



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Decoding the Alzheimer’s Continuum: Interpretable Multi-Gate Routing for Diagnosis and Transition Prediction

Yufeng Jiang^{*1}, Hexiao Ding^{*1}, Hongzhao Chen^{*1}, Jing Lan¹, Xinzhi Teng¹,
Gerald W.Y. Cheng¹, Yunlin Mao¹, Zongxi Li², Haoran Xie², Jung Sun Yoo^{†1},
and Jing Cai^{†1}

¹ Department of Health Technology and Informatics,
Hong Kong Polytechnic University, Hong Kong SAR, China
{yufeng.jiang, hexiao.ding, hongzhao.chen, jing-hti.lan,
yunlin.mao}@connect.polyu.hk
{xinzhi.x.teng, wai-yeung.cheng, jungsun.yoo, jing.cai}@polyu.edu.hk

² Division of Artificial Intelligence, School of Data Science,
Lingnan University, Hong Kong SAR, China
{zongxili, hrxie}@ln.edu.hk

^{*}Co-first author. [†]Corresponding author.

Abstract. Alzheimer’s disease (AD) manifests as a continuous progression from normal cognition (NC) through mild cognitive impairment (MCI) to dementia. However, most deep learning approaches reduce this continuum to disjointed classification tasks, largely ignoring dynamic stage transitions. To decode this complex progression, we propose M³AD, a unified framework that jointly addresses three-class diagnosis classification and diagnosis stage transition prediction using only T1-weighted sMRI. M³AD leverages an interpretable multi-gate mixture of experts architecture, employing specialized routing mechanisms to dynamically capture both diagnosis-specific pathological patterns and shared structural features across the continuum. It further integrates clinical priors (age, sex, eTIV) via adaptive attention fusion to enhance generalization. M³AD achieves 95.13% accuracy, compared to 90.44% reported by MCLNC under its original experimental setting, and 94.87% for transition prediction. Crucially, analyzing the multi-gate routing reveals distinct expert activation signatures distinguishing stable from progressive MCI, providing a mechanistic basis for individual-level progression risk stratification. Code is available at <https://github.com/csyfjiang/M3AD>.

Keywords: Neuroimaging · Multi-task Learning · Alzheimer’s Disease · Mixture of Experts · Diagnosis Stage Transition.

1 Introduction

Neuro-degeneration along the Alzheimer’s disease (AD) continuum, which spans from normal cognition (NC) through mild cognitive impairment (MCI) to AD

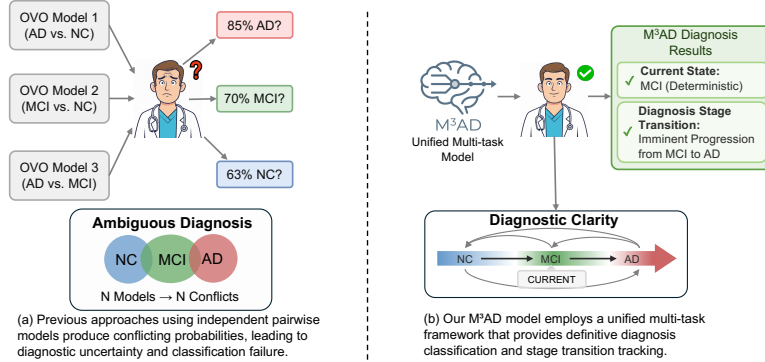


Fig. 1. Comparison between the conventional one-versus-one (OVO) strategy and M^3AD . While independent OVO classifiers produce conflicting probabilities leading to diagnostic ambiguity, M^3AD resolves this via a unified multi-task model that jointly predicts the current diagnosis (NC/MCI/AD) and stage transitions along the AD continuum.

