

Article

A Hybrid Architecture of CNN–Swin-T Integrated with Attention Mechanism and Explainable AI for Alzheimer’s Disease Classification

Saeed Mohsen ^{1,2,*} , Saada Khadragy ³, Norah Alnaim ⁴ , Noorah Albehajjan ⁴  and Ahmed F. Ibrahim ³

¹ Department of Electronics and Communications Engineering, Al-Madinah Higher Institute for Engineering and Technology, Giza 12947, Egypt

² Department of Artificial Intelligence Engineering, Faculty of Computer Science and Engineering, King Salman International University (KSU), El Tor, South Sinai 46511, Egypt

³ College of Arts and Sciences, Umm Al Quwain University, Umm Al Quwain P.O. Box 536, United Arab Emirates

⁴ Department of Computer Science, College of Sciences and Humanities in Jubail, Imam Abdulrahman Bin Faisal University, Dammam 34212, Saudi Arabia

* Correspondence: saidconnections@gmail.com

Abstract

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder that requires early and accurate diagnosis to improve patient outcomes. In this paper, an attention-enhanced hybrid deep learning (DL) framework is proposed that combines Convolutional Neural Network (CNN) and Swin Transformer (Swin-T) architectures for multi-class Alzheimer’s classification. The proposed model integrates an attention mechanism to enhance feature representation and improve classification performance. Experiments are conducted on a dataset containing three classes: Mild Demented, Very Mild Demented, and Non-Demented. To improve the model’s generalization, data augmentation techniques are applied to enhance the model’s performance. Additionally, three explainable artificial intelligence (XAI) techniques are employed, including Grad-CAM++, Integrated Gradients, and Saliency maps, to interpret the model’s predictions and to provide visual insights into decision-making processes. The proposed attention-enhanced hybrid CNN–Swin-T model achieves a testing accuracy of 99.92% and reaches 99.71%, 99.73%, and 99.72%, for precision, recall, and F1-score, respectively. The hybrid CNN–Swin-T with attention outperforms three implemented models: baseline CNN, standalone Swin-T, and hybrid CNN–Swin-T. The explainability results validate the proposed model’s focus on relevant regions, increasing trust in automated diagnosis systems. Finally, a comparative analysis with an ablation study is presented to demonstrate that the integration of the attention mechanism with a hybrid CNN–Swin-T architecture leads to the highest performance and more reliable predictions compared to the other three models.



Academic Editor: Paolo Bellavista

Received: 2 May 2026

Revised: 29 May 2026

Accepted: 1 June 2026

Published: 3 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Alzheimer’s disease; deep learning; CNN; swin transformer; attention mechanism; explainable AI; medical image classification; data augmentation

1. Introduction

AD is a progressive neurodegenerative disorder and one of the leading causes of dementia worldwide, affecting millions of individuals and posing a significant global health challenge. The disease is marked by a progressive deterioration in cognitive abilities, including memory impairment and reduced reasoning skills, and behavioural changes [1],

which severely impact patients' ability to perform daily activities independently. With the continuous aging of the global population, the prevalence of AD is projected to rise substantially, increasing the strain on healthcare systems. Early and accurate diagnosis is therefore essential for enabling timely intervention, improving patient management, and potentially slowing disease progression. In this regard, medical imaging techniques, particularly brain scans, play a crucial role in identifying structural and functional abnormalities associated with different stages of AD, providing valuable support for clinical diagnosis [2]. As shown in Figure 1, there are several causes that lead to AD for individuals, such as age, smoking, head injury, genetics, cardiovascular disease, diabetes, depression, and obesity.

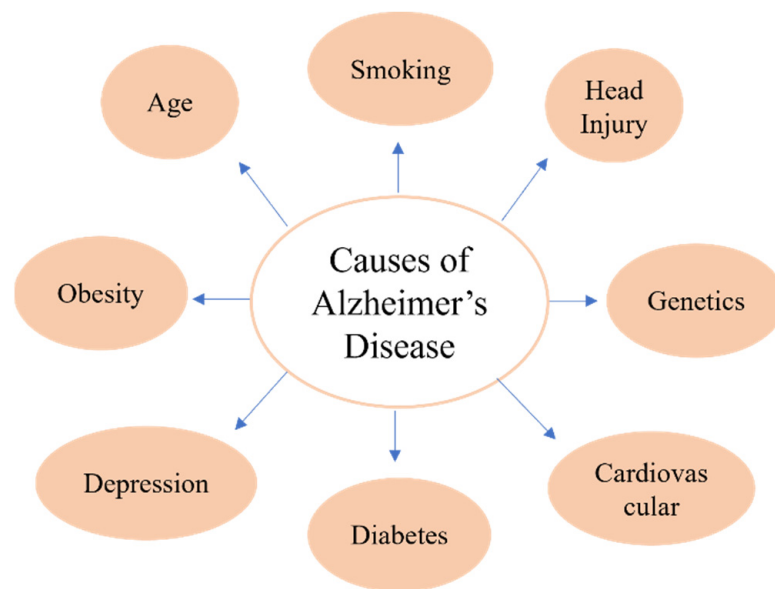


Figure 1. Alzheimer's disease causes.

In recent years, DL techniques have demonstrated remarkable success in medical image analysis [3,4], offering automated and highly accurate solutions for disease detection and classification. Convolutional Neural Networks (CNNs) have been extensively utilized because of their strong capability to extract local spatial features effectively and hierarchical patterns such as edges, textures, and shapes [5]. However, CNN-based models are inherently limited in modeling long-range dependencies and broader contextual relations within images. In other words, the CNNs are limited in capturing global contextual information. To address this limitation, transformer-based architectures, such as the Swin-T, have been presented. The Swin-T utilizes a hierarchical structure with shifted windows to efficiently capture global information [6]. As a result, hybrid approaches have emerged as a promising direction, enabling more comprehensive feature representation and improved performance in complex medical imaging tasks.

Despite these advancements, several challenges continue to hinder the development of reliable and clinically applicable models for AD classification. The limited availability of well-annotated medical datasets can negatively affect model generalization and robustness [7–9]. To address these challenges, this paper proposes a hybrid CNN–Swin-T framework enhanced with an attention mechanism. Additionally, XAI techniques are incorporated to provide visual and interpretable insights into the model's decision-making process. Experiments are conducted on a publicly available Alzheimer's dataset obtained from Kaggle [10].

Research Problem: Accurate classification of AD stages from medical images remains a highly challenging task due to the subtle and complex variations between different disease categories, such as Mild Demented, Very Mild Demented, and Non-Demented [11]. These

classes often exhibit overlapping visual characteristics, making it difficult for automated systems to distinguish between them with high precision. Furthermore, variations in image quality, noise, and limited dataset size introduce additional complexity, leading to inconsistent model performance across classes. In addition to these limitations, the lack of interpretability in many existing DL models presents a significant barrier to their clinical adoption. DL systems often provide predictions without clear explanations, reducing trust among healthcare professionals and restricting the adoption of such models in real-world clinical environments. Therefore, there is a need for a robust and interpretable framework that can effectively combine local and global feature learning while providing meaningful insights into its predictions.

Research Motivation: The motivation behind this paper arises from the need to develop accurate, reliable, and interpretable automated systems for AD classification. While DL approaches have achieved promising results, there is still a need for models that balance high performance with transparency in decision-making. In clinical practice, interpretability is essential, as medical professionals require clear explanations to support diagnostic decisions [12,13]. This paper is motivated by the potential of combining complementary DL paradigms into a unified hybrid framework. By integrating the CNN with Swin-T architectures, the proposed approach aims to leverage their strengths while overcoming their individual limitations. Furthermore, incorporating an attention mechanism improves the model's ability to concentrate on the most relevant regions within input images. Further, XAI methods are employed to provide visual interpretations of the model's predictions [14]. By combining hybrid modeling, attention mechanisms, data augmentation, and explainability, this paper aims to develop a comprehensive and clinically applicable solution for AD diagnosis.

Research Gap: Despite the significant progress in AD classification using DL techniques, several critical gaps still persist in the existing literature. Many studies depend on either CNNs or transformer-based models independently, without fully exploiting their complementary strengths through a unified hybrid architecture. The integration of attention mechanisms within CNN–Transformer hybrid models has been limited, reducing the ability of models to effectively prioritize the most relevant features. Also, most existing approaches don't focus on the model's interpretability, particularly through XAI techniques. There is also a shortage of holistic frameworks that simultaneously combine hybrid modeling, attention mechanisms, and XAI within a single pipeline. Moreover, the combined impact of data augmentation, hybrid architectures, and attention-based learning on classification performance has not been sufficiently explored. These limitations highlight the need for a more integrated and interpretable approach to AD classification.

Research Contribution: To address the aforementioned challenges and bridge the identified research gaps, this paper proposes a comprehensive framework for AD classification. The proposed approach integrates the strengths of convolutional and transformer-based architectures, enhanced with an attention mechanism to improve feature representation and classification performance. Also, it applies three XAI techniques. Also, data augmentation methods are employed to enhance model generalization on limited datasets.

The contributions of this paper are the following:

- Proposing a novel framework that integrates CNN and Swin-T architectures and incorporates an attention mechanism within this hybrid CNN–Swin-T to enhance feature selection and improve classification performance.
- Integrating multiple XAI techniques: Grad-CAM++, Integrated Gradients, and Saliency Maps to provide interpretability of the model's decisions, which is essential for clinical applicability.

- Developing a comprehensive pipeline that jointly combines hybrid modeling, attention mechanisms, and XAI, which has not been sufficiently explored in prior work.
- Applying data augmentation techniques to improve model generalization and robustness on limited medical imaging data. Further, demonstrating that the proposed approach achieves superior performance while maintaining interpretability makes it more suitable for real-world clinical applications.
- Performing a detailed comparative analysis of four models: CNN, Swin-T, CNN + Swin-T, and the proposed CNN + Swin-T + Attention model.

The rest of the paper is organized as follows. Section 2 presents related works in Alzheimer's disease classification using DL techniques. Section 3 describes the proposed methodology, including model architectures, data preprocessing, and training configuration. Section 4 introduces the performance evaluation methods used to assess the models. Section 5 presents the experimental results and performance analysis. Section 6 provides a detailed discussion of the findings, including model performance and interpretability. Finally, Section 7 concludes the paper and highlights potential directions for future research.

2. Related Works

Several recent studies have explored AD classification using DL and medical imaging techniques, highlighting the importance of accurate, reliable, and early-stage diagnosis. These approaches have utilized various models, including CNNs and transformer-based architectures, to improve classification performance. However, challenges such as limited interpretability, accuracy, subtle inter-class variations, and the need for more effective feature representation remain critical issues, motivating the development of more advanced and integrated frameworks.

Velu, K. and Jaisankar, N. [15] proposed a hybrid architecture based on convolution operations and transformers that integrated EfficientNet-B0 with a Swin-T and attention fusion mechanisms for AD classification using the OASIS MRI dataset. A DCGAN was employed to address class imbalance, and Grad-CAM was used to enhance the model's interpretability. However, it achieved a maximum accuracy of 95.3%, demonstrating a robust capability in capturing both local and global features, and the approach relied on 2D MRI slices rather than volumetric data and introduced significant computational complexity, which may limit its clinical applicability.

Mohsen, S. et al. [16] proposed a hybrid DL model combining ResNext101 with a Swin-T and incorporating multiple XAI methods, for instance, Grad-CAM, LIME, and saliency maps. Using a Kaggle MRI dataset of 6400 images, the model reached a testing accuracy of 98.83%. The study demonstrated the effectiveness of this model in improving its classification performance and interpretability. Nevertheless, the model remained computationally expensive, and issues related to class imbalance and limited external validation were identified.

Alhumaid et al. [17] presented a combination of EfficientNet-B0, and Swin-T with an attention fusion mechanism to classify medical images using a clinical CT dataset. This model obtained a total accuracy of 95.83%, showing robust feature fusion capabilities. However, the study was not Alzheimer-specific and relied on CT imaging rather than MRI, which limited its direct relevance to AD diagnosis. The dataset size was relatively small, potentially affecting generalizability.

In [18], a comprehensive survey was conducted covering CNNs, Vision Transformers, hybrid architectures, and XAI techniques for medical image analysis. The study provided a detailed taxonomy of existing methods and highlighted challenges, e.g., model interpretability, data scarcity, and computational complexity. While this survey offered valuable insights, it didn't propose a model or experimental validation.

Erdoğan, P. et al. [19] proposed a hybrid of CNN and Vision Transformer (ViT) framework integrated with a LightGBM classifier and Grad-CAM interpretability using MRI data combined with demographic information. The model reached a testing accuracy of 96.83% and illustrated improved classification efficiency through multimodal data integration. However, the approach introduced additional pipeline complexity and depended on external demographic data, which may limit its scalability and generalization.

Anzum, H. et al. [20] presented a transformer-based model incorporating LIME explainability using multimodal MRI and RNA data. The model achieved a reported accuracy of up to 99%, demonstrating the effectiveness of combining imaging and biological data. Despite its robust performance, the reliance on the RNA data reduced its practicality in real-world clinical settings, where such data may not be readily available.

Hussain et al. [21] applied a hybrid CNN–Swin-T model to MRI data for Parkinson’s disease classification, achieving an accuracy of 96%. The study demonstrated effective fusion of local and global features. However, the model was not specifically designed for AD, limiting its direct applicability.

Alorf [22] implemented a hybrid CNN and transformer-based architecture using ADNI rs-fMRI data combined with clinical features. The model achieved a 96% accuracy and demonstrated robust performance in multiclass classification tasks. However, the requirement for multimodal data increased the complexity and may limit real-world deployment.

Almansour et al. [23] proposed a deformable attention vision transformer (DA-ViT) for AD classification using MRI data. This model reached a total testing accuracy of 80.31% and focused on improving computational efficiency through deformable attention mechanisms. However, the absence of CNN-based feature extraction limited its ability to capture fine-grained local features, resulting in lower performance compared to hybrid models.

Saqib et al. [24] presented a Bidirectional Encoder representation from Image Transformers (BEiT) combined with generative adversarial networks (GANs) and XAI techniques for AD classification utilizing MRI data. The model achieved an accuracy of approximately 96%, with robust F1-scores indicating high classification performance. The integration of GANs helped in mitigating class imbalance, while the transformer backbone improved global feature extraction. However, the inclusion of GAN significantly increased training complexity and computational cost and introduced instability during optimization.

Hussain et al. [25] introduced EffResNet-ViT, a fusion-based model combining EfficientNet-B0 and ViT for explainable medical image classification. The model achieved a high accuracy of 99.31% on MRI datasets and 92.54% on retinal scans. The use of Grad-CAM enhanced the interpretability by highlighting critical regions influencing predictions. Despite its high performance, the model was not Alzheimer-specific and introduced additional computational overhead due to dual-architecture fusion.

In [26], Qezelbasha-Chamak and Hicklin proposed ConvTransFusion, which was a hybrid architecture combining ConvNeXt and Swin-T using a learnable dual-attention mechanism. Although the exact accuracy was not explicitly stated, the model demonstrated superior performance across multiple medical imaging datasets. The approach effectively integrated channel and spatial attention for improved feature representation. However, the increased architectural complexity and computational requirements may limit real-time deployment.

Yao et al. [27] developed an interpretable transformer (IT) model integrating PET and MRI data for AD prediction. While exact accuracy values were not explicitly stated, the model demonstrated improved diagnostic performance across evaluation benchmarks. The use of multimodal data enhanced both functional and structural representation of

brain activity. However, it depended on PET imaging, which increased cost and reduced scalability in clinical settings.

Alsubaie et al. [28] presented a hybrid 3D CNN–ViT with a Bidirectional Feature Pyramid Network (BiFPN), achieving an accuracy of 92% on ADNI MRI datasets. The model effectively captured volumetric features and multiscale representations, improving classification robustness. However, the use of 3D data significantly increased computational requirements and memory consumption.

Almalki et al. [29] conducted a comparative study analyzing CNN, ViT, and Swin-T models under both balanced and imbalanced datasets. The study reported accuracies up to 99% for ViT and 97% for CNN models, highlighting the impact of data imbalance on performance. The findings emphasized the importance of model selection and data preprocessing; however, no hybrid solution was proposed.

Gan et al. [30] introduced a Dense Attention Network (DANet), which was a CNN-based architecture incorporating linear attention mechanisms for effective AD classification using 3D MRI data. The model achieved high AUC performance with above 90% while maintaining a lower parameter count. Despite its efficiency, the absence of transformer-based global modeling limited its ability to extract long-range dependencies.

