Discovering Heterogeneous Neurodegenerative Disease Patterns From MRI Data for Improved Prediction

Yuanwang Zhang[1], Hongming Li[2], and Yong Fan[2,✉]

[1] Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia PA 19104, USA
ywzhang3@seas.upenn.edu
[2] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA 19104, USA
{Hongming.Li,Yong.Fan}@pennmedicine.upenn.edu

Abstract. Neurodegenerative diseases exhibit substantial heterogeneity, complicating both diagnosis and prognosis. Identifying clinically meaningful subtypes is crucial for understanding disease mechanisms and improving diagnostic precision and prognostic accuracy. Existing subtyping approaches primarily rely on unsupervised learning of patient data for capturing inter-individual variability, but they may fail to uncover subtypes that are directly informative for diagnosis or prognosis. To address this limitation, we propose a novel mixture-of-experts (MoE) framework that integrates predictive modeling with subtype discovery. Unlike traditional subtyping methods, our approach learns a router to assign individuals to specialized expert networks, each corresponding to a distinct subtype, for improving predictive accuracy. By coupling subtype identification with prediction, the proposed MoE framework encourages the discovery of subtypes that are not only statistically distinct but also clinically informative. We evaluate the framework on a real-world dataset of mild cognitive impairment (MCI) subjects and a semi-simulated dataset, demonstrating superior performance in predicting progression from MCI to Alzheimer's disease while revealing distinct, clinically meaningful MCI subtypes. Code is available at https://github.com/Kateridge/MoESubtyping.

Keywords: Disease Heterogeneity · Prediction-driven · Mixture-of-experts

1 Introduction

Neurodegenerative diseases, such as Alzheimer's disease, are complex disorders characterized by long prodromal phases and progressive deterioration. Disease progression is typically associated with changes in various biomarkers, including volumetric measurements from brain imaging and behavioral or cognitive measures from psychometric assessments [1]. These biomarkers have been extensively studied for their diagnostic and prognostic value. However, neurodegenerative diseases are biologically heterogeneous, and affected individuals may

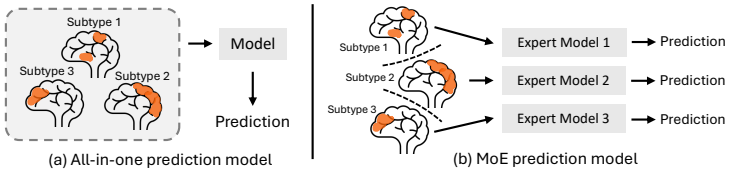

Fig. 1. (a) Most existing studies adopt a single all-in-one model to make predictions from heterogeneous data. (b) Our proposed MoE framework clusters heterogeneous data and routes samples to specialized expert models, improving prediction performance while providing interpretable insights into data heterogeneity.

exhibit distinct patterns in in-vivo biomarkers [11], posing significant challenges for their effective use in disease diagnosis and prognosis of disease progression. As illustrated in Fig. 1(a), most existing diagnosis and prognosis methods use an all-in-one model to make predictions from heterogeneous data, which may lead to suboptimal predictive performance.

To identify disease subtypes, many studies have adopted unsupervised clustering methods, such as K-Means and Gaussian Mixture Models (GMMs), to group individuals based on similarity of measurements derived from brain imaging data [21,6]. However, variability in imaging measures does not necessarily correlate with diagnosis or prognosis, meaning that these clusters often capture inter-subject differences without providing informative insights for these downstream tasks. Several studies have proposed semi-supervised approaches [23,24], which cluster diseased individuals relative to a reference group, such as cognitively unimpaired (CU) subjects, using generative modeling techniques. Although these methods achieve promising clustering performance, they are not explicitly designed for prediction tasks. In particular, they can cluster diseased individuals based on deviations from controls, but do not necessarily optimize diagnostic discrimination between diseased and control individuals. Alternative approaches integrate classification and clustering within a unified framework by employing ensembles of multiple SVM classifiers [9,22]. In these models, each linear hyperplane represents a potential subtype, and the ensemble of hyperplanes collectively forms a nonlinear decision boundary for distinguishing diseased individuals from the reference group. Although these methods enable simultaneous subtype identification and classification, the clustering remains primarily driven by feature similarity rather than predictive relevance.

