Multi-Stage Alzheimer's Disease Detection: A Comparative Study of CNN, ResNet50, and Hybrid ResNet50+ViT with Explainable AI

1st Dr. Manish Rai
dept. Artificial Intelligence and Machine Learning
Manipal University Jaipur
Jaipur, India

2nd Mr. Abhay Sharma
dept. of Artificial Intelligence and Machine Learning
Manipal University Jaipur
jaipur, India

***Abstract*—Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia worldwide. In 2021, 57 million people had dementia worldwide. The modest morphological changes across adjacent dementia stages make accurate automated staging of AD severity from Magnetic Resonance Imaging (MRI) difficult despite being clinically crucial.In this study, a systematic comparative study of three deep learning architectures, namely a Custom Convolutional Neural Network (CNN) trained from scratch, a fine-tuned ResNet50 with selective layer freezing and layer-wise different learning rates, and a Hybrid ResNet50+Vision Transformer (HybridRViT) for four-class Alzheimer's disease classification across Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented stages. each model was trained on 33,984 augmented MRI images and tested on a original(without augmented) test set of 6,400 images under identical experimental conditions using PyTorch on an NVIDIA Tesla T4 GPU. Experimental results reveals that ResNet50 achieves the highest performance across all important metrics, achieve 95.58% accuracy, with macro F1-Score 0.9708, macro AUC 0.9972, and Quadratic Weighted Kappa 0.9658, outperforming the Custom CNN (89.03%) and HybridRViT (92.84%). To enhance transparency and human-understandablility of the best model, three gradient-based Explainable AI (XAI) techniques namely, Grad-CAM, Grad-CAM++, and Score-CAM are applied to the final convolutional layer, generating heatmaps that consistently localize neuro-anatomically meaningful regions across all four dementia stages. The results establish ResNet50 with targeted fine-tuning as the optimal architecture for MRI-based Alzheimer's staging at the current data scale, while the multi-method XAI analysis provides the clinical transparency required for trustworthy deployment.**

***Index Terms*—Alzheimer's Disease, Multi-Stage Classification, Deep learning, Convolutional Neural Network, ResNet50, Vision Transformer, Hybrid Architecture, Transfer Learning, Explainable Artificial Intelligence, Grad-CAM, Grad-CAM++, Score-CAM.**

I. Introduction

Alzheimer's disease (AD) is a progressive, irreversiable brain disorder and the most common cause of dementia worldwide, which is around 60–70% of all dementia cases. The World Health Organization (WHO) estimates that over 55 million people worldwide suffer from dementia, with approximately 10 million new cases diagnosed annually. By 2050, this number is expected to increase to 139 million [1]. From sub-clinical alterations through mild cognitive impairment to complete dementia, AD is characterised clinically by a steady, irreversible loss of brain memory, language, reasoning abilities, and basic everyday tasks. In the absence of a curative treatment, timely and accurate staging of disease severity is the single most impactful clinical decision, it determines eligibility for disease-modifying interventions, informs care planning, and enables the monitoring of therapeutic response in clinical trials.

In the last decade, deep learning and convolutional neural networks (CNNs) have demonstrated revolutionary potential in medical image analysis, reaching radiologist-level performance on a wide variety of diagnostic tasks, including tumour detection, diabetic retinopathy grading, and pathology classification. In the Alzheimer's detection fields, CNN-based approaches ranging from lightweight custom architectures to deep transfer learning models have reported high classification accuracies on benchmark MRI datasets. More recently, the Vision Transformer (ViT), which replaces spatial convolutions with global self-attention over image patch sequences, has attracted considerable research interest in medical imaging for its capacity to model long-range spatial dependencies that CNNs are structurally constrained from capturing. Hybrid architectures that combine CNN feature extractors and Transformer encoders have emerged as a promising design paradigm, aiming to exploit the complementary inductive biases of the two families.

Despite this previous studies, there are three critical gaps in the literature. First, no past study has conducted a rigorous, controlled three-way evaluation of a custom CNN trained from scratch, a fine-tuned ResNet50, and a Hybrid ResNet50+ViT model under identical training conditions, preprocessing pipelines, and evaluation protocols on the same dataset and test partition. Second, existing hybrid CNN-ViT studies for Alzheimer's classification predominantly employ ResNet101 or GoogLeNet as the CNN backbone, leaving the specific ResNet50+ViT combination underexplored despite ResNet50's widespread practical adoption. Third, while gradient-based Class activation mapping methods Grad-CAM (Gradient-weighted Class Activation Mapping), Grad-CAM++ (Generalized Gradient-weighted Class Activation Mapping), and Score-CAM (Score-Weighted Class Activation Mapping) have each been applied individually to medical imaging tasks,

no prior work applies all three simultaneously to the same Alzheimer's classification model, preventing direct comparative assessment of their clinical localization quality and practical complementarity.

To address these gaps, this study proposes a unified comparative deep learning framework evaluated on the publicly available Augmented Alzheimer MRI Dataset [2], comprising 33,984 augmented training images and 6,400 original test images across four stages: non-demented, very mild demented, mild demented, and moderate demented. All three architectures are trained using the AdamW optimizer with layer-wise learning rates, Cross-Entropy Loss, and ReduceLROnPlateau scheduling on an NVIDIA Tesla T4 GPU under identical hyperparameter settings. Performance is evaluated using accuracy, macro and weighted Precision(positive predictive value), Recall(Sensitivity), F1-Score, AUC-ROC, and Quadratic Weighted Kappa (QWK) on the held-out test set. Explainability analysis using Grad-CAM, Grad-CAM++, and Score-CAM is subsequently applied to the best performing architecture to produce clinically interpretable heatmap visualizations aligned with established neuro-anatomical biomarkers of Alzheimer's disease.
The main contributions of this work are:

- A three way comparative evaluation of a Custom CNN, fine-tuned ResNet50, and Hybrid ResNet50+ViT (HybridRViT) for multi-class Alzheimer's disease staging from MRI, performed under identical training conditions and evaluated on a unseen original test data.
- This study proved that ResNet50 with selective layer freezing and layer-wise differential learning rates outperforms both from Custom CNN and the HybridRViT model, achieving 95.58% accuracy, macro F1-Score of 0.9708, macro AUC of 0.9972, and QWK of 0.9658, it is the best architecture for this task at the present data scale.
- An analytical explanation of why the HybridRViT model underperforms than ResNet50, because of the data-scale sensitivity of the Vision Transformer's self-attention mechanism and the absence of convolutional inductive biases, providing actionable guidance for future hybrid model design in medical imaging.
- The con-current application of three gradient-based Explainable AI (XAI) methods, gradient-based attribution methods specifically Grad-CAM and its variant Grad-CAM++, alongside the gradient-free visualization approach of Score-CAM on the ResNet50 in four clinical scenarios (correct predictions, misclassifications, disease progression, and prediction confidence levels), with neuro-anatomically grounded interpretation of the resulting heatmaps.
- A comprehensive per-class performance analysis that reveals the clinical implications of inter-model differences at the individual dementia stage level, with particular attention to the Mild and Moderate Demented classes where staging accuracy has the greatest direct impact on clinical decision-making.

II. Literature Review

A. *Alzheimer's Disease Detection using Convolutional Neural Network*

From the last decades a Convolutional Neural Networks(CNNs) have been the dominant approach to classify AD's from MRI images. sarraf and Tofighi [3] proved that deep CNNs based on the LeNet-5 architecture could classify Alzheimer's brain MRI from healthy controls with 98.84% accuracy using fMRI data, establishing the foundational viability of CNN-based AD diagnosis. Following this, A.M. Amer et al. [4] introduced a new custom CNN architecture trained on MRI scans that obtained high classification accuracy with the explainability of Grad-CAM to identify discriminative brain regions, demonstrating that CNNs can be effective in classification performance and clinical interpretability. Most recently, published paper El-Latif et al. [5] proposed a low resourced CNN framework trained on the Kaggle Alzheimer's MRI dataset, with competitive accuracy while significantly reducing the computational overhead, thus addressing the deployment constraints of resource-limited clinical environments. Dardouri S (2025) [6] also designed a custom three-branch CNN architecture with 6,026,324 parameters by using different kernel sizes in parallel convolutional branches for better multi-scale feature extraction in early AD detection from MRI images, and claimed excellent classification performance among four dementia stages

B. *Transfer Learning and ResNet-based Methods*

Transfer learning from large-scale image datasets has become as a powerful strategy to addressed the limited availability of annotated medical MRI data. In Al Shehri [7] paper did a well structured comparison of ResNet50 and DenseNet169 for four-class Alzheimer's classification on the OASIS dataset, and show that deep residual networks consistently outperform shallower CNNs due to their skip-connection mechanism that helps to preserves gradient flow through many layers. Shahid SB et al. [8] proposed a transfer learning based techniques for multi-class AD classification across four categories non-demented, very mild, mild, and moderate demented using ResNet152V2, VGG16, InceptionV3, and MobileNet as feature extractors on the ADNI and OASIS datasets. Their proposed Incep+Res fusion model achieved accuracies of 96.96%, 98.35%, and 97.13% respectively on the ADNI, OASIS, and merged datasets, respectively, with ResNet-based extractors consistently ranking among the top performers. Amine and Mourad [9] demonstrated the utility of ResNet50 for AD detection on the ADNI and MIRIAD benchmark datasets through a transfer learning pipeline, obtaining 99% accuracy on ADNI and 96% on MIRIAD by combining extracted deep features with Softmax, SVM and Random Forest classifiers. F. Mostafa et al. [10] also proposed a hybrid ensemble framework by combining ResNet50, NASNet, and MobileNet with a stacked meta-learner and achieved an accuracy of 99.21% for the AD vs mild cognitive impairment classification on the ADNI dataset

and Grad-CAM heatmaps were used to provide interpretable attribution maps for gray and white matter regions.

