A Confounder-aware Representation Learning Framework for Alzheimer's Disease Classification

Yuanhao Chen[1], Jingwen Chen[1], Jinyi Xu[2], Haoqi Yu[1]
Bin Wang[1], Chunzhong Li[4,*], Yongquan Zhang[1,*], Fenglei Fan[5]
Ahmed Elazab[3,], Xiang Wan[2,], and Changmiao Wang[2,*]

[1] Zhejiang University of Finance and Economics
[2] Shenzhen Research Institute of Big Data
[3] Tsinghua Shenzhen International Graduate School
[4] Anhui University of Finance and Economics
[5] City University of Hong Kong
zyq@zufe.edu.cn, cmwangalbert@gmail.com

Abstract. Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder, and current treatments cannot reverse its progression, making early diagnosis and intervention essential. Deep learning methods have shown promise for diagnosing AD from neuroimaging data, but these models are often affected by both visible and hidden sources of noise. Visible noise typically reflects systematic imaging differences, such as scanner-related biases and acquisition artifacts, whereas hidden noise is linked to individual variation that can lead models to depend on spurious patterns that do not generalize well. Most existing approaches emphasize statistical associations and do not adequately account for these confounding influences. Motivated by these limitations, we propose a multimodal, confounder-aware framework for AD diagnosis inspired by causal reasoning. The model includes a region-masking component (Mask Decoupling Module) that probes the image by making small, targeted changes to localized areas, then highlights the regions that most strongly affect the prediction, encouraging the network to prioritize structurally informative features. We also introduce a covariate-guided feature modulation mechanism that uses clinical and demographic variables to adjust local image features, thereby reducing variation attributable to these factors. Although the framework does not estimate causal effects directly, it is designed to learn more stable representations by limiting reliance on shortcuts, improving robustness and generalization in AD diagnostic models. Our code is available at https://github.com/redteacode/Causal_fusion.

Keywords: Alzheimer's Disease · Causal Learning · Multimodal Fusion.

1 Introduction

As the most common neurodegenerative disorder, Alzheimer's disease (AD) currently lacks curative treatment, making early identification and timely interven-

[1] * Corresponding author: zyq@zufe.edu.cn, cmwangalbert@gmail.com

tion essential for slowing disease progression [2,5]. However, pathological alterations in the early stages are often subtle in neuroimaging and difficult to detect consistently [11]. Moreover, substantial inter-individual variability in brain structure and function further increases diagnostic uncertainty [4]. Therefore, developing diagnostic models that are sensitive to subtle disease-related signals while remaining robust to individual variability constitutes a critical clinical objective in AD research.

Deep learning approaches have recently shown strong potential for neuroimaging-based AD diagnosis. Both convolutional neural network (CNN) visual models [22] and multimodal fusion methods that combine imaging with auxiliary clinical or demographic data have reported promising results [8,3]. However, these models are often affected by multiple sources of variation that distort predictions [18,19]. Some of this variation is observable, for example, scanner-related biases and acquisition artifacts [21]. Other sources are less directly observable and reflect individual characteristics, such as age- and sex-related differences in brain structure [15]. Together, these factors can lead models to exploit non-disease cues that do not generalize well across cohorts or imaging sites, reducing robustness when data conditions change [23]. Most existing AD diagnostic methods primarily learn statistical associations between imaging features and labels and give limited attention to these confounding influences [17]. Consequently, learned representations may reflect spurious patterns rather than disease-relevant signals, weakening performance across populations and clinical settings [6].

Causal representation learning provides a principled direction for addressing this limitation by encouraging models to capture relationships more closely tied to underlying biological mechanisms. For example, Tang *et al.* [16] separated disease-relevant features for Parkinson's disease diagnosis by learning causal scores from pretrained graph convolutional network representations. Wang *et al.* [20] used adversarial objectives to suppress redundant information across modalities, improving diagnostic stability. Despite these advances, many causal methods rely on multi-stage training pipelines, which can complicate practical deployment. Related work has attempted to reduce confounding using dedicated attention-based designs [23] or causal attention modules that identify confounding signals without manual annotation [19]. Nevertheless, these approaches are largely tailored to visual inputs and do not transfer easily to other data types, These limitations underscore the need for a structurally simple, training-stable, and multimodal confounder-aware diagnostic framework.

