BRAINM^3: A MULTI-TASK LEARNING FRAMEWORK BASED ON A MULTI-LEVEL MIXTURE-OF-EXPERTS FOR CROSS-DISEASE AND CROSS-DOMAIN DEMENTIA DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate differential diagnosis of dementia subtypes is crucial due to their distinct clinical trajectories and treatment responses. However, rare subtypes such as Lewy Body Dementia (LBD) suffer from data scarcity, and domain shifts across institutions further hinder model generalization. To address these challenges, we propose **BrainM**^{3}, a **M**ulti-task learning framework based on a **M**ulti-level **M**ixture-of-Experts (MoE) architecture for cross-domain and cross-disease **Brain** modeling. Our model jointly learns Alzheimer's disease (AD), mild cognitive impairment (MCI), and LBD diagnosis by disentangling disease-shared and specific brain connectivity features. At the domain level, a domain-aware Soft-MoE combined with adversarial training captures domain-invariant foundation brain representations, effectively mitigating scanner and cohort variability. At the task level, task-shared and task-specific Soft-MoEs enable mutual knowledge transfer and facilitate fine-grained pathological feature modeling. Experiments on multi-institutional datasets demonstrate that BrainM^3 consistently outperforms baselines under data heterogeneity. Moreover, our model offers interpretable insights into disease-relevant brain networks, offering potential clinical utility. Our work highlights the promise of an accurate and interpretable model for robust dementia diagnosis in real-world, cross-institution settings. Our code will be published based on acceptance.

1 INTRODUCTION

Dementia poses a significant threat to human health and presents substantial clinical and socioeconomic challenges (Arvanitakis et al., 2019). Among the various subtypes, Alzheimer's disease (AD) is the most common, followed by Lewy body dementia (LBD) ranks second in prevalence (Outeiro et al., 2019; Erkkinen et al., 2018; Hugo & Ganguli, 2014; Orad & Shiner, 2022). Because each subtype follows a distinct clinical trajectory and responds differently to available therapies, accurate differential diagnosis is essential for guiding effective, personalized treatment strategies and for slowing disease progression (Xue et al., 2024).

In recent years, deep learning models have emerged as a powerful tool for automating the detection of various neurological disorders, with a variety of modeling strategies proposed for AD diagnosis. For instance, graph-based models (Song et al., 2019; Ma et al., 2020; Zhang et al., 2023; Zhou et al., 2024) capture inter-regional brain connections; transformer-based models (Chen et al., 2024; Zhang et al., 2024; 2025a) effectively model long-range dependencies; and more recently, state-space models (Cao et al., 2024; Chen et al., 2025; Ren et al., 2025) have been introduced as a computationally efficient alternative for learning global spatiotemporal representations. While AD has been extensively studied and benefits from large-scale public datasets, LBD remains underexplored due to data scarcity, which also limits the applicability of complex deep learning models. Despite recent efforts (Falaschetti et al., 2024; Wang et al., 2025; Zhang et al., 2025b) developing models for LBD diagnosis, a common limitation is that they are trained on data-scarce tasks, which may hinder generalization and robustness.

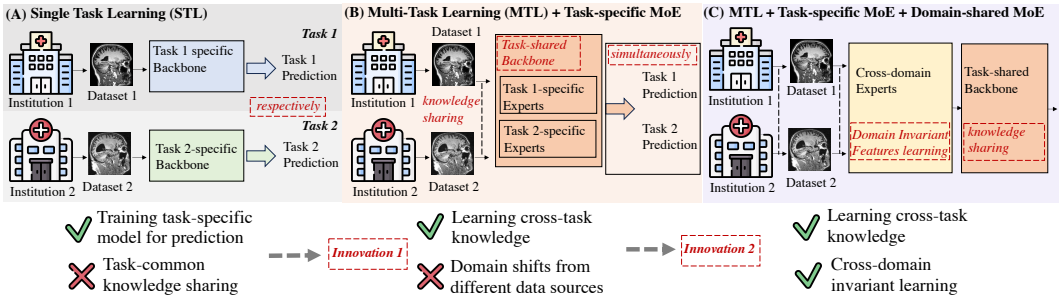

Figure 1: **Illustration of our innovations.** (A) Traditional single-task learning methods fail to capture task-shared knowledge. (B) *Our first innovation:* jointly learning multi-task in an unifed MTL framework and introducing task-specific MoE modules, learning both task-shared and task-specific features to enhance the diagnostic performance on both tasks. (C) *Our second innovation:* incorporating a cross-domain MoE, facilitating the learning of domain-invariant representations and mitigating domain shift caused by heterogeneous data sources.

Given that AD and LBD share common pathological features yet also exhibit disease-specific traits (Yousaf et al., 2019), a natural intuition is that shared brain representations can be leveraged across tasks, enabling mutual knowledge transfer between data-rich and data-scarce settings. Multi-task learning (MTL) (See Fig. 1(B)) offers a promising solution, as it enables simultaneous learning of multiple related tasks to model cross-task relationships and take advantage of task-shared representations. (Zhang & Yang, 2021; Liu et al., 2024; Zhou et al., 2011), which can effectively improve model generalization and reduce data scarcity in data-scarce tasks. However, another major challenge arises from data heterogeneity. Brain imaging data collected from different sites often differ due to scanners, protocols, and patient populations, leading to significant domain shifts. Training a unified MTL model on such heterogeneous data potentially biases learning and causes negative knowledge transfer (Aoki et al., 2022). In real-world clinical diagnostic protocols, this heterogeneous-featured MTL setting is more realistic but remains underexplored.

To address both the challenges of cross-task knowledge transfer and cross-domain heterogeneity, we propose **BrainM**^{3}, a **M**TL framework based on a **M**ulti-level **M**ixture-of-Experts (MoE) architecture (See Fig. 1(C)). At the domain level, a domain-aware MoE module and a domain adversarial training strategy are introduced to learn domain-invariant brain foundational representations. At the task level, we incorporate task-shared and task-specific MoEs to disentangle disease-related representations from domain-dependent foundational representations, enabling synergistic learning across data-rich and data-scarce diagnosis tasks. This hierarchical design allows the model to adaptively capture disease-related pathological patterns in a collaborative learning way while maintaining robustness across varied cohorts and imaging protocols. We validate our proposed model by comparing it with several established baselines on disease prediction tasks from diverse institutions. Experimental results demonstrate that our BrainM^3 consistently outperforms baseline methods, highlighting the effectiveness. To the best of our knowledge, this is the first work to explore the heterogeneous-feature multi-task learning problem in brain disorder research. The main contributions of this work are summarized as follows:

Domain-aware Adaptation: We introduce a domain-level MoE and a domain adversarial training strategy into a unified MTL framework to eliminate domain shifts caused by heterogeneous data, learning domain-invariant yet informative brain foundational representations.

Multi-task MoE Framework: We propose a unified MTL framework that incorporates both task-shared and task-specific MoE modules to jointly learn from data-rich and data-scarce diagnosis tasks. A residual fusion strategy integrates shared and specific features, capturing both general and disease-specific patterns.

Performance and Interpretability: We conduct extensive experiments on data-rich and data-scarce datasets, demonstrating the model's superior performance and generalizability, particularly under data scarcity and domain variability conditions. We also offer analysis of explainable insights into the model's decision-making process and the disease-related pathologies.

