A Hybrid Vision Transformer for T1-Weighted MRI-Based Alzheimer's Disease Staging with Biomarker Fusion

Sonali Deshpande
Computer Science & Engineering Department, MIT SoC, MIT Art, Design and Technology University, Pune, 412201, Maharashtra, India
sonali.deshpande@mituniversity.edu.in (corresponding author)

Nilima Kulkarni
Computer Science & Engineering Department, MIT SoC, MIT Art, Design and Technology University, Pune, 412201, Maharashtra, India
nilima.kulkarni@mituniversity.edu.in

Received: 29 December 2025 | Revised: 20 January 2026 and 3 February 2026 | Accepted: 7 February 2026

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.17151*

ABSTRACT

Magnetic Resonance Imaging (MRI) with deep learning is widely applied for computer-aided diagnosis of Alzheimer's Disease (AD); however, existing models generally struggle with small and imbalanced datasets and often fail to thoroughly utilize anatomically meaningful biomarkers, which are essential for the detection of AD in the early stages. In this work, we tackle these shortcomings by introducing Hybrid Transformer for Alzheimer's Diagnosis (HyTraAD), a hybrid transformer-based model for 4-stage AD classification, including Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Mild Cognitive Impairment (MCI), and AD, from T1-weighted structural MRIs. Our approach combines a Residual Network (ResNet) 50 feature extractor with a light-weight Vision Transformer (ViT) encoder and directly fuses three volumetric biomarkers: hippocampal volume, temporal parietal cortical thickness, and ventricular volume into the learned representation. To address dataset imbalance and improve robustness, a noise-tolerant preprocessing pipeline is introduced, combining Tomek Links for removing borderline samples with the Synthetic Minority Over-sampling Technique (SMOTE) for balancing underrepresented classes. The model was evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort consisting of 1,850 subjects. Experimental results demonstrate that HyTraAD achieves 99.81% overall accuracy, a macro F1-score of 0.99, and an EMCI recall of 0.95 on the test set, outperforming recent hybrid Convolutional Neural Network (CNN)-ViT architectures such as Hybrid ResNet-50 + ViT (RViT) and Visual Geometry Group (VGG)-based TSwinformer. Ablation studies further confirm that both biomarker integration and the proposed Tomek Links-SMOTE preprocessing strategy contribute significantly to the performance improvements, particularly in enhancing sensitivity to EMCI cases. Collectively, the results demonstrate that HyTraAD provides a flexible and interpretable framework for MRI-based staging of AD, which is promising for future deployment in clinically oriented decision-support systems, particularly in multi-center and multimodal diagnostic settings.

Keywords-Alzheimer's Disease (AD); attention mechanism; residual feature encoding; cognitive impairment detection; deep learning; Vision Transformer (ViT)

I. INTRODUCTION AND RELATED WORK

Progressive neurodegenerative Alzheimer's Disease (AD) is the most prevalent form of dementia, with more than 55 million people currently living with dementia globally, and about 60-70% of them being affected by AD, according to the World Health Organization. The global burden of dementia is projected to exceed 130 million by 2050, posing significant challenges for patients, caregivers, and healthcare systems [1]. The early stage of AD, known clinically as Mild Cognitive Impairment (MCI), is characterized by the presence of subtle memory complaints and constitutes a crucial period during which therapeutic approaches can be used to slow the progression of the disease and enhance quality of life [2, 3].

Neuroimaging and biomarker analysis have been increasingly used to study the progression from MCI to AD in clinical research. Authors in [4] proposed a biomarker-based model describing the temporal evolution of AD pathology, characterized by sequential amyloid-β accumulation, tau pathology, and neurodegeneration, which has since been

refined into the AT(N) research framework for biological classification of AD. Subsequent studies have demonstrated that accurate identification of prodromal stages, such as MCI, enables prediction of disease progression and supports informed clinical decision-making and therapeutic trial design [5]. These advances highlight the importance of integrating advanced computational approaches with established clinical and neuroimaging biomarkers to improve diagnostic accuracy and enable reliable disease monitoring in AD [6].