Li et al. [31] implemented a ViT-based model (Vi-ADiM) with SHAP and Grad-CAM++ explainability, achieving improved performance over baseline ViT models. The model reduced parameters by approximately 48.96% and computational cost (FLOPs) by 49.65%, while maintaining competitive accuracy. However, the lack of CNN integration limited the capability of local feature extraction.

Dessain et al. [32] designed a Swin-T-based model for AD classification using diffusion MRI data. This model obtained a maximum accuracy of 95.2% for dementia classification and approximately 79.6% for amyloid classification. The approach demonstrated robust hierarchical feature extraction and explainability using Grad-CAM. However, it relied on diffusion MRI, which limits accessibility, and the absence of CNN components reduced the sensitivity of local features.

3. The Proposed Methodology

In this Section, we present the overall workflow, including the architecture of the proposed attention-enhanced hybrid CNN–Swin-T model and the other three implemented models: CNN, Swin-T, and hybrid CNN–Swin-T. Also, this section offers the description of the utilized dataset, techniques of data preprocessing and augmentation, and XAI methods: Grad-CAM++, Integrated Gradients, and Saliency Maps, and experimental setup with the capabilities of the used hardware.

3.1. Overview and Workflow Architecture

Figure 2 illustrates the overall workflow of the proposed work, including data preprocessing, models, XAI techniques, and evaluation metrics. The AD classification framework follows a structured pipeline designed to ensure accurate and interpretable predictions. The process begins with data acquisition, where a publicly available Alzheimer’s dataset [10] obtained from the Kaggle repository is utilized. The dataset consists of brain images categorized into three classes: Mild Demented, Very Mild Demented, and Non-Demented. This workflow illustrates the end-to-end process, including data preprocessing and model training for evaluation and explainability analysis. In the initial stage of the workflow, all input images undergo preprocessing to ensure consistency and optimal model efficiency. Specifically, the images are resized to a fixed dimension of 224×224 pixels to fit the input requirements of both CNN and Swin-T architectures. The pixel values are normalized through rescaling to improve training stability and convergence. To enhance model general-

ization and mitigate the limitations of a relatively small dataset, various data augmentation techniques are applied, including transformations such as flipping, rotation, and scaling. These augmentation strategies increase data diversity and help reduce overfitting. Following preprocessing, the dataset is split into training and testing sets to assess the models' efficiency effectively. The proposed work includes four DL architectures: a standalone CNN model, a Swin-T model, a hybrid CNN–Swin-T model, and an attention-enhanced hybrid CNN–Swin-T model. The CNN component focuses on extracting local spatial features, while the Swin-T captures global contextual relationships within the images. In the hybrid CNN–Swin-T, these complementary features are combined to form a more comprehensive representation. To further improve feature learning, an attention mechanism is integrated into the hybrid CNN–Swin-T model. This attention mechanism enables the network to concentrate on the most relevant regions in the MRI images, enhancing discriminative feature extraction and improving classification accuracy. The final classification layer assigns each input image to one of the three AD categories. Model evaluation is performed via standard metrics, including accuracy, recall, precision, and F1-score, to guarantee a thorough evaluation of classification performance. Also, the system incorporates XAI methods: Grad-CAM++, Integrated Gradients, and Saliency maps, to provide visual interpretations of the model's predictions. These techniques highlight important regions in the images that contribute to the classification decision, enhancing model transparency and reliability. Finally, the proposed workflow includes a comparative analysis across all implemented models to demonstrate the effectiveness of the attention-enhanced hybrid CNN–Swin-T approach. The results show that the proposed hybrid CNN–Swin-T integrated with an attention mechanism achieves superior performance while maintaining interpretability, making it suitable for practical medical applications.

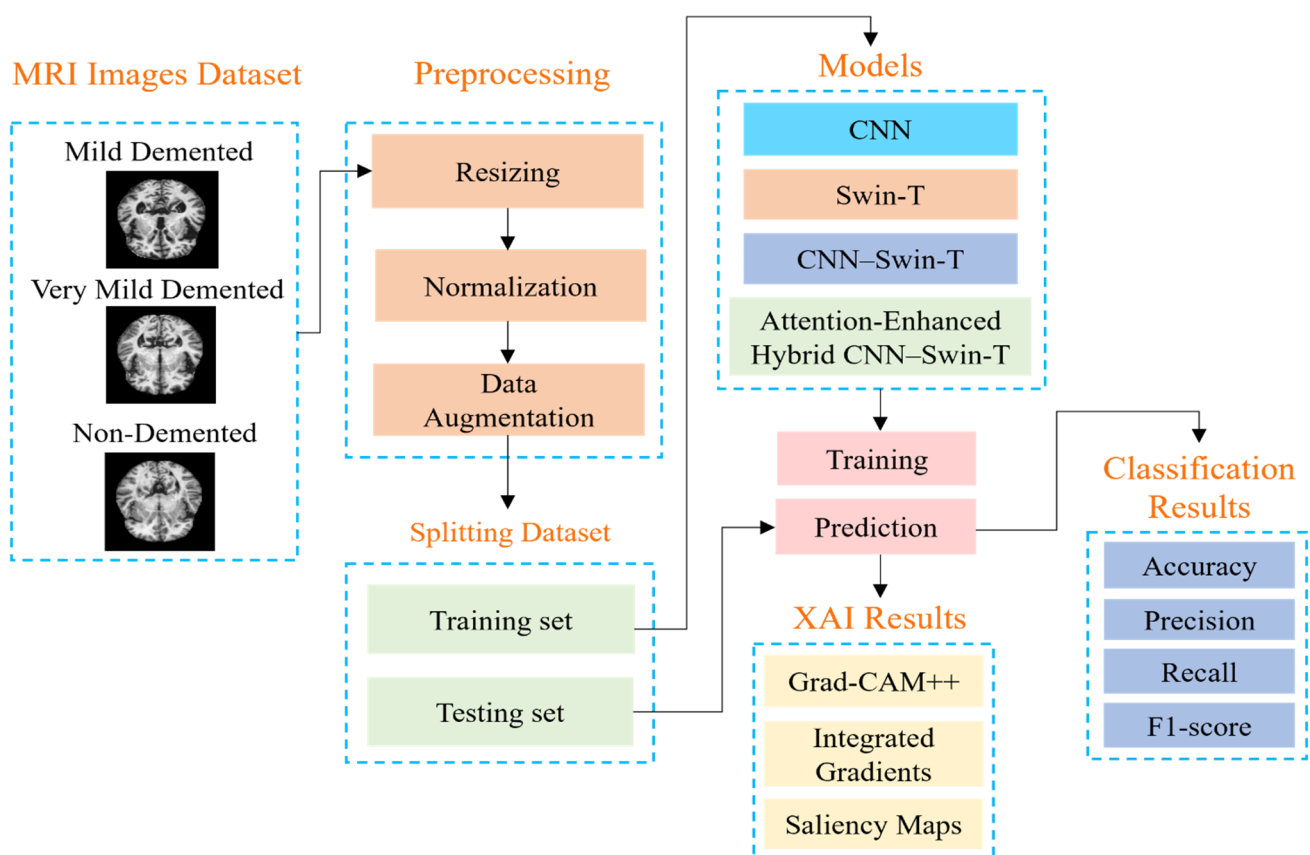


Figure 2. Overall workflow of the proposed work.

3.2. Dataset

The Alzheimer's dataset utilized in this paper is obtained from a publicly available source on Kaggle repository [10] and consists of brain MRI images categorized into three classes: Non-Demented, Mild Demented, and Very Mild Demented. The dataset includes a total of 6336 images, with 3200 images belonging to the Non-Demented class, 896 images to the Mild Demented class, and 2240 images to the Very Mild Demented class.

The dataset exhibits a moderate class imbalance, particularly in the Mild Demented category, which contains significantly fewer samples compared to the other classes. This imbalance introduces additional challenges for training DL models, as it may bias the learning process toward majority classes if not properly handled.

The images in the dataset represent brain structures associated with different stages of AD and include variations in anatomical patterns across classes. These variations are often subtle, especially between Mild Demented and Very Mild Demented cases, making the classification task more complex. Such characteristics make the dataset suitable for evaluating the performance of advanced DL models in multi-class medical image classification.

To ensure consistency and compatibility with the employed DL architectures, all images are resized to 224×224 pixels during preprocessing. The dataset is split into training and testing sets, where 80% of the samples are allocated to train models, and the remaining 20% are set aside for assessment. This split ensures that the models are trained on a representative portion of the dataset while being evaluated on unseen data to assess their generalization capability. Figure 3 illustrates the distribution of the dataset across the three classes, which shows class imbalance between Non-Demented, Mild Demented, and Very Mild Demented categories, while Figure 4 presents representative sample images from each category.

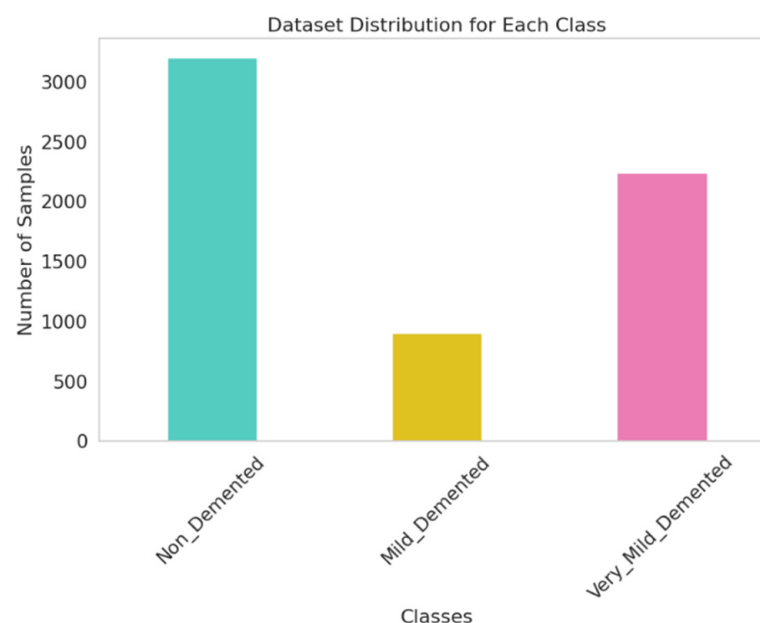


Figure 3. Distribution of samples across the three classes in the Alzheimer's dataset.

3.3. Data Processing and Augmentation

To enhance the performance and generalization capability of the proposed models, a comprehensive data preprocessing and augmentation strategy is employed. Initially, all input images are resized to a fixed resolution of 224×224 pixels to guarantee compatibility with both CNN and Swin-T architectures. The pixel values are normalized to improve training stability, accelerate convergence, and prevent issues related to varying intensity scales across images.

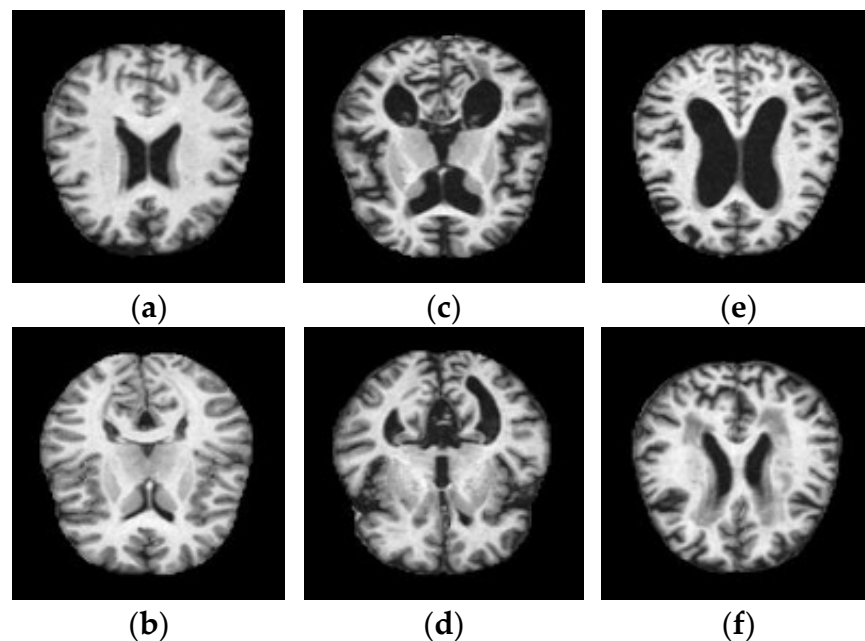


Figure 4. Representative samples of MRI images from the Alzheimer's dataset: (a,b) Non-Demented; (c,d) Mild Demented; and (e,f) Very Mild Demented.

Given the limited size and inherent class imbalance of the original dataset, data augmentation plays a critical role in improving model robustness. The original dataset exhibits class imbalance, with the Mild Demented category containing significantly fewer samples compared to the Non-Demented and Very Mild Demented categories. Such an imbalance can negatively impact model performance by biasing predictions toward majority classes. To address this issue, data augmentation methods are applied to the training set to increase the number of samples in each class and improve dataset balance.

So, these data augmentation techniques are utilized to artificially expand the dataset and introduce variability in the training samples. Specifically, for each original image, five additional augmented images are generated using a combination of geometric and spatial transformations. These transformations include controlled rotation, width and height shifting, shear transformation, zooming, and horizontal flipping. Such operations simulate real-world variations and enable the model to learn invariant and discriminative features, thereby reducing overfitting and enhancing generalization. This approach helps in mitigating bias toward majority classes and enhances the model's ability to correctly classify underrepresented categories, particularly the Mild Demented class.

The impact of the augmentation process on dataset size and class distribution is summarized in Table 1, where a significant increase in the number of training samples can be observed across all classes. This expansion not only improves data diversity but also helps in mitigating class imbalance, particularly for the Mild Demented category, which originally contained fewer samples. As a result, the augmented dataset provides a more balanced and representative training set for the DL models.

Table 1. Dataset distribution before and after augmentation.

Class	Number of Original Images	Number of Images After Data Augmentation
Non-Demented	3200	19,200
Mild Demented	896	5376
Very Mild Demented	2240	13,440
Total Samples of the Dataset	6336	38,016

The specific augmentation techniques and their corresponding parameter values are detailed in Table 2. These parameters are carefully selected to introduce meaningful variations without distorting the underlying medical features of the images. By maintaining a balance between augmentation intensity and data integrity, the proposed preprocessing pipeline ensures that the generated samples remain realistic and clinically relevant.

Table 2. Data augmentation parameters.

Data Augmentation Technique	Value
Rotation Range	20°
Width Shift Range	0.08
Height Shift Range	0.08
Shear Range	0.08
Zoom Range	0.08
Horizontal Flip	Enabled
Fill Mode	Nearest
Augmented Images Per Input	5

Overall, the combination of image resizing, normalization, and targeted data augmentation contributes significantly to improving the model's performance, stability, and generalization capability, making the proposed framework more effective for AD classification. The dataset is divided into training and test sets at the patient level, meaning that all MRI scans from a given patient are included entirely in either the training set or the test set, without any overlap between the two. This strategy ensures that the models are assessed on previously unseen patients.

3.4. The Architecture of Proposed Models

This paper presents four different DL architectures for AD classification: CNN, Swin-T, a hybrid CNN–Swin-T model, and the proposed attention-enhanced hybrid model. Each model is designed to capture different aspects of feature representation, including local spatial features, global contextual information, and attention-based feature prioritization.

3.4.1. Convolutional Neural Network (CNN) Model

The CNN is employed as a baseline model due to its effectiveness in extracting local spatial features from images [33]. CNNs utilize convolutional layers to learn hierarchical feature representations, where lower layers capture basic patterns such as edges and textures, and deeper layers extract more complex structures [34]. The convolution operation can be mathematically expressed as:

$$Y(i, j) = \sum_m \sum_n X(i + m, j + n) \cdot K(m, n) \quad (1)$$

where X represents the input image, K is the convolution kernel, and Y is the resulting feature map. After convolution, activation functions such as ReLU are applied:

$$f(x) = \max(0, x) \quad (2)$$

The pooling layers are then used to reduce spatial dimensions and retain important features:

$$Y = \max(X) \quad (3)$$

Finally, fully connected layers are used for classification, followed by a softmax function:

$$P(z_i) = \frac{e^{z_i}}{\sum_{j=1} e^{z_j}} \quad (4)$$

As illustrated in Figure 5, the implemented CNN architecture consists of three convolutional “Conv” layers, three batch normalization layers, three ReLU activation functions, two max-pooling layers, one adaptive average pooling layer, three dropout layers, and one fully connected layer “linear layer” for final classification. The kernel size of all three Conv layers is 3×3 , while the number of kernels is 32, 64, and 128 for the first, second, and third Conv layers, respectively. Each Conv-layer includes one zero-padding layer. The pool size of the two max-pooling layers is 2×2 , while the pool size of the adaptive average pooling layer is 1×1 . The adaptive average pooling layer automatically adjusts the pooling kernel size and stride according to the input feature-map dimensions and the desired output size. Unlike conventional average pooling, where the kernel and stride are fixed manually. The three dropout layers have different dropout rates of 0.25, 0.3, and 0.4 for the first, second, and third layers, respectively. The linear layer as output has three neurons, which are activated by the softmax activation function.

