Resolving the CN-MCI Boundary: A ResoNet Framework for Dynamic Multimodal Alzheimer's Classification

Karen Kuo[*1] d1229006@cgu.edu.tw
Wei-Yu Chen[*1] b1128020@cgu.edu.tw
Jenhui Chen[†1,2] jhchen@mail.cgu.edu.tw
[1] *Department of Computer Science and Information Engineering, Chang Gung University, Guishan Dist., Taoyuan 33302, Taiwan*

[2] *Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Guishan Dist., Taoyuan 33375, Taiwan*

Editors: Under Review for MIDL 2026

Abstract

Distinguishing mild cognitive impairment (MCI) from cognitively normal (CN) subjects remains a critical challenge, as most frameworks fail by treating tabular data as static covariates rather than dynamic modulators of MRI features, and by ignoring the ambiguous CN-MCI decision boundary. We propose ResoNet (Resolution Network), a novel multimodal framework that tackles these gaps. Critically, ResoNet achieves 93.55% accuracy on the difficult CN vs. MCI sub-task and 92.75% overall three-class (CN/MCI/AD) accuracy on the ADNI dataset. This effectiveness is driven by four core contributions:(1) A Cross-modal Feature-wise Linear Modulation (FiLM) Gating mechanism, where patient-specific tabular data generates parameters (γ, β) to dynamically modulate 3D MRI feature maps, creating personalized representations.(2) A joint boundary-aware objective combining a Supervised Contrastive (SupCon) Loss.(3) A dedicated Auxiliary CN-MCI Binary Head, which forces the model to learn finer discriminative features for the most ambiguous boundary.(4) A Class-aware Multimodal MixUp augmentation strategy that selectively increases mixing probability for MCI-involved pairs, further regularizing the boundary. ResoNet significantly improves diagnostic accuracy and model robustness, making it a powerful tool for resolving early-stage AD diagnostic ambiguity.

Keywords: Alzheimer's Disease, Mild Cognitive Impairment (MCI), Neuroimaging, Multimodal Fusion, Feature-wise Linear Modulation (FiLM), Boundary-Aware Learning, Deep Learning, Classification, 3D MRI, Supervised Contrastive Learning, ResoNet.

1. Introduction

A primary computational hurdle in neuroimaging is the accurate differentiation of early-stage Alzheimer's disease (AD) from mild cognitive impairment (MCI). This challenge is exacerbated by the subtle heterogeneity within the MCI cohort, which often blurs the clinical boundary with cognitively normal (CN) subjects, leading to unstable model predictions and reduced clinical reliability (Martinez-Murcia et al., 2020; Jo et al., 2019). While deep models struggle with this ambiguity, they particularly fail to capture the complex, patient-specific interactions between structural neuroimaging features and crucial cognitive indicators.

[*] Contributed equally
[†] Corresponding author

Structural MRI remains a standard for detecting AD-associated neuroatrophy (Jack et al., 2012), and its diagnostic power is substantially enhanced when fused with cognitive and clinical scores (Qiu et al., 2022). Consequently, deep learning has largely replaced hand-engineered features (Litjens et al., 2017), with Convolutional Neural Networks (CNNs) achieving strong baseline performance in AD classification (Islam et al., 2023; Elnaghi and Eltariny, 2024). However, traditional multimodal approaches often rely on naive feature concatenation (Zeng et al., 2019), which fails to model the nonlinear dependencies between modalities. This failure stems from the fundamental challenge of fusing heterogeneous data structures. Structural 3D MRI volumes are high-dimensional, continuous, and spatially rich, whereas clinical indicators (like MMSE or CDR-SB) are low-dimensional, numerical, and comprise attributes with diverse types and scales. Naive concatenation struggles to bridge this vast difference in dimensionality and data properties, often allowing the high-dimensional image features to overwhelm the sparse, yet critical, tabular data. (Radford et al., 2021) While recent Vision Transformer (ViT) frameworks (Shin et al., 2023), such as ADVIT (Xin et al., 2022) and SMIL-DeiT (Yin et al., 2022), offer alternatives by capturing long-range dependencies, a more fundamental challenge persists: most models are not explicitly trained to optimize the ambiguous decision boundary between CN and MCI, where diagnostic uncertainty is highest (Martinez-Murcia et al., 2020; Jo et al., 2019).

