Contents lists available at ScienceDirect

Medical Image Analysis



Meta Transfer of Self-Supervised Knowledge: Foundation Model in Action for Post-Traumatic Epilepsy Prediction

Wenhui Cui^a, Haleh Akrami^a, Ganning Zhao^a, Anand A. Joshi^a, Richard M. Leahy^{a,*} *a Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles 90089, United States* A R TICLE INFO *Article history:* A B S T R A C T Despite the impressive advancements achieved using deep-learnin, activity analysis, the heterogeneity of functional patterns and scale Despite the impressive advancements achieved using deep-learning for functional brain activity analysis, the heterogeneity of functional patterns and scarcity of imaging data still pose challenges in tasks such as prediction of future onset of Post-Traumatic Epilepsy (PTE) from data acquired shortly after traumatic brain injury (TBI). Foundation models pre-trained on separate large-scale datasets can improve the performance from scarce and heterogeneous datasets. For functional Magnetic Resonance Imaging (fMRI), while data may be abundantly available from healthy controls, clinical data is often scarce, limiting the ability of foundation models to identify clinically-relevant features. We overcome this limitation by introducing a novel training strategy for our foundation model by integrating meta-learning with self-supervised learning to improve the generalization from normal to clinical features. In this way we enable generalization to other downstream clinical tasks, in our case prediction of PTE. To achieve this, we perform self-supervised training on the control dataset to focus on inherent features that are not limited to a particular supervised task while applying meta-learning, which strongly improves the model's generalizability using bi-level optimization. Through experiments on neurological disorder classification tasks, we demonstrate that the proposed strategy significantly improves task performance on small-scale clinical datasets. To explore the generalizability of the foundation model in downstream applications, we then apply the model to an unseen TBI dataset for prediction of PTE using zero-shot learning. Results further demonstrated the enhanced generalizability of our foundation model.

© 2023 Elsevier B. V. All rights reserved.

1. Introduction

Deep learning based approaches have demonstrated success in analyzing brain connectivity based on functional magnetic resonance imaging (fMRI) Gadgil et al. (2020); AhmedtAristizabal et al. (2021), but the scarcity and heterogeneity of fMRI data still pose challenges in clinical applications such as predicting the future onset of Post-Traumatic Epilepsy (PTE) from acute data acquired shortly after traumatic brain injury (TBI) Akbar et al. (2022). Identification of subjects at high risk of developing PTE can eliminate the need to wait for spontaneous epileptic seizures to occur before starting treatment and enable the mitigation of risks to subjects whose seizures could

Preprint submitted to Medical Image Analysis





^{*}Corresponding author at: Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles 90089, United States. Email address: leahy@usc.edu

result in serious injury or death. fMRI plays a vital role in identifying biomarkers for PTE. The presence of lesions in TBI patients can alter resting-state brain dynamics Palacios et al. (2013). This will be reflected in fMRI data collected after injury, which can therefore provide valuable biomarkers for PTE. However, TBI datasets are usually characterized by high variability among subjects and limited numbers of subjects, presenting a significant challenge for training of deep learning methods to predict PTE. To tackle these challenges, we can employ a large pre-trained model. Developing a foundation model pre-trained on large-scale datasets has been exceptionally successful in natural language processing Radford et al. (2019) and computer vision tasks Yu et al. (2023). Typically, foundation models can generalize across domains and tasks, achieving promising performance even in few-shot and zero-shot learning scenarios. Foundation models are usually trained using a selfsupervised task Radford et al. (2019) involving extensive and diverse datasets. In medical data, it is common to have a large amount of healthy control data, while simultaneously facing a scarcity of clinical data collected for any particular neurological disorder. Simply aggregating all normal and clinical data and applying self-supervised learning may cause limited generalization and bias because of 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). Since our goal is to achieve superior performance on downstream clinical tasks, it is crucial to learn how to generalize to useful clinical features during the training of the foundation model.