Mixture-of-experts (MoE) architectures have recently emerged as a promising approach in large language models for scaling up the model capacity while reducing computational cost [3]. Recent works suggest that the MoE models can inherently learn underlying clustering structure in the input distribution, with routing decisions driven by the prediction objective [19,16,13]. Inspired by these findings, we develop an MoE framework that jointly performs prediction and clustering for heterogeneous neuroimaging data, as illustrated in Fig. 1(b). Unlike prior subtyping methods, our approach leverages predictive objectives to

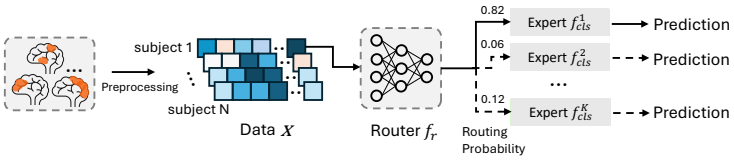

Fig. 2. Prediction-driven subtyping using an MoE framework. The MoE architecture assigns individuals to specialized expert networks, each representing a distinct subtype. These expert assignments naturally define cluster memberships, allowing the framework to perform prediction and subtyping simultaneously.

guide subtype discovery by assigning individuals to specialized expert networks that capture distinct disease patterns. As shown in Fig. 2, the framework consists of a router network and multiple expert networks. Each expert performs the prediction task and represents a disease subtype, while the router learns to assign each subject to the most appropriate expert. This design integrates prediction with clustering, enabling more accurate subtype identification while improving overall predictive performance. We validate the proposed framework and demonstrate substantial performance gains on two downstream tasks. We further show that these improvements stem from the inherent clustering capability of the MoE architecture. In addition, the proposed framework identifies four distinct patterns of brain atrophy in mild cognitive impairment (MCI) subjects, corresponding to different disease stages and subtypes.

2 Method

In this section, we present the proposed MoE framework, as shown in Fig. 2, along with additional regularization losses designed to improve robustness and prevent trivial solutions.

2.1 Mixture-of-experts (MoE) framework

Given N subjects, each represented by a vector of brain regional measurements $X_i \in \mathbb{R}^{\mathrm{M}}$ $(i = 1, 2, ..., \mathrm{N})$, our MoE model assigns each subject to one of K expert networks, denoted as $f_{\mathrm{ep}}^k(\cdot), k \in [1, \mathrm{K}]$, where M is the feature dimension and K is the number of experts. The assignment is determined by a router network $f_{\mathrm{r}}(\cdot)$, which outputs an assignment probability vector for each input X_i. Each subject is then routed to the expert with the highest probability. These expert assignments naturally define cluster memberships S_k over the input data.

In this study, we implement the MoE framework for two downstream applications: (1) classification of stable MCI versus progressive MCI and (2) predicting the progression risk of individuals from MCI to dementia. The MoE framework is trained using a task-specific prediction loss, denoted as $\mathcal{L}_{\mathrm{task}}$. For the classification task, each expert is a classifier, and the prediction loss $\mathcal{L}_{\mathrm{task}}$ is defined by the binary cross-entropy loss:

$$\mathcal{L}_{\text{task}} = -\frac{1}{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} \sum_{X_i \in S_k} \left[y_i \log(f_{\text{ep}}^k(X_i)) + (1 - y_i) \log(1 - f_{\text{ep}}^k(X_i)) \right], \quad (1)$$

where y_i denotes the binary label of subject i, and S_k denote the set of subjects in cluster k. For the prediction of progression risk, each expert is implemented as a Cox regression model, and the loss is defined by the partial likelihood of the Cox proportional hazards model [8,18]:

$$\mathcal{L}_{\text{task}} = \frac{1}{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} \sum_{X_i \in S_k} D_i \log \left(\sum_{j \in \mathcal{R}_i} \exp(r_j - r_i) \right), \quad (2)$$

where $r_j = f_{\text{ep}}^k(X_j)$ and $r_i = f_{\text{ep}}^k(X_i)$ represent the estimated risks of X_j and X_i, D_i is the event indicator for subject i, and $\mathcal{R}_i$ denotes the set of subjects which are uncensored and have not experienced the event prior to subject i.

2.2 Regularization losses

Training the model solely with the prediction loss may lead to degenerate routing behavior. To mitigate this issue, we introduce three regularization terms.