Two critical gaps, therefore, remain unresolved. First, existing multimodal fusion frameworks largely treat clinical indicators as static covariates rather than dynamic modulators of neural representations, limiting the personalization of learned features. Second, most architectures do not explicitly optimize for the ambiguous CN–MCI boundary, where diagnostic uncertainty is concentrated and model confidence is weakest (Martinez-Murcia et al., 2020; Jo et al., 2019). Bridging these gaps requires a framework capable of (i) context-aware cross-modal adaptation, (ii) boundary-sensitive learning objectives, and (iii) robust regularization for small clinical datasets.

To address these limitations, we propose ResoNet (Resolution Network), a boundary-aware multimodal deep learning framework explicitly designed to resolve diagnostic ambiguity for robust CN/MCI/AD classification. Our contributions are fourfold:

Cross-modal Feature-wise Linear Modulation (FiLM) Gating: A FiLM-based mechanism leverages patient-specific tabular data (e.g., MMSE) to adaptively modulate 3D MRI feature maps, ensuring that image representations are dynamically conditioned by individual cognitive states.

Joint Focal + Supervised Contrastive (SupCon) Loss: A composite loss enhances inter-class separation and mitigates hard-sample ambiguity. Focal Loss emphasizes challenging examples, while SupCon fosters discriminative feature clustering, improving CN–MCI separability.

Auxiliary CN–MCI Binary Head: A lightweight auxiliary classifier explicitly constrains the hardest boundary (CN–MCI), enabling sharper decision regions and reducing diagnostic uncertainty in the prodromal phase.

Class-aware MixUp Augmentation: A cross-domain MixUp scheme jointly interpolates MRI and tabular data, regularizing the multimodal manifold and improving generalization under limited data.

Collectively, these innovations enable boundary-aware multimodal reasoning, where patient-specific cues directly inform visual representations. Evaluated on the ADNI dataset,

ResoNet achieves 92.75% three-class accuracy, including 93.55% (CN vs. MCI), 96.67% (MCI vs. AD), and 100.00% (CN vs. AD)—surpassing prior multimodal fusion and manifold-learning baselines. These results demonstrate that personalized modulation and boundary optimization jointly enhance model robustness for early-stage AD diagnosis.

2. Related work

2.1. Multimodal Learning in Alzheimer's Disease and the Fusion Challenge

Multimodal deep learning approaches, integrating imaging data (e.g., MRI) with non-imaging clinical indicators, have become pivotal in Alzheimer's disease (AD) and mild cognitive impairment (MCI) research. The consensus is that multimodal fusion provides a significant diagnostic advantage over unimodal models for early-stage AD discrimination. For instance, frameworks that combine neuroimaging, clinical information, and functional assessments have achieved diagnostic accuracy comparable to that of practicing neurologists (Qiu et al., 2022). Recent work has also explored multimodal attention-based architectures for MCI/AD detection and models predicting cognitive decline using structural MRI and tabular data (Wang et al., 2024).

However, traditional multimodal fusion strategies often rely on naive feature concatenation (Zeng et al., 2019), which is ill-equipped to model the complex, non-linear dependencies between heterogeneous data structures. A more fundamental challenge is that existing frameworks largely treat these clinical indicators as static covariates rather than dynamic modulators of neural representations, limiting the personalization of learned features.