To address these gaps, we propose a causal-inspired, confounder-aware multimodal framework for AD diagnosis. Rather than explicitly separating "causal" from "non-causal" variables, our method aims to learn representations that are both predictive and robust through structured regularization. We introduce a Mask Decoupling Module (MDM) that estimates the prediction sensitivity of local image regions via targeted perturbations, guiding the model to emphasize structurally informative areas. We also design a confounder-aware feature modulation mechanism that conditions visual representations on tabular covariates, suppressing covariate-linked variation and reducing reliance on spu-

rious cues. The main contributions are as follows: 1) We propose a prediction-sensitivity–based mask learning strategy that promotes spatially robust representations by identifying regions that strongly influence the prediction function. 2) We introduce a covariate-conditioned feature modulation module that integrates tabular information to reduce confounder-correlated variation in local representations, improving multimodal robustness for AD diagnosis.

2 Methodology

We adopt the conceptual dependency diagram shown in Figure 1. In this formulation, the tabular covariates $\boldsymbol{Z}$ act as confounders that influence both the neuroimaging input $\boldsymbol{X}$ and the diagnostic outcome $\boldsymbol{Y}$. The MRI $\boldsymbol{X}$ contains a mix of signals: some regions reliably support prediction, while others may reflect cohort-specific artifacts or patterns tied to the underlying population and therefore may not generalize across cohorts.

Guided by this causal perspective, our framework (Figure 2) integrates two components: the MDM and the confounder-aware patch modulation (CAPM) module. The model is multimodal, using a ResNet-based backbone for imaging features and a multilayer perceptron (MLP) for tabular features. MDM estimates which spatial regions influence the prediction by masking individual patches and measuring the resulting change in the model output. This encourages the network to allocate capacity to structurally informative areas rather than incidental signals. CAPM then conditions patch-level features on the covariates $\boldsymbol{Z}$ to reduce image variations that are strongly associated with these covariates. Rather than performing a full causal adjustment, CAPM serves as a robustness-oriented normalization mechanism that discourages reliance on confounder-linked shortcuts.

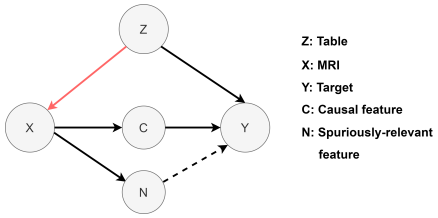


Fig. 1. Solid arrows denote hypothesized dependencies among variables; dashed arrow from $\boldsymbol{N}$ to $\boldsymbol{Y}$ represents the spurious relevant to be suppressed by MDM. The red edge $\boldsymbol{Z}$ to $\boldsymbol{X}$ indicates the confounding path learned via CAPM.

2.1 Mask Decoupling Module

In causal inference, the invariant prediction principle states that if variables that are not direct causes are perturbed while all direct causes are held fixed, the