2 PRELIMINARIES

2.1 PROBLEM DEFINITION

We begin by defining **heterogeneous-feature multi-task learning (MTL)** problem (Zhang & Yeung, 2011). Given a set of m related but distinct tasks $\{T_i\}_{i=1}^m$, the goal is to jointly learn all tasks to improve performance on each individual task by leveraging shared knowledge. Unlike traditional homogeneous MTL where all tasks share the same input space, in heterogeneous-feature MTL, each task operates on a distinct feature space $\mathcal{X}_k$, i.e.,

$$y_k = T_k(B_k), \quad B_k \in \mathcal{X}_k, \quad \mathcal{X}_i \neq \mathcal{X}_j \text{ for } i \neq j \tag{1}$$

This setting mimics real-world scenarios where input distributions vary across tasks, yet each task still benefits from joint representation learning.

2.2 SOFT MIXTURE-OF-EXPERT(MOE)

MoE enables domain adaptation of heterogeneous data (Guo et al., 2018; Zhong et al., 2022; Jain et al., 2023; Wu et al., 2025b; Mi et al., 2025) and flexibly captures task-specific patterns (Wang et al., 2022; Fan et al., 2022; Chen et al., 2023; Zhu et al., 2024; Ding et al., 2025; Wu et al., 2025a) in a data-dependent routing. Soft MoE (Puigcerver et al., 2023) replaces the top-K hard selection mechanism in the sparse MoE (Mustafa et al., 2022) with a softmax-based token-to-expert assignment, allowing each input token to contribute to all experts in a weighted manner. Formally, given an input feature $x \in \mathbf{R}^{B\times N\times D}$, where B is the batch size, N is the number of brain sub-networks, and D is the embedding dimension, a gating network computes logits $G(x) \in \mathbf{R}^{B\times N\times K}$ over K experts. The softmax function then yields gating weights $\mathbf{w} = \mathrm{softmax}(G(x)) \in \mathbf{R}^{B\times N\times K}$. Each expert E_k processes the input x independently to generate an output, and the final MoE product is a weighted combination of expert outputs: $\mathrm{MoE}(x) = \sum_{k=1}^{K} w_k(x) \cdot E_k(x)$, where $w_k(x)$ denotes the weighting assigned to the expert k.

Our approach employs a shared Soft-MoE to learn domain-invariant foundational brain features, while task-shared and task-specific Soft-MoEs are introduced to capture both common and unique pathological patterns for each task, enabling multi-task knowledge sharing across diverse diseases and heterogeneous clinical populations.

2.3 BRAIN SUB-NETWORK REPRESENTATION

Brain structural connectivity (SC) patterns reflect disease-related alterations and offer informative representation of whole-brain network organization (Farooq et al., 2019; Škoch et al., 2022; Yeh et al., 2021; Zhang et al., 2021). To encode SC in a clinically meaningful and interpretable way, we introduce a brain-inspired token representation. Specifically, for a given subject, we applied standard imaging preprocessing (Zhang et al., 2022), including eddy current correction, fiber tracking, and registration of T1-weighted images to DTI space. Cortical segmentation was performed to parcellate the brain into 148 regions of interest (ROIs) based on the Destrieux Atlas (Destrieux et al., 2010). Pairwise connectivity strength between ROIs was computed based on the number of reconstructed white matter fibers, resulting in a symmetric matrix $X \in \mathbf{R}^{148\times 148}$, where each entry X_{ij} denotes the connectivity strength between ROI i and j. Here, we define each column of the SC matrix as a ***brain sub-network***, capturing the connectivity fingerprint of a single ROI. This design preserves anatomical topology and enables region-wise expert selection within the Soft-MoE module.

3 METHOD

As illustrated in Fig. 2, our goal is to jointly model multi-task diagnosis from diverse institutions by leveraging shared and specific brain SC patterns. To this end, we propose BrainM^3, a framework built on a shared backbone that consists of three key components: (1) a domain-shared Soft-MoE module to learn domain-invariant foundation brain representations (See Fig. 2 (B-I)), (2) task-specific and task-shared Soft-MoE modules to capture both common and unique pathologies for each diagnosis task (See Fig. 2 (B-II)), and (3) a domain adversarial component to further enforce cross-domain generalization.

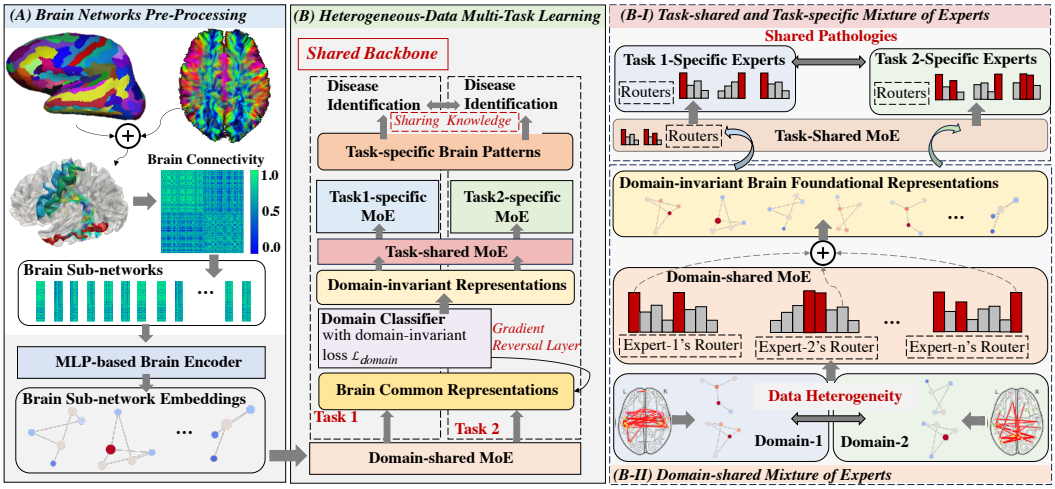

Figure 2: **Overview of our proposed framework**, which performs joint optimization across tasks under a unified shared backbone, including (A) brain networks processing and encoding, and (B) heterogeneous-feature MTL learning via hierarchical Soft-MoE.

3.1 Domain-Shared Feature Learning via Soft-MoE

To extract domain-invariant and task-free foundation features from brain SC, we introduce a domain-shared soft-MoE module, which serves as the core of the shared backbone and is jointly optimized across tasks. Specifically, as shown in Fig. 2(A), let $X \in \mathrm{R}^{R\times R}$ denote the SC matrix of a subject, where R is the number of brain regions. Each column vector $x_j \in \mathrm{R}^R$ reflects the connectivity profile of brain region j, capturing its relationship with all other regions. Each sub-network x_j is first projected through a token-wise MLP encoder to yield an embedded representation:

$$h_j = \mathrm{MLP}_{\mathrm{enc}}(x_j) \quad j = 1, \ldots, R \tag{2}$$

This produces an encoded sequence $H = [h_1, h_2, \ldots, h_R] \in \mathrm{R}^{R\times D}$, where D is the embedding dimension. To further obtain compact global brain network representations, we apply attention pooling over the sub-network sequence. Specifically, a learnable query vector $q \in \mathrm{R}^{1\times D}$ attends to all tokens using multi-head attention:

$$z_{\mathrm{global}} = \mathrm{MultiHeadAttn}(q, H, H) \tag{3}$$

where the query attends to the sequence H to produce a weighted global summary $z_{\mathrm{global}} \in \mathrm{R}^D$. This attention-based mechanism allows the model to focus on the most informative brain regions.

