

An Interpretable Dual-Path Convolutional Network for Multi-Class Alzheimer's Disease Staging from Structural MRI

Seba Aziz Sahy ^{1,*}, Sura hameed mahdi², Ayat Fadel Saddam³

¹ Middle Technical university, Institute of Medical Technology Al-Mansour, Baghdad, 10081 IRAQ, Saba.aziz@mtu.edu.iq.

² Polytechnic college of engineering specializations, Middle technical university, Baghdad, 10081 IRAQ, sura-hameed@mtu.edu.iq.

³ Intelligent Medical Systems Department, Science College, Al-Esraa University, Baghdad, 10081 IRAQ, ayatfadal097@gmail.com.

Abstract Alzheimer disease (AD) is a developmental neurodegenerative disorder impacting more than 55 million individuals worldwide and early, accurate staging is an elusive clinical necessity. This paper introduces a new deep learning system that is capable of staging Alzheimer disease in multiple classes automatically using structural Magnetic Resonance Imaging (MRI). Our architecture (Dual-Path Convolutional Neural Network (DP-CNN)) constitutes parallel ResNet-50 and VGG-16 feature extractors with a spatial squeeze-and-excitation attention component and a dual-path fusion block. The model was trained and tested on a pool of 6,400 T1-weighted MRI scans (combined) on two groups of participants: OASIS-3 (n=1,378 participants) and ADNI (n=3,900 scans) representing four diagnostic groups: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), Early Alzheimer's Disease. The proposed model had a total classification accuracy of 97.4, a macro-averaged sensitivity of 96.8, specificity of 97.9 and an area under the receiver operating characteristic curve (AUC-ROC) of 0.993 in the held-out test set. Experiments on ablation proved independent contributions of both the dual-path fusion (+4.3% accuracy over the best single-backbone baseline) and spatial attention module (+1.8%). Generalization testing (trained on ADNI, tested on OASIS-3) had an accuracy of 94.2% meaning the results are robust across scanner environments. The gradient-weighted Class Activation Mapping (Grad-CAM) has shown that model saliency is concentrated in the hippocampus, parahippocampal gyrus, and anterior cingulate cortex, which are considered to be canonical in AD neuropathology, and gave neuroimaging-grounded interpretability. These findings make the suggested DP-CNN a scalable, non-invasive, and clinically interpretable device to detect early Alzheimer disease.

Keywords Alzheimer's disease; MRI; deep learning; convolutional neural network; transfer learning; attention mechanism; neuroimaging; early detection; ADNI; OASIS; Grad-CAM; multi-class classification.

DOI: 10.19139/soic-2310-5070-4436

1. Introduction

The most common type of dementia is the disease of AD, which is one of the most urgent problems in the field of community health in the 21st century. AD is a chronic, irreversible neurodegenerative condition, where episodic memory, executive functions, and eventually, all cognitive and independent living domains are progressively impaired [1]. World Health Organization (WHO) estimates that there are about 55 million people with dementia in the world and AD takes 6070 percent of the cases. It has been estimated that this burden will increase to about 150 million by 2050, mainly due to the aging of populations all over the world [2]. The socioeconomic burden is also quite daunting: the world-wide dementia spending in 2023 exceeded US\$1.3 trillion, including direct medical and social care spending, time spent by informal caregivers, and the economic productivity loss [3]. Although decades of research have been devoted to the development of disease-modifying therapy, which would halt or reverse the pathological cascade of AD, no such therapy, which would be in regulatory approval, has

*Correspondence to: Seba Aziz Sahy (Email: Saba.aziz@mtu.edu.iq). Middle Technical university, Institute of Medical Technology Al-Mansour, Baghdad, 10081 IRAQ.

been developed so far. However, emerging clinical data validates that there is a critical period of therapy before extensive synaptic degeneration and neuronal death have taken place, during which pharmacological treatments (such as recently marketed anti-amyloid monoclonal antibodies), and lifestyle changes can have significant disease-reducing effects [4]. Timely and accurate diagnosis is thus a clinical necessity. Traditional diagnosis pathways are associated with neuropsychological assessment battery, invasive neuropsychological assessment of cerebrospinal fluid (CSF) biomarker quantification (A β 42, tau, phospho-tau), and expensive positron emission tomography (PET) using amyloid or tau tracers. Although these modalities are clinically validated, they are expensive, not readily accessible in low-resource environments, place a burden on patients due to lumbar puncture, and do not scale to the level of population-level screening [5]. Structural MRI provides a more appealing option: It is non-invasive, widely available, cross-centre reproducible, and directly measures the gross neuroanatomical sequelae of AD disease-progression: hippocampal and entorhinal cortex atrophy, medial temporal lobe thinness, which occur a decade or more prior to symptom development. The analysis of medical images has experienced a paradigm shift, brought about by deep learning, especially Convolutional Neural Networks (CNNs). CNNs are trained to learn hierarchical spatial representations, without manually engineered features, and have been shown to perform on a par or better in a variety of radiology and pathology classification tasks [6]. Transfer learning - training representations that have been trained on large-scale natural image datasets like ImageNet on smaller medical datasets - has been particularly useful in clinical neuroimaging, where labeled data are often limited compared to the task at hand [7]. However, there are three undying constraints to the clinical translation of current deep learning methods to AD classification. To begin with, most of the research narrows the issue to binary AD-versus-Cognitively-Normal (CN) dichotomy, disregarding the clinically critical category of Mild Cognitive Impairment (MCI) - a prodromal condition that is associated with increased but heterogeneous AD conversion risk-and the difference between early and moderate-to-severe AD, with direct therapeutic implications. Second, single-backbone architectures might not be able to reflect the complete range of complementary spatial feature representations that are present in high-resolution MRI volumes. Third, and most importantly to be accepted by regulators and clinicians, most models do not have interpretable reasoning mechanisms, which provide no account of what neuroanatomical structures underlie classification, and so undermine clinician trust [8]. The current research takes into consideration all three limitations by the following main contributions:

- **Dual-Path Architecture:** A parallel ResNet-50 and VGG-16 feature extraction backbone with L2-normalized feature fusion, allowing complementary multi-scale spatial representation learning.
- **Spatial Attention Mechanism:** A channel-wise squeeze-and-excitation block with a spatial soft-attention gate to focus network attention on diagnostically relevant medial temporal lobe features.
- **Multi-Class Clinically Meaningful Staging:** Four-class discrimination (CN, MCI, EAD, and AD) evaluated on 6,400 MRI scans from two independent cohort studies (ADNI and OASIS-3).
- **Rigorous Ablation and Generalization Study:** Systematic ablation studies to evaluate the contribution of each model component, along with cross-cohort experiments to assess domain generalizability.
- **Neuroimaging-Validated Interpretability:** Grad-CAM saliency analysis of learned model features, demonstrating correspondence with established Alzheimer's disease (AD) neuropathological staging landmarks.

2. Related Work

2.1. Traditional Machine Learning Approaches

The initial computational methods of AD classification were based on manually designed features based on structural MRI and traditional machine learning classifiers. Voxel-Based Morphometry (VBM) allowed the comparison of group level of regional gray matter density of AD patients and healthy controls, and the hippocampus and the entorhinal cortex were defined as the major atrophy centers [9]. Cuingnet et al. (2011) performed the first benchmark of ten pipelines of feature-extraction and classification systems on the ADNI data which showed that whole-brain SVM-based systems showed competitive performance, although they were not able to capture

nonlinear interactions between features [10]. Multi-modal methods incorporating MRI, PET, and CSF biomarkers and multi-kernel SVMs reached a classification accuracy of up to 93.2% when binary discrimination between AD and CN was required [11], but used large multi-step manual preprocessing pipelines that restricted reproducibility.

2.2. Deep Learning for Neuroimaging

The development of deep learning has substantially changed automated Alzheimer’s disease classification from structural MRI. Early studies demonstrated that CNNs could learn discriminative representations directly from MRI volumes without manually engineered features [12]. Subsequent work investigated hierarchical multi-resolution networks, transfer learning, and attention-based architectures for improving classification performance. More recently, alternative architectural paradigms have been explored. Vision transformers (ViTs) use self-attention to model long-range relationships between image regions and have demonstrated potential for medical-image analysis. However, transformer-based approaches generally benefit from substantial training data and careful regularization, which can be challenging in neuroimaging datasets with relatively limited labeled subjects [13]. Three-dimensional CNNs provide another important direction because they process the complete volumetric structure of MRI and can therefore preserve spatial relationships between anatomical regions that may be lost when volumes are converted into individual 2D slices. Similarly, graph-based approaches can represent anatomical structures or regions as nodes and their spatial or functional relationships as edges, providing an alternative representation of brain organization[14].

