Volume3, Number 2, 2026 Pages 438-450

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

A HYBRID CNN-TRANSFORMER FRAMEWORK FOR MULTI-CLASS ALZHEIMER'S DISEASE CLASSIFICATION USING MRI: A ROBUST AND INTERPRETABLE APPROACH

Muhammad Ali[1], Saddaqat Ali[*2], Rameesha Javed[3], Muhammad Usman Hameed[4]

[1]*Riphah International University, Sahiwal, 57000, Pakistan*
[*2]*Deputy Assistant Director in Nadra MPhil Computer Science NCBAE Lahore (Rahim Yarkhan Campu)*
[3]*Lecturer Computer Science NCBAE Lahore (Rahim Yar Khan Campus)*
[4]*Lecturer Computer Science Superior University Lahore (Rahim Yar khan Campus)*

[1]m.ali@riphahsahiwal.edu.pk, [*2]saddaqatali105@gmail.com, [3]jvedrameesha9@gmail.coma
[4]mrusmanphd@gmail.com

Article Information

Article History

Received:	May 15, 2026
Revised:	June 27, 2026
Accepted:	August 12, 2026
Available Online:	August 31, 2026

Keywords:
Alzheimer's disease, convolutional neural networks, vision transformers, hybrid deep learning, magnetic resonance imaging, medical image classification

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder for which timely and reliable assessment is important. Structural magnetic resonance imaging (MRI) provides useful anatomical information, but manual review is time-consuming and subject to reader variability. This study presents a hybrid CNN-Transformer framework for four-class classification of 2D MRI slices. A ResNet-18 backbone extracts local anatomical features, while a Transformer encoder models global relationships among the resulting spatial tokens. A feature-fusion layer combines the pooled CNN representation with the Transformer's derived class-token embedding before prediction. The model was evaluated on a Kaggle OASIS-derived dataset containing NonDemented, Very Mild Demented, Mild Demented, and Moderate Demented classes. Because subject identities were unavailable, the result is interpreted within the stated slice-level protocol rather than as evidence of subject-level clinical generalization. The model achieved 99.81% validation accuracy. Class-wise precision and recall were high, and Grad-CAM visualizations were used to inspect spatial evidence associated with predictions. These findings support the value of combining local and global representations for MRI-slice classification. However, the slice-level split may yield optimistic estimates because slices from the same subject can occur in both subsets. Subject-level and external validation are therefore required before conclusions can be drawn about patient-level generalization or clinical use.

INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that gradually impairs memory, cognition, and behavior. It is the most common cause of dementia and creates a substantial burden for

patients, caregivers, and health systems. As populations age, the number of people living with AD is expected to rise, making timely diagnosis and disease monitoring increasingly important. Early identification can support treatment planning, risk management, and access to appropriate care. Structural magnetic resonance imaging (MRI) is widely used to assess AD because it provides detailed anatomical information without ionizing radiation. MRI can reveal patterns of atrophy in regions such as the hippocampus and temporal lobes, but manual interpretation is time-consuming, requires specialist expertise, and may vary between readers. Automated decision-support systems could therefore assist clinicians by providing faster and more consistent image assessment.

Deep learning, particularly convolutional neural networks (CNNs), has achieved strong results across medical image analysis tasks [1] and in neuroimaging-based AD classification [2]. CNNs are well suited to learning local patterns such as edges, textures, and anatomical structures from MRI slices. Nevertheless, their reliance on local receptive fields can limit the direct modeling of relationships between distant brain regions. This limitation matters in disease staging, where discriminative changes may be subtle and distributed rather than confined to a single structure.

Vision Transformers (ViTs) offer a complementary approach by using self-attention to model interactions among image patches. Their ability to represent long-range dependencies has produced encouraging results in medical image classification and segmentation. Pure Transformer models, however, generally require substantial training data and can be computationally demanding at high image resolutions. These constraints are especially relevant in medical imaging, where labeled datasets are often limited. Hybrid CNN-Transformer architectures address this trade-off by using convolutional layers to extract local features and attention mechanisms to place those features in a broader spatial context.

This study develops a hybrid CNN-Transformer model for four-class classification of MRI slices as NonDemented, Very Mild Demented, Mild Demented, or Moderate Demented. A ResNet-18 backbone extracts local anatomical features, while a Transformer encoder models global relationships among the resulting spatial tokens. Their representations are fused before classification. On a stratified 80/20 slice-level split of the Kaggle OASIS-derived dataset, the model achieved 99.81% validation accuracy. Because subject identifiers were unavailable, this result is interpreted within the stated slice-level protocol rather than as evidence of subject-level clinical generalization.

Contributions

The main contributions of this study are as follows:

1.A hybrid architecture that combines ResNet-18 feature extraction with Transformer-based contextual modeling for four-class Alzheimer's disease classification.

2. A feature-fusion strategy that joins the globally pooled CNN representation with the Transformer's classification-token embedding before prediction.