C. *Vision Transformer Architectures*

Introduction of the Vision Transformer (ViT) by Dosovitskiy et al (2021). [11], this attention-based architectures have gained significant interest in medical imaging due to their capacity for global context modeling through self-attention. Berroukham et al (2023). [12] given a detailed study of ViT architectures, their variants, and their application domains, concluded that while Transformers offer superior long-range dependency modeling, their performance in data-starved domains such as medical imaging is limited by the lack of convolutional inductive biases and resulting high data requirements. Almufareh et al.(2023) [13] applied a pure ViT model to the Kaggle Alzheimer's MRI dataset and achieved accuracy = 99.06%, precision = 99.06%, recall = 99.14%, and F1-score = 99.10%, demonstrating that ViT models could be competitive on Kaggle dataset with appropriate training strategies. However, the authors highlighted some challenges related to the statistical reliability of near-perfect scores in minority classes, especially the Moderate Demented category.A recent study, an Information (MDPI) study Hechkel et al. (2026) [14] proposed a ViT-based framework for multi-class AD classification on the Kaggle dataset using a subject-wise data-splitting strategy to avoid data leakage. The authors questioned the statistical reliability of high scores achieved in under-represented classes and conducted ablation studies to measure the contribution of self-attention and patch embedding components. Şener et al. (2025) [15] explored ViT-based models for early Alzheimer's detection with a focus on MRI slice selection, showed that proper slice curation significantly impacts ViT classification performance, particularly for early Mild Cognitive Impairment detection.

D. *Hybrid CNN–Transformer Architectures*

The complementary strengths of CNNs in local feature extraction, translation equivariance and Transformers in global attention, long-range dependency modeling, has been explored hybrid architectures that combine both paradigms. Pantelaios et al. (2024) [16] revealed that hybrid CNN-ViT models for general medical image classification at IEEE ISBI 2024, proving that CNN backbones used as patch token generators for ViT encoders outperform pure CNNs on several medical imaging benchmarks by capturing both local texture features and global structural relationships. An other research paper [17] presented a comprehensive review of CNN-ViT hybrid design strategies including sequential, parallel, and cross-attention fusion architectures show that sequential hybrids approach, where CNN features are tokenized and fed into Transformer encoders, consistently achieve competitive performance across various visual tasks. In Senan et al. paper (2015) [18] proposed a hybrid ResNet101-ViT and GoogLeNet-ViT methodology for four-class AD classification using the OASIS dataset, incorporating adaptive median and Laplacian filters for preprocessing and modifying the ViT architecture to reduce computational cost. Their ResNet101-ViT hybrid achieved 98.7% accuracy, 96.4% precision, 99.68% sensitivity, and 97.78% specificity, outperforming the GoogLeNet-ViT variant. A comparative study published by Author in JISKA (Jurnal Informatika Sunan Kalijaga) [19] tested hybrid CNN-ViT versus pure CNN models for brain tumor classification, finding that while hybrid models suggested that improved global feature capture, their advantage over CNNs decreased in limited-data policy.

E. *Explainable AI (XAI) in Alzheimer's Detection*

The clinical adoption of deep learning models for Alzheimer's diagnosis is critically contingent on model interpretability. The adoption of deep learning models in medical fields such as Alzheimer disease detection is critically important of model Interpretability. Deep learning models predictions, however accurate, but insufficient for clinical decision support without enclosed explanations of the spatial brain regions driving the classification. Singhal et al (2025). [20] proposed ADNet, an optimized CNN architecture trained on 4,000 ADNI MRI scans, achieved a classification accuracy of 99.40% and AUC of 99.85%, and applied Grad-CAM XAI technique to produce heatmaps visually highlighting brain regions integrated on each dementia stage, help in the clinical utility of gradient-based visualization for Alzheimer's staging. In the Jahan et al paper. [21] conducted a comparative evaluation of multiple XAI algorithms such as Grad-CAM, SHAP, and LIME applied on EfficientNet models for AD prediction using MRI data presented at the Brain Informatics, proving that Grad-CAM provides the most spatially coherent and clinically interpretable activations for brain MRI, while SHAP offers superior feature-level attribution. In Mostafa et al. [8] further integrated Grad-CAM heatmaps into ensemble ResNet50-NASNet-MobileNet framework, demonstrating that XAI visualizations aligned with known structural biomarkers including hippocampal regions, no matter of which ensemble parts dominated the prediction.

III. Dataset and Pre-processing

*1) **Dataset Description**:* The study was performed on Augmented Alzheimer MRI Dataset available on kaggle, a complete set of brain MRI scan images in .jpg format, categorised into four clinically relevant stages of Alzheimer diesease: 1. Non-Demented, 2. Very Mild-Demented, 3. Mild-Demented, and 4. Moderate-Demented [2]. To enable a rigorous and unbiased evaluation, the dataset is split into two disjoint sets: an augmented training validation set, and an original hold-out test set.

The augmented partition contains a total of 33,984 images and it was used only for model training and validation. It comprises 9,600 images of the Non-Demented class, 8,960 images each for the Very Mild Demented and Mild Demented classes, and 6,464 images of the Moderate Demented class. The original partition dataset, reserved alone for final performance evaluation, consists of 6,400 images distributed as 3,200 Non-Demented images, 2,240 Very Mild-Demented images, 896 Mild-Demented images, and 64 Moderate-Demented images. The complete class-wise distribution for both partitions is summarized in Table II.

TABLE I
LITERATURE REVIEW SUMMARY: METHODS, DATASETS, ACCURACY, AND RESEARCH GAPS

S.No.	Authors (Year)	Dataset	Architecture / Method	Accuracy	XAI Used	Key Limitation
[1]	Sarraf & Tofighi (2016)	ADNI (fMRI)	LeNet-5 CNN	98.84%	No	Binary only (AD vs. NC); no staging
[2]	El-Assy et al. (2024)	ADNI	Novel custom CNN	High (reported)	Grad-CAM	Single dataset; no Transfer Learning
[3]	Abd El-Latif et al. (2023)	Kaggle MRI	Lightweight CNN	Competitive	No	No comparison with TL architectures
[4]	Dardouri (2025)	Kaggle MRI	Multi-branch custom CNN	High	No	No ViT or hybrid comparison; no XAI
[5]	Al Shehri (2022)	OASIS	ResNet50 vs DenseNet169	High	No	No multi-stage; no XAI integration
[6]	Khan et al. (2025)	ADNI, OASIS	IncepRes (ResNet152V2 + Inception)	98.35%	No	No ViT component; no XAI
[7]	Amine & Mourad (2025)	ADNI, MIRIAD	ResNet50 + Softmax/SVM/RF	99.00%	No	Feature extraction only; no end-to-end fine-tuning
[8]	Mostafa et al. (2025)	ADNI	ResNet50 + NASNet + MobileNet	99.21%	Grad-CAM	Ensemble complexity; no ViT; single XAI method
[9]	Dosovitskiy et al. (2021)	ImageNet	Vision Transformer (ViT)	88.55% (ViT-H)	Attention maps	Requires massive data; limited medical validation
[10]	Berroukham et al. (2023)	—	ViT Survey / Review	—	—	Review only; no empirical AD experiments
[11]	Almufareh et al. (2023)	Kaggle MRI	Pure ViT	99.06%	No	No CNN comparison; no XAI; class imbalance concern
[12]	MDPI Information (2026)	Kaggle MRI	ViT with subject-wise split	Reported	No	No hybrid or CNN comparison; no XAI
[13]	Şener et al. (2025)	ADNI	ViT + slice selection	Reported	No	Focuses on EMCI only; no multi-stage classification
[14]	Pantelaios et al. (2024)	Multi-domain	Hybrid CNN-ViT (general)	Varies by task	No	Not AD-specific; no XAI integration
[15]	Long (2024)	—	CNN-ViT Hybrid Review	—	—	Review only; no AD-specific evaluation
[16]	Senan et al. (2025)	OASIS	ResNet101-ViT Hybrid	98.70%	No	ResNet101, not ResNet50; no XAI applied
[17]	JISKA (2026)	Brain Tumor	Hybrid CNN-ViT vs CNN	Reported	No	Brain tumor, not Alzheimer's; no XAI
[18]	Bortty et al. (2025)	OASIS (Kaggle)	X-ViTCNN (ViT+DenseNet+MobileNet)	97.31%	No	Complex ensemble; no XAI; no ResNet50 comparison
[19]	Singhal et al. (2025)	ADNI (4000)	ADNet (optimized CNN)	99.40%	Grad-CAM	Small dataset; binary stages; no ViT comparison
[20]	Jahan et al. (2023)	MRI (Kaggle)	EfficientNet + XAI comparison	Reported	Grad-CAM, SHAP, LIME	EfficientNet only; no ResNet50 or ViT comparison
[21]	Shaukat et al. (2024)	—	Systematic Review (DL for AD)	—	—	Review; identifies XAI gap; no empirical study

TABLE II
CLASS-WISE IMAGE DISTRIBUTION OF THE ALZHEIMER MRI DATASET

Dementia Stage	Train+Val Set	Test Set	Total
Non-Demented	9,600	3,200	12,800
Very Mild Demented	8,960	2,240	11,200
Mild Demented	8,960	896	9,856
Moderate Demented	6,464	64	6,528
Total	**33,984**	**6,400**	**40,384**

It is important to note that while the augmented training set exhibits a relatively balanced class distribution facilitated by augmentation techniques applied by the dataset authors the original test set reflects a natural clinical imbalance, particularly in the Moderate Demented class (n = 64). This imbalance in the test partition provides a realistic evaluation scenario consistent with real-world Alzheimer's prevalence distributions.

*2) **Train-Validation Split**:* The augmented training set was divided using a stratified random split, allocating 80% of the data (27,187 images) for training and the remaining 20% (6,797 images) for validation. Stratification was applied to preserve the class proportions across both splits. The original dataset partition (6,400 images) was kept entirely separate and used only during the final test-set evaluation phase to provide an unbiased estimate of generalization performance.