For the early detection of AD and longitudinal monitoring, the most commonly employed tool used is Magnetic Resonance Imaging (MRI), because neuroanatomical features of AD-related neurodegeneration, including hippocampal atrophy, ventricular enlargement, and cortical thinning, can be observed. Nevertheless, visual interpretation of MRI is time-consuming, inconsistent between different interpreters, and challenging to implement in clinical practice. Consequently, artificial intelligence and deep learning techniques have been extensively explored to automate and enhance the diagnosis of AD using neuroimaging data, demonstrating promising performance in both classification and prognostic prediction tasks [6, 7]. Convolutional Neural Networks (CNNs) are widely used for MRI-based AD classification due to their ability to learn hierarchical spatial features. Nonetheless, CNNs primarily capture local spatial patterns and typically require large, balanced datasets to generalize effectively, which poses significant challenges when applied to limited and imbalanced AD datasets [8, 9].

Transformer-based Feature Attention (TFA) models, in contrast, leverage self-attention mechanisms to learn long-range dependencies across the entire brain volume, unlike CNNs that focus on limited spatial patterns through a predefined window-size receptive field [10]. This ability is particularly useful for AD, which causes neurodegeneration in non-adjacent brain regions (e.g., hippocampus, temporal parietal cortices) within the ventricular system. Specifically, Vision Transformer (ViT) models can effectively capture global atrophy patterns and nuanced inter-regional interactions in early and progressing AD by dedicating each image-patch to all other patches [11, 12], thus offering a more accurate representation of Alzheimer's-related brain alterations. Additionally, hybrid architectures combining CNNs and ViT have shown improved performance for AD classification from MRI; however, most existing approaches do not explicitly integrate anatomically informed biomarkers such as hippocampal volume, cortical thickness, or ventricular enlargement, which are critical for differentiating early disease stages [13, 14]. Furthermore, important practical issues such as class imbalance, presence of borderline or noisy samples, and under-representation of early-stage cases, such as Early Mild Cognitive Impairment (EMCI), remain insufficiently addressed in many current methods.

To address these limitations, the present work introduces Hybrid Transformer for Alzheimer's Diagnosis (HyTraAD), a transformer-based framework that combines anatomically informed biomarkers with a hybrid Residual Network (ResNet)-ViT model for four-stage AD classification from T1-weighted MRI scans. Without relying on structural segmentation, HyTraAD seeks to augment clinically meaningful brain areas (e.g., the hippocampus and the temporal and parietal-cortex) by combining convolutional feature maps from 3D MRI with biomarker vectors learned to model hippocampal atrophy, cortical thinning, and ventricular enlargement. The preprocessing pipeline uses Tomek Links for removing noisy samples and Synthetic Minority Over-sampling Technique (SMOTE) for treatment of class imbalance that enhances the robustness to underrepresented classes like EMCI. The proposed method achieves 99.81% of test accuracy and shows excellent generalization performance over AD, MCI, EMCI, and Cognitively Normal (CN) categories [15].

A. CNN-Based and Hybrid Architectures for AD Classification

CNN-based techniques have long dominated AD detection due to their efficiency in acquiring hierarchical spatial features from brain MRI scans. Authors in [16] utilized a CNN model on Diffusion Tensor Imaging (DTI) data and obtained excellent binary AD versus control classification performance. Several CNN-based approaches have demonstrated strong performance for AD classification using structural MRI, particularly through the use of 2D/3D slice representations, data augmentation, and transfer learning strategies [17, 18]. However, CNNs have inherent limitations in modeling long-range spatial dependencies across brain regions, which is a key challenge in understanding complex neurological conditions.

In contrast, authors in [19] employed ViTs, which employ self-attention mechanisms to capture both local and long-range spatial dependencies by representing images as sequences of patches rather than pixel grids.

In order to combine their strengths, hybrid models that harness CNNs' local feature extraction and ViTs' global context modeling have also been proposed. Notable examples include EfficientNetV2 + ViT + Generative Adversarial Network (GAN) [20], which employs data augmentation for improved generalization, and Self-supervised Multi-Instance Learning Transformer (SMIL-DeiT) [21], which incorporates multi-instance self-supervised learning for robust early-stage classification. Additionally, Explainable hybrid models like Hybrid ResNet-50 and ViT (RViT) [14] provide clinical insights using attention visualization and Shapley Additive Explanations (SHAP) analysis, achieving 95% accuracy on a four-class AD classification task using the Open Access Series of Imaging Studies (OASIS) dataset.

B. Preprocessing and Class Imbalance Handling

Researchers have also investigated the impact of preprocessing techniques such as skull stripping, intensity normalization, and data balancing to improve model performance. For instance, authors in [22, 23] used data augmentation methods (e.g., CutMix and MixUp) for training ViT to enhance the generalization of the model. However, class imbalance, particularly for minority early-stage cases, remained a persistent issue. Thus, advanced SMOTE variants that help oversample informative minority samples and Tomek Links that remove noisy borderline cases, improving accuracy, precision, and recall, have been suggested.