Subsequently, the encoded brain sub-networks are passed through a shared Soft-MoE layer consisting of K experts $\{E_k(\cdot)\}_{k=1}^K$, each implemented as an independent MLP. A gating network computes a soft assignment weight $\alpha_{j,k} \in [0, 1]$ for each pair of brain sub-network and expert:

$$\alpha_{j,k} = \frac{\exp(g_k(h_j))}{\sum_{k'=1}^K \exp(g_{k'}(h_j))} \tag{4}$$

The output of the Soft-MoE is a weighted sum of expert outputs:

$$\hat{h}_j = \sum_{k=1}^K \alpha_{j,k} \cdot E_k(h_j) \quad j = 1, \ldots, R \tag{5}$$

This results in a refined token representation $\hat{H} = [\hat{h}_1, \ldots, \hat{h}_R] \in \mathrm{R}^{R\times D}$ that captures domain-robust brain structural features.

3.2 Task-Shared and Task-Specific Feature Learning

To capture both shared and specific disease patterns, we introduce a dual-branch feature specialization design comprising task-shared and task-specific Soft-MoEs, which operate sequentially on domain-shared brain representations.

Let $\hat{H} = [\hat{h}_1, \ldots, \hat{h}_R] \in \mathrm{R}^{R \times D}$ be the output of the domain-shared MoE, where each $\hat{h}_j \in \mathrm{R}^D$ is the embedding of brain sub-network j. We first feed this sequence into a task-shared MoE module, which is shared across tasks and designed to extract generalizable pathological representations:

$$s_j = \sum_{k=1}^{K_s} \alpha_{j,k}^{(s)} \cdot E_k^{(s)}(\hat{h}_j) \quad j = 1, \ldots, R \tag{6}$$

where $\alpha_{j,k}^{(s)}$ are the gating weights for the shared experts, and $E_k^{(s)}(\cdot)$ denotes the k-th task-shared expert.

The output $s_j \in \mathrm{R}^D$ is further processed by task-specific MoEs for each task t. Each task-specific MoE has its own set of experts and gating functions:

$$z_j^{(t)} = \sum_{k=1}^{K_t} \alpha_{j,k}^{(t)} \cdot E_k^{(t)}(s_j) \quad j = 1, \ldots, R \tag{7}$$

where $\alpha_{j,k}^{(t)}$ are the task-specific gating weights and $E_k^{(t)}(\cdot)$ is the k-th expert for task t.

To effectively integrate shared and task-specific information, we apply a residual fusion mechanism by summing the outputs of the task-shared and task-specific MoEs:

$$\tilde{z}_j^{(t)} = s_j + z_j^{(t)} \quad j = 1, \ldots, R \tag{8}$$

The fused sequence $\tilde{Z}^{(t)} = [\tilde{z}_1^{(t)}, \ldots, \tilde{z}_R^{(t)}] \in \mathrm{R}^{R \times D}$ is then aggregated into a global representation via attention pooling:

$$z_{\text{global}}^{(t)} = \mathrm{MultiHeadAttn}(q, \tilde{Z}^{(t)}, \tilde{Z}^{(t)}) \tag{9}$$

followed by a task-specific classification head:

$$\hat{y}^{(t)} = \mathrm{Classifier}^{(t)}(z_{\text{global}}^{(t)}) \tag{10}$$

where $\hat{y}^{(t)} \in \mathrm{R}^{C_t}$ is the prediction for task t with C_t classes.

3.3 Domain Adversarial Learning

Despite the use of a shared feature extractor, subtle domain-specific variations may persist. To further promote domain-invariant representation learning, we introduce a domain adversarial training strategy based on a gradient reversal mechanism. Specifically, Let $H_{\text{shared}} \in \mathrm{R}^{B \times R \times D}$ denote the output token sequence from the domain-shared Soft-MoE, where B is the batch size, R is the number of brain regions, and D is the embedding dimension. To obtain global representations, we apply attention pooling over the sequence:

$$z_{\text{shared}} = \mathrm{MultiHeadAttn}(q, H_{\text{shared}}, H_{\text{shared}}) \tag{11}$$

A domain classifier $f_{\text{dom}}(\cdot)$ is trained to predict the domain label $d \in \{0, 1\}$. During training, we apply a gradient reversal layer (GRL) (Ganin et al., 2016) before the domain classifier:

$$\hat{d} = f_{\text{dom}}(\mathrm{GRL}(z^{\text{shared}})) \tag{12}$$

where GRL multiplies the incoming gradients by a negative scalar λ, encouraging the feature extractor to produce domain-invariant embeddings. The domain classification loss is computed as:

$$\mathcal{L}_{\text{domain}} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{CE}(\hat{d}_i, d_i) \tag{13}$$

where $\mathrm{CE}(\cdot, \cdot)$ is the cross-entropy loss between the predicted label $\hat{d}_i$ and the ground truth d_i of domain i.

This adversarial training paradigm formulates a minimax game, where in the domain classifier $f_{\text{dom}}(\cdot)$ is optimized to distinguish between data domains, while the feature encoder is simultaneously trained to produce domain-invariant representations that confuse the classifier.

3.4 Training Objectives

The proposed framework is trained end-to-end by jointly optimizing classification and domain adversarial losses. Let $\mathcal{L}_{\text{dr}}$ and $\mathcal{L}_{\text{ds}}$ denote the cross-entropy classification losses for the data-rich and data-scarce tasks, respectively. $\mathcal{L}_{\text{domain}}$ is the domain adversarial loss defined in section 3.3. The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dr}} + \mathcal{L}_{\text{ds}} + \lambda \cdot \mathcal{L}_{\text{domain}} \tag{14}$$

where λ is a hyperparameter for domain regularization.

During training, mini-batches from both data sources are alternately fed into the model. The shared encoder and domain-shared Soft-MoE are optimized jointly across tasks, while the task-specific MoEs and classifiers are trained independently per task. The domain classifier is trained adversarially via a gradient reversal layer to encourage the extraction of domain-invariant features. As shown in Fig. 2(B), this joint optimization strategy enables the model to leverage shared knowledge across tasks while maintaining task-specific discriminability and robustness to domain shift.

4 Experiments

4.1 Experimental Settings

Datasets In this study, we evaluated our proposed BrainM^3 model on two datasets: a public dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al., 2008), which offers a relatively large sample size, and an in-house dataset from an anonymized institution, characterized by limited data availability. The ADNI dataset is used for distinguishing normal controls (NC) from individuals with mild cognitive impairment (MCI), a prodromal stage of AD. After quality control, 418 subjects were included (301 NC and 117 MCI). The in-house dataset contains 147 subjects (23 NC, 77 LBD, and 47 AD). Data preprocessing followed the procedures described in section 2.3 *Brain Sub-network Representation*. We used 80% of the data for training and 20% for testing on each dataset. Additional preprocessing details and ADNI subject demographics are provided in the supplementary material.

Implementation Details In our method, we set the embedding dimension of each brain sub-network to 64. The MoE modules consist of 8 domain-shared experts, 4 task-shared experts, and 4 task-specific experts for each diagnosis task. The balancing hyperparameter λ for the $\mathcal{L}_{\text{domain}}$ is set to 1. The BrainM^3 is trained with a batch size of 16 for 128 epochs using the Adam optimizer with a learning rate of 1×10^{-4}. All experiments are conducted on a workstation equipped with an NVIDIA RTX 6000 GPU.

