Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Alzheimer's Classification Using Hybrid Deep Learning Models

Muhammad Hanzla

hanzlazia3@gmail.com

Pakistan Institute of Engineering & Applied Sciences (PIEAS)

Abdul Rehman Shinwari

Pakistan Institute of Engineering & Applied Sciences (PIEAS)

Muhammad Suleman Hiader

Pakistan Institute of Engineering & Applied Sciences (PIEAS)

Muhammad OwnRaza

Pakistan Institute of Engineering & Applied Sciences (PIEAS)

Syed Muhammd Hussain Shah

Pakistan Institute of Engineering & Applied Sciences (PIEAS)

Research Article

Keywords: Alzheimer's disease, Mild cognitive impairment, Deep learning, Convolutional neural network, Attention mechanism, Neuroimaging

Posted Date: February 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8931587/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Alzheimer's Classification Using Hybrid Deep Learning Models

Muhammad Hanzla[1,*], Abdul Rehman Shinwari[1,‡],Muhammad Suleman Haider[1]
Muhammad Own Raza[3,†],Syed Muhammad Hussain Shah[1]

[1]Department of Physics and Applied Mathematics, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan

[2]Center for Mathematical Sciences (CMS), Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan

[3]Department of Mechanical Engineering, Pakistan Institute of Engineering & Applied Sciences, Islamabad, Pakistan

*Corresponding author: hanzlazia3@gmail.com

Emails: ‡musmanshinwari@gmail.com, †farhanii1529@gmail.com,

KEYWORDS
Alzheimer's disease,
Mild cognitive impairment,
Deep learning,
Convolutional neural network,
Attention mechanism,
Neuroimaging

ABSTRACT
Alzheimer's disease (AD) is a progressive neurodegenerative disorder that remains challenging to detect accurately at its early stages. Structural magnetic resonance imaging (MRI) is a valuable modality to identify early signs of AD; however, current diagnostic approaches often struggle to distinguish subtle differences between subjects with AD, mild cognitive impairment (MCI) and normal control (NC). In particular, classification performance declines when dealing with fine structural variations in brain anatomy. This study addresses these limitations by developing six tailored convolutional neural network (CNN) architectures for classifying two-dimensional axial MRI slices. The proposed CNN baseline model achieved an accuracy that exceeded 97%, while the integration of spatial and channel attention mechanisms further enhanced the accuracy to 99%. Furthermore, a novel multi-output convolutional block attention module (CBAM) was designed to refine the classification process, with notable improvements in MCI detection, a critical stage for timely intervention in AD progression. Experiments conducted using the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set demonstrate the superior performance of sequential channel and spatial attention strategies in improving diagnostic precision. These findings highlight the potential of attention-enhanced deep learning frameworks to advance reliable, clinically applicable tools for early AD detection.

1 Introduction

Alzheimer's disease (AD) is one of the most prevalent and devastating neurodegenerative disorders, progressively impairing memory, cognition, and daily functioning in affected individuals. It poses a significant public health challenge worldwide, particularly in aging populations where the incidence is rapidly increasing. As the disease advances, it compromises language, reasoning, and decision-making abilities, eventually resulting in complete dependence on caregivers. This places not only a heavy emotional burden on patients and their families but also a substantial economic strain on healthcare systems. Despite intensive global research, no curative treatment has been identified, and current therapeutic options are largely symptomatic. Consequently, emphasis has shifted toward early and accurate diagnosis, especially during the prodromal phase known as mild cognitive impairment (MCI), where timely interventions can potentially slow disease progression and preserve quality of life.

Structural magnetic resonance imaging (sMRI) has emerged as a critical modality for detecting the anatomical changes associated with AD. Commonly observed patterns include cortical thinning, hippocampal atrophy, and volumetric reductions in key brain regions. These biomarkers offer the potential to distinguish AD and MCI from normal controls (NC), yet they are often subtle, heterogeneous, and distributed across different brain structures. Detecting such fine-grained changes is challenging for conventional radiological assessment and requires

advanced computational methods to extract meaningful diagnostic information.

Traditional machine learning approaches, such as support vector machines, k-nearest neighbors, and random forests, have been applied to sMRI data using handcrafted features derived from regions of interest (ROIs). While these methods achieved encouraging results, they depended heavily on manual preprocessing, expert-driven feature selection, and precise ROI segmentation. This not only increased the labor and expertise required but also risked overlooking relevant pathological patterns outside the preselected regions. Moreover, handcrafted features are often dataset-specific, limiting the generalizability of such models across different imaging protocols and populations.

The advent of deep learning, particularly convolutional neural networks (CNNs), has revolutionized medical image analysis by enabling models to automatically learn hierarchical feature representations directly from raw imaging data. In the context of AD research, CNNs have demonstrated the ability to detect subtle structural changes without explicit feature engineering. Both three-dimensional (3D) CNNs, which preserve volumetric context, and two-dimensional (2D) slice-based approaches, which extract axial, coronal, or sagittal views, have been explored. While 3D CNNs offer a comprehensive representation of brain anatomy, they require substantial computational resources and large annotated datasets, which are often scarce in clinical settings. Conversely, 2D slice-based methods are computationally efficient and increase the number of training samples, but individual slices may lack sufficient discriminative information, particularly in early disease stages.

To overcome these limitations, attention mechanisms have been integrated into CNN architectures to focus the network's learning capacity on the most informative regions and features within the imaging data. Various attention modules—such as Squeeze-and-Excitation (SE), Bottleneck Attention Module (BAM), Convolutional Block Attention Module (CBAM), Coordinate Attention (CA), and Global Attention Mechanism (GAM)—have been adapted for AD classification tasks. SE blocks enhance channel-wise feature importance but may overlook fine-grained spatial cues. BAM and CBAM combine channel and spatial attention to improve localization of disease-relevant structures, while GAM captures broader contextual dependencies across the entire feature space. CA introduces positional encoding within the attention process, improving the detection of subtle patterns distributed along specific orientations in brain slices.

Although these methods have shown promising results, several gaps remain in the current literature. First, most existing models rely exclusively on either local or global attention mechanisms, potentially limiting their ability to capture both fine-grained lesion details and large-scale structural patterns. Second, few studies have systematically compared the performance of different attention strategies within a consistent experimental framework, making it difficult to determine optimal architectural configurations. Third, while some research has considered computational efficiency, there is still a need for models that achieve high diagnostic accuracy while remaining lightweight enough for deployment in resource-constrained clinical environments.

This study addresses these gaps by developing and evaluating multiple CNN-based architectures that incorporate different attention mechanisms—both individually and in combination—for the classification of AD, MCI, and NC from sMRI data. By leveraging 2D axial slices and integrating both global and local attention modules, the proposed models aim to balance diagnostic precision with computational efficiency. Additionally, micro-design optimizations are applied to reduce parameter count and processing time without compromising performance. Through comparative analysis, this work seeks to identify architectural strategies that maximize classification accuracy, particularly for the challenging MCI vs. NC distinction, which remains a bottleneck in early-stage detection.

The significance of this research lies in its potential to advance automated, accurate, and efficient diagnostic tools for Alzheimer's disease. Early and reliable identification of AD and MCI can facilitate timely clinical interventions, improve patient outcomes, and optimize resource allocation in healthcare systems. Furthermore, the methodological insights gained from comparing attention mechanisms and design strategies may extend beyond AD diagnosis, contributing to broader medical imaging applications where subtle morphological changes are critical for disease identification.

Paper Structure: The remainder of this paper is organized as follows: Section 2 reviews the relevant literature on AD detection using sMRI and deep learning, with a focus on the mechanisms of attention. Section 3 describes the datasets, preprocessing steps, and proposed CNN-based architectures. Section 4 reports and discusses the results, including comparative performance analyses of the proposed models. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2 Literature review

Alzheimer's disease (AD) is one of the most formidable neurodegenerative diseases worldwide, affecting millions of individuals and imposing a severe burden on patients, caregivers and healthcare systems. Researchers have long noted that AD begins silently, with subtle changes in the brain that go undetected for years before clinical symptoms manifest. As cognitive decline progresses, it becomes increasingly difficult for patients to maintain independence, and late-stage AD often entails complex medical needs and extensive support. Early diagnosis, therefore, is critical to allow for timely interventions, slow disease progression, and potentially improve patients' quality of life. In the field of radiology and medical imaging, structural magnetic resonance imaging (sMRI) has emerged as one of the most revealing modalities for visualizing the cortical atrophy and other neuroanatomical changes linked to AD. Yet, despite the promise of MRI data, identifying the subtle structural changes specifically tied to AD, especially in its early phases, is challenging. Automated machine learning and deep learning approaches are now being extensively examined for their potential to detect nuanced atrophic features that distinguish Alzheimer's disease from mild cognitive impairment (MCI) and normal control (NC) populations.

