Multistage Alignment and Fusion for Multimodal Multiclass Alzheimer's Disease Diagnosis

Shuo Huang[1,2,*], Lujia Zhong[1,3,*], and Yonggang Shi[1,2,3]

[1] Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California (USC), Los Angeles, CA 90033, USA
[2] Alfred E. Mann Department of Biomedical Engineering, Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA 90089, USA
[3] Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA 90089, USA
[*] S. Huang and L. Zhong contributed equally to this work.
yshi@loni.usc.edu

Abstract. For the early diagnosis of Alzheimer's disease (AD), it is essential that we have effective multiclass classification methods that can distinct subjects with mild cognitive impairment (MCI) from cognitively normal (CN) subjects and AD patients. However, significant overlaps of biomarker distributions among these groups make this a difficult task. In this work, we propose a novel framework for multimodal, multiclass AD diagnosis that can integrate information from diverse and complex modalities to resolve ambiguity among the disease groups and hence enhance classification performances. More specifically, our approach integrates T1-weighted MRI, tau PET, fiber orientation distribution (FOD) from diffusion MRI (dMRI), and Montreal Cognitive Assessment (MoCA) scores to classify subjects into AD, MCI, and CN groups. We introduce a Swin-FOD model to extract order-balanced features from FOD and use contrastive learning to align MRI and PET features. These aligned features and MoCA scores are then processed with a Tabular Prior-data Fitted In-context Learning (TabPFN) method, which selects model parameters based on the alignment between input data and prior data during pre-training, eliminating the need for additional training or fine-tuning. Evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset ($n = 1147$), our model achieved a diagnosis accuracy of 73.21%, outperforming all comparison models ($n = 10$). We also performed Shapley analysis and quantitatively evaluated the essential contributions of each modality.

Keywords: Multimodal · Multiclass · Alzheimer's disease · Diagnosis · Feature alignment.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that advances from cognitively normal (CN) to mild cognitive impairment (MCI) before

reaching AD [1]. With no effective treatment available, automated multiclass diagnosis holds the potential of early detection of the MCI status, which is crucial for delaying or preventing disease progression. However, significant overlap in biomarker values and imaging characteristics between MCI, CN, and AD makes accurate diagnosis challenging. The integration of multimodal data can potentially enhance multiclass AD diagnosis by leveraging complementary clinical and imaging information [1, 2], but it is still a challenging task to effectively align features from complex and diverse modalities.

Various multimodal data analysis methods have been developed for both image and tabular data. In image analysis, deep neural networks such as CoAtNet [3], ConvNeXt [4], MaxViT [5], SwinUNETR [6] and contrastive learning approaches [7] have shown strong performance by effectively capturing local and global image features. These models have been proved to be efficient in multimodal multiclass AD diagnosis [8]. Embedding images into tabular features can help combine the information in images with tabular test results, and algorithms such as AdaBoost [9], XGBoost [10], TabNet [11], AutoGluon [12], LightGBM [13], and Tabular Prior-data Fitted Network (TabPFN) [14] have achieved state-of-the-art results for multiclass classifications on tabular medical data.

In the context of AD diagnosis, multimodal analysis methods have proven valuable. For instance, Ou et al. [1] integrated MRI and PET data, and Qiu et al. [15] combined MRI data with mini-mental state examination (MMSE) tabular data for AD diagnosis. However, the performance of these methods are still limited on multiclass classification ($\sim 60\%$ accuracy for 3-class diagnosis). To further enhance multicalss AD diagnosis, it is essential to incorporate more diverse and complex imaging modalities together with tabulated information from clinical evaluations. However, if not handled effectively, the misalignment of data from heterogeneous modalities can obscure modality relationships and limit diagnostic accuracy [16, 17].

To address the challenges in the alignment and fusion of data from heterogeneous and complex modalities, we propose here a novel framework for multimodal and multiclass AD diagnosis. Our framework integrates scalar images (T1-weighted MRI (T1w MRI), tau PET), high-dimensional fiber orientation distribution (FOD) from diffusion MRI, and tabular data (age, sex, Montreal Cognitive Assessment (MoCA) scores). First, we developed a SWIN-FOD model to process the complex 4D FODs efficiently. For fusing MRI and PET, we adapted the ALBEF model to handle 3D volumes. To capture relationships between features, we employed the pretrained priors in TabPFN, avoiding the need for additional feature alignment. Tested on the ADNI dataset ($n = 1147$), our model achieved 73.21% accuracy, surpassing all comparison methods. Additionally, we analyzed the impact of each modality on the final diagnosis by Shapley analysis.