4.2 Comparison with Baselines

We evaluate the proposed BrainM^3 on multi-type dementia identification tasks using both the ADNI dataset and a private dataset. The performance is compared against several baseline methods, including two traditional machine learning models: Support Vector Machine (SVM) and XGBoost, two CNN/GNN-based approaches: BrainNetCNN (Kawahara et al., 2017) and FBNETGEN (Kan et al., 2022a), and two Transformer-based methods: VanillaTF (Kan et al., 2022b) and BrainNetTF (Kan et al., 2022b). All baseline models are trained on ***single diagnosis task*** separately, with hyperparameters adopted from their original papers. In the row "BrainM^3 (Single-task)" of Table 1, we ablate both the domain-shared and task-shared modules and ***train each single task independently*** to enables a fair comparison.

From a single-task perspective, CNN/GNN-based approaches consistently outperform traditional machine learning models by capturing complex brain topological patterns, and transformer-based methods further improve performance on both tasks by learning long-range global dependencies. Our model outperforms all baselines on the private dataset across all evaluation metrics, demonstrating its advantage in data-scarce scenarios. Compared to baseline methods, BrainM^3 integrates both local sub-network modeling and flexible expert routing, enabling data-adaptive representation learning that is more robust to limited learning samples. In contrast, baseline deep learning models may suffer from overfitting due to their high complexity.

Table 1: Performance comparison of different baselines on the ANDI and Private datasets. The best results within single-task learning are highlighted in underline, the overall best results across all methods are highlighted with **bold**.

Methods	ANDI (data-rich)				Private (data-rare)			
	ACC	AUROC	SEN	SPE	ACC	AUROC	SEN	SPE
SVM	60.00	75.01	50.19	57.50	63.33	75.15	53.89	55.56
XGBoost	63.33	75.32	63.89	63.53	70.00	75.47	55.97	75.59
BrainNetCNN	64.29	73.13	63.33	76.67	73.33	78.52	58.52	81.00
FBNETGNN	70.24	62.30	63.75	75.00	74.78	81.84	60.14	82.54
VanillaTF	71.43	73.78	66.77	73.33	76.33	82.50	62.50	83.92
BrainNetTF	73.62	75.07	67.17	77.00	75.97	81.97	60.88	84.13
BrainM^3 (Single-task)	70.16	79.81	67.50	73.33	76.67	84.54	66.04	82.46
BrainM^3 (Multi-task)	**84.38**	**85.69**	**69.34**	**88.68**	**80.00**	**88.13**	**72.12**	**87.30**

Notably, when ***both tasks are trained jointly***, performance improves significantly across the board, particularly on the ADNI dataset, where accuracy increases by 14.19%. Moreover, both tasks achieve approximately 5% improvements in AUROC, as illustrated in the row "BrainM^3 (Multi-task)" of Table 1. These highlight the effectiveness of cross-task representation learning. Through mutual learning between data-rich and data-scarce neurological-related tasks, our BrainM^3 improves performance by positive knowledge transfer and offers a promising direction for maximizing data use in real-world data-rare medical settings.

To further understand the model's behavior, we analyze the confusion matrix for the CN vs AD vs LBD diagnosis task. As shown in Fig. 4. the model correctly identifies most LBD and AD samples, but struggles with CN. This is potentially due to limited CN cases, which may lead to insufficient pattern learning. In contrast, it performs well on both AD and LBD diagnosis, which have relatively more samples. Notably, our BrainM^3 achieves clear separation between AD and LBD, a clinically challenging task due to their overlapping symptoms. This demonstrates that our BrainM^3 is capable of learning fine-grained disease-specific features, even in challenging, data-rare scenarios.

4.3 Interpretability

Beyond predictive accuracy, an ideal diagnostic model should also offer interpretability by uncovering disease-relevant brain patterns and decision-making process, which is essential for clinical trust. As shown in Fig. 3, we visualize the top-5 discriminative brain sub-networks for each diagnosis task based on expert activation scores from the task-specific MoE module, highlighting task-specific brain SC patterns that contribute most significantly to the model's predictions (See Fig. 3(A-B)). In addition, we identify the top-5 task-shared brain sub-networks by analyzing expert activation scores in the task-shared MoE, providing evidence of common pathological substrates across different dementia types (See Fig. 3(C)).

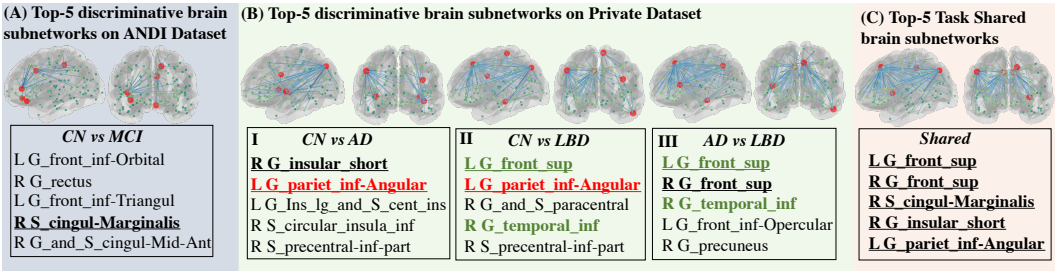

Figure 3: Top-5 discriminative brain sub-networks identified for each task, and the Top-5 most activated task-shared sub-networks. Shared sub-networks across diseases are highlighted in **bold** with different colors, while across task-specific and task-shared sets are additionally underlined.

Explainable Result on ADNI Dataset Figure 3 (A) illustrates the top-5 discriminative brain SC sub-networks distinguishing between CN and MCI, which are high aligned with previous studies. For instance, both the *L G_front_inf-Orbital* and the *R G_rectus* have been reported to show significantly reduced gray matter volume in individuals with MCI compared to CN (Han et al., 2012; Xie et al., 2015). The *L G_fron_inf-Triangul*, has been widely implicated in neurodegenerative diseases, especially those impacting language-related functions (Whitwell et al., 2015; Reyes et al., 2018; Mandelli et al., 2016). The *R G_and_S_cingul-Mid-Ant* belongs to the anterior cingulate cortex, a region frequently affected by structural, functional, and metabolic alterations in neurodegenerative conditions (Jones et al., 2006; Yuan et al., 2022). Additionally, the *R S_cingul-Marginalis* is located near a key hub of the default mode network (DMN), where AD-related pathological changes such as amyloid-β deposition and cortical atrophy often occur (Leech & Sharp, 2014). Notably, this region is also identified as a task-shared sub-network (see Fig. 3 (C)).

Explainable Result on Private Dataset Figure 3 (B) highlights the top-5 discriminative SC brain sub-networks that differentiate CN, AD, and LBD. As expected, our model effectively captures disease-related brain alterations. Frist, the *R G_insular_short*, located in the anterior insula, is a critical hub linking sensory perception, emotional processing and autonomic regulation (Uddin et al., 2017). This region is known to undergo significant structural atrophy in AD (Fathy et al., 2020) and is successfully identified in our method (see Fig. 3 (B-I)), and is also recognized as a task-shared sub-network (see Fig. 3 (C)). Moreover, the *L G_pariet_inf-Angular*, functions as a multimodal convergence zone and is a key hub within the brain's DMN(Wagner & Rusconi, 2023; Wang et al., 2019). In both AD and LBD, this region shows pronounced glucose metabolism reduce (Lim et al., 2009). Our model successfully identifies this sub-network as discriminative between both CN vs. AD and CN vs. LBD (see Fig. 3 (B-I, B-II)), and it is also recognized as a task-shared sub-network (see Fig. 3 (C)).