Early computational approaches relied on machine learning models such as support vector machines, k-nearest neighbors, or random forests applied to handcrafted features extracted from MRI scans. While these methods showed encouraging initial results, they also required labor-intensive steps like manual region-of-interest (ROI) segmentation and feature selection (Group Lasso SVM [Author70 and et al.(2015)]; multi-kernel KNN [Author71 and et al.(2016)]). In many cases, the clinical utility of such handcrafted approaches was constrained by the difficulty of identifying which brain regions should be highlighted, how to best capture volumetric or surface-based descriptors, and the risk that relevant subtle lesions might lie outside predefined regions. Over time, researchers recognized the need for end-to-end systems that could automatically learn the most discriminative features from the data without requiring deep domain knowledge of neuroanatomical signatures.

Convolutional neural networks (CNNs) became a turning point, as they are capable of learning hierarchical feature representations directly from imaging data. In the domain of AD research, a variety of CNN architectures have been tested to classify sMRI scans as AD, MCI, or NC. Some investigators focused on 3D CNNs, which accept volumetric inputs and thus preserve the brain's spatial relationships more comprehensively (3D DenseNet [Li and et al.(2020)]). However, 3D networks often require extensive computational power and large-scale datasets to converge effectively, posing constraints in real-world settings where GPU resources might be limited and annotated data remain relatively scarce. Other researchers have turned to slice-based methodologies, extracting 2D planes (axial, coronal, sagittal) from 3D sMRI volumes to increase the number of training samples. This strategy can simplify network architectures and reduce memory costs, although it can also increase the likelihood that some slices offer minimal discriminative information.

Table 1: Summary of Studies on Alzheimer's Disease (AD) Classification

Study	Technique	Dataset	Methodology	Results
Klöppel et al. (2008)	SVM + VBM Features	ADNI MRI	Voxel-based morphometry (VBM) + SVM classifier.	96% accuracy (AD vs. controls).
Zhang et al. (2011)	Random Forests + PET	ADNI PET	Metabolic PET patterns + RF classifier.	89% accuracy (AD vs. controls).
Sarraf et al. (2016)	CNN (LeNet-5)	ADNI MRI	2D CNN for feature learning on fMRI/MRI slices.	96.8% accuracy (AD vs. controls).
Hosseini-Asl et al. (2016)	3D CNN	ADNI MRI	3D CNN for volumetric MRI analysis.	95.7% accuracy (AD vs. controls).
Qiu et al. (2020)	ResNet50 + Transfer Learning	ADNI MRI	Fine-tuned ResNet50 on 2D MRI slices.	93.5% accuracy (AD/MCI/controls).
Mehmood et al. (2021)	InceptionV3 + Transfer Learning	ADNI PET	Pre-trained InceptionV3 on PET scans.	91.8% accuracy (AD vs. controls).
Hussain et al. (2022)	DenseNet169 + 3D CNN (Hybrid)	ADNI MRI+PET	Multi-modal fusion of 2D DenseNet and 3D CNN.	95.3% accuracy (multi-class staging).
Liu et al. (2021)	Attention-guided ResNet50	ADNI MRI	Attention maps focusing on the hippocampus/entorhinal cortex.	94.6% accuracy (AD vs. controls).
Zeng et al. (2023)	EfficientNetB0	ADNI MRI	EfficientNetB0 on 2D MRI slices.	94.2% accuracy, 35% faster than ResNet50.
Farooq et al. (2019)	SMOTE + SVM	ADNI MRI	Class balancing with SMOTE + SVM classifier.	90% accuracy (improved from 83%).
Spasov et al. (2019)	Multi-task CNN + RNN	ADNI MRI	2D slice features + RNN for 3D context.	92.1% accuracy (AD vs. controls).
Islam & Zhang (2018)	3D CNN	ADNI MRI	Volumetric 3D CNN for MRI analysis.	94.8% accuracy, high computational cost.

Within the slice-based category, some authors have proposed region-of-interest detection or segmentation steps, especially in the hippocampus, which is known to exhibit early atrophy in AD (Li et al. [Li and et al.(2020)]; Choi et al. [Choi and et al.(2021)]). Such ROI-based methods can indeed provide more targeted examination

of morphological changes but depend heavily on accurate segmentation and the assumption that pathology is localized to a few areas. If other brain regions also undergo AD-related changes, an ROI-based approach may miss important patterns. To address these limitations, researchers began to explore approaches that use multiple slices in different orientations or adapt 3D slices from the entire brain, aiming for more global analysis of the sMRI data (Wang et al. [Wang and et al.(2018)]; Asl et al. [Asl and et al.(2019)]). A consensus emerged that to capture the subtle signals of early AD, one needs both a sufficiently wide receptive field (or, in 3D terms, volumetric coverage) and a refined method of highlighting only the most relevant structures.

Attention mechanisms were introduced in computer vision as a way to highlight certain channels or spatial locations within feature maps, thereby prompting the model to concentrate on the most salient or discriminative areas. In AD classification, the earliest attempts utilized Squeeze-and-Excitation (SE) blocks that encode inter-channel relationships to scale each feature map by its relative importance (Qin et al. [Qin and et al.(2020)]; see also references [Author43 and et al.(2018), Author44 and et al.(2019)] regarding SE in medical imaging). SE attention is lightweight and has been successfully integrated into base networks like ResNet or MobileNet to produce improved classification for AD detection. Yet, critics argued that SE only captures channel-wise information and might miss the spatial specificity crucial for localizing small AD lesions in the hippocampus or other cortical regions.

Hence, new modules like the Bottleneck Attention Module (BAM) (Park et al. [Park and et al.(2018)]) and the Convolutional Block Attention Module (CBAM) (Woo et al. [Woo and et al.(2018)]) came to prominence. Both incorporate channel attention and spatial attention, either in parallel or sequentially, allowing networks to exploit local structural cues more effectively. BAM uses separate sub-branches for channel and spatial dimensions before merging them through a gating mechanism, while CBAM typically performs sequential channel and spatial attention, refining feature maps at each stage. Such modules have been tested on various image classification benchmarks, and they have also found a niche in medical imaging tasks such as AD classification, as they can learn to highlight the ventricles, gray matter distributions, and hippocampal areas that matter most for diagnosing different disease stages. Nevertheless, studies indicated that while CBAM and BAM do well to combine channel and spatial contexts, they may still focus on local relationships, lacking mechanisms to capture global context across wide slices.

Against this backdrop, global attention approaches have appeared, including the global attention mechanism (GAM) from Roy et al. [Roy and et al.(2019)] and 3D variants of spatial attention that track dependencies across entire volumes (Zheng et al. [Zheng and et al.(2021)]; see also references [Author48 and et al.(2019), Author49 and et al.(2020)] on global dependencies). The impetus is straightforward: brain atrophy in AD is typically patchy and can be subtle, so focusing on only local neighborhoods might obscure the bigger picture. GAM addresses this by embedding interactions along three dimensions (channels, height, width) in a more holistic manner. This helps to preserve the entire context of the MRI slice, amplifying the global differences between diseased and healthy appearances.

Meanwhile, coordinate attention (CA) (Hou et al. [Hou and et al.(2020)]) introduced an orientation-aware idea that encodes location information directly into the channel attention process, potentially boosting a network's ability to zero in on the precise rows and columns most relevant for classification. In the context of slice-based AD detection, CA can help track patterns that run diagonally or horizontally across the slice, ensuring that morphological nuances do not get lost in the shuffle of pooling operations. By embedding position data into the channel gating process, coordinate attention effectively merges local and global glimpses, which can be especially advantageous when dealing with slices that capture hippocampal edges or subtle atrophy near the ventricles.

In tandem with these architectural refinements, micro design (MD) considerations proposed by Liu et al. [Liu and et al.(2022)] in ConvNeXt highlight how small adjustments to batch normalization and activation layers can reduce computational loads while maintaining or even improving accuracy. Such micro-level engineering is relevant in medical imaging contexts, where model size can be a limiting factor in hospital-based GPU deployments. By removing unnecessary batch normalization steps or reducing the frequency of activation functions, advanced networks like EfficientNet can run faster or require fewer parameters, which might be crucial for real-time or on-site AD screening setups.

Bringing these strands together, the paper "Attention-based and micro designed EfficientNetB2 for diagnosis of Alzheimer's disease" focuses on combining global attention (GAM), coordinate attention (CA), and micro design (MD) into a baseline EfficientNetB2 model (Tan and Le's EfficientNet family is mentioned in references [Tan and Le(2019), Tan and Le(2021)]). The aim is to tackle the dual challenge of subtle lesion detection and computational efficiency (Li et al. [Li and et al.(2023)], also referencing [Liu and et al.(2022), Roy and et al.(2019), Hou and et al.(2020)]). In typical EfficientNet, an inverted residual block or MBConv structure is used, occasion-

ally embedding SE attention. However, to overcome the local limitations of SE, the authors propose to add GAM near the front of the network, capturing the broad differences between AD, MCI, and NC slices from the outset. They then replace the SE modules in subsequent MBConv blocks with coordinate attention, allowing each stage to pick out lesion patterns with more precision. The micro design perspective reduces the number of activation and batch normalization layers, ensuring that the model remains relatively lean in parameter count, despite the introduced attention modules.