2 Method

Fig. 1 shows the workflow of the proposed framework. We will introduce each part of the method in detail in the following sections.

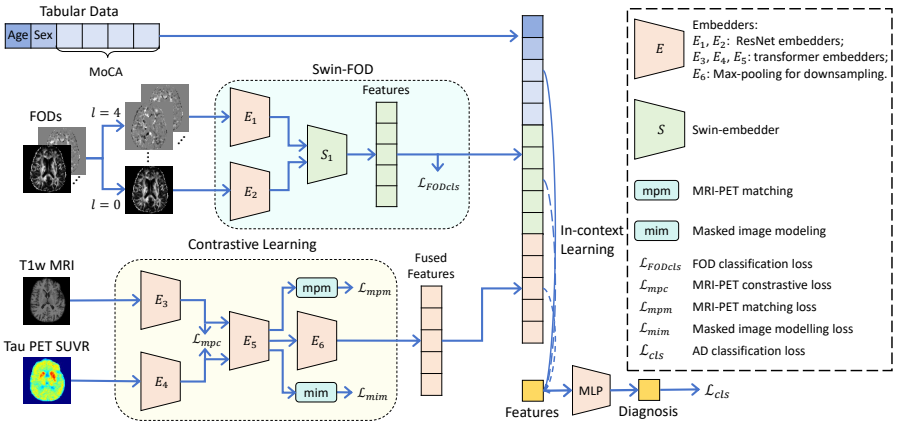

Fig. 1. Overview of the proposed framework. The Swin-FOD model and contrastive learning model are first trained separately. Their extracted features are then concatenated with tabular data, and in-context learning is applied to generate the diagnosis.

2.1 Swin-FOD

Our first alignment balances FOD information across different orders. Following Ref. [18], FODs are represented using spherical harmonics (SPHARM) up to a maximum order L ($L = 0, 2, 4, ...$), forming complex 4D images. The number of 3D images at order l is $(2l+1)$, totaling $(L+1)(L+2)/2$. Lower-order FODs carry low-frequency information with higher signal-to-noise ratios and fewer volumes, while high-order FODs capture finer brain connectivity details.

To efficiently process both low- and high-order FODs while reducing memory consumption in transformer-based models [19], we propose an order-balanced Swin encoder [6, 20], Swin-FOD, as shown in Fig. 2a. FOD volumes are grouped by order and embedded using a ResNet-based module [21], ensuring uniform latent representations across orders. These features are concatenated and passed through the Swin encoder to extract multi-resolution features.

To enhance information flow, we introduce long-range skip connections and use a convolution block to align feature sizes at each resolution. Finally, a multilayer perceptron (MLP) generates 1D tabular features.

Swin-FOD is trained by minimizing cross-entropy loss ($\mathcal{L}_{FODcls}$) for three-class AD diagnosis to extract features that closely related with AD diagnosis, as $\mathcal{L}_{FODcls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{3} y_{i,c}\log(\hat{y}_{i,c})$, where $\hat{y}_{i,c}$ is the predicted probability of data i for class c, $y_{i,c}$ is the ground truth diagnosis.

2.2 Fusing T1w MRI and Tau PET by Contrastive Learning

Inspired by the ALBEF model [7], we propose a contrastive learning-based method to align PET and T1w MRI data after a pre-process of image registration and tau PET Standardized Uptake Value Ratio (SUVR) calculation, generating a fused 1D tabular representation (Fig. 2b).

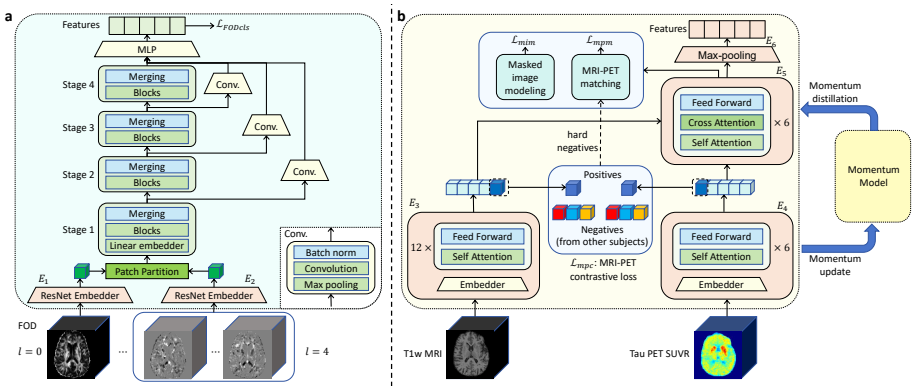

Fig. 2. (a) Flowchart of the Swin-FOD model. (b) Contrastive learning-based data fusion between T1-weighted MRI and Tau PET.