2.2. Conditional Modulation and Boundary-Aware Learning

To overcome the limitations of static fusion, conditional modulation strategies show significant promise for creating patient-specific image representations. Feature-wise Linear Modulation (FiLM), introduced by (Perez et al., 2017), serves as a foundational conditioning layer. This method allows auxiliary data (e.g., tabular features) to dynamically modulate the image feature extraction pipeline. Although FiLM is not yet widely adopted in AD diagnostics, related hypernetwork applications fusing clinical data and imaging confirm the value of conditioning image processing on tabular values (Duenias et al., 2025).

Furthermore, the high diagnostic uncertainty and ambiguous clinical boundary between the cognitively normal (CN) and MCI boundary requires explicit optimization (Martinez-Murcia et al., 2020; Jo et al., 2019). Contrastive Learning has emerged as a robust technique for acquiring discriminative feature representations, particularly by aligning and disentangling representations in heterogeneous clinical data contexts (Xu et al., 2025). Specifically, Supervised Contrastive Loss ($\mathcal{L}_{SupCon}$) (Khosla et al., 2020) is critical for separating ambiguous boundary regions. Yet, few existing works explicitly integrate both contrastive and classification objectives within a unified framework designed to target the ambiguous CN vs. MCI decision boundary.

2.3. Gaps and Positioning of This Work

In summary, three key gaps in current multimodal AD research remain: (1) The reliance on static feature fusion rather than dynamic, patient-specific modulation; (2) The lack of archi-

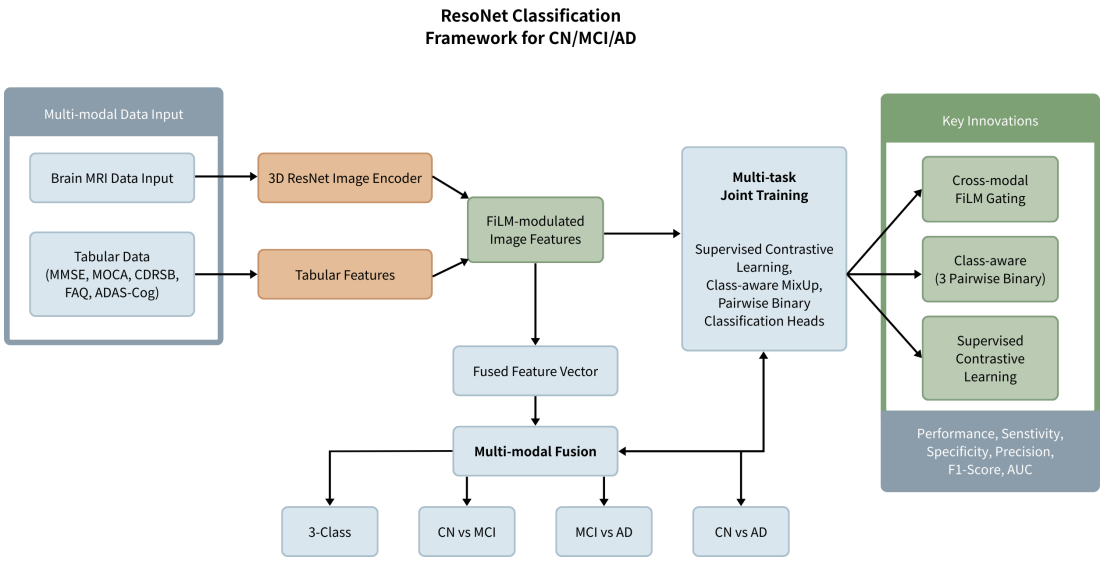

Figure 1: Overview of the ResoNet Framework. The framework features a dual-branch encoder with Cross-modal FiLM Gating to dynamically modulate MRI features using tabular embeddings. The model is optimized via a multi-task objective including SupCon loss, an auxiliary CN-MCI head, and Class-aware MixUp.

tectures and objectives explicitly trained to optimize the CN-MCI boundary ; and (3) The absence of a unified framework that combines conditional modulation and joint boundary-aware objectives for robust three-class discrimination. Our proposed ResoNet framework addresses these limitations by introducing FiLM-based cross-modal feature modulation and a joint boundary-aware objective featuring an auxiliary CN-MCI head and Supervised Contrastive Loss.