3. A controlled ablation analysis comparing the proposed model with CNN-only, Transformer-only, and unfused hybrid configurations under the same evaluation protocol.

4. Transparent reporting of the slice-level split and its potential to produce optimistic estimates when slices from the same subject occur in different subsets.

5. A qualitative Grad-CAM analysis that examines the spatial evidence used by the model and identifies areas requiring further clinical validation.

The remainder of the paper is organized as follows. Section 2 reviews CNN-, Transformer-, and hybrid approaches to AD classification. Section 3 describes the dataset, preprocessing pipeline, architecture, and training objective. Section 4 presents the experimental results and their interpretation. Section 5 summarizes the conclusions, limitations, and directions for future work.

LITERATURE REVIEW

Deep learning has become an important component of neuroimaging research on Alzheimer's disease. As MRI datasets and computational resources have expanded, studies have moved from handcrafted features toward learned representations that support diagnosis and disease staging. The literature now includes CNN-based methods, attention-driven models, and hybrid architectures designed to combine local anatomical detail with global context

.

CNN-Based Approaches for Alzheimer's Disease Classification

CNNs remain widely used for MRI-based AD classification because they learn hierarchical spatial features directly from images. Early work focused largely on binary discrimination between cognitively normal participants and patients with AD. More recent studies have considered multi-class settings that include intermediate stages such as mild cognitive impairment or very mild dementia [2]. These settings are more challenging because adjacent stages may differ only in subtle structural patterns.

Deeper networks, including ResNet, DenseNet, and EfficientNet variants, have reported strong performance on structural MRI. Their convolutional filters can capture localized changes such as cortical thinning and hippocampal atrophy. Helaly et al. [3], for example, evaluated a deep-learning approach for early AD detection and highlighted the value of learned MRI features. At the same time, unusually high results on slice-based datasets must be interpreted carefully: if slices from one subject are divided across training and evaluation subsets, the measured performance may overstate generalization to unseen patients.

A further limitation is that convolution alone does not explicitly model relationships among distant regions. Although stacking layers enlarges the effective receptive field, long-range interactions remain indirect. This may be important when distinguishing NonDemented from Very Mild Demented cases, for which structural differences can be small and spatially distributed. Attention mechanisms have therefore been explored as a way to complement local convolutional features.

Transformer and Attention-Based Models in Neuroimaging

Transformers were introduced for natural language processing and later adapted to images through the Vision Transformer [4]. In contrast to the local operations of a CNN, self-attention allows each image token to interact with every other token. This property makes Transformers attractive for neuroimaging tasks that may depend on distributed anatomical changes.

Recent studies have applied Transformer variants to MRI, functional MRI, retinal imaging, and multimodal data for AD-related prediction [4]-[8]. These models can capture broad spatial or temporal dependencies, particularly when supported by pretraining, augmentation, or an auxiliary convolutional backbone. Ensemble and multimodal designs have also been investigated to improve stability across disease categories [5], [8].

The main difficulty with a Transformer-only design is its demand for data and computation. Medical imaging datasets are typically much smaller than natural-image corpora, increasing the risk of overfitting. In addition, the cost of self-attention grows rapidly with the number of tokens. Hybrid architectures are therefore appealing because convolution can reduce and structure the feature space before attention is applied.

Hybrid CNN-Transformer Architectures

Hybrid CNN-Transformer models divide representation learning into two complementary tasks. The CNN extracts local anatomical patterns and produces a compact feature map, while the Transformer models relationships among spatial locations in that map. This arrangement can retain fine-grained image information without applying self-attention directly to every input pixel.

Several studies have combined convolutional and attention-based components for AD prediction. Pixel-level fusion with a Vision Transformer has been explored for early detection [9], while Kadri et al. [12] reported a CNN-Transformer encoder design for progression-stage classification. Related work has used graph Transformers for resting-state fMRI [7] and improved Transformer modules for multimodal diagnosis [8]. Although these studies differ in modality and evaluation protocol, they collectively suggest that local feature extraction and global contextualization can be complementary.

Existing hybrid models also leave several practical questions unresolved. Some architectures are computationally complex, many focus on binary or three-class tasks, and reporting at the class level is inconsistent. Comparisons are further complicated by differences in preprocessing, dataset composition, and subject-level versus slice-level splitting. A useful evaluation should therefore report the split strategy, class-wise behavior, and limitations alongside aggregate accuracy.

Emerging Trends: Robustness, Explainability, and Multi-Class Classification

Clinical decision-support models require more than high aggregate accuracy. Their behavior should remain stable under realistic variation, and clinicians need evidence about the image regions that influence a prediction. Grad-CAM, saliency maps, and attention visualization are commonly used for qualitative inspection, including in explainable Transformer-based AD research [6]. These methods can reveal where a model attends, but they do not by themselves establish clinical validity or causal relevance. Systematic reviews show continuing interest in multi-class AD classification and in models that distinguish early disease stages rather than relying only on binary diagnostic labels [2], [10]. They also identify persistent concerns about heterogeneous evaluation protocols, limited external validation, and the interpretability of high-performing models. Hybrid CNN-Transformer designs are promising in this context, provided that their gains are established through controlled comparisons and appropriately cautious reporting.