Our model has two alignment stages. The first stage minimizes the MRI-PET contrastive loss $\mathcal{L}_{mpc}$, reducing distances between MRI and PET embeddings. MRI and PET data are embedded using separate encoders, with deeper layers for MRI due to its complexity [7]. The loss function is:

$$\mathcal{L}_{mpc} = (\mathcal{L}_{M2P} + \mathcal{L}_{P2M})/2 \tag{1}$$

where $\mathcal{L}_{M2P}$ (MRI to PET loss) is:

$$\mathcal{L}_{M2P} = \mathbb{E}_{j\sim D}\, H(\boldsymbol{y}_j^{M2P}, \hat{\boldsymbol{s}}_j^{M2P}), \tag{2}$$

with $\mathbb{E}$ is the average function, $H()$ is the cross entropy, $\boldsymbol{y}_j^{M2P}$ is the ground truth similarity (1 for same case, 0 otherwise). $\hat{\boldsymbol{s}}^{M2P} = (\boldsymbol{e}_{\text{mri}} \cdot \boldsymbol{e}_{\text{pet}}^{'\top})/\tau$ is the predicted similarity between normalized MRI $\boldsymbol{e}_{\text{mri}}$ and PET $\boldsymbol{e}_{\text{pet}}^{'}$. $\tau = 0.07$ is the temperature parameter controlling similarity sharpness.

The second stage generates fused features by minimizing masked image modeling loss ($\mathcal{L}_{mim}$) and MRI-PET matching loss ($\mathcal{L}_{mpm}$). $\mathcal{L}_{mim}$ measures the cross-entropy loss of the recovered 15% randomly masked PET patches. Since PET is strongly associated with AD, this guides MRI features toward relevant features in AD diagnosis. $\mathcal{L}_{mpm}$ ensures that features from different subjects remain distinguishable by preserving discriminative information:

$$\mathcal{L}_{mim} = \mathbb{E}_{j\sim D}\, H(\boldsymbol{y}_j, \hat{\boldsymbol{p}}_j^{mask}),\ \mathcal{L}_{mpm} = \mathbb{E}_{j\sim D}\, H(\boldsymbol{y}_j^{itm}, \hat{\boldsymbol{p}}_j^{itm}) \tag{3}$$

where $\boldsymbol{y}_j$ and $\boldsymbol{y}_j^{itm}$ are the ground truth, and $\hat{\boldsymbol{p}}_j^{mask}$ and $\hat{\boldsymbol{p}}_j^{itm}$ is the predicted probability. The full loss function is:

$$\mathcal{L} = \mathcal{L}_{mpc} + \mathcal{L}_{mim} + \mathcal{L}_{mpm}. \tag{4}$$

To stabilize training under noisy data, we use a momentum encoder that maintains a slowly updated copy of the feature extractor, following Ref. [7]. This

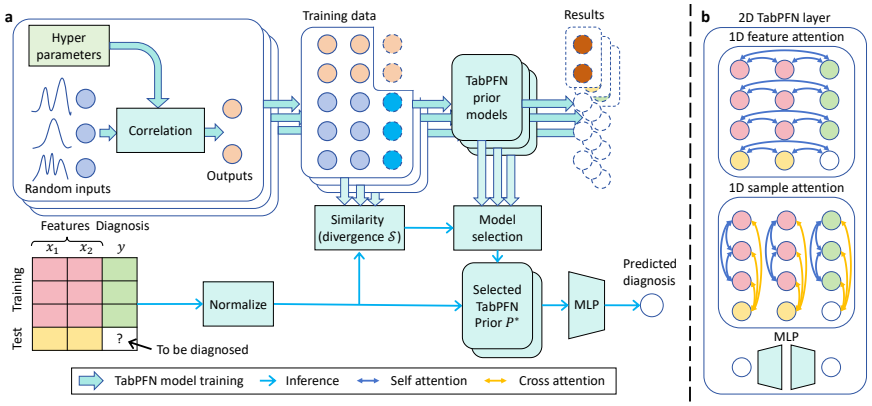

Fig. 3. Flowchart of AD diagnosis using TabPFN: (a) Prior selection and model inference for real-world data. (b) TabPFN layer structure.