C. Summary of Key Studies

Table I summarizes transformer-based AD classification approaches, highlighting their preprocessing methods, datasets, reported performance, and limitations. Unlike prior work that relies heavily on image-only features, our method incorporates clinically relevant biomarkers and noise-tolerant preprocessing for robust multi-stage classification.

TABLE I. SUMMARY OF TRANSFORMER-BASED ARCHITECTURES FOR AD CLASSIFICATION

Ref.	Model/Architecture	Preprocessing Techniques	Dataset Used (Classes)	Reported Accuracy	Limitation /Comparison
[6]	Hybrid CNN-ViT	Adaptive Median Filter (AMF) + Laplacian Filter for enhancement.	OASIS MRI Dataset (ND, VMD, MD, AD)	ResNet101-ViT: 98.7% GoogLeNet-ViT: 97.5% ResNet101 alone: 86.5% GoogLeNet alone: 85%	Focuses on image enhancement; no imbalance or biomarker modeling.
[14]	Hybrid-RViT (ResNet-50 + ViT)	Resizing, OASIS preprocessing, and normalization.	OASIS (4-class)	95.0	No explicit biomarker integration or noise handling.
[20]	EfficientNetV2 + ViT + GAN	Modality extraction, Self-Attention GAN.	ADNI + OASIS (AD vs. CN)	96%	Binary classification, no biomarker integration, no noise-aware refinement.
[24]	Ensemble of 4 ViTs (ViT, Swin, DeiT, BEiT)	Reorientation, registration, skull-stripping.	ADNI and OASIS (ND, VMD, MD, AD, EMCI, LMCI, CN)	99.29%	High complexity; lacks noise removal and class balancing.

Bidirectional Encoder Representations from Transformers Pre-Training of Image Transformers (BEiT), Non-Demented (ND), Very Mild Demented (VMD), Mild Demented (MD), Late Mild Cognitive Impairment (LMCI).

II. PROPOSED METHODOLOGY

Figure 1 illustrates the HyTraAD framework, which integrates anatomically informed biomarker features with a lightweight ViT encoder for four-stage, robust classification of AD from structural MRI. The pipeline consists of three main parts: i) data preprocessing and purification, ii) biomarker-driven feature extraction, and iii) hybrid transformer-based encoding and classification.

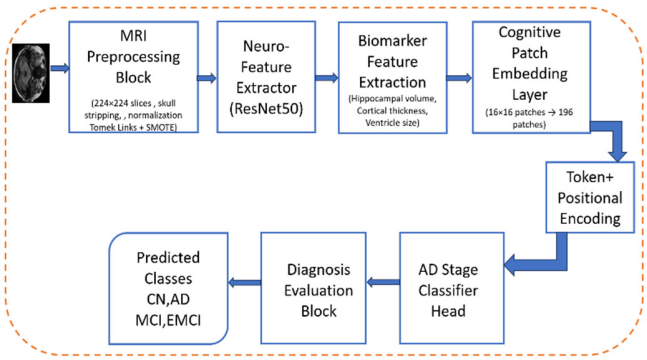

Fig. 1. Overall architecture of the proposed HyTraAD framework.

A. Dataset and Preprocessing

The experiments were conducted using the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [25]. After quality control, which consisted of removing motion-affected scans, incomplete clinical records, or failed skull-stripping, 1,850 subjects from ADNI-1, ADNI-2, and ADNI-GO were included. The cohort consists of 4 diagnostic groups according to the ADNI consensus criteria:

- CN, with 620 subjects.
- EMCI, with 380 subjects.
- MCI, with 550 subjects.
- AD, with 300 subjects.

For each subject, three axial slices were extracted at standardized anatomical locations centered on the hippocampus in T1-weighted Neuroimaging Informatics Technology Initiative (NIfTI) images, yielding 5,550 images.

Furthermore, to ensure data integrity, each individual was represented only once using a unique ADNI participant ID, ensuring no duplicates even across longitudinal scans. This guarantees independence between training, validation, and test sets and prevents data leakage.

It is important to note that even though the model utilizes slice-level inputs during both training and testing, final subject-level diagnoses are obtained via majority voting over slice predictions. This provides the model with the ability to capture local anatomical differences between slices before producing clinically relevant subject-level predictions.