To address the limited generalization, we adopt metalearning as a novel approach for developing foundation models that leverage features from large-scale normal datasets and small-scale clinical datasets. 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. (2020); Finn et al. (2017) and has shown success in few-shot learning tasks. Meta-learning enhances the model's generalization, even when trained on smaller-scale datasets. One of the most popular meta-learning algorithms, the Model Agnostic Meta-Learning method (MAML) Finn et al. (2017), is a gradientbased approach that uses a bi-level optimization scheme to enable the model to learn how to generalize on an unseen domain during training. However, in the context of fMRI, the availability of diverse datasets is typically limited. Instead of learning to generalize from multiple source tasks to multiple target tasks in MAML, Liu et al. (2020) propose a meta representation learning approach to learn generalizable features from one source domain and improve the generalization to one target domain. To combine data acquired from healthy (control) populations with clinical data from patients, we consider a source domain with abundant control data and a target domain with limited clinical data during upstream training of the foundation model. Through meta-learning Liu et al. (2020), the model is enabled to generalize from control features to clinical features. For downstream applications, we focus on a TBI/PTE dataset characterized by extreme heterogeneity and scarcity of clinical fMRIs, where traditional deep learning models often over-fit and fail to generalize. Our goal is to apply the meta-learning pre-trained model to this PTE dataset during downstream adaptation. By leveraging the learned generalization from normal to clinical features, we aim to enhance the model performance on this challenging clinical dataset.

Self-supervised learning has shown the ability to improve the generalization of features in foundation models Ortega Caro et al. (2023); Thomas et al. (2022); Azizi et al. (2023). 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 intraclass variances Wang et al. (2022). For our foundation model, we apply contrastive self-supervised learning, known to be effective in representation learning Azizi et al. (2023), to the control data (the source domain in the meta-learning framework) to learn more generalizable features.

We propose a novel training strategy for the foundation model: Meta Transfer of Self-supervised Knowledge (MeTSK), which harnesses meta-learning to facilitate the transfer of selfsupervised features from large-scale control to scarce clinical datasets. This training strategy is designed to enhance the foundation model's capacity to generalize from normal features to clinical features, which will also facilitate the generalization to new and unseen clinical features in downstream applications. The proposed network architecture consists of a feature extractor that learns general 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. To further validate the generalization improvement achieved by MeTSK, we adopt domain similarity, representing the least amount of work required to transform features into a different domain, as our generalization metric. A larger domain similarity implies that the features are more transferable. Our experimental results demonstrate the effectiveness of MeTSK on neurological disorder classification tasks by improving both inter-domain and intra-domain generalization. This finding underscores the potential of MeTSK in effectively bridging the gap between normal and clinical datasets.

Beyond the typical approach of fine-tuning the entire foundation model for downstream adaptation, linear probing is a crucial method for evaluating the quality of features learned by the foundation 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 features should be linearly separable between classes Chen et al. (2020a). For our downstream application, we perform linear probing on the PTE prediction task. We apply our foundation model trained using MeTSK to directly generate features for the PTE fMRI data without any fine-tuning. We then input these features to a linear classifier and achieved superior classification performance compared to using functional connectivity features as input. In summary, our contribution is two-fold:

- We propose a novel training strategy for developing a foundation model for fMRI data by learning how to generalize from control to clinical features;
- We address the heterogeneity and scarcity of clinical fMRI data by improving the generalization of the model through the integration of meta-learning and self-supervised learning.

2. Related Work

2.1. Foundation Models for fMRI

Foundation models pre-trained on large-scale data have shown remarkable performance in tasks including image and video generation Yu et al. (2023), speech recognition Rubenstein et al. (2023), and medical question answering Singhal et al. (2023). Recently, Thomas et al. (2022) adapted several prominent models in natural language processing including BERT Devlin et al. (2018) and GPT Radford et al. (2019), to learn the dynamics of brain activity in fMRIs. The models are trained on massive fMRI data from 11,980 experimental scans of 1,726 individuals across 34 datasets. A self-supervised task is adopted during training. The trained model is then fine-tuned on benchmark mental state decoding datasets and achieved improvements compared to the same model trained from scratch. BrainLM Ortega Caro et al. (2023) is a recently published foundation model for brain activity dynamics trained on 6,700 hours of fMRI recordings. The model consists of a Transformer-based Vaswani et al. (2017) masked auto-encoder architecture adapted from BERT Devlin et al. (2018) and Vision Transformer Dosovitskiy et al. (2020). During pre-training, BrainLM incorporates a self-supervised task that predicts the masked segments of time series in fMRI data, which is similar to the pre-training task in Thomas et al. (2022). They fine-tuned the model to predict metadata variables acquired from the UK BioBank dataset Allen et al. (2014) and achieved superior performance. In contrast to previous work that developed foundation models for fMRI using extensive datasets comprising vast

fMRI recordings, we propose a training strategy for a foundation model with relatively limited data and focus on improving the generalization of the model to downstream clinical applications.