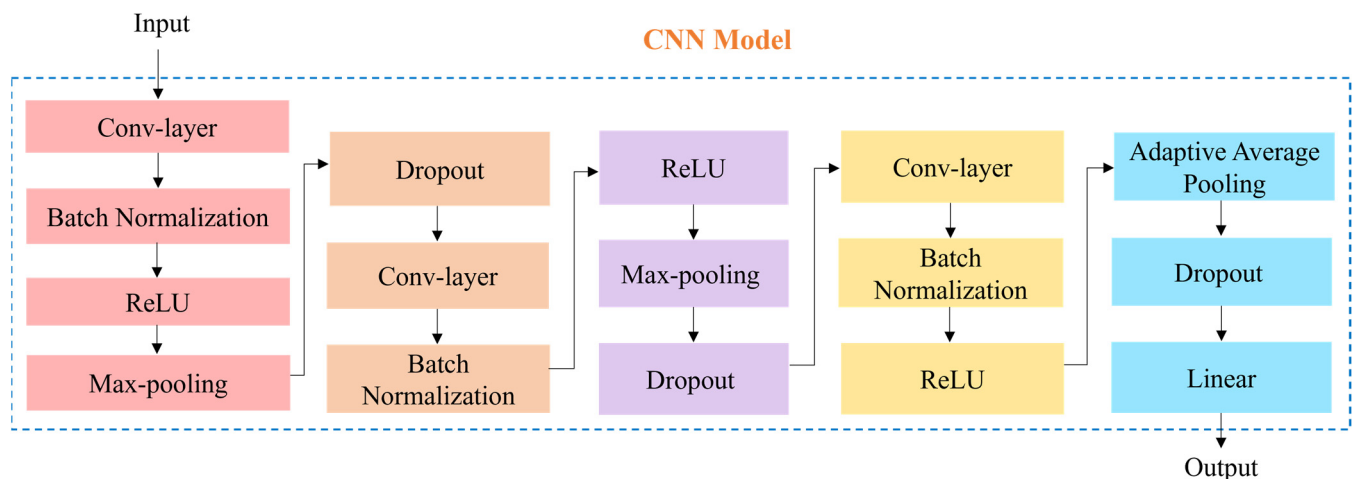


Figure 5. The implemented CNN architecture.

3.4.2. Swin Transformer (Swin-T) Model

The Swin-T is a hierarchical vision transformer that captures both global and local dependencies using a shifted window mechanism [35]. Unlike standard transformers, it computes self-attention within local windows and shifts these windows across layers to enable cross-window interactions. The self-attention mechanism is represented as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where K , Q , and V represent key, query, and value matrices, respectively, while d_k is the dimension of key vectors. Dividing by a square root of the d_k to normalize attention scores, which makes more stable gradients during training, therefore improving the model’s performance. The Swin-T introduces window-based multi-head self-attention (W-MSA) and shifted window attention (SW-MSA), enabling efficient computation while preserving global context [36]. As shown in Figure 6, the Swin-T architecture processes input images through patch partitioning, linear embedding, and hierarchical transformer blocks to generate feature representations.

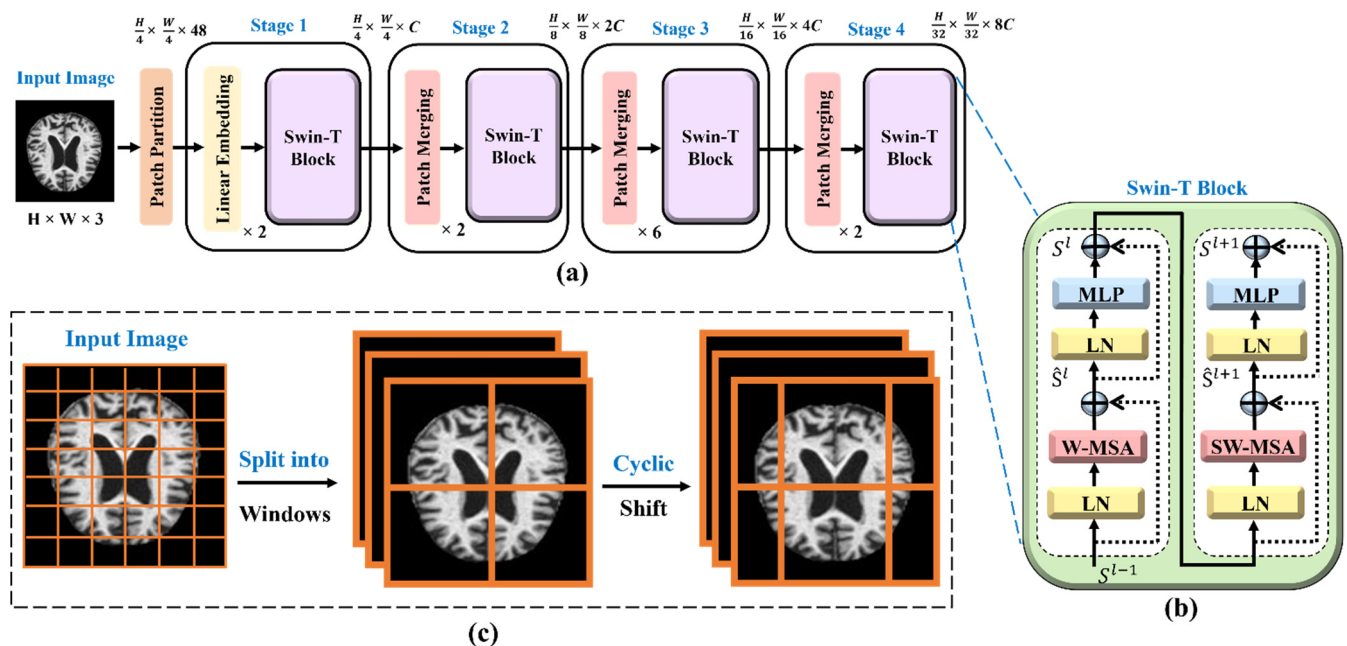


Figure 6. The implemented Swin-T architecture: (a) Swin-T architecture; (b) internal structure of a Swin-T block; and (c) mechanisms of windows partitioning and shifting.

The Swin-T architecture processes the medical input images by dividing them into non-overlapping small patches, which are flattened and embedded into feature vectors for transformer input. As shown in Figure 6a, it consists of four hierarchical stages, each including a Swin-T block followed by patch merging. The model uses self-attention inside local windows, improving efficiency for high-resolution images, while the patch merging reduces spatial size and increases feature depth. As the network progresses, it learns increasingly abstract representations, with the final stage capturing high-level semantic features for classification.

Figure 6b presents the internal structure of the Swin-T Block. Each block contains two self-attention mechanisms: W-MSA and SW-MSA. The W-MSA calculates attention within fixed local windows, reducing computational cost, while the SW-MSA shifts window positions to enable cross-window interactions. Both are combined with Layer Normalization, MLP layers with GELU activation, and residual connections to enhance learning stability and non-linearity. This dual-attention design allows the Swin-T model to extract both global and local dependencies efficiently. The feature extraction process across successive Swin-T blocks is mathematically formulated in Equations from (6) to (9) [16], where S_{l+1} and S_l represent the outputs of the W-MSA and SW-MSA modules with respective MLP layers. By integrating both SW-MSA and W-MSA, each Swin-T block efficiently extracts fine-grained local features as well as long-range dependencies, while preserving computational efficiency.

As shown in Figure 6c, the Swin-T employs window partitioning and cyclic shifting, where windows are shifted between layers to connect neighboring regions. Cyclic padding ensures proper alignment after shifting. This mechanism enables effective extraction of multi-scale features along with modeling of long-range dependencies, making the Swin-T well-suited for high-resolution medical imaging tasks such as AD analysis.

$$S^l = \text{MLP}\left(\text{LN}\left(\hat{S}^l\right)\right) + \hat{S}^l \quad (6)$$

$$S^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{S}^{l+1}\right)\right) + \hat{S}^{l+1} \quad (7)$$

$$\hat{S}^l = W-MSA\left(LN\left(S^{l-1}\right)\right) + S^{l-1} \quad (8)$$

$$\hat{S}^{l+1} = SW-MSA\left(LN\left(S^{l-1}\right) + S^{l-1}\right) \quad (9)$$

3.4.3. Hybrid CNN–Swin-T Model

To leverage the strengths of both CNN and transformer architectures, a hybrid CNN–Swin-T model is implemented. In this architecture, the CNN component extracts local spatial features, while the Swin-T captures global contextual information. Let F_{CNN} and F_{Swin-T} denote the feature representations from CNN and Swin-T, respectively. The combined feature representation F_{hybrid} is defined as follows:

$$F_{hybrid} = F_{CNN} \oplus F_{Swin-T} \quad (10)$$

where $\oplus$ represents feature concatenation. The fused features are then passed through a fully connected output layer for classification, activated by Softmax to generate the output y :

$$y = \text{Softmax}\left(W \cdot F_{hybrid} + b\right) \quad (11)$$

This hybrid approach enables the model to extract both fine-grained local details and long-range dependencies [37]. As illustrated in Figure 7a, the hybrid architecture integrates parallel CNN and Swin-T branches, followed by feature fusion and classification layers. This architecture is designed in parallel to extract complementary feature representations from the same input image. The feature fusion process is performed after the independent feature extraction stages of both branches. The output feature maps from the CNN encoder and the Swin-T encoder are transformed into compatible dimensions. The resulting fused feature vector combines both local discriminative information and global contextual representations.

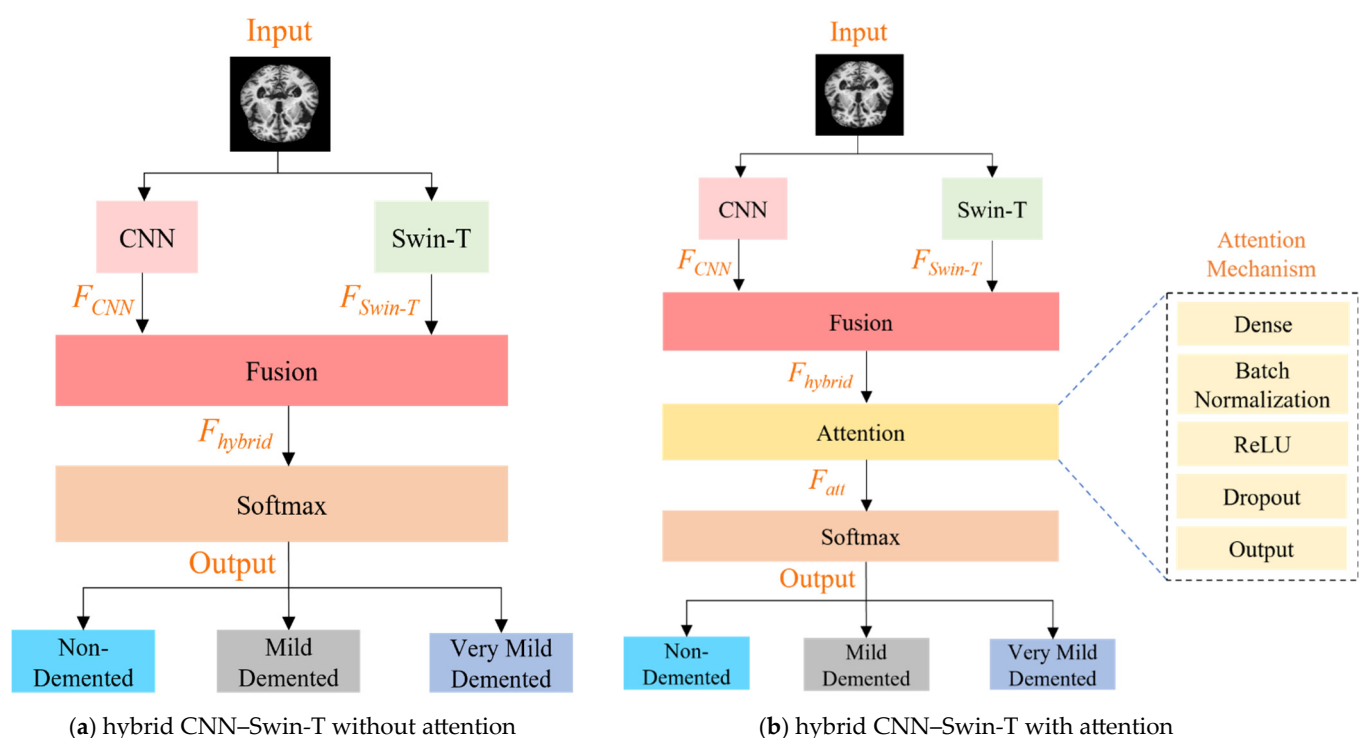


Figure 7. (a) The implemented hybrid CNN–Swin-T architecture; and (b) the proposed attention-enhanced hybrid CNN–Swin-T architecture.

3.4.4. Attention-Enhanced Hybrid CNN–Swin-T Model

The proposed model extends the hybrid CNN–Swin-T architecture by incorporating an attention mechanism to enhance feature selection, as shown in Figure 7b. The attention module assigns weights to features, enabling the model to concentrate on the most relevant regions of the input image. The attention mechanism can be formulated as follows:

$$F_{att} = \alpha \cdot F_{hybrid} \quad (12)$$

F_{att} is the output of the attention mechanism, while α represents the attention weights learned during training. These weights are computed using:

$$\alpha = \text{Softmax}(W_a \cdot F_{hybrid} + b_a) \quad (13)$$

The attention-enhanced features are then used for final classification:

$$y = \text{Softmax}(W \cdot F_{att} + b) \quad (14)$$

This attention mechanism enhances the model's ability to prioritize discriminative features, particularly in challenging cases with subtle differences between classes. As shown in Figure 7b, the proposed architecture integrates CNN and Swin-T components with an attention module, followed by classification layers. The attention mechanism consists of a dense layer, followed by a batch normalization with a ReLU activation, then a dropout layer, and finally the output. The feature extraction process in the proposed architecture culminates in a classification head designed to map the learned representations into multi-class predictions corresponding to AD stages. The extracted feature maps from the hybrid CNN–Swin-T architecture are first transformed into a compact feature vector through global pooling operations, which aggregate spatial information while reducing dimensionality.

The resulting feature vector is then passed through fully connected layers to perform high-level feature integration. Specifically, a dense layer with 512 neurons is employed to combine local and global features extracted from the previous stages. To enhance the training stability and prevent overfitting, batch normalization is applied, followed by L2 regularization to constrain the magnitude of learned weights.

To enhance generalization, dropout regularization with a probability of $p = 0.5$ is introduced, randomly skipping neurons during learning. This approach decreases the model's reliance on specific features and helps prevent overfitting, especially given the limited size of medical imaging datasets. For classification, the final output layer is represented by a fully connected layer with three output neurons corresponding to the three classes: Non-Demented, Mild Demented, and Very Mild Demented. A softmax activation function is added to transfer the output logits into class probabilities, as defined in Equation (4), where $C = 3$ represents the number of classes, and z_i denotes the output of the final layer for the class i .

The predicted class is determined by selecting the class with the highest probability:

$$\hat{y} = \underset{i}{\operatorname{argmax}} P(z_i) \quad (15)$$

This classification framework ensures that the model effectively integrates the extracted features and produces reliable multi-class predictions for AD classification.

3.4.5. Model Comparison Strategy

To assess the effectiveness of the proposed work, a comparative analysis is conducted across the four models: CNN, Swin-T, hybrid CNN–Swin-T, and the proposed attention-enhanced hybrid CNN–Swin-T model. Each model is trained and evaluated under the same conditions to ensure a fair comparison. The comparison focuses on: (1) Classification performance (accuracy, recall, precision, and F1-score); (2) Model robustness; and (3) the Interpretability through XAI techniques.

3.5. Optimization Strategy and Training Configuration

The proposed four models are optimized via the Adaptive Moment Estimation (Adam) optimizer, which is widely adopted in DL due to its ability to combine the advantages of both momentum and adaptive learning rate methods. A learning rate of $\alpha = 1 \times 10^{-5}$ is selected to ensure stable convergence while maintaining sufficient learning capacity. This relatively low learning rate helps prevent oscillations during training and allows the model to gradually refine its parameters, which is particularly important for complex hybrid architectures.