*3) **Pre-processing Pipeline**:* A standardized preprocessing pipeline was applied to all images across the training, validation,

and test sets. Each image was resized to 224 × 224 pixels to satisfy the spatial input requirements of the ResNet50 and Hybrid ResNet50+ViT architectures, and to maintain consistency across all three models under comparison. Pixel values were subsequently converted from the integer range [0, 255] to normalized floating-point values in [0.0, 1.0] via PyTorch's *transforms.ToTensor()* operation.

Channel-wise mean and standard deviation normalization was then applied. For ResNet50 and the Hybrid ResNet50+ViT model, the canonical ImageNet statistics were employed *mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]* to align with the pretrained weight distributions. For the Custom CNN, which was trained from scratch without leveraging pretrained weights, the normalization statistics were computed directly from the training dataset, yielding *mean = [0.2956, 0.2956, 0.2956] and standard deviation = [0.3184, 0.3184, 0.3184]*, reflecting the grayscale-dominant nature of MRI images replicated across three channels.

*4) **Data Augmentation**:* Since the training partition already incorporates augmented samples generated by the original dataset authors, only minimal additional augmentation transforms were applied during training to introduce variability and reduce overfitting. Specifically, *transforms.RandomHorizontalFlip()* and *transforms.RandomRotation(10°)* were used. These normal transforms are well-suited for brain MRI images, as they preserve the anatomical plausibility of the scans while introducing sufficient diversity to regularize the models. No augmentation was applied during validation or testing to ensure evaluation integrity.

IV. Methodology

*1) **Overall Framework Details**:* Three distinct architectural paradigms are benchmarked against one another within a unified experimental protocol for four-category dementia severity grading from structural brain scans: (1) a shallow feature learner built entirely from randomly initialised weights, referred to as the Custom CNN, (2) a residual network with 50 layers whose parameters were initialised from large scale natural image training and subsequently adapted to the neuroimaging domain — designated ResNet50, and (3) a dual-component model that routes convolutional spatial features through a self-attention encoder, collectively termed HybridRViT. Each architecture receives a standardised 224 × 224 three-channel tensor as its sole input and maps it to one of four ordinal severity categories defined by the dataset. Following the comparative evaluation, gradient-based saliency mapping tools are deployed exclusively on the top-ranked architecture to generate spatially interpretable overlays that reveal which anatomical regions drive each staging decision. The end-to-end pipeline is summarised visually in Fig. 1

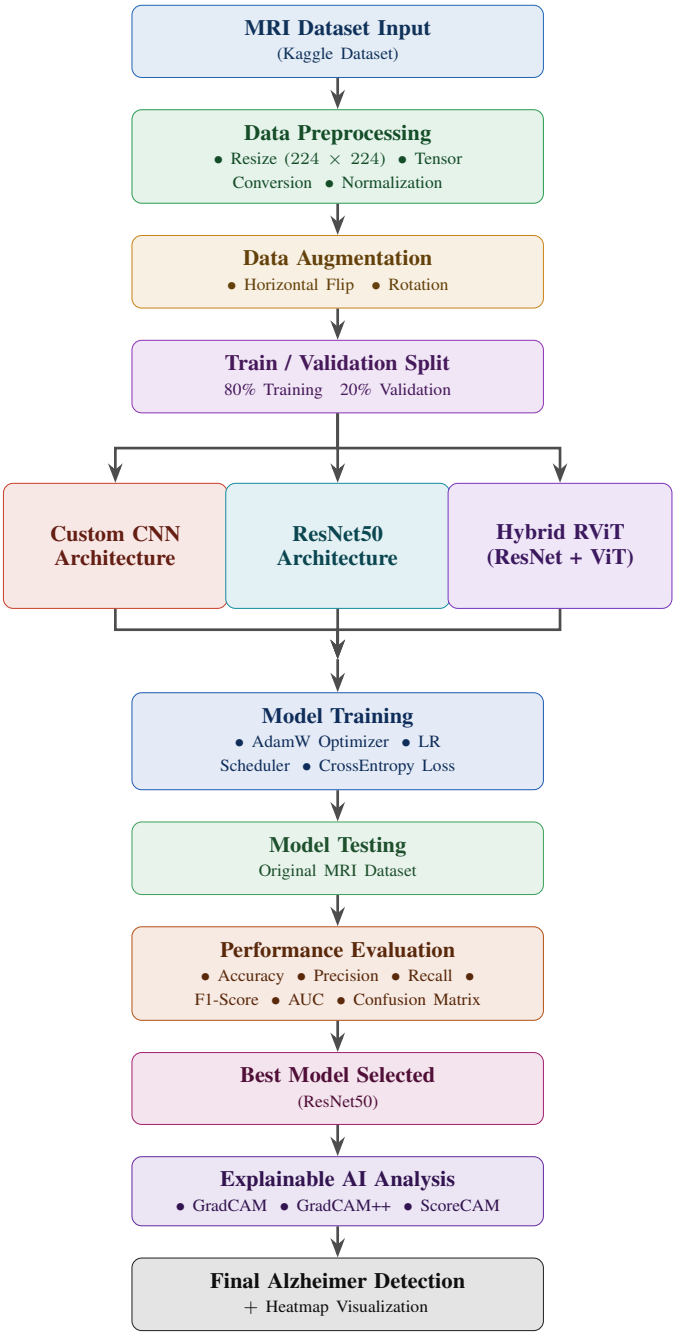

Fig. 1. Proposed methodology flow diagram for multi stage AD's classification.

*2) **Model 1 Custom CNN**:* The Custom CNN works as the baseline model, designed and trained from scratch without any pretrained weights initialization. The network is composed of 4 sequential convolutional blocks followed by a fully connected classifier, as shown in Fig 2. Each convolutional block applies two consecutive 3×3 convolution layers with padding=1, each block followed by Batch Normalization and ReLU activation. A Max Pooling layer with stride 2 halves the spatial dimensions after each block, and a spatial Dropout layer (rate = 0.1) provides regularization. The channel progression across the four blocks is $364, 64 \rightarrow 128, 128 \rightarrow 256, and 256 \rightarrow 512$,

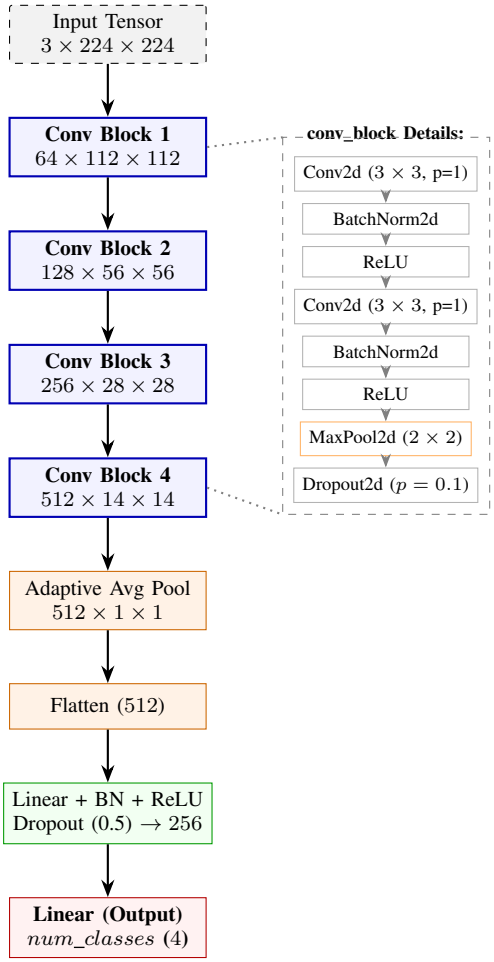

Fig. 2. Proposed Custom CNN Architecture for multi stage AD's classification.

progressively compressing the spatial resolution from 224 × 224 to 14 × 14 while expanding feature representations.

The classifier head employs Adaptive Average Pooling to reduce the 14×14 feature maps to a 1×1 spatial output, followed by flattening and two fully connected layers (512→256→4) with Batch Normalization, ReLU, and a 50% Dropout between them. All convolutional weights are initialized using Kaiming Normal initialization with fan-out mode, appropriate for layers followed by ReLU activations. Batch Normalization parameters are initialized with weight=1 and bias=0 for numerical stability. The configuration of the Custom CNN Architecture is summarized in Table III

TABLE III
CUSTOM CNN ARCHITECTURE SUMMARY

Stage	Layer Configuration	Output Shape	Dropout
Block 1	Conv3×3 (3→64)×2, BN, ReLU, MaxPool	64×112×112	0.1
Block 2	Conv3×3 (64→128)×2, BN, ReLU, MaxPool	128×56×56	0.1
Block 3	Conv3×3 (128→256)×2, BN, ReLU, MaxPool	256×28×28	0.1
Block 4	Conv3×3 (256→512)×2, BN, ReLU, MaxPool	512×14×14	0.1
Pooling	AdaptiveAvgPool2d(1,1) + Flatten	512	—
FC-1	Linear(512→256), BN, ReLU	256	0.5
FC-2 (Output)	Linear(256→4)	4	—

*3) **Model 2 ResNet50 with Layer-wise Fine-tuning**:* The second model is built on the ResNet50 architecture [2] pretrained on the ImageNet-1K dataset. ResNet50 used residual connections strategy, A shortcut paths that bypass one or more convolutional layers, which enables gradient flow through very deep networks and avoid the vanishing gradient problem. The pretrained backbone is decomposed into five named stages: a stem block consisting of a 7×7 convolution, with Batch Normalization, ReLU, and Max Pooling, followed by four residual stage groups (from layer1 to layer4) containing bottleneck blocks with channel dimensions 256, 512, 1024, and 2048, respectively as shown in Fig 3.