2.2. Prediction of Post-Traumatic Epilepsy

Survivors of Traumatic Brain Injury (TBI) often experience significant disability due to their injuries Parikh et al. (2007). These injuries can lead to a range of physical and psychological effects, with some symptoms appearing immediately and others developing over time. Post-traumatic epilepsy (PTE) refers to recurrent and unprovoked post-traumatic seizures occurring after 1 week Verellen and Cavazos (2010). Identifying individual prognostic markers for PTE is crucial Engel Jr et al. (2013), as it can reduce the time and cost for TBI patients to begin clinical trials and decrease the risk of severe injury or death due to seizures. The prediction and prevention of PTE development remains a significant challenge. Animal studies in adult male Sprague-Dawley rats have shown the potential of MRI-based image analysis in identifying biomarkers for PTE Immonen et al. (2013); Pitkänen et al. (2016). These studies indicate the involvement of the perilesional cortex, hippocampus, and temporal lobe in PTE Pitkänen and Bolkvadze (2012). Despite progress, brain imaging is still not fully leveraged in PTE biomarker research. Various human neuroimaging studies have provided insights into TBI Dennis et al. (2016); Farbota et al. (2012); Kim et al. (2008) and epilepsy Li et al. (2009); Mo et al. (2019); Sollee et al. (2022), but fMRI-based PTE prediction is limited.

Clinical and research studies in epilepsy often include both anatomical (MRI, CT) and functional (PET, EEG, MEG, ECoG, depth electrodes, fMRI) mapping. While epileptogenic zones can be found in almost any location in the brain, the temporal lobe and the hippocampus are the most common sites causing focal epileptic seizures Sollee et al. (2022). Multimodal MRI and PET imaging has been used to predict the laterality of temporal lobe epilepsy Pustina et al. (2015); Sollee et al. (2022). Extensive changes in brain networks due to epilepsy were reported using PET, fMRI, and diffusion imaging Li et al. (2009); Pitkänen et al. (2016); Pustina et al. (2015); Akrami et al. (2021); Sollee et al. (2022). Recent studies employing machine learning to identify potential PTE biomarkers Rocca et al. (2019); Akrami et al. (2021, 2022) have primarily focused on pairwise correlation patterns in resting fMRI signals. However, the heterogeneity of PTE functional activity and data scarcity often lead to over-fitting and limited generalization in deep-learning approaches. Here we similarly focus on the use of only fMRI in PTE prediction, but with the novel use of a foundation-model approach for this problem.

3. Methods

Here we introduce our proposed strategy, MeTSK, which improves the generalization of self-supervised fMRI features from a control dataset to a clinical dataset. Assume there exists a source domain (healthy controls) S with abundant training data X_S and a target domain (clinical) T, where the training data X_T is limited. A feature extractor $f(\phi)$, a target head $h_T(\theta_t)$, and a source head $h_S(\theta_s)$ are constructed to learn source features $h_S(f(X_S; \phi); \theta_s)$ as well as target features $h_T(f(X_T; \phi); \theta_t)$, where ϕ , θ_t , and θ_s are model parameters. The overall framework of MeTSK and the foundation model pipeline is illustrated 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 sub-sequences from the whole fMRI time



Fig. 1. An illustration of the proposed MeTSK strategy for upstream training and downstream applications. 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. For downstream applications, we directly apply the pre-trained foundation model without any fine-tuning and generate zero-shot features for the downstream dataset. The zero-shot features are then used to train a simple classifier and generate final classification results.

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 and Welling (2016), applied to the spatial graph at time point *l* in the *i*-th layer, can be expressed as follows.

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

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. Meta Knowledge Transfer

We introduce a bi-level optimization strategy to perform gradient-based update of model parameters Finn et al. (2017); Liu et al. (2020). The model first backpropagates the gradients through the target head only in several fast adaptation steps, and then backpropagates 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 $X_{\mathcal{T}_{tr}} \cap X_{\mathcal{T}_{val}} = \emptyset$, $X_{\mathcal{T}_{tr}} \cup X_{\mathcal{T}_{val}} = X_{\mathcal{T}}$.

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.

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

$$\{\theta_{s}^{i+1}, \phi^{i+1}\} = \{\theta_{s}^{i}, \phi^{i}\} - \beta(\nabla_{\theta_{s}^{i}, \phi^{i}}\mathcal{L}_{\mathcal{S}}(h_{\mathcal{S}}(f(X_{\mathcal{S}}; \phi^{i}); \theta_{s}^{i})) + \nabla_{\phi^{i}}\lambda\mathcal{L}_{\mathcal{T}}(h_{\mathcal{T}}(f(X_{\mathcal{T}_{wi}}; \phi^{i}); \theta_{s}^{k})))$$

$$(3)$$

The target head, source head and feature extractor are updated in an alternating fashion. The target head is first trained on $X_{\mathcal{T}_{ir}}$ 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. (2020) to perform a self-



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

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. (2020).

