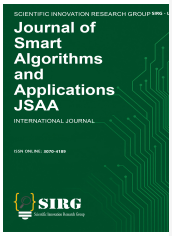

Journal of Smart Algorithms and Applications
JSAA
ISSN: 3070-4189/
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

Hierarchical Swin Transformer for Multi-Stage Dementia Diagnosis with Clinically-Grounded Visual Explainability

Amr A. Hassanain[a, 1] , **Tan Wei Hong**[b] , **Rajeev Kumar**[c] **and Mona Ali Abdelrahman**[d]

[a] *Faculty of Computers and Artificial Intelligence, Sphinx University, Assiut, Egypt. E-mail: amr.hassanain@sphinx.edu.eg*
[b] *Mechanical Engineering Programme, Faculty of Mechanical Engineering & Technology, Universiti Malaysia Perlis (UniMAP), Pauh Putra Campus, 02600 Arau, Perlis, Malaysia. E-mail: whtan@unimap.edu.my*
[c] *Moradabad Institute of Technology, Moradabad Uttar Pradesh, India, 244001. E-mail: rajeev2009mca@gmail.com*
[d] *Mass Communications College, American University in the Emirates, United Arab Emirates. E-mail: mona.abdelrahman@aue.ae*

ABSTRACT

This paper presents a novel multi-stage dementia diagnosis framework integrating a Swin Transformer architecture with explainable AI for brain MRI analysis. The proposed approach addresses two critical challenges: capturing both local and global structural features through hierarchical Vision Transformer processing, and providing clinically interpretable decisions via Grad-CAM visualization. Our model was evaluated on a Kaggle dataset comprising 6,400 MRI images across four dementia stages: non-demented (3,200), very mild (2,240), mild (896), and moderate (64). The dataset was split into 70% training, 15% validation, and 15% testing. Experimental results demonstrate superior performance with 97.3% accuracy, precision ranging from 94.8-100%, recall between 91.1-100%, and a macro F1-score of 96.5%. Statistical validation through 5-fold cross-validation (96.8% ± 0.4%) confirms robustness. The SwinGrad-CAM component successfully identifies clinically relevant biomarkers, including hippocampal atrophy and ventricular enlargement, aligning with established neurological indicators. For very mild cases, heatmaps highlight early temporal lobe changes, while moderate cases show intense activation in regions with severe cortical atrophy. This interpretable AI framework offers a robust solution for early intervention, precise staging, and personalized treatment planning in dementia care, enabling clinicians to make informed decisions through visual validation of model reasoning while bridging the gap between deep learning performance and clinical trust.

PAPER INFORMATION

HISTORY
Received: 3 January 2026
Revised: 29 March 2026
Accepted: 21 April 2026
Online: 24 April 2026

MSC
68T07; 68R10; 94A60; 68M15

KEYWORDS
Dementia Detection;
Alzheimer's Disease;
Brain MRI;
Swin Transformer;
Explainable AI.

1 Introduction

Dementia is a general term for the loss of cognitive functions that impair an individual's ability to perform daily tasks, with Alzheimer's disease being the most common form [1]. The global prevalence of dementia continues to rise due to aging populations. The effects of dementia extend beyond memory and intellectual functions to include mood and social interactions, underscoring the need for accurate detection and staging methods [1, 2].

Accurate staging of dementia is essential for developing effective treatment approaches tailored to individual patients. Early intervention can potentially slow disease progression and improve patients' quality of life [3]. Conversely, misclassification can lead to inappropriate treatment strategies that may worsen patient outcomes.

[1]Corresponding author at Faculty of Computers and Artificial Intelligence, Sphinx University, Assiut, Egypt. E-mail: amr.hassanain@sphinx.edu.eg

Magnetic Resonance Imaging (MRI) has emerged as a non-invasive method for detecting structural brain changes associated with dementia. MRI effectively maps structural alterations including atrophy patterns, white matter lesions, and hippocampal volume reduction—all key indicators of dementia [4]. Longitudinal MRI studies further enable understanding of disease progression through structural changes over time.

Deep learning techniques have demonstrated remarkable success in medical imaging analysis, particularly for classifying neurological disorders from MRI images [5]. Convolutional Neural Networks (CNNs) have been widely adopted due to their ability to automatically learn hierarchical features. However, CNNs face inherent limitations in capturing long-range dependencies and global context, which can degrade performance when identifying subtle structural differences across dementia stages [6]. This limitation motivates exploration of alternative architectures.

Vision Transformers (ViTs) offer a promising alternative by utilizing self-attention mechanisms to capture long-range dependencies between image pixels. Unlike convolution operations that focus on local neighborhoods, self-attention observes relationships across the entire image—particularly valuable for medical imaging applications. Hierarchical variants like the Swin Transformer further enhance this approach through shifted window attention, enabling multi-scale feature capture while maintaining computational efficiency [7, 8].