Since the task involves multi-class classification of AD stages, the categorical cross-entropy loss function L is employed as the optimization objective. This loss function is convenient for multi-class issues and provides smooth gradients for effective optimization. It is defined as follows:

$$L = - \sum_{i=1}^C y_i \log_2 \hat{y}_i + (1 - y_i) \log_2 (1 - \hat{y}_i) \quad (16)$$

where C represents the number of classes, $\hat{y}_i$ is the ground truth label, and y_i is the predicted probability for the class i . The training process is conducted over 30 epochs, with continuous monitoring of validation performance to prevent overfitting and ensure optimal model generalization. The dataset is divided into training and testing subsets, enabling the model to be evaluated on unseen data at each epoch.

To prevent overfitting and improve generalization, an early stopping mechanism is employed based on validation performance, with a patience of 5 epochs. Training is terminated when no improvement in validation loss is observed over consecutive epochs. The batch size of images is 32 during the training phase. The hyperparameter settings used for training all models are summarized in Table 3, ensuring reproducibility and transparency of the experimental setup. The computational complexity analysis revealed variations in both training time and model size across the evaluated architectures. The standalone CNN model required approximately 20 min for training with data augmentation and contained 94,851 trainable parameters. In comparison, the Swin-T model exhibited a higher computational cost, requiring approximately 180 min of training time with 27,522,090 parameters. The hybrid CNN–Swin-T architecture slightly increased the computational complexity, resulting in a training time of approximately 190 min and 27,619,757 trainable parameters. The proposed attention-enhanced hybrid CNN–Swin-T model achieved the highest computational efficiency, requiring approximately 200 min of training time while increasing the number of trainable parameters to 27,754,182.

To comprehensively evaluate the models' performance, multiple assessment metrics are applied, including accuracy, recall, precision, and F1-score. Accuracy measures the total correctness of predictions, while recall and precision provide insights into class-wise performance. The F1-score balances precision and recall, offering a more robust evaluation in the presence of class imbalance.

Table 3. Hyperparameter settings used in the training phase.

Parameter	Value
Batch Size	32
Number of Epochs	30
Learning Rate	1×10^{-5}
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Input Image Size	224×224
Dropout Rate	0.5
Weight Regularization	L2 Regularization
Early Stopping	Patience = 5
Data Augmentation	Enabled
Augmented Images for Each Original Image	5

This optimization strategy ensures stable training, efficient convergence, and reliable performance across all evaluated models, supporting the development of an accurate and generalizable AD classification framework.

3.6. Explainable Artificial Intelligence (XAI)

XAI techniques play a critical role in enhancing the transparency and reliability of DL models, particularly in medical imaging applications [16]. Although DL models have demonstrated high performance in classification tasks, they are often considered black-box systems, making it difficult to interpret how predictions are generated. In clinical environments, interpretability is essential to ensure trust and support decision-making by healthcare professionals. Therefore, this paper incorporates multiple XAI methods to gain insights into the internal decision-making process of the proposed models.

To achieve this, three widely used XAI methods are employed: Grad-CAM++, Integrated Gradients, and Saliency maps. These techniques are applied to all three implemented models, including CNN, Swin-T, and hybrid CNN-Swin-T, and are also applied to the proposed attention-enhanced hybrid CNN-Swin-T model, enabling a comprehensive interpretability analysis. Each method provides a different perspective on feature importance, allowing for a more thorough understanding of how the models process and classify input images.

3.6.1. Gradient-Weighted Class Activation Mapping++ (GRAD-CAM++)

Grad-CAM++ is a visualization technique designed to highlight the regions of input images that are most influential in determining the model's prediction [38]. It extends the original Grad-CAM approach by incorporating higher-order gradient information, resulting in more precise localization of important regions.

The Grad-CAM++ computes the importance weights of feature maps by analysing the gradients of the target class score with respect to the feature maps extracted from the final convolutional layer. The class activation map is defined as follows:

$$L_{\text{Grad-CAM++}}^c = \sum_{k=0}^l \alpha_k^c A_{ij}^k \quad (17)$$

$$\alpha_k^c = \frac{\partial y^c}{\partial A_{ij}^k} \quad (18)$$

where A_{ij}^k represents the k -th feature map and α_k^c denotes the corresponding importance weight for the class c . These weights are derived using partial derivatives of the class score, enabling improved localization compared to traditional gradient-based methods.

This technique provides a coarse localization map that identifies the most relevant regions contributing to the classification decision, making it particularly useful for understanding model focus in medical images.

3.6.2. Integrated Gradients

Integrated Gradients (IG) is a feature attribution method that assigns importance scores to individual input features by accumulating gradients along a trajectory from a reference (baseline) input to the actual input [39]. This approach addresses limitations of standard gradient methods by providing more stable and theoretically grounded attributions. The attribution for each input feature is computed as follows:

$$IG_i(x) = (x_i - x'_i) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (19)$$

where x is the input image, x' is a baseline reference (such as a zero image), and $F(x)$ represents the model's prediction function. The parameter α defines the interpolation between the baseline and the input. The IG provides fine-grained pixel-level importance, allowing for a detailed understanding of how each part of the input contributes to the model's prediction.

3.6.3. Saliency Maps

Saliency maps are one of the most fundamental gradient-based XAI techniques used to identify the importance of individual input pixels with respect to the model's prediction [16]. This method provides a straightforward way to visualize which regions of images have the greatest influence on the output of a DL model. The core idea behind saliency maps is to compute a gradient of the model's output relative to input images. These gradients indicate how sensitive the prediction is to small changes in each pixel. Formally, the saliency map $S(x)$ is represented as follows:

$$S(x) = \left| \frac{\partial F_c(x)}{\partial x} \right| \quad (20)$$

where $F(x)$ represents the model's output score for the target class, while x denotes the input image, and $\frac{\partial F(x)}{\partial x}$ is the gradient of the output with respect to the input. The absolute value of the gradient is taken to capture the magnitude of influence regardless of direction. Pixels with higher gradient values are considered more important, as small changes in these regions can significantly affect the model's prediction.

In practice, the saliency map is computed by performing a backward pass through the network, propagating gradients from the output layer to the input layer. The resulting gradient map is then visualized as a heatmap, highlighting the most influential regions in the image. This process does not require modification of the model's architecture and can be applied to any differentiable model. The saliency maps are particularly useful for providing fine-grained, pixel-level explanations, enabling a detailed understanding of how the model responds to local variations in the input. Unlike region-based methods, saliency maps emphasize edges, textures, and subtle structural patterns that contribute to classification decisions. However, it is important to note that saliency maps can be sensitive to noise and may sometimes produce less smooth or less interpretable visualizations compared to more advanced methods such as Grad-CAM++. Despite this limitation, they remain valuable tools for understanding models' behavior and complement other explainability techniques within the proposed framework.

3.6.4. The Proposed Explainability Framework

To ensure a comprehensive and systematic interpretation of model predictions, an explainability framework is developed by integrating multiple XAI techniques within the proposed pipeline. This framework is designed to provide consistent, reliable, and comparative insights into how different models make classification decisions for AD.

The explainability process is applied after the training phase of each model, including the CNN, Swin-T, hybrid CNN–Swin-T model, and the proposed attention-enhanced hybrid CNN–Swin-T model. For consistency and fair comparison, the same set of input images is used across all models when generating explainability visualizations. This ensures that differences in interpretation are attributed to the models' behavior rather than variations in input data.

Within this framework, three complementary XAI techniques, Grad-CAM++, Integrated Gradients, and Saliency maps, are employed to capture different aspects of models' interpretability. The Grad-CAM++ provides coarse localization of important regions by highlighting class-discriminative areas in the feature maps. The Integrated Gradients offer fine-grained, pixel-level attribution by measuring the contribution of each input feature relative to a baseline. The Saliency maps further complement this analysis by identifying regions where small input variations significantly influence the model's output.

By combining these three methods, the framework enables both global and local interpretability. The global interpretability is achieved through region-based visualization methods such as Grad-CAM++, while the local interpretability is supported by pixel-level techniques such as Integrated Gradients and Saliency maps. This multi-perspective analysis provides a more complete understanding of the models' decision-making process.

The proposed framework is designed to support both model-wise and class-wise analysis. In model-wise analysis, explainability outputs are compared across different architectures to evaluate how each model focuses on relevant regions. In class-wise analysis, representative samples from each class are analyzed to understand how the models differentiate between disease stages. This dual-level analysis enhances the reliability of the interpretability assessment. Overall, the proposed explainability framework ensures that the models' predictions are not only accurate but also transparent and interpretable. By systematically integrating multiple XAI techniques and enabling consistent comparison across the models, the framework contributes to building trust in DL-based systems for medical image analysis and supports their potential adoption in clinical practice.

3.7. Experimental Setup

The development and evaluation of the proposed hybrid CNN–Swin-T workflow with attention mechanisms are conducted within the Kaggle computational environment using the Python programming language. The experimental setup leverages modern DL frameworks and data processing libraries to ensure effective model training, reliable performance evaluation, and reproducibility of results. The implementation requires integrating multiple tools for data handling, visualization, and model development, enabling a comprehensive and structured experimental workflow. The detailed workflow configurations used to conduct the experiments are summarized in Table 4.

Table 4. The configurations of workflow for the proposed work.

Component	Specifications
IDE/Platform	Kaggle Notebook Environment
Programming Language	Python (version 3.10.0)
DL Framework	PyTorch (version 2.10.0)

Table 4. Cont.

Component	Specifications
Libraries	NumPy, OpenCV (cv2), PyTorch, TensorFlow, Keras, Matplotlib, scikit-learn, and tqdm
Data Augmentation	Keras ImageDataGenerator
GPU	NVIDIA Tesla P100 (16 GB VRAM)
CPU	Intel Xeon CPU (2.3 GHz, 46 MB cache)
RAM	16 GB System Memory
Input Image Size	224 × 224 pixels
Dataset Source	Kaggle Alzheimer MRI Dataset

4. Performance Evaluation Methods

The performance of the proposed models is assessed utilizing a set of standard machine learning metrics to ensure a thorough and reliable assessment. These metrics are derived from confusion matrix analysis and provide quantitative measures of classification effectiveness across multiple dimensions. Specifically, accuracy, precision, F1-score, and recall are employed to evaluate the models' ability to correctly classify the AD stages. This multi-metric evaluation framework enables detailed analysis of models' behavior, facilitates a fair comparison among different architectures, and provides insights into class-wise performance, particularly in the presence of class imbalance.

4.1. Confusion Matrix

The confusion matrix provides a comprehensive visual representation of classification performance, serving as the foundation for computing evaluation metrics such as accuracy, precision, recall, and F1-score [3]. It illustrates the relationship between actual class labels and predicted class labels, enabling detailed analysis of correct and wrong predictions across all classes.

In the context of multi-class classification, the confusion matrix is extended to a $C \times C$ matrix, where C denotes the total number of classes. Each row corresponds to the true class, while each column represents the predicted class. Correct predictions are located along the diagonal, whereas the off-diagonal entries represent classification errors. For each class, the following components are defined:

- True Positives (TPs) are the instances correctly predicted as belonging to the target class.
- False Positives (FPs) represent the instances incorrectly predicted as belonging to the target class.
- False Negatives (FNs) are the instances of the target class incorrectly predicted as another class.
- True Negatives (TNs) represent the instances correctly predicted as not belonging to the target class.

These components are computed for each class in a one-vs-all manner to evaluate class-wise performance.

4.2. Accuracy

Accuracy evaluates the overall performance of the model by determining the ratio of correctly predicted instances to the total number of samples [4]. It is defined as follows:

$$Accuracy = \frac{TP_i + TN_i}{Total\ Testing\ data\ Samples} \quad (21)$$

where TP_i represents the correctly classified instances for the class i and TN_i is the correctly classified instances not belonging to the class i . The Accuracy provides a general

measure of performance; however, it may not fully reflect the model's effectiveness in imbalanced datasets.

4.3. Precision

It evaluates the model's ability to correctly identify instances of a particular class among all instances predicted as that class [4]. It is defined as follows:

$$Precision = \frac{TP_i}{TP_i + FP_i} \quad (22)$$

High precision means a low false positive rate, meaning the model makes fewer incorrect positive predictions.

4.4. Recall

Recall, also referred to as sensitivity, assesses the model's ability to correctly detect all true samples of a given class [4]. It is expressed as follows:

$$Recall = \frac{TP_i}{TP_i + FN_i} \quad (23)$$

A high recall value indicates that the model successfully captures most of the true instances, which is particularly important in medical applications.

4.5. F1-Score

It is a harmonic mean of the recall and precision, giving a balanced assessment metric that considers both false negatives and false positives [4]. It is defined as follows:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (24)$$

This metric is especially useful in scenarios involving class imbalance, as it ensures that both precision and recall are taken into account when assessing model performance.

5. Experimental Results

This section presents the experimental evaluation of the four models for AD classification. The primary objective is to evaluate the efficiency of the proposed attention-enhanced hybrid CNN-Swin-T model in comparison with baseline architectures, including CNN, Swin-T, and the hybrid CNN-Swin-T model. The evaluation is conducted using the augmented dataset described in Section 3, ensuring a comprehensive analysis of model performance under consistent conditions.

All models are trained and tested using the same data split, preprocessing steps, and optimization settings to ensure a fair and unbiased comparison. The experiments are performed using the Kaggle computational environment with GPU acceleration, enabling effective training and evaluation. The performance is evaluated via some evaluation metrics, for instance, precision, accuracy, recall, and F1-score, as described in Section 4.

A qualitative analysis is conducted using XAI techniques, including Grad-CAM++, Integrated Gradients, and Saliency maps. These methods provide insights into the decision-making behavior of each model by highlighting the most relevant regions in the input images. This combined evaluation strategy ensures that both performance and interpretability are considered.

The following subsections present a detailed analysis of the experimental results, including performance comparison across the models, the impact of data augmentation, and the interpretability analysis using XAI techniques.

5.1. Model Performance Comparison

To assess the effectiveness of the proposed framework, a comprehensive comparison is conducted among the four DL models: CNN, Swin-T, a hybrid CNN–Swin-T model, and the proposed attention-enhanced hybrid CNN–Swin-T model. All models are evaluated under identical conditions to ensure a fair and unbiased comparison.

The performance of each model is assessed via the previously mentioned four metrics. The results are presented for both scenarios: without data augmentation and with data augmentation, as shown in Tables 5 and 6, respectively. The results presented in Table 5 demonstrate that, without data augmentation, the CNN model achieves significantly lower performance compared to the other architectures, with an accuracy of 97.58%. This indicates that CNN alone struggles to capture the complex patterns required for accurate AD classification. In contrast, the Swin-T and hybrid models achieve substantially higher performance, highlighting the importance of incorporating global contextual information.

Table 5. Performance comparison of the models without data augmentation.

Model	Accuracy %	Precision %	Recall %	F1-Score %
CNN	97.58	97.52	97.46	97.48
Swin-T	98.42	98.36	98.00	98.20
CNN + Swin-T	98.34	98.12	98.02	98.06
CNN + Swin-T + Attention	98.50	98.33	98.24	98.28

Table 6. Performance comparison of the models with data augmentation.

Model	Accuracy %	Precision %	Recall %	F1-Score %
CNN	98.32	98.24	98.06	98.14
Swin-T	99.67	99.61	98.03	98.81
CNN + Swin-T	99.88	99.62	99.53	99.57
CNN + Swin-T + Attention	99.92	99.71	99.73	99.72

Among the models evaluated without augmentation, the proposed attention-enhanced hybrid CNN–Swin-T model achieves the highest testing accuracy of 98.50%, demonstrating the effectiveness of integrating attention mechanisms to refine feature selection and improve classification performance.

As shown in Table 6, the application of data augmentation leads to a significant improvement in performance across all four models. The CNN model shows a modest improvement, increasing its accuracy to 98.32%, while the Swin-T and hybrid models achieve high performance. Notably, the proposed hybrid model integrated with an attention mechanism reaches the highest accuracy of 99.92%, along with the highest precision, recall, and F1-score, 99.71%, 99.73%, and 99.72%, respectively. It is clear that the implemented data augmentation techniques significantly improved the model's accuracy.

The results clearly indicate that data augmentation plays an important role in improving the generalization of models and reducing overfitting. The superior performance of the proposed attention-enhanced hybrid CNN–Swin-T model confirms that combining CNN, Swin-T, and attention mechanisms provides a powerful and effective solution for AD classification. Overall, the comparative analysis demonstrates that the proposed model consistently outperforms baseline and hybrid architectures under both experimental settings, validating its robustness and effectiveness.

