 34 datasets. During pre-training, the GPT model performs a self-supervised task to predict the next masked time point in the fMRI time-series. Their pre-trained model is publicly available at `https://github.com/athms/learning-from-brains`. We directly applied their pre-trained model to generate features. BrainLM (Ortega Caro et al., 2023), one of the latest foundation models developed for fMRI, was also incorporated into our experiments for comparison. BrainLM, consisting of a masked auto-encoder and a vision transformer, was trained on 6,700 hours of fMRI recordings. During pre-training, BrainLM incorporates a self-supervised task that predicts the randomly masked segments of time series in fMRI data, which is similar to the pre-training task in (Thomas et al., 2022). Similarly, we directly applied their pre-trained model available at `https://github.com/vandijklab/BrainLM` to generate features. We also pre-trained a ST-GCN model on both HCP and ADHD-Peking datasets using only the proposed contrastive self-supervised learning (SSL). From this pre-trained SSL model, we again directly generated features for unseen clinical data.

5.2.2 Evaluation Results

We compare the linear probing performance across foundation model approaches as well as the performance of classifiers trained with functional connectivity features extracted from raw fMRI data. We employed the same machine learning classifiers as used in the previous experiments, including a linear SVM, RF, and MLP. The same 5-fold cross-validation was applied and AUCs for classification tasks were computed. Although complex classifiers could be used, the use of SVM, RF and MLP isolates the quality of the learned representations from the model's complexity.

Our proposed MeTSK achieved the best performance on all four datasets, outperforming the two latest fMRI foundation models, as shown in Table 2. The features generated by MeTSK, the SSL model and BrainLM (Ortega Caro et al., 2023) all achieved better performance than functional connectivity features, owing to the knowledge learned from their extensive pre-training datasets. However, due to the domain gap between their pre-training data and the clinical datasets used here, the foundation models may require additional fine-tuning to further improve their performance. Notably, MeTSK achieved statistically significant improvements in every dataset according to the p-values calculated using a paired t-test. These results highlight MeTSK's potential in enhancing representation learning for clinical diagnostic purposes.

5.2.3 Interpretation of Learned Representations

To gain further insights and enhance the interpretability of the representations learned by MeTSK, we computed a feature importance map based on the positive coefficients from the SVM trained for PTE

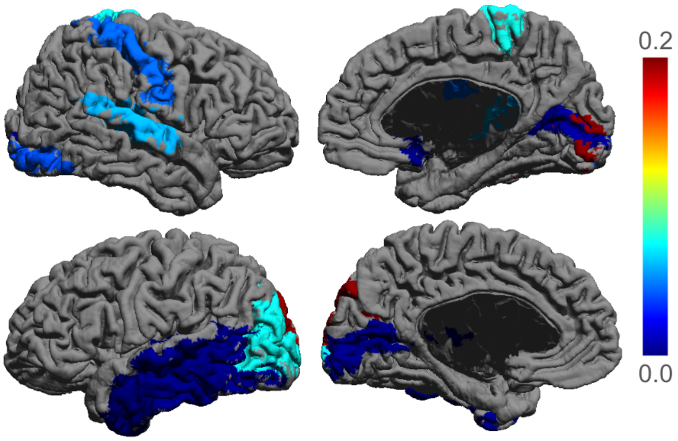

Figure 3: Feature importance map of PTE features generated from MeTSK shown as color-coded ROIs overlaid on the AAL atlas. The numbers represent the absolute value of coefficients from the trained SVM.

Table 2: Generalization evaluation results using 5-fold cross-validation: Mean and std of AUCs for PTE, AD, and PD classification using representations generated from different models (denoted as "linear probing" in the Table) as well as directly from functional connectivity features. The best performance achieved on each dataset is highlighted in bold. Asterisks (*) indicate that the difference between our model MeTSK and the comparing methods is statistically significant using a paired Student's t-test. Significance levels are denoted by $^*(p \leq 0.05)$, $^{**}(p \leq 0.01)$, $^{***}(p \leq 0.001)$.