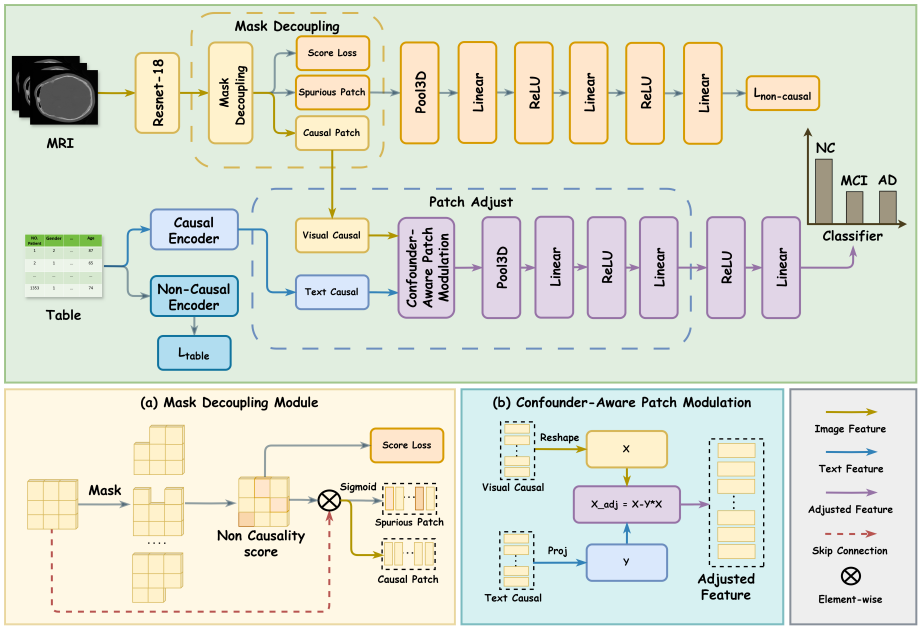


Fig. 2. The overview of our model. (a) introduces localized perturbations by masking individual patches and estimates prediction sensitivity based on the resulting changes in model output (b) Adjusting patch-level representations on tabular covariates to attenuate covariate-associated variations and enhance representation robustness.

target variable remains unchanged [13,14]. Motivated by this idea, we propose a MDM to separate patch features that drive the prediction from those that do not.

Given an input image $\boldsymbol{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$, we first partition it into non-overlapping 3D patches:

$$\boldsymbol{P} = \text{Patchify}(\boldsymbol{X}) \in \mathbb{R}^{\boldsymbol{B}\times \boldsymbol{N_p}\times \boldsymbol{C}\times \boldsymbol{D}\times \boldsymbol{H}\times \boldsymbol{W}}, \tag{1}$$

where $\boldsymbol{N_p}$ denotes the number of patches. To quantify the contribution of each patch to the classification outcome, we apply a local perturbation that removes the p-th patch:

$$\boldsymbol{X}^{(-\boldsymbol{p})} = \text{Mask}(\boldsymbol{X}, \boldsymbol{p}). \tag{2}$$

This operation changes only a localized region while keeping the rest of the image intact. Each masked image is then processed by a shared scoring network $F(\cdot)$, implemented with residual blocks and pooling layers, to produce a score:

$$\boldsymbol{s_p} = F(\boldsymbol{X}^{(-\boldsymbol{p})}). \tag{3}$$

Intuitively, $\boldsymbol{s_p}$ reflects how dependent the prediction is on patch p: if removing the patch substantially weakens class-related evidence, the score should drop. We

therefore treat $\boldsymbol{s_p}$ as a patch-wise estimate of prediction relevance. We map these scores through a sigmoid function to obtain a soft mask $\boldsymbol{M} \in [0, 1]$, and use it to decompose the original patch features into two components:

$$\boldsymbol{P}^{\text{non-causal}} = \boldsymbol{P} \odot \boldsymbol{M}, \quad \boldsymbol{P}^{\text{causal}} = \boldsymbol{P} \odot (1 - \boldsymbol{M}), \tag{4}$$

where $\odot$ denotes element-wise multiplication; $\boldsymbol{P}^{\text{non-causal}}$ captures patch features whose removal has little effect on the prediction, whereas $\boldsymbol{P}^{\text{causal}}$ retains features that strongly influence the classification result.

Because $\boldsymbol{P}^{\text{non-causal}}$ is intended to contain minimal class information, predictions based on it should be unreliable. Accordingly, we maximize its classification loss during training:

$$\boldsymbol{L}_{\text{non-causal}} = -\frac{1}{\boldsymbol{B}} \sum_{i=1}^{\boldsymbol{B}} CE\Big(f_{\text{non-causal}}(\boldsymbol{P}^{\text{non-causal}}), \boldsymbol{y_i}\Big), \tag{5}$$

where CE is the cross-entropy (CE) loss, and $f_{\text{non-causal}}$ denotes a lightweight classifier composed of pooling and linear layers.