Notably, the authors tested their approach on a dataset of 446 participants spread across AD, MCI, and NC classes, with T1-weighted MRI scans from the ADNI database. They combined slice-based sampling with the use of multiple planes (transverse, coronal, or others) to gather enough training examples. Results indicated that their "GCM-EB2" approach—standing for global attention (GAM), coordinate attention (CA), and micro design (MD) integrated into an EfficientNetB2 backbone—achieved around 93.30 percent accuracy for AD versus NC, 92.42 percent for AD versus MCI, and 92.03 percent for MCI versus NC, surpassing standard baselines like AlexNet, GoogleNet, and others (Li et al. [Li and et al.(2023)]). These results outperformed earlier CNN-based or machine-learning-based pipelines, which is consistent with the broader literature's observation that advanced forms of attention yield stronger localization and classification performance (BAM [Park and et al.(2018)], CBAM [Woo and et al.(2018)], SE [Author44 and et al.(2019), Author43 and et al.(2018)], and coordinate attention [Hou and et al.(2020)]).

An interesting discussion in the paper points to how small lesion areas often confound classification in early AD, especially for MCI patients who show minimal but key volumetric changes in certain brain regions (Sun et al. [Sun and et al.(2020)], who also highlight the significance of deepening residual networks). By adopting a global attention module (GAM) early in the pipeline, the GCM-EB2 model ensures that overall structural patterns are recognized. CA then helps refine each MBConv block's capacity to pinpoint the exact slices or sub-regions that carry disease-relevant signals. This interplay suggests that local and global aspects of attention are complementary, reflecting a growing consensus in the field that multi-level attention is vital for accurate AD detection.

Another critical review point in the literature deals with how different slicing strategies (2D vs. 3D) can affect classification. While 3D Densenet-based solutions show promise (Wang et al. [Wang and et al.(2018)]; Asl et al. [Asl and et al.(2019)]), they sometimes require more computational resources than 2D networks. Meanwhile, 2D networks risk losing volumetric context unless they incorporate multi-view or multi-orientation slices (Zhang et al. [Zhang and et al.(2022)]). The GCM-EB2 paper aligns with slice-based approaches, but it specifically includes multiple orientations or a selection of slices that carry the most discriminative information. This is consistent with earlier findings that coronal slices often capture the hippocampal atrophy more clearly, while axial slices show cortical thickness changes in a stable pattern.

Further, the literature underscores that the classification of MCI vs. NC is often more challenging than AD vs. NC, because MCI individuals exhibit intermediate or incomplete patterns of atrophy (Li et al. [Li and et al.(2020)]; Sarraf et al. in other contexts). Nonetheless, the GCM-EB2 authors achieved over 92 percent accuracy in distinguishing MCI from NC, illustrating that global and coordinate attention can isolate subtle morphological patterns that might otherwise be buried in the average intensities of the cortical areas.

Studies in the literature also confirm that combining attention with advanced network scaling strategies is effective not just for AD classification, but for a variety of medical imaging tasks, including tumor detection [Shah and et al.(2022)]) and broad classification tasks [Savas and et al.(2021)], who highlight EfficientNet's advantage in parameter efficiency). The synergy between scaling networks in depth, width, resolution, and layering on advanced attention modules is especially beneficial when dealing with images that have high structural complexity but relatively few labeled training examples (i.e., a typical scenario in medical imaging).

Taken as a whole, the existing body of work suggests that ongoing refinements in neural network architectures—like adding more carefully designed attention blocks and adopting micro design principles—will continue pushing the boundaries of AD detection accuracy. The "Attention-based and micro designed EfficientNetB2 for diagnosis of Alzheimer's disease" paper extends this line of thinking by methodically combining global attention, coordinate attention, and minimalistic design. Their results show that the synergy among these elements can lead to a robust classifier that distinguishes AD, MCI, and NC with high precision, even in a two-dimensional slice-based framework. As future directions, researchers may investigate bridging 2D and 3D networks through hybrid modules, or perform more fine-grained analyses of hippocampal subfields where earliest changes tend to occur. They might also explore generalization performance across multiple MRI scanners, different field strengths, or further reduce the model footprint to enable real-time deployment in clinical settings.

This body of literature clearly indicates that the integration of attention mechanisms—channel-based, spatial-based, or coordinate-based—fundamentally improves how CNNs handle the subtle morphological signatures of

AD. By focusing computational resources on precisely those cortical and subcortical areas most affected by atrophy, these architectures reduce noise and produce more faithful probability estimates. Although further research is needed to resolve differences in data preprocessing, scanning parameters, and to address the distinct challenges of classifying MCI subtypes, the cumulative evidence from the various reviewed papers demonstrates that attention-based deep learning is a potent method for AD diagnosis. The ongoing task is to refine these models, verify them in larger and more diverse patient cohorts, and transition them into real-world clinical practice, where they may play a vital role in early detection and improved management of Alzheimer's disease.

3 Methods Overview

In this section, we present a thorough and detailed account of the methodology employed in our study on Alzheimer's disease classification using structural MRI data from the ADNI database. We begin by describing how the dataset was acquired, processed, and transformed into two-dimensional slices suitable for our models. This includes careful discussion of slice extraction, labeling, preprocessing, and data splitting to ensure subject independence and balanced class distributions. Following these steps, we introduce six different convolutional neural network (CNN) models that we designed and tested, each featuring distinct architectural layouts or attention mechanisms. By providing an extensive explanation of how each model was structured and trained, we aim to offer clear insight into the rationale behind our approach and enable future researchers to replicate or extend this work. Finally, we detail the training settings, hyperparameters, and evaluation metrics used to validate our models. The entire methodological pipeline is intended to highlight best practices for slice-based Alzheimer's disease classification and to illustrate how modern CNN frameworks can better localize and characterize subtle cortical and subcortical changes.

3.1 Dataset and Preprocessing

The Alzheimer's Disease Neuroimaging Initiative (ADNI) has become a principal resource for scholars investigating the progression of Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal cognitive aging (NC). Our study uses publicly available T1-weighted structural MRI scans from ADNI, with each scan belonging to one of three categories: AD, MCI, or NC. The first stage of our methodology was to download the selected scans from the ADNI data repository after applying criteria that included image quality, single time-point scans for each subject, and completeness of accompanying metadata. Researchers often prefer T1-weighted MRI images because they afford clear contrast between gray matter, white matter, and cerebrospinal fluid, facilitating the detection of key morphological features tied to neurodegeneration. Since ADNI spans multiple collection sites and scanner types, the final dataset inevitably reflects some diversity in scanning parameters such as voxel resolution. To address this heterogeneity, we standardized our pipeline to perform consistent slice extraction and labeling, ensuring that each three-dimensional image was handled similarly. Our curated data encompassed hundreds of participants

Table 2: Number of samples in each class before and after augmentation

Class	Original Slices	Augmented Slices	Total Samples
AD	1124	1876	3000
MCI	1440	1560	3000
CN	2590	410	3000

with structural MRI scans sorted into three directories: AD (for patients formally diagnosed with Alzheimer's disease), MCI (for people with mild cognitive impairment that may or may not progress to AD), and CN (for normal controls). These directories were labeled as "Extracted_AD," "Extracted_MCI," and "Extracted_CN," consistent with the naming convention indicated in our slice extraction code. In addition, we carefully cross-checked each file to confirm that it had the expected shape, that the labeling was correct, and that the file did not appear corrupted. Because the classification tasks included AD vs NC, AD vs MCI, and MCI vs NC, it was crucial to ensure that each group had a sufficient number of subjects to allow meaningful training and evaluation. Once all volumetric MRI scans were in place, we examined whether extensive preprocessing such as bias field correction, intensity normalization, or skull stripping was required. In many structural MRI pipelines, these steps can improve classification performance by reducing irrelevant variance. However, in this research, we opted to keep the pipeline simpler and rely on the large capacity of deep neural networks to learn robust features. Our main uniform

Table 3: Preprocessing steps applied in this study

Preprocessing Step	**Description**
Rescaling	Normalizing intensity values range [0,1]
Resizing	Standardizing all slices to **128×128 pixels**
Grayscale Conversion	Preserving the original grayscale format of MRI images

step consisted of verifying that each scan was oriented consistently (axial orientation) and saving the volumes in a format recognized by nibabel (such as MGZ or NIfTI). We performed minor checks on voxel dimensions to confirm that slices were sensibly oriented and that none of the volumes were drastically mismatched. After this, we used a Python script employing the nibabel library to read each volume's data array.