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

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

where $\tilde{X}_{S} = h_{S}(f(X_{S}; \phi); \theta_{s})$ 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, and such features are generalizable across domains. 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 strategy, MeTSK, is

$$\mathcal{L}_{meta} = \mathcal{L}_{S} + \lambda \mathcal{L}_{\mathcal{T}}$$
$$\mathcal{L}_{\mathcal{T}} = -\sum_{\text{classes}} Y_{\mathcal{T}} \log(h_{\mathcal{T}}(f(X_{\mathcal{T}}; \phi); \theta_{t}))$$
(6)

3.4. Domain Similarity

To evaluate the generalization 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 and 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}_{S} = Flatten $(\frac{1}{N}\sum_{n=1}^{N}\tilde{X}_{S}), \bar{X}_{T}$ = 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}_{S} and \bar{X}_{T} , respectively. Domain similarity (DS) is defined in Eq. 7 and Eq. 8. A larger domain similarity indicates better transferability and generalizability from the source domain to the target domain because the amount of work needed to transform source features into target features is smaller.

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

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

where $B_s(i)$ is the i-th bin of the histogram and $|B_s|$ is the total number of bins, $|\bar{X}_S \in B_s(i)|$ is the number of features in $B_s(i)$, $|\bar{X}_S|$ 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 $\gamma = 0.01$.

4. Datasets

In this section, we introduce the datasets used to build the foundation model. The HCP Van Essen et al. (2013) and ADHD Bellec et al. (2017) datasets described below are used during the upstream training of the foundation model, the ABIDE dataset Craddock et al. (2013) is used in the ablation study of the proposed MeTSK strategy. We then introduce the PTE dataset that is used for evaluation of downstream performance.

4.1. Foundation Model Datasets

HCP dataset: The healthy control data for the foundation model 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 restingstate-fMRI data collected at a total of 1200 time-points per session. 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.

ADHD-Peking: 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 investigate the scenario where clinical data is scarce, we use only the subset of the larger ADHD database that was collected from the Peking site. 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 bandpass 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.

ABIDE-UM: The Autism Brain Imaging Data Exchange I (ABIDE I) Craddock et al. (2013) collects restingstate 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.

4.2. Downstream Clinical PTE Dataset

We use the Maryland TBI MagNeTs dataset Gullapalli (2011) for downstream 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 receiveronly 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.

5. Experiments and Results

5.1. Upstream Results

We first trained the foundation model using the proposed MeTSK strategy on the HCP data (healthy controls) and ADHD-Peking data (clinical data). To investigate the effectiveness of MeTSK, we designed an experiment for an upstream task that performs ADHD v.s. TDC classification. We evaluate different strategies and compare their effectiveness in enhancing the generalization from a healthy dataset to a clinical dataset.

For comparison, we designed (i) a baseline model using a ST-GCN with a supervised task directly trained on the ADHD-Peking data (Baseline),(ii) a ST-GCN model fine-tuned on ADHD-Peking data after pre-training on HCP data (FT), (iii) a model performing multi-task learning on HCP data and ADHD-Peking data 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 generalization to ADHD data. For the 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. 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 data and use all HCP data for training. For meta-learning, the ADHD training set in each fold is further divided into a meta-training set $X_{\mathcal{T}_{tr}}$ of 157 subjects and a metavalidation set $X_{\mathcal{T}_{val}}$ of 39 subjects. Model performance is evaluated on the test ADHD data set using the average area-under-

Method	HCP	ADHD-Peking	AUC	ACC
SVM	X	✓	0.6182 ± 0.0351	0.6086 ± 0.0412
RF	X	\checkmark	0.6117 ± 0.0503	0.6102 ± 0.0564
MLP	X	\checkmark	0.6203 ± 0.0468	0.6092 ± 0.0507
LSTM Gadgil et al. (2020)	X	1	0.5913 ± 0.0510	0.5652 ± 0.0539
STAGIN Kim et al. (2021)	X	1	0.5638 ± 0.0468	0.5279 ± 0.0511
Baseline (ST-GCN)	X	1	0.6215 ± 0.0435	0.6171 ± 0.0556
FT	1	1	0.6243 ± 0.0483	0.6367 ± 0.0501
MTL	1	1	0.6518 ± 0.0428	0.6316 ± 0.0513
MeTSK (ours)	1	\checkmark	0.6981 ± 0.0409	0.6775 ± 0.0443

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

the-ROC-curve (AUC) and classification accuracy (ACC) as evaluation metrics as shown in Table 1. MeTSK achieved the best mean AUC of 0.6981, which is a significant improvement compared to the baseline model trained only on ADHD data. MeTSK also surpassed the performance of fine-tuning and multi-task learning, providing evidence for overcoming limited generalization. The results from upstream training demonstrate that the MeTSK strategy possesses a clear capability to enhance generalization from healthy data to clinical data.