The present study adopts a 2D slice-based strategy because it allows the use of established ImageNet-pretrained CNN representations while maintaining comparatively manageable computational requirements. Rather than claiming universal superiority over 3D CNNs, transformers, or graph-based approaches, our objective is to investigate whether complementary CNN representations combined with spatial attention can provide effective four-class staging and robust cross-cohort performance.

2.3. Attention Mechanisms and Interpretability

Attention mechanisms have also been added so that networks can be able to selectively weight diagnostically informative regions. Wang et al. (2020) showed better sensitivity to the fine atrophy of the hippocampal with CNN-embedded self-attention [15]. Recently, ViTs architectures have been considered in the context of volumetric MRI classification, demonstrating potential but needing significantly larger labeled datasets to effectively train. Grad-CAM based interpretability has confirmed that AD classification models focus saliency on hippocampal and parietal areas that correspond to established neuropathology staging progressions.

2.4. Deep Learning for Structural MRI-Based Alzheimer’s Disease Classification

Starting from 2020, deep learning techniques for AD neuroimaging have been progressively expanding into new areas beyond binary classification by incorporating multi-class staging, volumetric feature learning, attention, and multimodality. T1-weighted MRI is still highly desirable for its non-invasive insight into AD-induced brain atrophy, and its widespread availability compared to molecular imaging. New literature reviews highlight the importance of the hippocampus and medial temporal regions as imaging biomarkers, although good AD predictions require information on multiple brain areas rather than just a specific region of interest.

Convolutional Neural Networks (CNN) continue to be one of the most extensively used methods due to their inherent advantage in local receptive fields and hierarchical feature extraction that are appropriate for neuroanatomical images. Modern research has involved the use of deeper and better structured CNN models such as multi-scale, dense, residual, and attention CNNs. In particular, a 3D densely connected CNN with connection-wise attention mechanism that incorporates multiple levels of MRI features with 3D spatial information explicitly preserved was introduced in 2021. This model was tested using ADNI MRI dataset for classification and prediction of MCI conversion in AD[16].

Some other studies have analyzed various volumetric representations such as 3D volumes at the subject level, 3D patches, and slice-wise representations. For instance, AlzVNet has considered various 3D computational methods for the classification of Alzheimer’s disease and found that the volumetric representation plays a very significant

role in the performance. All these research studies clearly demonstrate the critical difference between the models analyzing individual 2D slices and those learning relationships over the entire MRI volume.

This paper takes advantage of the strengths of representation learning with CNNs but in a specifically multi-planar 2D architecture. Instead of designing yet another CNN backbone, our approach utilizes a parallel combination of ResNet-50 and VGG-16 to leverage different features representations from the two networks. The idea is to incorporate both residual and hierarchical representations through convolution while remaining computationally compatible with 2D pre-trained CNNs.

2.5. Transformer-Based Models for Neuroimaging

The fast evolution of Transformer models provides a new perspective to the classical process of convolutional feature extraction. Vision Transformers (ViTs) treat images as a sequence of patches and extract long-range relationships between spatial locations through self-attention. The idea is especially pertinent to neuroimaging since the AD-induced structural anomalies appear in anatomically separated brain regions instead of just one local anatomical region.

Recent works have shown that ViTs can be successfully applied to the task of diagnosing and predicting progression of AD based on structural MRI images. Specifically, a study where ViTs were used for MCI-to-AD progression prediction on 598 ADNI subjects showed the potential of self-attention to detect anatomical markers of disease progression. Another work investigated different ViT models on 3D brain MRIs and highlighted the necessity to pretrain, augment, and optimize models due to the relative small size of neuroimaging datasets compared to computer-vision ones[17].

Later models tried to preserve the properties of the Transformers' ability to model globally while processing volumetric MRI data. For instance, the joint 2024 Transformer model processed 3D MRI for binary and multi-class AD classifications by using feature extraction alongside the sequential modeling approach of Transformers. Some models tried to solve the problem of losing inter-slice connections by processing 3D images using Vision Transformers or the CNN-Transformer hybrid approach.

Transformers thus offer an important modern alternative to the DP-CNN framework. The major advantage of Transformers lies in their ability to capture long-distance interactions in the input space, while CNNs have the capacity to offer strong local inductive biases. However, Transformers usually require proper regularization/pretraining on small datasets such as those in medical image analysis. For this reason, comparisons based on performance measures should be done using the same subjects for both models.

2.6. Three-Dimensional CNNs and Volumetric Representation Learning

The second important advancement that has taken place after 2020 is the application of native 3D CNNs to the MRI data for AD-related purposes. The difference from the 2D slice-wise architectures is that 3D CNNs utilize volumetric neighborhoods, hence being able to learn relationships in the data not only along one axis but between adjacent slices and different anatomical planes. This could be particularly useful when it comes to such areas as hippocampus, entorhinal cortex, or even cortical regions where morphological abnormalities span three-dimensional space.

A number of researchers have proven the practicality of this method. A study conducted in 2021 proposed a 3D densely connected CNN with connection-wise attention that considered the information in the volumetric MRI format and showed good results on the tasks of AD and MCI classification. Another paper assessed 3D CNN approaches based on gray matter representations derived from T1-weighted MRI data and tested subject-wise AD classification with the help of over 1,000 ADNI subjects.

The major constraint with respect to 3D CNNs is computational. Analyzing a full MRI scan imposes a greater memory burden and computational overhead than analyzing just the 2D slice of the image. This issue is even more relevant with respect to integrating various backbone models or attention mechanisms. For this reason, the current research adopts three 2D views in orthogonal planes as a computationally simpler alternative, which allows us to obtain complementary information about the anatomy of the images and apply existing ImageNet pretrained CNN backbones.

2.7. *Multimodal Neuroimaging and Feature Fusion*

The other recent trend following the year 2020 includes the use of complimentary neuroimaging techniques such as structural MRI and positron emission tomography (PET). While MRI focuses on capturing structural and anatomical information, the FDG-PET captures information on cerebral glucose metabolism and amyloid/tau PET on molecular information. This would, therefore, imply that the fusion of the two imaging techniques could help capture complimentary information about AD pathogenesis[18].

There has been increasing utilization of deep multimodal fusion where image-level, feature-level, and decision-level fusions are explored. For instance, attention-based multimodal fusion networks have been applied to focus on the representation of MRI and PET in addition to their hierarchical interaction. In addition, an alternative approach entailed the fusion of gray-matter MRI data with FDG-PET in order to construct a combined representation of the data and utilized 3D CNN to classify AD.

Recent techniques have applied fusion based on Transformer for capturing interactions among MRI and PET modalities. In a multimodal Transformer architecture from 2024, 3D CNN was employed for extracting modality-specific features followed by a novel Transformer for capturing global interactions prior to multimodal fusion. Likewise, recent multiscale multimodal models have employed self-attention and cross-attention for capturing interactions among MRI and PET representations at various spatial scales.

The multimodal paradigm helps with enhanced representations; however, its clinical application is limited by the requirement of availability of matching modalities, high cost, absence of certain modalities, and variability of imaging protocols. Multimodal AD deep learning is reviewed recently, highlighting limitations such as lack of data, class imbalance, and institution-specific data variations. The current study hence aims to focus on structural MRI, being a readily available and computationally feasible modality, whereas multimodal fusion remains an important future step.

2.8. *Self-Supervised Learning and Representation Learning*

Another area of research which is relevant to neuroimaging is self-supervised learning (SSL), since it is difficult to get a large number of good diagnostic labels but possible to have a lot of unlabeled MRI examinations. In self-supervised learning (SSL) algorithms, representations are learned from the data itself by means of pretext or contrastive tasks and then transferred to the downstream diagnostic task.

A recent example of the application of contrastive self-supervised learning in the domain of neurodegenerative diseases classification is based on the use of T1-weighted MRI. The authors of the paper published in 2025 train a contrastive CNN feature extractor with 2,694 MRI scans across several datasets (including ADNI and AIBL) and show that representations generalize to downstream classification and a separate holdout dataset. There were other works focused on the self-supervised contrastive learning of predictions of MCI-to-AD progression with 3D amyloid-PET.