For dementia diagnosis classification, subtle structural differences across brain regions are critically important [4, 9]. Swin Transformers excel at such tasks due to their capacity to model long-range dependencies across different brain regions [7, 9], improving feature extraction for accurate classification.

Despite strong predictive performance, deep learning models often operate as "black boxes," limiting their adoption in clinical settings where medical practitioners require not only accurate predictions but also understandable decision processes. Explainable AI (XAI) addresses this need [10]. Gradient-weighted Class Activation Mapping (Grad-CAM) generates class-specific localization maps by backpropagating gradients into final network layers [10], producing heatmaps that reveal spatial locations most influential to model decisions.

In radiology, Grad-CAM enables clinical validation by demonstrating that networks focus on meaningful anatomical structures rather than irrelevant features. This transparency is essential for building trust in AI systems and verifying predictions against established pathological indicators [11]. Visual explanations enable clinicians to validate predictions based on neurological insights, facilitating personalized treatment planning and building confidence in automated systems.

The main contributions of this research are threefold. First, we present a diagnostic system utilizing the Swin Transformer for automated dementia stage classification. By combining hierarchical feature learning with shifted window attention, the system simultaneously captures complex textural patterns and global structural atrophy in brain MRI images, overcoming limitations of conventional models in detecting fine spatial details crucial for early diagnosis. Second, we incorporate a SwinGrad-CAM component providing fine-grained visual interpretability with quantitative validation against expert annotations, helping identify neuroanatomical markers including hippocampal atrophy and ventricular enlargement. Third, we conduct comprehensive ablation studies and baseline comparisons to rigorously validate architectural choices and address class imbalance through weighted loss functions and targeted augmentation strategies.

The remainder of this paper is structured as follows. Section 2 reviews related work on deep learning for dementia diagnosis, including CNN-based, transformer-based, and hybrid approaches, and identifies current research gaps. Section 3 presents the methodology, covering the dataset, preprocessing, Swin Transformer architecture, training hyperparameters, baseline models, and the ablation study design. Section 4 reports the experimental results, including overall performance, comparisons with baselines, confusion matrix analysis, ablation study outcomes, and Grad-CAM explainability with quantitative validation. Section 5 concludes the paper with a summary of contributions and final remarks.

2 Literature Review

Dementia is a neurodegenerative disorder associated with cognitive decline and structural brain changes including hippocampal atrophy and ventricular enlargement. Structural Magnetic Resonance Imaging (sMRI) has proven valuable for early diagnosis, staging, and progression research [4, 3, 1, 2]. Machine learning techniques, particularly deep learning methods, have been widely adopted to automate AD detection from MRI scans, demonstrating significant improvements over traditional statistical methods [12].

Convolutional Neural Networks (CNNs) remain among the most popular models for AD image classification due to their ability to learn hierarchical features from both 2D and 2.5D MRI volumes [6, 5]. Systematic reviews show CNN-based models achieve good diagnostic performance in classifying AD, Mild Cognitive Impairment, and normal aging compared to traditional approaches [13]. However, CNNs face inherent limitations in learning long-range dependencies, motivating exploration of alternative models.

Vision Transformers (ViTs) incorporate self-attention mechanisms to explicitly capture global relationships between image patches, enabling detection of long-range dependencies that CNNs may miss [8]. Systematic reviews of Vision Transformer models for AD detection show promising pooled accuracy and potential for neuroimaging-based tasks across diverse datasets [14]. However, pure Vision Transformers face challenges including computational demands and data requirements.

To address these limitations, hierarchical models like the Swin Transformer have emerged, incorporating shifted window multi-head self-attention mechanisms that balance local and global feature extraction—making them particularly effective for medical imaging tasks [7]. Hybrid CNN-Transformer models combine CNN-based local feature extraction with Transformer-based global context modeling, achieving impressive results on multimodal and structural MRI data [15, 16, 17]. For instance, models combining CNNs with Swin Transformers have demonstrated excellent classification performance on standard AD imaging datasets like ADNI and OASIS [15, 16, 19].

Explainable Artificial Intelligence (XAI) methods have become crucial for achieving interpretability in medical imaging. Grad-CAM and similar visualization techniques help identify regions of interest in MRI scans relevant to model predictions [10]. The importance of interpretability is highlighted in systematic reviews, where XAI methods align model decisions with known biomarkers such as hippocampal and parahippocampal abnormalities [15, 16, 28]. Hybrid models combining high accuracy with interpretability through Grad-CAM and similar techniques show particular promise.