A selective layer-freezing strategy is adopted to adapt the pre-trained model to the MRI domain without losing low-level generic features. Specifically, the stem, layer1, and layer2 stages are frozen (parameters fixed), as these layers encode low-level features such as edges and textures that are domain-agnostic. Layer3 and Layer4, which encode higher-level semantic and structural representations, are kept trainable to allow domain-specific adaptation to brain MRI morphology. This fine-tuning strategy reduces the risk of catastrophic forgetting while enabling the network to adapt its deep representations to the multi stage Alzheimer's disease classification task.

The original model of ResNet50 classification head is replaced with a three-layer fully connected head: $2048 \rightarrow 512 \rightarrow 256 \rightarrow 4$, with Batch Normalization, ReLU activation, and Dropout (rates 0.4 and 0.2 respectively) inserted between layers. This deeper implementation provides improved regularization over a single linear projection. A layer-wise learning rate schedule is applied via the AdamW optimizer, with lower learning rates assigned to early-stage layers (1×10^{-6} from starting layer through layer 2) and progressively higher rates for deeper and newly initialized layers (1×10^{-5} for layer3 and layer 4; 1×10^{-4} for the classification head). The ResNet50 stage-wise Setup has given in the Table IV.

TABLE IV
RESNET50 STAGE CONFIGURATION AND TRAINING STRATEGY

Stage	Output Shape	Status	Learning Rate
Stem (Conv1+BN+ReLU+MaxPool)	64×56×56	Frozen	1×10^{-6}
Layer1 (3 bottleneck blocks)	256×56×56	Frozen	1×10^{-6}
Layer2 (4 bottleneck blocks)	512×28×28	Frozen	1×10^{-6}
Layer3 (6 bottleneck blocks)	1024×14×14	Trainable	1×10^{-5}
Layer4 (3 bottleneck blocks)	2048×7×7	Trainable	1×10^{-5}
AdaptiveAvgPool + Flatten	2048	Trainable	—
FC Head (2048→512→256→4)	4	Trainable	1×10^{-4}

*4) **Model 3 Hybrid ResNet50 + Vision Transformer (HybridRViT)**:* The third model, HybridRViT, integrates the local feature extraction ability of ResNet50 with the global self-attention mechanism of the Vision Transformer (ViT). The motivation for such a hybrid design is that CNNs are good at extracting local texture and edge patterns, but do not have explicit mechanisms to model long-range spatial dependencies, whereas Transformers model global context via self-attention but require large training datasets for effective generalisation. The HybridRViT architecture aims at leveraging the complementary strengths of both paradigms.

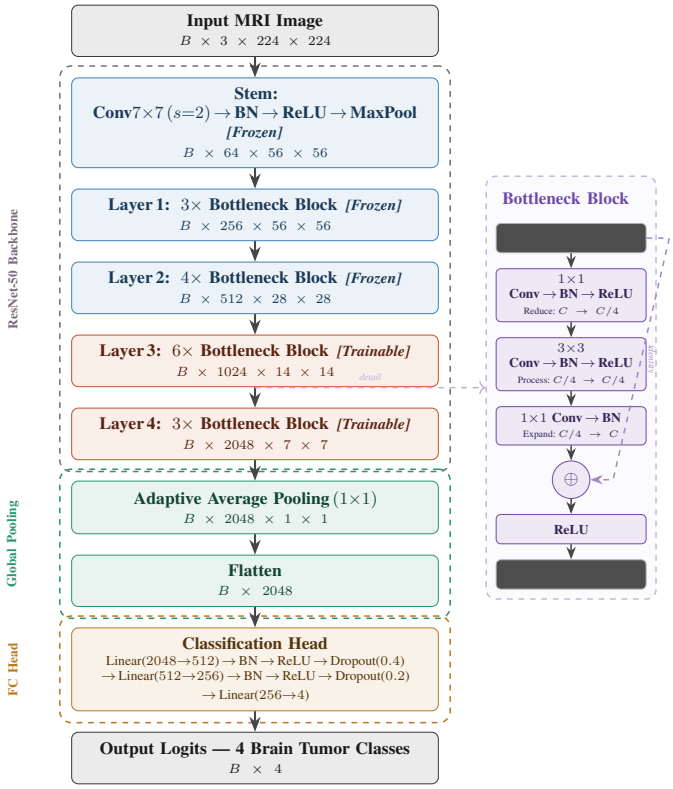

Fig. 3. Proposed ResNet50 Architecture for Multi stage AD's classification.

The ResNet50 backbone (stem through Layer4) is retained with the same freezing strategy as the ResNet50 baseline — stem, Layer1, and Layer2 are frozen, while Layer3 and Layer4 remain trainable. The 7×7 spatial feature map produced by Layer4, with 2048 channels, is treated as a sequence of 49 patch tokens. Each token is linearly projected to an embedding dimension of 768 using a learned projection layer followed by Layer Normalization, matching the standard ViT-Base embedding dimension. A learnable [CLS] token is prepended to the 49 patch tokens, and learned positional embeddings of shape (1, 50, 768) are added to encode spatial position information. A Dropout layer (p = 0.1) is applied to the resulting token sequence.

The token sequence is then processed by a Transformer Encoder consisting of 6 stacked Transformer layers, each with 12 attention heads, a feed-forward dimension of 3072 (4× the embedding dimension), GELU activation, and Pre-Layer Normalization (normfirst=True). The output associated with from the [CLS] token and passed through a multi-layer classification head: 768→512→256→4, with Batch Normalization, GELU activation, and dropout between each linear layer. Learnable parameters (CLS token, positional embeddings, and projection weights) are initialized using truncated normal distributions (std = 0.02) following standard ViT initialization practices. The HybridRViT architecture is summarized in Table V and in model architecure diagram Fig 4 .

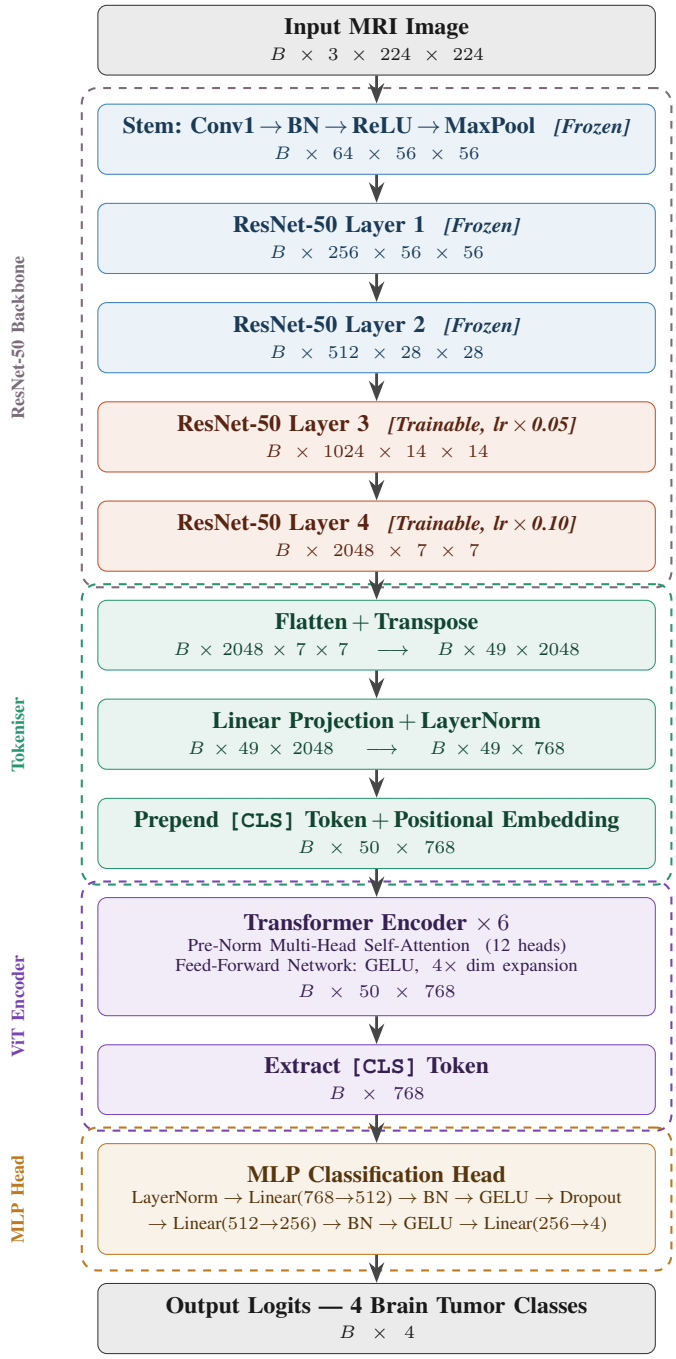

Fig. 4. Proposed HybridRViT Architecture for multi stage AD's classification.

TABLE V
HYBRIDRVIT ARCHITECTURE SUMMARY

Component	Configuration	Output Shape	Status
ResNet50 Backbone (Stem–Layer4)	Pretrained ImageNet; partial freeze	2048×7×7	Partial
Patch Tokenization	Flatten spatial dims → 49 tokens	49×2048	Trainable
Token Projection + Norm	Linear(2048→768), LayerNorm	49×768	Trainable
CLS Token + Pos. Embedding	Learnable; shape (1, 50, 768)	50×768	Trainable
Transformer Encoder	6 layers, 12 heads, dim_ff=3072, GELU	50×768	Trainable
CLS Token Extraction	Index 0 from encoder output	768	—
MLP Head	Linear(768→512→256→4), BN, GELU, Dropout	4	Trainable

*5) **Training Configuration**:* All three models architecture code were written in PyTorch and trained on nvidia's Tesla

T4 GPU 16 GB memory on a Kaggle Notebook environment. A training configuration was applied to all architectures were same, where applicable: 15 epochs, a small batch size of 64, and Cross-Entropy Loss as the objective function. The AdamW optimizer was used for all models, incorporating L2 weight decay ($\lambda = 1 \times 10^{-4}$) to regularize model parameters. Learning rate adaptation was controlled by a ReduceLROnPlateau scheduler monitoring validation loss, with a patience of 3 epochs and a reduction factor of 0.5, down to a minimum learning rate apply of 1×10^{-6} for(Custom CNN) and 1×10^{-7} for (ResNet50 and HybridRViT).