5.2. Training Behavior Analysis

To analyse the learning dynamics and stability of the four models, the testing and training accuracy and loss curves are examined. These curves provide valuable insights into model convergence, generalization capability, and potential overfitting during the training process. The training behaviour of all models is illustrated in Figures 8 and 9, where Figure 8 presents the training and testing accuracy curves, and Figure 9 shows the corresponding loss curves over the training epochs. The accuracy curves reflect how effectively each model learns to correctly classify the input data, while the loss curves indicate how efficiently the optimization process minimizes prediction errors. As observed in Figures 8a and 9a, the CNN model exhibits relatively unstable learning behaviour, with slower convergence and noticeable fluctuations in both training and testing accuracy. Similarly, the loss curves show irregular patterns, indicating less efficient optimization. This behaviour reflects the CNN model's limited ability to capture complex features in the dataset, which is consistent with its lower performance reported in Section 5.1.

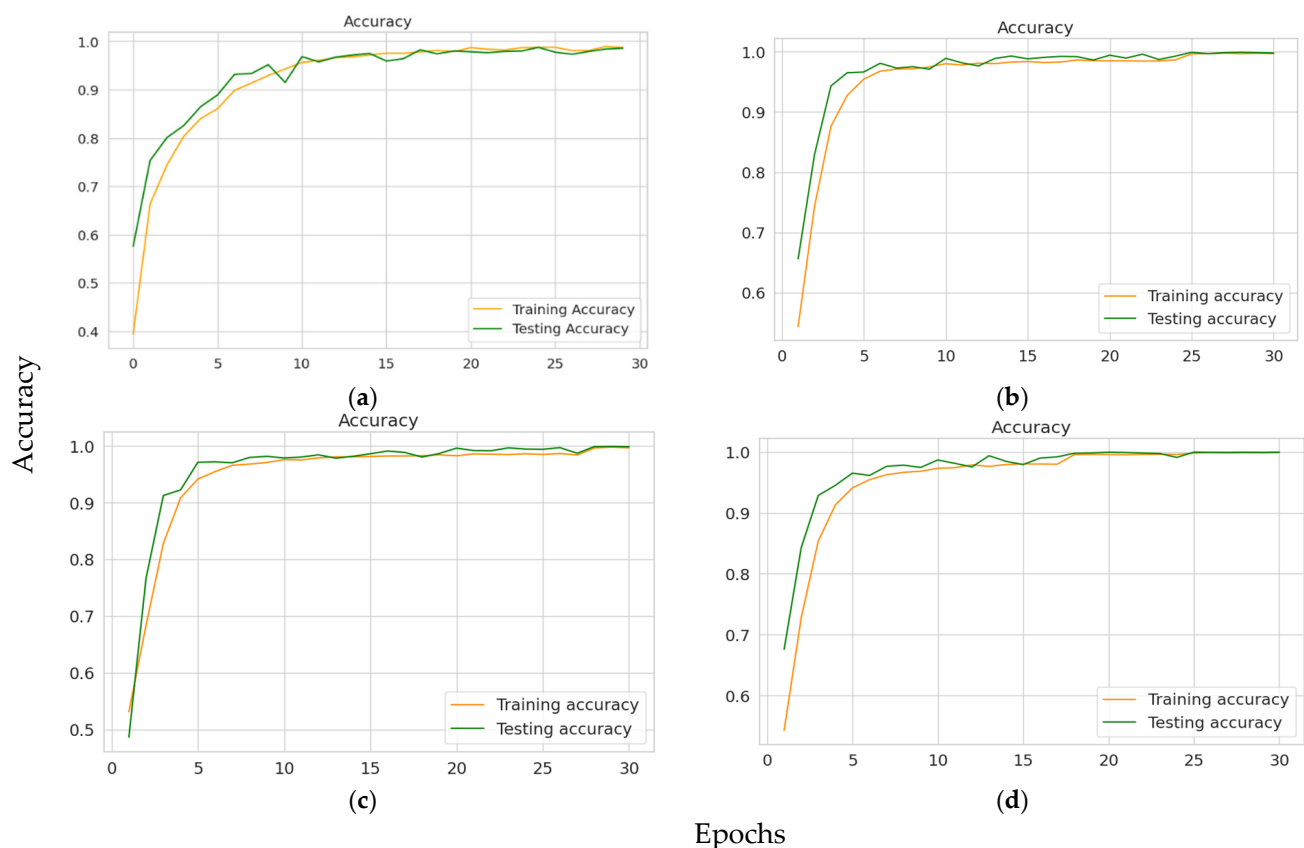


Figure 8. Training and testing accuracy curves: (a) CNN; (b) Swin-T; (c) hybrid CNN-Swin-T; and (d) the proposed model.

In contrast, the Swin-T model, shown in Figures 8b and 9b, demonstrates more stable and consistent learning. The accuracy curves exhibit rapid convergence with a small gap between training and testing performance, while the loss curves demonstrate a smooth and steady decline. This indicates strong generalization capability and effective feature learning through global contextual modelling.

The hybrid CNN-Swin-T model further enhances training stability, as illustrated in Figures 8c and 9c. This model achieves faster convergence compared to the standalone CNN, with smoother accuracy progression and a more consistent reduction in loss values. This improvement highlights the benefit of combining local and global feature representations.

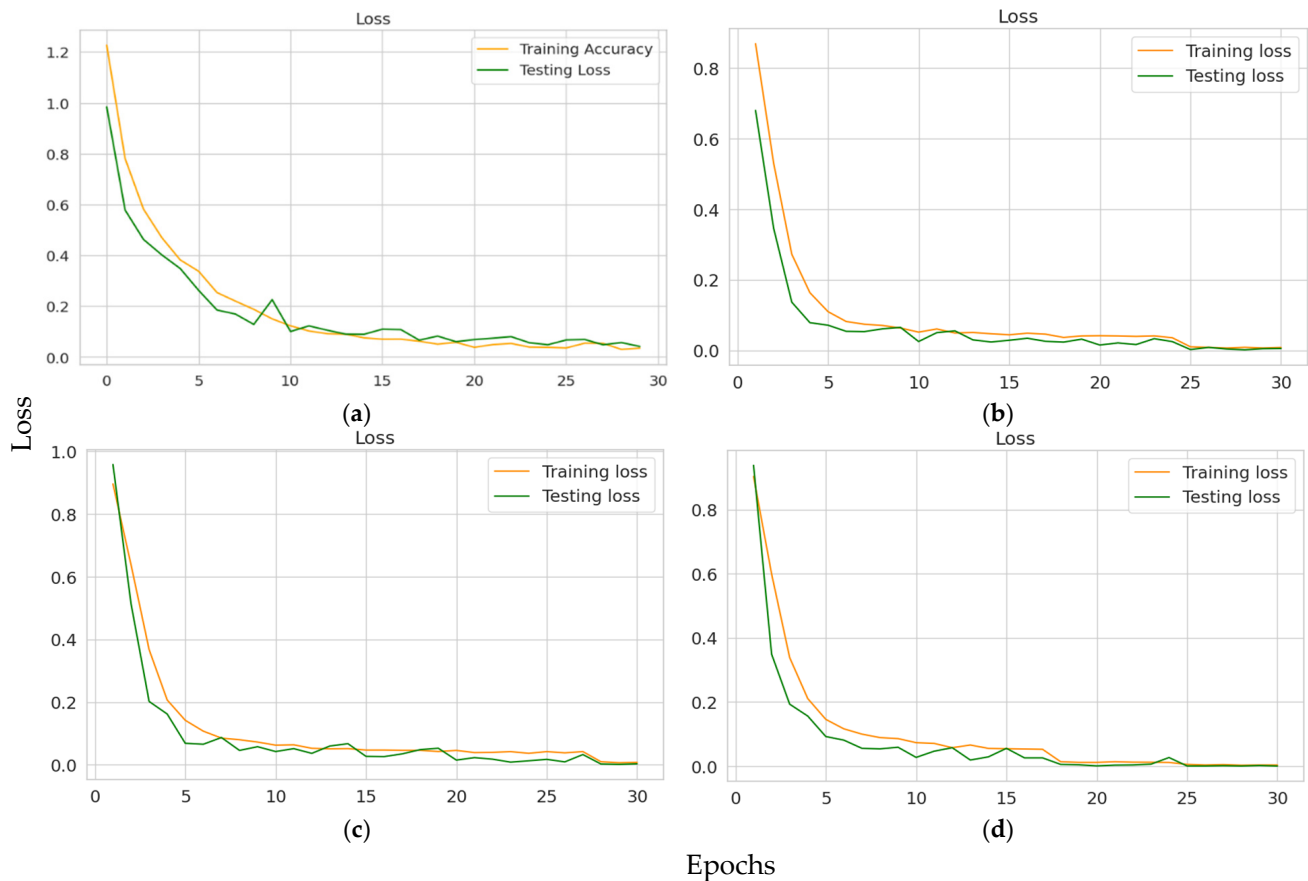


Figure 9. Training and testing loss curves: (a) CNN; (b) Swin-T; (c) hybrid CNN-Swin-T; and (d) the proposed model.

The proposed attention-enhanced hybrid CNN-Swin-T model achieves the most stable and efficient training performance, as shown in Figures 8d and 9d. This model demonstrates rapid convergence, high training accuracy, and minimal divergence between training and testing curves. The loss curves exhibit a smooth and consistent decline, indicating effective optimization and reduced overfitting. The integration of the attention technique enables the hybrid model to concentrate on the most important features, resulting in improved learning efficiency and generalization.

Overall, the analysis of training behaviour confirms that the proposed attention-enhanced hybrid CNN-Swin-T model not only achieves superior classification performance but also demonstrates stable and reliable learning characteristics. The combination of a hybrid CNN-Swin-T architecture and attention mechanism contributes significantly to improved convergence, reduced overfitting, and enhanced generalization capability.

5.3. Confusion Matrix Analysis

To assess the classification performance of the four models, confusion matrices and normalized confusion matrices are analysed. These matrices provide a detailed view of the models' predictions by illustrating the relationship between actual and predicted class labels, enabling identification of correctly classified instances as well as misclassification patterns across different classes.

The confusion matrices for all models are presented in Figure 10, while the corresponding normalized confusion matrices are shown in Figure 11. The normalized matrices provide a clearer interpretation by representing values as percentages, allowing for better comparison across classes regardless of sample size.

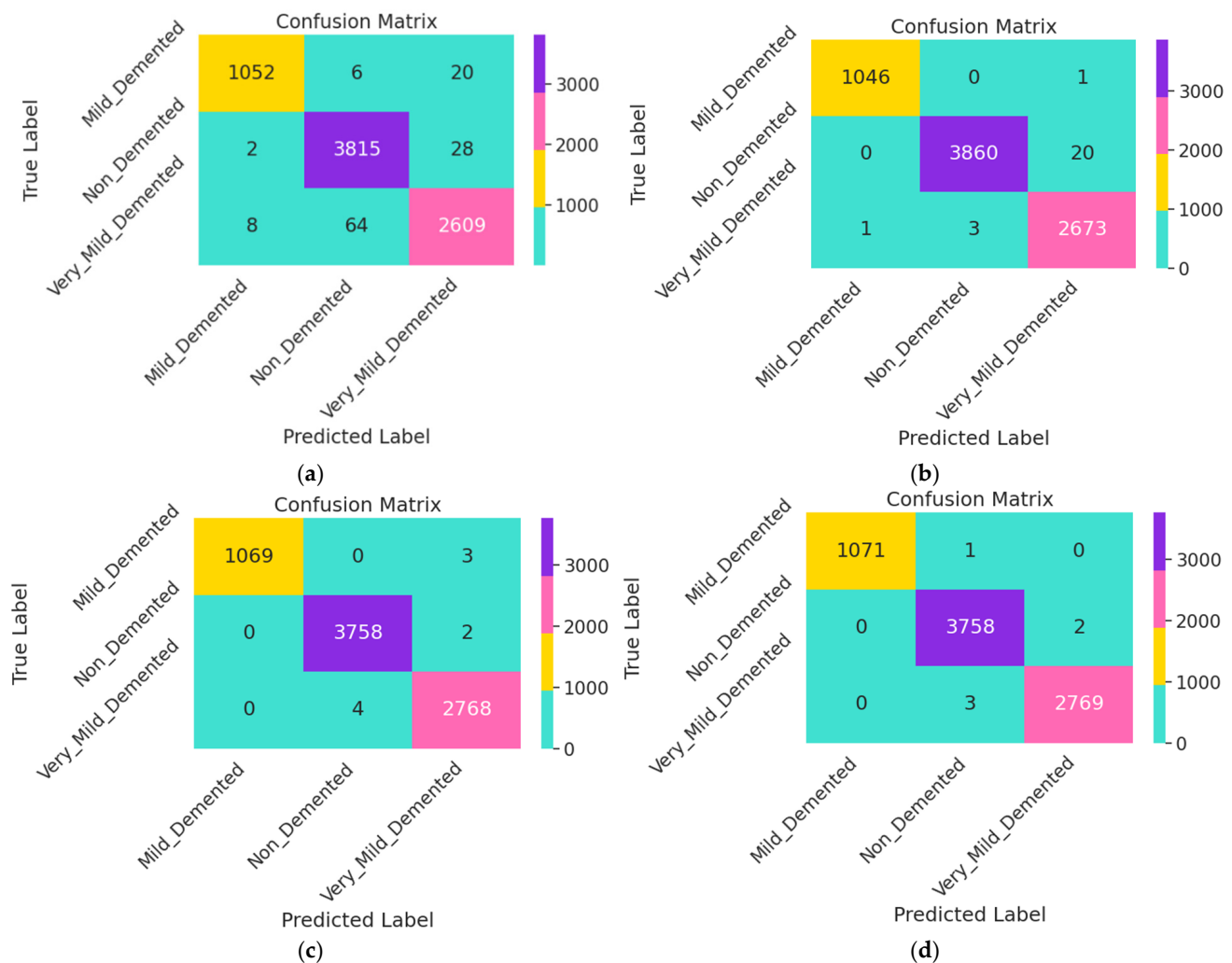


Figure 10. Confusion matrices: (a) CNN; (b) Swin-T; (c) hybrid CNN-Swin-T model; and (d) proposed attention-enhanced hybrid CNN-Swin-T model.

As illustrated in Figures 10a and 11a, the CNN model exhibits significant misclassification across all classes, particularly between the Mild Demented and Very Mild Demented categories. This indicates that the CNN struggles to distinguish subtle differences between similar disease stages, which is consistent with its low accuracy reported in Section 5.1.

In contrast, the Swin-T model, illustrated in Figures 10b and 11b, demonstrates a substantial improvement in classification performance. The majority of predictions are concentrated along the diagonal, indicating correct classification of most samples. However, minor misclassifications may still occur between closely related classes, reflecting the inherent difficulty of the task.

Also, the hybrid CNN-Swin-T model improves classification performance, as shown in Figures 10c and 11c. The combination of local and global feature extraction results in a higher concentration of correct predictions, with reduced confusion between classes. This confirms the effectiveness of integrating complementary feature representations.

The proposed attention-enhanced hybrid CNN-Swin-T model achieves the best performance, as illustrated in Figures 10d and 11d. The confusion matrix shows perfect classification, with almost all predictions correctly aligned along the diagonal. The normalized confusion matrix further confirms that the hybrid model integrated with an attention mechanism achieves highly accurate classification across the three classes, with minimal or no misclassification. This demonstrates that the model has a strong ability to distinguish between subtle variations in AD stages.