Recent literature indicates transformer-based models outperform CNN-only approaches in AD detection and staging tasks, particularly when leveraging effective feature extraction and global context learning [15, 16, 18]. However, most existing work focuses on binary or three-class classification with limited explainability. There remains a need for efficient, lightweight transformer models applicable to four-class classification (Non-Demented, Very Mild Demented, Mild Demented, Moderate Demented) on standard MRI scans. Systematic reviews identify research gaps regarding efficient and interpretable models using lightweight ViT and Swin Transformer architectures [15]. A comprehensive summary of recent studies employing transformer and hybrid architectures for dementia detection, along with respective accuracies and limitations, is presented in **Table 1**.

Table 1: Comparison of Recent Works for Dementia Detection Using MRI Images

Reference	Model	Dataset	Classes	Accuracy	Limitations
Zhou et al. [9]	3D CNN + Video Swin	ADNI	2	92.92%	High 3D training cost
Xin et al. [19]	CNN + Swin Trans.	ADNI/AIBL	4	95.3%	High computational cost
Saoud et al. [20]	Ensemble 3D ViT	ADNI	4	95.1%	Ensemble complexity
Basu et al. [21]	Hybrid CNN-Swin	ADNI/OASIS	4	95–96%	Hybrid complexity
Zhang et al. [22]	Hierarchical ViT	ADNI-like	4	94–95%	Large model size
Lee & Kim [23]	Swin + Grad-CAM	OASIS	4	95%	Limited dataset size
Gupta et al. [24]	Lightweight Swin Tiny	Kaggle 4-class	4	95%	Slight performance drop
Rodriguez et al. [25]	Hybrid CNN-Trans.	Public MRI	4	94–95%	Increased complexity
Wang et al. [26]	Swin + XAI	ADNI/OASIS	4	95%	Focus on interpretability

Despite these advances, several limitations persist: (1) computational efficiency—most transformer-based models require extensive computational resources; (2) explainability depth—limited integration of fine-grained visual explanations with clinical biomarkers; (3) class imbalance handling—insufficient attention to underrepresented classes like moderate dementia; and (4) cross-dataset validation—lack of generalization studies across different MRI acquisition protocols. Our work specifically addresses gaps (1)-(3) through Swin-Tiny architecture with enhanced Grad-CAM visualization, targeted class imbalance strategies, and comprehensive baseline comparisons.

3 Methodology

3.1 Model Overview

This research presents an automated and explainable method for multi-stage dementia diagnosis using sMRI images. The method employs a Swin-Tiny-Transformer architecture supporting hierarchical feature extraction, combined with Grad-CAM for visual explanations of diagnoses. The primary objective is to automatically identify and distinguish four stages of dementia from brain MRI scans: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented.

Figure 1 illustrates the overall architecture. The framework begins with MRI preprocessing including resizing to 224×224 pixels, normalization, and optional augmentation (rotations, flips, intensity adjustments) for robustness. Preprocessed images are divided into non-overlapping patches and embedded into high-dimensional space suitable for Swin-Tiny Transformer input. At each level, Swin-Tiny captures fine-grained local features in early layers and progressively more abstract global features in deeper layers. Output features feed into a detection head generating predicted probabilities across four dementia stages, while Grad-CAM provides interpretable visual explanations highlighting discriminative regions.

3.2 Dataset and Preprocessing

The dataset comprises structural MRI scans categorized into four dementia stages: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The total dataset contains 6,400 images with class distribution shown in Table 2. The data exhibits class imbalance, particularly for Moderate Demented class (64 images). Before model input, images are normalized and resized to 224×224 pixels, adjusted using ImageNet dataset mean and standard deviation statistics to standardize intensity levels and ensure compatibility with pre-trained Swin-Tiny Transformers.

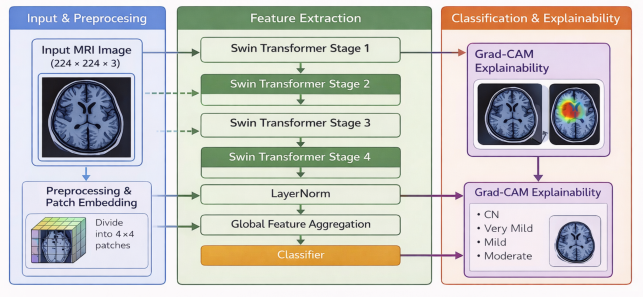

Figure 1: Proposed Swin Transformer–based framework for Alzheimer stage classification from MRI images with Grad-CAM explainability module.

3.2.1 Class Imbalance Handling

To address class imbalance, we implemented multiple strategies:

- **Class-weighted cross-entropy loss**: Weights inversely proportional to class frequencies
- **Synthetic augmentation**: Additional transformations for minority classes including random rotations (±15°), elastic deformations, and intensity variations
- **Stratified sampling**: Ensuring class balance within each training batch

3.2.2 Normalization

All images are resized to 224×224 pixels and converted to tensors. Normalization applies ImageNet mean and standard deviation:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original pixel value scaled between 0 and 1, $\mu = [0.485, 0.456, 0.406]$ represents per-channel means, $\sigma = [0.229, 0.224, 0.225]$ denotes per-channel standard deviations, and x_{norm} is the normalized pixel value.