For the Custom CNN, an only global learning rate of 1×10^{-3} was applied to all parameters. For ResNet50 and HybridRViT, layer-wise learning rates were set to account for the changing degrees of pre-training and task adaptation required at each stage of the network (see Tables III and IV). This differential learning rate strategy stop the uncontroled updates to lower-level pretrained features while allowing the task-specific upper layers to adapt more fastly and efficiently. The complete training hyperparameter setting has been given in Table VI.

TABLE VI
TRAINING HYPERPARAMETER CONFIGURATION FOR ALL THREE MODELS

Hyperparameter	Custom CNN	ResNet50	HybridRViT
Framework	PyTorch	PyTorch	PyTorch
Epochs	15	15	15
Batch Size	64	64	64
Optimizer	AdamW	AdamW	AdamW
Base Learning Rate	1×10^{-3}	Layer-wise	Layer-wise
Weight Decay	1×10^{-4}	1×10^{-4}	1×10^{-4}
LR Scheduler	ReduceLROnPlateau	ReduceLROnPlateau	ReduceLROnPlateau
Scheduler Patience	3 epochs	3 epochs	3 epochs
Reduction Factor	0.5	0.5	0.5
Min. Learning Rate	1×10^{-6}	1×10^{-7}	1×10^{-7}
Loss Function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Hardware	NVIDIA Tesla T4 (16 GB), Kaggle Notebook		

A. *Evaluation Metrics*

To quantify classification behaviour across all three architectures, six complementary performance measures were derived entirely from the held-out original test partition: overall correctness rate (Accuracy), positive predictive value (Precision), sensitivity(Recall), harmonic mean score(F1-score), diagnostic discriminability (AUC), and ordinal agreement coefficient (QWK). Throughout this section, TP, FP, FN, and TN refer to the count of true positives, false positives, false negatives, and true negatives respectively for an individual severity stage, while N denotes the total number of test observations. Given the unequal representation of dementia stages in the test partition — particularly the severely underrepresented advanced stage (n = 64) — class-unweighted macro averages are adopted as the principal reporting standard, since they assign equal importance to every stage irrespective of how frequently it appears. Frequency-proportional weighted averages are additionally included for completeness. Numerical computation of all reported figures was performed via the *sklearn.metrics* module .

Accuracy: The overall correctness rate captures what fraction of all test observations received the right severity label:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision(Positive Predictive Value) and Recall(Sensitivity): When a model assigns a sample to a particular dementia stage, positive predictive value quantifies how often that assignment is correct. Sensitivity, by contrast, measures how many of the genuinely belonging samples the model successfully retrieves:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1-Score: Rather than treating sensitivity and positive predictive value as independent measures, the F1-Score consolidates both into a single figure by computing their harmonic mean. This makes it especially informative when class sizes are unequal, since it penalises models that sacrifice one measure to inflate the other:

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

AUC: For multi-class settings, each dementia stage is evaluated independently against all remaining categories under a one vs rest decomposition. The resulting per-class curve quantifies how consistently a model separates a target stage from all others as the decision threshold is varied continuously. A macro-average is subsequently computed by weighting each class equally:

$$\text{AUC}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \text{AUC}_c \tag{5}$$

where C = 4 is the total number of severity stages, and AUCc represents the individual score for the c-th stage, estimated geometrically via the trapezoidal approximation applied to its receiver operating characteristic curve.

Quadratic Weighted Kappa (QWK): Standard accuracy treats all misclassifications as equivalent, which is inappropriate for ordinal tasks where severity levels carry a natural ordering. QWK corrects for this by assigning heavier penalties to predictions that deviate further from the true stage, and additionally adjusts for the level of correspondence expected by chance alone:

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij} \cdot O_{ij}}{\sum_{i,j} w_{ij} \cdot E_{ij}} \tag{6}$$

where Oij is the observed frequency matrix, Eij is the expected frequency matrix under chance agreement, and wij is the quadratic weight matrix defined as:

$$w_{ij} = \frac{(i - j)^2}{(N_{\text{classes}} - 1)^2} \tag{7}$$

Standard accuracy considers all misclassifications as equal. This is not appropriate for ordinal tasks where the severity

levels have a natural order. QWK compensates for this by penalising predictions further away from the true stage more heavily, and also correcting for the level of correspondence expected by chance alone

V. Experiments and Results

A. *Overall Performance Comparison*

The given Table VII consolidates the test-partition outcomes for all three competing architectures. Among the candidates evaluated, the selectively fine-tuned ResNet50 attains top-ranked scores on every reported indicator simultaneously — recording a correctness rate of 95.58%, a class-unweighted harmonic mean score of 0.9708, a macro-averaged diagnostic discriminability of 0.9972, and an ordinal agreement coefficient of 0.9658. Positioned second overall, the HybridRViT architecture delivers a correctness rate of 92.84% alongside a class-unweighted harmonic mean score of 0.9484, while the from-scratch trained Custom CNN registers the lowest figures at 89.03% and 0.9277 on the same two indicators. The performance separation between ResNet50 and its two counterparts is consistent and substantial — approximately 6.5 percentage points over the scratch-trained baseline and 2.7 points over the hybrid model when measured by correctness rate alone — collectively affirming that transfer learning via residual feature reuse is the most effective strategy for this four-stage dementia classification task at the present data scale.

TABLE VII
Overall Performance Comparison on the Test Set (n = 6,400)

Metric	Custom CNN	ResNet50	HybridRViT
Accuracy	0.8903	0.9558	0.9284
Quadratic Weighted Kappa	0.9128	0.9658	0.9164
Macro Precision	0.9251	0.9762	0.9613
Macro Recall	0.9306	0.9667	0.9389
Macro F1-Score	0.9277	0.9708	0.9484
Macro AUC	0.9799	0.9972	0.9946
Weighted Precision	0.8896	0.9572	0.9318
Weighted Recall	0.8903	0.9558	0.9284
Weighted F1-Score	0.8897	0.9555	0.9277
Weighted AUC	0.9663	0.9954	0.9916

B. *Per-Class Performance Breakdown*

The four reported metrics positive predictive value, sensitivity, harmonic mean, and AUC were used to evaluate each model's performance per class across the four stages of AD. Attached Fig. 5, 6 and 7 show that several clinically significant findings. First, all three architectures achieve a score (Precision = Recall = F1 = AUC = 1.000) in detect of Moderate Demented class, it suggesting that this class occupies a highly distinct feature space within the dataset reflecting the severe and visually unambiguous cortical atrophy characteristic of late-stage Alzheimer's disease. Second, the Very Mild Demented class proves to be the most challenging across all models, with the lowest F1-Scores recorded Custom CNN = 0.8463, ResNet50 = 0.9372, and Hybrid RViT = 0.8967 . This is clinically expected, as Very Mild Dementia occupies the most ambiguous boundary with Non-Demented cases. Third, the Non-Demented class shows a outstanding recall advantage for ResNet50 = 0.9875 and HybridRViT = 0.9869 over the Custom CNN = 0.9025, indicating that transfer learning significantly improves sensitivity for telling difference of healthy brains from early-stage degeneration. Fourth, the Mild Demented class achieves high Precision across all three models Custom CNN = 0.9476, Resnet50 = 0.9955, Hybrid RViT = 0.9928, confirming that when any model predicts Mild Dementia, it is highly likely to be correct.

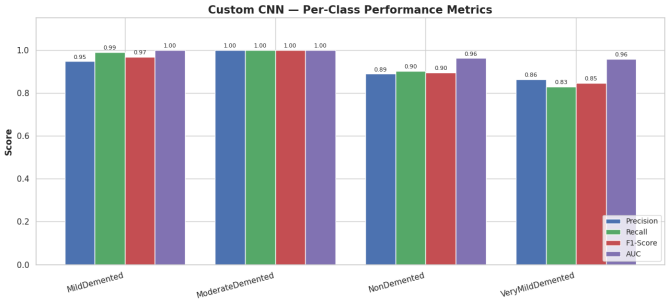

Fig. 5. Custom CNN Per Class Evaluation

Fig. 6. ResNet50 Per Class Evaluation

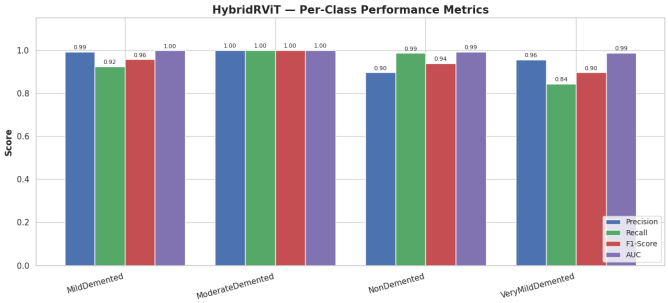

Fig. 7. HybridRViT Per Class Evaluation

C. *Confusion Matrix Analysis*

Fig. 8, 9, and 10 respectively present the classification confusion matrices obtained from the held-out test partition for the Custom CNN, ResNet50, and HybridRViT architectures,

respectively. Across all three models, the dominant failure pattern emerges at the boundary separating cognitively normal subjects from those exhibiting negligible cognitive decline — specifically, the two stages that share the greatest structural overlap in T1-weighted MRI appearance. The Custom CNN produces 286 misassignments from the first of these adjacent categories into the second, and 358 in the reverse direction, collectively reflecting substantial inter-class ambiguity at this clinically sensitive boundary. Fine-tuning ResNet50 on ImageNet representations substantially reduces these errors: only 38 and 223 incorrect assignments are observed in the same two directions — improvements of 86.7% and 37.7% over the baseline, respectively. The HybridRViT architecture records 39 and 345 such boundary errors, placing its performance between the other two models at this stage transition. A separate pattern worth noting involves the HybridRViT producing 49 incorrect stage assignments from the third severity category into the second — an error occurring only 7 and 8 times in ResNet50 and the Custom CNN, respectively — which may reflect the tendency of self-attention to over-generalise spatially distributed features under constrained training conditions. Across all three architectures, the most advanced severity category achieves flawless classification with all 64 test samples correctly identified, demonstrating that the structural hallmarks of late-stage neurodegeneration are unambiguously captured regardless of architectural choice.