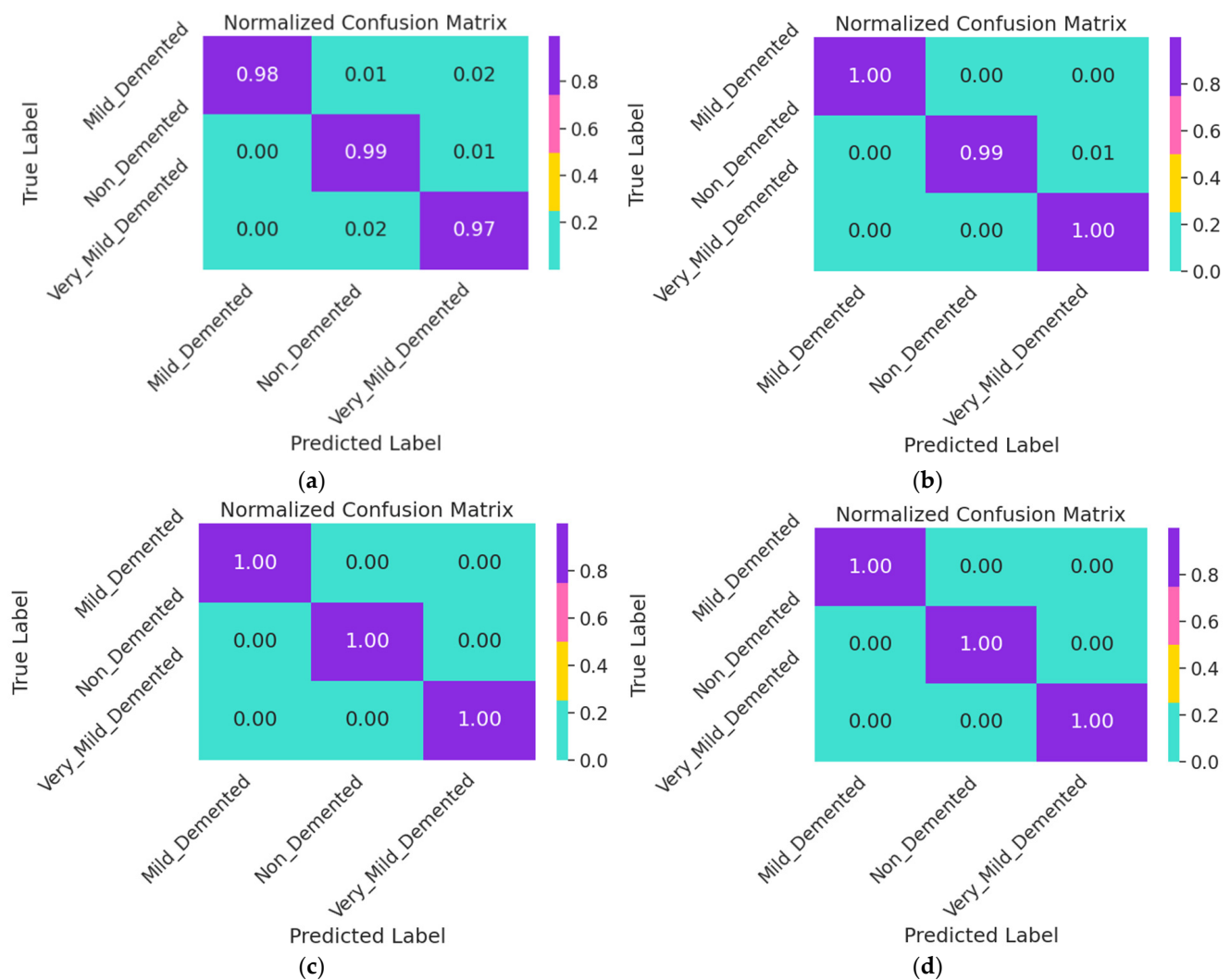


Figure 11. Normalized confusion matrices: (a) CNN; (b) Swin-T; (c) hybrid CNN-Swin-T model; and (d) proposed attention-enhanced hybrid CNN-Swin-T model.

Overall, the confusion matrix analysis highlights the progressive improvement in classification performance for the CNN, Swin-T, the hybrid CNN-Swin-T model, and the proposed attention-enhanced hybrid CNN-Swin-T model. The results confirm that the integration of attention mechanisms significantly enhances the hybrid model's ability to concentrate on relevant patterns and reduce misclassification, particularly in challenging cases involving similar classes.

5.4. Classification Report and Cross-Validation Analysis

To provide a more detailed evaluation of the proposed model's performance across individual classes, a classification report is analysed. This report includes the percentage of precision, recall, and F1-score for each class, offering deeper insights into how well the model distinguishes between different stages of AD. Table 7 presents the classification performance of the proposed attention-enhanced hybrid CNN-Swin-T model across the three classes: Non-Demented, Mild Demented, and Very Mild Demented.

These results indicate that the hybrid model integrated with attention achieves high classification performance across all classes, with precision, recall, and F1-score values. This demonstrates the model's ability to correctly detect all samples of each class without misclassification.

Table 7. Classification performance per class for the proposed attention-enhanced hybrid CNN–Swin–T Model.

Class	Precision (%)	Recall (%)	F1-Score (%)
Non-Demented	99.72	99.71	99.73
Mild Demented	99.71	99.73	99.72
Very Mild Demented	99.73	99.72	99.71

In particular, the attention-enhanced hybrid CNN–Swin–T model shows strong performance in distinguishing between Mild Demented and Very Mild Demented classes, which are typically challenging due to their subtle visual differences. The balanced performance across all classes confirms that the proposed model does not favor any specific class and is robust against class imbalance.

These findings further validate the effectiveness of the hybrid CNN–Swin–T architecture combined with the attention mechanism, as it enables the model to capture both fine-grained and global features while concentrating on the most important regions of the input images. Table 8 presents the k -fold cross-validation for the proposed attention-enhanced hybrid CNN–Swin–T model. The model attained an average accuracy of 99.93%, demonstrating robust performance across the 5-fold cross-validation process.

Table 8. k -Fold cross-validation for the proposed attention-enhanced hybrid CNN–Swin–T Model.

k -Fold Value	Accuracy (%)
1	99.91
2	99.94
3	99.88
4	99.97
5	99.95
Average	99.93

5.5. Explainability Results

To validate the reliability and interpretability of the four models, XAI techniques are applied to analyse the decision-making behaviour of each architecture. Three complementary methods: Grad-CAM++, Integrated Gradients, and Saliency maps are used to visualize the parts of the input images that contribute most significantly to the classification results. These techniques are applied consistently across all models, including CNN, Swin-T, hybrid CNN–Swin-T, and the proposed attention-enhanced hybrid CNN–Swin–T model. The goal is to provide a qualitative comparison of how each model focuses on relevant regions and to evaluate the efficiency of the proposed architecture in learning meaningful representations.

5.5.1. Explainability Comparison Across Models

The explainability results for all models are shown in Figures 12–14, corresponding to Grad-CAM++, Integrated Gradients, and Saliency maps, respectively.

Each figure includes visualizations generated for the same input samples to ensure a fair comparison across models. From the Grad-CAM++ visualizations in Figure 12, it can be observed that the CNN model produces relatively diffuse and less focused activation maps, indicating a limited ability to localize important regions.

The Swin-T model improves localization by capturing broader contextual information; however, some irrelevant regions may still be highlighted.

The hybrid CNN–Swin–T model demonstrates more concentrated activation regions, reflecting improved feature representation through the combination of local and global features.

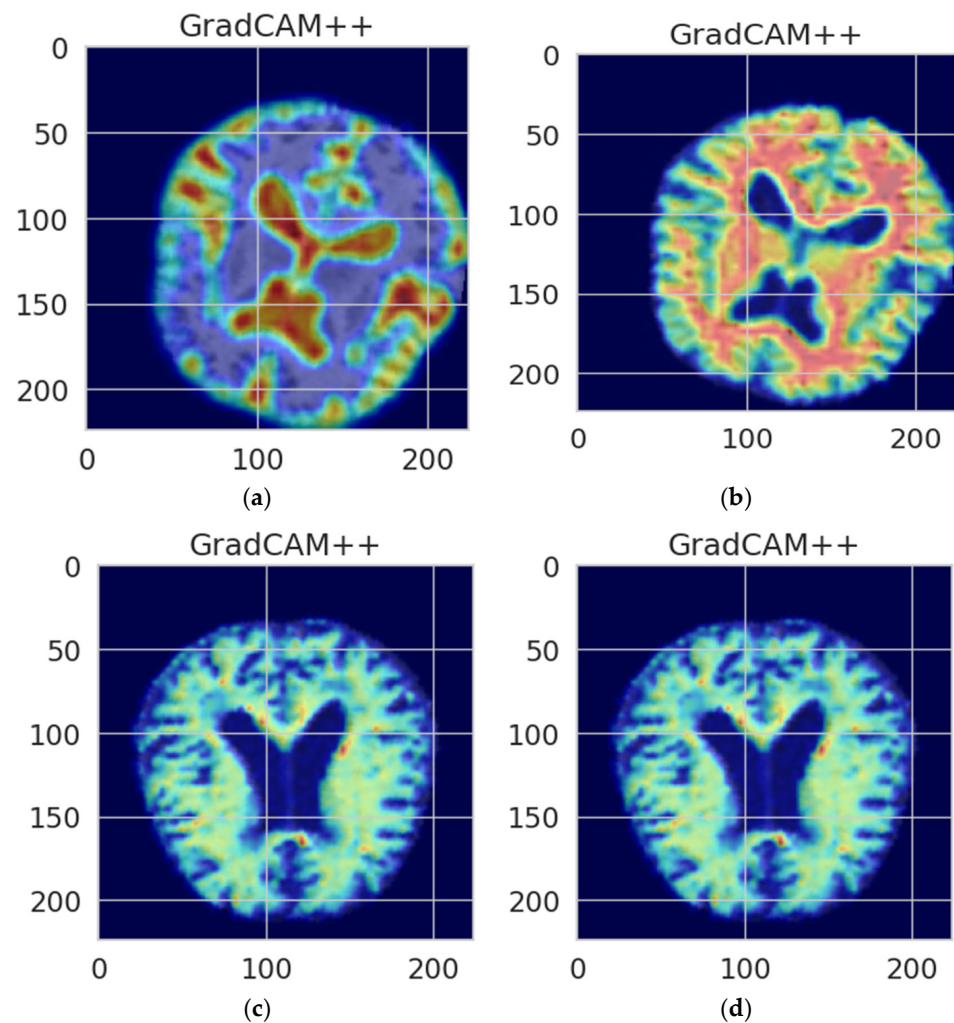


Figure 12. Grad-CAM++ visualizations: (a) CNN; (b) Swin-T; (c) hybrid CNN–Swin-T model; and (d) proposed attention-enhanced hybrid CNN–Swin-T model, highlighting the most influential regions contributing to classification decisions.

The proposed attention-enhanced hybrid CNN–Swin-T model shows the most precise and focused activation patterns, highlighting only the most relevant regions associated with AD. This indicates that the attention mechanism effectively enhances feature prioritization and suppresses irrelevant information.

Similarly, the Integrated Gradients results shown in Figure 13 provide fine-grained attribution maps, revealing pixel-level contributions to the model’s predictions.

The proposed attention-enhanced hybrid CNN–Swin-T model exhibits clearer and more meaningful patterns compared to the other three models, indicating improved sensitivity to subtle structural variations in the input images.

The Saliency maps presented in Figure 14 further support these observations. While the CNN model shows scattered and noisy gradients, the proposed attention-enhanced hybrid CNN–Swin-T model produces more structured and interpretable saliency patterns, emphasizing important edges and regions relevant to classification.

Overall, the comparison across models demonstrates that the proposed attention-enhanced hybrid CNN–Swin-T model achieves superior interpretability by focusing more accurately on clinically relevant regions.

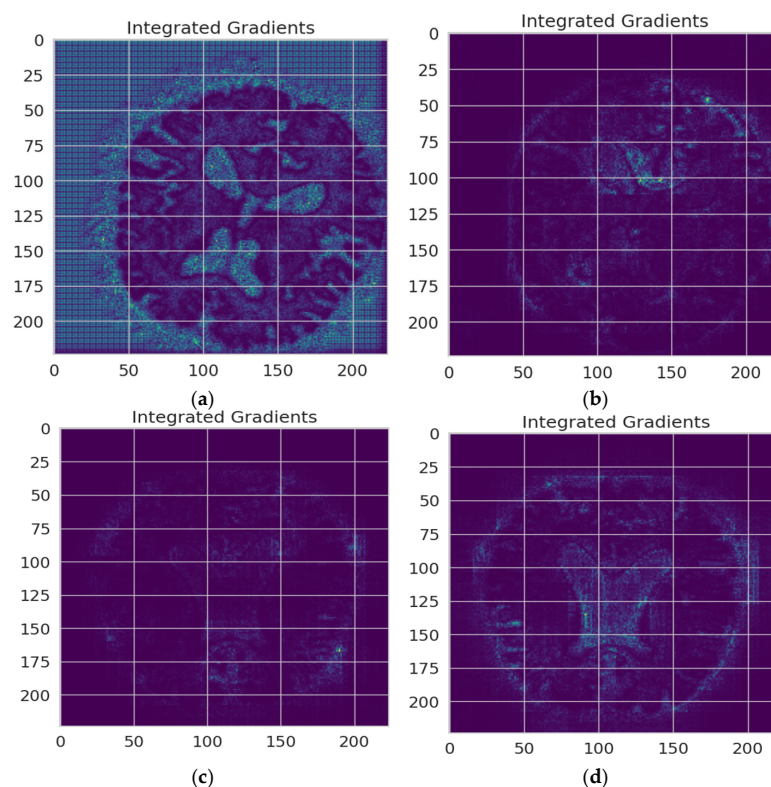


Figure 13. Integrated Gradient visualizations: (a) CNN; (b) Swin-T; (c) hybrid CNN–Swin-T model; and (d) proposed attention-enhanced hybrid CNN–Swin-T model, showing pixel-level feature attribution across input images.

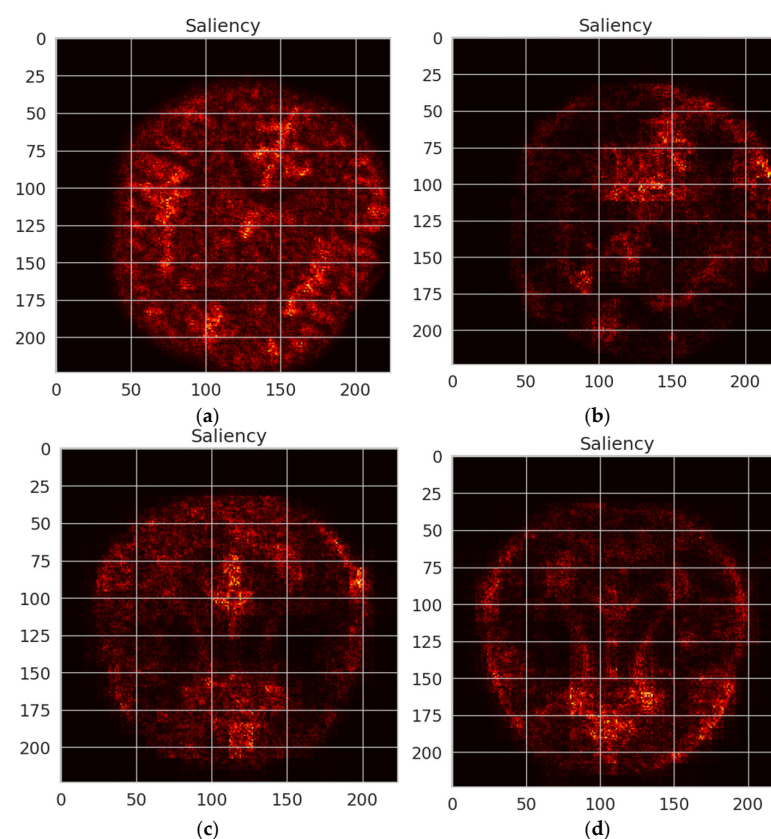


Figure 14. Saliency map visualizations: (a) CNN; (b) Swin-T; (c) hybrid CNN–Swin-T model; and (d) proposed attention-enhanced hybrid CNN–Swin-T model, emphasizing important edges and regions relevant to classification.

5.5.2. Class-Wise Explainability Analysis

Further, to investigate the interpretability of the proposed attention-enhanced hybrid CNN–Swin-T model, a class-wise analysis is conducted using representative samples from each category: Non-Demented, Mild Demented, and Very Mild Demented. The results are illustrated in Figure 15. The visualizations show that the proposed model effectively adapts its attention depending on the class of the input image. For Non-Demented samples, this model focuses on normal brain structures with distributed attention patterns. In contrast, for Mild Demented and Very Mild Demented cases, the model concentrates on specific regions associated with early structural changes and brain atrophy.