dementia, manifests as quantifiable structural brain alterations [22] that are effectively captured and tracked via structural MRI (sMRI) [8]. Beyond static diagnostic neuroimaging biomarkers, sMRI also detects latent indicators of stage transition patterns driven by the nature of neuroplasticity [4]. However, most deep learning frameworks reduce this continuum to simple two-class or three-class classification tasks [1, 24], failing to model these clinically valuable dynamic trajectories. Furthermore, independent one-versus-one (OVO) classifiers [5] often generate conflicting probabilities due to unconstrained label overlap (Fig. 1), fundamentally misrepresenting the mutually exclusive nature of real-world clinical assessments while remaining blind to dynamic stage transitions along the disease continuum. To address these limitations, the multi-task Multi-gate Mixture-of-Experts (MMoE) framework leverages shared and task-specific features jointly, reducing gradient interference and improving convergence stability [3, 17]. Its effectiveness in neuroimaging has been demonstrated across diverse applications [12, 14]. Moreover, to fully ground such models in clinical reality, it is crucial to incorporate clinical priors. Age, sex, and estimated total intracranial volume (eTIV) are established correlates of brain atrophy and AD risk, yet rarely integrated in a principled manner within existing frameworks [15].

To address these gaps, we propose M^3AD , a **M**ulti-task **M**ulti-gate **M**ixture of experts framework built upon Swin V2 [16] with Tok-MLP [25]. M^3AD jointly addresses diagnosis stage classification (NC/MCI/AD) and dynamic stage transition prediction, incorporating age, sex, and eTIV as clinical priors via adaptive attention fusion. Specifically, SimMIM pretraining first initializes diagnosis-specific expert specialization, followed by multi-task fine-tuning where specialized experts capture pathological patterns and shared experts model common structural features. Validated across six datasets comprising 12,037 sMRI scans, M^3AD achieves 95.13% accuracy for three-class diagnosis, outperforming the

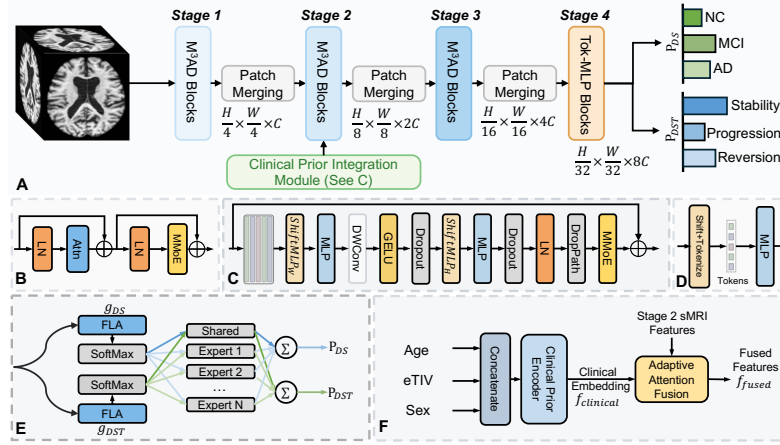


Fig. 2. Overview of the M³AD framework. (A) Four-stage hierarchical architecture processing sMRI inputs with clinical prior integration at Stage 2, producing dual outputs P_{DS} (NC/MCI/AD) and P_{DST} (Stability/Progression/Reversion). (B) M³AD block with MSA and MMoE. (C) Tok-MLP block with dual ShiftMLP operations. (D) ShiftMLP module. (E) Dual gate MMoE routing via feature-level attention (FLA). (F) Clinical Prior Integration Module with adaptive attention fusion.

state-of-the-art [29] by 4.69% and surpassing standard vision transformer baselines including ViT (89.54%). Beyond performance, interpretability analysis of the dual-gate routing mechanism reveals distinct expert activation signatures between MCI stability cases and MCI to AD progression cases, offering a mechanistic basis for individual-level progression risk stratification.

The contributions of this work are summarized as follows. (1) We introduce a unified M³AD framework that jointly optimizes three-class diagnosis stage classification and diagnosis stage transition pattern prediction within a single end-to-end MMoE architecture, validated on 12,037 sMRI scans across six datasets. (2) We design a clinical prior integration module incorporating age, sex, and eTIV via adaptive attention fusion, enabling context-sensitive balancing of imaging and demographic information across diverse patient populations. (3) We present an interpretation analysis of the dual-gate routing mechanism, unveiling distinct expert activation signatures and demonstrating that our imaging–demographic fusion is consistent with established theories of functional brain networks.