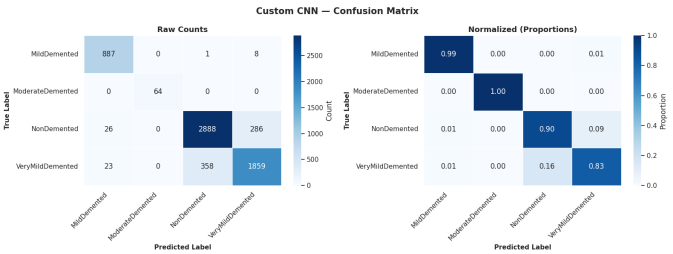

Fig. 8. Custom CNN Confusion Matrix

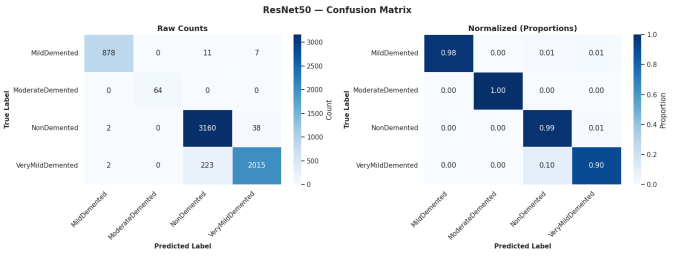

Fig. 9. ResNet50 Confusion Matrix

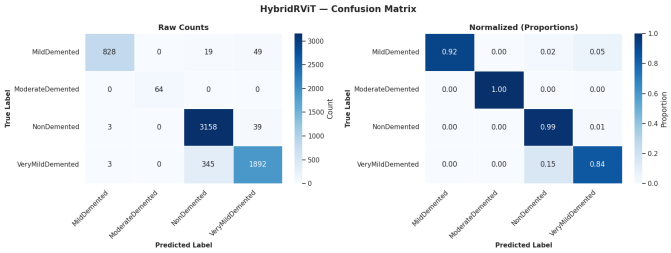

Fig. 10. HybridRViT Confusion Matrix

D. *Training and Validation Loss Convergence*

Fig. 11, 12, and 13, respectively shows that epoch-wise training and validation losses for all three models over 15 epochs. The ResNet50 demonstrated the most consistent and smooth convergence trajectory, with training loss declining from 1.0729 at 1st epoch to 0.1177 at 15th epoch, and validation loss following closely from 0.8122 to 0.1059, that indicating minimal overfitting and excellent generalization. The Custom CNN shows a slightly less stable convergence, particularly at epoch 2 where validation loss temporarily increases to 1.2092 before recovering, suggesting sensitivity to the randomly initialized weights at early training stages. The HybridRViT demonstrates stable convergence overall but show the marginal validation loss fluctuations around epochs 11–13 (0.2128, 0.2132, 0.1810), consistent with the optimization challenges typically associated with joint CNN-Transformer training on limited medical imaging datasets.

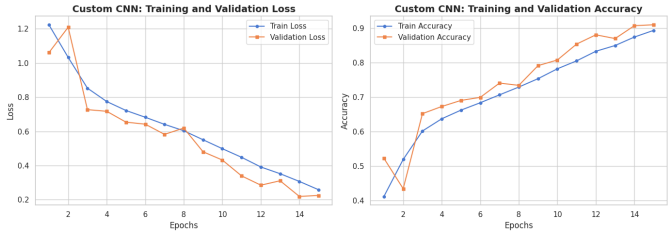

Fig. 11. Custom CNN Accuracy & Loss

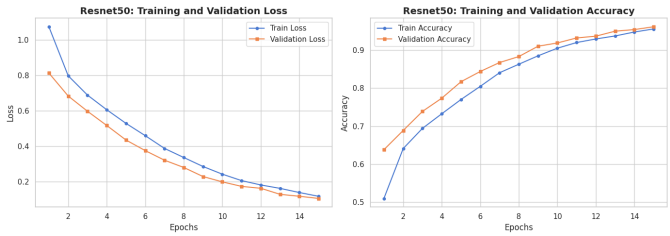

Fig. 12. ResNet50 Accuracy & Loss

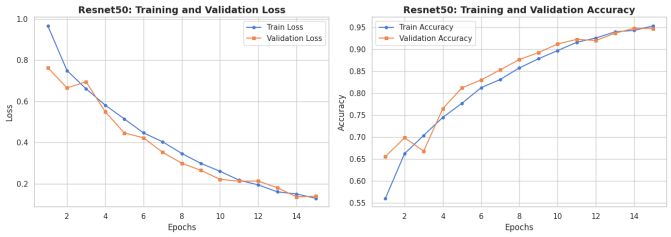

Fig. 13. HybridRViT Accuracy & Loss

E. Explainable AI Results — Grad-CAM, Grad-CAM++, and Score-CAM Visualisations

On the best performing model Resnet50 the three gradiend based Class activation mapping techniques were applied CAM-based methods to the final convolutional layer to interpret the model decision making. The resulting heatmaps were overlaid on the original MRI images and analyzed in four ways: correctly classified predictions Fig. 14, misclassified predictions Fig.15 , dementia stage progression Fig. 16, and high- versus low-confidence predictions Fig. 17. For correctly classified samples, all three XAI methods consistently produce activations confined to the medial temporal lobe and hippocampus neuro-anatomical structures known to be among the earliest and most severely affected in Alzheimer's disease. Grad-CAM yields coarse localisation maps that typically encompass the affected region. On the other hand, Grad-CAM++ produces activations that are sharper and more precisely bounded, especially for smaller lesion regions in early stage cases. Overall, Score-CAM provides the smoothest heatmaps with activations nicely distributed across the relevant anatomical boundaries avoiding gradient-saturation artefacts sometimes observed in Grad-CAM outputs.

Fig 16 presents the disease progression visualisation, demonstrating a clinically plausible trend with progressive increase in activation intensity and spatial extent from the Non-Demented class, characterised by sparse and diffuse activations, through the Very Mild-Demented and Mild-Demented stages, to the Moderate-Demented class, where large, densely activated areas are identified covering the hippocampus, entorhinal cortex, and adjacent temporal areas. This monotonic progression of saliency is consistent with the known neurodegeneration trajectory of Alzheimer's disease and confirms the clinical plausibility of the model.

Analysis of misclassified predictions Fig. 15 shows that errors are mainly made at the Mild-Moderate boundary, where the model's activations suggest high spatial overlap with next-stage patterns. In high-confidence correct predictions Fig. 17, activations are sharply localised and anatomically specific, while low-confidence predictions show fragmented or laterally displaced activations, suggesting that classification uncertainty is associated with ambiguous or atypical activation patterns. Together these results demonstrate that the ResNet50 model not only outperforms in quantitative terms, but also offers interpretable and clinically meaningful decision regions.

VI. Discussion

A. Superiority of ResNet50 over the Custom CNN

The fine-tuned ResNet50 beats the Custom CNN by a margin of 6.55% in overall accuracy 95.58% vs. 89.03% and 4.31 points in macro F1-Score 0.9708 vs. 0.9277. This performance gap can be ascribed to three compounding advantages inherent to the transfer learning paradigm. First, ResNet50's pretrained weights, derived from training on over 1.2 million ImageNet images, encode hierarchical feature representations ranging from low-level edge detectors and texture filters in early layers to high-level structural patterns in deeper layers that works as a powerful initialization for medical image feature extraction. The Custom CNN, initialized from random weights via Kaiming Normal initialization, must learn all such representations entirely from the training data, which, despite the augmented set size of 33,984 samples, remains considerably smaller than the ImageNet corpus.

Second, the residual connections in ResNet50 enable stable gradient flow through 50 layers without the vanishing gradient degradation that can afflict deep networks trained from scratch. The Custom CNN's four-block architecture, while sufficient to capture coarse MRI features, lacks the representational depth required to discriminate subtle inter-stage morphological differences particularly at the clinically ambiguous Mild–Moderate boundary. This is directly evidenced in the confusion matrices: the Custom CNN produces 286 Mild $\rightarrow$ Moderate and 358 Moderate $\rightarrow$ Mild errors, compared to ResNet50's substantially lower 38 and 223 errors in the same cells.

Third, the selective layer-freezing strategy applied to ResNet50 freezing the stem, Layer1, and Layer2 while fine-tuning Layer3 and Layer4 effectively prevents destructive forgetting of generic low-level features while allowing the network's upper layers to specialize toward MRI-specific morphology. The layer-wise learning rate implementation 1×10^{-6} for frozen stages, 1×10^{-5} for deeper layers, 1×10^{-4} for the classification head further ensures that pretrained representations are preserved while the task-specific head adapts rapidly. This design choice is reflected in the smooth, decreasing validation loss of ResNet50 (0.8122 $\rightarrow$ 0.1059 over 15 epochs, in compare to the Custom CNN's validation spike at epoch 2 1.2092, indicating early instability from random initialization.

B. Why HybridRViT Underperforms ResNet50

HybridRViT model, despite its architectural sophistication, achieves the second position accuracy 92.84%, macro F1 = 0.9484, being surpassed by ResNet50 by 2.74% in accuracy and 0.0224 point in macro F1-Score. This result, while apparently counter-intuitive given the global self-attention capacity of the Transformer, aligns with well-established findings in the medical imaging literature on huge data-scale requirements for Vision Transformers [3].