5.2. Downstream Results on PTE Dataset

For the downstream application we performed zero-shot evaluation on the PTE dataset. This involved initially extracting features from the PTE dataset using the pre-trained foundation model without any further fine-tuning. These extracted features are "zero-shot" features, as they are generated directly from the model trained on different datasets. Subsequently, we input these zero-shot features into a classifier to differentiate between PTE and non-PTE subjects, thereby assessing the model's ability to generalize and apply learned patterns to the downstream clinical applications.

Training a foundation model with only self-supervised learning is a typical approach. To compare different pre-training strategies for the foundation model, 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 generated zeroshot features for PTE data. We also compared our proposed foundation model to a large pre-trained fMRI model, as detailed in Thomas et al. (2022). This model involves pretraining 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 pretrained model is publicly available at https://github.com/ athms/learning-from-brains. We directly applied their pre-trained model to generate zero-shot PTE features.

Finally, we compare the zero-shot features generated from different foundation models with functional connectivity features extracted from raw fMRI data. We employed the same machine learning classifiers as used in the upstream experiments, including a linear SVM, RF, and MLP. The same 5-fold cross-validation was applied and AUCs for PTE v.s. non-PTE classification were computed.

The zero-shot features generated by the foundation model pre-trained using the MeTSK strategy achieved the best performance among all features in every classifier, as shown in Table 2, indicating superior generalization of the foundation model on the heterogeneous PTE dataset. The zero-shot features generated by the SSL model also achieved better performance than functional connectivity features, owing to the generalizable knowledge learned from upstream datasets. However, Thomas et al. (2022) achieved the worst performance, possibly because this pre-trained model needs further fine-tuning to boost its optimal performance. Notably, the best performance achieved by Linear SVM suggests that these zero-shot features are linearly separable. This outcome not only demonstrates MeTSK's ability to produce discriminative features for an unseen dataset like PTE but also highlights its potential in enhancing feature learning for clinical diagnostic purposes.

To gain further insights and improve the interpretability of the zero-shot PTE features from MeTSK, we computed a feature importance map derived from the positive SVM coefficients. In a linear SVM, each feature in each ROI is assigned a coefficient, indicating its significance in the decision-making process of the model. The higher the absolute value of a coefficient, the more impact that feature has on the model's predictions. We derived the coefficients for features of each ROI from the trained SVM and visualized these coefficients in the form of a feature importance map overlaid on the brain, which is shown in Fig. 3. Through observing the feature importance map, we can identify and interpret the most significant brain regions for PTE classification, which are mainly located in the temporal lobe. Given that epilepsy most commonly occurs in the temporal lobe, these significant brain regions identified from the zero-shot features offers potentially meaningful insights into the prediction of PTE. Interestingly, the other areas of high feature importance are in primary sensory (visual and somatomotor) regions.

5.3. Implementation Details

Upstream: To optimize model performance, we follow the training setting in Gadgil et al. (2020) for the ST-GCN model. 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 and 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 temperature parameter $\tau = 30$ to adjust the scale of losses following Liu et al. (2020); You et al. (2020). Since contrastive loss converges slowly Jaiswal et al. (2020), a warmup phase is applied to train the model only on HCP data using the graph contrastive loss for the first half of total training steps.

Downstream: We use the pre-trained feature extractor for

generating zero-shot features. The generated features are graphlevel 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 experiments 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 another 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, which is then averaged within each time-point and fed into classifiers. We follow the other detailed settings of the pre-trained GPT model in Thomas et al. (2022). We ran 100 iterations of stratified cross-validation on the PTE data for each method.