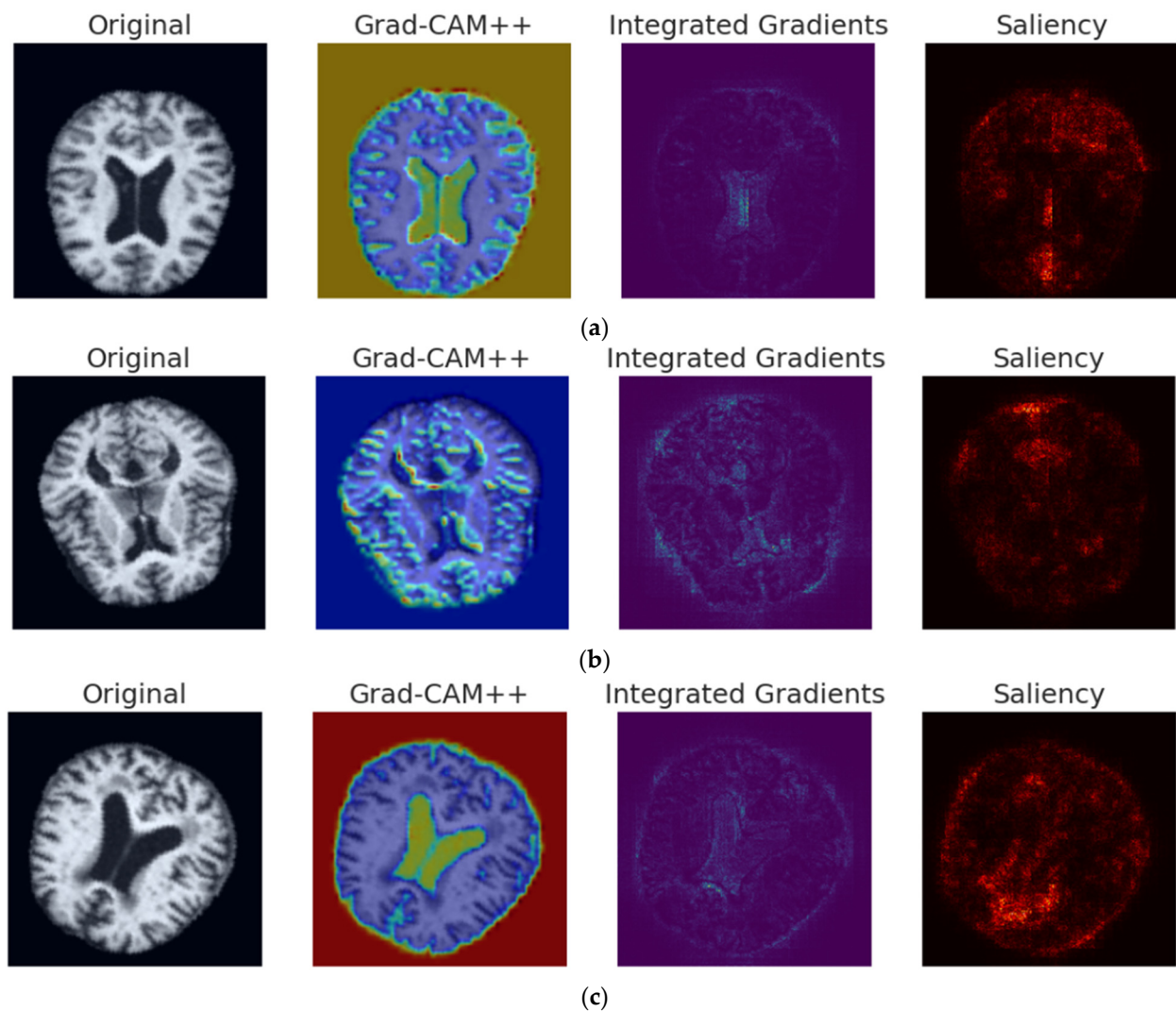


Figure 15. Explainability results for the proposed attention-enhanced hybrid CNN–Swin-T model on representative samples from each class: (a) Non-Demented; (b) Mild Demented; and (c) Very Mild Demented, using Grad-CAM++, Integrated Gradients, and Saliency maps.

This behaviour indicates that the proposed attention-enhanced hybrid CNN–Swin-T model is capable of capturing subtle differences between disease stages and adjusting its focus accordingly. The ability to highlight meaningful and class-specific regions enhances the interpretability of the model and provides additional confidence in its predictions.

6. Discussion

The experimental results that are presented in the previous section provide comprehensive evidence of the efficiency of the proposed attention-enhanced hybrid CNN–Swin-T

model for AD classification. This section discusses the key findings, analyzes the performance improvements, and highlights the significance of the proposed approach in a comparison with baseline models.

6.1. Impact of Model Architecture

The results show a clear improvement in performance from the standalone CNN model to the Swin-T, hybrid CNN–Swin-T model, and the proposed attention-enhanced hybrid CNN–Swin-T model. The CNN model, while effective in capturing local spatial features, exhibits limited performance due to its inability to model global contextual relationships. This limitation is particularly evident in its difficulty in distinguishing between the Mild Demented and Very Mild Demented classes, which require fine-grained feature discrimination.

The Swin-T model significantly improves the performance by capturing global dependencies through self-attention mechanisms. However, relying solely on transformer-based representations may not fully exploit local structural details essential for medical image analysis. The hybrid CNN–Swin-T model addresses this limitation by combining local and global feature extraction, resulting in improved classification performance and stability.

The proposed hybrid CNN–Swin-T model is enhanced by incorporating an attention mechanism, which enables the network to focus on the most relevant regions of the input images. This leads to improved feature selection and more discriminative representations, ultimately resulting in superior classification accuracy and robustness.

6.2. Effect of Data Augmentation

The results clearly indicate that data augmentation plays a critical role in enhancing the models' performance. The augmentation process significantly increases dataset size and diversity, allowing the models to generalize better to unseen data. This is particularly important in medical imaging applications, where datasets are often limited.

As observed in the experimental results, all models benefit from data augmentation; however, the improvement is more pronounced in the hybrid CNN–Swin-T and proposed attention-enhanced hybrid CNN–Swin-T models. The CNN model shows only modest improvement, indicating that data augmentation alone cannot compensate for architectural limitations. In contrast, the proposed attention-enhanced hybrid CNN–Swin-T model effectively leverages the augmented data, achieving near-perfect performance. This highlights the importance of combining robust architectures with appropriate data augmentation strategies.

6.3. Training Stability and Convergence

The analysis of training and testing curves reveals that the proposed attention-enhanced hybrid CNN–Swin-T model exhibits superior convergence behaviour and training stability. Compared to the three implemented models: CNN, Swin-T and CNN–Swin-T, which show slower convergence, the proposed hybrid model with attention achieves rapid and stable convergence with minimal divergence between learning and testing performance.

This stability can be attributed to the combined effect of hybrid feature extraction and attention mechanisms, which facilitate efficient learning and reduce overfitting. The smooth decline in loss values and consistent increase in accuracy further confirm the effectiveness of the optimization strategy and model design.

6.4. Analysis of Misclassification Patterns

The confusion matrix analysis offers valuable insights into the model's performance at the class level. The CNN model shows significant misclassification, particularly between

Mild Demented and Very Mild Demented categories, highlighting its inability to capture subtle differences.

The Swin-T and hybrid CNN–Swin-T models reduce these misclassifications, indicating improved feature representation. The proposed attention-enhanced hybrid CNN–Swin-T model achieves near-perfect classification, with minimal or no confusion between classes. This demonstrates its strong capability to distinguish between closely related disease stages, which is critical for clinical applications.

6.5. Interpretability and Clinical Relevance

One of the strengths of this paper is the integration of XAI methods. The XAI results demonstrate that the proposed attention-enhanced hybrid CNN–Swin-T model focuses on meaningful and relevant regions of the input images, providing visual evidence of its decision-making process.

Compared to baseline models, the proposed attention-enhanced hybrid CNN–Swin-T model produces more precise and less noisy activation maps, indicating improved feature localization. This enhanced interpretability is essential in medical applications, where transparency and trust are critical for adoption by healthcare professionals. The ability of this model to highlight clinically relevant regions not only validates its predictions but also provides additional insights that can assist clinicians in understanding disease progression.

6.6. Discussion on Interpretability

The integration of multiple XAI techniques provides a comprehensive understanding of the models' decision-making process. By combining region-based and pixel-level explanations, the proposed framework offers both global and local interpretability.

The results demonstrate that the proposed attention-enhanced hybrid CNN–Swin-T model not only achieves high classification performance but also provides more reliable and interpretable explanations compared to baseline models. This is particularly important in medical applications, where transparency and trust are essential for practical deployment. Overall, the explainability analysis confirms that the proposed framework is capable of producing accurate and interpretable predictions, making it a promising solution for AD classification in clinical settings.

6.7. Ablation Study

The improved performance of the four models can be largely attributed to careful hyperparameter tuning, including the number of learning epochs, choice of the loss function, batch size of data during the training phase, learning rate, and the type of optimizer. For the CNN model, setting the batch size to 128, learning rate to 0.0001 with Adam optimizer, and training epochs to 10, along with a categorical cross-entropy loss function, reached a testing accuracy of 97.89%. However, when the batch size is decreased to 32, the learning rate is reduced to 0.00001, and the number of epochs is increased to 30 (while keeping the same loss function and optimizer), the testing accuracy increased to 98.32%. Similarly, for the Swin-T and hybrid CNN–Swin-T models, using a batch size of 128, a learning rate of 0.0001 with Adam, and 10 training epochs with a categorical cross-entropy loss function yielded testing accuracies of 99.32% and 99.43%, respectively. When these hyperparameters are adjusted to a batch size of 32, a learning rate of 0.00001, and 30 epochs, the accuracies reached to 99.67% and 99.88%, respectively. Eventually, for the proposed attention-enhanced hybrid CNN–Swin-T model, setting the batch size to 128, learning rate to 0.0001 with Adam optimizer, and training epochs to 10, along with a categorical cross-entropy loss function, resulted in a testing accuracy of 99.47%. However, when the batch size is decreased to 32, the learning rate is reduced to 0.00001, and the number of epochs is

increased to 30 (while keeping the same loss function and optimizer), the testing accuracy increased to 99.92%. These results demonstrate that proper optimization and selection of hyperparameters play a critical role in improving the models' performance. Overall, the best results for all four models were achieved using a batch size of 32, 30 training epochs, and a learning rate of 0.00001, combined with the Adam optimizer and a categorical cross-entropy. Figure 16 shows a bar chart showing the effect of the hyperparameters on the models' performance.

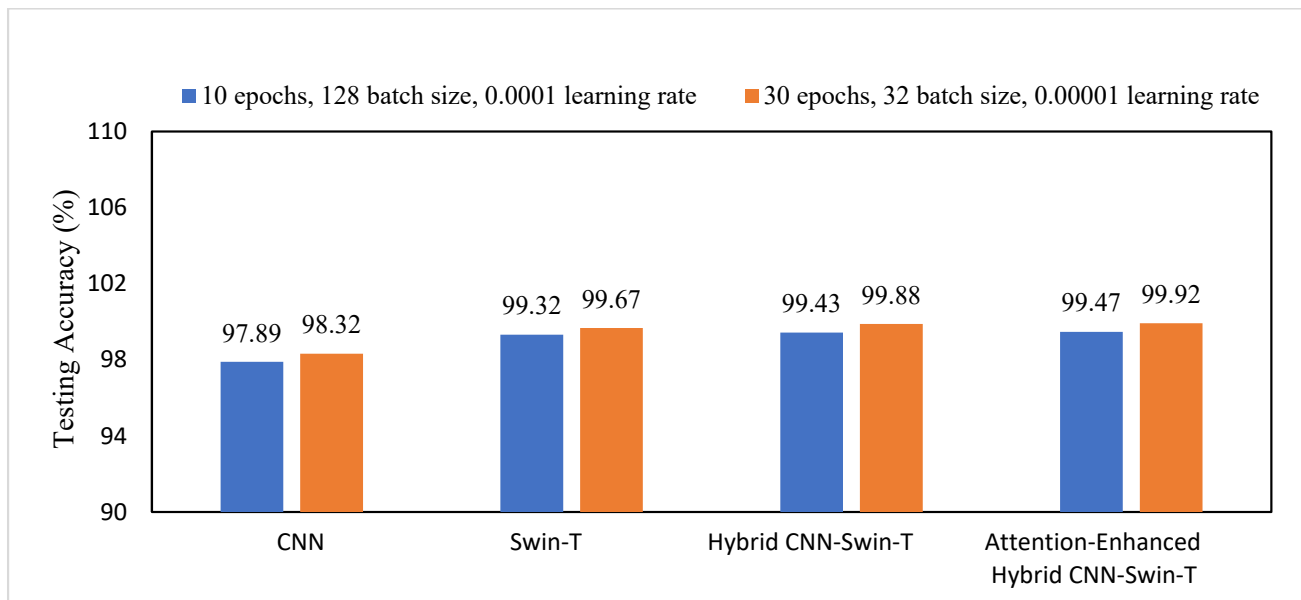


Figure 16. A bar chart showing the effect of the hyperparameters on the models' performance.

It is essential that the proposed model has high performance to predict the classes. So, the significance of improvement for the model is very important in the generalization phase. The improvement in the accuracy percentage is due to the utilized techniques and the optimal fine-tuning of the model's hyperparameters.

6.8. Overall Performance and Practical Implications

The proposed attention-enhanced hybrid CNN–Swin-T model achieves the highest performance across all evaluation metrics, with a testing accuracy of 99.92% and high precision, recall, and F1-score, 99.71%, 99.73%, and 99.72%, respectively. These results demonstrate the effectiveness of combining hybrid architectures, attention mechanisms, and data augmentation. From a practical perspective, the proposed framework offers a reliable and interpretable solution for AD classification. Its high accuracy and transparency make it suitable for potential deployment in clinical decision support systems, where both performance and interpretability are essential. Table 9 presents a comparison between the previously published studies and the proposed work. This comparison is in terms of dataset, model, accuracy, strengths, and weaknesses. The proposed attention-enhanced hybrid CNN–Swin-T model achieved a testing accuracy of 99.92%, outperforming the models reported in [15–32]. The 99.92% accuracy is due to the utilized data augmentation techniques and the optimal fine-tuning of the model's hyperparameters.

Table 9. Comparison of the proposed work and previous published studies on AD classification using DL techniques.

Reference	Year	Dataset	Model	Accuracy	Strengths	Weaknesses
[15]	2025	OASIS MRI dataset	EfficientNet-B0 + Swin-T + Attention + DCGAN + Grad-CAM	95.3%	Hybrid local/global feature learning, attention fusion, DCGAN for imbalance, and Grad-CAM interpretability.	Higher computational complexity, used 2D slices, and had limited clinical validation.
[16]	2026	Kaggle MRI (6400 images)	ResNeXt101_32 × 8d + Swin-T + XAI	98.83%	Very high performance, hybrid CNN-transformer design, multiple XAI methods, cross-validation.	Computationally expensive, class imbalance remains a concern, and external clinical validation was limited.
[17]	2025	Clinical CT dataset	EfficientNet + Swin-T	95.83%	Learnable attention fusion, Grad-CAM, balanced four-class setup, multimodal medical-imaging relevance.	CT rather than MRI, not Alzheimer-specific, relatively small dataset.
[18]	2025	Multiple datasets	Survey (CNN, ViT, Hybrid, and XAI)	N/A	Comprehensive taxonomy, discussed CNN/ViT/hybrid/XAI methods, highlighted data scarcity and interpretability challenges.	No proposed experimental model.
[19]	2025	MRI + demographic	CNN + ViT + LightGBM	96.83%	Hybrid deep feature extraction plus ML classifier, demographic fusion, Grad-CAM, and statistical validation.	Pipeline complexity and dependence on extra non-imaging data may reduce generalizability.
[20]	2025	MRI + RNA	Transformer + LIME	99%	Multimodal fusion, transformer-based classification, explicit XAI, robust reported performance.	Relied on RNA data; architecture details were limited on the provided page, which might be harder to deploy broadly.
[21]	2025	MRI (Parkinson's)	CNN + Swin Fusion	96%	Parallel local/global feature fusion, attention-based integration, and good MRI relevance.	Parkinson's rather than Alzheimer's, no explicit XAI component.
[22]	2025	ADNI (rs-fMRI)	CNN + Transformer	96%	Multimodal ADNI setup, attention-guided CNN-transformer fusion, strong multiclass results, computationally efficient relative to benchmarks.	Depended on rs-fMRI and clinical data, integration complexity was higher than in single-modality models.
[23]	2025	MRI	Deformable ViT	80.31%	Efficient transformer design with deformable attention, lower parameter count, suitable for resource-limited settings.	No CNN branch, lower accuracy than several hybrid models, and no explicit XAI.
[24]	2026	MRI	BEiT + GAN + XAI	96%	Transformer-based AD prediction, GAN-assisted class balancing, explicit explainability, strong F1/MCC/Kappa.	No CNN integration, pipeline was complex, and the interpretability method was not compared in detail on the provided page.

Table 9. Cont.