provides more consistent pseudo-targets for learning and improves convergence. We add KL-divergence regularization to each term of the loss $\mathcal{L}$:

$$\mathcal{L}_{new} = (1-\alpha)\mathcal{L} + \alpha \mathbb{E}_{j\sim D}\,\mathrm{KL}(\boldsymbol{q}_j||\hat{\boldsymbol{p}}_j), \tag{5}$$

where $\boldsymbol{q}_j$ is the pseudo-target generated by the momentum model, $\hat{\boldsymbol{p}}_j$ is the model's prediction, and $\alpha = 0.4$, following the settings in Ref. [7].

2.3 Prior-data Fitted In-context Learning

Features extracted from images along with tabular data, including age, sex, and MoCA scores varies from each other in value and distribution, including continuous values (image features), binary values (sex), and categorical values (MoCA scores). To better measure the correlation between each feature and each data sample, we employed the pre-trained TabPFN model [14]. Inference using the TabPFN does not require training or fine tuning, only an alignment between our data and the prior data in pre-training is needed, as shown in Fig. 3a.

TabPFN is pre-trained on $\sim$ 100 million synthetic datasets generated via structural causal models, capturing real-world tabular data characteristics. The model consists of 12 $\times$ 2D TabPFN layers (Fig. 3b), which use two 1D attentions to extract both within-sample and cross-sample correlations [22]. TabPFN selects the prior model whose training data aligns with our inputs best. Given features $F(X, y)$, TabPFN finds the optimal prior P^* by minimizing divergence $\mathcal{S}$:

$$P^* = \arg\min_{P_i} \mathcal{S}(F, P_i). \tag{6}$$

We then use Bayesian posterior prediction to estimate y_{test} for new input X_{test} using the selected prior, computing a weighted average of predictions from different task hypotheses [23]:

$$P(y_{test}|X_{test}, F) = \int_t P(y_{test}|X_{test}, t)P(t|F) \approx \int_t P(y_{test}|X_{test}, t)P(t|P^*), \tag{7}$$

Table 1. Diagnosis results of the proposed method on the test set.

Group	Number	Acc. ↑	AUC ↑	F1 ↑	Prec. ↑	Recall ↑	MCC ↑	Confusion matrix		
CN	243			0.8310	0.8038	0.8601		209	34	0
MCI	143	0.7321	0.8625	0.5959	0.5838	0.6084	0.5299	49	87	7
AD	62			0.6336	0.8205	0.5161		2	28	32

where t denotes the function mapping inputs X to outputs y under a given prior distribution P. Finally, an MLP is used to generate the AD diagnosis.

3 Results

Our model is trained and tested using the ADNI [24, 25] dataset ($n = 1147$), which contains 643 CN, 345 MCI, and 159 AD cases. The dataset includes 597 female (72.92 ± 7.68 years) and 550 male (75.44 ± 7.27 years) subjects. We randomly split the data into 60% training, 20% validation, and 20% test sets, ensuring the same training, validation and test sets across all models to prevent data leakage. As TabPFN does not require training, we test it on the validation and test sets. To address class imbalance, we oversampled MCI and AD cases in the training set. Data from the same case are aligned to the same space before using. All models were trained and tested on NVIDIA A5000 GPU workstations.

Our model is evaluated using mean Accuracy (Acc.), Area Under the Curve (AUC), mean F1 score, mean Precision (Prec.), mean Recall, and Matthews Correlation Coefficient (MCC). To interpret feature contributions, we employ SHapley Additive exPlanations (SHAP) [26, 27], which quantify the impact of individual features on the model's predictions.

3.1 Comparison between Different Methods

Table 1 presents the classification results of the proposed method. Our model achieved a high diagnostic performance with an accuracy of 0.73 and an AUC of 0.86. The MCC of 0.5299 indicates a strong correlation between the predicted diagnosis and the ground truth.