6. Ablation Study and Generalization Analysis

6.1. Experiments on ABIDE-UM

To investigate the robustness of our proposed pre-training strategy, MeTSK, across various clinical datasets, we also conducted experiments using the ABIDE-UM dataset as the target clinical dataset during upstream training. The same methods were compared and same experimental settings were applied to the ABIDE-UM data as for ADHD-Peking. We performed ASD v.s. TDC binary classification using the same 5fold cross-validation. As shown in Fig. 4, the performance on the ABIDE-UM dataset aligns with our findings for the ADHD-Peking dataset, with MeTSK consistently achieving the highest mean AUC among all compared methods. The results on ABIDE-UM illustrate MeTSK's applicability in different clinical datasets. When the downstream clinical task shares more similarities with ASD features or other clinical features, the training strategy of the foundation model can be adjusted to leverage different clinical features, demonstrating the flexibility of MeTSK in accommodating varying clinical datasets.

Table 2. Downstream results using 5-fold cross-validation: Mean and std of AUCs for PTE classification using zero-shot features generated from different foundation models as well as functional connectivity features.

		Connectivity Features		
	MeTSK	SSL	Thomas et al. (2022)	j
SVM	0.6415 ± 0.0312	0.5972 ± 0.0492	0.5369 ± 0.0451	0.5697 ± 0.0477
RF	0.5392 ± 0.0553	0.5253 ± 0.0486	0.4814 ± 0.0664	0.5081 ± 0.0612
MLP	0.5813 ± 0.0504	0.5216 ± 0.0329	0.5278 ± 0.0643	0.5111 ± 0.0402



Fig. 3. Feature importance map of zero-shot PTE features shown as color-coded ROIs overlaid on the AAL atlas.



Fig. 4. AUCs of 5-fold cross-validation on the ABIDE-UM dataset: a comparison of baseline (ST-GCN), fine-tuning (FT), multi-task learning (MTL), and other baseline methods.

6.2. Ablation Study of MeTSK

We examine the individual contributions of self-supervised learning and meta-learning to the model performance during upstream training 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 0.6215 to 0.6562 for ADHD classification, and from 0.6085 to 0.6675 for ASD classification without source domain knowledge. This finding is consistent with an increased generalization achieved by meta-learning on the clinical datasets.

Furthermore, 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. Finetuning, 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.

6.3. Generalization Analysis Using Domain Similarity

To further investigate the generalization enabled by MeTSK, domain similarity was computed to evaluate the generalizability 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 generalizability 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 metalearning, not only the inter-domain generalization of features is boosted, but also the effect of heterogeneous data within the same domain is alleviated.

7. Discussion and Conclusion

Our proposed strategy opens up new possibilities for enabling data-efficient generalization to downstream applications and handling extremely heterogeneous and scarce datasets that eluded traditional deep-learning approaches. According to Kumar et al. (2022), fine-tuning can distort good pre-trained features and degrade downstream performance under large distribution shifts. So unlike the common fine-tuning methods used in other foundation model approaches for fMRI analysis Ortega Caro et al. (2023); Thomas et al. (2022), we explored zero-shot features and linear probing for downstream adaptation, which achieved superior performance on the challenging PTE prediction task. Despite the improvements achieved, exciting future work still remains to be explored. We trained the foundation model on one healthy control dataset and one clinical dataset, an approach that is sensitive to the cost of data collection and expert annotation. Without these constraints, multiple datasets could be combined to learn generalizable functional activity patterns from a diverse span of subjects and clinical conditions.

To tackle the heterogeneity and scarcity of fMRI data, we propose a novel training strategy for developing a foundation model by learning from both clinical and healthy fMRI data. We integrate meta-learning with self-supervised learning to improve the generalization from normal features to clinical fea-

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	



Fig. 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.

tures during upstream training, and thus enhance the generalization to other unseen clinical features in a downstream task for predicting post-traumatic epilepsy. Specifically, we perform a self-supervised task on the healthy control dataset and apply meta-learning to transfer self-supervised knowledge to the clinical dataset. To explore the generalizability of the foundation model to a post-traumatic epilepsy (PTE) dataset, we compared zero-shot features generated by different foundation models for PTE classification. The features from MeTSK demonstrated the best performance. Additionally, the interpretation of the zeroshot PTE features may contribute to our understanding of PTE, offering insights into the identification of PTE via functional brain activity patterns in different brain regions. To summarize, the improved generalization of our foundation model in predicting PTE is attributed to: (i) the application of meta-learning, which bolsters the model's generalization to clinical features, and (ii) the use of self-supervised features that are inherently more task-agnostic and more generalizable.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

This work is supported by NIH grants: R01EB026299, R01NS074980 and DoD grants: W81XWH181061, HT94252310149.

References

- Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L., 2021. Graph-based deep learning for medical diagnosis and analysis: past, present and future. Sensors 21, 4758.
- Akbar, M.N., Ruf, S.F., Singh, A., Faghihpirayesh, R., Garner, R., Bennett, A., Alba, C., Imbiriba, T., La Rocca, M., Erdogmus, D., et al., 2022. Post traumatic seizure classification with missing data using multimodal machine learning on dmri, eeg, and fmri. medRxiv.