B. Preprocessing Pipeline

The preprocessing pipeline, which was implemented in Python 3.9 with NiBabel, SimpleITK, and SciPy, included:

- Skull stripping brain extraction, which was completed using a community-validated U-Net segmentation model (trained on 500 manually labeled brain masks from ADNI) to strip the skull and non-brain tissues with >98% accuracy.
- Spatial normalization, where all slices were rigid-body registered to the MNI152 1mm template using SimpleITK's affine transform with mutual information metric and 12

degrees of freedom, followed by resampling to 224×224 pixels (isotropic 1 mm resolution).

- Intensity normalization, which consisted of N4 bias field correction (ITK implementation) and z-score normalization:

$$I_{norm} = \frac{I_{raw} - \mu_{brain}}{\sigma_{brain}} \quad (1)$$

where μ_{brain} and σ_{brain} are computed from brain-masked voxels only.

- Quality assurance, where all normalized slices were visually inspected; 47 slices (0.8%) with failed skull-stripping or artifacts were discarded.
- Data splits: To prevent data leakage, subject-level stratified splitting was enforced: i) training set: 1,295 subjects (70%), ii) validation set: 278 subjects (15%), and iii) test set: 277 subjects (15%).

C. Data Refinement: Tomek Links and SMOTE

The raw training set exhibited severe imbalance, with CN having 1,302 slices, EMCI having 798 slices, MCI having 1,155 slices, and AD having 630 slices.

After applying Tomek Links cleaning (Imbalanced-Learn v0.10.1's TomekLinks), 1,247 borderline noisy samples were removed, primarily from CN-EMCI and EMCI-MCI boundaries, reducing ambiguity near class margins.

Then, applying SMOTE (k_neighbors=3) resulted in oversampling EMCI, MCI, and AD classes to 1,300 slices each, achieving near-perfect balance (final training distribution: 5,200 slices). SMOTE interpolations were performed only within the same diagnostic class using MRI slices aligned in MNI space to ensure anatomical plausibility.

Validation and test sets were not augmented to preserve real-world class ratios.

D. Feature Extraction and Biomarker Integration

For feature extraction, a neuro-feature extractor is utilized. Preprocessed MRI slices are fed into a ResNet-50 backbone (ImageNet-pretrained, TensorFlow/Keras v2.13). The final convolutional block (conv5_x) outputs feature maps of size 7×7×2048, which are subsequently reduced via global average pooling into a 2,048-dimensional vector. This vector captures spatial patterns relevant to AD, including cortical thinning, ventricular dilation, and hippocampal atrophy.

In parallel, three clinically validated volumetric biomarkers are extracted from the ADNI UPenn biomarker files (UPENNBIOMK.csv) provided in the ADNI data repository:

- Hippocampal volume (left + right, mm^3).
- Mean cortical thickness (temporal and parietal regions, mm).
- Lateral ventricle volume (mm^3).

To account for inter-individual differences in head size, each biomarker is normalized by Intracranial Volume (ICV) using:

$$\text{Bio}_{norm} = \frac{\text{Bio}_{raw}}{\text{ICV}} \cdot 10^3 \quad (2)$$

The resulting 3-dimensional biomarker vector is concatenated with the ResNet-50 feature vector, producing a 2,051-dimensional hybrid representation (2,048 CNN features + 3 biomarkers). This explicit biomarker integration provides the model with direct anatomical context, addressing a key limitation of pure CNN or ViT approaches.

E. Cognitive Patch Embedding and Transformer Encoding

The 2,051-dimensional hybrid vector is then reshaped into a 43×43×1 2D grid and partitioned into 16×16 non-overlapping patches, yielding 196 patches. Each patch is linearly projected into a 768-dimensional embedding space using a trainable matrix $W_p \in \mathbb{R}^{16^2 \times 768}$. Next, a learnable [CLS] classification token is prepended to the sequence, resulting in 197 tokens. Learnable positional embeddings (197×768) are added to each token to retain both spatial and biomarker positional information.

This token sequence is processed by a Cognitive Transformer Module (CTM) consisting of six identical transformer encoder layers (ViT-Base configuration). Each layer comprises: i) a multi-head self-attention (8 heads, 96-dimensional each), ii) a 3,072-neuron Multilayer Perceptron (MLP) block with Gaussian Error Linear Unit (GELU) activation and 0.1 dropout, iii) LayerNorm applied pre-attention and pre-MLP, and iv) residual skip connections. These layers iteratively refine the feature representations, enabling the model to fuse local CNN features with global biomarker context while suppressing extraneous noise.