Table 2 presents a comparison between the proposed method and other approaches. We first evaluated our model using volumetric images only (MRI, PET, and FOD), on comparison with existing image-based classification models [28] which have demonstrated effectiveness in multimodal multiclass AD diagnosis [8]. All models are trained, validated and tested using the same data as our method. Next, we compared our results with other models that are efficient in tabular data analysis, using the features generated by our models. Our model outperforms all comparison methods. Although CoAtNet [3] and ConvNeXt [4] achieve higher recall when using image features, their lower precision indicates an increased false positive rate.

Table 2. Comparison of different methods based on classification metrics. **Bold**: best value, underline: second best value.

T1	PET	FOD	Tab.	Method	Acc. ↑	AUC ↑	F1 ↑	Prec. ↑	Recall ↑	MCC ↑
✓	✓	✓		ResNet[21]	0.5381	0.5912	0.4373	0.4732	0.4656	0.2127
✓	✓	✓		MaxViT[5]	0.5391	0.6217	0.4471	0.4626	0.4351	0.2216
✓	✓	✓		CoAtNet[3]	0.5348	0.6234	0.4483	0.3888	0.5328	0.2078
✓	✓	✓		ConvNeXt[4]	0.5165	0.6216	0.4225	0.3622	**0.5413**	0.1916
✓	✓	✓		SwinUNETR[6]	0.5429	0.6379	0.4575	0.4321	0.4634	0.2165
✓	✓	✓		Proposed	**0.6138**	**0.6889**	**0.4789**	**0.5761**	0.4717	**0.2841**
✓	✓	✓	✓	AdaBoost[9]	0.7054	0.7584	0.6402	0.6954	0.6190	0.4850
✓	✓	✓	✓	XGBoost[10]	0.7143	0.8317	0.6676	0.6913	0.6596	0.4991
✓	✓	✓	✓	TabNet[11]	0.6295	0.7215	0.5697	0.5677	0.5830	0.3607
✓	✓	✓	✓	AutoGluon[12]	0.6942	0.8367	0.6441	0.6703	0.6280	0.4608
✓	✓	✓	✓	LightGBM[13]	0.7076	0.8338	0.6701	0.6834	0.6610	0.4893
✓	✓	✓	✓	Proposed	**0.7321**	**0.8625**	**0.6868**	**0.7361**	**0.6615**	**0.5299**

Table 3. Comparison of ablation studies based on classification metrics.

T1	PET	FOD	Tab.	Acc. ↑	AUC ↑	F1 ↑	Prec. ↑	Recall ↑	MCC ↑
✓	✓			0.5915	0.7039	0.5094	0.5270	0.5015	0.2733
		✓		0.5535	0.5545	0.3560	0.4248	0.3822	0.1408
✓	✓	✓		0.6138	0.6889	0.4789	0.5761	0.4717	0.2841
			✓	0.6786	0.8402	0.5852	0.6603	0.5633	0.4224
		✓	✓	0.7031	0.8446	0.6174	0.7097	0.5902	0.4670
✓	✓		✓	0.7120	0.8548	0.6689	0.6946	0.6532	0.5023
✓	✓	✓	✓	**0.7321**	**0.8625**	**0.6868**	**0.7361**	**0.6615**	**0.5299**

3.2 Ablation Studies

Table 3 presents the results of the ablation studies. The model using multimodal features outperforms each single modality, with the inclusion of image features yielding more accurate diagnoses than using only MoCA scores. The best performance across all metrics is achieved by combining all modalities.

Table 4 analyzed the contrastive learning method. Specifically, we evaluated the performance of alignment at each stage by comparing features before and after the merging encoder E_5 in Fig. 2b, named as "not fused" and "fused", respectively. Alignment using $\mathcal{L}_{mpc}$ improves diagnosis accuracy of both MRI and PET data. However, directly concatenating the features results in performance between that of the single modalities. After the two-stage aliment, the accuracy exceeds that of each single modality's aligned features, demonstrating that the second stage effectively fuses information from MRI and PET.

Table 4 also compares the classification metrics between the original Swin-UNETR and Swin-FOD. The addition of the order balance embedder and long-range skip connections both contribute to improved performance.

Table 4. Ablation studies of contrastive learning and Swin-FOD.