		Linear Probing				Connectivity Features
		MeTSK	SSL	Thomas et al. (2022)	Ortega Caro et al. (2023)	
PTE	SVM	$\mathbf{0.6415 \pm 0.0312}$	$0.5972 \pm 0.0492^*$	$0.5369 \pm 0.0451^{**}$	0.6011 ± 0.0465	$0.5697 \pm 0.0477^*$
	RF	0.5392 ± 0.0553	0.5253 ± 0.0486	$0.4814 \pm 0.0664^*$	0.5589 ± 0.0611	0.5081 ± 0.0612
	MLP	0.5813 ± 0.0504	$0.5216 \pm 0.0329^*$	0.5278 ± 0.0643	0.5290 ± 0.0674	$0.5111 \pm 0.0402^*$
OASIS-3	SVM	0.6407 ± 0.0741	$0.5992 \pm 0.0867^*$	0.6115 ± 0.0582	0.6028 ± 0.0831	$0.5541 \pm 0.0677^{**}$
	RF	$\mathbf{0.6750 \pm 0.0753}$	$0.6055 \pm 0.0682^*$	0.6233 ± 0.0672	$0.5604 \pm 0.0607^{**}$	$0.5329 \pm 0.0721^{**}$
	MLP	0.6034 ± 0.0725	$0.5113 \pm 0.0591^{**}$	$0.5593 \pm 0.0576^*$	$0.5523 \pm 0.0711^{**}$	$0.5231 \pm 0.0689^{**}$
TaoWu	SVM	$\mathbf{0.6831 \pm 0.1431}$	0.6371 ± 0.1578	$0.5528 \pm 0.1586^{***}$	$0.4937 \pm 0.1506^{***}$	$0.5725 \pm 0.1502^{**}$
	RF	0.6553 ± 0.1701	0.6273 ± 0.1679	$0.4843 \pm 0.1885^{***}$	$0.4891 \pm 0.1710^{***}$	0.6031 ± 0.1631
	MLP	0.6208 ± 0.1582	$0.5013 \pm 0.1621^{**}$	$0.5325 \pm 0.1672^*$	$0.5078 \pm 0.1592^{**}$	0.5875 ± 0.1631
Neurocon	SVM	$\mathbf{0.7529 \pm 0.1579}$	$0.5643 \pm 0.1068^{***}$	$0.6433 \pm 0.1535^{***}$	$0.6476 \pm 0.1686^{**}$	$0.6599 \pm 0.1807^*$
	RF	0.6230 ± 0.1658	$0.5219 \pm 0.1627^{***}$	0.5813 ± 0.1818	0.6774 ± 0.1676	$0.5427 \pm 0.1715^*$
	MLP	0.6100 ± 0.1635	$0.5073 \pm 0.1516^{**}$	$0.5401 \pm 0.1802^*$	0.5744 ± 0.1629	$0.5313 \pm 0.1721^*$

prediction. In a linear SVM, each feature within each ROI (brain region) is assigned a coefficient, indicating its importance in the model's decision-making process. The higher the absolute value of a coefficient, the greater its impact on the model's predictions. The coefficient for each ROI is calculated as the average of the coefficients for all features within that ROI. We extracted these coefficients for each ROI from the trained SVM and visualized them as a feature importance map overlaid on the brain, as shown in Fig. 3. Through observing the feature importance map, we identified that the most significant regions for PTE prediction are located in the temporal, parietal, and occipital lobes. Since epilepsy most commonly occurs in the temporal lobes, it is not surprising to see that they are among the regions identified from the feature important maps. A recent study on PTE (Akrami et al., 2024) also reported statistically significant differences between PTE and non-PTE groups in the parietal and occipital lobes. This alignment between our interpretation and clinical evidence demonstrate that the representations learned by MeTSK are not only predictive but also clinically meaningful.