The main drawback of the HybridRViT in the experimental setting is that its data-hungry nature of the Transformer encoder. In each of the 6 Transformer layers, the self-attention mechanism has to learn all pairwise relations at all 50 token

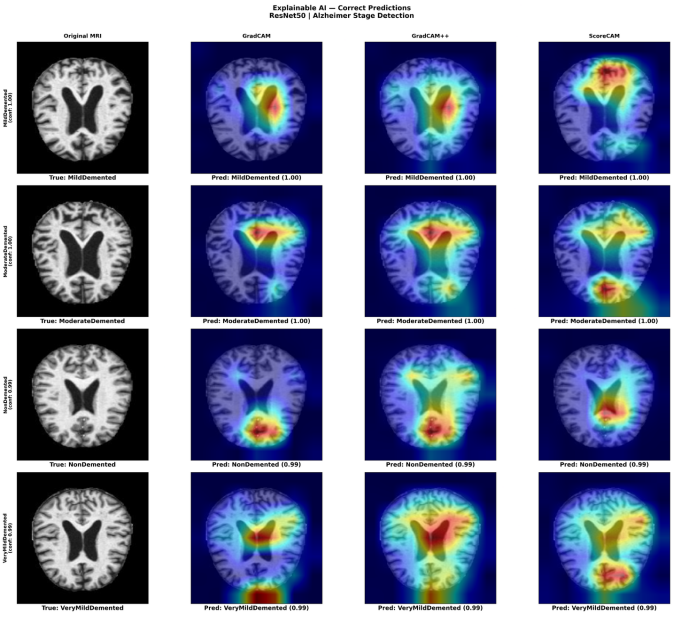

Fig. 14. Visualization of Grad-CAM, Grad-CAM++, and Score-CAM heatmaps generated from ResNet50 for correctly classified MRI samples

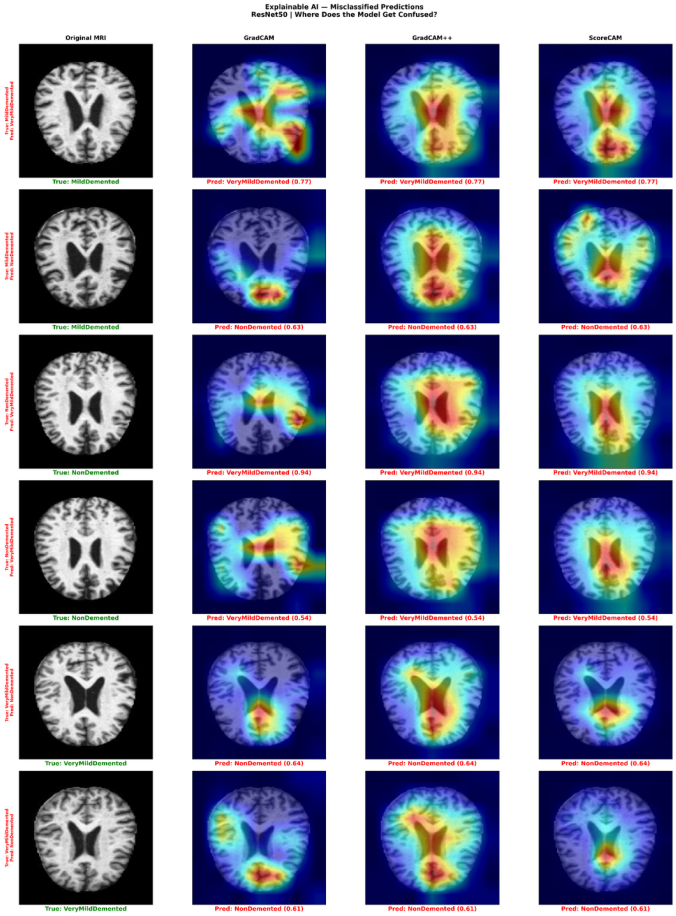

Fig. 15. Heatmaps of misclassified MRI samples using ResNet50 model with Grad-CAM, Grad-CAM++ and Score-CAM techniques.

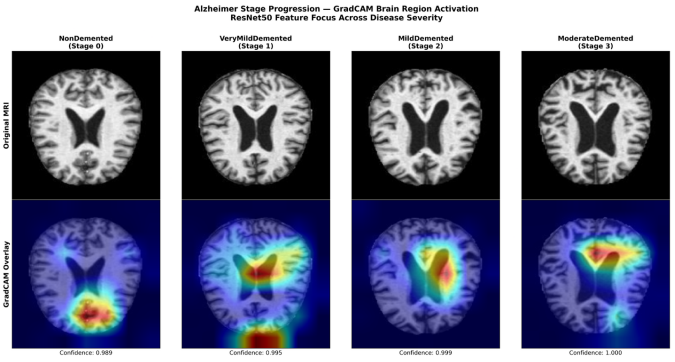

Fig. 16. Grad-CAM activation heatmaps depicting spatial evolution of class-discriminative regions across all four dementia severity stages for example MRI samples

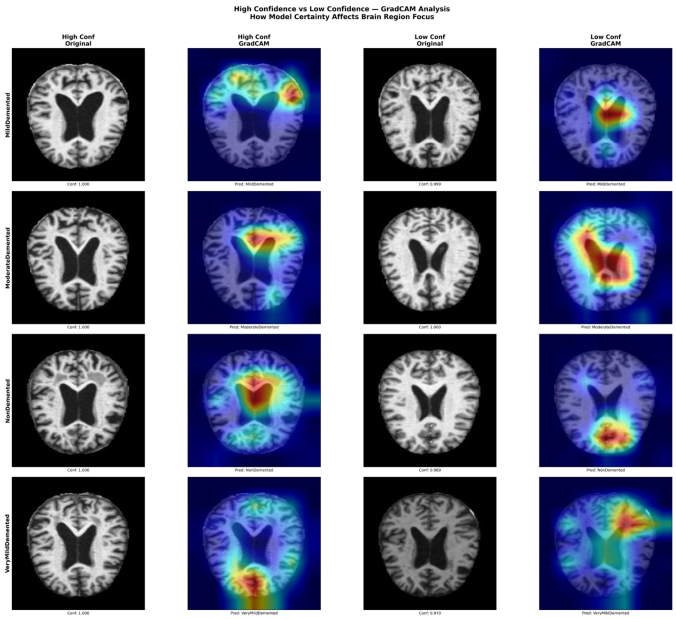

Fig. 17. Comparison of Grad-CAM++ heatmaps for high-confidence vs low-confidence ResNet50 predictions across dementia stages

positions (49 patch tokens + 1 CLS token), which produce 2,500 attention coefficients per layer per sample. In contrast, while the convolutional inductive biases inherently impose local spatial connectivity and translational equivariance, the self-attention mechanism of the Transformer has to learn these spatial priors from data. However, the Transformer encoder is not sufficiently exposed to the global relationships in a training set of 27,187 samples to reliably generalise these, especially for the morphologically subtle differences between AD's disease stages.

This limitation in data scale translates into two patterns that can be observed. First, the HybridRViT shows inactive and small oscillation in validation loss between epochs 11 to 13 (0.2128, 0.2132, 0.1810), a signature of the optimiser struggling to navigate the high-dimensional attention weight space without sufficient gradient signal diversity. Secondly, the most severe misclassification error in HybridRViT is the prediction of 49 Non-Demented samples as Moderate Demented, which is a clinically severe and anatomically incredible error, which is almost absent in ResNet50 only 7 such errors and the Custom CNN only 8 such errors. This suggests that the global attention mechanism at times attends to globally distributed but spurious features, classifying healthy brain anatomy as late-stage degeneration patterns a failure mode based on insufficient regularisation of the attention weights at this dataset scale.

C. *Clinical Importance of ResNet50's Per-Class Evaluation*

The per-class analysis show findings with direct clinical relevance. For Mild Demented class, ResNet50 achieve recall 0.9875 and for Moderate Demented class achieve 0.8996, both are higher than the corresponding Custom CNN values 0.9025 and 0.8299. In clinical screening contexts, recall or sensitivity is the more critical metric, as false negatives (missed Alzheimer's cases) carry more clinical cost than false positives. The 8.5 points recall improvement for the Mild Demented class is particularly significant, as mild-stage detection directly determines eligibility for disease-modifying intervention.

The Moderate Demented class presents the greatest challenge across all three models due to the severe imbalance in the test partition n = 64, representing only 1% of the test set. Despite this, ResNet50 achieves a Precision of 0.9782 and AUC of 0.9941 for this class, showing that the model's learned representations for late-stage dementia are highly discriminative even under extreme data scarcity. In Very Mild Demented class classification the all models are perfectly classify with Precision = Recall = F1 = AUC = 1.0000. While this result is convincing numerically, it likely reflects the fact that the Very Mild Demented class in the Kaggle dataset occupies a distinct feature space potentially due to the augmentation methodology applied by the dataset authors rather than implying that this stage is trivially detectable in clinical practice. This observation should be validated against external datasets before being generalized.

D. *Interpretation of XAI Findings and Clinical Alignment*

The use of the three XAI techniques on the ResNet50 model provides a multi-view of the spatial decision regions of the model that significantly improves clinical trustworthiness. The consistent activation of medial temporal lobe structures especially the hippocampus and entorhinal cortex across all three XAI methods for correctly classified samples aligns directly with the established neuropathological trajectory of AD's, where hippocampal atrophy is one of the earliest and most reliable biomarkers of neurodegeneration [9].