2 Methodology

Study Cohort Design Diagnosis stage classification used diagnosis stage labels NC, MCI, and AD. Diagnosis stage transition prediction used only the baseline first visit T1 weighted structural MRI scan and a 36 month observation window, with the endpoint defined as the last available follow up diagnosis within

that window, and the diagnosis stage transition label derived from baseline and endpoint diagnoses as Stability, Progression, or Reversion.

Datasets This study uses six neuroimaging cohorts with 12,037 T1-weighted structural MRI scans. The primary cohort is ADNI³, contributing 8,243 scans with diagnosis stage label set {NC, MCI, AD} and diagnosis stage transition labels for multi-task learning. DLBS⁴, IXI⁵, NKI-RS [23], OASIS-1 [19], and OASIS-2 [18] contribute 3,794 additional scans to broaden acquisition variability and reduce domain shift. Among the six cohorts, only ADNI provides longitudinal transition annotations; the remaining five datasets contribute exclusively to the diagnosis stage classification task. We formulate diagnosis stage transition prediction with a 3-class label set (C3: Stability, Progression, and Reversion) and diagnosis stage transition pattern prediction with a theoretically complete 9-class label set (C9) enumerating all specific inter-stage transitions. Due to the absence of AD Reversion cases (AD to MCI; and AD to NC) in the original dataset, C9 is empirically reduced to 7 observable classes in our implementation. All scans undergo standardized preprocessing with MONAI [2], N4 bias field correction, skull stripping with HD-BET [11], and SyN registration to MNI152 [7]. Intensity normalization uses z-score standardization within the brain foreground. All datasets were accessed under their respective data use agreements and institutional approvals.

M³AD Framework Overview Figure 2A illustrates the M³AD architecture. The model is built on a hierarchical vision transformer structure inspired by Swin V2 [16]. It processes sMRI data through four stages, where the first three stages utilize our proposed M³AD Blocks (Figure 2B). These blocks integrate an MMoE mechanism to better handle multi-task learning. In the final stage (Stage 4), we employ Tok-MLP Blocks [25] to integrate global contextual information for effective semantic mapping, facilitating the subsequent multi-task predictions. To enhance the model’s clinical relevance, we introduce a Clinical Prior Integration Module (Figure 2F). This module is inserted specifically at Stage 2, where it fuses clinical data (Age, Sex, eTIV) with the image features. The final output branches into two tasks: Diagnosis Stage Classification (P_{DS}) and Diagnosis Stage Transition Prediction (P_{DST}).

Base Architecture Enhanced Swin V2 As illustrated in Fig. 2B, the M³AD block integrates the MMoE mechanism into the Swin V2 [16] architecture by replacing the standard feed-forward network, with the computation defined as follows:

$$\tilde{z}^l = \text{MSA}(\text{LN}(z^{l-1})) + z^{l-1}, \quad z_t^l = \text{MMoE}_t(\text{LN}(\tilde{z}^l)) + \tilde{z}^l, \quad t \in \{\text{DS}, \text{DST}\}. \quad (1)$$

³ <https://adni.loni.usc.edu/>

⁴ https://fcon_1000.projects.nitrc.org/indi/retro/dlbs.html

⁵ <https://brain-development.org/ixi-dataset/>

In the last stage, standard Swin V2 blocks are replaced by Tok-MLP [25] as shown in Fig. 2C and Fig. 2D. Tok-MLP applies shifted MLP operations along spatial axes to model local atrophy patterns in sMRI

$$T_W = \text{Tokenize}(\text{Shift}_W(z^{l-1})), \quad \tilde{z}^l = f(\text{DWConv}(\text{MLP}(T_W))), \quad (2)$$

$$T_H = \text{Tokenize}(\text{Shift}_H(\tilde{z}^l)), \quad z^l = f(\text{LN}(T_W + \text{MLP}(\text{GELU}(T_H)))), \quad (3)$$

where DWConv provides positional encoding, and $\text{Shift}_W/\text{Shift}_H$ partition the features into h groups and cyclically shift them by j positions along the width and height axes respectively, inducing axial locality prior to tokenization [25].