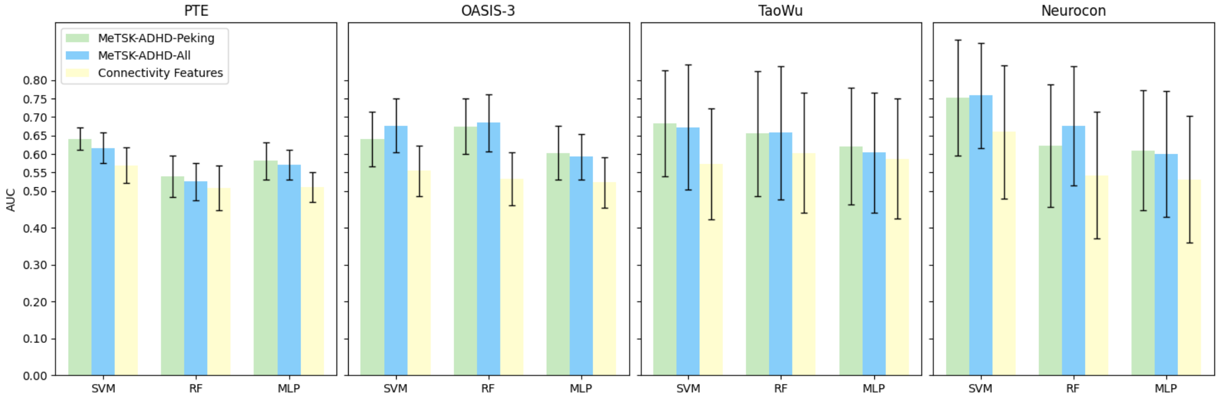

Figure 4: A comparison of the linear probing performance between MeTSK model trained with a single target site (MeTSK-ADHD-Peking, results from Table 2) and MeTSK model trained using all the target sites from ADHD-200 dataset (MeTSK-ADHD-All). Also shown are results from linear classifiers trained directly using connectivity features for baseline comparison. The height of each bar indicates the average AUC computed from a 5-fold cross-validation, while the error bars denote the standard deviations. MeTSK-ADHD-All achieved similar performance to MeTSK-ADHD-Peking on all the four downstream datasets. The p-values computed using a paired Student's t-test indicate that there is no significant difference in performance.

6 Ablation Study

6.1 Training with More Clinical Data

In this paper we have explored a strategy MeTSK which trains a model to generalize from abundant normal features to scarce clinical features. To model this situation, we employed only a single site (Peking) from the ADHD-200 dataset to train our model. MeTSK emphasizes the model's ability to extrapolate from abundant healthy control features to scarce clinical features, equipping it with generalization capabilities for unseen real-world applications where clinical data may also be limited.

However, while MeTSK is intentionally tailored for training with small-scale clinical data, we also explored the impact of using more clinical target data to train the model. We used the data from all of the sites in the ADHD-200 datasets for training. The entire dataset comprised a total of 362 ADHD subjects and 585 TDCs, all pre-processed using the same steps as described in Section 4.1.2. The same linear probing was performed to evaluate the models on the four clinical datasets: PTE, OASIS-3, TaoWu, and Neurocon. As shown in Fig 4, MeTSK consistently demonstrated similar performance for various downstream tasks when trained using the larger clinical dataset. We computed the p-value using a paired Student's t-test, and showed that there is no significant difference between the downstream performance of the model trained with Peking site only (MeTSK-ADHD-Peking) and the model trained with the entire ADHD-200 dataset (MeTSK-ADHD-All). This underscores the model's robustness and its capacity to achieve effective generalization without relying on a large amount of training data.

6.2 Ablation Study of MeTSK

We examine the individual contributions of self-supervised learning and meta-learning to the model performance on both target clinical datasets (ADHD-Peking, ABIDE-UM) in this section. To explore the effect of meta-learning, we designed an experiment using only the target (clinical) dataset in meta-learning (MeL). This approach involves removing the source head and the source loss during bi-level optimization. The target head is first trained on the ADHD/ABIDE meta-training set in the inner loop, followed by feature extractor learning to generalize on a held-out validation set in the outer loop. Our results, as shown in the last two rows of Table 3, reveal that the mean AUC improved from the baseline performance of 0.6215 to 0.6562 for ADHD classification, and from 0.6051 to 0.6675 for ASD classification without source domain knowledge.

Table 3: Ablation study on ADHD-Peking and ABIDE-UM dataset. The FT, MTL, and MeTSK methods are compared for two cases - transferring features from (i) a self-supervised source task and (ii) a sex classification source task, respectively. The last two rows are models trained only on target clinical data: a meta-learning model without source task and a baseline model.

Dataset	ADHD-Peking		ABIDE-UM	
Source Task	Self-supervision	Sex Classification	Self-supervision	Sex Classification
FT	0.6213 ± 0.0483	0.6150 ± 0.0497	0.6368 ± 0.0454	0.6071 ± 0.0742
MTL	0.6518 ± 0.0428	0.6377 ± 0.0512	0.6345 ± 0.0663	0.6240 ± 0.0711
MeTSK	0.6981 ± 0.0409	0.6732 ± 0.0579	0.6967 ± 0.0568	0.6786 ± 0.0749
MeL	0.6562 ± 0.0489		0.6675 ± 0.0505	
Baseline	0.6215 ± 0.0435		0.6051 ± 0.0615	

To assess the contribution of self-supervised learning, we compared the impact of using a self-supervised task versus a sex classification task on the HCP dataset. Fine-tuning, multi-task learning, and MeTSK were implemented using sex classification (female vs male) as the source task. The same 5-fold cross-validation method was applied to compare the average AUC. As detailed in Table 3, all three methods: FT, MTL, and MeTSK, showed a degraded performance when transferring knowledge from the sex classification task. This suggests that the sex-related features of the brain may be less relevant to ADHD/ASD classification, negatively affecting the model's performance.