Load balance. MoE models often suffer from expert collapse, where most or all subjects are routed to a single expert. To prevent this, we adopt a load-balancing loss [7] that promotes balanced utilization of experts:

$$\mathcal{L}_{\text{lb}} = \sum_{k=1}^{\mathrm{K}} p_{\mathrm{k}} \log p_k, \quad (3)$$

where $p_k = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} [f_{\mathrm{r}}(X_i)]_k$ denotes the average assignment probability for expert k. Minimizing the negative entropy of the average assignment distribution across experts encourages the router to distribute subjects more uniformly among experts and thereby reduces expert collapse.

Sparse assignments. Another degenerate case occurs when the router network produces similar assignment probabilities across all experts, resulting in ambiguous routing decisions. To encourage confident expert assignments, we adopt a sparse assignment loss that minimizes the entropy of the assignment distribution for each subject:

$$\mathcal{L}_{\text{sp}} = -\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \sum_{k=1}^{\mathrm{K}} q_{ik} \log q_{ik}, \qquad q_i = f_{\mathrm{r}}(X_i), \quad (4)$$

where q_{ik} denotes the assignment probability of subject i to expert k. Minimizing this loss encourages each subject to be strongly associated with a single expert, leading to more distinct and interpretable expert assignments.

Clustering guidance. At the start of optimization, both the router and expert networks are randomly initialized, making it difficult for the router to learn meaningful patterns based solely on the prediction loss. To facilitate convergence and avoid degenerated situations, we introduce a clustering guidance loss $\mathcal{L}_{\text{guide}}$, which uses K-Means assignments as pseudo-labels to guide the router during early training. This loss is defined as the cross-entropy between the router's routing probabilities and the K-means cluster assignments. The weight of this loss is gradually decayed to zero over the first 20 epochs.

The overall objective is defined as a weighted sum of the task loss and three regularization terms: $\mathcal{L} = \lambda_1\mathcal{L}_{\text{task}} + \lambda_2\mathcal{L}_{\text{lb}} + \lambda_3\mathcal{L}_{\text{sp}} + \lambda_4\mathcal{L}_{\text{guide}}$, where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are hyperparameters that control the relative contributions of the corresponding loss terms.

3 Experiments

3.1 Experiments setup

Datasets and preprocessing. We collected 3,458 subjects with 11,602 longitudinal sessions from three cohorts: the Alzheimer's Disease Neuroimaging Initiative [20], the Open Access Series of Imaging Studies [17], and the Australian Imaging Biomarkers and Lifestyle Study of Aging [10]. Each session included valid diagnostic information, with 5,937 sessions labeled as cognitively unimpaired (CU), 4,703 as mild cognitive impairment (MCI), and 962 as dementia.

To construct the progression risk prediction dataset, we selected 1,243 subjects with baseline MRI scans and follow-up diagnostic records. Each subject had a baseline diagnosis of MCI and at least two sessions. To construct the classification dataset, we selected 668 MCI subjects with baseline MRI scans. They were categorized as stable MCI (no progression to dementia within a follow-up period longer than three years) or progressive MCI (progressing to dementia within three years). Subjects without sufficient follow-up time were excluded.

Raw T1-weighted MR images were preprocessed using the standard FreeSurfer[3] pipeline to extract cortical thickness measures from 68 cortical regions of interest (ROIs) based on the Desikan–Killiany atlas, as well as volumes from 28 subcortical ROIs based on the Aseg atlas.

Semi-simulated dataset. Since ground-truth subtype labels are unavailable in real-world datasets, we further evaluated clustering performance using a semi-simulated dataset. This dataset was generated by imposing simulated brain atrophy patterns on real CU data. Specifically, all CU sessions were randomly divided into a reference group of 2,937 sessions and a simulated disease group of 3,000 sessions. Within the simulated disease group, we created three subtypes, each containing 1,000 sessions, by manually introducing distinct atrophy patterns. For the first subtype, 15 cortical ROIs were randomly selected; for the second subtype, 5 subcortical ROIs were randomly selected; and for the third subtype, 15 cortical and 5 subcortical ROIs were randomly selected. The values

[3] https://surfer.nmr.mgh.harvard.edu/

Table 1. Prediction performance for classification and progression risk prediction. Mean metrics across folds with 95% confidence interval were reported.