Multi-gate Mixture of Experts (MMoE) Framework The MMoE uses $E = 8$ experts as shown in Fig. 2E. Two experts are shared across tasks, and the remaining six are organized as two diagnosis-aligned experts per class in NC, MCI, AD. Each expert maps the input by $f_e(\tilde{z}^l) = \text{MLP}_e(\text{LN}(\tilde{z}^l))$. For each task, a task-specific gate produces expert weights via feature-level attention:

$$g^t(\tilde{z}^l) = \text{Softmax}\left(\frac{\mathbf{W}_g^t \cdot \text{FLA}(\text{LN}(\tilde{z}^l))}{\tau}\right); \quad z_t^l = \sum_{e=1}^E g_e^t(\tilde{z}^l) \cdot f_e(\tilde{z}^l) + \tilde{z}^l, \quad (4)$$

where $t \in \{\text{DS}, \text{DST}\}$ denotes diagnosis stage classification and diagnosis stage transition prediction.

Clinical Prior Integration Module Figure 2F shows clinical priors $\mathbf{p} = [p_{\text{age}}, p_{\text{sex}}, p_{\text{eTIV}}] \in \mathbb{R}^3$. A Clinical Prior Encoder maps $\mathbf{p}$ to $\mathbf{p}_{\text{encoded}} \in \mathbb{R}^{C_{\text{fusion}}}$ and projects it to a feature tensor $\mathbf{X}_{\text{clinical}}$ compatible with Stage 2 features. Adaptive Attention Fusion computes $\mathbf{X}_{\text{fused}} = \mathbf{W}_{\text{proj}}(w_0\mathbf{X} + w_1\mathbf{X}_{\text{clinical}})$, where $(w_0, w_1) = \text{Softmax}(\text{MLP}(\text{AvgPool}([\mathbf{X} \parallel \mathbf{X}_{\text{clinical}}])))$.

Two-Stage Training Strategy (1) SimMIM Pretraining. Following SimMIM [27], training fold sMRI inputs are masked with ratio 0.6 and reconstructed. Pretraining is performed exclusively within each training fold, with no access to validation or test subjects, ensuring no label leakage across folds. Expert selection is guided by diagnosis stage labels, and each diagnosis-aligned expert is trained on its assigned subset. This structured initialization encourages each expert to encode distinct anatomical representations, enabling the dual gates in subsequent fine-tuning to dynamically compose shared and stage-specific knowledge, which also provides the basis for interpretable expert activation analysis

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{expert}}, \quad \mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \|I_{i,j} - \hat{I}_{i,j}\|_1, \quad (5)$$

where $\mathcal{L}_{\text{expert}}$ measures the L1 reconstruction error for each diagnosis-aligned expert on its corresponding samples, encouraging stage-specific specialization.

(2) Supervised fine-tuning. Dual gates are enabled to dynamically route and compose the pretrained expert representations, and all parameters are optimized with a multi-task objective over diagnosis stage classification and diagnosis stage transition prediction

$$\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{DS}} + \mathcal{L}_{\text{DST}}, \quad (6)$$

where $\mathcal{L}_{\text{DS}}$ is a cross entropy loss over the diagnosis stage label set {NC, MCI, AD} and $\mathcal{L}_{\text{DST}}$ is a cross entropy loss over the diagnosis stage transition label set {*Stability*, *Progression*, *Reversion*}.

3 Experiments

Implementation Details M³AD is implemented in PyTorch 2.7.1. Training runs on one server with four NVIDIA H800 GPUs with 80 GB memory each. Optimization uses AdamW with learning rate 10^{-4} for pretraining and 5×10^{-5} for fine-tuning, cosine annealing, weight decay 0.05, gradient clipping 1.0, and mixed precision; the MMoE gating temperature τ is set to 1.0 during pretraining and annealed to 0.5 during fine-tuning. Full hyperparameter configurations are provided in the code repository. Training lasts 200 epochs with batch size 256 per GPU. We employ five-fold cross-validation with early stopping under multiple random seeds, ensuring that all longitudinal scans from the same subject are confined to a single fold to prevent data leakage. Sagittal slices are extracted and resampled to 256×256 resolution with z-score intensity normalization, and augmentation is restricted to neuroanatomy-preserving transformations. At validation, predictions from all sagittal slices of a subject are aggregated via majority voting to produce the final subject-level label. We note that slice-level labels constitute weak supervision derived from subject-level diagnoses; while majority voting mitigates the impact of unreliable individual slices at inference, training-time label noise from uninformative slices remains a limitation and will be addressed in future work.