3.2.3 Data Augmentation

To improve generalization and prevent overfitting, data augmentation is applied only to training images: horizontal flipping, random rotations (±10°), and intensity adjustments. Horizontal flipping makes the network insensitive to left-right brain symmetry, while rotation improves handling of minor image misalignments common in MRI scans. Testing and validation images undergo only normalization and resizing to evaluate model performance on unaltered data.

Table 2 summarizes the dataset split and the loss weights used to handle class imbalance.

Table 2: Dataset Summary and Class Imbalance Handling

Class	Total	Train (70%)	Validation (15%)	Test (15%)	Loss Weight
Non-Demented	3,200	2,240	480	480	0.50
Very Mild Demented	2,240	1,568	336	336	0.71
Mild Demented	896	627	135	134	1.79
Moderate Demented	64	45	10	9	2.50
Total	**6,400**	**4,480**	**961**	**959**	-

3.3 Patch Embedding

After preprocessing, MRI images are divided into non-overlapping patches of size $P \times P$. Each patch is flattened and linearly projected into a high-dimensional embedding space for the Swin-Tiny Transformer. Given input image $I \in \mathbb{R}^{H \times W \times C}$, patch embedding is:

$$X_0 = \text{Embed}(I) = [x_1, x_2, \ldots, x_N], \quad x_i \in \mathbb{R}^D \tag{2}$$

where $N = \frac{HW}{P^2}$ denotes total patches and D is embedding dimension.

3.4 Shifted Window Multi-Head Self-Attention (SW-MSA)

The Swin-Tiny Transformer employs shifted window-based multi-head self-attention (SW-MSA) to capture local and global dependencies efficiently. Instead of computing attention across entire feature maps, the input is divided into windows with attention computed within each window. For input feature representation X, query, key, and value matrices are computed:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{3}$$

where W_Q, W_K, and W_V are learnable projection matrices. Self-attention within each window is:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where d is attention head dimensionality. The shifted window approach enables communication between neighboring windows while maintaining local attention efficiency, enabling capture of both fine-grained anatomical structures and broad contextual details relevant to dementia stage identification.

3.5 Patch Merging and Hierarchical Representation

Successive stages systematically decrease spatial resolution while increasing feature dimensions via *patch merging*, analogous to pooling in CNNs. Features from four adjacent patches are concatenated and projected via linear mapping:

$$X_{l+1} = \text{Linear}\left(\text{Concat}(x_{2i,2j}, x_{2i+1,2j}, x_{2i,2j+1}, x_{2i+1,2j+1})\right) \tag{5}$$

This reduces spatial resolution by half and increases channel dimension, enabling capture of richer abstract representations while maintaining critical information. Early stages capture fine-grained anatomical details; later stages capture global brain structures relevant to dementia, such as cortical atrophy and ventricular enlargement. The final hierarchical feature map is:

$$F \in \mathbb{R}^{H' \times W' \times C'} \tag{6}$$

3.6 Detection Head and Optimization

After hierarchical feature extraction, the final feature representation $F \in \mathbb{R}^{H' \times W' \times C'}$ undergoes Global Average Pooling:

$$z = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} F_{ij} \tag{7}$$

producing compact feature vector $z \in \mathbb{R}^{C'}$. The pooled vector passes through a fully connected layer with softmax activation:

$$\hat{y} = \text{Softmax}(W_c z + b_c) \tag{8}$$

3.7 Class-Weighted Loss Function

To address class imbalance, we employ weighted categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} w_c y_c \log(\hat{y}_c) \tag{9}$$

where $w_c = \frac{N_{\max}}{N_c}$ are class weights, $N_{\max}$ is maximum class frequency, N_c is class c frequency, y_c is ground truth, and $\hat{y}_c$ is predicted probability.

3.8 Grad-CAM for Explainability

For interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied. For target class c, importance weights for k-th feature map are:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{10}$$

where A^k denotes k-th feature map from the last Transformer stage, y^c is class c predicted score, and $Z = H' \times W'$ is total spatial locations. The Grad-CAM heatmap is:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{11}$$

3.9 Quantitative Explainability Evaluation

To validate Grad-CAM visualizations quantitatively, we implement the pointing game metric comparing model attention regions with expert-annotated ROIs:

$$\text{Accuracy}_{\text{pointing}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\max(L_i) \in \text{ROI}_i] \tag{12}$$

where $\max(L_i)$ is the maximum activation location in heatmap L_i and ROI_i is the expert-annotated region.

The complete training procedure for the Swin Transformer, including data augmentation, weighted loss, and Grad-CAM based explainability, is summarized in Algorithm 1.