7 Discussion

Our proposed method, MeTSK, addresses the critical need for generalizable representations in clinical fMRI applications, particularly in scenarios with data scarcity and heterogeneity. Traditional deep learning approaches often struggle in these scenarios due to the lack of sufficient labeled data and high variability across subjects, as shown in our results (Appendix A.1) where simple machine learning classifiers, like SVM and RF, outperformed deep learning models when trained directly on limited clinical data. The high-dimensional nature and complex spatial-temporal dynamics of fMRI data make it even more challenging to extract diagnostic features given limited training data. These challenges underscore the importance of developing a generalizable representation learning strategy for clinical applications.

By employing meta-learning, we enhance the generalization capabilities of self-supervised features from the source domain (control) to target clinical domains, ensuring that the learned representations capture intrinsic patterns in functional brain activity, such patterns are shared across datasets. The learned representations enable effective classification even with simple linear classifiers. This demonstrates the quality of the learned features and their generalizability across unseen clinical datasets.

References

David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. Graph-based deep learning for medical diagnosis and analysis: past, present and future. *Sensors*, 21(14): 4758, 2021.

Haleh Akrami, Andrei Irimia, Wenhui Cui, Anand A Joshi, and Richard M Leahy. Prediction of posttraumatic epilepsy using machine learning. In *Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11600, pp. 424–430. SPIE, 2021.

Haleh Akrami, Wenhui Cui, Paul E Kim, Christianne N Heck, Andrei Irimia, Karim Jebri, Dileep Nair, Richard M Leahy, and Anand Joshi. Prediction of post traumatic epilepsy using mri-based imaging markers. *bioRxiv*, pp. 2024–01, 2024.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, pp. 1–24, 2023.

Liviu Badea, Mihaela Onu, Tao Wu, Adina Roceanu, and Ovidiu Bajenaru. Exploring the reproducibility of functional connectivity alterations in parkinson's disease. *PLoS One*, 12(11):e0188196, 2017.

Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27):5, 2013.

Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Soham Gadgil, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Ehsan Adeli, and Kilian M Pohl. Spatio-temporal graph convolution for resting-state fmri analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 528–538. Springer, 2020.

Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Rao P Gullapalli. Investigation of prognostic ability of novel imaging markers for traumatic brain injury (tbi). Technical report, BALTIMORE UNIV MD, 2011.

Zhi-An Huang, Rui Liu, and Kay Chen Tan. Multi-task learning for efficient diagnosis of asd and adhd using resting-state fmri data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Nikhil Ketkar. Stochastic gradient descent. In *Deep learning with Python*, pp. 113–132. Springer, 2017.

Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021.

Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization (2014). *arXiv preprint arXiv:1412.6980*, 22, 2014.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, pp. 2019–12, 2019.

Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

Hong Liu, Jeff Z HaoChen, Colin Wei, and Tengyu Ma. Meta-learning transferable representations with a single target domain. *arXiv preprint arXiv:2011.01418*, 2020a.

Hong Liu, Jeff Z. HaoChen, Colin Wei, and Tengyu Ma. Meta-learning transferable representations with a single target domain, 2020b.

Xingdan Liu, Jiacheng Wu, Wenqi Li, Qian Liu, Lixia Tian, and Huifang Huang. Domain adaptation via low rank and class discriminative representation for autism spectrum disorder identification: A multi-site fmri study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:806–817, 2023.

Jaehoon Oh, Sungnyun Kim, Namgyu Ho, Jin-Hwa Kim, Hwanjun Song, and Se-Young Yun. Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty, 2022.

Josue Ortega Caro, Antonio Henrique Oliveira Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pp. 2023–09, 2023.

Elena N Pitsik, Vladimir A Maximenko, Semen A Kurkin, Alexander P Sergeev, Drozdstoy Stoyanov, Rositsa Paunova, Sevdalina Kandilarova, Denitsa Simeonova, and Alexander E Hramov. The topology of fmri-based networks defines the performance of a graph neural network for the classification of patients with major depressive disorder. *Chaos, Solitons & Fractals*, 167:113041, 2023.