3. Method

Our framework, ResoNet, is an end-to-end multimodal deep learning architecture designed for robust CN/MCI/AD classification. The design is guided by two core principles: (i) achieving personalized modulation of imaging features by clinical scores via FiLM gating, and (ii) explicitly optimizing the CN-MCI boundary through a multi-task joint learning strategy. A detailed overview of the complete ResoNet architecture is presented in Figure 1.

3.1. Model Architecture: Multimodal Fusion with FiLM

ResoNet employs a dual-branch structure: a 3D Image Backbone, a Tabular Encoder, and a FiLM Module for dynamic conditioning. The resulting fused features are then processed by three independent output heads for disentangled multi-task supervision.

Image and Tabular Encoders. The image backbone, $\mathcal{M}_{\text{img}}$, utilizes a MONAI 3D ResNet-50 outputting $C_{\text{img}} = 512$ features. For an input 3D MRI volume $x \in \mathbb{R}^{1\times 80\times 80\times 80}$,

it outputs a global feature vector $h \in \mathbb{R}^{C_{\text{img}}}$. Concurrently, a two-layer MLP, $\mathcal{E}_{\text{tab}}$ (Tabular Encoder), maps the 5-dimensional clinical feature vector $t \in \mathbb{R}^5$ (comprising MMSE, MOCA, CDR-SB, FAQ and ADAS) into a low-dimensional embedding $t_{\text{emb}} \in \mathbb{R}^{D_{\text{tab}}}$ (where $D_{\text{tab}} = 16$).

To be precise, the $\mathcal{E}_{tab}$ is a two-layer MLP. The architecture consists of a hidden layer with 32 units followed by a ReLU activation, mapping the 5-D input vector t to the $D_{tab} = 16$ embedding t_{emb} (i.e., Linear(5) $\rightarrow$ ReLU $\rightarrow$ Linear(16)). No batch normalization or dropout was applied to this sub-network.

Cross-modal Feature-wise Linear Modulation (FiLM). To enable dynamic conditioning, the FiLM module, $\mathcal{F}_{\text{FiLM}}$, is conditioned on the tabular embedding t_{emb} and outputs two vectors, γ (scaling) and β (shifting), both matching the dimension C_{img} of the image feature h

$$(\gamma, \beta) = \mathcal{F}_{\text{FiLM}}(t_{\text{emb}}). \tag{1}$$

The FiLM module ($\mathcal{F}_{FiLM}$) is implemented as a single, lightweight linear layer. It takes the 16-D embedding t_{emb} as input and directly projects it to a 1024-dimensional vector. This vector is then deterministically split into the γ (scaling) and β (shifting) parameters, each of dimension $C_{img} = 512$.

The FiLM gating mechanism then applies a channel-wise affine modulation to the image feature h

$$\tilde{h} = (1 + \gamma) \odot h + \beta \tag{2}$$

where $\odot$ denotes the element-wise product. This allows the model to dynamically amplify or suppress MRI channels based on individual cognitive scores.

Feature Fusion and Multi-Head Output. The modulated image feature $\tilde{h}$ is concatenated with the tabular embedding t_{emb} to form the final fused feature vector $f = [\tilde{h}; t_{\text{emb}}]$. This vector f is then passed to three independent output heads: (1) a Main Classification Head ($\mathcal{H}_{\text{main}}$) outputting $K = 3$ class logits (CN/MCI/AD); (2) an Auxiliary CN–MCI Head ($\mathcal{H}_{\text{aux}}$) acting as a boundary regularizer for the CN vs. MCI distinction; and (3) a Supervised Contrastive Head ($\mathcal{H}_{\text{SupCon}}$) outputting an L2-normalized 128-D projection vector z. To define the output heads, we first denote the input fused feature vector $f = [\tilde{h}; t_{emb}]$ which has a dimension of 528 (i.e., $512 + 16$).