Furthermore, the *L G_front_sup* serves as a critical hub in multiple brain networks including DMN (Li et al., 2013). Both AD and LBD exhibit cortical atrophy and reduced glucose metabolism in this region, with LBD showing greater posterior involvement and a distinct metabolic pattern (Yousaf et al., 2019; Mistur et al., 2009). Similarly, the *R G_temporal_inf*, a region essential for high-level visual recognition and multimodal integration (Onitsuka et al., 2004), shows structural and functional abnormalities in LBD, while its relatively preserved metabolism compared to AD highlights its diagnostic relevance (Mak et al., 2014; Shivamurthy et al., 2015; Barber et al., 2000). Our model identifies these sub-network as discriminative in both CN vs LBD and AD vs LBD comparisons (see Fig. 3 (B-II, B-III)), and the *L G_front_sup* is identified as a task-shared sub-network (see Fig. 3 (C)). More importantly, the *R G_front_sup* is identified as both a discriminative sub-network between AD and LBD (see Fig. 3, (B-III)) and a task-shared sub-network (see Fig. 3 (C)). Prior studies have reported significant frontal lobe atrophy, reduced glucose metabolism, and disrupted functional connectivity in this region among individuals with AD, whereas these alterations are generally milder in LBD (Yousaf et al., 2019; Valdés Hernández et al., 2018; Tang et al., 2021; Roquet et al., 2016). Such distinctions are clinically important for differentiating between AD and LBD. Our method effectively captures these sub-network, highlighting the clinical interpretability of our model in capturing disease-specific brain alterations.

4.4 Ablation Studies

Number of Experts We categorize experts into three types: *domain-shared*, *task-shared*, and *task-specific*. In this ablation study, we vary the number and composition of experts to examine (1) the necessity of each expert type and (2) the effect of expert count in each MoE block on overall model performance. Experiments are conducted based on the BrainM^3 and the results are summarized in Table 2. Overall, models incorporating domain-shared experts consistently outperform those with the same total number of experts but without domain-shared ones, with 8 domain-shared experts yielding the best performance. Moreover, setting the number of experts in any individual MoE block to zero leads to a performance drop, underscoring the essential role of each MoE block in identifying brain diseases.

Effect of $\mathcal{L}_{\text{domain}}$ To investigate the impact of the domain adversarial loss $\mathcal{L}_{\text{domain}}$, we conduct a sensitivity analysis by varying its balancing weight λ in the overall objective. Fig. 5 presents the classification performance on both datasets under different λ values. As λ increases from 0 to 1, performance on both datasets improves significantly, with the best result achieved at $\lambda = 1$. When

λ exceeds 1, the performance on the ADNI dataset drops slightly, indicating that an overly strong domain alignment may suppress task-discriminative features. Additionally, we ablate the GRL in the domain classifier and the result shows that when it is removed, performance on both tasks drops. These results validate the effectiveness of the proposed domain adversarial mechanism in mitigating domain shift.

Table 2: Ablation Study on Expert Number and Structure. Here, m denotes the total number of experts, m_d denotes the number of domain-shared experts, m_s denotes the task-shared experts, and m_t denotes the task-specific experts. The best results under the same m are shown in underlined, while the overall best results across all settings are **bold**.

m	m_d	m_s	m_t	**ANDI (data-rich)**				**Private (data-rare)**			
				ACC	AUROC	SEN	SPE	ACC	AUROC	SEN	SPE
2	2	0	0	77.37	80.22	67.44	82.35	70.09	85.87	63.54	83.15
2	0	2	0	75.00	67.43	57.14	82.15	70.00	82.87	61.02	80.16
2	0	0	2	75.10	68.57	68.12	82.33	70.00	83.97	65.60	82.25
4	4	0	0	78.12	79.33	56.10	88.51	73.33	82.02	64.72	84.38
4	0	4	0	75.10	67.51	54.29	82.19	70.94	86.98	64.19	82.19
4	0	0	4	77.34	69.22	63.44	84.23	72.67	**89.32**	63.54	83.46
4	2	2	0	78.91	85.12	69.31	85.24	71.79	87.26	60.70	82.52
4	0	2	2	76.56	80.44	69.75	87.87	70.94	86.98	69.19	84.19
8	8	0	0	82.03	85.61	67.26	83.33	73.50	86.29	63.76	83.88
8	0	8	0	75.00	66.86	54.29	83.15	66.77	81.97	54.35	78.03
8	0	0	8	76.07	69.19	68.01	85.48	70.00	81.97	58.06	80.16
8	4	4	0	81.25	85.21	63.54	83.20	74.36	87.52	61.05	83.53
8	4	0	4	82.81	83.78	65.71	83.77	75.21	87.91	65.84	84.82
8	4	2	2	82.03	**86.57**	69.19	83.19	76.67	87.05	71.39	85.97
16	8	0	8	83.59	85.58	67.06	86.81	74.36	88.84	62.34	83.87
16	0	8	8	79.69	80.41	66.13	85.65	71.70	87.06	67.69	82.78
16	8	8	0	82.87	83.43	67.14	86.77	73.33	81.16	64.72	83.59
16	8	4	4	**84.38**	85.69	**69.34**	**88.68**	**80.00**	88.13	**72.12**	**87.30**

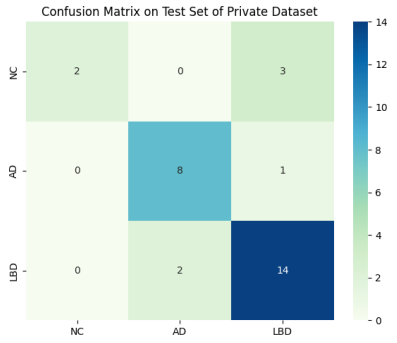

Figure 4: Confusion matrix for three-class diagnosis task.

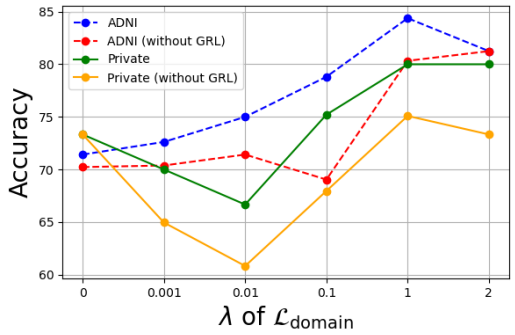

Figure 5: The impact of the hyperparameter λ on both tasks.

5 Conclusion

We propose BrainM^3, a novel MTL framework with a Multi-level Mixture-of-Experts architecture, designed to address both cross-domain heterogeneity and cross-task knowledge transfer in brain disorder diagnosis. The domain-shared MoE effectively mitigates domain shift, while task-specific and task-shared MoEs benefits knowledge transfer between date-rich and date-rare neurological tasks. Beyond superior performance, our model offers interpretable insights into disease-relevant brain sub-networks, providing potential clinical relevance. To our best knowledge, our work represents the first exploration of heterogeneous MTL in brain disorder research, establishing a foundation for robust, cross-institutional dementia diagnosis.