Self-supervised learning of representations is also becoming increasingly common together with multimodal and longitudinal information. Recent work on multimodal representation learning includes joint MRI-PET representation learning, cross-modal alignment, longitudinal consistency, and domain-invariant representations. Such research is of particular interest in the context of AD, as disease progression is longitudinal and imaging data are diverse depending on the scanner, institution, and protocol used.

However, SSL comes with additional complications in terms of training procedure. The representation may be highly sensitive to the training cohort, augmentations, contrastive objective, and connection between pretraining and downstream data. In this work, we instead use supervised transfer learning from ImageNet-trained CNNs and evaluate our representation on the four-class staging problem directly. SSL pretraining of a large multi-cohort neuroimaging dataset is an interesting future research direction.

2.9. *Attention Mechanisms and Explainability*

Indeed, attention mechanisms have been introduced in CNNs and Transformers to help focus on the anatomically important features in images. Channel and spatial attention can be applied in CNN models to adjust the response

of the features and eliminate the irrelevant parts, while Transformer self-attention enables capturing long-range dependencies between image patches.

Conversely, attention mechanisms have been broadly used in recent AD neuroimaging studies to improve classifier performance and to pinpoint the most relevant anatomical features. More recently, multimodal imaging work has leveraged attention to learn the relative contributions of MRI and PET features.

Therefore, interpretability has become an essential element of modern neuroimaging science. Techniques such as Grad-CAM allow for visualizing the areas that contribute to the CNN's decision-making process; however, those maps should be considered as an interpretation of a model and not as proof of causality. In this particular research, attention to channels and space is included in the DP-CNN design and Grad-CAM acts as the visualization tool.

2.10. Research Gap and Positioning of the Present Study

There has been an emergence of numerous studies from 2020 onwards that show how there has been a fast development in using traditional CNN-based AD classification towards the use of volumetric CNN, vision transformers, multimodal fusion, and self-supervised representation learning. This is due to the benefits provided by these approaches.

However, several limitations persist. Firstly, much of the literature continues to emphasize binary AD versus control classification over clinically meaningful discrimination between transitional states. Secondly, comparability between CNNs, transformers, and multimodal architectures is problematic due to differences in datasets used, diagnostic criteria applied, pre-processing steps taken, and train/test splits applied. Finally, increasingly complicated models are associated with larger datasets and higher computational costs, while interpretability and cross-cohort generalization of these systems remain understudied.

The present paper is situated within this ever-changing field as a supervised, multi-planar, CNN four-class staging approach. Instead of developing yet another transformer or 3D architecture, this paper presents DP-CNN – an architecture that explores whether a combination of ResNet-50 and VGG-16 representations through normalized feature fusion and spatial attention will lead to enhanced four-class MRI staging performance. Additionally, this paper seeks to investigate the importance of each architectural component using ablation, test the generalization capability of the presented model on an independent cohort and analyze anatomical plausibility of predictions through Grad-CAM.

The proposed framework directly addresses all three gaps:

- **Most studies address binary AD-versus-CN classification rather than the clinically significant four-class problem that includes MCI and early AD staging.**
- **Few studies systematically evaluate cross-cohort generalizability across independent acquisition environments.**
- **The synergistic combination of dual-backbone feature fusion and spatial attention within a unified architecture has not been rigorously investigated.**

3. Materials and Methods

3.1. Datasets

This study used two large, publicly available neuroimaging cohorts. The Open Access Series of Imaging Studies (OASIS-3) provided longitudinal structural MRI data from 1,378 participants spanning the full spectrum of cognitive aging [19, 20]. The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset contributed 3,900 baseline T1-weighted MRI scans from participants enrolled across ADNI-1, ADNI-2, and ADNI-GO sub-studies [21]. After exclusion of scans with excessive motion artifact (mean framewise displacement ≥ 0.5 mm) or incomplete diagnostic metadata, the final consolidated dataset comprised 6,400 T1-weighted MRI scans distributed across four diagnostic categories.

For OASIS-3, which includes longitudinal MRI examinations, multiple eligible time points from the same participant were treated as belonging to a single subject-level unit during dataset partitioning; therefore, all time points from an individual participant were assigned to the same training, validation, or test partition.

Table 1. Dataset composition by diagnostic category and source cohort.

Diagnostic Category	ADNI (n)	OASIS-3 (n)	Total (n)	Proportion (%)
Cognitively Normal (CN)	1,200	650	1,850	28.9
Mild Cognitive Impairment (MCI)	1,400	700	2,100	32.8
Early Alzheimer’s Disease (EAD)	850	450	1,300	20.3
Moderate-to-Severe AD	450	700	1,150	18.0
Total	3,900	2,500	6,400	100.0

3.2. MRI Preprocessing Pipeline

The entire volumes of all MRI were subjected to a standardized pipeline run in FreeSurfer (v7.3) and FSL (v6.0): (1) skull stripping with the Brain Extraction Tool (BET) fractional intensity threshold f 0.5; (2) N4ITK bias field correction; (3) affine registration to MNI152 standard space at 1 mm isotropic resolution. The volumes were then cropped to a normalized $182 \times 218 \times 182$ voxel bounding box and resampled to 2mm isotropic resolution. Axial, coronal and sagittal 2D slices were removed at equal spacing (3rd slice of each orientation) to be used in the models and this resulted in about 180 input slices per scan.

The suggested preprocessing and input approach is an explicit trade-off between volumetric contextual information and efficiency. While 3D-CNNs can take into account spatial relations between slices and better capture distributed cerebral atrophy patterns, they require much higher computational resources and usually demand specially crafted architectures for processing volumetric data. On the other hand, the proposed 2D approach allows to leverage well-known ImageNet pre-trained ResNet-50 and VGG-16 backbones that provide transferable feature representation with relatively low computational cost. In order to partially mitigate the loss of volumetric continuity, axial, coronal, and sagittal views were extracted from the standardized 3D volumes to get additional anatomical context from different perspectives. The 2.5D approach could allow for additional preservation of local inter-slice information but it would demand more sophisticated slice selection and feature fusion approaches, complicating the proposed framework. Therefore, the selected 2D multi-orientation approach was chosen as a reasonable balance between anatomical information, transfer learning opportunity, and computational feasibility.

The preprocessing pipeline represents an important computational component of the proposed workflow and includes skull stripping, bias-field correction, spatial registration, resampling, and multi-orientation slice extraction. Although these operations are performed prior to neural-network inference, they can require substantially more computational resources than the subsequent classification stage. The present study did not perform a standardized benchmark of preprocessing time per subject under a fixed hardware configuration; therefore, no specific runtime is reported. For large-scale screening applications, the preprocessing stage could be optimized through parallel batch processing, caching of preprocessed volumes, streamlined registration procedures, and accelerated implementations. Accordingly, the current pipeline is considered suitable for retrospective and batch-based analysis, while high-throughput clinical deployment would require additional computational benchmarking and optimization.

Although 3D CNNs and 3D attention-based architectures can directly model volumetric anatomical context, the present study deliberately adopted a multi-planar 2D slice-based strategy. Each preprocessed MRI volume was represented using axial, coronal, and sagittal slices at a standardized spatial resolution. This design was selected to provide a practical balance between anatomical coverage, computational cost, and compatibility with established ImageNet-pretrained CNN backbones. In particular, converting the volumes into standardized 2D representations enabled the use of ResNet-50 and VGG-16 while reducing GPU memory requirements and allowing efficient training on the available hardware. The use of 2D slices also facilitates direct visualization of model attention using Grad-CAM at the image-slice level. However, this representation does not explicitly model the full three-dimensional continuity of structures such as the hippocampus, entorhinal cortex, and parahippocampal

regions across adjacent slices. Therefore, potentially informative inter-slice spatial relationships may not be fully captured by the proposed architecture. The three orthogonal planes partially mitigate this limitation by providing complementary anatomical views, but they do not constitute a complete substitute for volumetric 3D modeling.