Results Among the baselines, ResNet-50 [9], DenseNet [10], and ViT [6] are re-implemented under our experimental protocol, while MCAD [28], Stacked DAE [26], PDMML [15], and MCLNC [29] are cited from their respective original publications for reference. As reported in Table 1 on the left side, ResNet-50 (89.21%), DenseNet (88.67%), and ViT (89.54%) achieve accuracies below 90% on diagnosis stage classification, which suggests limited sensitivity to the subtle structural differences separating NC, MCI, and AD. Among prior specialized methods, MCAD and Stacked DAE remain weaker despite multi-modal inputs. MCAD applies cross-modal attention to unaligned heterogeneous features and reduces volumetric representations from 128^3 to 4^3 through pooling, which removes localized atrophy cues required for diagnosis stage discrimination. Stacked DAE concatenates reconstruction-oriented autoencoder features across modalities, yielding representational mismatch that degrades diagnosis stage label separability. MCLNC reaches 90.44% accuracy, the highest among

Table 1. Comparison with prior methods. **Left:** Three-class diagnosis stage classification {NC, MCI, AD} on pooled datasets with five-fold cross-validation. **Right:** Cross-cohort generalization evaluated as NC vs AD binary classification, trained on ADNI and tested on OASIS-1&2. Results for other methods are cited from their original publications on respective datasets for reference. [†] denotes SimMIM-pretrained initialization. Best values are shown in **bold**. “–” indicates results not reported.

Method	Acc	Rec	Pre	Spe	F1
ResNet-50 [9]	89.21	88.45	87.93	92.13	88.18
DenseNet [10]	88.67	87.32	88.01	91.45	87.66
ViT [6]	89.54	88.91	89.12	92.67	89.01
MCAD [28]	64.03	63.85	–	82.00	61.85
Stacked DAE [26]	78.00	78.00	77.00	–	78.00
PDMML [15]	80.80	81.00	81.00	–	81.00
MCLNC [29]	90.44	86.29	88.97	93.47	87.47
MT-M ³ AD-C3 [†]	95.13	94.84	94.15	97.54	94.48
MT-M ³ AD-C9 [†]	94.72	93.82	95.23	97.03	94.47

Method	Acc	Rec	Pre	Spe	F1
ResNet-50 [9]	91.34	89.76	90.23	92.18	89.99
DenseNet [10]	92.47	91.83	91.56	93.21	91.69
ViT [6]	93.12	92.45	92.87	93.78	92.66
LA-GMF [22]	93.02	–	–	–	91.00
MC-CL [13]	93.16	95.00	94.44	94.44	94.72
SSH&LSH [20]	93.65	91.22	–	96.25	–
MCLNC [29]	98.60	98.86	97.82	98.86	98.29
MT-M ³ AD-C3 [†]	98.75	96.18	95.82	99.47	95.99
MT-M ³ AD-C9 [†]	98.08	97.21	95.34	99.38	96.27

prior methods, but its instance-level contrastive objective suppresses category-level commonalities within {NC, MCI, AD} and complicates optimization. MT-M³AD-C3 achieves 95.13% accuracy and 94.48% F1, improving over MCLNC by 4.69% and exceeding the best general backbone by 5.59%, while the specificity of 97.54% reduces false positive diagnoses critical for screening. MT-M³AD-C9 attains 94.72% accuracy with 95.23% precision, showing that the gain persists under a nine-class diagnosis stage transition label set. On the cross-cohort binary task (Table 1 on the right side), general backbones trained on ADNI transfer poorly to OASIS, with ViT reaching only 93.12% despite the simplified NC vs. AD protocol, confirming that backbone capacity alone does not ensure domain generalization. Prior specialized methods improve upon this, yet MCLNC, while achieving the highest recall (98.86%) and F1 (98.29%) among competitors, underlines that subject-level contrastive objectives remain sensitive to cohort shift. MT-M³AD-C3 achieves 98.75% accuracy and 99.47% specificity, surpassing all baselines on both metrics and demonstrating that the learned representations generalize across acquisition protocols without domain adaptation.