REFERENCES

Raquel Aoki, Frederick Tung, and Gabriel L Oliveira. Heterogeneous multi-task learning with expert diversity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6): 3093–3102, 2022.

Zoe Arvanitakis, Raj C Shah, and David A Bennett. Diagnosis and management of dementia. *Jama*, 322(16):1589–1599, 2019.

R Barber, C Ballard, IG McKeith, A Gholkar, and JT O'brien. Mri volumetric study of dementia with lewy bodies: a comparison with ad and vascular dementia. *Neurology*, 54(6):1304–1309, 2000.

Tangwei Cao, Xin Liu, Zuyu Du, Jiankui Zhou, Jie Zheng, and Lin Xu. A diagonal-structured-state-space-sequence-model-based deep learning framework for effective diagnosis of mild cognitive impairment. *IEEE Sensors Journal*, 24(10):16734–16743, 2024.

Minheng Chen, Xiaowei Yu, Jing Zhang, Tong Chen, Chao Cao, Yan Zhuang, Yanjun Lyu, Lu Zhang, Tianming Liu, and Dajiang Zhu. Core-periphery principle guided state space model for functional connectome classification. *arXiv preprint arXiv:2503.14655*, 2025.

Qiuhui Chen, Qiang Fu, Hao Bai, and Yi Hong. Longformer: longitudinal transformer for alzheimer's disease classification with structural mris. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3575–3584, 2024.

Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17346–17357, 2023.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1):1–15, 2010.

Rizhi Ding, Hui Lu, and Manhua Liu. Denseformer-moe: A dense transformer foundation model with mixture of experts for multi-task brain image analysis. *IEEE Transactions on Medical Imaging*, 2025.

Michael G Erkkinen, Mee-Ohk Kim, and Michael D Geschwind. Clinical neurology and epidemiology of the major neurodegenerative diseases. *Cold Spring Harbor perspectives in biology*, 10 (4):a033118, 2018.

Laura Falaschetti, Giorgio Biagetti, Michele Alessandrini, Claudio Turchetti, Simona Luzzi, and Paolo Crippa. Multi-class detection of neurodegenerative diseases from eeg signals using lightweight lstm neural networks. *Sensors*, 24(20):6721, 2024.

Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457, 2022.

Hamza Farooq, Yongxin Chen, Tryphon T Georgiou, Allen Tannenbaum, and Christophe Lenglet. Network curvature as a hallmark of brain structural connectivity. *Nature communications*, 10(1): 4937, 2019.

Yasmine Y Fathy, Susanne E Hoogers, Henk W Berendse, Ysbrand D van der Werf, Pieter J Visser, Frank J de Jong, and Wilma DJ van de Berg. Differential insular cortex sub-regional atrophy in neurodegenerative diseases: a systematic review and meta-analysis. *Brain Imaging and Behavior*, 14(6):2799–2816, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Jiang Guo, Darsh J Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. *arXiv preprint arXiv:1809.02256*, 2018.

Ying Han, Su Lui, Weihong Kuang, Qi Lang, Ling Zou, and Jianping Jia. Anatomical and functional deficits in patients with amnestic mild cognitive impairment. *PloS one*, 7(2):e28664, 2012.

Julie Hugo and Mary Ganguli. Dementia and cognitive impairment: epidemiology, diagnosis, and treatment. *Clinics in geriatric medicine*, 30(3):421, 2014.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

Yash Jain, Harkirat Behl, Zsolt Kira, and Vibhav Vineet. Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. *Advances in Neural Information Processing Systems*, 36:69625–69637, 2023.

Bethany F Jones, Josephine Barnes, Harry BM Uylings, Nick C Fox, Chris Frost, Menno P Witter, and Philip Scheltens. Differential regional atrophy of the cingulate gyrus in alzheimer disease: a volumetric mri study. *Cerebral cortex*, 16(12):1701–1708, 2006.

Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International conference on medical imaging with deep learning*, pp. 618–637. PMLR, 2022a.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022b.

Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

Robert Leech and David J Sharp. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32, 2014.

Wei Li, Wen Qin, Huaigui Liu, Lingzhong Fan, Jiaojian Wang, Tianzi Jiang, and Chunshui Yu. Subregions of the human superior frontal gyrus and their connections. *Neuroimage*, 78:46–58, 2013.

Seok Ming Lim, Andrew Katsifis, Victor L Villemagne, Rene Best, Gareth Jones, Michael Saling, Jennifer Bradshaw, John Merory, Michael Woodward, Malcolm Hopwood, et al. The 18f-fdg pet cingulate island sign and comparison to 123i-β-cit spect for diagnosis of dementia with lewy bodies. *Journal of Nuclear Medicine*, 50(10):1638–1645, 2009.

Jin Liu, Xu Tian, Hanhe Lin, Hong-Dong Li, and Yi Pan. Multi-task learning for alzheimer's disease diagnosis and mini-mental state examination score prediction. *Big Data Mining and Analytics*, 7 (3):828–842, 2024.

Junbo Ma, Xiaofeng Zhu, Defu Yang, Jiazhou Chen, and Guorong Wu. Attention-guided deep graph neural network for longitudinal alzheimer's disease analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 387–396. Springer, 2020.

Elijah Mak, Li Su, Guy B Williams, and John T O'Brien. Neuroimaging characteristics of dementia with lewy bodies. *Alzheimer's research & therapy*, 6(2):18, 2014.

Maria Luisa Mandelli, Eduard Vilaplana, Jesse A Brown, H Isabel Hubbard, Richard J Binney, Suneth Attygalle, Miguel A Santos-Santos, Zachary A Miller, Mikhail Pakvasa, Maya L Henry, et al. Healthy brain connectivity predicts atrophy progression in non-fluent variant of primary progressive aphasia. *Brain*, 139(10):2778–2791, 2016.

Zhenxing Mi, Ping Yin, Xue Xiao, and Dan Xu. Learning heterogeneous mixture of scene experts for large-scale neural radiance fields. *arXiv preprint arXiv:2505.02005*, 2025.

Rachel Mistur, Lisa Mosconi, Susan De Santi, Marla Guzman, Yi Li, Wai Tsui, and Mony J de Leon. Current challenges for the early detection of alzheimer's disease: brain imaging and csf studies. *Journal of clinical neurology*, 5(4):153–166, 2009.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

Toshiaki Onitsuka, Martha E Shenton, Dean F Salisbury, Chandlee C Dickey, Kiyoto Kasai, Sarah K Toner, Melissa Frumin, Ron Kikinis, Ferenc A Jolesz, and Robert W McCarley. Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: an mri study. *American Journal of Psychiatry*, 161(9):1603–1611, 2004.

Rotem Iris Orad and Tamara Shiner. Differentiating dementia with lewy bodies from alzheimer's disease and parkinson's disease dementia: an update on imaging modalities. *Journal of Neurology*, 269(2):639–653, 2022.

Tiago Fleming Outeiro, David J Koss, Daniel Erskine, Lauren Walker, Marzena Kurzawa-Akanbi, David Burn, Paul Donaghy, Christopher Morris, John-Paul Taylor, Alan Thomas, et al. Dementia with lewy bodies: an update and outlook. *Molecular neurodegeneration*, 14(1):5, 2019.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

Ziyin Ren, Meng Zhou, Sadia Shakil, and Raymond Kai-Yu Tong. Alzheimer's disease recognition via long-range state space model using multi-modal brain images. *Frontiers in Neuroscience*, 19: 1576931, 2025.