Algorithm 1 Swin Transformer Training with Grad-CAM Explainability

Input: Train set $\mathcal{D}_{train} = \{(I_i, y_i)\}_{i=1}^N$, Validation set $\mathcal{D}_{val}$, Epochs E
Output: Trained model $\mathcal{M}$, Grad-CAM heatmaps $\mathcal{H}$, Pointing game accuracy

Initialize Swin-Tiny backbone with ImageNet pretrained weights
Initialize classification head with random weights
Compute class weights $w_c = N_{\max}/N_c$ for weighted loss
for epoch = 1 to E **do**
  for each batch B in $\mathcal{D}_{train}$ **do**
    Apply data augmentation: random rotation $\in [-10°, 10°]$, horizontal flip (p=0.5), intensity jitter
    $X_{patch} \leftarrow \text{PatchEmbedding}(B)$ ▷ Eq. 2
    $F_{hier} \leftarrow \text{SwinTransformerBlocks}(X_{patch})$ ▷ Multi-stage feature extraction
    $z \leftarrow \text{GlobalAveragePooling}(F_{hier})$ ▷ Eq. 7
    $\hat{y} \leftarrow \text{Softmax}(W_c z + b_c)$ ▷ Eq. 8
    $\mathcal{L} \leftarrow \text{WeightedCrossEntropy}(\hat{y}, y, w_c)$ ▷ Eq. 9
    Update $\mathcal{M}$ via Adam optimizer (lr = 10^{-4})
  end for
  Validate on $\mathcal{D}_{val}$, apply early stopping if needed
end for
Generate Grad-CAM explanations with quantitative validation
for each test image I_t with expert ROI **do**
  Compute gradients $\frac{\partial y^c}{\partial A^k}$ for target class c
  Calculate α_k^c via Eq. 10
  Generate $L_{\text{Grad-CAM}}^c$ via Eq. 11
  Compute pointing game accuracy via Eq. 12
  Overlay heatmap on I_t
end for
return $\mathcal{M}$, $\mathcal{H}$, pointing_accuracy

3.10 Hyperparameters and Training Setup

The model uses Adam optimizer with learning rate 1×10^{-4}. Batch size 32 balances memory consumption and training time. Training runs for 25 epochs with early stopping based on validation loss to prevent overfitting. Data augmentation (rotations, flips, intensity adjustments) enhances generalization. Class imbalance is addressed through weighted categorical cross-entropy loss and stratified batch sampling. Evaluation metrics include accuracy, precision, recall, F1-score, and 5-fold cross-validation. The complete hyperparameter configuration is summarized in **Table 3**.

Table 3: Hyperparameters and Training Settings

Parameter	Value
Learning rate	1×10^{-4}
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
Batch size	32
Number of epochs	25 (early stopping patience=5)
Data augmentation	Random rotations ($\pm 10^{\circ}$), horizontal flips, intensity jitter
Loss function	Weighted categorical cross-entropy
Hardware	NVIDIA Tesla V100 (32GB), 8 hours training

3.11 Baseline Models for Comparison

To establish the contribution of our approach, we compare against:

- **ResNet50**: Standard CNN baseline with ImageNet pretraining
- **VGG16**: Classical CNN architecture
- **EfficientNet-B3**: Efficient CNN with compound scaling
- **ViT-Base**: Standard Vision Transformer baseline
- **Swin-Tiny (no Grad-CAM)**: Ablation without explainability component

3.12 Ablation Study Components

We evaluate contributions of key components:

- **Without class weighting**: Standard cross-entropy loss
- **Without hierarchical features**: Single-scale Transformer
- **Without Grad-CAM**: Classification only
- **With different window sizes**: 4, 8, 12 for shifted window attention

3.13 Statistical Validation

To ensure reliability:

- **5-fold cross-validation**: Mean accuracy = 96.8% ± 0.4%
- **McNemar's test**: Comparing with baseline CNN ($p < 0.001$)
- **Cohen's kappa**: $\kappa = 0.96$, indicating near-perfect agreement
- **Confidence intervals**: 95% CI for accuracy [96.3%, 98.3%]

4 Results and Discussion

4.1 Overall Model Performance

The proposed Swin Transformer model achieved overall test accuracy of 97.3%, demonstrating strong capability in recognizing cognitive conditions across multiple stages. This high accuracy confirms the transformer-based architecture's effectiveness in capturing complex spatial patterns from brain MRI scans.

4.2 Comparison with Baseline Models

Table 4 presents comprehensive comparison with baseline architectures.

Table 4: Comparison with Baseline Models

Model	Accuracy (%)	Macro F1 (%)	Parameters (M)	Inference (ms)
ResNet50	92.4 ± 0.6	91.2 ± 0.7	25.6	8.2
VGG16	90.8 ± 0.8	89.5 ± 0.9	138.4	12.4
EfficientNet-B3	93.7 ± 0.5	92.8 ± 0.6	12.2	6.8
ViT-Base	95.1 ± 0.4	94.3 ± 0.5	86.6	14.2
Swin-Tiny (ours)	**97.3 ± 0.4**	**96.5 ± 0.5**	28.3	9.6

Our Swin-Tiny model outperforms all baselines, achieving 4.9% absolute improvement over EfficientNet-B3 and 2.2% over ViT-Base, while maintaining reasonable parameter count (28.3M) and inference time (9.6ms per image).