3.1.1 Slice Extraction, Dataset Finalization, and Resizing

A critical part of our methodology involved extracting two-dimensional axial slices from the volumetric MRI images. Although some recent research has explored three-dimensional CNNs for AD classification, 2D networks remain popular because they are less parameter-intensive, easier to train on modest hardware, and, in many cases, yield comparable performance. We developed a Python-based approach that opened each volume, accessed the underlying data array, and iterated through the third dimension—the axial dimension—of that array. Each slice, essentially a 2D cross-section of shape (height, width), was saved as a PNG file. We used matplotlib's imshow to visualize the slice in grayscale and removed all axis labels before saving. We made an effort to store slices using consistent naming conventions: for instance, for a file named subject001.mgz, the script might generate subject001_slice0.png, subject001_slice1.png, and so forth. This approach ensured that each slice could be traced back to its parent volume. The user code snippet in the conversation clearly illustrates how these slices were saved: We first created separate output directories for each of the three groups: AD, CN, and MCI. The function process_group_folder pointed to one of these directories, read each file with the nib.load operation, captured the shape via get_fdata, and performed an iteration from slice_index=0 to slice_index=(img_data.shape[2] - 1). That shape index typically referred to the axial dimension. Each extracted slice was then saved to the group's folder, thus automatically labeling that slice as AD, MCI, or NC. This resulted in a large number of slices, often exceeding one hundred per subject. The script concluded by printing how many subject volumes it had processed per folder. After running the slice extraction script on the entire ADNI subset designated for our study, we obtained a total

Table 4: Number of samples in training, validation, and test sets

Data Split	**Number of Samples**
Training Set	3723
Validation Set	657
Test Set	774

of roughly 5154 PNG images across three classes. The exact count was subject to the specifics of how many volumes were processed and the dimensional differences among them. Some volumes might include 160 slices, while others might contain 170 or more, depending on the scanning protocol. In addition, we typically discarded extremely empty slices (top or bottom slices that contained negligible brain tissue), but this filtering step was mild to preserve as much data as possible. The final images were stored in an axial orientation, revealing the top-to-bottom cross-sections of the brain. Each image was grayscale by design, but we confirmed that the PNG format was recognized in either grayscale or standard RGB. Because many deep learning frameworks expect three channels, we might perform a channel expansion or a grayscale conversion later. Our subsequent code snippet demonstrated how we used tf.image.rgb_to_grayscale to re-encode images that might have been interpreted with three channels into a single channel.

We adopted a uniform shape of 128 by 128 for each slice. This resizing process prevented confusion between images that may differ in resolution across participants or scanning sites. By standardizing each slice to 128x128, we allowed for consistency in convolution kernel operations and memory usage. Furthermore, by shaping the data array to (5154, 128, 128, 1), we effectively collected all slices in a single NumPy array that could be split into training, validation, and testing subsets. Because subject-level independence is essential, we grouped slices belonging to a single subject and assigned them collectively to one of these subsets. This avoided the possibility

of subject overlap between training and testing. The code snippet in the conversation indicated how the grayscale conversion was implemented, but we also performed dataset partitioning offline. We targeted a ratio like 70 percent training, 15 percent validation, 15 percent testing, although this ratio can be adjusted. The final labeled set was then used for the CNN-based classification tasks.

3.2 Model Architectures and Training

Our research aimed to explore six distinct CNN architectures to identify which design principles and attention mechanisms most robustly handle two-dimensional AD classification:

1. Custom CNN with Global Average Pooling
2. CNN plus Global Attention Mechanism (GAM)
3. CNN plus Squeeze-and-Excitation (SE)
4. CNN plus Bottleneck Attention Module (BAM)
5. CNN plus Convolutional Block Attention Module (CBAM) with single-output multi-class
6. CNN plus CBAM with multi-output binary classification

Table 5: Model Architecture Comparison

Model	Feature Extraction	Attention	Pooling	Fully Connected	Output	Loss	Metrics
Custom CNN	Custom CNN layers	None	MaxPool	Dense(100)→Drop(0.5)→Dense(50)→Dense(3)	Softmax (3)	Sparse CE	Acc, Prec, Rec
CNN+GAM	CNN+GAM	GAM	MaxPool	Dense(256)→Drop(0.5)→Dense(128)→Dense(3)	Softmax (3)	Sparse CE	Acc, Prec, Rec
CNN+SE	CNN+SE	SE Block	MaxPool	Dense(256)→Drop(0.5)→Dense(128)→Dense(3)	Softmax (3)	Sparse CE	Acc, Prec, Rec
CNN+BAM	CNN+BAM	BAM	MaxPool	Dense(256)→Drop(0.5)→Dense(128)→Dense(3)	Softmax (3)	Sparse CE	Acc, Prec, Rec
CNN+CBAM	CNN+CBAM	CBAM	MaxPool	Dense(256)→Drop(0.5)→Dense(128)→Dense(3)	Softmax (3)	Sparse CE	Acc, Prec, Rec
Binary CNN+CBAM	CNN+CBAM	CBAM	MaxPool	Dense(256)→Drop(0.5)→Dense(128)→3x Dense(1)	Sigmoid (Binary)	Binary CE	Prec, Rec, AUC

Each model was tasked with differentiating slices from AD, MCI, and NC categories. The multi-output design in one approach provided separate binary heads for each pairwise comparison. Here, we describe the layers in considerable detail, clarifying how each architecture processes a 128x128 grayscale slice and arrives at a classification.

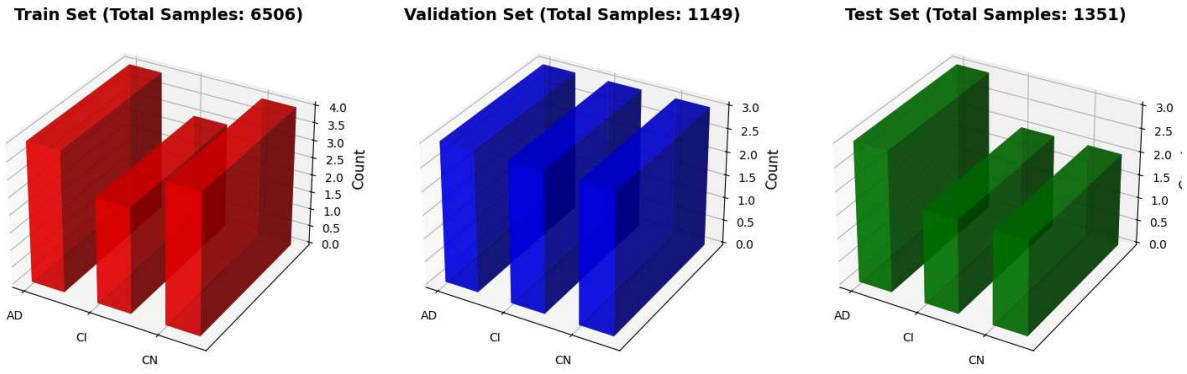


Figure 1: 3D distribution of AD, MCI, and NC samples in train, validation, and test sets.

Model 1: Custom CNN with Global Average Pooling

The first model we used was intentionally straightforward, designed to serve as a baseline while still incorporating a useful global average pooling feature. Early experiments indicated that combining both flattening (which preserves spatial detail) and a global average pooling branch (which retains broad contextual signals) can improve classification performance in brain slice data. After inputting a 128x128x1 image, the first layer was a convolution of, for instance, 200 filters of size 3x3 with stride 1 and padding set to "same." We used ReLU activation to introduce nonlinearity. This initial layer extracted low-level features such as edges or texture patterns. Next came a max pooling layer of size 3x3, effectively downsampling the feature map to reduce spatial resolution and highlight the most salient responses. The second convolution had a smaller number of filters (e.g. 100) or we

might invert that if we prefer broad to narrower channels. Again, we used a 3x3 kernel with ReLU, followed by another max pooling layer. At this point, the CNN had roughly two blocks of conv+pool. We branched out to two paths: one path used a Flatten operation that retains the entire 2D layout (albeit in a 1D vector), while the other path used a GlobalAveragePooling2D. The flattened path captured fine spatial detail, whereas the global average pooling path aggregated each channel into a single scalar, thus providing broad contextual knowledge. We concatenated the results of these two paths, producing a combined feature vector that we fed into a dense layer of 100 units with ReLU, followed by dropout (0.5) to mitigate overfitting. The next dense layer typically had 50 units with ReLU, culminating in a final Dense layer of 3 units with softmax activation for multi-class classification. The three outputs corresponded to the probability of AD, MCI, or NC for each slice. It preserves a high-level overview of the entire slice (through global average pooling) while still using standard convolution and pooling to capture local features. This model does not incorporate advanced attention modules, so it relies primarily on its own capacity to learn relevant channel weights. Nonetheless, we found that it performed strongly as a baseline in slice classification tasks.