Svetlozar T Rachev. The monge–kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging, 2019.

Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2584–2594, 2022.

Chunlei Shi, Xianwei Xin, and Jiacai Zhang. Domain adaptation using a three-way decision improves the identification of autism patients from multisite fmri data. *Brain Sciences*, 11(5):603, 2021.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems*, 33:18158–18172, 2020.

Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in Neural Information Processing Systems*, 35:21255–21269, 2022.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1): 273–289, 2002.

Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Xuesong Wang, Lina Yao, Islem Rekik, and Yu Zhang. Contrastive functional connectivity graph learning for population-based fmri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pp. 221–230. Springer, 2022.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020a.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. When does self-supervision help graph convolutional networks? In *international conference on machine learning*, pp. 10871–10880. PMLR, 2020b.

Z. Yu and G. Herman. On the earth mover's distance as a histogram similarity metric for image retrieval. In *2005 IEEE International Conference on Multimedia and Expo*, pp. 4 pp.–, 2005. doi: 10.1109/ICME.2005.1521516.

Hao Zhang, Ran Song, Liping Wang, Lin Zhang, Dawei Wang, Cong Wang, and Wei Zhang. Classification of brain disorders in rs-fmri via local-to-global graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2):444–455, 2023. doi: 10.1109/TMI.2022.3219260.

Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2487–2495, 2019.

Yongxia Zhou, Michael P Milham, Yvonne W Lui, Laura Miles, Joseph Reaume, Daniel K Sodickson, Robert I Grossman, and Yulin Ge. Default-mode network disruption in mild traumatic brain injury. *Radiology*, 265 (3):882, 2012.

Qi-Hong Zou, Chao-Zhe Zhu, Yihong Yang, Xi-Nian Zuo, Xiang-Yu Long, Qing-Jiu Cao, Yu-Feng Wang, and Yu-Feng Zang. An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff. *Journal of neuroscience methods*, 172(1):137–141, 2008.

A Appendix

A.1 Applying Deep Learning Models to Small-Scale Clinical Datasets

We present the classification results on the clinical datasets mentioned in Section 5.2 using some popular deep learning models. These models demonstrated severe over-fitting and poor generalization, achieving lower

Table 4: Mean and std of AUCs of classification results using 5-fold cross-validation

Methods	PTE	OASIS-3	TaoWu	Neurocon
SVM	0.5697 ± 0.0477	0.5541 ± 0.0677	0.5725 ± 0.1502	0.6599 ± 0.1807
RF	0.5081 ± 0.0612	0.5329 ± 0.0721	0.6031 ± 0.1631	0.5427 ± 0.1715
MLP	0.5111 ± 0.0402	0.5231 ± 0.0689	0.5875 ± 0.1631	0.5313 ± 0.1721
ST-GCN (Gadgil et al., 2020)	0.5108 ± 0.0718	0.5079 ± 0.0833	0.5377 ± 0.1739	0.5089 ± 0.1890
LSTM (Gadgil et al., 2020)	0.5031 ± 0.0712	0.4675 ± 0.0880	0.5019 ± 0.1721	0.4914 ± 0.1799
STAGIN (Kim et al., 2021)	0.4517 ± 0.0806	0.4891 ± 0.0904	0.4332 ± 0.1859	0.4470 ± 0.1822

Table 5: Summary of additional target datasets, which are different imaging sites from the same cohort ADHD/ABIDE.

Dataset	Total Subjects	Healthy Controls	Condition Subjects	Time Length
ADHD-NYU	216	98	118 ADHD	172
ADHD-NI	48	23	25 ADHD	231
ABIDE-NYU	175	100	75 ASD	176
ABIDE-Leuven	63	34	29 ASD	246

Table 6: Performance on additional sites from ADHD and ABIDE dataset (complimentary for Section 5.1).