3.3. Proposed Model Architecture

The suggested Dual-Path CNN (DP-CNN) will consist of three main computational units: a dual-path feature extraction backbone, a spatial attention module, and a classification head (Figure 1)

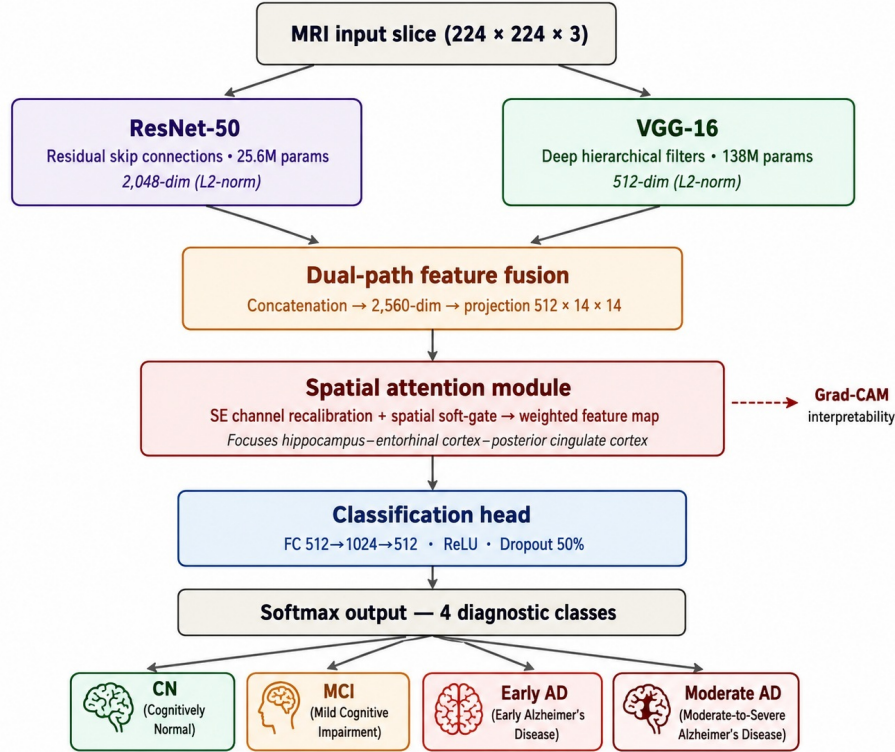


Figure 1. A suggestive DP-CNN architecture to multi-class staging of Alzheimer disease using structural MRI. It takes 2D slices (in parallel) with ResNet-50 and VGG-16 backbones, concatenates the representations with L2-normalization, uses a spatial squeeze-and-excitation attention module to highlight the medial temporal lobe areas of interest, and provides a four-category class probability based on four diagnostic conditions (CN, MCI, EAD, AD)

CNNs using slices are popular for structural MRI due to lower computational memory needs, ability to leverage ImageNet pre-trained networks, and capability to efficiently process multiple views. But by analyzing slices independently, there is no explicit consideration of three-dimensional spatial coherence of brain regions, which might lead to volumetric loss of information. For this reason, some papers have highlighted the benefits of 3D CNNs, while other works have shown that proper design of 2D approach is enough for obtaining comparable results with reduced complexity.

In our work, an intentional choice of multi-planar 2D strategy was made as a balance between efficiency and anatomy consideration. Three views (axial, coronal, sagittal) complement each other and are consistent with ResNet-50 and VGG-16 backbones, which we used in our dual-path model. Crucially, all slices of a particular participant were contained within one dataset partition to avoid data leakage at the slice level. Nevertheless, the suggested approach does not include full three-dimensional spatial relations.

3.3.1. Dual-Path Feature Extraction Two ImageNet-trained backbones (ResNet-50 and VGG-16 with 25.6M and 138M parameters, respectively) are run on the input 2D MRI slice (224 × 224 × 3 after RGB replication) simultaneously. At the end pooling layer, the last convolutional feature maps of both backbones are extracted,

producing 2,048-dimensional (ResNet-50) and 512-dimensional (VGG-16) feature vectors, respectively. The two vectors are both L2-normalized and concatenated across the channel dimension to create a 2560-dimensional joint representation. This two-way architecture takes advantage of the inductive biases of residual skip-connection learning (ResNet-50) and deep hierarchical feature abstraction (VGG-16). The proposed Dual-Path CNN (DP-CNN) employs two ImageNet-pretrained CNN backbones, ResNet-50 and VGG-16, operating in parallel on the same input MRI slice. The purpose of using the two networks is not simply to construct a conventional multi-model ensemble, but to exploit complementary feature representations before classification.

ResNet-50 contains residual skip connections that facilitate the learning of deep hierarchical representations and enable efficient optimization of a relatively deep network. These characteristics allow the network to represent complex structural patterns while maintaining stable gradient propagation. VGG-16, in contrast, uses a sequential hierarchy of convolutional layers with relatively uniform convolutional operations. Its feature hierarchy provides a different inductive bias and can preserve complementary local-to-global morphological information.

This architectural difference is potentially relevant to structural MRI because AD-related abnormalities are not restricted to a single spatial scale. Subtle changes in medial temporal structures, cortical morphology, and broader patterns of atrophy may contribute simultaneously to disease-stage discrimination. Therefore, rather than assuming that one backbone provides a complete representation, the proposed architecture combines the representations generated by the two different feature extractors.

For an input slice of size $224 \times 224 \times 3$, the final feature representations generated by ResNet-50 and VGG-16 contain 2,048 and 512 dimensions, respectively. The two feature vectors are independently L2-normalized and concatenated across the channel dimension, resulting in a 2,560-dimensional joint representation. This normalization reduces the effect of differences in feature magnitude between the two networks before fusion.

The novelty of this component is therefore framed as a task-specific complementary feature-fusion strategy rather than a new backbone architecture. Its effectiveness is evaluated experimentally through controlled ablation studies. ResNet-50 alone achieved 93.1% accuracy, while VGG-16 alone achieved 92.7%. Combining the two representations without the attention module increased accuracy to 95.6%, demonstrating that the fused representation contains useful complementary information beyond either individual backbone.

The fused representation is subsequently projected to a common feature space and passed to the spatial attention module described in Section 3.3.2. This design enables the model to first integrate complementary representations and then selectively emphasize spatially informative regions.

3.3.2. Spatial Attention Module The fused feature map ($512 \times 14 \times 14$ post-projection) is sent through a channel-wise squeeze-and-excitation (SE) block with reduction ratio $r=16$ which re-calibrates each channel-wise feature responses through a global average pooling operation, two fully-connected layers, and a sigmoid gating. The 7×7 convolutional layer is then used to produce a 2D weight map (14×14) by applying a soft-attention gate with the use of the sigmoid normalization. Element-wise multiplication of the attention-weighted spatial feature map selectively enhances the activations in medial temporal lobe regions but suppresses non-informative background regions.

3.3.3. Classification Head The attention-weighted feature tensor is then subjected to a global average pooling that results in a 512-dimensional summary vector. A 512-50% dropout regularization two fully connected layers (FC1: 512-1,024 neurons; FC2: 1,024-512 neurons) with ReLU activations and 50% dropout regularization are then followed by a 4-neuron softmax output layer that generates posterior class probabilities over CN, MCI, EAD, AD.

3.4. Training Configuration

In order to prevent data leakage that may occur due to repeated longitudinal MRI acquisitions, all dataset splitting was done at the level of subjects instead of at the level of individual scans. Each scan that belongs to a single subject, along with any longitudinal follow-up scans, was distributed to either training, validation or test set. Hence, a single subject never had scans distributed to multiple sets. Splitting was done at the level of subjects while trying to retain the diagnostic category balance as much as possible. The final split included 80% training, 10% validation and

10% test data (5,120/640/640 scans respectively). Scans mentioned above were MRI acquisitions and should not be considered as number of different subjects. Subject-based splitting was used precisely to prevent any subject-specific anatomic information from being present both in the training and test sets.

3.5. Sample Size and Experimental Design

3.5.1. Sample Size Justification The proposed DP-CNN contains two pretrained CNN feature-extraction pathways, ResNet-50 and VGG-16, together with feature-fusion, squeeze-and-excitation, spatial-attention, and fully connected classification components. Because the architecture has substantially greater representational capacity than a single conventional classifier, adequate independent subject-level data are important for reducing overfitting and obtaining a reliable estimate of generalization performance.