The intensity of activation is noticiable across dementia stages from sparse and diffuse activations in Non-Demented samples to dense, widespread temporal lobe activation in Moderate Demented cases provides a visually coherent and medically plausible representation of disease progression. This monotonic saliency gradient supports the model's internal consistency and suggests that the ResNet50 has learned a representation space that respecting the ordinal severity ordering of Alzheimer's stages, a property also reflected quantitatively in the high QWK of 0.9658. The three XAI methods combine complementary strengths, and so a richer diagnostic picture than any single technique. Grad-CAM's coarser activations that are useful for rapid, global region identification and would be most interpretable to radiologists unfamiliar with deep learning outputs.For sharp localisation Grad-CAM++'s is more helpful for identifying precise lesion boundaries and is particularly valuable for distinguishing adjacent-stage cases where the discriminative region may be spatially compact.As Score-CAM does not use gradients, the heatmaps it generates are smoother and less likely to be misinterpreted because of gradient saturation artifacts. Score-CAM may be preferred in high-stakes clinical reporting where heatmap stability is essential.

E. *Limitations*

- **Severe class imbalance in the test set**: The Moderate Demented class has only 64 test samples 1% of the test set, making per-class metrics for this stage are statistically less reliable. Performance on this class should be interpreted with proper caution.
- **Single-source dataset**: All experiments were performed on a single publicly available Kaggle dataset that was created from 2D axial MRI scan images.
- **Limited training epochs**: All models were trained on a fixed 15 epochs. While the ResNet50 validation loss converges well within this limited epochs, but the HybridRViT model might need longer training since the convergence of the Transformer-based architectures is slower.
- **Future XAI evaluation**: The XAI analysis applied in this study relies on visual inspection of heatmaps and qualitative alignment with neuroanatomical priors. Quantitative XAI evaluation metrics such as faithfulness, sensitivity, and localization accuracy against expert-annotated segmentation masks are not included and represent an important direction for future implementation.

VII. Conclusion

This study evaluated three deep learning architectures a Custom CNN, a fine-tuned ResNet50, and a Hybrid ResNet50+ViT (HybridRViT) on the four dementia classes from brain MRI scans. Experiments were conducted under strictly identical conditions on the Kaggle Augmented Alzheimer MRI Dataset that contains (33,984 training images; 6,400 test images). ResNet50 apperead as the best performing architecture across all ten evaluation metrics, achieving 95.58% accuracy, macro F1-Score of 0.9708, macro AUC of 0.9972, and QWK of 0.9658 outperforming the Custom CNN model by 6.55 percentage points and HybridRViT model by 2.74 points. Its advantage stems from ImageNet-pretrained weights which helped in feature extraction, residual skip-connection for gradient stability, and a selective layer-freezing strategy with layer-wise differential learning rates that control transfer learning with task adaptation. The HybridRViT, despite its architectural advantages, it underperformed ResNet50 due to the data scale sensitivity of its Transformer encoder, producing higher cross-stage confusion particularly at the Non-Demented and Very Mild Demented boundary confirming that dataset size remains a decisive constraint for hybrid CNN-Transformer models in medical imaging.

To improve clinical interpretability, three XAI techiniques were applied simultaneously to the ResNet50 model across four analytical scenarios that was correct predictions, misclassifications, disease progression, and prediction confidence levels. All three methods consistently activated the neuroanatomical structures of hippocampus, entorhinal cortex and medial temporal lobe with activation saliency intensifying monotonically from Non-Demented to Moderate Demented cases reflecting a clinically coherent representation of disease progression. Further it will extend to 3D MRI images, with multi-site datasets such as ADNI and OASIS, longitudinal progression modeling, and quantitative XAI evaluation against expert-annotated segmentation masks, with the longer-term goal of integrating the ResNet50 model and its explainability outputs into a radiologist-in-the-loop clinical decision support system.

References

[1] World Health Organization "Dementia" https://www.who.int/news-room/fact-sheets/detail/dementia

[2] Uraninjo, "Augmented Alzheimer MRI Dataset," Kaggle Available:[link]

[3] S. Sarraf and G. Tofighi, "Classification of Alzheimer's disease structural MRI data by deep learning convolutional neural networks," arXiv preprint arXiv:1607.06583, Jul. 2016. https://arxiv.org/abs/1607.06583

[4] El-Assy, A.M., Amer, H.M., Ibrahim, H.M. et al. A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data. Sci Rep 14, 3463 (2024). https://doi.org/10.1038/s41598-024-53733-6

[5] El-Latif, A.A.A.; Chelloug, S.A.; Alabdulhafith, M.; Hammad, M. Accurate Detection of Alzheimer's Disease Using Lightweight Deep Learning Model on MRI Data. Diagnostics 2023, 13, 1216. https://doi.org/10.3390/diagnostics13071216

[6] Dardouri S (2025) An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks. Front. Artif. Intell. 8:1563016. doi: 10.3389/frai.2025.1563016

[7] Al Shehri W. 2022. Alzheimer's disease diagnosis and classification using deep learning techniques. PeerJ Computer Science 8:e1177 https://doi.org/10.7717/peerj-cs.1177

[8] Shahid SB, Kaikaus M, Kabir MH, Yousuf MA, Azad AKM, Al-Moisheer AS, Alotaibi N, Alyami SA, Bhuiyan T and Moni MA (2025) Novel deep learning for multi-class classification of Alzheimer's in disability using MRI datasets. Front. Bioinform. 5:1567219. doi: 10.3389/fbinf.2025.1567219

[9] Amine JM and Mourad M (2025) Toward accurate Alzheimer's detection: transfer learning with ResNet50 for MRI-based diagnosis. Front. Neurosci. 19:1664418. doi: 10.3389/fnins.2025.1664418

[10] Mostafa, F.; Hossain, K.; Das, D.; Khan, H. Deep Learning Approaches with Explainable AI for Differentiating Alzheimer's Disease and Mild Cognitive Impairment. AppliedMath 2025, 5, 171. https://doi.org/10.3390/appliedmath5040171

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020. https://doi.org/10.48550/arXiv.2010.11929

[12] A. Berroukham, K. Housni and M. Lahraichi, "Vision Transformers: A Review of Architecture, Applications, and Future Directions," 2023 7th IEEE Congress on Information Science and Technology (CiSt), Agadir - Essaouira, Morocco, 2023, pp. 205-210, doi: 10.1109/CiSt56084.2023.10410015.

[13] Almufareh, M.F.; Tehsin, S.; Humayun, M.; Kausar, S. Artificial Cognition for Detection of Mental Disability: A Vision Transformer Approach for Alzheimer's Disease. Healthcare 2023, 11, 2763. https://doi.org/10.3390/healthcare11202763

[14] Hechkel, W., Leo, M., Carcagnì, P., Del-Coco, M., Helali, A. (2026). Enhancing Early Detection of Alzheimer's Disease via Vision Transformer Machine Learning Architecture Using MRI Images. Information, 17(2), 163. https://doi.org/10.3390/info17020163

[15] Şener, B., Açıcı, K. Sümer, E. Improving early detection of Alzheimer's disease through MRI slice selection and deep learning techniques. Sci Rep 15, 29260 (2025). https://doi.org/10.1038/s41598-025-14476-0

[16] D. Pantelaios, P.-A. Theofilou, P. Tzouveli, and S. Kollias, "Hybrid CNN-ViT models for medical image classification," in Proc. IEEE Int. Symp. Biomed. Imaging (ISBI), Athens, Greece, May 2024, pp. 1–4, doi: 10.1109/ISBI56570.2024.10635205.

[17] Hanhua Long. 2024. Hybrid Design of CNN and Vision Transformer: A Review. In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence (CISAI '24). Association for Computing Machinery, New York, NY, USA, 121–127. https://doi.org/10.1145/3703187.3703208

[18] Almalki, H., Khadidos, A.O., Alhebaishi, N. et al. Early detection of Alzheimer's disease progression stages using hybrid of CNN and transformer encoder models. Sci Rep 15, 16799 (2025). https://doi.org/10.1038/s41598-025-01072-5

[19] Authors (JISKA), "Comparative analysis of hybrid CNN-ViT and CNN for brain tumor classification," JISKA (Jurnal Informatika Sunan Kalijaga), vol. 11, no. 1, pp. 127–142, Jan. 2026, doi: 10.14421/jiska.5860.

[20] Singhal, A., Kumari, A.C. Srinivas, K. Transforming Alzheimer's diagnosis: ADNet deep learning with explainable AI framework. J Ambient Intell Human Comput 16, 1059–1072 (2025). https://doi.org/10.1007/s12652-025-04999-9

[21] Jahan, S., Saif Adib, M.R., Mahmud, M., Kaiser, M.S. (2023). Comparison Between Explainable AI Algorithms for Alzheimer's Disease Prediction Using EfficientNet Models. In: Liu, F., Zhang, Y., Kuai, H., Stephen, E.P., Wang, H. (eds) Brain Informatics. BI 2023. Lecture Notes in Computer Science(), vol 13974. Springer, Cham. https://doi.org/10.1007/978-3-031-43075-6$_3$1

[22] S. G. Mueller et al., "The Alzheimer's disease neuroimaging initiative," Neuroimaging Clin. N. Am., vol. 15, no. 4, pp. 869–877, Nov. 2005, doi: 10.1016/j.nic.2005.09.008.

[23] C. R. Jack Jr. et al., "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," Lancet Neurol., vol. 9, no. 1, pp. 119–128, Jan. 2010, doi: 10.1016/S1474-4422(09)70299-6.

[24] Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." International Journal of Computer Vision, edited by , vol. 128, no. 2, Oct. 2019, pp. 336–59. Crossref, https://doi.org/10.1007/s11263-019-01228-7.

[25] Chattopadhyay, Aditya Sarkar, Anirban Howlader, Prantik Balasubramanian, Vineeth. (2017). Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. 10.48550/arXiv.1710.11063.

[26] He, Y., Zheng, J. Zou, E. SHARPEN-CAM: efficient hierarchical SHAP-based visual explanation for deep convolutional neural networks. Multimedia Systems 32, 3 (2026). https://doi.org/10.1007/s00530-025-02025-8