Dataset	ADHD		ABIDE	
Site	NYU	NI	NYU	Leuven
SVM	0.6202 ± 0.0662	0.6887 ± 0.0476	0.6944 ± 0.0755	0.6672 ± 0.0883
RF	0.5605 ± 0.0586	0.6631 ± 0.0591	0.6785 ± 0.0739	0.6228 ± 0.0579
ST-GCN	0.5722 ± 0.0694	0.6013 ± 0.0781	0.6889 ± 0.0621	0.6786 ± 0.0734
MeTSK	$\mathbf{0.6704 \pm 0.0782}$	$\mathbf{0.8020 \pm 0.0914}$	$\mathbf{0.7268 \pm 0.0687}$	$\mathbf{0.7116 \pm 0.0502}$

AUCs than traditional machine learning classifiers. The deep learning models struggled to learn from the limited and highly variable clinical data, where the high-dimensional and complex nature of fMRI data and small sample sizes amplified the over-fitting issues. These results emphasized our motivation for developing a representation learning approach capable of generating generalizable features for such challenging datasets.

A.2 Experiment Results Using Additional Target Datasets

We present additional results by training the MeTSK model with different target datasets and subsequently evaluating its generalization performance on the same clinical tasks described in Section 5.2. Table 6 reports the knowledge transfer performance when evaluated on additional target datasets (other sites from the ADHD and ABIDE datasets). Table 7 shows the generalization performance on four unseen clinical datasets using MeTSK models pre-trained on different target datasets. The results consistently demonstrate performance improvements compared to using functional connectivity features, highlighting the robustness of the proposed representation learning strategy, which further validates the effectiveness of MeTSK in extracting transferable and generalizable representations across different clinical applications.

A.3 Implementation Details

Training: To optimize model performance, we follow the training setting in (Gadgil et al., 2020) for the ST-GCN model. The length of input sub-sequences for ST-GCN is fixed at 128. We generate one meta-training batch by randomly selecting an equal number of samples from each class. The batch size is 32, both for the meta-training and the meta-validation set. We use an Adam optimizer (Kingma & Ba, 2014) with learning rate $\beta = 0.001$ in the outer loop, and an SGD optimizer (Ketkar, 2017) with learning rate $\alpha = 0.01$ in the inner loop. The number of inner loop update steps is 25. We set the hyper-parameter $\lambda = 30$ and the

Table 7: A comparison of MeTSK models pre-trained with different datasets as the target domain. Linear probing was performed for evaluation of models on the four clinical datasets.

		ADHD			ABIDE			Connectivity Features
		Peking (proposed)	NYU	NI	UM	NYU	Leuven	
PTE	SVM	**0.6415 ± 0.0312**	0.5779 ± 0.0469	0.5589 ± 0.0392	0.5753 ± 0.0459	0.5661 ± 0.0483	0.5928 ± 0.0475	0.5697 ± 0.0477
	RF	0.5392 ± 0.0553	0.5395 ± 0.0492	0.5199 ± 0.0508	0.5156 ± 0.0517	0.5836 ± 0.0558	0.5783 ± 0.0489	0.5081 ± 0.0612
	MLP	0.5813 ± 0.0504	0.5283 ± 0.0397	0.5449 ± 0.0517	0.5038 ± 0.0478	0.5469 ± 0.0527	0.5592 ± 0.0531	0.5111 ± 0.0402
OASIS-3	SVM	0.6407 ± 0.0741	0.6050 ± 0.0812	0.6255 ± 0.0806	0.6340 ± 0.0889	0.6567 ± 0.1133	0.6503 ± 0.1097	0.5541 ± 0.0677
	RF	**0.6750 ± 0.0753**	0.5812 ± 0.0711	0.6718 ± 0.0745	0.6233 ± 0.0769	0.6066 ± 0.1019	0.5906 ± 0.1083	0.5329 ± 0.0721
	MLP	0.6034 ± 0.0725	0.5325 ± 0.0791	0.5563 ± 0.0762	0.5998 ± 0.0824	0.5277 ± 0.0819	0.5166 ± 0.0729	0.5231 ± 0.0689
TaoWu	SVM	**0.6831 ± 0.1431**	0.5597 ± 0.1419	0.6807 ± 0.1597	0.6281 ± 0.1766	0.6073 ± 0.1750	0.6219 ± 0.1476	0.5725 ± 0.1502
	RF	0.6553 ± 0.1701	0.5375 ± 0.1645	0.5718 ± 0.1821	0.6328 ± 0.1642	0.5525 ± 0.1725	0.5810 ± 0.1651	0.6031 ± 0.1631
	MLP	0.6208 ± 0.1582	0.5625 ± 0.1578	0.5825 ± 0.1577	0.5750 ± 0.1521	0.5925 ± 0.1593	0.6050 ± 0.1422	0.5875 ± 0.1631
Neurocon	SVM	**0.7529 ± 0.1579**	0.6794 ± 0.1772	0.6721 ± 0.1842	0.6874 ± 0.1717	0.6943 ± 0.1635	0.6760 ± 0.1859	0.6599 ± 0.1807
	RF	0.6230 ± 0.1658	0.5883 ± 0.1807	0.5756 ± 0.1719	0.5850 ± 0.1826	0.6084 ± 0.1879	0.5893 ± 0.1551	0.5427 ± 0.1715
	MLP	0.6100 ± 0.1635	0.5202 ± 0.1473	0.5478 ± 0.1683	0.5755 ± 0.1709	0.5497 ± 0.1537	0.5344 ± 0.1681	0.5313 ± 0.1721