Research Gap and Motivation

Despite recent progress, three gaps remain. First, CNN-only models do not explicitly capture global dependencies, whereas Transformer-only models can be data- and computation-intensive. Second, relatively few studies examine four-class staging with detailed class-wise results. Third, very high performance on slice-level datasets is often reported without sufficient discussion of leakage risk or its implications for clinical generalization.

To address these gaps, this study presents a hybrid CNN-Transformer network for four-class classification of AD from 2D MRI slices. The architecture combines a ResNet-18 feature extractor, Transformer-based contextual modeling, and late feature fusion. The evaluation includes class-wise metrics, an ablation study, a comparison with published results, Grad-CAM visualizations, and an explicit account of the limitations imposed by the available dataset.

MATERIALS AND METHODS

This section describes the dataset, preprocessing pipeline, hybrid architecture, feature-fusion mechanism, and optimization procedure used in the study.

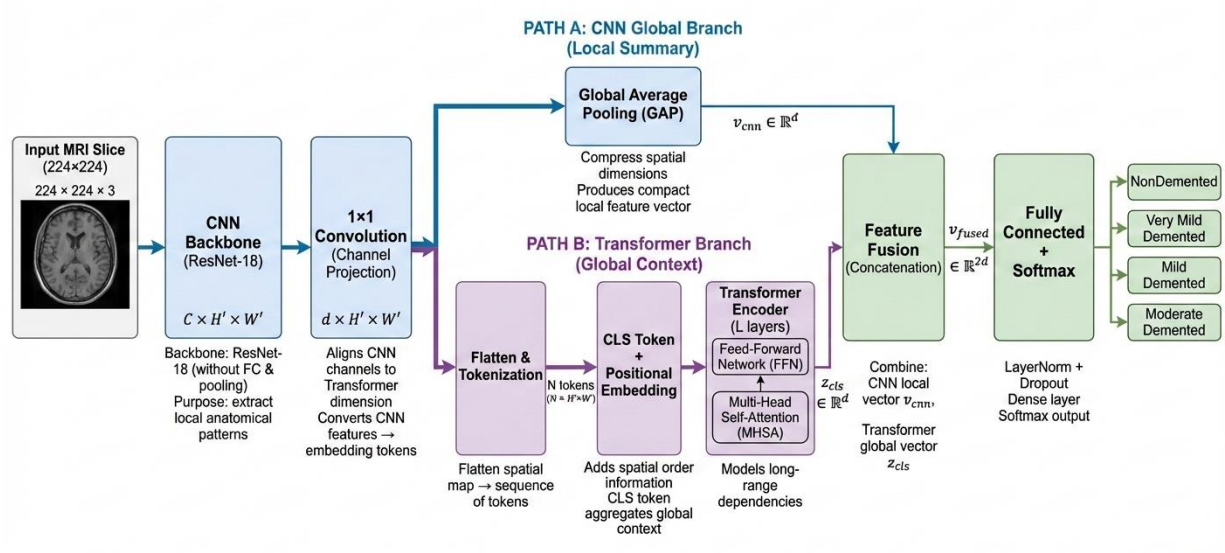


Fig. 1. Proposed hybrid CNN-Transformer architecture for four-class Alzheimer's disease classification.

Dataset Description

The experiments use the Kaggle OASIS Alzheimer's disease MRI dataset, a collection of preprocessed T1-weighted 2D MRI slices labeled as NonDemented, Very Mild Demented, Mild Demented, or Moderate Demented. The available files do not include subject identifiers. The images were therefore divided using

a stratified 80/20 slice-level split for training and validation, respectively, while preserving the class distribution. This choice enables reproducible experiments with the released data, but it may allow slices from the same individual to appear in both subsets. The reported results should consequently be interpreted as slice-level performance.