Reference	Year	Dataset	Model	Accuracy	Strengths	Weaknesses
[25]	2025	MRI + retinal	EfficientNet + ViT	99.31%	Fusion of CNN and transformer features, Grad-CAM, and good cross-domain validation.	Not Alzheimer-specific; architecture appears relatively complex.
[26]	2025	Multi imaging	ConvNeXt + Swin-T	N/A	Learnable channel/spatial fusion, CNN-Swin integration, and multi-modality validation.	Not AD-specific, abstract did not quantify performance precisely, potentially high computational cost.
[27]	2025	PET/MR	Transformer	N/A	Interpretable transformer, structural + functional multimodal fusion, clinically relevant design.	Depended on PET/MR availability, no CNN branch, metrics not clearly quantified on the provided page.
[28]	2026	ADNI MRI (3D)	3D CNN + ViT + BiFPN	92%	Used 3D volumetric context, multiscale fusion via BiFPN, transformer refinement, and attention visualization.	High computational complexity, 3D processing burden, and lower accuracy than some 2D hybrid studies.
[29]	2025	Balanced/imbalanced MRI	ViT, CNN, Swin-T	99%	Explicitly studies data imbalance, large datasets, and compares CNN and transformer families.	No hybrid architecture, no XAI, focused on evaluation rather than proposing a new model.
[30]	2025	3D MRI	CNN + Attention	N/A	Lightweight CNN-attention model, efficient parameter usage, activation-map interpretability, 3D MRI relevance.	No transformer integration, limited global-context modeling compared with hybrid transformer methods.
[31]	2025	MRI (AD + MCI)	ViT + XAI	N/A	Focused on early AD/MCI differentiation, effective transformer design, multi-granular explainability, and reduced parameters/FLOPs.	Without CNN fusion, the abstract did not provide a single explicit overall accuracy.
[32]	2023	Diffusion MRI	Swin-T	95.2%	Swin-T on advanced diffusion MRI, hierarchical representation learning, Grad-CAM, and biomarker-oriented evaluation.	Without the CNN branch, the specialized MRI modality limited practicality, and some tasks showed lower performance.
The proposed Work	2026	Alzheimer's MRI dataset of 6400 images available on Kaggle	CNN,	98.32%	Extracting the combination of local and global, and important features.	Deployment of the proposed models in a real-time environment
			Swin-T,	99.67%	Applying multiple AXI techniques: Grad-CAM++, Integrated Gradients, and Saliency maps. Applying, Data Augmentation	
			Hybrid CNN–Swin-T,	99.88%		
			Attention-Enhanced Hybrid CNN–Swin-T	99.92%		

6.9. Limitations of the Proposed Work

Despite the promising performance achieved by the proposed hybrid CNN–Swin-T framework with attention mechanisms and XAI integration, several limitations should be presented.

6.9.1. Dependence on Dataset Size and Diversity

Although data augmentation techniques are applied to improve generalization, the performance of the proposed models may still be affected by the limited size and diversity of the available medical imaging datasets. Models trained on restricted datasets may not generalize effectively to unseen clinical data collected from different imaging devices or patient populations.

6.9.2. High Computational Complexity

The integration of CNN, Swin-T, attention mechanisms, and multiple XAI techniques increases the overall computational cost of the framework. This leads to higher training time, increased memory consumption, and greater hardware requirements, which may limit deployment in resource-constrained clinical environments.

6.9.3. Risk of Overfitting

Due to the complexity and large number of trainable parameters in the hybrid architecture, the proposed model may still be susceptible to overfitting, particularly when trained on relatively small medical datasets despite the use of augmentation strategies.

6.9.4. Limited Clinical Validation

The proposed framework is primarily evaluated on benchmark datasets under experimental conditions. Further validation on large-scale real-world clinical datasets and multi-center studies is necessary to confirm the robustness, reliability, and clinical applicability of the model.

6.9.5. Lack of Multi-Modal Data Integration

The proposed framework focuses mainly on imaging data and does not incorporate other clinically relevant information, such as patient history, laboratory results, or genetic data, which could further enhance diagnostic accuracy and robustness. Also, it is essential to evaluate the framework on well-established clinical datasets such as ADNI and OASIS would further strengthen the clinical reliability and external validity of the proposed models.

6.9.6. Generalization to Other Medical Imaging Modalities

The proposed framework is evaluated on a specific medical imaging task, and its effectiveness for other imaging modalities or diseases remains to be investigated.

6.9.7. Trade-Off Between Accuracy and Time

While the proposed CNN + Swin-T + Attention framework achieved superior diagnostic performance, the increased architectural complexity may reduce inference speed, which could affect real-time clinical deployment scenarios.

7. Conclusions and Future Work

This paper presents an attention-enhanced hybrid DL framework that combines CNN and Swin-T architectures for multi-class AD classification. The proposed model leverages the strengths of CNNs in capturing local spatial features and transformer-based models in modelling global contextual relationships, while the integration of an attention mechanism enhances feature representation by concentrating on the most relevant regions of the input images. The results reveal that the hybrid CNN–Swin-T model with attention outperforms baseline architectures, including standalone CNN, Swin-T, and hybrid CNN–Swin-T models. The proposed attention-enhanced hybrid CNN–Swin-T model achieves a classification accuracy of 99.92%, along with high precision, F1-score, and recall, 99.71%, 99.73%, and

99.72%, respectively, indicating its robustness and effectiveness in distinguishing between Non-Demented, Mild Demented, and Very Mild Demented classes.

The results highlight the importance of data augmentation in improving the model's generalization and efficiency, especially when working with limited medical MRI images. Besides high classification accuracy, the integration of XAI techniques, including Grad-CAM++, Integrated Gradients, and Saliency maps, provides valuable insights into the model's decision-making process. The explainability analysis confirms that the proposed attention-enhanced hybrid CNN–Swin-T model focuses on meaningful and clinically relevant regions, enhancing transparency and increasing trust in automated diagnostic systems. The efficiency of the four models is strongly influenced by their hyperparameters, including the loss function, learning rate, type of optimizer, number of learning epochs, and the batch size of data during the learning phase. The high testing accuracy achieved is closely associated with the careful selection and optimal tuning of these hyperparameters.

Despite these promising results, several directions for future work can be explored. Firstly, the proposed framework can be evaluated on larger and more diverse datasets, including real-world clinical data, to further assess its generalizability. Secondly, integrating multi-modal data, for instance, merging medical images with clinical or genetic information, could enhance diagnostic accuracy and provide a better understanding of AD progression. Thirdly, optimizing the model's complexity and its computational efficiency would facilitate the deployment in real-time clinical environments. Fourthly, advancements in XAI can be incorporated to provide more detailed and user-friendly interpretations, enhancing the practical applicability of the system in clinical settings by applying other XAI techniques such as LIME and SHAP. Fifthly, one could apply standard attention techniques such as SE-Net or CBAM.

Author Contributions: S.M., S.K., N.A. (Norah Alnaim), N.A. (Noorah Albehaijan) and A.F.I. contributed to writing the original draft of the manuscript, conceptualization, methodology, resources, visualization, similarity reduction, the editing of the manuscript, the review of the writing and grammatical errors of the manuscript, and supervision of the proposed work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the paper are openly available in [Kaggle Repository] at [<https://www.kaggle.com/datasets/ahamedfarouk/alzheimer>] (accessed on 15 March 2026)].

Conflicts of Interest: All authors declare that there are no conflicts of interest.

References

1. Chaddad, A.; Desrosiers, C.; Niazi, T. Deep radiomic analysis of MRI related to Alzheimer's disease. *IEEE Access* **2018**, *6*, 58213–58221. [[CrossRef](#)]
2. Kaya, M.; Çetin-Kaya, Y. A novel deep learning architecture optimization for multiclass classification of Alzheimer's disease level. *IEEE Access* **2024**, *12*, 46562–46581. [[CrossRef](#)]
3. Sar, A.; Kumar, A.; Roy, S.; Kaushish, A.; Choudhury, T.; Minakshi; Alqahtani, S.S.; Abraham, A. Quantum machine learning in medical image analysis: From diagnostics to surgery planning. *IEEE Access* **2026**, *14*, 27690–27732. [[CrossRef](#)]
4. Mohsen, S.; Oraby, S.; Abdel-Aziz, M. Deep learning and machine learning for brain tumor detection: A review, challenges, and future directions. *Arch. Comput. Methods Eng.* **2025**, *33*, 3931–3955. [[CrossRef](#)]
5. Mienye, I.D.; Swart, T.G.; Obaido, G.; Jordan, M.; Ilono, P. Deep convolutional neural networks in medical image analysis: A review. *Information* **2025**, *16*, 195. [[CrossRef](#)]
6. Xu, Y.; Ying, W.; Du, Y.; Wang, Q.; Zheng, J.; Yang, Z. Swin Transformer-based intelligent fracture classification and radiographic assessment to support clinical decision-making. *BMC Med. Inform. Decis. Mak.* **2026**, *26*, 42. [[CrossRef](#)]
7. Nawaz, K.; Zani, A.; Wang, Y.; Ayesha, N.; Bahaj, S.A.O.; Mahmood, T. Optimized deep learning-based multi-stage Alzheimer's disease characterization with temporal modeling and soft-NMS for neuroimaging. *Discov. Comput.* **2026**, *29*, 36. [[CrossRef](#)]

8. Sharma, V.K.; Rout, R.K.; Tanwar, R.; Singh, P.K. An efficient approach to synthesize medical images for data augmentation with constrained training data. *Discov. Comput.* **2026**, *29*, 82. [CrossRef]
9. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
10. Alzheimer Dataset. Kaggle Repository. Available online: <https://www.kaggle.com/datasets/ahamedfarouk/alzheimer> (accessed on 25 May 2026).
11. Shaikh, M.R.; Jeyabose, A.; Arjunan, R.V. Deep learning for Alzheimer’s disease: Advances in classification, segmentation, subtyping, and explainability. *Biomed. Eng. Online* **2025**, *24*, 150. [CrossRef] [PubMed]
12. Bricaud, I.; Masala, G.L. Multi-architecture deep learning for early Alzheimer’s detection in MRI: Slice- and scan-level analysis. *Int. J. Environ. Res. Public Health* **2026**, *23*, 322. [CrossRef]
13. Abbas, Q.; Jeong, W.; Lee, S.W. Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare* **2025**, *13*, 2154. [CrossRef]
14. Hasan Saif, F.; Al-Andoli, M.N.; Bejuri, W.M.Y.W. Explainable AI for Alzheimer detection: A review of current methods and applications. *Appl. Sci.* **2024**, *14*, 10121. [CrossRef]
15. Velu, K.; Jaisankar, N. Design of a CNN–Swin Transformer model for Alzheimer’s disease prediction using MRI images. *IEEE Access* **2025**, *13*, 149409–149429. [CrossRef]
16. Mohsen, S.; Yousef, A.; Abdel-Aziz, M. A hybrid of ResNeXt101_32x8d and Swin transformer networks with XAI for Alzheimer’s disease detection. *Computers* **2026**, *15*, 95. [CrossRef]
17. Alhumaid, M.; Fayoumi, A.G. Hybrid CNN–Swin transformer model to advance the diagnosis of maxillary sinus abnormalities on CT images using explainable AI. *Computers* **2025**, *14*, 419. [CrossRef]
18. Hosny, K.M.; Mohammed, M.A. Explainable AI and vision transformers for detection and classification of brain tumor: A comprehensive survey. *Artif. Intell. Rev.* **2025**, *58*, 259. [CrossRef]
19. Erdoğmuş, P.; Kabakuş, A.T. Early diagnosis of Alzheimer’s disease using hybrid CNN–transformer models with Grad-CAM interpretability. *Gümüşhane Univ. J. Sci. Technol.* **2025**, *15*, 829–853. [CrossRef]
20. Anzum, H.; Sammo, N.S.; Akhter, S. Leveraging transformers and explainable AI for Alzheimer’s disease interpretability. *PLoS ONE* **2025**, *20*, e0322607. [CrossRef] [PubMed]
21. Hussain, S.S.; Shah, P.M.; Dawood, H.; Degang, X.; Alshamayleh, A.; Khan, M.A.; Ghazal, T.M. A Swin transformer and CNN fusion framework for accurate Parkinson disease classification in MRI. *Sci. Rep.* **2025**, *15*, 15117. [CrossRef] [PubMed]
22. Alorf, A. Transformer and convolutional neural network: A hybrid model for multimodal data in multiclass classification of Alzheimer’s disease. *Mathematics* **2025**, *13*, 1548. [CrossRef]
23. Almansour, A.G.M.; Alshomrani, F.; Almutairi, A.T.M.; Alalwany, E.; Alshuhri, M.S.; Alshaari, H.; Alfahaid, A. DA-ViT: Deformable attention vision transformer for Alzheimer’s disease classification from MRI scans. *Comput. Model. Eng. Sci.* **2025**, *144*, 2395–2418. [CrossRef]
24. Saqib, S.M.; Alkhatabi, M.A.; Khan, M.A.; Mazhar, T.; Iqbal, M.; Saudagar, A.K.J.; Paracha, W.T.; Hamam, H. Explainable bidirectional encoder representations from image transformers for Alzheimer’s disease prediction. *Digit. Health* **2026**, *12*, 20552076261416823. [CrossRef]
25. Hussain, T.; Shouno, H.; Hussain, A.; Hussain, D.; Ismail, M.; Mir, T.H.; Hsu, F.R.; Alam, T.; Akhy, S.A. EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access* **2025**, *13*, 54040–54068. [CrossRef]
26. Qezelbash-Chamak, J.; Hicklin, K. A hybrid learnable fusion of ConvNeXt and Swin transformer for optimized image classification. *IoT* **2025**, *6*, 30. [CrossRef]
27. Yao, Z.; Xie, W.; Chen, J.; Zhan, Y.; Wu, X.; Dai, Y.; Pei, Y.; Wang, Z.; Zhang, G. IT: An interpretable transformer model for Alzheimer’s disease prediction based on PET/MR images. *NeuroImage* **2025**, *311*, 121210. [CrossRef]
28. Alsubaie, M.G.; Luo, S.; Shaikat, K.; Zhang, W. A novel Alzheimer’s disease classification framework integrating 3D CNN with bidirectional feature pyramid network and vision transformers. *IEEE Access* **2026**, *14*, 15241–15259. [CrossRef]
29. Almalki, H.; Khadidos, A.O.; Alhebaishi, N. Assessing model performance in Alzheimer’s disease classification: The impact of data imbalance on fine-tuned vision transformers and CNN architectures. *J. Intell. Syst.* **2025**, *34*, 20240406. [CrossRef]
30. Gan, Y.; Lan, Q.; Huang, C.; Su, W.; Huang, Z. Dense convolution-based attention network for Alzheimer’s disease classification. *Sci. Rep.* **2025**, *15*, 5693. [CrossRef]
31. Li, Y.; Xu, B.; Bai, Q.; Liu, Z.; Zhu, J.; Chen, Q. Vision transformer-based identification for early Alzheimer’s disease and mild cognitive impairment. *Information* **2026**, *17*, 129. [CrossRef]
32. Dessain, Q.; Delinte, N.; Hanseeuw, B.; Dricot, L.; Macq, B. Leveraging Swin transformer for enhanced diagnosis of Alzheimer’s disease using multi-shell diffusion MRI. *IEEE Trans. Biomed. Eng.* **2025**, Online ahead of print.
33. Mohsen, S.; Abo-Zahhad, M.M.; Abdel-Aziz, M. A deep super-resolution GAN framework integrating ResNet-18 and CNN for enhanced Alzheimer’s disease diagnosis. *Discov. Appl. Sci.* **2026**, *8*, 573. [CrossRef]

34. Reza, H.; Sibly, N.B.; Rabbani, S.G.; Sadi, S.H.; Ahamed, F.; Shafi, F.B.; Sarmun, R.; Chowdhury, M.E.H. A comprehensive review of convolutional neural networks: Foundations, enhancements and applications. *Neural Comput. Appl.* **2026**, *38*, 56. [[CrossRef](#)]
35. Li, H.; He, W.; Lan, A. Swin transformer-based real-time multi-tasking image detection in industrial automation production environments. *Machines* **2025**, *13*, 972. [[CrossRef](#)]
36. El-Den, N.N.; Elsharkawy, M.; Saleh, I.; Mahmoud, A.H.; Ghazal, M.; Khalil, A.; Sewelam, A.; Mahdi, H.; El-Baz, A. Seg-Swin: A dual-attention transformer model for advanced AMD classification and lesion detection using color fundus imaging. *IEEE Access* **2025**, *13*, 70489–70511. [[CrossRef](#)]
37. Chen, L.; Bai, Y.; Cheng, Q.; Wu, M. Swin transformer with local aggregation. In Proceedings of the 2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS), Guangzhou, China, 22–24 July 2022; pp. 77–81.
38. Soomro, S.; Niaz, A.; Choi, K.N. Grad++ScoreCAM: Enhancing visual explanations of deep convolutional networks using incremented gradient and score-weighted methods. *IEEE Access* **2024**, *12*, 61104–61112. [[CrossRef](#)]
39. Boardman, J.; Alam, M.S.; Huang, X.; Xie, Y. Integrated gradients is a nonlinear generalization of the industry standard approach to variable attribution for credit risk models. In Proceedings of the IEEE Big Data, Osaka, Japan, 17–20 December 2022; pp. 5012–5023.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.