temperature parameter $\tau = 30$ to adjust the scale of losses following (Liu et al., 2020a; You et al., 2020a). Since contrastive loss converges slowly (Jaiswal et al., 2020), a warm-up phase is applied to train the model only on HCP data using the graph contrastive loss for the first half of total training steps.

For the multi-task learning (MTL) implementation, we simply remove the inner loop in MeTSK and use all the training data to update the target head. Both heads and the feature extractor are updated simultaneously in one loop. For meta-learning, the target training set in each fold is further divided into a meta-training set $X_{\mathcal{T}_{tr}}$ of 157 subjects and a meta-validation set $X_{\mathcal{T}_{val}}$ of 39 subjects.

Evaluation: For SSL and MeTSK, we use the pre-trained feature extractor for generating features. The generated features are graph-level representations, having a two dimensional feature matrix at each node (brain region). We averaged the features along the first dimension and applied Pinciple Component Analysis (PCA) to reduce the dimensionality before feeding the features into classifiers. The MLP used in the downstream experiments here consists of three linear layers, with hidden dimensions of 32, 16, 16. The SSL model trained on both HCP and ADHD-Peking data used the same contrastive loss. In our comparative analysis with a foundation model for fMRI (Thomas et al., 2022), we flatten the brain signals at each time-point and input the whole time-series without masking into the pre-trained GPT model. This generates a feature embedding for each time-point. Since there is no class token in the pre-trained model, we averaged the feature embeddings across time-points (tokens) to get the output features for downstream datasets. We follow the other detailed settings of the pre-trained GPT model in (Thomas et al., 2022). For another foundation model BrainLM (Ortega Caro et al., 2023), we followed the instructions provided on `https://github.com/vandijklab/BrainLM/tree/main`. Specifically, we directly input the fMRI time-series into the pre-trained model without masking. Then we extracted the output class token as the generated features for clinical datasets. We ran 100 iterations of stratified cross-validation on all clinical datasets for each method.

A.4 Domain Similarity

To evaluate the transferability of learned features, we measure the distance between features extracted from different domains using Domain Similarity (Cui et al., 2018; Oh et al., 2022). We first compute the Earth Mover's Distance (EMD) (Yu & Herman, 2005), which is based on the solution to the Monge-Kantorovich problem (Rachev, 1985), to measure the cost of transferring features from the source to target domain. We define $\bar{X}_{\mathcal{S}} = \text{Flatten}(\frac{1}{N}\sum_{n=1}^{N}\tilde{X}_{\mathcal{S}})$, $\bar{X}_{\mathcal{T}} = \text{Flatten}(\frac{1}{N}\sum_{n=1}^{N}\tilde{X}_{\mathcal{T}})$ as the flattened vectors of the output graph features averaged over all subjects, and then define B_s and B_t as the set of bins in the histograms representing feature distribution in $\bar{X}_{\mathcal{S}}$ and $\bar{X}_{\mathcal{T}}$, respectively. A larger Domain Similarity indicates better transferability from the source domain to the target domain because the amount of work needed to transform source features into target features is smaller. Domain similarity (DS) is defined as:

$$\text{DS} = \exp\left(-\gamma\, \text{EMD}(\bar{X}_{\mathcal{S}}, \bar{X}_{\mathcal{T}})\right) \tag{7}$$