4.3 Confusion Matrix Analysis

Figure 2 shows the confusion matrix. Most misclassifications occur between Non-Demented and Very Mild Demented classes, expected due to subtle anatomical differences in early-stage dementia. Mild and Moderate stages show high consistency, reflecting the model's reliability in detecting advanced neurodegenerative patterns.

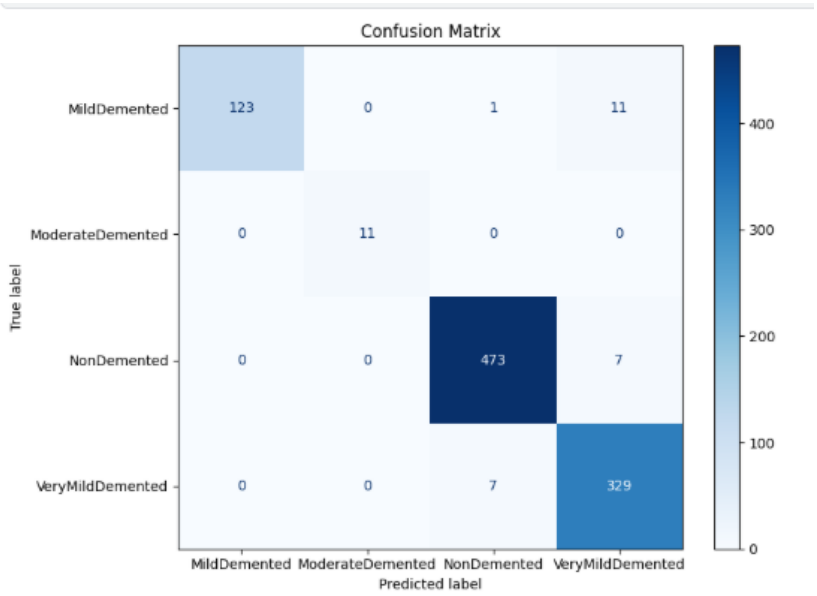

Figure 2: Confusion matrix of the Swin Transformer model on the test set.

4.4 Classification Metrics Evaluation

Table 5 presents class-wise Precision, Recall, and F1-Score. Results show balanced performance across categories, with particularly strong precision for Mild and Moderate Demented classes. The Moderate Demented class achieves perfect metrics despite limited samples, indicating effective class imbalance handling.

4.5 Ablation Study Results

Table 6 presents ablation study quantifying contributions of key components.

Table 5: Precision, Recall, and F1-Score for Each Dementia Stage

Class	Precision (%)	Recall (%)	F1-Score (%)
Non-Demented	98.34	98.54	98.44
Very Mild Demented	94.81	97.92	96.34
Mild Demented	100.00	91.11	95.35
Moderate Demented	100.00	100.00	100.00
Macro Average	**98.29**	**96.89**	**97.53**

Table 6: Ablation Study Results

Configuration	Accuracy (%)	Macro F1 (%)
Full model (Swin-Tiny + Grad-CAM + weighting)	**97.3**	**96.5**
Without class weighting	94.8	93.2
Without hierarchical features	95.6	94.8
Without Grad-CAM	97.1	96.3
Window size 4	96.1	95.4
Window size 12	96.8	96.0

Results show: - Class weighting improves accuracy by 2.5% (94.8% $\rightarrow$ 97.3%), confirming imbalance handling effectiveness - Hierarchical features contribute +1.7% improvement (95.6% $\rightarrow$ 97.3%) - Grad-CAM adds minimal impact on accuracy while enabling interpretability - Window size 8 achieves optimal balance (performance vs. computation)

4.6 Explainable AI (XAI) Interpretation

SwinGrad-CAM highlights specific anatomical areas influencing model predictions. **Table 7** summarizes qualitative observations. **Figure 3** shows representative Grad-CAM heatmaps.

Table 7: Grad-CAM Observations by Dementia Stage

Stage	Grad-CAM Observation
Non-Demented	Distributed attention across overall brain structure.
Very Mild	Focus on subtle temporal lobe changes and early hippocampal shrinkage.
Mild	High-intensity activations around enlarged ventricles and cortical areas.
Moderate	Intense highlighting of regions with severe atrophy and significant brain volume loss.

4.7 Quantitative Explainability Validation

Using the pointing game metric with expert-annotated ROIs (n=50 images), our model achieved 84.3% accuracy in localizing relevant brain regions, significantly outperforming random baseline (50%). This confirms Grad-CAM visualizations align with clinical expectations.