Methods	Classification		Progression Risk Prediction	
	Accuracy ↑	Sensitivity ↑	C-Index ↑	AUC ↑
Single	0.729 (0.694-0.763)	0.710 (0.649-0.770)	0.725 (0.696-0.753)	0.787 (0.754-0.821)
Single*	0.736 (0.714-0.759)	0.699 (0.649-0.750)	0.738 (0.712-0.764)	0.803 (0.740-0.832)
MoE (2 experts)	0.750 (0.716-0.783)	0.720 (0.667-0.774)	0.750 (0.722-0.779)	0.813 (0.784-0.842)
MoE (3 experts)	**0.769 (0.741-0.797)**	0.726 (0.676-0.776)	0.757 (0.733-0.781)	0.817 (0.792-0.843)
MoE (4 experts)	0.765 (0.737-0.794)	**0.761 (0.699-0.823)**	**0.763 (0.740-0.786)**	**0.830 (0.802-0.858)**
MoE (5 experts)	0.761 (0.736-0.786)	0.737 (0.675-0.800)	0.756 (0.730-0.782)	0.819 (0.790-0.847)
Random Forests	0.732 (0.700-0.763)	0.671 (0.607-0.734)	0.749 (0.727-0.770)	0.814 (0.789-0.839)
TabFPN [12]	0.768 (0.732-0.803)	0.722 (0.659-0.786)	-	-
DeepSurv [15]	-	-	0.729 (0.707-0.752)	0.796 (0.764-0.828)

of the selected ROIs were then reduced by 10–30%, 10–20%, or 5–15% across different experimental settings, with the atrophy rate for each subject uniformly sampled within the specified range.

Model architecture and experimental setting. Each expert and the router network were implemented as a three-layer MLP with an input dimension of 96 and a hidden dimension of 128. Experimental results were obtained through 10-fold cross-validation to ensure robust performance evaluation. For the classification task, accuracy (ACC) and sensitivity (SEN) were reported. For the progression risk prediction task, the concordance index (C-index) and the averaged time-dependent AUC [2] were used as evaluation metrics. For clustering performance on the semi-simulated dataset, the adjusted Rand index (ARI) and normalized mutual information (NMI) were reported.

3.2 MoE improves prediction performance

Comparison with baselines. The baseline all-in-one prediction model, denoted as Single, was implemented using the same MLP architecture with a single expert network. To ensure a fair comparison with the MoE model, which has more parameters, we also implemented a higher-capacity baseline, denoted as Single*, with a similar model structure but approximately 5× parameters by increasing the hidden layer sizes. As shown in Table 1, the MoE framework substantially outperformed both the standard and higher-capacity all-in-one models on both tasks. These results highlight the benefits of routing heterogeneous inputs to specialized expert networks.

Effects of cluster number. We further investigated the impact of the cluster/expert numbers, as shown in Table 1. Prediction performance improved as the number of experts increased and reached its best performance with 4 experts. We also evaluated the subtyping results in terms of clustering stability, measured by adjusted Rand index (ARI), and clustering quality, measured by the Calinski–Harabasz (CH) index. For K = 2–5, the ARI values were 0.896, 0.840, 0.822, and 0.525, and the CH indices were 195.2, 113.4, 116.1, and 98.8, respectively. Although K = 2 achieved the highest clustering stability, it mainly separated subjects with mild versus severe atrophy. Therefore, we used K = 4 in

Table 2. Clustering performance on the semi-simulated dataset. Results are reported as the agreement between predicted subtypes and the ground-truth subtype labels.

Atrophy Rate	10%-30%		10%-20%		5%-15%	
	ARI ↑	NMI ↑	ARI ↑	NMI ↑	ARI ↑	NMI ↑
K-Means	0.937	0.916	0.640	0.636	0.289	0.367
MoE+K-Means	**0.998**	**0.997**	**0.717**	**0.719**	**0.419**	**0.476**
GMM	0.998	0.997	0.677	0.692	0.408	0.478
MoE+GMM	**1.000**	**1.000**	**0.713**	**0.720**	**0.451**	**0.504**
DecoNet [14]	0.998	0.997	0.647	0.674	0.513	0.579
MoE+DecoNet	**0.999**	**0.997**	**0.656**	**0.688**	**0.529**	**0.592**

Table 3. Ablation study. Clustering performance was evaluated on the semi-simulated dataset with 10%–20% atrophy rates, and prediction performance was assessed on both the classification and progression risk prediction tasks using 4 experts.