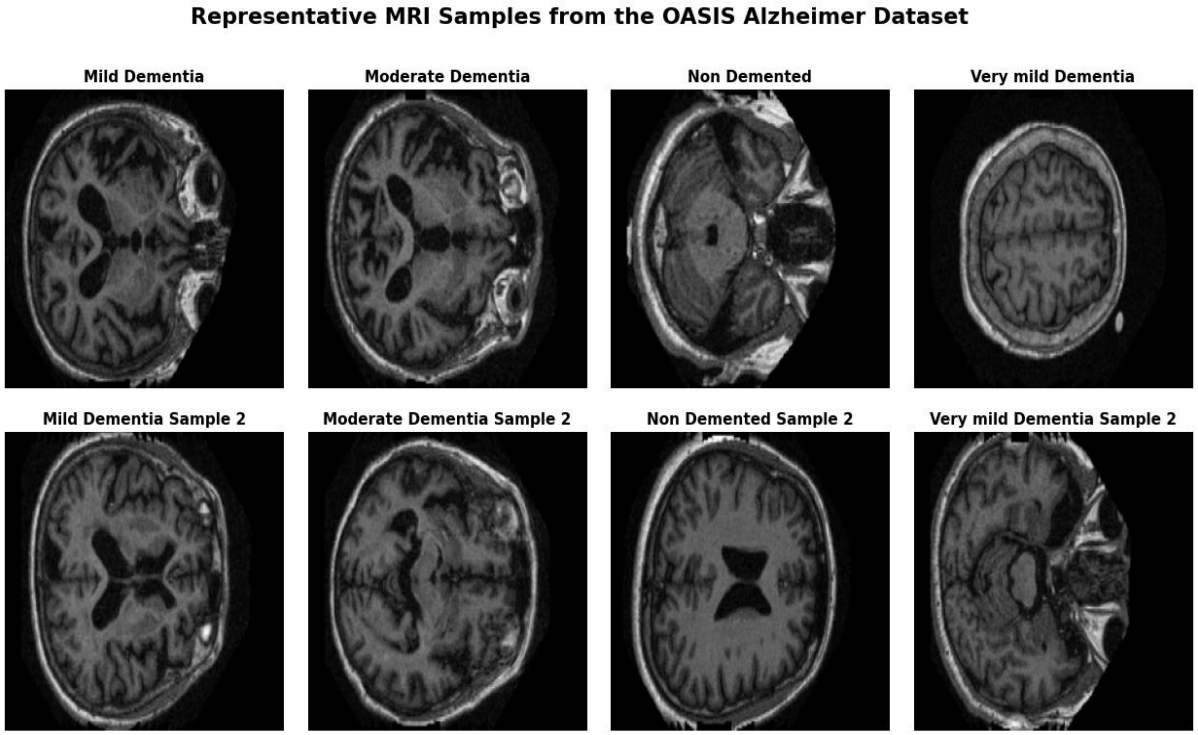


Fig. 2. Representative MRI slices from the four diagnostic categories in the OASIS-derived dataset.

Image Processing and Data Augmentation

Each MRI slice was resized to 224 x 224 pixels to match the input dimensions of the CNN backbone. Pixel values were normalized using the ImageNet mean and standard deviation. During training, random rotation, horizontal flipping, and mild intensity perturbations were applied to increase variation and reduce overfitting. Augmentation was not applied to the validation subset.

For an input MRI slice, the preprocessing pipeline is expressed as:

$$\sim I = N(A(R(I)))$$

where R denotes resizing, A denotes the training-time augmentation operations, and N denotes normalization.

Overview of the Hybrid CNN-Transformer Architecture

The proposed model combines a CNN feature extractor with a Transformer encoder. The CNN learns local anatomical patterns, and the Transformer models contextual relationships among spatial tokens derived from the CNN feature map. The architecture contains four main components:

1. CNN-based feature extraction
2. Tokenization and positional embedding
3. Transformer encoding for global context
4. Feature fusion and multi-class classification

Together, these components allow the classifier to use both fine-scale image structure and broader spatial relationships.

CNN-Based Local Feature Extraction

ResNet-18 was selected as the CNN backbone because it offers a practical balance between representation capacity and computational cost. Its global-pooling and fully connected layers were removed so that the convolutional blocks produce a spatial feature map. For an input image, the CNN output is defined as:

$$F_{cnn} = f_{cnn(I\sim)}, \; F_{cnn} \in R^{\{C\times H'\times W'\}}$$

where f_cnn denotes the convolutional feature extractor, C is the number of output channels, and H' and W' are the reduced spatial dimensions.

A 1 x 1 convolution then projects the CNN channels to the embedding dimension used by the Transformer:

$$F_{proj} = W_p * F_{cnn}$$

where W_p is the learnable projection and * denotes convolution.

Tokenization and Transformer Encoding

The projected feature map is flattened across its spatial dimensions to create a sequence of tokens:

$$T = \{t_1, t_2, \dots, t_N\}, \; t_i \in R^d$$

where N = H' x W' is the number of spatial tokens and d is the embedding dimension.

A learnable classification token is prepended to the sequence, and positional embeddings are added to retain spatial order:

$$Z_0 = [t_{cls}; T] + E_{pos}$$

The resulting sequence is processed by a Transformer encoder composed of multi-head self-attention (MHSA) and position-wise feed-forward networks (FFNs). For layer l, the updates are written as:

$$Z_l = MHSA(Z_{\{l-1\}}) + Z_{\{l-1\}}$$
$$Z_l = FFN(Z_l) + Z_l$$

MHSA allows information to be exchanged across tokens, while the FFN transforms each token representation independently.

After the final encoder layer, the representation associated with the classification token summarizes global contextual information from the MRI slice.

Feature Fusion and Classification

In parallel with the Transformer branch, global average pooling is applied to the original CNN feature map to obtain a compact local representation:

$$v_{cnn} = GAP(F_{proj})$$

The pooled CNN vector and the Transformer's classification-token vector are concatenated:

$$v_{fused} = [v_{cnn} \| z_{cls}]$$

where || denotes vector concatenation.