- Akrami, H., Irimia, A., Cui, W., Joshi, A.A., Leahy, R.M., 2021. Prediction of posttraumatic epilepsy using machine learning, in: Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging, SPIE. pp. 424–430.
- Akrami, H., Leahy, R., Irimia, A., Kim, P., Heck, C., Joshi, A., 2022. Neuroanatomic markers of posttraumatic epilepsy based on mr imaging and machine learning. American Journal of Neuroradiology 43, 347–353.
- Allen, N.E., Sudlow, C., Peakman, T., Collins, R., biobank, U., 2014. Uk biobank data: come and get it.
- Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al., 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering, 1–24.
- Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D.S., Craddock, R.C., 2017. The neuro bureau adhd-200 preprocessed repository. Neuroimage 144, 275–286.
- Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2020a. Generative pretraining from pixels, in: International conference on machine learning, PMLR. pp. 1691–1703.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.
- Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M., et al., 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics 7, 5.
- Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S., 2018. Large scale fine-grained categorization and domain-specific transfer learning. arXiv:1806.06193.
- Dennis, E.L., Hua, X., Villalon-Reina, J., Moran, L.M., Kernan, C., Babikian, T., Mink, R., Babbitt, C., Johnson, J., Giza, C.C., et al., 2016. Tensorbased morphometry reveals volumetric deficits in moderate/severe pediatric traumatic brain injury. Journal of neurotrauma 33, 840–852.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
- Engel Jr, J., Pitkänen, A., Loeb, J.A., Edward Dudek, F., Bertram III, E.H., Cole, A.J., Moshé, S.L., Wiebe, S., Jensen, F.E., Mody, I., et al., 2013. Epilepsy biomarkers. Epilepsia 54, 61–69.
- Farbota, K.D., Sodhi, A., Bendlin, B.B., McLaren, D.G., Xu, G., Rowley, H.A., Johnson, S.C., 2012. Longitudinal volumetric changes following traumatic brain injury: a tensor-based morphometry study. Journal of the International Neuropsychological Society 18, 1006–1018.
- Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR. pp. 1126–1135.
- Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Adeli, E., Pohl, K.M., 2020. Spatio-temporal graph convolution for resting-state fmri analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 528–538.
- Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178.
- Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al., 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124.
- Gullapalli, R.P., 2011. Investigation of Prognostic Ability of Novel Imaging Markers for Traumatic Brain Injury (TBI). Technical Report. BALTIMORE UNIV MD.
- Immonen, R., Kharatishvili, I., Gröhn, O., Pitkänen, A., 2013. Mri biomarkers for post-traumatic epileptogenesis. Journal of neurotrauma 30, 1305–1309.
- Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F., 2020. A survey on contrastive self-supervised learning. Technologies 9, 2.
- Ketkar, N., 2017. Stochastic gradient descent, in: Deep learning with Python. Springer, pp. 113–132.
- Kim, B.H., Ye, J.C., Kim, J.J., 2021. Learning dynamic graph representation

of brain connectome with spatio-temporal attention. Advances in Neural Information Processing Systems 34, 4314–4327.