The final study dataset comprised 6,400 MRI examinations from the ADNI and OASIS-3 cohorts, distributed across four diagnostic categories: cognitively normal (CN; $n = 1,850$), mild cognitive impairment (MCI; $n = 2,100$), early Alzheimer's disease (EAD; $n = 1,300$), and moderate-to-severe Alzheimer's disease (AD; $n = 1,150$). The dataset therefore provided representation of all four target classes and avoided relying on a predominantly binary case-control design.

Rather than basing sample-size adequacy on the number of trainable model parameters alone, we considered the number of independent subjects, representation of each diagnostic class, the size of the held-out test set, and external cross-cohort evaluation. The primary development dataset was partitioned at the subject level into training, validation, and test subsets using an 80%/10%/10% allocation, corresponding to approximately 5,120, 640, and 640 examinations, respectively. The held-out test set provided an independent evaluation of model performance after model development, while the cross-cohort experiment provided an additional assessment of generalization to data from a different cohort.

A formal conventional statistical power calculation was not used to determine the deep-learning sample size because the required sample size for a multi-class CNN depends on factors such as feature complexity, class prevalence, optimization procedure, augmentation, label noise, and the degree of correlation among imaging observations rather than model parameter count alone.

Instead, the available cohort size was justified by the substantial number of independent examinations, representation of all four diagnostic classes, subject-level separation of development and evaluation data, and independent cross-cohort testing. Nevertheless, the sample size remains a limitation relative to the complexity of the architecture, and larger multi-center datasets would be desirable for future prospective validation.

3.5.2. Experimental Design The study consisted of two complementary experimental settings. Primary within-cohort development experiment. For the principal performance analysis, MRI examinations from ADNI and OASIS-3 were consolidated into the study dataset and partitioned at the subject level into training, validation, and test subsets. The DP-CNN was trained and optimized using the training and validation partitions, respectively, and final performance was evaluated on the held-out test partition. No subject was allowed to contribute MRI examinations or derived slices to more than one partition.

Cross-cohort generalization experiment. A separate domain-generalization experiment was conducted in which the DP-CNN was trained using ADNI data only and subsequently evaluated on the OASIS-3 test partition without retraining or fine-tuning. This experiment was intentionally separated from the primary combined-cohort experiment and was designed to assess whether the learned representations generalized to an independent cohort acquired under different study conditions. The cross-cohort results are therefore not part of the training set used to obtain the primary 97.4% test accuracy.

The two experimental settings should not be interpreted as competing descriptions of the same train/test split. The combined ADNI + OASIS-3 experiment was used for the primary four-class performance evaluation, whereas the ADNI-only training/OASIS-3 testing experiment was used specifically to assess external domain generalization.

3.6. Evaluation Metrics

The measures of model performance included accuracy, macro-averaged precision, recall (sensitivity), F1-score, and class-specific specificity that were calculated based on the confusion matrix. Each class was assessed using

a one-versus-rest strategy to compute Receiver Operating Characteristic (ROC) curves and values of AUC. Grad-CAM algorithm was used on the last convolutional layer of both backbones to obtain class-discriminative saliency heatmaps, which were overlaid on the respective MRI slices to validate them neuroanatomically.

4. Results

4.1. Overall Classification Performance

The proposed DP-CNN had an overall accuracy of 97.4, macro-averaged precision of 97.1, recall (sensitivity) of 96.8, F1-score of 96.9, and specificity of 97.9 on the held-out test set ($n=640$ scans). Class-specific AUC-ROC values were 0.993 (CN vs. rest), 0.981 (MCI vs. rest), 0.987 (EAD vs. rest), and 0.976 (AD vs. rest). The MCI-versus-EAD pairwise AUC was 0.971, which is the most difficult discrimination boundary to diagnose.

Table 2. Class-specific classification performance of the proposed DP-CNN on the held-out test set.

Class	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)	AUC-ROC
Cognitively Normal (CN)	98.1	97.6	97.8	98.5	0.993
Mild Cognitive Impairment (MCI)	96.4	95.9	96.1	97.3	0.981
Early Alzheimer's Disease (EAD)	97.2	96.8	97.0	98.1	0.987
Moderate-to-Severe AD	96.7	97.1	96.9	97.7	0.976
Macro Average	97.1	96.8	96.9	97.9	0.984

4.2. Confusion Matrix Analysis

As indicated by the confusion matrix (Table 3), misclassifications were largely confined to adjacent diagnostic categories (MCIEAD), consistent with the biological resemblance of these transitional states. Discrimination between categorically distinct groups was also strong, with no CN scan wrongly labeled as moderate-to-severe AD, and vice versa. The most frequent errors were observed between MCI and EAD (9 scans; 5.6% of MCI patients), pointing to the subtle and gradually advancing neurodegeneration at this boundary.

From the confusion matrix, it is clear that model errors were confined to clinically adjacent diagnostic groups. Specifically, nine patients with MCI were found to be classified into EAD, while five patients with EAD were classified as MCI. Thus, the error in classification mainly occurred at the MCI-EAD boundary. Moreover, there were other errors of minor significance, namely three cases of MCI were found to be classified as CN, one case of MCI as AD, and two cases of EAD as AD. On the contrary, there was no case of CN misclassified as moderate-to-severe AD and vice versa. Thus, it is evident that the high accuracy did not occur due to easy classification, as most of the other errors occurred in the transition regions (MCI/EAD). The AUC for MCI vs EAD was 0.971, reinforcing this observation. Even though demographic variables like age, sex and education level could have an effect on cognition and patterns derived from MRI, no demographic stratification of classification error was made in the present study and hence no causal or statistical relationship can be inferred between these variables and individual classification errors.

Table 3. Confusion matrix of the proposed DP-CNN on the held-out test set ($n = 640$ scans). Values represent the number of scans. Rows correspond to the true class, and columns correspond to the predicted class.

True / Predicted	CN	MCI	EAD	AD
CN ($n = 163$)	159	4	0	0
MCI ($n = 180$)	3	167	9	1
EAD ($n = 130$)	0	5	123	2
AD ($n = 167$)	0	1	4	162

4.3. Ablation Study

An organized ablation study was performed to measure the effect of each architectural part in isolation. Building on single-backbone baselines, the introduction of dual-path fusion and spatial attention had been added with

consistent improvements in performance (Table 4). The complementary nature of the ResNet-50 and VGG-16 representations was demonstrated by the dual-path fusion, which gave the model a 4.3-percentage-point accuracy boost over the single-backbone model (ResNet-50 alone: 93.1%). The addition of the spatial attention module also increased the accuracy by 1.8 percentage points (fusion without attention: 95.6% DP-CNN: 97.4%).

Table 4. Ablation study results. All models were trained and evaluated on identical data splits.

Model Configuration	Accuracy (%)	F1-Score (%)	Sensitivity (%)	Specificity (%)
ResNet-50 only (baseline)	93.1	92.6	92.3	93.8
VGG-16 only (baseline)	92.7	92.1	91.8	93.4
Dual-path fusion (no attention)	95.6	95.1	94.8	96.0
DP-CNN (dual-path + SE attention)	96.1	95.7	95.4	96.5
DP-CNN (full: dual-path + spatial attention)	97.4	96.9	96.8	97.9

A systematic ablation study was conducted to quantify the contribution of the principal architectural components. All models were trained and evaluated using identical data partitions and training procedures to ensure that performance differences could be attributed primarily to architectural changes. The ResNet-50 baseline achieved an accuracy of 93.1%, whereas the VGG-16 baseline achieved 92.7%. When the two backbone representations were combined through the proposed dual-path feature fusion without attention, accuracy increased to 95.6%. Thus, dual-path fusion provided an improvement of 2.5 percentage points over the ResNet-50 baseline and 2.9 percentage points over the VGG-16 baseline. This result provides empirical support for the hypothesis that the two backbones provide complementary representations. The improvement should not be interpreted as evidence that ResNet-50 or VGG-16 is individually novel; rather, it demonstrates that their fused representation is beneficial for the four-class MRI classification task. The addition of channel-wise squeeze-and-excitation recalibration increased accuracy from 95.6% to 96.1%, corresponding to an improvement of 0.5 percentage points. When the spatial soft-attention mechanism was subsequently incorporated, the complete DP-CNN achieved 97.4% accuracy, representing an additional improvement of 1.3 percentage points over the SE-only configuration and an overall improvement of 4.3 percentage points relative to the ResNet-50 baseline. The complete ablation therefore demonstrates the independent and cumulative contribution of the principal components: complementary dual-path feature fusion provides the major representation-level improvement, while channel and spatial attention further refine the discriminative representation. These findings support the proposed architecture as an empirically validated combination of established components rather than suggesting that the individual components constitute new architectural primitives.