To ease optimization and improve convergence, we further introduce a weak structural prior. Specifically, we compute the voxel-wise mean image of the training set, $\bar{\boldsymbol{X}} = \frac{1}{\boldsymbol{N}} \sum_{i=1}^{N} \boldsymbol{X}$. Features (and corresponding patch scores) extracted from $\bar{\boldsymbol{X}}$ provide guidance for mask learning via a ranking-consistency objective:

$$\boldsymbol{L}_{\text{rank}} = \text{RankLoss}(\boldsymbol{s}, \boldsymbol{s}^{\text{prior}}), \tag{6}$$

where $\boldsymbol{s}^{\text{prior}}$ denotes the prior scores. Concretely, for randomly sampled patch pairs (i, j) satisfying $\boldsymbol{s}_i^{\text{prior}} > \boldsymbol{s}_j^{\text{prior}}$, we use:

$$\mathbb{E}(i, j)[\max(0, \boldsymbol{m} - (\boldsymbol{s}_i - \boldsymbol{s}_j))], \tag{7}$$

where $\boldsymbol{m}$ denotes the margin hyperparameter.

2.2 Confounder-Aware Patch Modulation

Confounding factors can introduce spurious correlations that distort imaging feature learning. In causal inference, a common way to address this issue is backdoor adjustment, which controls for variables that create non-causal associations. Motivated by this idea, we propose a Confounder-Aware Patch Modulation (CAPM) module to attenuate confounder-related variation in patch-level representations.

Given patch features produced by the MDM, $\boldsymbol{X}^c \in \mathbb{R}^{\boldsymbol{B} \times \boldsymbol{N_p} \times \boldsymbol{C} \times \boldsymbol{d} \times \boldsymbol{h} \times \boldsymbol{w}}$, and tabular features $\boldsymbol{z} \in \mathbb{R}^{\boldsymbol{B} \times \boldsymbol{d_z}}$, we use a projection network $G(\cdot)$ (implemented with linear layers and nonlinear activations) to map the tabular input to a patch-wise adjustment vector:

$$\boldsymbol{\gamma} = G(\boldsymbol{z}) \in \mathbb{R}^{\boldsymbol{B} \times \boldsymbol{P}}. \tag{8}$$

Each entry $\boldsymbol{\gamma_p}$ represents how strongly the p-th patch is affected by confounding factors. The vector is then expanded along the channel and spatial

dimensions to align with the visual feature tensor. We modulate each patch by subtracting its confounding component from the corresponding visual features:

$$\boldsymbol{X}_p^{adj} = \boldsymbol{X}_p^c - \boldsymbol{\gamma_p} \boldsymbol{X}_p^c = (1 - \boldsymbol{\gamma_p}) \boldsymbol{X}_p^c. \tag{9}$$

This operation suppresses patches that are strongly associated with confounders while leaving weakly affected regions largely unchanged. To improve optimization stability and retain useful visual information, we incorporate a skip connection and apply normalization:

$$\boldsymbol{X}^{out} = \mathrm{Norm}(\boldsymbol{X}^{adj} + \boldsymbol{X}^c), \tag{10}$$

where Norm denotes a normalization layer.

In addition, we include a regularization term to prevent the tabular branch from dominating the overall prediction and to encourage balanced multimodal fusion. Concretely, we train the tabular-only classifier to perform poorly by maximizing its classification loss $\boldsymbol{L_{table}}$, implemented by minimizing the negative cross-entropy.

2.3 Loss Functions

Combining the loss terms from the MDM and CAPM modules, we define the overall objective function for the proposed framework as:

$$\boldsymbol{L}_{total} = \boldsymbol{L}_{CE} - \boldsymbol{\alpha} \boldsymbol{L}_{non-causal} - \boldsymbol{\beta} \boldsymbol{L}_{table} + \boldsymbol{\lambda_t} \boldsymbol{L}_{rank}, \tag{11}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are balancing coefficients, and $\boldsymbol{\lambda_t}$ follows a decay schedule. This objective encourages the model to base its predictions on the causal representation, discourages reliance on confounding signals, and improves the stability of mask estimation.