- Kim, J., Avants, B., Patel, S., Whyte, J., Coslett, B.H., Pluta, J., Detre, J.A., Gee, J.C., 2008. Structural consequences of diffuse traumatic brain injury: a large deformation tensor-based morphometry study. Neuroimage 39, 1014– 1026.
- Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980 22.
- Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
- Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P., 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054.
- Li, W., He, H., Lu, J., Lv, B., Li, M., Jin, Z., 2009. Detection of whole-brain abnormalities in temporal lobe epilepsy using tensor-based morphometry with dartel, in: MIPPR 2009: Medical Imaging, Parallel Processing of Images, and Optimization Techniques, International Society for Optics and Photonics. p. 749723.
- Liu, H., HaoChen, J.Z., Wei, C., Ma, T., 2020. Meta-learning transferable representations with a single target domain. arXiv preprint arXiv:2011.01418
- Mo, J., Liu, Z., Sun, K., Ma, Y., Hu, W., Zhang, C., Wang, Y., Wang, X., Liu, C., Zhao, B., et al., 2019. Automated detection of hippocampal sclerosis using clinically empirical and radiomics features. Epilepsia 60, 2519–2529.
- Oh, J., Kim, S., Ho, N., Kim, J.H., Song, H., Yun, S.Y., 2022. Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty. arXiv:2202.01339.
- Ortega Caro, J., Oliveira Fonseca, A.H., Averill, C., Rizvi, S.A., Rosati, M., Cross, J.L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R.M., et al., 2023. Brainlm: A foundation model for brain activity recordings. bioRxiv, 2023–09.
- Palacios, E.M., Sala-Llonch, R., Junque, C., Roig, T., Tormos, J.M., Bargallo, N., Vendrell, P., 2013. Resting-state functional magnetic resonance imaging activity and connectivity and cognitive outcome in traumatic brain injury. JAMA neurology 70, 845–851.
- Parikh, S., Koch, M., Narayan, R.K., 2007. Traumatic brain injury. International anesthesiology clinics 45, 119–135.
- Pitkänen, A., Bolkvadze, T., 2012. Head trauma and epilepsy. Jasper's Basic Mechanisms of the Epilepsies [Internet]. 4th edition.
- Pitkänen, A., Löscher, W., Vezzani, A., Becker, A.J., Simonato, M., Lukasiuk, K., Gröhn, O., Bankstahl, J.P., Friedman, A., Aronica, E., et al., 2016. Advances in the development of biomarkers for epilepsy. The Lancet Neurology 15, 843–856.
- Pustina, D., Avants, B., Sperling, M., Gorniak, R., He, X., Doucet, G., Barnett, P., Mintzer, S., Sharan, A., Tracy, J., 2015. Predicting the laterality of temporal lobe epilepsy from pet, mri, and dti: a multimodal study. NeuroImage: clinical 9, 20–31.
- Rachev, S.T., 1985. The monge–kantorovich mass transference problem and its stochastic applications. Theory of Probability & Its Applications 29, 647– 676.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.
- Rocca, M.L., Garner, R., Jann, K., Kim, H., Vespa, P., Toga, A.W., Duncan, D., 2019. Machine learning of multimodal MRI to predict the development of epileptic seizures after traumatic brain injury. URL: https: //openreview.net/forum?id=Bye0tkLNcV.
- Rubenstein, P.K., Asawaroengchai, C., Nguyen, D.D., Bapna, A., Borsos, Z., de Chaumont Quitry, F., Chen, P., Badawy, D.E., Han, W., Kharitonov, E., Muckenhirn, H., Padfield, D., Qin, J., Rozenberg, D., Sainath, T., Schalkwyk, J., Sharifi, M., Ramanovich, M.T., Tagliasacchi, M., Tudor, A., Velimirović, M., Vincent, D., Yu, J., Wang, Y., Zayats, V., Zeghidour, N., Zhang, Y., Zhang, Z., Zilka, L., Frank, C., 2023. Audiopalm: A large language model that can speak and listen. arXiv:2306.12925.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., y Arcas, B.A., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S.S., Barral, J., Webster, D., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A., Natarajan, V., 2023. Towards expert-level medical question answering with large language models. arXiv:2305.09617.
- Sollee, J., Tang, L., Igiraneza, A.B., Xiao, B., Bai, H.X., Yang, L., 2022. Artificial intelligence for medical image analysis in epilepsy. Epilepsy Research

, 106861.

- Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3d self-supervised methods for medical imaging. Advances in Neural Information Processing Systems 33, 18158–18172.
- Thomas, A., Ré, C., Poldrack, R., 2022. Self-supervised learning of brain dynamics from broad neuroimaging data. Advances in Neural Information Processing Systems 35, 21255–21269.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage 15, 273–289.
- Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al., 2013. The wu-minn human connectome project: an overview. Neuroimage 80, 62–79.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.
- Verellen, R.M., Cavazos, J.E., 2010. Post-traumatic epilepsy: an overview. Therapy 7, 527.
- Wang, X., Yao, L., Rekik, I., Zhang, Y., 2022. 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, Springer. pp. 221–230.
- You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y., 2020. Graph contrastive learning with augmentations. Advances in Neural Information Processing Systems 33, 5812–5823.
- Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., et al., 2023. Language model beats diffusion-tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737.
- Yu, Z., Herman, G., 2005. 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.–. doi:10.1109/ICME.2005.1521516.
- Zhang, X.S., Tang, F., Dodge, H.H., Zhou, J., Wang, F., 2019. Metapred: Metalearning 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.
- Zhou, Y., Milham, M.P., Lui, Y.W., Miles, L., Reaume, J., Sodickson, D.K., Grossman, R.I., Ge, Y., 2012. Default-mode network disruption in mild traumatic brain injury. Radiology 265, 882.