The fused representation is passed to a fully connected classification head followed by softmax:

$$\hat{y} = Softmax(W_c v_{fused} + b_c)$$

where W_c and b_c are learnable classifier parameters.

Training Objective and Optimization

The model was optimized using categorical cross-entropy:

$$L = -\Sigma_{\{i=1\}}^{K} y_i log(\hat{y}_i)$$

where K = 4 is the number of classes, y_i is the ground-truth indicator for class i, and the corresponding model output is the predicted class probability.

AdamW was used for optimization, with weight decay for regularization. Gradient clipping was applied to limit unstable updates. Training continued for a fixed number of epochs, and the checkpoint with the highest validation accuracy was retained for evaluation.

Methodological Summary
The complete pipeline first extracts a spatial feature map with ResNet-18, then processes projected feature tokens with a Transformer encoder. A classification head combines the pooled CNN representation with the Transformer summary vector. This design provides a direct mechanism for testing whether local and global representations contribute complementary information.

RESULTS
This section reports the experimental configuration, evaluation measures, training behavior, class-wise results, ablation analysis, comparison with published methods, and qualitative Grad-CAM assessment.

Experimental Setup
Experiments were conducted in the Kaggle GPU environment using the four-class OASIS-derived MRI dataset. Images were divided into 80% training and 20% validation subsets using a stratified slice-level split. All slices were resized to 224 x 224 pixels and normalized with ImageNet statistics. Random rotation, horizontal flipping, and mild brightness and contrast variation were used during training. The model was trained for 20 epochs with a batch size of 32 using AdamW, a learning rate of 3 x 10^-4, weight decay of 1 x 10^-4, and categorical cross-entropy loss. The checkpoint with the highest validation accuracy was selected.

Evaluation Metrics
Performance was evaluated using accuracy and class-wise precision, recall, and F1-score. Macro averages were also calculated so that each diagnostic category contributed equally, regardless of its frequency.

1. Accuracy
2. Precision
3. Recall
4. F1-score

For each class, precision, recall, and F1-score are defined as:

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1 = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.
Macro-averaged scores were calculated by taking the unweighted mean across the four classes.

Training Behavior
Figure 3 shows the training-accuracy trajectory over 20 epochs. Accuracy exceeded 94% by the third epoch and reached 99.61% in the final epoch. The gradual increase indicates stable optimization on the training data; generalization is assessed separately using the held-out validation subset.

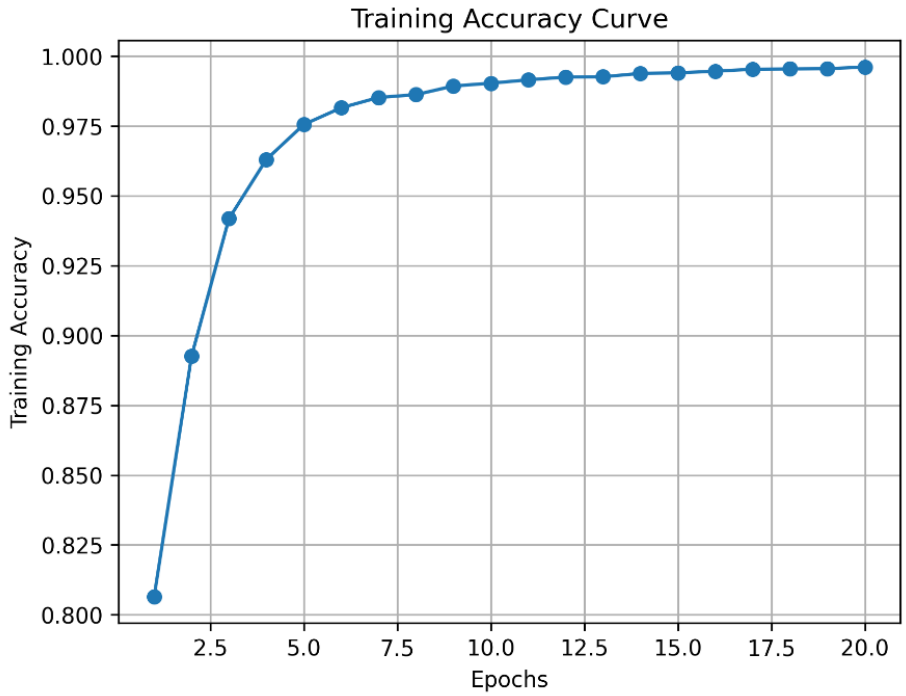


Fig. 3. Training-accuracy curve for the proposed hybrid CNN-Transformer model.

Overall Classification Performance

Table 1 summarizes precision, recall, F1-score, and support for each class on the validation subset.

Table 1. Class-wise performance of the proposed model on the validation subset.