F. AD Stage Classifier Head

The final [CLS] token output from the transformer is passed through a two-layer MLP classifier. The first layer maps 768 → 256 neurons (Rectified Linear Unit (ReLU) activation, 0.1 dropout), and the second layer maps 256 → 4 neurons with a softmax activation, producing a probability distribution over the diagnostic categories: CN, EMCI, MCI, and AD.

G. Training Configuration and Hyperparameters

- Software and Hardware Environment: TensorFlow v2.13.0 and Keras v2.13.1 on an NVIDIA RTX 4090 (CUDA v12.1, cuDNN v8.9). Random seeds were fixed using numpy.random.seed (42) and tensorflow.random.set_seed (42) to ensure reproducibility.
- Optimizer and Learning Rate**:** Adam optimizer (beta_1=0.9, beta_2=0.999, epsilon=1e-7) with a cosine annealing schedule from 5e-5 to 1e-6 over 10 epochs.
- Loss Function: Categorical cross-entropy with class weighting to mitigate dataset imbalance (CN=1.0, EMCI=1.2, MCI=1.1, AD=1.0).
- Training Schedule: Batch size of 16 (max fit on 24 GB Video Random Access Memory (VRAM)), maximum 10 epochs with early stopping (patience = 3 epochs) based on validation macro-F1 score.
- Regularization: L2 weight decay (1e-4) applied to all dense layers, dropout of 0.1 in MLP and transformer layers, and

stochastic depth (survival probability = 0.9) in transformer blocks to reduce overfitting. These measures are critical for stabilizing training and ensuring generalizable performance on limited and imbalanced medical imaging datasets.

- Data Augmentation (training set only): Random horizontal flips (p=0.5), rotations ±10°, and zoom range [0.9, 1.1].
- Training time per run on the RTX 4090 was approximately 4.5 hours.

III. RESULTS AND DISCUSSION

To highlight the effectiveness of the proposed HyTraAD framework, we compared its classification performance with a closely related model, Hybrid-RViT [14] (Table II). The proposed HyTraAD outperformed Hybrid-RViT in terms of precision, recall, and F1-score across all classes, including a notably improved recall for the EMCI class (0.95 vs. 0.79 with Hybrid-RViT). For each evaluation metric, robustness was assessed by estimating 95% Confidence Intervals (CI) using non-parametric bootstrapping with 1,000 iterations on the held-out ADNI test set.

Overall, the HyTraAD model achieved an overall classification accuracy of 99.81% (95% CI: [99.4%, 100.0%]), and a macro-averaged F1-score of 0.99 (95% CI: [0.98, 0.99]).

TABLE II. EVALUATION METRICS OF THE PROPOSED SYSTEM AND THE PREVIOUS STUDY

	HyTraAD [Our]			**Hybrid-RViT [14]**		
Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
CN	1.00	1.00	1.00	0.9750	0.9300	0.9500
EMCI	1.00	0.95	0.97	0.9525	0.7875	0.8575
MCI	1.00	1.00	1.00	0.8950	0.9725	0.9300
AD	1.00	1.00	1.00	0.9275	0.9450	0.9375

A detailed class-wise analysis indicates that the CN, MCI, and AD classes were classified almost perfectly across all metrics. A slightly lower recall value (0.95) was observed for the EMCI class, suggesting that a small number of EMCI samples were misclassified. The estimated 95% CI for EMCI recall was [0.91, 0.98], indicating that the reduction in sensitivity is minor and statistically consistent with the inherent overlap between EMCI and MCI stages. Such behavior is generally expected in medical imaging applications, where clinical boundaries between consecutive disease stages are often subtle and overlapping.

Despite this, the F1-score for EMCI remained high at 0.97, indicating that when the model predicts EMCI, it does so with high precision. The remaining misclassifications may be attributed to inter-class similarity, limited representation of EMCI samples, or overlapping neuroanatomical patterns in MRI scans. Importantly, the weighted average F1-score of 1.00 demonstrates that the model maintains highly consistent performance even under class imbalance conditions.

These results highlight the model's robustness and its potential utility in real-world clinical settings, where high sensitivity and specificity are critical for early diagnosis. Furthermore, the hybrid use of convolutional and transformer-based attention mechanisms appears to effectively capture both local structural features and global contextual patterns from MRI scans, offering a comprehensive representation of the underlying neuropathology.