P Reyes, MP Ortega-Merchan, A Rueda, F Uriza, Hernando Santamaria-García, N Rojas-Serrano, J Rodriguez-Santos, MC Velasco-Leon, JD Rodriguez-Parra, DE Mora-Diaz, et al. Functional connectivity changes in behavioral, semantic, and nonfluent variants of frontotemporal dementia. *Behavioural neurology*, 2018(1):9684129, 2018.

Daniel Roquet, Marion Sourty, Anne Botzung, Jean-Paul Armspach, and Frédéric Blanc. Brain perfusion in dementia with lewy bodies and alzheimer's disease: an arterial spin labeling mri study on prodromal and mild dementia stages. *Alzheimer's research & therapy*, 8(1):29, 2016.

Veeresh KN Shivamurthy, Abdel K Tahari, Charles Marcus, and Rathan M Subramaniam. Brain fdg pet and the diagnosis of dementia. *American Journal of Roentgenology*, 204(1):W76–W85, 2015.

Antonín Škoch, Barbora Rehák Bučková, Jan Mareš, Jaroslav Tintěra, Pavel Sanda, Lucia Jajcay, Jiří Horáček, Filip Španiel, and Jaroslav Hlinka. Human brain structural connectivity matrices–ready for modelling. *Scientific Data*, 9(1):486, 2022.

Tzu-An Song, Samadrita Roy Chowdhury, Fan Yang, Heidi Jacobs, Georges El Fakhri, Quanzheng Li, Keith Johnson, and Joyita Dutta. Graph convolutional neural networks for alzheimer's disease classification. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pp. 414–417. IEEE, 2019.

Fanyu Tang, Donglin Zhu, Wenying Ma, Qun Yao, Qian Li, and Jingping Shi. Differences changes in cerebellar functional connectivity between mild cognitive impairment and alzheimer's disease: a seed-based approach. *Frontiers in Neurology*, 12:645171, 2021.

Lucina Q Uddin, Jason S Nomi, Benjamin Hébert-Seropian, Jimmy Ghaziri, and Olivier Boucher. Structure and function of the human insula. *Journal of clinical neurophysiology*, 34(4):300–306, 2017.

Maria del C Valdés Hernández, Stuart Reid, Shadia Mikhael, Cyril Pernet, and Alzheimer's Disease Neuroimaging Initiative. Do 2-year changes in superior frontal gyrus and global brain atrophy affect cognition? *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10(1):706–716, 2018.

Jennifer Wagner and Elena Rusconi. Causal involvement of the left angular gyrus in higher functions as revealed by transcranial magnetic stimulation: a systematic review. *Brain Structure and Function*, 228(1):169–196, 2023.

Di Wang, Nicolas Honnorat, Jon B Toledo, Karl Li, Sokratis Charisis, Tanweer Rashid, Anoop Benet Nirmala, Sachintha Ransara Brandigampala, Mariam Mojtabai, Sudha Seshadri, et al. Deep learning reveals pathology-confirmed neuroimaging signatures in alzheimer's, vascular and lewy body dementias. *Brain*, 148(6):1963–1977, 2025.

Junkai Wang, Jianghong Liu, Zhiqun Wang, Pei Sun, Kuncheng Li, and Peipeng Liang. Dysfunctional interactions between the default mode network and the dorsal attention network in subtypes of amnestic mild cognitive impairment. *Aging (Albany NY)*, 11(20):9147, 2019.

Sinan Wang, Yumeng Li, Hongyan Li, Tanchao Zhu, Zhao Li, and Wenwu Ou. Multi-task learning with calibrated mixture of insightful experts. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pp. 3307–3319. IEEE, 2022.

Jennifer L Whitwell, David T Jones, Joseph R Duffy, Edythe A Strand, Mary M Machulda, Scott A Przybelski, Prashanthi Vemuri, Brian E Gregg, Jeffrey L Gunter, Matthew L Senjem, et al. Working memory and language network dysfunctions in logopenic aphasia: a task-free fmri comparison with alzheimer's dementia. *Neurobiology of aging*, 36(3):1245–1252, 2015.

Chenwei Wu, Zitao Shuai, Zhengxu Tang, Luning Wang, and Liyue Shen. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Junxian Wu, Minheng Chen, Xinyi Ke, Tianwang Xun, Xiaoming Jiang, Hongyu Zhou, Lizhi Shao, and Youyong Kong. Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5144–5153, 2025b.

Yunyan Xie, Zaixu Cui, Zhongmin Zhang, Yu Sun, Can Sheng, Kuncheng Li, Gaolang Gong, Ying Han, and Jianping Jia. Identification of amnestic mild cognitive impairment using multimodal brain features: a combined structural mri and diffusion tensor imaging study. *Journal of Alzheimer's Disease*, 47(2):509–522, 2015.

Chonghua Xue, Sahana S Kowshik, Diala Lteif, Shreyas Puducheri, Varuna H Jasodanand, Olivia T Zhou, Anika S Walia, Osman B Guney, J Diana Zhang, Serena Poésy, et al. Ai-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 30(10):2977–2989, 2024.

Chun-Hung Yeh, Derek K Jones, Xiaoyun Liang, Maxime Descoteaux, and Alan Connelly. Mapping structural connectivity using diffusion mri: challenges and opportunities. *Journal of Magnetic Resonance Imaging*, 53(6):1666–1682, 2021.

Tayyabah Yousaf, George Dervenoulas, Polytimi-Eleni Valkimadi, and Marios Politis. Neuroimaging in lewy body dementia. *Journal of neurology*, 266(1):1–26, 2019.

Qianqian Yuan, Xuhong Liang, Chen Xue, Wenzhang Qi, Shanshan Chen, Yu Song, Huimin Wu, Xulian Zhang, Chaoyong Xiao, and Jiu Chen. Altered anterior cingulate cortex subregional connectivity associated with cognitions for distinguishing the spectrum of pre-clinical alzheimer's disease. *Frontiers in Aging Neuroscience*, 14:1035746, 2022.

Jianjia Zhang, Xiaotong Wu, Xiang Tang, Luping Zhou, Lei Wang, Weiwen Wu, and Dinggang Shen. Asynchronous functional brain network construction with spatiotemporal transformer for mci classification. *IEEE Transactions on Medical Imaging*, 2024.

Jing Zhang, Yanjun Lyu, Xiaowei Yu, Lu Zhang, Chao Cao, Tong Chen, Minheng Chen, Yan Zhuang, Tianming Liu, and Dajiang Zhu. Classiffication of mild cognitive impairment based on dynamic functional connectivity using spatio-temporal transformer. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2025a.

Jing Zhang, Xiaowei Yu, Tong Chen, Chao Cao, Mingheng Chen, Yan Zhuang, Yanjun Lyu, Lu Zhang, Li Su, Tianming Liu, et al. Brainnet-moe: Brain-inspired mixture-of-experts learning for neurological disease identification. *arXiv preprint arXiv:2503.07640*, 2025b.

Lu Zhang, Li Wang, Jean Gao, Shannon L Risacher, Jingwen Yan, Gang Li, Tianming Liu, Dajiang Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Deep fusion of brain structure-function in mild cognitive impairment. *Medical image analysis*, 72:102082, 2021.