$$
\begin{aligned}
\text{EMD}(\bar{X}_{\mathcal{S}}, \bar{X}_{\mathcal{T}}) = {} & \frac{\sum_{i=1}^{|B_s|}\sum_{j=1}^{|B_t|} f_{i,j} d_{i,j}}{\sum_{i=1}^{|B_s|}\sum_{j=1}^{|B_t|} f_{i,j}}, \\
s.t. \quad & f_{ij} \geq 0, \\
& \sum_{j=1}^{|B_t|} f_{ij} \leq \frac{|\bar{X}_{\mathcal{S}} \in B_s(i)|}{|\bar{X}_{\mathcal{S}}|}, \\
& \sum_{i=1}^{|B_s|} f_{ij} \leq \frac{|\bar{X}_{\mathcal{T}} \in B_t(j)|}{|\bar{X}_{\mathcal{T}}|}, \\
& \sum_{i=1}^{|B_s|}\sum_{j=1}^{|B_t|} f_{ij} = 1
\end{aligned}
\tag{8}
$$

where $B_s(i)$ is the i-th bin of the histogram and $|B_s|$ is the total number of bins, $|\bar{X}_{\mathcal{S}} \in B_s(i)|$ is the number of features in $B_s(i)$, $|\bar{X}_{\mathcal{S}}|$ is the total number of features, $d_{i,j}$ is the Euclidean distance between the averaged features in $B_s(i)$ and $B_t(j)$, $f_{i,j}$ is the optimal flow for transforming $B_s(i)$ into $B_t(j)$ that minimizes the EMD. Following the setting in (Cui et al., 2018), we set $\gamma = 0.01$.

A.4.1 Experiments Using Domain Similarity

To further investigate the transferability enabled by MeTSK, domain similarity was computed to evaluate the knowledge transfer from control data (source) to clinical data (target) as well as from the training set to the testing set of target data. We conducted domain similarity analysis on both ADHD-Peking and ABIDE-UM datasets to further validate the robustness and versatility of MeTSK. Fig. 5 illustrates that the self-supervised source features have a higher similarity with the target features, indicating better inter-domain transferability and thus improved performance on the target classification task. Moreover, compared to the baseline, both intra-ADHD-class/intra-ASD-class and intra-TDC-class domain similarities between the training and testing sets of ADHD/ABIDE data are increased by MeL. This enhancement provides evidence to explain the improved classification performance on training with only target data achieved by meta-learning. By applying meta-learning, not only the inter-domain generalization of features is boosted, but also the effect of heterogeneous data within the same domain is alleviated.

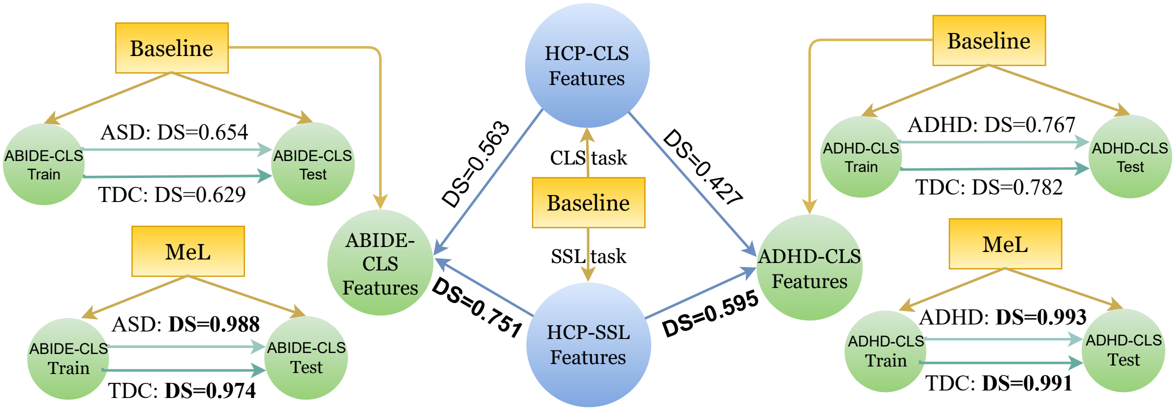

Figure 5: A comparison of the domain similarity between HCP self-supervised features (HCP-SSL, from Baseline ST-GCN trained on HCP data with a self-supervised task) and ADHD/ASD classification features (ADHD-CLS, ABIDE-CLS, from Baseline trained using all ADHD/ABIDE data), the domain similarity between HCP sex classification features (HCP-CLS, from Baseline trained on HCP data with a sex classification task) and ADHD-CLS/ABIDE-CLS, the intra-class (ADHD; TDC and ASD; TDC) domain similarities between training and testing set of ADHD/ASD data from Baseline and MeL (a meta-learning model trained only on target data), respectively.