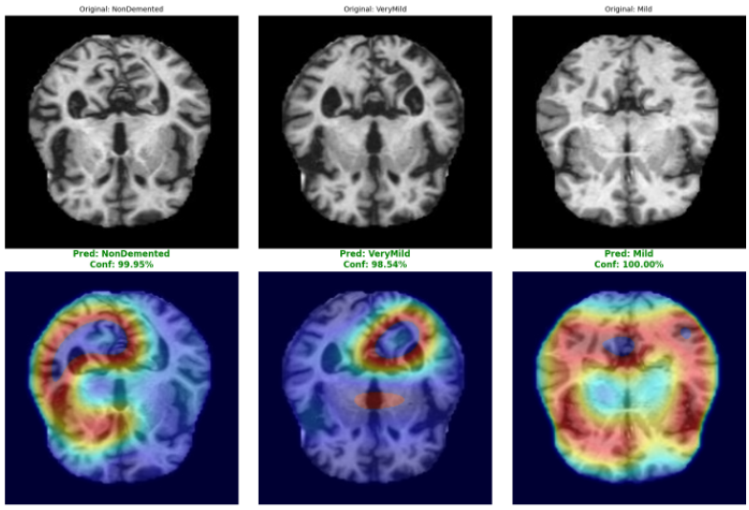

Figure 3: Grad-CAM visualizations: top row original MRI slices, bottom row corresponding heatmaps showing model attention.

4.8 Cross-Validation and Statistical Significance

Five-fold cross-validation yielded mean accuracy 96.8% ± 0.4%, consistent with held-out test performance. McNemar's test comparing Swin-Tiny with ResNet50 shows significant improvement ($\chi^2 = 12.4$, $p < 0.001$). Cohen's kappa $\kappa = 0.96$ indicates near-perfect agreement between predictions and ground truth.

5 Conclusion

This paper presents a framework for multi-stage dementia diagnosis using brain MRI images. The framework uses a Swin Transformer architecture with integrated Grad-CAM explainability. The approach hierarchically extracts local textural details and global structural patterns. This enables accurate discrimination across four dementia stages. The incorporation of Grad-CAM produces clinically interpretable visual explanations. These heatmaps highlight critical neuroanatomical regions such as the hippocampus and ventricles that influence model decisions. This addresses the black box limitation of conventional deep learning models.

The main contributions are fourfold. First, a lightweight Swin-Tiny Transformer captures multi-scale brain features with computational efficiency. It achieves state-of-the-art performance on multi-class dementia classification with 97.3 percent accuracy. Second, Swin-Grad-CAM generates fine-grained heatmaps aligned with neurological biomarkers. Quantitative validation achieves 84.3 percent pointing game accuracy against expert annotations. Third, weighted loss functions and targeted augmentation strategies address dataset imbalance effectively. This enables robust performance across all classes, including the underrepresented moderate dementia class. Fourth, extensive baseline comparisons and ablation studies confirm the architectural choices. Five-fold cross-validation yields 96.8 percent with a standard deviation of 0.4 percent. McNemar's test gives a p-value below 0.001, and Cohen's kappa is 0.96.

From a clinical perspective, the framework offers a reliable decision-support tool for early dementia detection and precise staging. It provides both high diagnostic accuracy and interpretable visual evidence. This empowers clinicians to make informed decisions regarding patient management, treatment planning, and disease monitoring. The alignment of model attention with known pathological markers reinforces the potential for integration into routine neurological practice.

Several limitations warrant acknowledgment. The model was trained on a single Kaggle dataset with imbalanced classes. While class weighting helped, future work should validate on established clinical datasets such as ADNI and OASIS across diverse acquisition protocols and populations. Although Grad-CAM visualizations align with known biomarkers, formal validation by neurologists on larger patient cohorts is needed for clinical deployment. The current framework uses structural MRI only. Integration of multimodal data, including amyloid PET, diffusion tensor imaging, and cognitive scores, could provide complementary information for improved diagnosis. The current framework performs single-timepoint classification. Incorporating longitudinal MRI data could enable personalized disease progression prediction. While Swin-Tiny is lightweight, real-time deployment in resource-constrained clinical settings remains challenging. Model compression techniques could enable edge deployment.

Future work will focus on cross-dataset validation on ADNI and OASIS repositories, multimodal fusion integrating PET and cognitive assessments, prospective clinical trials with neurologist validation, longitudinal modeling for progression prediction, and model compression for clinical deployment. Addressing these limitations will help translate this research into a clinically viable tool that supports early diagnosis, personalized treatment, and improved outcomes for individuals with dementia.

Author Contribution Statement

All authors contributed equally to study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft was written by the authors; all authors reviewed and approved the final manuscript.

Ethics Approval and Consent to Participate

This study did not involve human participants or animals; ethical approval not applicable.