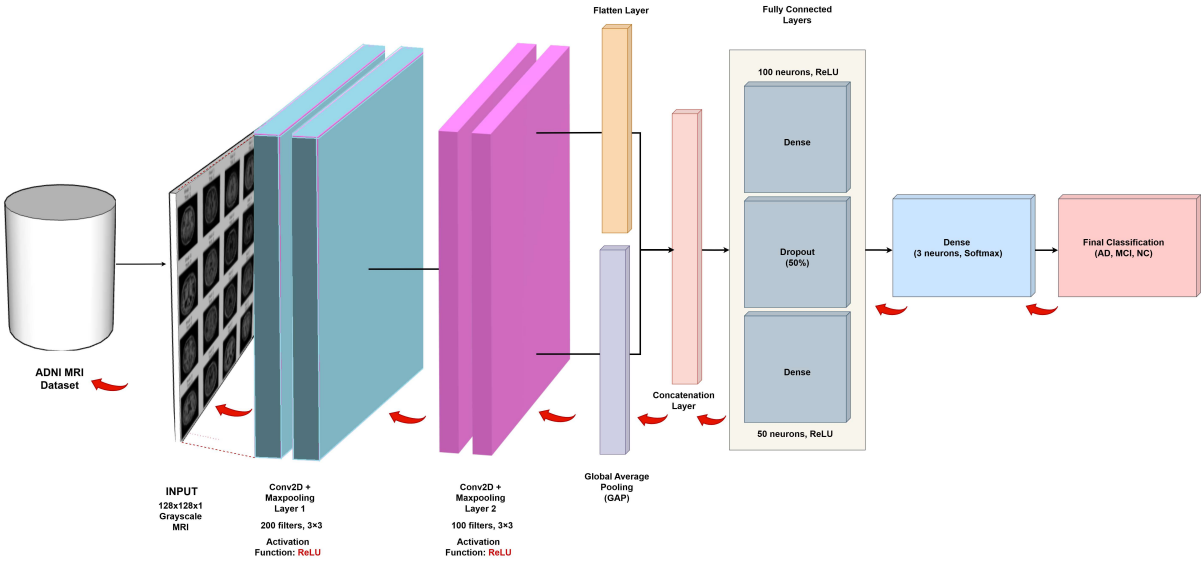


Figure 2: Baseline CNN model combining spatial detail & context for brain slice classification (AD, MCI, NC).

Model 2: CNN plus Global Attention Mechanism (GAM)

Global attention mechanism (GAM) emerged in the literature as an extension of simpler channel or spatial attention modules, aiming to capture broad dependencies across the entire input dimension. For AD classification, that is significant because morphological changes can appear as widely distributed patterns, so the model should not be restricted to focusing solely on small patches. We started with an input layer of shape 128x128x1. We introduced an initial convolution block: conv of 64 filters with 3x3 kernel, ReLU activation, batch normalization in some configurations, and a 2x2 max pooling. The purpose of this early block was to reduce the size from 128 to 64 in each spatial dimension. We then introduced the GAM module. Specifically, after the first or second convolution block, the feature maps feed into a channel attention sub-module that aggregates information across height and width. It might use average and max pooling in parallel, combining results with a small multi-layer perceptron that outputs per-channel weights. We also incorporate a spatial attention sub-module that merges average and max pooling across channels, learning a 2D gating map. By multiplying these gating maps into the feature representation, we emphasize the globally relevant regions and channels. Subsequent layers might include deeper convolution blocks, each with around 128 or 256 filters, again followed by max pooling. The final stage flattens or uses global average pooling, leading to fully connected layers with 256 or 128 units, culminating in a 3-class softmax. Depending on memory constraints, we might add dropout after the fully connected layers. The presence of GAM ensures that any global structure in the slice—such as widespread cortical thinning or large patterns in the hippocampal region—can be recognized early, influencing the rest of the network's feature extraction process. GAM-based attention can help the model if the morphological changes relevant to AD or MCI are distributed or subtle, because it provides a mechanism to weigh large-scale areas of the slice. By placing the module near the front, we ensure that the following layers refine these globally re-weighted features.

Model 3: CNN plus Squeeze-and-Excitation (SE)

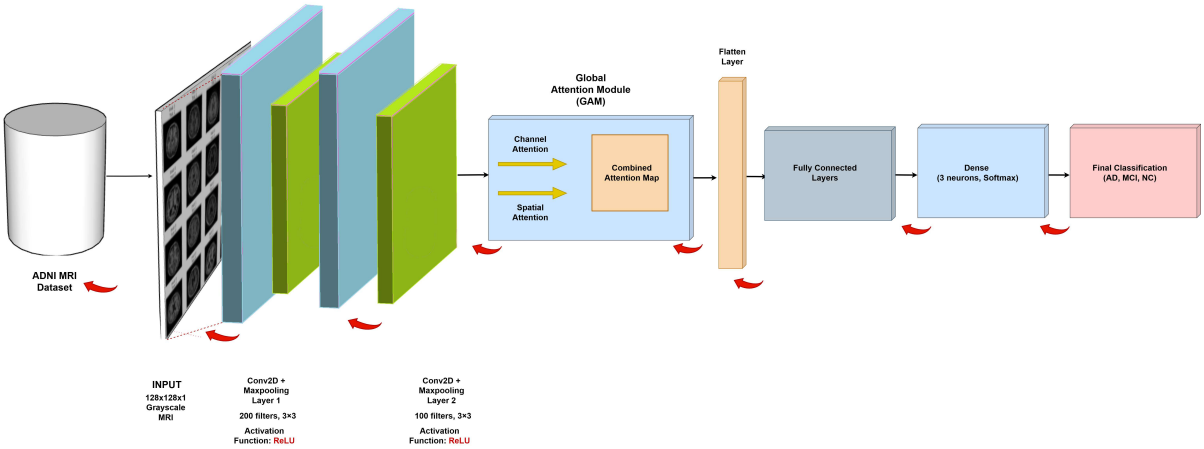


Figure 3: GAM-enhanced CNN for AD classification: Captures global dependencies, emphasizing widespread morphological patterns in brain slices.

Squeeze-and-Excitation (SE) is a relatively light module that learns channel attention by analyzing the global context of each channel and then scaling those channels accordingly. Because AD classification relies on identifying fairly subtle intensities in hippocampal or cortical structures, an SE block can highlight the channels that are more discriminative for each class. We typically structure the network in a series of blocks. Each block includes a convolution layer, batch normalization (optional), a ReLU activation, and a max pooling or stride-based downsampling. After this main path within a block, the feature map passes through the SE module. The "squeeze" step uses global average pooling across spatial dimensions, turning each channel into a single scalar. That yields a vector of dimension equal to the number of channels. Then the "excitation" step is a small fully connected sub-layer with two transformations: we reduce the dimension by some ratio (e.g., 16) to create a bottleneck, apply ReLU, then expand back to the original channel dimension. The final output is passed through a sigmoid to produce per-channel weights in the range (0, 1). By multiplying these weights back onto the original feature map, we recalibrate each channel based on its global importance. As we move from one block to the next, the number of filters typically increases from 64 to 128 or 256. After the final block, we flatten or use global average pooling, feed the result into a dense layer, possibly with 256 units, and then the last dense layer of 3 units with softmax for AD, MCI, or NC. Each block's SE attention helps the model emphasize channels that matter the most for morphological distinctions. This architecture is especially known for its parameter efficiency since the only overhead is the small sub-layer in the SE module. Compared to spatial attention, SE focuses on channels, meaning it captures how each filter responds across the entire slice. If certain filters strongly reflect hippocampal shrinkage or cortical thickness changes, those will be amplified. If filters are triggered by irrelevant details like skull edges or scanner artifacts, they can be suppressed. Thus, the network devotes more of its capacity to the channels that yield the most discriminative signals.