T1	PET	FOD	Method	Acc. ↑	AUC ↑	F1 ↑	Prec. ↑	Recall ↑	MCC ↑
✓			ResNet Encoder	0.5044	0.5564	0.3513	0.3839	0.3672	0.0593
	✓		ResNet Encoder	0.5446	0.6798	0.4598	0.4750	0.4657	0.1769
✓			Contr. (not fused)	0.5250	0.6111	0.3871	0.4090	0.3914	0.1277
	✓		Contr. (not fused)	0.5708	0.7027	0.4951	0.5142	0.4859	0.2392
✓	✓		Contr. (not fused)	0.5468	0.6616	0.4570	0.4852	0.4483	0.2035
✓	✓		Contr. (fused)	**0.5915**	**0.7039**	**0.5094**	**0.5270**	**0.5015**	**0.2733**
		✓	SwinUNETR	0.5000	0.5282	0.3279	0.3341	0.3667	0.0964
		✓	+ Order balance	0.5326	0.5504	0.3419	0.3659	**0.3919**	0.1205
		✓	Swin-FOD	**0.5535**	**0.5545**	**0.3560**	**0.4248**	0.3822	**0.1408**

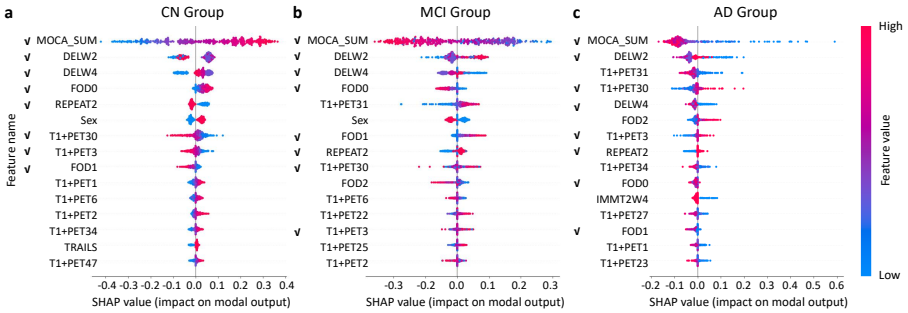

Fig. 4. Top 15 features with the highest impact on diagnosis for CN (a), MCI (b), and AD (c) groups. Features marked with a tick have a high impact across all three groups.

3.3 Impact of Features

Fig. 4 shows the top 15 features with the highest mean absolute SHAP values [26, 27] contributing to the diagnosis of each group. Eight features have high impact in all three groups (marked with a tick). The MoCA score has the highest impact across all groups. More features from MRI and PET data than FOD data are among the top 15 features.

Table 5 shows the total mean absolute SHAP values for each modality. Our model extracts the most information for the CN group, followed by the MCI group, then for the AD group. The tabular data (including age, sex and MoCA scores) has the greatest impact, with MRI and PET features providing additional information. Although the total impact of FOD is less than 10%, its inclusion still improves model performance, as shown in Table 3. While FOD contributes less to AD diagnosis, its information remains beneficial.

4 Conclusion

We propose a framework for multiclass AD diagnosis by integrating multimodal data. Our approach includes a Swin-FOD model for extracting order-balanced

Table 5. Comparison of SHAP scores for each modality across each group.

Modality	CN / Percentage	MCI / Percentage	AD / Percentage
Tabular data	0.4185 / 59.90%	0.2861 / 53.70%	0.2605 / 52.83%
T1w MRI+Tau PET	0.2307 / 33.02%	0.1946 / 36.53%	0.2038 / 41.34%
FOD	0.0495 / 7.08%	0.0521 / 9.77%	0.0288 / 5.83%
All	0.6987 / 100.00%	0.5328 / 100.00%	0.4930 / 100.00%

FOD features, a contrastive learning model for aligning MRI and PET images, and a pre-trained TabPFN model for tabular data analysis. The model achieved 73.21% accuracy in classifying CN, MCI, and AD groups, outperforming all comparison methods. Additionally, we performed SHAP analysis to assess the contribution of each modality to the final diagnosis. Our code is available at: https://github.com/huangshuo343/multimodalAD.

Future work will explore additional imaging modalities, extend analysis to multi-site [29–31] and multi-tracer datasets to enhance model robustness, integrate all modalities within one contrastive learning framework, analyze the contribution of each component, and make comparison with more methods.

Acknowledgments. This work was supported by the National Institute of Health (NIH) under grants R01EB022744, RF1AG077578, RF1AG064584, U19AG078109, and P30AG066530. Authors thank the ADNI investigators (https://adni.loni.usc.edu).