Regularization	Clustering (ARI)	Clustering (NMI)	Classification (ACC)	Progression Risk Prediction (C-Index)
$\mathcal{L}_{\text{task}}$ only	0.000	0.000	0.732	0.745
$\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{lb}}$	0.145	0.129	0.739	0.750
$\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{lb}} + \mathcal{L}_{\text{sp}}$	0.312	0.307	0.751	0.753
$\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{lb}} + \mathcal{L}_{\text{sp}} + \mathcal{L}_{\text{guide}}$	**0.717**	**0.719**	**0.765**	**0.763**

the following experiments, as it achieved the best predictive performance while yielding more fine-grained subtypes.

Comparison with other methods. We further compared our approach with several established methods, including Random Forests for both tasks, a zero-shot tabular foundation model, TabFPN [12], for the classification task, and a widely used survival analysis model, DeepSurv [15], for the progression risk prediction task. As shown in Table 1, our method outperformed Random Forests and DeepSurv, while achieving performance comparable to TabFPN.

3.3 MoE inherently characterizes heterogeneous patterns of MRI

We further demonstrate that the performance improvements of the MoE model are closely associated with its inherent ability to cluster heterogeneous inputs.

Clustering performance on the semi-simulated dataset. We evaluated clustering results using the semi-simulated dataset. The prediction task was formulated as binary classification between the reference group and the simulated disease group. To ensure a robust evaluation, we used clustering results from different methods as target labels of the clustering guidance loss, including K-means, Gaussian mixture models (GMM), and DecoNet [14]. As shown in Table 2, the MoE framework consistently improved clustering performance across different atrophy rate settings. These results indicate that the MoE model can inherently discover meaningful clusters in heterogeneous neuroimaging data.

Ablation study of regularization losses. We further evaluated the impact of each regularization loss on the MoE model, as summarized in Table 3. Without any regularization, the MoE model may converge to degenerate solutions that fail

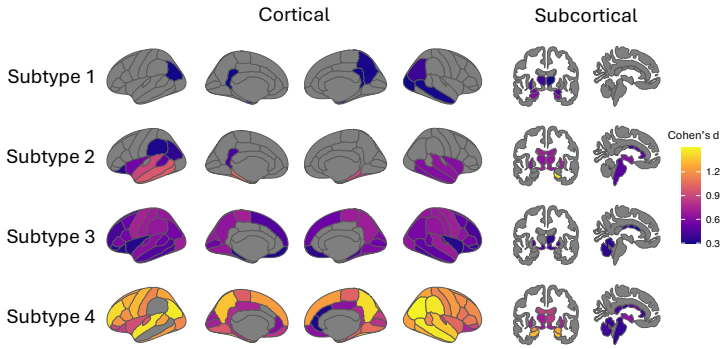

Fig. 3. Cortical and subcortical visualization of effect sizes (Cohen's d) for different MCI subtypes relative to the CU population. Higher values indicate greater differences compared to the CU population.

to meaningfully cluster the inputs, causing its predictive performance to collapse toward that of a standard all-in-one model. Introducing proposed regularization losses improved both clustering quality and prediction performance. These results demonstrate the effectiveness of each loss component and support that the MoE's predictive gains are closely linked to its inherent clustering ability.

3.4 Four subtypes identified in MCI population

Finally, we set the number of experts to K = 4 for progression risk prediction and grouped MCI subjects into four subtypes based on router assignments. For each subtype, we computed ROI-wise effect sizes relative to the CU population and visualized atrophy patterns on cortical Desikan–Killiany and subcortical Aseg atlases (Fig. 3). Subtype 4 exhibits severe atrophy across both cortical and subcortical regions, consistent with a late-stage MCI pattern associated with dementia-related pathology. In contrast, Subtype 1 showed only mild atrophy, consistent with an early-stage MCI pattern. Subtype 3 demonstrated widespread cortical atrophy, while Subtype 2 presented prominent atrophy in the medial temporal lobe, consistent with an amnestic MCI pattern. Analysis of clinical outcomes showed that Subtypes 1 and 3 had similar progression rates ($\sim 20\%$) and comparable cognitive decline despite exhibiting distinct spatial atrophy patterns. Similarly, Subtypes 2 and 4 had comparable progression rates ($\sim 50\%$) and cognitive decline. These findings suggest that similar clinical outcomes may arise from different neuroanatomical patterns, potentially reflecting heterogeneous disease trajectories underlying MCI progression.