Both the Main Classification Head ($\mathcal{H}_{main}$) and the Auxiliary CN-MCI Head ($\mathcal{H}_{aux}$) are simple linear classifiers, implemented as Linear(528, 3) and Linear(528, 2), respectively.

The Supervised Contrastive Head ($\mathcal{H}_{SupCon}$) follows the standard design for contrastive learning, employing a 2-layer MLP projection head. It maps the fused vector f to the 128-D L2-normalized projection z via a [Linear(528) $\rightarrow$ ReLU $\rightarrow$ Linear(128)] architecture.

We highlight two key design choices within this architecture.

First, we adopt the residual formulation $\tilde{h} = (1+\gamma) \odot h + \beta$ for the FiLM operation. This specific form, as opposed to $h = \gamma h + \beta$, allows the network to learn an identity mapping by default if γ and β are initialized to zero. This stabilization technique ensures that the model can first learn robust unimodal image features while progressively incorporating the learned modulatory effect from the tabular data.

Second, the tabular embedding t_{emb} is utilized in two capacities. It is first used by the FiLM module to modulate the image features, resulting in $\tilde{h}$. It is then explicitly concatenated with $\tilde{h}$ to form the final fused vector f. This dual-use design ensures that the final

classification heads have direct, unadulterated access to both the original clinical/cognitive scores (via t_{emb}) and their contextual effect on the image representations (via $\tilde{h}$), preventing the tabular information from being lost or "washed out" during the modulation process.

3.2. Boundary-Aware Learning Objectives

We optimize a joint loss $\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda_{SupCon}\mathcal{L}_{SupCon} + \lambda_{aux}\mathcal{L}_{aux}$, with empirically set weights $\lambda_{SupCon} = 0.2$ and $\lambda_{aux} = 0.5$.

Main Classification Loss ($\mathcal{L}_{main}$). We employ a Class-balanced Focal Loss to address imbalance. With a focusing parameter $\gamma = 2.0$ and a class balancing factor α_t (boosted by 1.5 for MCI), the loss is:

$$\mathcal{L}_{Focal}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{3}$$

During Class-aware MixUp, this is replaced by a soft-target Cross-Entropy loss.

Supervised Contrastive Loss ($\mathcal{L}_{SupCon}$). Following (Khosla et al., 2020), we apply $\mathcal{L}_{SupCon}$ to the L_2-normalized projection z with a temperature $\tau = 0.07$ to enforce intra-class compactness:

$$\mathcal{L}_{SupCon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i) \setminus \{i\}} \exp(z_i \cdot z_a/\tau)} \tag{4}$$

Auxiliary CN-MCI Binary Head ($\mathcal{L}_{aux}$). We utilize Binary Cross-Entropy (BCE) masked to exclude AD samples, strictly optimizing the CN-MCI boundary (N_{sub} is the count of non-AD samples):

$$\mathcal{L}_{aux} = -\frac{1}{N_{sub}} \sum_{j=1}^{N} \mathbb{1}_{\{y_j \neq AD\}} \cdot [y_j \log(\hat{y}_{aux,j}) + (1 - y_j) \log(1 - \hat{y}_{aux,j})] \tag{5}$$

3.3. Class-aware Multimodal MixUp Augmentation

To densify the manifold around the sparse MCI class, we introduce a biased sampling strategy for MixUp interpolation ($\lambda \sim Beta(0.4, 0.4)$). The mixing probability is conditioned on the pair composition: MCI-involved Pairs: $P_{MCI} = 0.5$ (if pair includes MCI). Other Pairs: $P_{other} = 0.3$ (for CN-CN, AD-AD, CN-AD). This bias creates more virtual samples in the transition zones (CN $\leftrightarrow$ MCI), smoothing the decision boundary where uncertainty is highest.