As an additional validation check, model predictions were evaluated under a label permutation test, where diagnostic labels in the test set were randomly shuffled. Under this random labeling, the accuracy dropped to 24.9%, which is close to the 25% chance level for four classes, confirming that HyTraAD is learning meaningful disease-related patterns rather than memorizing noise.

Furthermore, a McNemar's statistical test was conducted on paired predictions from HyTraAD and Hybrid-RViT using the ADNI test set. The results showed that the improvement in EMCI detection was statistically significant ($p < 0.001$), while the overall accuracy improvement was also significant ($p = 0.01$). These findings indicate that HyTraAD provides better sensitivity for early-stage disease detection.

A. *Training Dynamics*

Figure 2 illustrates the training and validation accuracy curves across 10 epochs. During the initial epochs, the training accuracy increases rapidly as the model adapts to the data. After the third epoch, both curves stabilize near the maximum value, indicating that the model has learned generalized representations of the input data, while the close alignment between the training and validation curves suggests minimal overfitting and strong generalization capability.

Figure 3 presents the training and validation loss curves. The training loss decreases rapidly during the early epochs and then gradually converges. The validation loss behaves similarly, remaining relatively low and stable after the initial decline. The consistency between these curves indicates that the training process is stable and well-optimized.

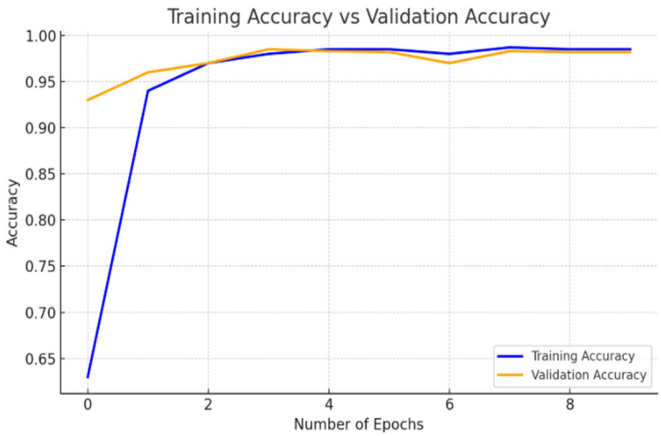

Fig. 2. Training and validation accuracy curve across 10 epochs.

B. *Confusion Matrix and Ablation Study*

The confusion matrix in Figure 4 illustrates how the model provides accurate four-stage classification of AD from MRI, with the vast majority of the samples being correctly assigned

to their true class, with only a small number of misclassifications between CN and EMCI cases.

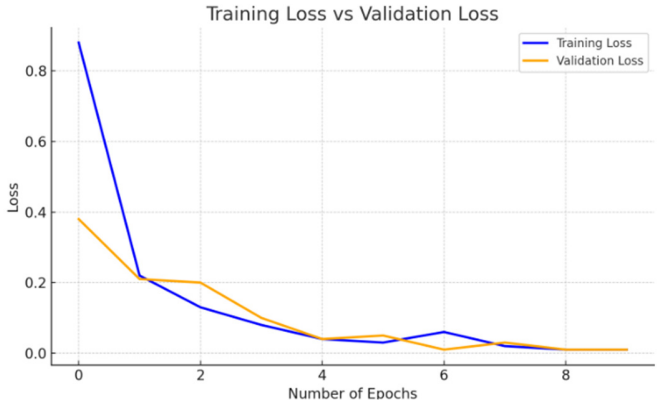

Fig. 3. Training and validation loss curve across 10 epochs.

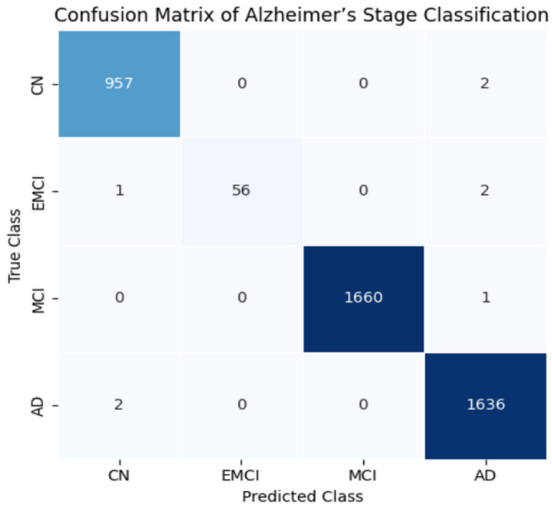

Fig. 4. Model classification results for Alzheimer's stages using the HyTraAD framework.