3 Experiments

3.1 Datasets and Implementation

Datasets. We conducted experiments on the ADNI [10] and NACC [1] datasets, which include 1,225 and 1,621 MRI scans, respectively. Each scan was paired with a follow-up record collected within three months, containing demographic information and clinical assessment scores. Table 1 summarizes the key dataset characteristics. ADNI provides scans that have already been preprocessed using a standardized pipeline, whereas NACC provides raw images. To harmonize the two datasets, we applied bias field correction to all scans and registered them to the Montreal Neurological Institute standard space.

Implementation Details. All experiments were implemented in PyTorch 2.11.0+cu130 and run on a workstation with an NVIDIA GeForce RTX 5090 GPU. We used the Adam optimizer with a learning rate of 1×10^{-4} and trained each model for 200 epochs. In the composite loss, the weighting coefficients α

and β were both set to 0.01, while λ_t was linearly decayed to zero within the first 100 epochs. We performed the CN/MCI/AD triple classification task on both the ADNI and the NACC datasets. Performance was evaluated using accuracy (ACC), recall (REC), F1-score (F1), area under the ROC curve (AUC), and specificity (SPE), computed with a macro-averaging strategy via TorchMetrics. We report results from five-fold cross-validation on both datasets.

Table 1. Demographic information of the publicly available datasets used for classification tasks.

DATASET	NUMBER(M/F)	AGE	NC:MCI:AD
ADNI [10]	687/538	75.15 ± 6.42	355:303:569
NACC [1]	718/902	71.67 ± 9.05	985:435:201

Table 2. Comparison of our model with other state-of-the-art models on the public ADNI and NACC dataset. I and T denote imaging and tabular modalities, respectively. Bold values denote the best-performing results.

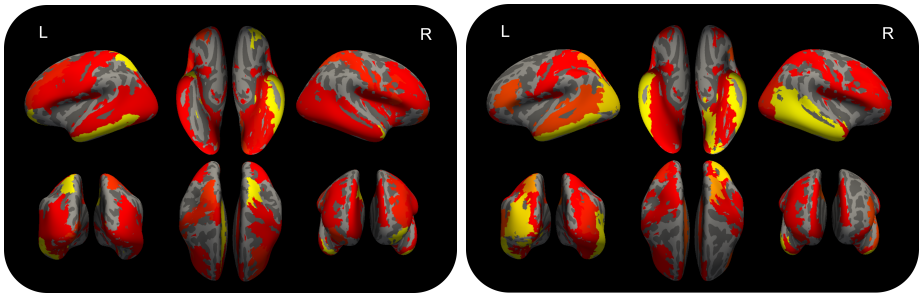
Dataset			ADNI [10]					NACC [1]				
Method	I	T	ACC(↑)	REC(↑)	F1(↑)	AUC(↑)	SPE(↑)	ACC(↑)	REC(↑)	F1(↑)	AUC(↑)	SPE(↑)
ResNet [7]	✓		$0.547_{\pm 0.009}$	$0.465_{\pm 0.012}$	$0.440_{\pm 0.015}$	$0.671_{\pm 0.020}$	$0.740_{\pm 0.007}$	$0.649_{\pm 0.025}$	$0.466_{\pm 0.041}$	$0.473_{\pm 0.047}$	$0.742_{\pm 0.023}$	$0.747_{\pm 0.017}$
TabFormer [12]		✓	$0.656_{\pm 0.020}$	$0.577_{\pm 0.019}$	$0.558_{\pm 0.033}$	$0.795_{\pm 0.010}$	$0.807_{\pm 0.011}$	$0.691_{\pm 0.036}$	$0.533_{\pm 0.080}$	$0.526_{\pm 0.092}$	$0.805_{\pm 0.077}$	$0.762_{\pm 0.033}$
Causalmixnet [23]	✓		$0.579_{\pm 0.070}$	$0.559_{\pm 0.072}$	$0.554_{\pm 0.074}$	$0.720_{\pm 0.050}$	$0.783_{\pm 0.034}$	$0.614_{\pm 0.015}$	$0.403_{\pm 0.049}$	$0.390_{\pm 0.073}$	$0.646_{\pm 0.045}$	$0.711_{\pm 0.024}$
CAAM [19]	✓		$0.526_{\pm 0.024}$	$0.448_{\pm 0.018}$	$0.422_{\pm 0.019}$	$0.653_{\pm 0.020}$	$0.731_{\pm 0.011}$	$0.640_{\pm 0.007}$	$0.456_{\pm 0.028}$	$0.459_{\pm 0.029}$	$0.725_{\pm 0.034}$	$0.749_{\pm 0.012}$
MAD-former [22]	✓		$0.479_{\pm 0.028}$	$0.389_{\pm 0.039}$	$0.342_{\pm 0.059}$	$0.593_{\pm 0.048}$	$0.697_{\pm 0.020}$	$0.602_{\pm 0.021}$	$0.375_{\pm 0.013}$	$0.347_{\pm 0.022}$	$0.658_{\pm 0.032}$	$0.704_{\pm 0.015}$
Hyperfusion [3]	✓	✓	$0.645_{\pm 0.046}$	$0.597_{\pm 0.041}$	$0.583_{\pm 0.047}$	$0.789_{\pm 0.036}$	$0.813_{\pm 0.022}$	$0.652_{\pm 0.008}$	$0.571_{\pm 0.036}$	$0.564_{\pm 0.031}$	$0.757_{\pm 0.022}$	$0.792_{\pm 0.011}$
Ma *et al* [9]	✓	✓	$0.754_{\pm 0.010}$	$0.702_{\pm 0.011}$	$0.691_{\pm 0.011}$	$0.853_{\pm 0.007}$	$0.870_{\pm 0.005}$	$0.823_{\pm 0.007}$	$0.764_{\pm 0.024}$	$0.770_{\pm 0.016}$	$0.909_{\pm 0.014}$	$0.887_{\pm 0.005}$
Ours	✓	✓	$\mathbf{0.833_{\pm 0.021}}$	$\mathbf{0.818_{\pm 0.016}}$	$\mathbf{0.818_{\pm 0.020}}$	$\mathbf{0.939_{\pm 0.009}}$	$\mathbf{0.915_{\pm 0.008}}$	$\mathbf{0.831_{\pm 0.012}}$	$\mathbf{0.783_{\pm 0.028}}$	$\mathbf{0.805_{\pm 0.047}}$	$\mathbf{0.910_{\pm 0.021}}$	$\mathbf{0.900_{\pm 0.008}}$