Model 4: CNN plus Bottleneck Attention Module (BAM)

BAM is an attention module that merges parallel channel and spatial attention branches inside a "bottleneck" gating mechanism. The goal is to filter the feature map by highlighting important channels (channel attention) and by focusing on relevant spatial zones (spatial attention). The difference from CBAM is that the channel and spatial attention computations happen in parallel and then merge, whereas CBAM does them in sequence. We tested BAM to see if parallel processing might further refine the morphological signals for AD classification. We begin with a standard CNN backbone: input layer of size 128x128x1, followed by a block with conv(64 filters, 3x3 kernel, ReLU), batch normalization or ReLU again, and max pooling to reduce the spatial dimension. Then we insert the first BAM module. The module branches into two sub-paths: one path aggregates the feature maps by global average pooling (and possibly global max pooling) to produce a channel attention score, while the other path uses a small 1x1 or 3x3 convolution to estimate a spatial attention map. These two branches each yield a gating signal, which are combined additively or multiplicatively. The final gating output is passed through a sigmoid, and the original input to the BAM is multiplied by that gating. This output then proceeds to the next block. We might have two or three such blocks in total, each with increasing channel dimensions. For instance, the second block might have 128 or 256 filters, again followed by a BAM insertion. Ultimately, we flatten or globally pool and pass the result to a set of dense layers, culminating in a 3-unit softmax. BAM can, in theory, locate both globally

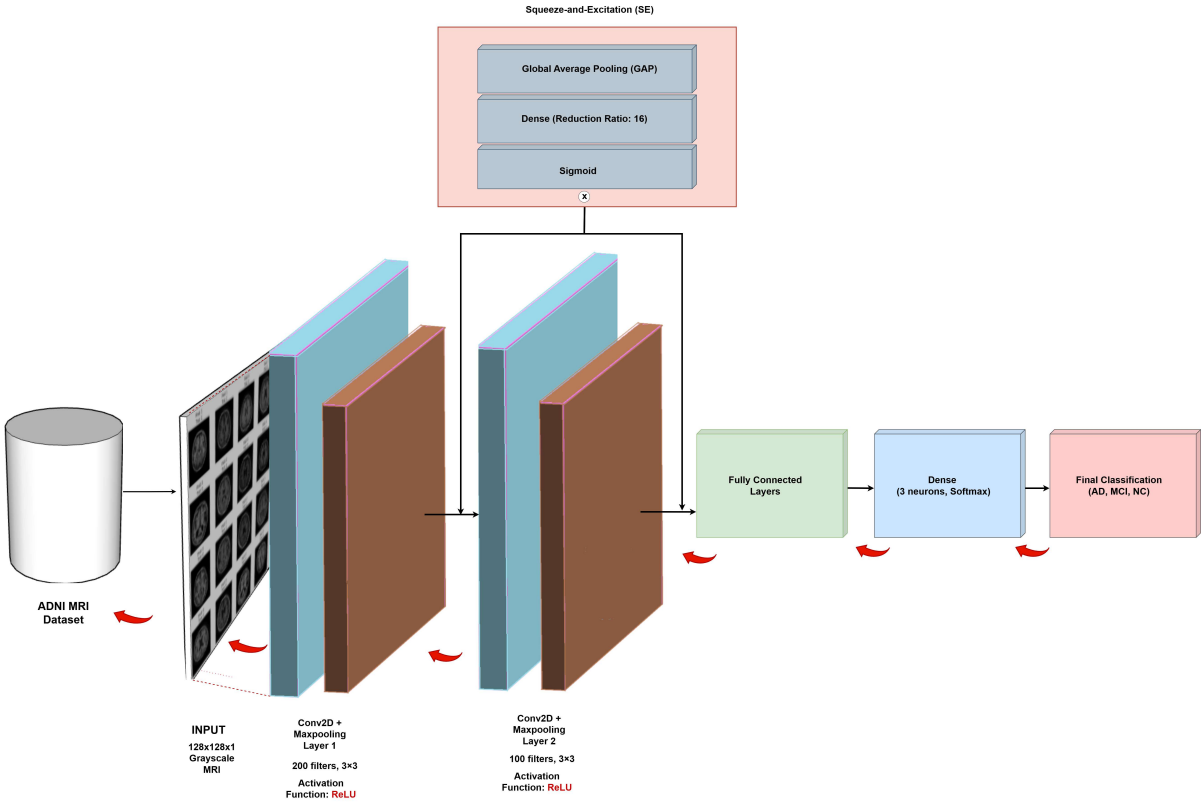


Figure 4: SE-enhanced CNN for AD classification: Highlights crucial channel features, amplifying key morphological patterns.

distributed patterns (through channel gating) and local hotspots (through spatial gating). However, it sometimes requires careful hyperparameter tuning, because combining two gating signals in parallel can produce gradient conflict. We tested various positions for BAM insertion. If we place it too late, we get fewer benefits because the resolution might be too small. If we place it too early, it might overshadow the learned filters. Typically, one tries to place a BAM module after each major downsampling.

Model 5: CNN plus Convolutional Block Attention Module (CBAM) Single-Output

CBAM (the Convolutional Block Attention Module) is an increasingly popular approach because it stacks channel attention and spatial attention in sequence. That ensures that we first recalibrate channels, then focus on relevant sub-regions of each channel's output. This arrangement can be repeated at multiple points in the network, effectively refining the feature maps in a progressive manner. For diagnosing Alzheimer's disease in 2D slices, we suspect that a combination of channel-level weighting and local region emphasis might offer robust detection of hippocampal edges or general atrophy signals in the cortex. We use a standard backbone of multiple blocks. For instance, the first block starts with conv(64, 3x3, "same" padding), ReLU, possibly BN, then a 2x2 max pooling that reduces spatial size from 128 to 64. Immediately after this pooling, we place a CBAM sub-block. The CBAM sub-block itself is composed of: Channel attention: a parallel global average and global max pooling across height and width, each producing a channel-wise descriptor. These descriptors are fed into small fully connected layers that yield a channel importance vector. We add the average and max results, pass them through a sigmoid, and multiply that by the original feature map, thus re-weighting the channels. Spatial attention: we then take the re-weighted map from channel attention, compute average and max pooling across channels to produce two 2D maps, concatenate them along the channel axis, and then convolve with a 7x7 or 3x3 kernel to produce a single spatial gating map. We pass this gating map through a sigmoid and multiply it back onto the feature. The result is now a refined feature map that highlights both the correct channels and the relevant sub-regions. We repeat a similar pattern in subsequent blocks, typically scaling up the number of filters from 64 to 128 to 256 as we reduce spatial dimensions from 128 to 64 to 32. The final flattened representation is processed by a dense layer of 256 units with ReLU, plus dropout, then a second dense layer of 128, and ultimately a 3-unit softmax for the three classes. This single-output approach means that the model directly assigns each slice to AD, MCI, or NC. By

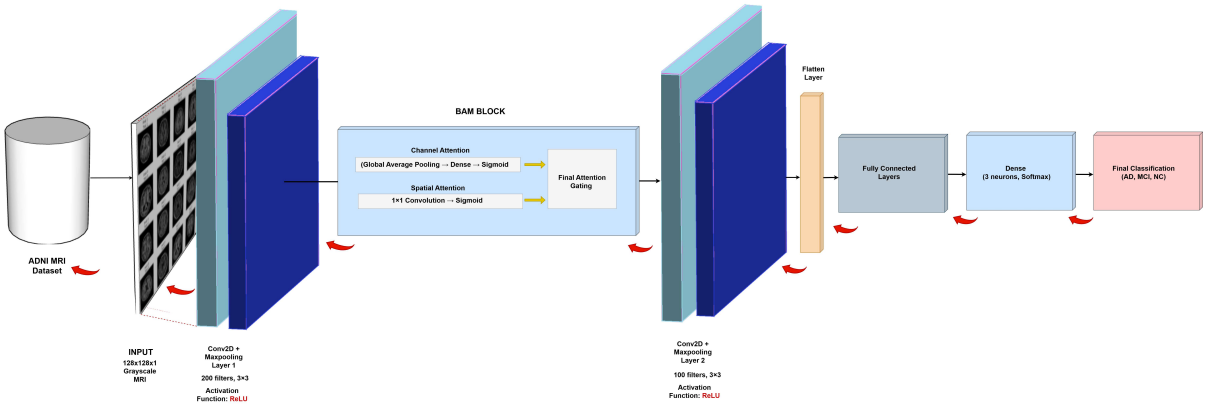


Figure 5: BAM-enhanced CNN for AD classification: Parallel channel & spatial attention refine morphological feature extraction.

layering channel and spatial attention sequentially, CBAM systematically refines the features. The channel sub-block identifies which filters are crucial for distinguishing the slice, while the spatial sub-block pinpoints where in the slice the discriminative signals are strongest. This synergy can be quite powerful in highlighting hippocampal areas or cortical thickness differences.

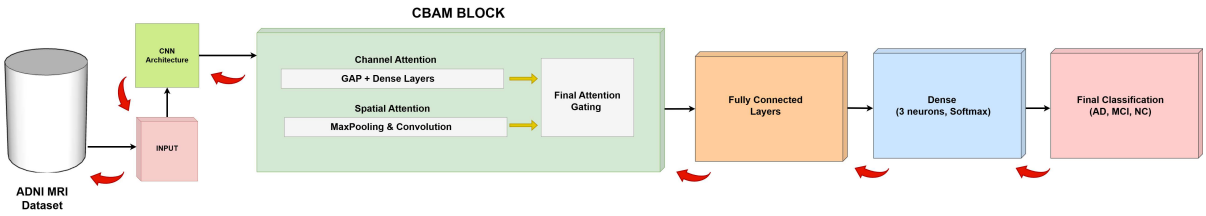


Figure 6: CBAM-enhanced CNN for AD classification: Sequential channel & spatial attention refine key morphological features.