Class	Precision	Recall	F1-score	Support
Mild Demented	1.00	0.99	0.99	1000
Moderate Demented	1.00	1.00	1.00	98
Non Demented	1.00	1.00	1.00	13445
Very Mild Demented	1.00	0.99	1.00	2745
Macro Avg	1.00	0.99	1.00	–
Weighted Avg	1.00	1.00	1.00	–

The model achieved 99.81% validation accuracy. Because several class-wise values round to 1.00 at two decimal places, the support counts and confusion matrix provide additional context for the remaining errors.

Confusion Matrix Analysis

Figure 4 presents the validation confusion matrix.

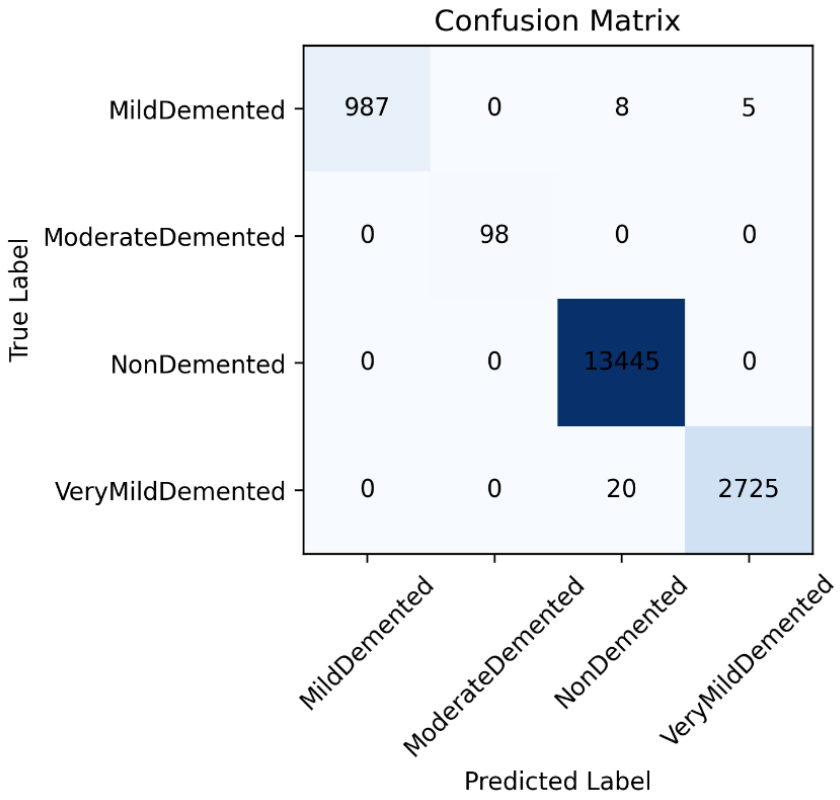


Fig. 4. Validation confusion matrix for the proposed hybrid CNN-Transformer model.

Predictions are concentrated along the main diagonal. The remaining errors occur mainly between the Very Mild Demented and NonDemented categories, with a smaller number involving Mild Demented cases. Distinguishing adjacent stages is expected to be more difficult than separating advanced disease from non-demented scans; however, clinical plausibility cannot be established from the confusion matrix alone.

Ablation Study

To examine the contribution of each architectural component, the CNN branch, Transformer branch, and fusion mechanism were evaluated in controlled model variants.

Table 2. Ablation results for the CNN, Transformer, and fusion components.

Model Configuration	Accuracy	F1-score
CNN (ResNet-18) only	96.8%	0.96
Transformer only	95.9%	0.95
CNN + Transformer (no fusion)	98.7%	0.98
Proposed Hybrid Model	99.81%	0.99

The ResNet-18 baseline reached 96.8% accuracy, while the Transformer-only model reached 95.9%. Combining the two branches without the proposed fusion increased accuracy to 98.7%, and the complete model reached 99.81%. Under this experimental protocol, the results support the view that the local CNN representation and global Transformer representation provide complementary information.

Comparison with Published Methods

Table 3 places the proposed result alongside recently reported MRI-based AD classification methods. The comparison is descriptive because the studies differ in dataset composition, preprocessing, class definitions, and evaluation protocol.

Table 3. Comparison with recently reported MRI-based Alzheimer's disease classification methods.

Reference	Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score	Split strategy
Helaly et al., 2022 [3]	Custom deep CNN	OASIS (Kaggle)	96.25	—	—	—	80/20 slice-level
Maqsood et al., 2019 [11]	AlexNet (transfer learning)	OASIS	92.85	74.27	74.27	0.83	Slice-level
Hu et al., 2023 [4]	VGG-TSwinformer	ADNI + OASIS	97.40	97.10	97.30	0.97	Subject-level
Odusami et al., 2023 [9]	ViT + pixel-level fusion	OASIS (Kaggle)	95.40	95.20	95.30	0.95	Slice-level
ResNet101-ViT, 2024 [4]	ResNet101 + ViT (hybrid)	OASIS (Kaggle)	98.70	98.60	98.70	0.987	Slice-level
Latif et al., 2025 [5]	Deep ensemble + Transformer	OASIS (Kaggle)	99.10	99.00	99.10	0.991	Slice-level
Dardouri, S., 2025 [13]	Multi-branch deep CNN	OASIS (Kaggle)	99.68	—	—	—	Slice-level
Kaur et al., 2025 [14]	Ensemble (ResNet50 + EfficientNet-B3)	OASIS (Kaggle)	99.32	99.00	99.00	0.990	Slice-level
Proposed (Ours)	**Hybrid CNN (ResNet-18) + Transformer encoder**	**OASIS (Kaggle)**	**99.81**	**100.00**	**99.00**	**0.99**	**80/20 slice-level**