4. Experiments

4.1. Experimental Setup

We evaluate our framework on the public Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, using T1-weighted structural MRI scans and 5 clinical/tabular features (MMSE, MOCA, CDR-SB, FAQ, ADAS). Our final dataset consists of N=920 participants, including 521 CN, 308 MCI, and 91 AD. Standard preprocessing (N4, skull-stripping, registration) was applied.

Data Handling & Augmentation. The dataset is divided using a 70/15/15 stratified split (Training/Validation/Testing). All preprocessing steps were applied within each split to avoid information leakage. To address the class imbalance (Subsection 4.1), we use a WeightedRandomSampler for the training set, with weights boosted for the MCI class by a factor of MCI_BOOST=1.5. Training-time augmentation includes random 3D flipping (p=0.5) and Gaussian noise (p=0.2). For inference, we employ Test-Time Augmentation (TTA), averaging logits from the original and three flipped volumes.

Training & Evaluation. The model is implemented in PyTorch/MONAI and trained for 60 epochs using the AdamW optimizer ($lr = 2e-4$, $wd = 1e-4$) with a Cosine Annealing scheduler. We utilize AMP (mixed precision), Gradient Clipping ($\ell_2 \leq 2.0$), and Gradient Accumulation (ACCUM_STEPS = 2) with a BATCH_SIZE = 4 for an effective batch size of 8. We report 3-class Accuracy and Macro-F1. For binary sub-tasks (CN vs AD, MCI vs AD, CN vs MCI), we conduct detailed analysis using ROC-AUC, AUPRC, Sensitivity, Specificity, and MCC. The best model weights are saved based on validation accuracy.

Table 1: Performance on Binary Classification Sub-tasks. We report the detailed performance of our full ResoNet model on the three binary sub-tasks. The model achieves perfect accuracy on CN vs AD and demonstrates exceptionally high accuracy (93.55%) on the difficult CN vs MCI task, validating its effectiveness in resolving the ambiguous boundary.

Task	**Acc.**	**Prec.**	**Recall**	**F1**
CN vs AD	1.0000	1.0000	1.0000	1.0000
MCI vs AD	0.9667	0.8750	1.0000	0.9333
CN vs MCI	0.9355	0.9318	0.8913	0.9111
3-Class (CN/MCI/AD)	0.9275	0.9229	0.9229	0.9229

In addition to the 3-class task, we conducted a detailed analysis of the model's performance on the binary sub-tasks, as detailed in Table 1. Our framework achieves perfect separation (1.0000 Acc.) for the CN vs AD task. More importantly, it achieves 93.55% accuracy and 0.9231 F1-score on the highly challenging CN vs MCI sub-task, directly supporting our claim of resolving the ambiguous CN-MCI boundary.

To validate the contribution of each proposed component, we conducted a detailed ablation study. The results, presented in Table 2, show the stepwise improvement as each innovation is integrated. The baseline model, using simple feature concatenation, achieves 81.16% 3-class accuracy. The introduction of our Cross-modal FiLM Gating (+FiLM) provides the most significant boost, increasing accuracy to 88.41%. Subsequent additions of Supervised Contrastive loss (+SupCon) and the Auxiliary CN-MCI Head (+Aux) further refine the decision boundaries, improving CN vs MCI accuracy to 91.13%. Our full model, incorporating Class-aware MixUp, achieves the state-of-the-art performance of 92.75% (3-class) and 93.55% (CN vs MCI).

To further analyze the per-class performance, Figure 2 shows the confusion matrix for the 3-class task. The matrix confirms the model's robustness. For instance, out of 46 true MCI samples, 41 are correctly identified, while only 4 are misclassified as CN, demonstrating a strong ability to correctly identify the MCI cohort.

Table 2: Ablation Study of ResoNet Components. We evaluate the incremental contribution of each key component. Starting from a "Baseline(Concat)" model, we progressively add FiLM gating (+FiLM), Supervised Contrastive Loss (+SupCon), the Auxiliary CN–MCI Head (+Aux Head), and Class-aware MixUp (+MixUp). The full model demonstrates the best performance on both the 3-class task (92.75%) and the critical CN vs MCI task (93.55%). Bold indicates the best performance.