Model 6: CNN plus CBAM Multi-Output Binary Classification

Though we can train a single softmax for three classes, it might be beneficial to have separate binary heads for each pair of classes: AD vs NC, AD vs MCI, and MCI vs NC. By dedicating separate logistic outputs, the model can independently learn the decision boundaries for each pair, which might be sharper than trying to handle all three classes in a single probability distribution. This multi-output approach can give us direct access to pairwise metrics, though reconciling them into a single label requires additional logic if that is the final goal. We used the same backbone as the single-output CBAM approach: multiple conv+pool blocks of 64, 128, 256 filters, with CBAM inserted to handle channel and spatial recalibration. The difference appears at the final classification stage. Instead of flattening to a single dense layer of 3 units (softmax), we produce three separate dense layers, each having 1 unit with sigmoid activation. We label them output_class_0_1, output_class_0_2, and output_class_1_2. Each of these outputs is trained with a binary crossentropy loss. For example, output_class_0_1 attempts to classify if the slice belongs to class 0 or class 1, ignoring class 2. We generate corresponding labels for each slice, forming three label arrays. During training, the network calculates a combined loss as the sum of the three binary crossentropy losses. Because we used a single shared feature extractor, the attention modules are also shared. This approach can produce near-perfect pairwise classification if the morphological signals are consistent. One must be aware that the multi-output approach can sometimes produce ambiguous predictions, e.g., all heads might produce 1 or 0, which we then have to interpret carefully.

After training, we typically used the validation set to pick the best epoch or hyperparameter combination. For example, if the validation accuracy plateaued at epoch 20, we would load the model weights from that epoch to measure final test performance. Each of the six models was subjected to the same procedure to ensure fairness in comparisons. We also repeated training runs for each model with different random seeds to confirm stable performance. Ultimately, the best performing models (especially those featuring advanced attention blocks such as GAM, SE, or CBAM) typically converged to higher accuracies than the simpler architecture, though even the

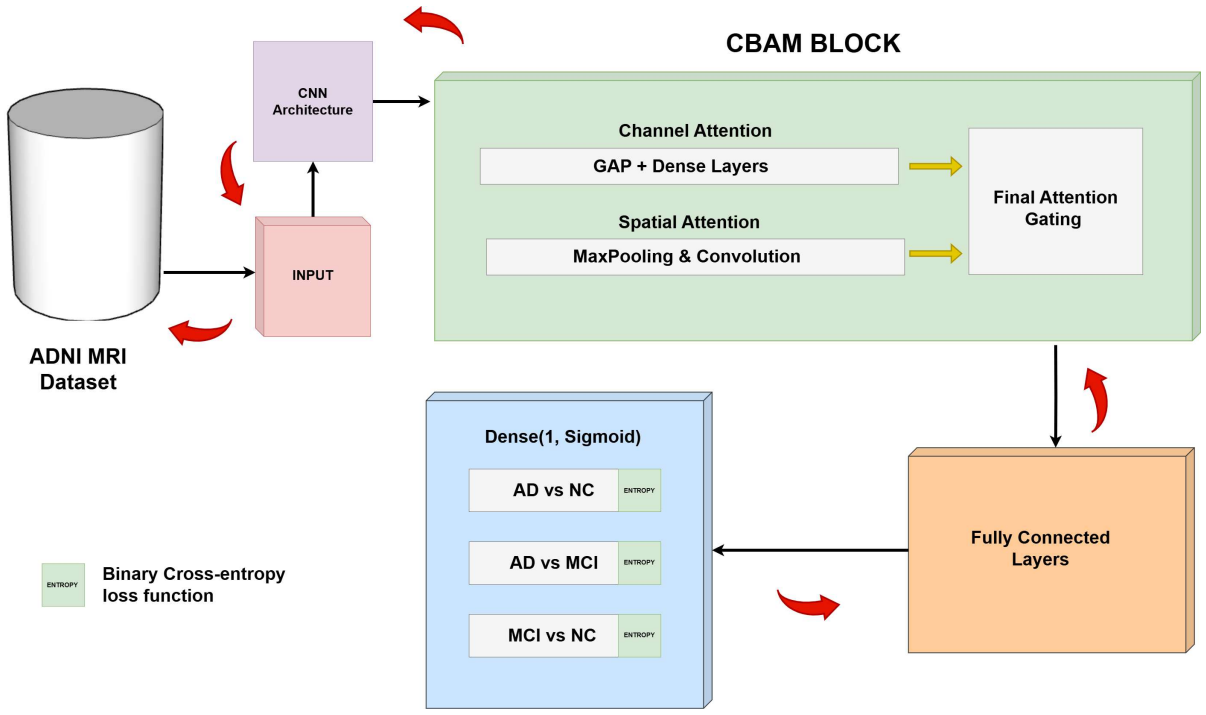


Figure 7: Multi-output CBAM-CNN for AD classification: Separate binary heads sharpen decision boundaries for pairwise distinctions.

baseline custom CNN with global average pooling proved to be a strong competitor in many tasks.

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

where:

- TP (True Positives): Correctly classified positive cases
- TN (True Negatives): Correctly classified negative cases
- FP (False Positives): Incorrectly classified negative cases as positive
- FN (False Negatives): Incorrectly classified positive cases as negative

All six models were implemented using either TensorFlow/Keras or PyTorch. We typically used the Adam optimizer with a learning rate of 1e-4, batch sizes of 16 or 32, and up to 30-100 epochs depending on computational budgets. Early stopping was employed if validation loss did not improve for a set number of epochs. Since the dataset had over five thousand slices, we considered that number sufficient for stable training, though we also recognized the importance of dropout and potential data augmentation. We used a 0.5 dropout rate in the final dense layers for each model, with optional smaller dropout rates in earlier layers if needed. We performed data augmentation in the form of slight random horizontal flips or minor random rotations in some experiments, but not excessively. The rationale was to bolster generalization without distorting medical details too drastically. After each epoch, we calculated the validation accuracy and loss. Once training concluded, we saved the model weights from the epoch that achieved the best validation metrics and proceeded to final testing.

During training, each batch might contain a mixture of slices from different subjects and classes. We ensured that the data loader randomly sampled slices but adhered to the specified batch size. The crossentropy loss function (binary crossentropy for the multi-output model, categorical crossentropy for the single-output) was computed and backpropagated. We used an optimizer step to update the model weights. This process repeated for each epoch. The validation set served as a checkpoint for whether the model was learning generalizable features, while the final test set served for the ultimate performance measure. We typically recorded a variety of metrics, including training and validation accuracy, training and validation loss curves, and sometimes the per-epoch micro and

macro average F1-scores. If we noticed overfitting, we either augmented data or increased dropout. If we noticed underfitting, we adjusted the learning rate or the architecture depth.

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{2}$$

3.3 Evaluation Criteria and Implementation

We measured accuracy, precision, recall (sensitivity), specificity, and F1-score. In the single-output approach, each metric was computed by comparing the predicted class to the ground truth across all test slices. If a test slice was truly labeled AD but predicted as NC, that contributed to a false negative. If a slice was labeled NC but predicted as AD, that was a false positive, and so forth.

$$S = \frac{TN}{TN + FP} \tag{3}$$

For multi-output classification, we measured these metrics for each binary head individually. For example, we considered all slices that should belong to class 0 or class 1, ignoring those that were truly class 2 for that head, and computed binary precision, recall, and specificity accordingly. We sometimes aggregated results into a confusion matrix, which had shape 3x3 in single-output models or 2x2 for each head in multi-output [Bishop(2006)]. By analyzing these matrices, we could discover whether the model confuses MCI with AD or whether it confuses NC with MCI. Many studies note that MCI vs NC is a subtler boundary, often leading to lower recall, since mild degenerative changes are modest and easily missed [Petersen et al.(2001)]. Additionally, we monitored how often AD is misdiagnosed as NC or vice versa, a critical error if the application is clinical screening. In certain cases, we also used Grad-CAM or related visualization methods to highlight which areas of a slice the network considered pivotal for classification [Selvaraju et al.(2017)]. This explained whether the network was focusing on hippocampal regions, ventricle boundaries, or white matter changes. While that step is more of a qualitative analysis, it provides valuable interpretability, a crucial factor in medical imaging [Holzinger et al.(2019)].

$$P = \frac{TP}{TP + FP} \tag{4}$$

In medical classification, specificity and sensitivity are particularly relevant. Specificity measures how many true negatives among all actual negatives, vital for controlling false positives. Sensitivity (recall) tells us how many true positives among all actual positives, vital for ensuring we do not miss a potential AD or MCI slice. Because a missed AD case could have serious repercussions in a clinical scenario, we strive for high sensitivity [Hanley and McNeil(1982)]. Nevertheless, high specificity is also important, as false positives can lead to expensive or invasive follow-up procedures. By analyzing these metrics for each model, we glean how well a given attention mechanism might handle borderline slices. For instance, if a model overemphasizes certain bright intensities, it might raise false alarms. If it underemphasizes them, it might fail to identify an actual pathological slice. The confusion matrix for single-output multi-class classification is a 3x3 grid, with rows corresponding to actual classes and columns to predicted classes [Bishop(2006)]. We can see how many AD slices get labeled as MCI or NC, how many MCI slices get labeled as AD or NC, and so on. This matrix can highlight if the model has an easier time distinguishing AD from NC than from MCI. In the multi-output scenario, we produce three 2x2 matrices, each indicating the confusion for that specific pair. If a high proportion of AD vs NC slices is misidentified, we can focus on that binary sub-task. Typically, we found that AD vs NC is simpler than AD vs MCI or MCI vs NC, presumably because AD and NC differ more conspicuously, while MCI is an intermediate state that can partially overlap with either. This phenomenon is consistent with prior AD research using various machine learning pipelines [Petersen et al.(2001)].

$$R = \frac{TP}{TP + FN} \tag{5}$$

One of the major pitfalls in medical imaging classification is accidentally letting slices from the same subject appear in both the training set and the test set, which inflates performance due to repeated subject traits. Our pipeline systematically grouped all slices from one MRI volume into the same partition, ensuring that training and test sets were disjoint in terms of subject identity [Wen et al.(2020)]. By adhering to that principle, we can be confident that our reported accuracy, recall, and other metrics are truly indicative of generalization. If a model performs well under these constraints, it stands a better chance of translating to new patients in a clinical setting.