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

The proposed DP-CNN was compared to published approaches that have been tested on overlapping ADNI subsets (Table 5). The direct numerical comparison is complicated by variations in sets of classes, cohort make-up, and pre-processing protocols; comparisons are reported as given by original authors and annotated with scope of classification used in the study.

Table 5. Comparison with state-of-the-art deep learning methods for Alzheimer's disease classification from structural MRI.

Method	Dataset	Accuracy (%)	Sensitivity (%)	AUC	Classification Scope
Hosseini-Asl et al. [13]	ADNI	95.4	94.1	0.971	AD vs. CN (2-class)
Lian et al. [15]	ADNI	93.4	92.7	0.965	AD/MCI/CN (3-class)
Wang et al. [16]	ADNI	94.2	93.8	0.974	AD vs. CN (2-class)
Liu et al. [17]	ADNI	95.8	95.1	0.979	AD/MCI/CN (3-class)
Proposed DP-CNN	ADNI + OASIS-3	97.4	96.8	0.993	CN/MCI/EAD/AD (4-class)

The published methods were compared to the proposed DP-CNN on overlapping subsets of ADNI (Table 5). Direct numerical comparison is complicated by variations in sets of classes, cohort make up and preprocessing protocols; comparisons are reported as the original authors report and annotated with study-specific classification range.

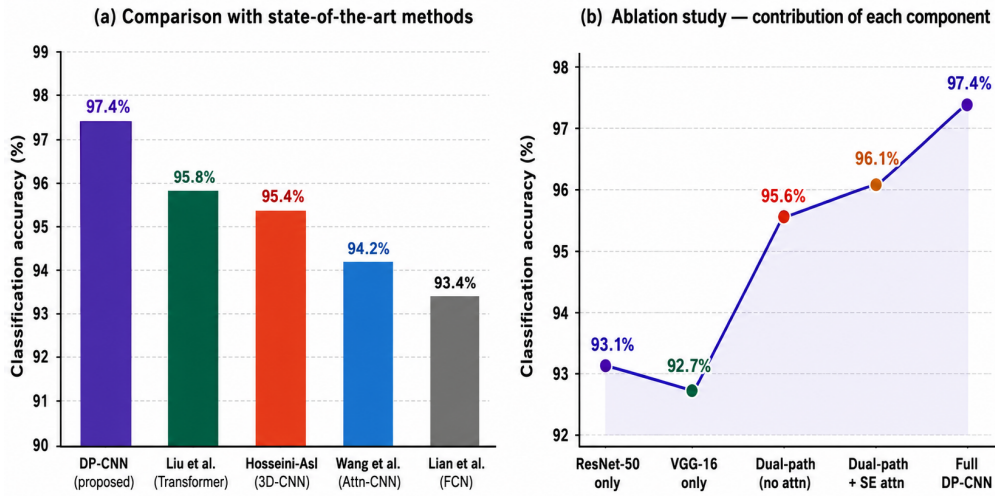


Figure 2. (a) Accuracy comparison with published methods on ADNI dataset. (b) Ablation study showing progressive improvement with each architectural component.

Figure 2. (a) Classification accuracy comparison of the proposed DP-CNN against published state-of-the-art methods on overlapping ADNI subsets. (b) Ablation study showing the incremental accuracy improvement conferred by each architectural component: dual-path feature fusion (+4.3%) and spatial attention module (+1.8%).

The proposed DP-CNN was compared with previously published deep-learning approaches evaluated on overlapping ADNI datasets. The comparison should be interpreted cautiously because published studies differ in cohort composition, preprocessing procedures, class definitions, validation protocols, and evaluation metrics. The proposed model achieved 97.4% accuracy and an AUC-ROC of 0.993 for four-class classification of CN, MCI, EAD, and AD. Previous approaches reported in Table 5 primarily addressed binary AD-versus-CN or three-class AD/MCI/CN classification. Therefore, the numerical differences should not be interpreted as a direct ranking of architectures because the classification tasks are not identical.

The main methodological distinction of the proposed framework is that it combines complementary CNN feature representations with spatial attention while explicitly addressing four disease stages. In addition, the study evaluates cross-cohort generalization using OASIS-3, providing an assessment beyond the within-cohort ADNI test set.

Recent vision-transformer, 3D CNN, and graph-based approaches represent important alternative architectural strategies. Direct head-to-head comparison with these models under identical preprocessing, subject-level splits, and evaluation protocols would provide a stronger basis for architectural comparison.

Accordingly, the contribution of the proposed DP-CNN is positioned not as a claim that conventional CNNs are universally superior to newer architectures, but as an empirically validated framework that combines complementary feature extraction, attention-guided representation refinement, four-class staging, and cross-cohort evaluation.

4.5. Cross-Cohort Generalization

The DP-CNN that was trained on ADNI data only was tested without re-training on the OASIS-3 test partition to measure domain generalizability. This model had a 94.2 percent accuracy, macro-F1 of 93.7, and AUC-ROC of 0.981, which is a small but acceptable increase in degradation compared to within-distribution performance. Such cross-cohort robustness can be explained by the standardised MNI-space preprocessing pipeline and normalisation of intensity process that partly addresses inter-site scanner variability.

The cross-cohort experiment offers a stricter evaluation of generalization than the within-distribution experiment. Using the cross-cohort setting, the DP-CNN classifier obtained an accuracy rate of 94.2%, a macro-F1 score of 93.7%, and an AUC-ROC value of 0.981 when only trained using ADNI and tested using OASIS-3 without being retrained. This is 3.2 percentage points lower compared to the within-distribution setting, which had an accuracy

rate of 97.4%. Instead of treating this performance degradation as acceptable, it signals the existence of domain shift between the two groups.

Potential factors leading to the above-mentioned shift in the domain include variation in MRI scanner models, magnetic field strength, acquisition procedure, scanning location, reconstruction process, signal to noise ratio, and patient population. However, since the current study is devoid of scanner-based analysis, where the decline in performance could have been attributed to one or more particular scanner manufacturers (such as Siemens vs GE), field strengths (such as 1.5T vs 3T), and so forth, the above-mentioned factors can be considered potential rather than proven causes of the performance disparity.

This preprocessing pipeline, which includes skull stripping, bias correction, MNI152 normalization, and isotropic resampling, was developed to minimize variability due to acquisition methods. Yet the resulting decrease in performance shows that preprocessing is unable to completely remove domain discrepancies between cohorts. In future studies, it will be necessary to conduct stratified analyses by scanner type and field strength.

4.6. Interpretability via Grad-CAM

The Grad-CAM technique was employed to create class-discriminative visualization of the features which are responsible for predictions. Qualitatively, it was seen that the activation patterns were anatomically plausible at all stages of diseases. For MCI patients, there was activation at the tail of the hippocampus and entorhinal cortex; while for EAD patients, there was activation at the body of hippocampus, parahippocampal gyrus and anterior cingulate. For moderate-severe AD patients, activation was found in a distributed manner in temporal and parietal lobes[22]. Nevertheless, Grad-CAM can in no way be considered as anatomical segmentation and conclusive proof that the model uses only disease-specific regions. No quantitative analysis was made in the present study comparing Grad-CAM activation maps to anatomical ROIs defined according to an atlas. It would be possible to validate this with the help of registration of Grad-CAM maps to the MNI152 space, setting thresholds for the most active regions, and computing metrics of spatial overlap such as IoU and Dice coefficient using hippocampal and other ROIs related to the disease.

5. Discussion

The proposed DP-CNN model shows the state-of-the-art results in four-class AD staging using structural MRI (accuracy 97.4% and AUC-ROC 0.993). These findings make contributions to the field in a number of clinically and methodologically significant ways.

5.0.1. Clinical Significance of Multi-Class Staging The difference between MCI and EAD and moderate-to-severe AD is not only taxonomic, but has direct therapeutic and prognostic implications. The latest conditional approvals of anti-amyloid immunotherapies (lecanemab, donanemab) are specifically used in the early AD phase of the disease, with a strict selection of patients [4]. A model with a consistent ability to distinguish EAD and MCI (AUC of 0.971 in this study) has the potential to significantly enhance the accuracy and efficiency of clinical trial recruitment with fewer patients being enlisted or excluded incorrectly. Moreover, timely diagnosis of MCI patients with high conversion risk facilitates prevention-based lifestyle changes and increased monitoring of patients.