Note: Values are reproduced from the cited studies. Differences in preprocessing, class distribution, and especially slice-level versus subject-level splitting prevent a direct ranking of the methods.

The proposed model records the highest accuracy among the slice-level results listed in Table 3. This comparison suggests competitive performance under a similar Kaggle-based setting, but it does not establish superiority over models evaluated with subject-level separation or external test data.

Explainability Analysis Using Grad-CAM

Grad-CAM was used to visualize image regions that contributed strongly to the model's predictions. Figure 5 pairs representative MRI slices with their heatmaps across the four classes. The activation patterns are spatially localized and vary across examples, including the early-stage categories. These visualizations offer a qualitative check on model behavior, but they should not be interpreted as confirmation that the

highlighted regions are clinically diagnostic. Expert review and region-based quantitative analysis would be needed for that conclusion.

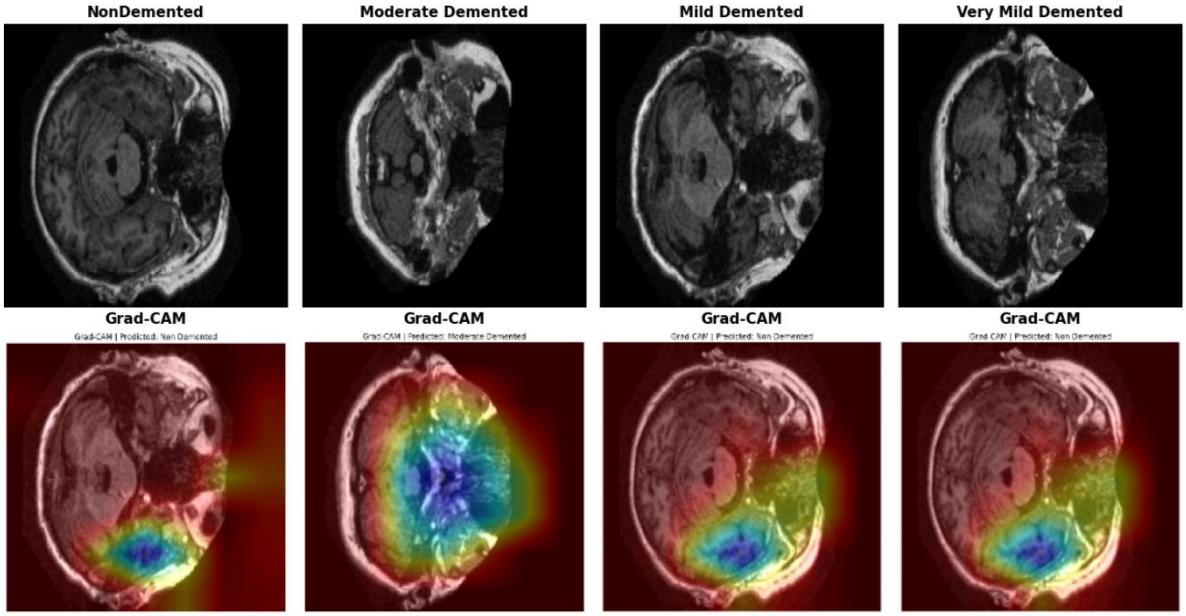


Fig. 5. Grad-CAM visualizations for representative predictions from the four diagnostic categories.

The first row contains the original MRI slices, and the second row shows the corresponding Grad-CAM heatmaps for the predicted classes.

Discussion

Within the stated slice-level protocol, the hybrid model performed better than each ablated configuration. The CNN branch provides a compact description of local anatomy, whereas the Transformer branch relates features across the spatial map. Their fusion appears to account for the improvement from 96.8% for the CNN-only baseline and 95.9% for the Transformer-only baseline to 99.81% for the complete model. The remaining errors are concentrated in neighboring diagnostic categories, where image differences are likely to be less pronounced.

The result should nevertheless be interpreted in light of the evaluation design. The validation set is dominated by NonDemented slices, while the Moderate Demented category has comparatively few examples. More importantly, the absence of subject identifiers required a slice-level split, so anatomically similar slices from the same person may occur in both subsets. The Grad-CAM analysis provides useful qualitative evidence about model attention, but no radiologist annotations or independent clinical assessment were available to validate those regions.