Model	FiLM	SupCon	Aux Head	MixUp	3-class	CN vs MCI	CN vs AD	MCI vs AD
Baseline(Concat)					0.8116	0.8145	1.0000	0.9333
+FiLM	✓				0.8841	0.8871	1.0000	0.9667
+FiLM+SupCon	✓	✓			0.8986	0.8952	1.0000	0.9667
+FiLM+SupCon+Aux	✓	✓	✓		0.9058	0.9113	1.0000	0.9333
Our(Full Model)	✓	✓	✓	✓	**0.9275**	**0.9355**	**1.0000**	**0.9667**

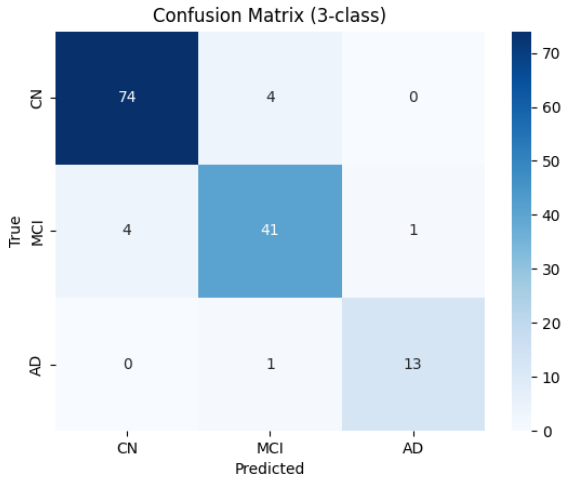

Figure 2: Confusion Matrix for 3-Class (CN/MCI/AD) Classification. The confusion matrix for our full ResoNet model on the test set. The model shows strong predictive accuracy across all classes, with very few misclassifications between the CN and AD groups.

4.2. Results and Ablation Study

The experimental results presented in this section provide strong, quantitative validation for our proposed ResoNet framework. Our central hypothesis—that dynamic, patient-specific feature modulation combined with explicit boundary-aware optimization is critical for resolving diagnostic ambiguity—is robustly supported by the data.

The ablation study (Table 2) clearly demonstrates the cumulative value of our contributions. While the baseline concatenation model achieves only 81.16% 3-class accuracy,

the introduction of our Cross-modal FiLM Gating mechanism provides the most significant performance leap to 88.41%, confirming the superiority of dynamic modulation over static fusion. Subsequent additions of the Supervised Contrastive loss, the dedicated Auxiliary CN-MCI Head, and Class-aware MixUp each progressively enhance the model's discriminative power, leading to our final 3-class accuracy of 92.75%.

Crucially, our framework excels at the most challenging clinical sub-task: the CN vs. MCI distinction. As shown in Table 1 the model achieves 93.55% accuracy on this specific pair, a significant improvement that directly validates the effectiveness of our boundary-aware learning objectives (i.e., the SupCon loss and the auxiliary head). This, combined with the perfect 1.0000 accuracy on the CN vs. AD task and the detailed breakdown in the confusion matrix (Figure 2), confirms that ResoNet not only achieves high overall accuracy but also successfully sharpens the decision boundaries for the most ambiguous and clinically relevant cohorts.

4.3. Ethical and Data Disclosure

This study utilized exclusively data from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI data collection was previously approved by the Institutional Review Boards (IRBs) of all participating institutions , and all participants provided written informed consent. All data were fully anonymized before analysis in accordance with the ethical guidelines outlined in the Declaration of Helsinki.