4 Discussion and Conclusions

We developed a generalized MoE framework for joint prediction and subtype identification in neurodegenerative diseases. By leveraging multiple specialized

expert networks, the framework simultaneously enhances predictive performance and uncovers meaningful disease subtypes. The identified subtypes revealed distinct patterns of brain atrophy associated with different disease stages and phenotypes, underscoring the potential of prediction-driven subtyping to advance precision medicine in neurodegenerative disorders.

Our study could be further improved to enhance its generalizability and applicability. First, the framework is readily adaptable to voxel-wise imaging data by replacing the expert networks with image-based backbones and substituting the clustering guidance model with deep learning–based clustering methods [5,4]. This can potentially improve the performance and generalization because voxel-wise images may contain richer spatial information. Second, the clustering guidance loss was primarily designed to provide an effective initialization for the router, thereby improving convergence and preventing degenerate solutions. Future work could explore fully optimizable clustering objectives that are directly integrated into the framework, which may further enhance performance if the prediction and clustering objectives are appropriately balanced.

Acknowledgments. This work was supported in part by the National Institutes of Health under grants of R01AG066650 and R01AG054409.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Bateman, R.J., Xiong, C., Benzinger, T.L., Fagan, A.M., Goate, A., Fox, N.C., Marcus, D.S., Cairns, N.J., Xie, X., Blazey, T.M., et al.: Clinical and biomarker changes in dominantly inherited Alzheimer's disease. New England Journal of Medicine **367**(9), 795–804 (2012)
2. Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H.: Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biometrical journal **55**(5), 687–704 (2013)
3. Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., Huang, J.: A survey on mixture of experts in large language models. IEEE Transactions on Knowledge and Data Engineering (2025)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
6. Chen, I.Y., Krishnan, R.G., Sontag, D.: Clustering interval-censored time-series for disease phenotyping. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 6211–6221 (2022)
7. Chen, T., Li, H., Zheng, H., Fan, Y.: dfcexpert: Learning dynamic functional connectivity patterns with modularity and state experts. IEEE Transactions on Medical Imaging **45**(3), 1088–1098 (2026)

8. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
9. Eavani, H., et al.: Capturing heterogeneous group differences using mixture-of-experts: Application to a study of aging. Neuroimage **125**, 498–514 (2016)
10. Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., et al.: The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
11. Ferreira, D., Nordberg, A., Westman, E.: Biological subtypes of Alzheimer disease: a systematic review and meta-analysis. Neurology **94**(10), 436–448 (2020)
12. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate predictions on small data with a tabular foundation model. Nature **637**(8045), 319–326 (2025)
13. Hou, B., Li, H., Jiao, Z., Zhou, Z., Zheng, H., Fan, Y.: Deep clustering survival machines with interpretable expert distributions. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2023)
14. Kang, S., Kim, S.W., Seong, J.K.: Disentangling brain atrophy heterogeneity in Alzheimer's disease: A deep self-supervised approach with interpretable latent space. NeuroImage **297**, 120737 (2024)
15. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology **18**(1), 24 (2018)
16. Kawata, R., Matsutani, K., Kinoshita, Y., Nishikawa, N., Suzuki, T.: Mixture of experts provably detect and learn the latent cluster structure in gradient-based learning. In: International Conference on Machine Learning. pp. 29390–29448. PMLR (2025)
17. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., et al.: OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv pp. 2019–12 (2019)
18. Li, H., Habes, M., Wolk, D.A., Fan, Y.: A deep learning model for early prediction of alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's & Dementia **15**(8), 1059–1070 (2019)
19. Nielsen, S., Teo, R., Abdullaev, L., Nguyen, T.M.: Tight clusters make specialized experts. In: The Thirteenth International Conference on Learning Representations (2025)
20. Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack Jr, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., et al.: Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. Neurology **74**(3), 201–209 (2010)
21. Poulakis, K., Pereira, J.B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A., Wahlund, L.O., et al.: Heterogeneous patterns of brain atrophy in Alzheimer's disease. Neurobiology of aging **65**, 98–108 (2018)
22. Varol, E., et al.: HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. Neuroimage **145**, 346–364 (2017)
23. Yang, Z., et al.: A deep learning framework identifies dimensional representations of Alzheimer's disease from brain structure. Nature communications **12**(1), 7065 (2021)
24. Yang, Z., et al.: Brain aging patterns in a large and diverse cohort of 49,482 individuals. Nature medicine **30**(10), 3015–3026 (2024)