Limitations

This study has several limitations. First, the slice-level split may produce optimistic performance estimates because slices from one subject can potentially appear in both training and validation subsets. Second, the experiment uses 2D slices from a single Kaggle dataset and does not include an independent test cohort. Third, the class distribution is highly imbalanced, particularly for the Moderate Demented category. Fourth, checkpoint selection and reporting use the same validation subset. Finally, Grad-CAM is evaluated qualitatively and has not been compared with expert annotations. These factors limit conclusions about patient-level generalization and clinical utility.

CONCLUSIONS AND FUTURE WORK

This paper presented a hybrid CNN-Transformer framework for four-class classification of Alzheimer's disease from 2D MRI slices. ResNet-18 extracts local anatomical features, a Transformer encoder models

relationships among spatial tokens, and a fusion layer combines both representations for prediction. On the stratified 80/20 slice-level split, the model achieved 99.81% validation accuracy and a macro F1-score of approximately 0.99. The ablation results indicate that both branches contribute to performance under the same protocol, while the confusion matrix shows that most residual errors involve adjacent early-stage categories. Grad-CAM provides a qualitative view of the regions influencing individual predictions.

The reported performance is encouraging but should not be treated as evidence of patient-level or clinical generalization. Future work should evaluate the model on subject-separated datasets such as ADNI and OASIS-3, include an independent external test cohort, and extend the analysis to 3D volumes. Multimodal inputs, including PET, cognitive scores, and clinical variables, may provide additional diagnostic information. Further work should also report computational cost, assess calibration and robustness, and validate explanations with clinical experts before considering deployment in decision-support settings.

REFERENCES

[1] M. Ali, M. F. Mushtaq, S. N. Khosa, N. Kiran, H. Arshad, and U. Akram, "Deep learning-based bone age growth disease detection (BAGDD) using RSNA radiographs," International Journal of Advanced Computer Science and Applications, vol. 16, no. 6, Jun. 2025.

[2] M. G. Alsubaie, S. Luo, and K. Shaukat, "Alzheimer's disease detection using deep learning on neuroimaging: A systematic review," Machine Learning and Knowledge Extraction, vol. 6, no. 1, pp. 464-505, Feb. 2024.

[3] H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of Alzheimer's disease," Cognitive Computation, vol. 14, no. 5, pp. 1711-1727, Sep. 2022.

[4] Z. Hu, Z. Wang, Y. Jin, and W. Hou, "VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction," Computer Methods and Programs in Biomedicine, vol. 229, Art. no. 107291, Feb. 2023.

[5] S. Latif, N. U. Islam, Z. Uddin, K. M. Cheema, S. S. Ahmed, and M. F. Khan, "Deep ensemble learning with Transformer models for enhanced Alzheimer's disease detection," Scientific Reports, vol. 15, Art. no. 24720, Jul. 2025.

[6] S. Jamshidiha, A. Rezaee, F. Hajati, M. Golzan, and R. Chiong, "An explainable Transformer model for Alzheimer's disease detection using retinal imaging," Scientific Reports, vol. 15, Art. no. 26773, Jul. 2025.

[7] P. He, Z. Shi, Y. Cui, R. Wang, D. Wu, and the Alzheimer's Disease Neuroimaging Initiative, "A spatiotemporal graph Transformer approach for Alzheimer's disease diagnosis with rs-fMRI," Computers in Biology and Medicine, vol. 178, Art. no. 108762, Aug. 2024.

[8] Y. Tang, X. Xiong, G. Tong, Y. Yang, and H. Zhang, "Multimodal diagnosis model of Alzheimer's disease based on improved Transformer," Biomedical Engineering Online, vol. 23, no. 1, Art. no. 8, Jan. 2024.

[9] M. Odusami, R. Maskeliunas, and R. Damasevicius, "Pixel-level fusion approach with Vision Transformer for early detection of Alzheimer's disease," Electronics, vol. 12, no. 5, Art. no. 1218, Mar. 2023.

[10] V. Mubonanyikuzo, H. Yan, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Detection of Alzheimer disease in neuroimages using Vision Transformers: Systematic review and meta-analysis," Journal of Medical Internet Research, vol. 27, Art. no. e62647, Feb. 2025.

[11] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, and O.-Y. Song, "Transfer learning-assisted classification and detection of Alzheimer's disease stages using 3D MRI scans," Sensors, vol. 19, no. 11, Art. no. 2645, 2019.

[12] R. Kadri, M. Bouguessa, I. Rekik, et al., "Early detection of Alzheimer's disease progression stages using a hybrid of CNN and Transformer encoder models," Scientific Reports, vol. 15, Art. no. 16482, 2025.

[13] S. Dardouri, "An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1563016, 2025.

[14] A. Kaur, F. S. Alshammari, A. U. Rehman, and S. Bharany, "Intelligent Alzheimer's diagnosis and disability assessment: Robust medical imaging analysis using ensemble learning with ResNet-50 and EfficientNet-B3," Frontiers in Medicine, vol. 12, Art. no. 1619228, 2025.