Table 3. Ablation experiments on the ADNI and NACC datasets. Bold values denote the best-performing results.

Dataset	ADNI [10]					NACC [1]				
Method	ACC(↑)	REC(↑)	F1(↑)	AUC(↑)	SPE(↑)	ACC(↑)	REC(↑)	F1(↑)	AUC(↑)	SPE(↑)
Wo MDM&CAPM	$0.592_{\pm 0.063}$	$0.590_{\pm 0.056}$	$0.585_{\pm 0.055}$	$0.787_{\pm 0.035}$	$0.801_{\pm 0.028}$	$0.574_{\pm 0.070}$	$0.503_{\pm 0.043}$	$0.490_{\pm 0.050}$	$0.726_{\pm 0.032}$	$0.761_{\pm 0.019}$
Wo MDM	$0.819_{\pm 0.006}$	$0.796_{\pm 0.011}$	$0.795_{\pm 0.009}$	$0.932_{\pm 0.013}$	$0.907_{\pm 0.004}$	$0.802_{\pm 0.005}$	$0.784_{\pm 0.009}$	$0.769_{\pm 0.005}$	$0.898_{\pm 0.020}$	$0.887_{\pm 0.003}$
Wo CAPM	$0.660_{\pm 0.075}$	$0.641_{\pm 0.095}$	$0.622_{\pm 0.104}$	$0.818_{\pm 0.083}$	$0.823_{\pm 0.047}$	$0.617_{\pm 0.014}$	$0.500_{\pm 0.041}$	$0.506_{\pm 0.040}$	$0.732_{\pm 0.032}$	$0.771_{\pm 0.013}$
Ours	$\mathbf{0.833_{\pm 0.021}}$	$\mathbf{0.818_{\pm 0.016}}$	$\mathbf{0.818_{\pm 0.020}}$	$\mathbf{0.939_{\pm 0.009}}$	$\mathbf{0.915_{\pm 0.008}}$	$\mathbf{0.831_{\pm 0.012}}$	$\mathbf{0.783_{\pm 0.028}}$	$\mathbf{0.805_{\pm 0.047}}$	$\mathbf{0.910_{\pm 0.021}}$	$\mathbf{0.900_{\pm 0.008}}$