Ablation Study As shown in Table 2, multi-task learning consistently outperforms single-task models. For MT-M³AD-C3, pretraining improves diagnosis accuracy to 95.13% and transition accuracy to 94.87%, while removing the Tok-MLP reduces these to 93.47% and 93.56%, respectively. Integrating clinical priors at Stage 2 achieves optimal diagnosis accuracy, surpassing Stage 0 (93.87%), Stage 1 (94.23%), and Stage 3 (94.01%). Excluding all priors drops diagnosis accuracy by 1.92%. Among single priors, eTIV (94.45%) provides the largest gain, followed by age (94.21%) and sex (93.56%). Notably, as most transition patterns concentrate on stability, a long-tail issue persists; future work will focus on addressing this class imbalance.

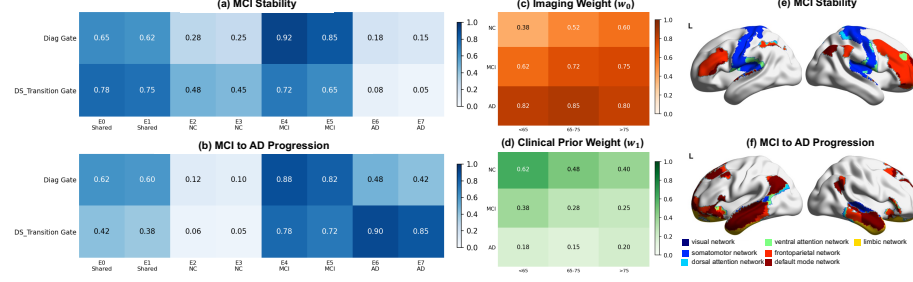


Fig. 3. Interpretation analysis of M³AD. (a)(b) Expert-gate activation heatmaps contrasting the MCI Stability subgroup and the MCI to AD progression subgroup, indicating increased activation of the AD aligned expert under the diagnosis stage (DS) transition gate in the progression subgroup. (c)(d) Adaptive fusion weights across diagnosis stage groups and age cohorts, where the imaging weight w_0 increases from NC to AD and the clinical prior weight w_1 decreases accordingly. (e)(f) GradCAM [21] visualizations of discriminative brain regions for the MCI Stability subgroup and the MCI to AD progression subgroup, mapped into seven functional brain networks.

Interpretation Analysis As shown in Fig. 3(a)(b), the diagnosis stage transition gate in the MCI to AD progression subgroup increases activation of the AD aligned experts E6 to E7 from 0.08 to 0.90 relative to the MCI Stability subgroup. This gating shift indicates cross stage feature sharing that supports multi-task training and yields an individual-level progression risk signal. Fig. 3(c)(d) shows adaptive fusion weights across diagnosis stage groups and age cohorts. The imaging weight w_0 increases with disease severity, with NC in the range 0.38 to 0.60 and AD in the range 0.82 to 0.85, while the clinical prior weight w_1 is larger in younger NC subjects at 0.62 where structural differences are limited. Fig. 3(e)

Table 2. Performance of M³AD under single-task, multi-task, and ablation settings for diagnosis stage classification and transition pattern prediction on ADNI dataset. C3 and C9 denote ternary and nine-class transition annotations, respectively. Results are reported as mean $\pm$ std across cross-validation folds. $\dagger$ denotes SimMIM-pretrained initialization.