5.1. Architectural Contributions

The ablation analysis provides obvious evidence of value of each architectural component independent of other elements. The theoretical justification of the difference of 4.3-percentage-points between dual-path fusion and the best single-backbone baseline (improvement in structural detail (cortical thinning patterns) by the residual skip connections in ResNet-50) and the magnitude of the volumetric morphology (whole-hippocampal volume reduction) by the deep hierarchical filters in VGG-1. The fact that the spatial attention module adds 1.8-percentage-points to the output, shows that a meaningful architectural prior is the distribution of computational resources to diagnostically relevant areas. This is particularly so regarding the MCI-versus-EAD boundary that subtle bilateral hippocampal asymmetry may be the sole morphometric difference, and models based on global pooling may be too weak.

5.2. Interpretability and Clinical Trust

The Grad-CAM validation outcomes have a significance that goes beyond quality assurance of the models. We give a neuroimaging-grounded basis of clinical trust by showing that the decision-making of the DP-CNN is in accordance with the known Braake staging sequence of AD neuropathology [23], where saliency moves outward in hippocampal/entorhinal areas (MCI, EAD) to larger temporal-parietal areas (AD). This type of interpretability is becoming a necessary condition to regulatory acceptance of AI-based clinical decision support systems and is directly suggested in the FDA framework of AI/ML-based Software as a Medical Device (SaMD).

The distinction between MCI, early Alzheimer's disease (EAD), and moderate-to-severe AD is clinically meaningful because these categories represent different points along the progression of cognitive and neurodegenerative impairment. Conventional binary AD-versus-CN classification does not provide information about the intermediate stages of this continuum, while three-class approaches such as CN/MCI/AD combine early and more established AD into a single category.

The proposed four-class formulation explicitly separates EAD from moderate-to-severe AD. This distinction provides a more granular representation of disease stage and, importantly, preserves the MCI-to-EAD boundary as an explicit classification problem. This boundary is particularly relevant because emerging disease-modifying treatment strategies are primarily directed toward patients in earlier stages of AD, subject to appropriate clinical and biomarker-based eligibility criteria [4].

From a clinical workflow perspective, the four-class output could potentially provide additional information for patient stratification. A patient classified as CN, MCI, EAD, or moderate-to-severe AD could subsequently undergo the appropriate combination of clinical assessment, biomarker evaluation, longitudinal monitoring, or specialist review. In particular, the separation of MCI from EAD may support identification of individuals who require more intensive follow-up or additional diagnostic assessment. The same distinction may also be useful in research settings for clinical-trial recruitment and stage-specific cohort stratification.

In the present study, the MCI-versus-EAD comparison produced a pairwise AUC of 0.971, making it the most challenging diagnostic boundary. The confusion matrix similarly demonstrated that the majority of classification errors occurred between adjacent MCI and EAD categories. This result is important because it demonstrates that the model is addressing precisely the boundary that is expected to be difficult because of the similarity of structural abnormalities during early disease progression.

However, the present study does not directly demonstrate improved clinical outcomes or improved physician decision-making. Therefore, the four-class classification should be interpreted as providing potentially clinically useful stage information, rather than as a validated replacement for clinical diagnosis. Prospective studies incorporating clinical decision analysis, longitudinal outcomes, and evaluation by clinicians will be necessary to determine whether the additional staging information leads to measurable improvements in patient management.

5.3. Dataset Bias and Generalizability

Despite the advantage of using both ADNI and OASIS-3 to conduct a much more extensive analysis compared to the use of only one neuroimaging database, both have sampling and acquisition features that make the generalization of the current results to clinical settings difficult. Thus, the high accuracy achieved through the testing on ADNI+OASIS-3 should be taken as a sign of the performance on these research databases, rather than clinical generalizability.

To begin with, there is a presence of selection bias in both datasets since the subjects in each case were not chosen randomly, but were enrolled into the study according to specific research procedures. Specifically, ADNI uses strict enrollment criteria and diagnosis criteria and serves as a research cohort aimed at longitudinal assessment of aging, mild cognitive impairment and Alzheimer's disease. The second dataset, OASIS-3, is a retrospectively compiled database of participants recruited through several research projects at Washington University Knight Alzheimer's Disease Research Center in about 30 years. This particular form of recruitment of subjects results in the selection of healthier and more motivated subjects who are characterized in more detail than the ones who might have been found during regular clinical practice.

Secondly, there is the limitation of racial and ethnic representation. Research studies on ADNI recruitment show

that just 11% of participants recruited were of ethnoculturally underrepresented groups compared to about 25% in the U.S. adult population aged 60 years or above, with less than 12 years of education being another group of underrepresented individuals. This has prompted the researchers to state that findings from ADNI might not be fully generalizable to ethnoculturally diverse and less educated population groups. In general, in a systematic review of U.S. Alzheimer's neuroimaging studies, underrepresentation of Black/African Americans, Hispanics/Latinos, Asian Americans and other marginalized groups was found in comparison to the overall U.S. population. It is especially critical when MRI imaging is used as different population groups have varying characteristics which could influence brain morphology, prevalence of disease, vascular load and comorbidities, and thus the visual features learnt by the convolutional model.

Thirdly, there might be educational and socioeconomic biases that could impact the discriminability of the classes. Education and context affect the cognitive performance and disease stages; on the other hand, research cohorts usually select only participants capable of undergoing multiple tests and imaging. According to the analysis of ADNI cohort recruiting, people who had received less than 12 years of education were underrepresented as well as less likely to pass the screening. In spite of the fact that the introduced classification scheme relies on structural MRI but not cognitive testing, the labels used in supervised learning are also obtained within a research environment where the selection is nonrandom.

Fourth, scanner effects, acquisition effects, preprocessing effects, and site effects may impact model performance. There is longitudinal MRI data in ADNI collected using different scanners and settings, and in the documentation for ADNI, it is stated that longitudinal scans may be affected by various changes in the hardware of the scanner, coils, software versions, and other technical aspects. Similarly, in OASIS-3, imaging is pooled from multiple studies and multiple decades. While this is useful in assessing robustness, it can also create systematic differences between the datasets which do not have anything to do with Alzheimer's disease. This means that a model may learn both biological and acquisition signals. The current preprocessing pipeline minimizes technical variability due to the use of bias correction and spatial normalization, among other things.

Finally, there is a specific longitudinal structure in the OASIS-3 dataset. The OASIS-3 dataset comprises repeat MRI scans of the same subjects and is compiled retrospectively from multiple studies with varying visit frequencies. Such longitudinal structure is a valuable asset when analyzing disease progression; however, it creates dependency among observations of the same subject. For this reason, in the current paper, subject-level partitioning was employed, so that both MRI scans and slices derived from them of the same subject belonged only to one partition. At the same time, the fact of longitudinal data makes the number of independent subjects less than the number of MRI scans.

It is notable that cross-cohort experiment gives better proof of generalization than internal test results; however, it cannot solve the mentioned problems. The model, which was trained on ADNI and tested on OASIS-3 without retraining, provided the accuracy of 94.2% and the macro-F1 of 93.7%, indicating that the obtained representations have good discriminative power after a shift in cohort. However, the transfer from ADNI to OASIS-3 is a comparison of two U.S.-based cohorts, and not an evaluation in different healthcare systems, regions, populations, or imaging setting. Differences in the cohorts themselves can also affect this result. The recent work comparing ADNI and OASIS-3 cohorts showed the differences in demographics, clinical profiles, and even the acquisition parameters of structural MRI data sets, which explains the necessity of understanding the difference between the performance of the models on different datasets and their generalizability.

These factors give rise to a key difference between generalization of a dataset and generalization clinically. The findings reported herein lend support to the assertion that the DP-CNN algorithm generalizes fairly well when applied from the ADNI cohort to an independent research cohort OASIS-3, while the 97.4% performance rate on the combined data set confirms good performance within the conditions of the study. These findings do not prove that the same performance rates can be achieved by applying the algorithm in other clinical settings where MRI machines are produced by other companies and operate under a different protocol or with community-based populations or underrepresented demographics.