3.2 Experiments Results

Comparative experiments. We compared our method against a range of baselines, including established architectures (ResNet [7] and TabFormer [12]);



(a) (b)

Fig. 3. Visualization of important patches identified by ours model, where darker shades indicate greater weights, with red appearing darker than yellow. Subfigure (a) is derived from the ADNI dataset, while (b) originates from the NACC dataset.

methods that learn representations from a single data source (CAAM [19], CausalMixNet [23], and MAD-Former [22]); and multimodal fusion approaches (Hyperfusion [3] and the method by Ma et al. [9]).

Table 2 reports the performance of our model against existing methods. All experiments use 5-fold cross-validation. Among approaches that use a single data type, ResNet and TabFormer achieve the strongest overall results. By comparison, several feature-learning baselines appear to rely on restrictive assumptions, which can yield high precision but lower overall accuracy. Methods that combine multiple data types, including HyperFusion and the approach of Ma et al., show clear advantages. Our model further improves performance by integrating information across data types, with particularly strong gains on the ADNI dataset: +10.4% (ACC), +16.5% (REC), +18.3% (F1), +10% (AUC), and +5% (SPE). It also achieves state-of-the-art results on the NACC dataset.

Ablation Studies. To identify the sources of performance gains, we conducted ablation studies using a baseline model composed solely of ResNet and patch embedding. Each module and its loss were added separately to assess individual contributions. As reported in Table 3, incorporating either module leads to consistent performance improvements, while the integration of both modules yields the best overall results, suggesting that they contribute complementary advantages.

To examine whether our model learns clinically meaningful patterns, we additionally visualized the learned spatial weighting maps immediately before the classification layer. We resized these maps to the original image resolution and projected them onto the standard 90-region AAL atlas. As shown in Figure 3, the model consistently emphasizes regions in the medial temporal lobe. This focus aligns with the well-established neurodegenerative trajectory of Alzheimer's disease, supporting both the interpretability and clinical relevance of the proposed method.

4 Conclusion

This study introduces a multimodal framework to learn representations robust to non-pathological influences, guided by causal inference principles. To reduce the effect of imaging-related bias, the model uses MDM to learn feature masks that filter out redundant information and noise during feature extraction. Building on this, we add a patch-level cross-modal modulation module designed to simulate how confounding influences can affect specific local brain regions. Experiments and ablation studies on the ADNI and NACC datasets show that these two modules each contribute meaningful gains in classification performance. Looking ahead, we plan to extend the model to estimate causal effects, which may improve its usefulness for investigating disease pathology.

Acknowledgments. This work was supported by the Guangxi Science and Technology Program (No. FN2504240022), Guangxi Key R&D Project (No. AB24010167), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617), China Ministry of Education Humanities and Social Sciences Research-Planning Fund Project (No. 23YJA910007), Zhejiang Provincial Philosophy and Social Sciences Planning Project-Major Project (No. 23SYS12ZD).

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

References

1. Beekly, D.L., Ramos, E.M., van Belle, G., Deitrich, W., Clark, A.D., Jacka, M.E., Kukull, W.A., et al.: The national alzheimer's coordinating center (nacc) database: an alzheimer disease database. Alzheimer Disease and Associated Disorders **18**(4), 270–277 (2004)
2. Dodel, R., Frölich, L.: Donanemab for alzheimer's disease: from preclinical research to the clinical application. Expert Review of Neurotherapeutics pp. 1–13 (2025)
3. Duenias, D., Nichyporuk, B., Arbel, T., Raviv, T.R., et al.: Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. Medical Image Analysis **102**, 103503 (2025)
4. Dujardin, S., Commins, C., Lathuiliere, A., Beerepoot, P., Fernandes, A.R., Kamath, T.V., De Los Santos, M.B., Klickstein, N., Corjuc, D.L., Corjuc, B.T., et al.: Tau molecular diversity contributes to clinical heterogeneity in alzheimer's disease. Nature Medicine **26**(8), 1256–1263 (2020)
5. Elazab, A., Wang, C., Abdelaziz, M., Zhang, J., Gu, J., Gorriz, J.M., Zhang, Y., Chang, C.: Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: Achievements and future directions. Expert Systems with Applications **255**, 124780 (2024)
6. Guo, Z., Wu, Z., Xiao, T., Aggarwal, C., Liu, H., Wang, S.: Counterfactual learning on graphs: A survey. Machine Intelligence Research **22**(1), 17–59 (2025)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