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

References

1. Ou, Z., Jiang, C., Liu, Y., Zhang, Y., Cui, Z., Shen, D.: A graph-embedded latent space learning and clustering framework for incomplete multimodal multiclass alzheimer's disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 45–55. Springer (2024)
2. Odusami, M., Maskeliūnas, R., Damaševičius, R., Misra, S.: Machine learning with multimodal neuroimaging data to classify stages of alzheimer's disease: a systematic review and meta-analysis. Cognitive Neurodynamics **18**(3), 775–794 (2024)
3. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in neural information processing systems **34**, 3965–3977 (2021)
4. Woo, S., Debnath, S., Hu, R., et al.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
5. Tu, Z., Talebi, H., Zhang, H., et al.: Maxvit: Multi-axis vision transformer. In: European conference on computer vision. pp. 459–479. Springer (2022)
6. Hatamizadeh, A., Nath, V., Tang, Y., et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)

7. Li, J., Selvaraju, R., Gotmare, A., et al.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
8. Kadri, R., Bouaziz, B., Tmar, M., Gargouri, F.: Efficient multimodel method based on transformers and coatnet for alzheimer's diagnosis. Digital Signal Processing **143**, 104229 (2023)
9. Zhu, J., Zou, H., Rosset, S., Hastie, T., et al.: Multi-class adaboost. Statistics and its Interface **2**(3), 349–360 (2009)
10. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
11. Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 6679–6687 (2021)
12. Erickson, N., Mueller, J., Shirkov, A., et al.: Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505 (2020)
13. Ke, G., Meng, Q., Finley, T., et al.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
14. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., et al.: Accurate predictions on small data with a tabular foundation model. Nature **637**(8045), 319–326 (2025)
15. Qiu, S., Miller, M.I., Joshi, P.S., et al.: Multimodal deep learning for alzheimer's disease dementia assessment. Nature communications **13**(1), 3404 (2022)
16. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020)
17. Liu, Y., Liu, M., Zhang, Y., Sun, K., Shen, D.: A progressive single-modality to multi-modality classification framework for alzheimer's disease sub-type diagnosis. In: International Workshop on Machine Learning in Clinical Neuroimaging. pp. 123–133. Springer (2025)
18. Qiao, Y., Shi, Y.: Unsupervised deep learning for fod-based susceptibility distortion correction in diffusion mri. IEEE transactions on medical imaging **41**(5), 1165–1175 (2021)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
22. Garg, S., Tsipras, D., Liang, P.S., Valiant, G.: What can transformers learn in-context? a case study of simple function classes. Advances in Neural Information Processing Systems **35**, 30583–30598 (2022)
23. Müller, S., Hollmann, N., Arango, S.P., Grabocka, J., Hutter, F.: Transformers can do bayesian inference. arXiv preprint arXiv:2112.10510 (2021)
24. Weiner, M.W., Veitch, D.P., Aisen, P.S., et al.: The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. Alzheimer's & Dementia **13**(5), 561–571 (2017)

25. Weiner, M.W., Kanoria, S., Miller, M.J., et al.: Overview of alzheimer's disease neuroimaging initiative and future clinical trials. Alzheimer's & Dementia **21**(1), e14321 (2025)
26. Shapley, L.S.: A value for n-person games. Contribution to the Theory of Games **2** (1953)
27. Scott, M., Su-In, L., et al.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30**, 4765–4774 (2017)
28. Solovyev, R., Kalinin, A.A., Gabruseva, T.: 3d convolutional neural networks for stalled brain capillary detection. Computers in Biology and Medicine **141**, 105089 (2022). https://doi.org/10.1016/j.compbiomed.2021.105089
29. Sperling, R.A., Aisen, P.S.: Anti-amyloid treatment in asymptomatic alzheimer's disease (a4) and longitudinal evaluation of amyloid risk and neurodegeneration (learn) study longitudinal results. In: Alzheimer's Association International Conference. ALZ (2023)
30. Grober, E., Lipton, R.B., Sperling, R.A., et al.: Associations of stages of objective memory impairment with amyloid pet and structural mri: The a4 study. Neurology **98**(13), e1327–e1336 (2022)
31. Jann, K., Boudreau, J., Albrecht, D., et al.: Fmri complexity correlates with tau-pet and cognitive decline in late-onset and autosomal dominant alzheimer's disease. Journal of Alzheimer's Disease **95**(2), 437–451 (2023)