To further analyze the contribution of individual components (biomarker integration, Tomek Links, and SMOTE), ablation experiments were conducted by sequentially removing one key element of the proposed framework:

- Removing biomarker integration reduced EMCI recall from 0.95 to 0.91.
- Removing Tomek Links cleaning reduced the macro F1-score from 0.99 to 0.97.
- Training without SMOTE balancing reduced EMCI recall to 0.88.

These results demonstrate that biomarker-aware features and imbalance-handling techniques both play a critical role in improving classification performance. The complete HyTraAD configuration, combining biomarkers, Tomek Links cleaning, and SMOTE balancing, achieves the highest overall accuracy and strongest EMCI sensitivity.

C. Comparison with State-of-the-Art Methods

Table III compares the HyTraAD framework with previous state-of-the-art ADNI-based AD classification models from the literature.

TABLE III. PERFORMANCE COMPARISON OF HYTRAAD AND EXISTING METHODS ON THE ADNI DATASET

Model Variant	Bio mar kers	Tomek Links	SMO TE	Overall Accuracy	Classification Task
3D CNN [7]	✗	✗	✗	99.20%	AD vs HC (binary)
AHANet [9]	✗	✗	✗	98.53%	CN vs MCI vs AD (3-class)
CNN (DTI + MRI) [16]	✗	✗	✗	93.50%	HC vs AD (binary)
3D CNN [17]	✗	✗	✗	89.47%	AD vs MCI vs HC (3-class)
ViT (Swin) [24]	✗	✗	✗	99.52%	CN vs EMCI vs LMCI vs AD (4-class)
HyTraAD (proposed)	✓	✓	✓	99.81%	CN vs EMCI vs MCI vs AD (4-class)

Adaptive Hybrid Attention Network (AHANet), Healthy Controls (HC)

While previous CNN- and transformer-based approaches successfully learn spatial and global representations from MRI data, many focus on binary or three-class tasks. In contrast, HyTraAD integrates a hybrid CNN-Transformer architecture with explicit biomarker features and imbalance-aware preprocessing, enabling improved discrimination of the clinically challenging EMCI and MCI stages within a four-class AD classification framework.

IV. CONCLUSION

In this work, we propose a novel deep learning approach for the classification of Alzheimer's Disease (AD) from Magnetic Resonance Imaging (MRI), designed to capture both local structural features and global contextual patterns associated with neurodegeneration. By integrating convolutional feature extraction with transformer-based attention mechanisms, along with anatomically informed biomarker features and imbalance-aware preprocessing, the proposed framework effectively addresses several limitations commonly encountered in medical imaging datasets, including class imbalance, noisy samples, and subtle early-stage anatomical variations.

Experimental evaluation on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrated that the proposed Hybrid Transformer for Alzheimer's Diagnosis (HyTraAD) achieves 99.81% classification accuracy across four diagnostic stages. Overall, the proposed approach is a notable advance over prior approaches and can serve as a strong foundation for further advanced research in automatic diagnosis of AD.

Nevertheless, a limitation of the study the model was tested on the relatively homogeneous ADNI dataset with standardized imaging protocols, which may restrict applicability for external validation in heterogeneous clinical environments

Future work will focus on evaluating the proposed framework on multi-center clinical datasets to further assess its robustness and real-world applicability. Additionally, incorporating longitudinal imaging data and multimodal biomarkers, such as Positron Emission Tomography (PET) imaging or clinical cognitive scores, could further increase the clinical relevance of the proposed model.

REFERENCES

[1] H. Givian, J.-P. Calbimonte, and the Alzheimer's Disease Neuroimaging Initiative, "Early diagnosis of Alzheimer's disease and mild cognitive impairment using MRI analysis and machine learning algorithms," Discover Applied Sciences, vol. 7, no. 1, Dec. 2024, Art. no. 27, https://doi.org/10.1007/s42452-024-06440-w.

[2] Alzheimer's Disease International, "World Alzheimer Report 2019: Attitudes to dementia", Alzheimer's Disease International (ADI), London, United Kingdom, Sept. 2019.

[3] Alzheimer's Association, "2019 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, Mar. 2019, https://doi.org/10.1016/j.jalz.2019.01.010.

[4] C. R. Jack *et al.*, "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, Apr. 2018, https://doi.org/10.1016/j.jalz.2018.02.018.

[5] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild Cognitive Impairment: Clinical Characterization and Outcome," *Archives of Neurology*, vol. 56, no. 3, Mar. 1999, Art. no. 303, https://doi.org/10.1001/archneur.56.3.303.