Lu Zhang, Li Wang, Dajiang Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting brain structural network using functional connectivity. *Medical image analysis*, 79:102463, 2022.

Lu Zhang, Saiyang Na, Tianming Liu, Dajiang Zhu, and Junzhou Huang. Multimodal deep fusion in hyperbolic space for mild cognitive impairment study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 674–684. Springer, 2023.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

Yu Zhang and Dit-Yan Yeung. Multi-task learning in heterogeneous feature spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 574–579, 2011.

Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. *Advances in Neural Information Processing Systems*, 35:22243–22257, 2022.

Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 814–822, 2011.

Zhiheng Zhou, Qi Wang, Xiaoyu An, Siwei Chen, Yongan Sun, Guanghui Wang, and Guiying Yan. A novel graph neural network method for alzheimer's disease classification. *Computers in Biology and Medicine*, 180:108869, 2024.

Xun Zhu, Ying Hu, Fanbin Mo, Miao Li, and Ji Wu. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-moe. *Advances in Neural Information Processing Systems*, 37:81225–81256, 2024.

A APPENDIX

Subject Demographics and Preprocessing details of ADNI

The imaging protocol for each subject included structural MRI (T1-weighted) and diffusion tensor imaging (DTI). T1-weighted images were acquired with a field of view (FOV) of 240 mm × 256 mm × 208 mm, isotropic voxel size of 1.0 mm, and repetition time (TR) of 2.3 s. DTI data were obtained using a b-value of 1000 s/mm², 54 gradient directions, FOV of 232 mm × 232 mm × 160 mm, isotropic voxel size of 2.0 mm, TR of 7.2 s, and echo time (TE) of 56 ms. Preprocessing steps included skull stripping for both modalities, followed by registration of T1 images to DTI space using FSL. T1 images underwent tissue segmentation via FreeSurfer, with regions of interest (ROIs) defined according to the Destrieux Atlas. DTI preprocessing involved eddy current correction using FSL, followed by fiber tracking reconstruction using MedINRIA.

Demographic details of the ADNI subjects are presented in Table 3, including sample size, sex distribution, and age (mean ± standard deviation) for both NC and MCI.

Table 3: Demographic information of subjects from ADNI

	Mean ± standard deviation	
	NC	**MCI**
Sample size	301	117
Male/female	118/183	74/43
Male age (years)	71.37 ± 5.92	73.21 ± 6.86
Female age (years)	70.25 ± 5.91	70.17 ± 7.32

Brain Structural Connectivity Visualization Across Datasets

To qualitatively assess the structural differences across datasets, we visualize the brain structural connectivity matrices of three randomly selected subjects from each diagnostic group in both the ADNI and private datasets (See Figure 6).

In the ADNI dataset, both NC and MCI groups display structured and dense connectivity patterns, with symmetric topologies and moderate connection strengths. In contrast, the connectivity matrices from the private dataset (NC, AD, and LBD groups) are sparser. The distinct patterns highlight substantial distributional shifts between the two datasets, reinforcing the presence of cross-domain heterogeneity in brain structural connectivity. This observation motivates the integration of domain-invariant representation learning in our framework to enhance model robustness and generalizability.

The use of large language models (LLMs)

LLMs were used solely to improve the clarity and fluency of the language in this manuscript. All research ideas, methodology, experiments, analyses, and conclusions were conceived and carried out by the authors.

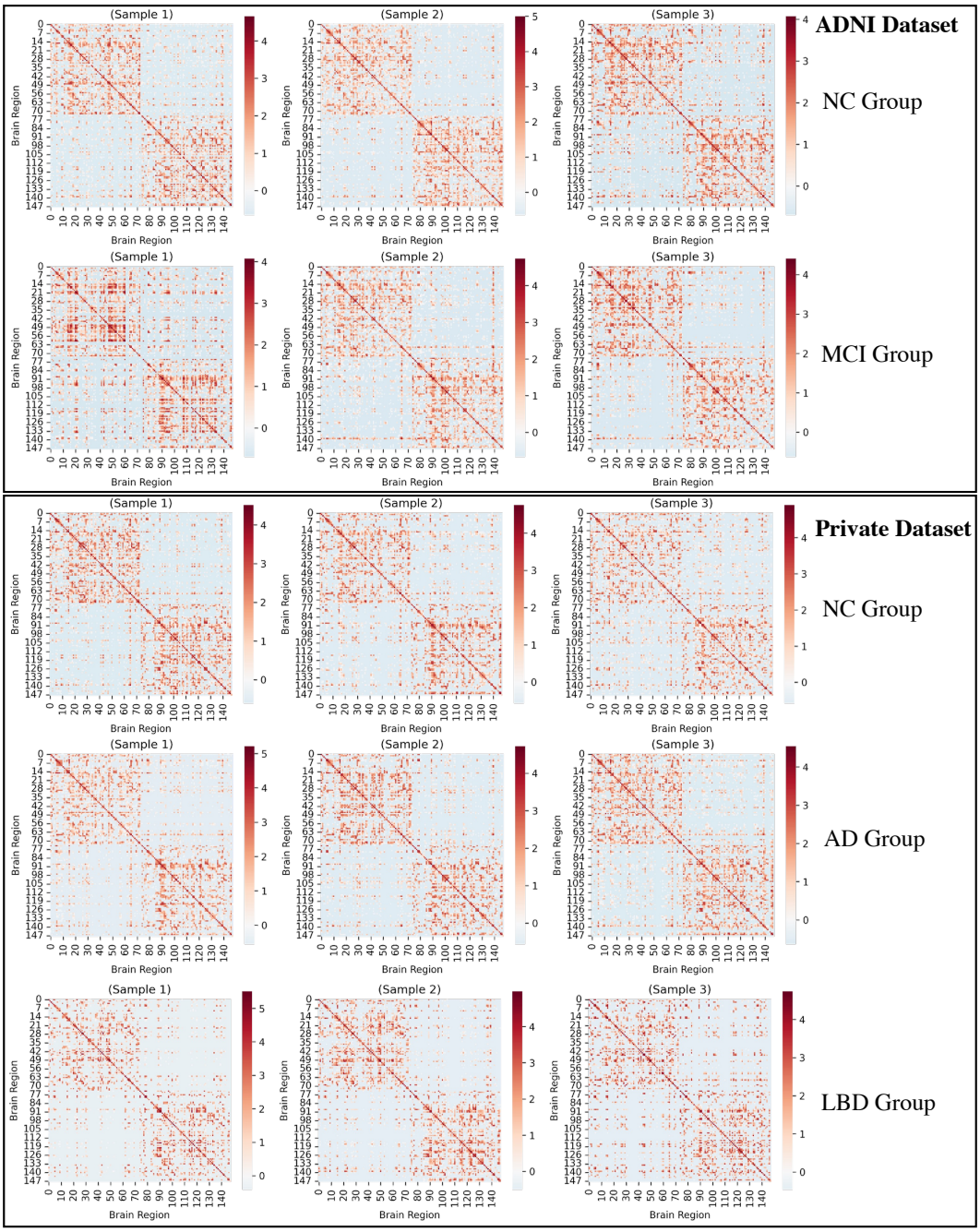

Figure 6: Brain structural connectivity matrices from ADNI and private datasets show distinct patterns across groups, revealing a clear domain shift in connectivity distributions.