5. Conclusion and Future Directions

We introduced ResoNet, a boundary-aware multimodal deep learning framework for early Alzheimer's disease diagnosis. Our work overcomes the fundamental limitations of conventional methods, which are hampered by static feature fusion and a failure to address ambiguous clinical boundaries. We tackle these challenges through two core innovations: first, a cross-modal conditioning mechanism based on Feature-wise Linear Modulation (FiLM) to dynamically personalize 3D MRI features conditioned on individual patient scores ; and second, a joint multi-task learning objective featuring a dedicated auxiliary head and supervised contrastive loss, acting as a powerful boundary regularizer.

Our experimental results on the ADNI dataset validate our central hypothesis. We not only achieve state-of-the-art 92.75% three-class accuracy but, critically, demonstrate a 93.55% accuracy specifically on the CN vs. MCI classification task. This provides direct evidence that the synergy of personalized feature adaptation and explicit boundary optimization is crucial for resolving diagnostic ambiguity and enhancing model robustness. This research pushes the paradigm of multimodal fusion in medical image analysis from 'static concatenation' toward 'dynamic, context-aware reasoning'.

While our framework demonstrates strong performance, future work will address existing limitations by focusing on model generalizability and predictive depth. We will prioritize external validation on independent clinical datasets (e.g., AIBL or OASIS) to confirm robustness across different acquisition protocols. The framework will be extended to longitudinal data to model disease progression and predict MCI-to-AD conversion risk.

References

D. Duenias, B. Nichyporuk, T. Arbel, and T. R. Raviv. HyperFusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. *Medical Image Analysis*, 2025.

L. M. Elnaghi and Y. M. Eltariny. Evaluation of deep learning models on Alzheimer's MRI dataset: AD-VGG16, AD-Resnet50, and AD-2DCNN. In *Proceedings of the 6th International Conference on Computing and Informatics (ICCI)*, pages 237–242, 2024.

F. Islam, M. H. Rahman, M. S. Hossain, and S. Ahmed. A novel method for diagnosing Alzheimer's disease from MRI using ResNet50 and SVM. *International Journal of Advanced Computer Science and Applications*, 14(6), 2023.

C. R. Jack et al. Magnetic resonance imaging in Alzheimer's disease. *Neuroimaging Clinics of North America*, 22(1):75–88, 2012.

T. Jo, K. Nho, and A. J. Saykin. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in Aging Neuroscience*, 11:220, 2019. doi: 10.3389/fnagi.2019.00220.

P. Khosla et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020.

G. Litjens et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

F. J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, and D. Castillo-Barnes. Studying the manifold structure of Alzheimer's disease: A deep learning approach using convolutional autoencoders. *IEEE Journal of Biomedical and Health Informatics*, 24(1):17–26, 2020. doi: 10.1109/JBHI.2019.2914970.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2017. Also available as arXiv:1709.07871.

S. Qiu, M. I. Miller, P. S. Joshi, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications*, 13:3404, 2022. doi: 10.1038/s41467-022-31037-5.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

H. Shin, S. Jeon, Y. Seol, S. Kim, and D. Kang. Vision transformer approach for classification of Alzheimer's disease using 18f-florbetaben brain images. *Applied Sciences*, 13: 3453, 2023. doi: 10.3390/app13063453.

C. Wang, S. Li, Y. Chen, et al. A multimodal deep learning approach for the prediction of cognitive decline and its effectiveness in clinical trials for Alzheimer's disease. *Translational Psychiatry*, 2024.

X. Xin et al. Advit: Vision transformer on multi-modality PET images for Alzheimer disease diagnosis. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2022.

X. Xu, J. Li, and Y. Wang. Contrastive learning in brain imaging: Methods and applications. *Computerized Medical Imaging and Graphics*, 2025.

Y. Yin, W. Jin, J. Bai, R. Liu, and H. Zhen. SMIL-DeiT: Multiple instance learning and self-supervised vision transformer network for early Alzheimer's disease classification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, Padua, Italy, 2022. doi: 10.1109/IJCNN55064.2022.9892524.

N. Zeng et al. A hybrid deep learning model for dementia diagnosis based on multimodal MRI. *Frontiers in Neuroscience*, 13:1199, 2019.