[6] 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*, vol. 11, Aug. 2019, Art. no. 220, https://doi.org/10.3389/fnagi.2019.00220.

[7] S. Basaia *et al.*, "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage: Clinical*, vol. 21, 2019, Art. no. 101645, https://doi.org/10.1016/j.nicl.2018.101645.

[8] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, https://doi.org/10.1016/j.media.2017.07.005.

[9] T. Illakiya, K. Ramamurthy, M. V. Siddharth, R. Mishra, and A. Udainiya, "AHANet: Adaptive Hybrid Attention Network for Alzheimer's Disease Classification Using Brain Magnetic Resonance Imaging," *Bioengineering*, vol. 10, no. 6, June 2023, Art. no. 714, https://doi.org/10.3390/bioengineering10060714.

[10] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it Time to Replace CNNs with Transformers for Medical Images?" arXiv, 2021, https://doi.org/10.48550/ARXIV.2108.09038.

[11] K. Kawadkar, "Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification." arXiv, 2025, https://doi.org/10.48550/ARXIV.2507.21156.

[12] Y. Shen *et al.*, "MoViT: Memorizing Vision Transformers for Medical Image Analysis." arXiv, 2023, https://doi.org/10.48550/ARXIV.2303.15553.

[13] A. Muhammad, Q. Jin, O. Elwasila, and Y. Gulzar, "Hybrid Deep Learning Architecture with Adaptive Feature Fusion for Multi-Stage Alzheimer's Disease Classification," *Brain Sciences*, vol. 15, no. 6, June 2025, Art. no. 612, https://doi.org/10.3390/brainsci15060612.

[14] H. Yan, V. Mubonanyikuzo, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Hybrid-RViT: Hybridizing ResNet-50 and Vision Transformer for Enhanced Alzheimer's disease detection," *PLOS ONE*, vol. 20, no. 2, Feb. 2025, Art. no. e0318998, https://doi.org/10.1371/journal.pone.0318998.

[15] S. Khanapur, J. S. Nayak, B. S. Rajeshwari, M. Namratha, C. B. Bharadwaj, and R. Bhardwaj, "SHAP-Based Explainability for Local and Global Insights in Alzheimer's Detection," *Engineering, Technology & Applied Science Research*, vol. 16, no. 1, pp. 30940–30947, Feb. 2026, https://doi.org/10.48084/etasr.13932.

[16] E. N. Marzban, A. M. Eldeib, I. A. Yassine, Y. M. Kadah, and for the Alzheimer's Disease Neurodegenerative Initiative, "Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks," PLOS ONE, vol. 15, no. 3, Mar. 2020, Art. no. e0230409, https://doi.org/10.1371/journal.pone.0230409.

[17] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks." arXiv, 2015, https://doi.org/10.48550/ARXIV.1502.02506.

[18] J. Wen *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, July 2020, Art. no. 101694, https://doi.org/10.1016/j.media.2020.101694.

[19] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, 2020, https://doi.org/10.48550/ARXIV.2010.11929.

[20] R. Kadri, B. Bouaziz, M. Tmar, and F. Gargouri, "Multimodal deep learning based on the combination of EfficientNetV2 and ViT for Alzheimer's disease early diagnosis enhanced by SAGAN data augmentation," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 14, pp. 313–325, May 2022.

[21] M. Baniata, S. Abuowaida, M. Aljaidi, M. Kharabsheh, A. Alsarhan, and A. A. Alsuwaylimi, "A Multi-Modal Attention-Guided Network for Alzheimer's Disease Classification Using Deep Learning," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27150–27158, Oct. 2025, https://doi.org/10.48084/etasr.12510.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002, https://doi.org/10.1613/jair.953.

[23] M. B. Khatooni and M. Soryani, "EffNetViTLoRA: An Efficient Hybrid Deep Learning Approach for Alzheimer's Disease Diagnosis." arXiv, Aug. 2025, https://doi.org/10.48550/arXiv.2508.19349.

[24] N. Shaffi, V. Viswan, and M. Mahmud, "Ensemble of vision transformer architectures for efficient Alzheimer's Disease classification," *Brain Informatics*, vol. 11, no. 1, Dec. 2024, Art. no. 25, https://doi.org/10.1186/s40708-024-00238-7.

[25] C. R. Jack *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, Apr. 2008, https://doi.org/10.1002/jmri.21049.