5.4. Limitations

A number of restrictions are worth noting. Originally, the dataset is relatively small given the extent of model complexity used, since it consists of 6,400 scans, two cohorts, and the data needs to be validated on larger, multi-site datasets in prospective. Second, both the ADNI and OASIS-3 are North American, mostly White cohorts, which castings doubt on the generalizability of the demographics to populations that are underrepresented. Third, the input strategy is a 2D slice-based strategy, which is computationally efficient, but might not utilize the three-dimensional spatial context of volumetric MRI, future research can consider 3D convolutional or attention-based architectures. Fourth, the existing model lacks the use of longitudinal imaging, which restricts its capability to model disease progression or forecast conversion[24].

Several limitations should be considered when interpreting the findings of this study. First, although the dataset includes 6,400 MRI examinations from ADNI and OASIS-3, the sample size remains relatively modest compared with the parameter complexity of the dual-backbone architecture. Larger and more diverse multi-center datasets would be valuable for further assessment of generalization.

Second, the proposed model uses a multi-planar 2D slice-based representation rather than a native 3D architecture. This was a deliberate methodological trade-off that allowed the use of ImageNet-pretrained ResNet-50 and VGG-16 backbones, reduced computational and GPU-memory requirements, and facilitated slice-level Grad-CAM visualization. The use of axial, coronal, and sagittal views provides complementary anatomical information and partially compensates for the absence of volumetric modeling. Nevertheless, a 2D representation cannot explicitly learn the spatial continuity and long-range relationships between neighboring slices within the complete MRI volume. This may be particularly relevant for neuroanatomical structures whose morphology and atrophy patterns extend across multiple planes and slices.

Accordingly, 3D CNNs and 3D attention-based architectures represent an important direction for future investigation. Such models could potentially learn volumetric representations of hippocampal and cortical atrophy directly from the complete MRI volume and may provide a more anatomically coherent representation of disease-related changes. However, direct claims of superiority cannot be made from the present study because these architectures were not evaluated under the same dataset, preprocessing, training, and testing conditions. A controlled head-to-head comparison between the proposed 2D DP-CNN, a 3D CNN baseline, and a 3D attention-based model would therefore be an appropriate future experiment.

Third, the ADNI and OASIS-3 cohorts are not fully representative of the broader global population, and differences in scanner characteristics, acquisition protocols, demographic composition, and disease prevalence may affect external validity. Although the cross-cohort experiment provides evidence of generalization, additional independent and geographically diverse cohorts are required.

Fourth, the current analysis is primarily cross-sectional and does not explicitly model longitudinal disease progression. Longitudinal MRI sequences could provide additional information about individual rates of structural change and potentially improve discrimination between transitional stages such as MCI and EAD.

Finally, although Grad-CAM provides a useful visual explanation of the regions contributing to model predictions, it does not establish causal relationships between anatomical findings and diagnostic decisions. Prospective evaluation involving clinicians and independent expert assessment of the generated explanations is required before such visualizations can be considered validated clinical decision-support tools.

While the suggested model yielded high AUC-ROC scores, AUC reflects only the discriminatory ability and cannot serve as an indicator of well-calibrated probabilities. Calibration curves, reliability plots, as well as metrics like Brier score or ECE have not been included in the current assessment. Consequently, the output probabilities cannot be considered reliable for use in the clinical setting. Further research is required to assess the calibration of the probabilities by using external validation data and applying post-hoc calibration methods like temperature scaling.

6. Conclusions

This paper introduces a Dual-Path CNN architecture with state-of-the-art four-class staging of Alzheimer disease on structural MRI, and an overall accuracy of 97.4, AUC-ROC of 0.993, and cross-cohort generalizability of 94.2. The independent roles of dual path feature fusion (+4.3% accuracy) and spatial attention (+1.8% accuracy) were measured in systematic ablation experiments and principled architectural justification of each

design decision was made. Grad-CAM interpretability analysis established congruity between model saliency and known AD neuropathological landmarks, indicating the biological plausibility of the framework and its potential to be translated to clinical use. The ability to accurately differentiate MCI and EAD, the diagnostically and therapeutically important boundary, with non-invasive, non-experimental structural MRI makes the proposed framework a useful tool in terms of early AD detection in clinical and screening settings. Future studies will expand the framework to multi-modal fusion with amyloid-PET and tau-PET imaging and federated learning models to train privacy-preserving multi-site models and longitudinal data integration to model disease trajectories and predict conversion.

REFERENCES

1. Alzheimer's Association. (2024). Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 20, 3708–3821.
2. Dhia Jaafar, H., Jabbar Hassan, T., Hussien Mohamed, M., Abdullateff Abad, B., & Salman Qasim, S. (2026). Artificial intelligence and ensemble learning for coronary artery disease prediction. *Statistics, Optimization & Information Computing*, 15(5), 4255–4276. <https://doi.org/10.19139/soic-2310-5070-3354>
3. Wimo, A., Guerchet, M., Ali, G. C., Wu, Y. T., Prina, A. M., Winblad, B., & Prince, M. (2023). The worldwide costs of dementia 2019. *Alzheimer's & Dementia*, 19, 1–13.
4. Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14, 535–562.
5. Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6, 67–77.
6. Rahrouh, M., Ali Salim, R., Salman Qasim, S., Hussien Mohamed, M., & Mahmoud, N. (2026). A super-efficient and compact CNN for tomato leaf disease detection on resource-constrained devices. *Statistics, Optimization & Information Computing*, 16(2), 1045–1059. <https://doi.org/10.19139/soic-2310-5070-3778>
7. Jenifer, D. P. I. R., Reddy, D. V. M. K., Chaitanya, D. D. E., Pithani, L. K., Rao, D. P. V., Gupta, D. V., et al. (2026). A hybrid stylometric transformer embedding framework for robust authorship verification against generative AI adversaries. *Journal of Theoretical and Applied Information Technology*, 104(1).
8. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.
9. Fadlallah, H. R., Mohamed, M. H., Alayash, W., Stephan, J. J., Qasim, S. S., Alqabany, T., & Ali, M., 'SECURE IOT COMMUNICATIONS USING SCRP-DRIVEN DYNAMIC QUASIGROUP CRYPTOGRAPHY', *Journal of Theoretical and Applied Information Technology*, vol. 104, no. 1, Jan. 2026, <https://doi:10.5281/zenodo.18259402>.
10. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.-O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods. *NeuroImage*, 56, 766–781.
11. Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative (ADNI). (2012). Multi-modal multi-task learning for joint prediction of regression and classification variables in Alzheimer's disease. *NeuroImage*, 59, 895–907.
12. Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: A neuroimaging study with 3D convolutional neural networks. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Barcelona, Spain.
13. Hosseini-Asl, E., Keynton, R., & El-Baz, A. (2016). Alzheimer's disease diagnostics by adaptation of 3D CNN with small data sizes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA.
14. Wen, J., et al. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducibility evaluation. *Medical Image Analysis*, 63, 101694.
15. Lian, C., Liu, M., Zhang, J., & Shen, D. (2020). Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42, 880–893.
16. Wang, S., et al. (2020). Automatic diagnosis of attention deficit hyperactivity disorder using a multi-feature convolutional neural network with attention mechanism. *Computers in Biology and Medicine*, 119, 103704.
17. Liu, M., et al. (2022). A transformer-based network for Alzheimer's disease classification using multi-scale features and multi-modal brain MRI. *Medical Image Analysis*, 77, 102369.
18. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy.
19. Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in Aging Neuroscience*, 11, 194.
20. LaMontagne, P. J., et al. (2019). OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer's disease. *medRxiv*.
21. Jack, C. R., et al. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27, 685–691.
22. Braak, H., & Braak, E. (1995). Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiology of Aging*, 16, 271–278.
23. El-Aziem, Ayman H. Abd, Marwa Hussien Mohamed, and Ahmed Abdelhafeez. High-Security Image Encryption Using Baker Map Confusion and Extended PWAM Chaotic Diffusion. *Computers*, 15(2):106, 2026. <https://doi.org/10.3390/computers15020106>
24. M. H. Mohamed, M. Elkholy, and M. A. Marzouk, *Innovative Machine Learning Approaches for Identifying Pre-diabetes in Patients*, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 16, no. 1, pp. 365–378, March 2025.