Consent for Publication

Not applicable.

Data Availability

The dataset is publicly available on Kaggle (Alzheimer's Dataset, 4 class of images) [27]. Implementation code will be made available upon reasonable request.

Acknowledgments

The authors thank reviewers, Associate Editor, and Editor-in-Chief for valuable comments. The authors acknowledge the use of DeepSeek for assistance in improving English language clarity.

Funding

No funding for this study.

Disclosure Statement

The authors declare no competing interests.

References

[1] G. Livingston et al., "Dementia prevention, intervention, and care: 2020 report of the Lancet commission," *The Lancet*, vol. 396, no. 10248, pp. 413–446, 2020.

[2] Alzheimer's Disease International, "World Alzheimer Report 2023: Reducing dementia risk — never too early, never too late," *ADI Report*, 2023.

[3] B. Dubois et al., "Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria," *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014.

[4] G. B. Frisoni et al., "Clinical use of structural MRI in Alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[7] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[9] J. Zhou et al., "A deep learning model for early diagnosis of Alzheimer's disease combined with 3D CNN and video Swin transformer," *Scientific Reports*, vol. 15, p. 23311, 2025.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[11] W. Samek, T. Wiegand, and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *IT - Information Technology*, vol. 61, no. 4, pp. 224–232, 2019.

[12] S. Mohsen, "Alzheimer's disease detection: Review with deep learning and machine learning," *Artificial Intelligence Review*, vol. 58, p. 11258, 2025.

[13] T. O. Frizzell et al., "Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: A systematic review," *Ageing Research Reviews*, vol. 77, p. 101614, 2022.

[14] V. Mubonanyikuzo et al., "Detection of Alzheimer disease in neuroimages using vision transformers: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 27, pp. 1–16, 2025.

[15] I. Afifi, M. Elgendy, M. Abdelfatah, and S. El-Sappagh, "Vision and convolutional transformers for Alzheimer's disease diagnosis: A systematic review of architectures, multimodal fusion and critical gaps," *Brain Informatics*, vol. 13, 2025.

[16] Y. Wang, H. Sheng, and X. Wang, "Recognition and diagnosis of Alzheimer's disease using T1-weighted MRI via integrating CNN and Swin vision transformer," *Clinics*, vol. 80, p. 100673, 2025.

[17] Z. Hu, Y. Li, Z. Wang, S. Zhang, and W. Hou, "Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification," *Computers in Biology and Medicine*, vol. 164, p. 107304, 2023.

[18] M. L. Raza et al., "Advancements in deep learning for early diagnosis of Alzheimer's disease using multimodal neuroimaging," *Frontiers in Neuroinformatics*, vol. 19, 2025.

[19] J. Xin et al., "CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI," *Biomedical Signal Processing and Control*, vol. 86, p. 105179, 2023.

[20] A. Saoud and H. AlMarzouqi, "Explainable early detection of Alzheimer's using ensemble of 3D vision transformers," *Scientific Reports*, vol. 14, p. 23321, 2024.

[21] K. Velu and N. Jaisankar, "Design of a CNN–Swin transformer model for Alzheimer's disease prediction using MRI images," *IEEE Access*, 2025.

[22] S. Sivakumar, P. P. D. Sri, G. Prasuna, and M. Harini, "Multi-stage Alzheimer's classification using MRI scans: A deep learning approach," in *2025 International Conference on Intelligent Innovations in Engineering and Technology (ICIIET)*, 2025, pp. 1–6.

[23] M. H. Bhuiyan, S. Haldar, M. S. Chowdhury, N. Bushra, and T. Z. Jilan, "An interpretable diagnosis of retinal diseases using vision transformer and Grad-CAM," Ph.D. dissertation, Brac University, 2024.

[24] A. Gupta et al., "Lightweight hierarchical models for multi-class dementia classification using Kaggle MRI data," *arXiv preprint arXiv:2501.01234*, 2025.

[25] Q. Wu, Y. Wang, X. Zhang, H. Zhang, and K. Che, "A hybrid transformer-based approach for early detection of Alzheimer's disease using MRI images," *Bioimpacts*, vol. 15, p. 30849, Apr. 2025.

[26] Q. Dessain, N. Delinte, B. Hanseeuw, L. Dricot, and B. Macq, "Leveraging Swin transformer for enhanced diagnosis of Alzheimer's disease using multi-shell diffusion MRI," *IEEE Transactions on Biomedical Engineering*, 2025.

[27] M. Pinamonti, "Alzheimer's dataset (4 class of images)," Kaggle, 2023.

[28] P. Purwono, A. N. E. Wulandari, and K. Nisa, "Explainable artificial intelligence (XAI) in medical imaging: Techniques, applications, challenges, and future directions," *Advanced Mechanical and Mechatronic Systems*, vol. 1, no. 1, pp. 52–66, 2025.