8. Hu, X., Shen, X., Sun, Y., Shan, X., Min, W., Su, L., Fan, X., Elazab, A., Ge, R., Wang, C., et al.: Itcfn: Incomplete triple-modal co-attention fusion network for mild cognitive impairment conversion prediction. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2025)
9. Ma, W., Chen, C., Gong, Y., Chan, N.Y., Jiang, M., Mak, C.H.K., Abrigo, J.M., Dou, Q.: Causal effect estimation on imaging and clinical data for treatment decision support of aneurysmal subarachnoid hemorrhage. IEEE Transactions on Medical Imaging **43**(8), 2778–2789 (2024)
10. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). Alzheimer's and Dementia **1**(1), 55–66 (2005)
11. Oh, H.S.H., Urey, D.Y., Karlsson, L., Zhu, Z., Shen, Y., Farinas, A., Timsina, J., Duggan, M.R., Chen, J., Guldner, I.H., et al.: A cerebrospinal fluid synaptic protein biomarker for prediction of cognitive resilience versus decline in alzheimer's disease. Nature Medicine **31**(5), 1592–1603 (2025)
12. Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., Altman, E.: Tabular transformers for modeling multivariate time series. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3565–3569 (2021)
13. Pearl, J.: Causality. Cambridge University Press (2009)
14. Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology **78**(5), 947–1012 (2016)
15. Pölsterl, S., Wachinger, C.: Estimation of causal effects in the presence of unobserved confounding in the alzheimer's continuum. In: International Conference on Information Processing in Medical Imaging. pp. 45–57. Springer (2021)
16. Tang, X., Guo, R., Zhang, C., Zhuang, X., Qian, X.: A causality-driven graph convolutional network for postural abnormality diagnosis in parkinsonians. IEEE Transactions on Medical Imaging **42**(12), 3752–3763 (2023)
17. Toumaj, S., Heidari, A., Shahhosseini, R., Jafari Navimipour, N.: Applications of deep learning in alzheimer's disease: a systematic literature review of current trends, methodologies, challenges, innovations, and future directions. Artificial Intelligence Review **58**(2), 44 (2024)
18. Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis **79**, 102470 (2022)
19. Wang, T., Zhou, C., Sun, Q., Zhang, H.: Causal attention for unbiased visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3091–3100 (2021)
20. Wang, Z., Ge, R., Yu, J., Qin, F., Tian, Y., Li, N., Min, W., Elazab, A., Wang, C.: Cclnet: Causal and contrastive learning framework for enhanced pulmonary embolism detection. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2591–2596. IEEE (2024)
21. Xiong, H., Wang, Y., Shen, Z., Sun, K., Fang, Y., Chen, Y., Shen, D., Wang, Q.: Learning contrast and content representations for synthesizing magnetic resonance image of arbitrary contrast. Medical Image Analysis **104**, 103635 (2025)
22. Ye, J., Zeng, A., Pan, D., Zhang, Y., Zhao, J., Chen, Q., Liu, Y.: Mad-former: a traceable interpretability model for alzheimer's disease recognition based on multi-patch attention. IEEE Journal of Biomedical and Health Informatics **28**(6), 3637–3648 (2024)

23. Zhang, Y., Huang, Y.A., Hu, Y., Liu, R., Wu, J., Huang, Z.A., Tan, K.C.: Causalmixnet: A mixed-attention framework for causal intervention in robust medical image diagnosis. Medical Image Analysis **103**, 103581 (2025)