Method	Diagnosis		Transition Pattern	
	Acc($\pm$ std)	F1($\pm$ std)	Acc($\pm$ std)	F1($\pm$ std)
Single Tasks				
M³AD-C9	89.12(1.8)	88.12(1.7)	89.34(1.8)	69.21(2.8)
M³AD-C3	89.12(1.6)	88.12(1.6)	91.23(1.6)	82.47(2.3)
M³AD-C9 †	94.80(0.8)	94.49(0.7)	91.56(1.2)	73.18(2.1)
M³AD-C3 †	94.80(0.7)	94.49(0.6)	93.45(1.0)	85.34(1.6)
Multiple Tasks				
M³AD-C9	89.23(1.9)	89.25(1.8)	90.67(1.7)	71.83(2.6)
M³AD-C3	90.15(1.7)	89.24(1.7)	92.34(1.5)	84.12(2.1)
M³AD-C9 †	94.72(0.9)	94.47(0.8)	93.21(1.0)	75.47(1.8)
M³AD-C3 †	95.13(0.8)	94.48(0.7)	94.87(0.8)	87.63(1.4)
w/o Tok-MLP (C3) †	93.47(1.0)	93.12(0.9)	93.56(1.1)	85.21(1.7)
Method	Diagnosis		Transition Pattern	
	Acc($\pm$ std)	F1($\pm$ std)	Acc($\pm$ std)	F1($\pm$ std)
<i>w/o Clinical Prior</i>				
M³AD-C9 †	92.45(1.2)	92.42(1.1)	91.34(1.3)	71.28(2.2)
M³AD-C3 †	93.21(1.0)	92.72(0.9)	92.56(1.1)	83.47(1.8)
<i>w/ Single Clinical Prior (M³AD-C3)</i>				
w/ Age †	94.21(0.9)	93.87(0.8)	93.78(0.9)	85.34(1.5)
w/ Sex †	93.56(1.0)	93.14(0.9)	93.12(1.0)	84.67(1.6)
w/ eTIV †	94.45(0.8)	94.02(0.8)	94.01(0.9)	85.89(1.4)
<i>Different Fusion Stages (M³AD-C3)</i>				
Stage 0 †	93.87(1.1)	93.28(1.0)	93.34(1.1)	84.56(1.7)
Stage 1 †	94.23(0.9)	93.82(0.8)	94.12(0.9)	85.78(1.5)
Stage 3 †	94.01(1.0)	93.60(0.9)	93.67(1.0)	85.12(1.6)

shows discriminative regions for the MCI Stability subgroup concentrated in the frontoparietal network (FPN) and somatomotor network (SMN). Fig. 3(f) shows discriminative regions for the MCI to AD progression subgroup involving the default mode network (DMN), limbic network (LMN), and FPN.

4 CONCLUSION

In this work, we present M³AD, a unified multi-task multi-gate mixture of experts framework that jointly optimizes diagnosis stage classification {NC, MCI, AD} and stage transition prediction from structural MRI, integrating clinical priors via adaptive attention fusion. M³AD achieves 95.13% diagnosis accuracy and 94.87% transition prediction accuracy on 12,037 scans, while its dual-gate routing mechanism reveals interpretable expert activation signatures for individualized progression risk estimation. Future work will extend cross-cohort validation and leverage routing patterns to identify structural markers of early stage transitions.

ACKNOWLEDGEMENTS

This work was fully supported by a Collaborative Research Fund (Project No. C5082-25E), a General Research Fund (Project No. 15101422) from the Research Grants Council of the Hong Kong SAR, P. R. China, an internal grant (P0051278) of The Hong Kong Polytechnic University, and the Theme-based Research Scheme (Project No. T45-401/22-N) from the Research Grants Council of the Hong Kong SAR, P. R. China. Zongxi Li has been supported by Lingnan University through Lam Woo Research Fund (LWP20040), Direct Grant (DR26E9), and Interdisciplinary and Strategic Research Grant (ISRG252601).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alorf, A., Khan, M.U.G.: Multi-label classification of alzheimer’s disease stages from resting-state fmri-based correlation connectivity data and deep learning. *Computers in Biology and Medicine* **151**, 106240 (2022)
2. Cardoso, M.J., Li, W., et al.: Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701* (2022)
3. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: *International conference on machine learning*. pp. 794–803. PMLR (2018)
4. Chou, Y.h., Sundman, M., et al.: Cortical excitability and plasticity in alzheimer’s disease and mild cognitive impairment: A systematic review and meta-analysis of transcranial magnetic stimulation studies. *Ageing research reviews* **79**, 101660 (2022)

5. Ding, R., Lu, H., Liu, M.: Denseformer-moe: A dense transformer foundation model with mixture of experts for multi-task brain image analysis. *IEEE TMI* (2025)
6. Dosovitskiy, A., Beyer, L., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929* (2020)
7. Fonov, V., Evans, A.C., et al.: Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage* **54**(1), 313–327 (2011)
8. Hamaide, J., De Groof, G., Van der Linden, A.: Neuroplasticity and mri: a perfect match. *NeuroImage* **131**, 13–28 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *CVPR* (2016). <https://doi.org/10.1109/cvpr.2016.90>
10. Huang, G., Liu, Z., et al.: Densely connected convolutional networks. In: *Proceedings of the IEEE conference on CVPR*. pp. 4700–4708 (2017)
11. Isensee, F., Schell, M., et al.: Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping* **40**(17), 4952–4964 (2019)
12. Jiang, Y., Shen, Y.: M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts . In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. vol. LNCS 15012. Springer Nature Switzerland (October 2024)
13. Li, J., Wei, Y., Wang, C., Hu, Q., Liu, Y., Xu, L.: 3-d cnn-based multichannel contrastive learning for alzheimer’s disease automatic diagnosis. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–11 (2022)
14. Li, J., Zhang, Y., Shu, W., Feng, X., Wang, Y., Yan, P., Li, X., Sha, C., He, M.: M4: Multi-proxy multi-gate mixture of experts network for multiple instance learning in histopathology image analysis. *Medical Image Analysis* **103**, 103561 (2025)
15. Liu, F., Yuan, S., Li, W., Xu, Q., Sheng, B.: Patch-based deep multi-modal learning framework for alzheimer’s disease diagnosis using multi-view neuroimaging. *Biomedical Signal Processing and Control* **80**, 104400 (2023)
16. Liu, Z., Hu, H., et al.: Swin transformer v2: Scaling up capacity and resolution. In: *Proceedings of the IEEE/CVF conference on CVPR*. pp. 11999–12009 (2022)
17. Ma, J., Zhao, Z., et al.: Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. pp. 1930–1939 (2018)
18. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience* **22**(12), 2677–2684 (2010)
19. Marcus, D.S., et al.: Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience* **19**(9), 1498–1507 (2007)
20. Pan, X., Initiative, A.D.N., et al.: Multiscale spatial gradient features for 18f-fdg pet image-guided diagnosis of alzheimer’s disease. *Computer Methods and Programs in Biomedicine* **180**, 105027 (2019)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision* **128**(2), 336–359 (2020)
22. Tahami Monfared, A.A., et al.: Alzheimer’s disease: epidemiology and clinical progression. *Neurology and therapy* **11**(2), 553–569 (2022)
23. Tobe, R.H., MacKay-Brandt, A., et al.: A longitudinal resource for studying connectome development and its psychiatric associations during childhood. *Scientific data* **9**(1), 300 (2022)

24. Upadhyay, P., Tomar, P., Yadav, S.P.: Comprehensive systematic computation on alzheimer's disease classification. *Archives of Computational Methods in Engineering* **31**(8), 4773–4804 (2024)
25. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: *International conference on MICCAI*. pp. 23–33. Springer (2022)
26. Venugopalan, J., Tong, L., et al.: Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports* **11**(1), 3254 (2021)
27. Xie, Z., Zhang, Z., et al.: Simmim: A simple framework for masked image modeling. In: *Proceedings of the IEEE/CVF conference on CVPR*. pp. 9653–9663 (2022)
28. Zhang, J., He, X., et al.: Multi-modal cross-attention network for alzheimer's disease diagnosis with multi-modality data. *Computers in biology and medicine* **162**, 107050 (2023)
29. Zhao, L., Zhang, J., et al.: Multimodal contrastive learning with neuroimaging and cognitive tests for alzheimer's disease diagnosis. In: *2024 IEEE International Conference on BIBM*. pp. 2971–2976. IEEE (2024)