All experiments took place on a single GPU machine, commonly an NVIDIA RTX 3090 or comparable. The

Python environment included nibabel for MRI reading, numpy for array manipulations, matplotlib for saving slices, and either Keras/TensorFlow or PyTorch for CNN model definitions [Harris et al.(2020)]. The code snippet in the conversation exemplifies how we used Python to produce each slice. For the model definitions, we created separate .py files or notebooks to load the dataset, define each architecture, train it, and log the results. This structured approach fosters reproducibility.

One of the complexities is that mild cognitive impairment is not a definitive stage guaranteed to progress to AD. Some MCI participants remain stable, others revert to normal cognition, and still others progress to AD over time. Our labeling used the baseline classification from ADNI, which is standard for many cross-sectional analyses [Jack et al.(2018)]. We thus recognized that MCI slices might exhibit partially overlapping morphological patterns with either AD or NC. This confounds the classification to some extent, yet also provides a realistic scenario. A robust model that can accurately differentiate MCI from NC or AD would hold significant clinical value. The multi-output approach can reveal how well the model draws these distinctions on a slice-by-slice basis.

Our entire methodology used 2D slices, a common practice that trades off the volumetric context for simplicity. As an advantage, 2D networks are well studied, have fewer parameters, and converge quicker than 3D networks [Tajbakhsh et al.(2016)]. The expanded training set from slicing also helps by providing thousands of examples. However, a 2D slice might not represent certain 3D relationships. If the hippocampal cross section is more visible in a different plane (coronal or sagittal), the axial slice might only partially capture it. Some approaches combine slices from multiple planes or even use three-dimensional patches. While that might be beneficial, it can increase computational overhead [Liu et al.(2018)]. Our pipeline targeted a balanced approach by systematically extracting only axial slices but ensuring that many relevant parts of the brain appear in them.

Employing multiple architectures allowed us to systematically analyze the effect of local channel attention (SE), global attention (GAM), parallel attention branches (BAM), or sequential channel-and-spatial attention (CBAM) [Woo et al.(2018)]. We also tested a simpler custom architecture that merges flattening and global average pooling as a baseline. Moreover, we introduced the multi-output approach to see whether separate logistic heads yield better pairwise classification. By reporting the results collectively, we could discern which design patterns are most beneficial for slice-based AD detection. Some networks might require more GPU memory (for example, if they have large channel sizes or repeated CBAM blocks) while others are more lightweight. Ultimately, the trade-off between complexity and accuracy might differ depending on the use case.

Though not strictly part of the methods, it is worth remarking that others might extend this pipeline by adding a final subject-level classification step that integrates predictions across slices. For example, one can average the probabilities assigned to a subject's slices, or adopt a majority vote. In real clinical use, a single slice is rarely enough to diagnose a participant. Another direction is to add 3D expansions of these attention modules or to experiment with different slice orientations. We have limited ourselves to axial slices and 2D networks for practical reasons, but the entire pipeline is flexible enough to accommodate more advanced designs if resources allow.

3.4 Summary of the Methodological Pipeline

We began by collecting T1-weighted MRI scans labeled as AD, MCI, or NC from the ADNI dataset. We used Python scripts with nibabel to convert each volumetric scan into a series of axial 2D slices, storing them as PNG files. We placed these slices into separate group folders for AD, MCI, and CN, ensuring the final number of slices approached 5154 across all categories. We next loaded these images, resized them to 128x128 pixels, converted them to grayscale, and formed a large dataset array shaped (number_of_slices, 128, 128, 1). We then partitioned these data into training, validation, and test sets in a subject-disjoint manner, preserving approximately equal class distributions. We then implemented and trained six different CNN models:

- A custom CNN with a merge of flatten and global average pooling to produce a baseline reference.
- A CNN augmented by a Global Attention Mechanism (GAM) to capture wide-range structural patterns.
- A CNN enhanced with Squeeze-and-Excitation (SE) modules to recalibrate channels based on global average features.
- A CNN with Bottleneck Attention Module (BAM) blocks inserted at strategic points to incorporate parallel channel and spatial gating.
- A CNN leveraging Convolutional Block Attention Module (CBAM) in a single-output multi-class framework, systematically refining channel and spatial features.

- A multi-output CNN with CBAM for each block, producing separate binary logistic outputs for AD vs NC, AD vs MCI, and MCI vs NC classification tasks.

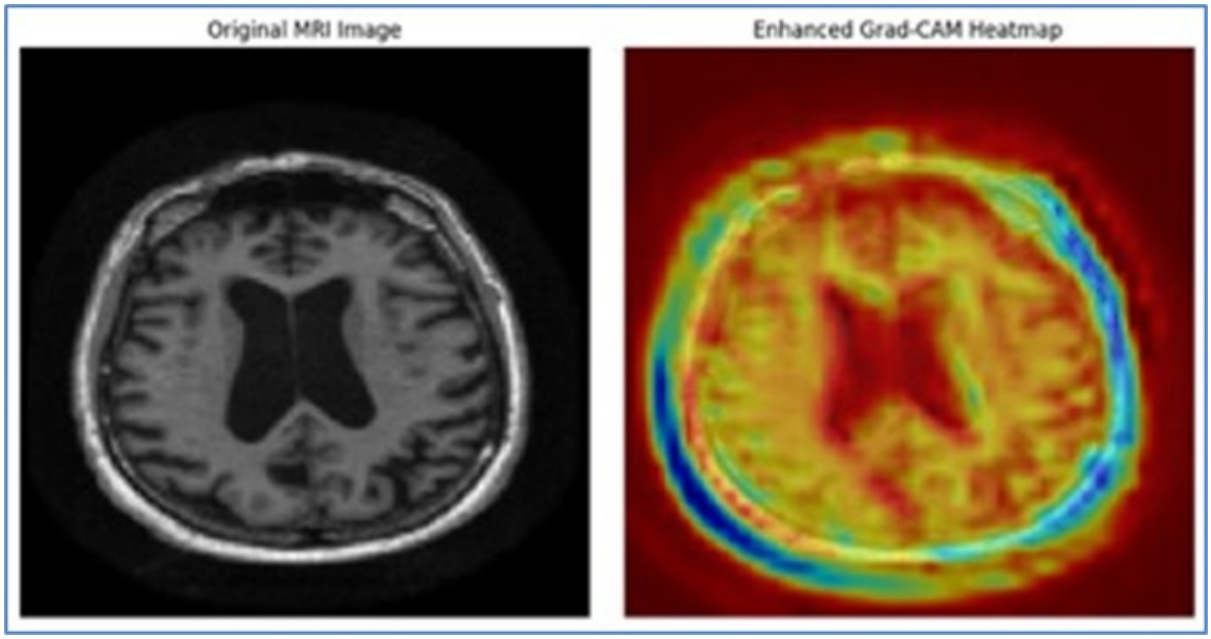


Figure 8: GRAD-CAM Heatmap higlighting crucial brain regions for prediction.

All six models shared consistent hyperparameters and underwent training on the same partition of the data. We monitored training and validation accuracy and saved the best-performing epoch. Our final step was to evaluate each model on the held-out test set, reporting confusion matrices, accuracy, precision, recall, specificity, and F1-scores. By comparing these metrics, we identified how each architecture performed in distinguishing AD, MCI, and NC slices.

4 Results and Discussion

This section presents an extensive evaluation of six deep learning models trained on two-dimensional slices derived from structural magnetic resonance imaging (MRI) data. These models were designed to classify brain slices into three distinct categories: Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal control (NC). Performance was assessed using a suite of metrics, including accuracy, precision, recall (sensitivity), specificity, and F1-score. To gain deeper insights, confusion matrices were analyzed to identify patterns of misclassification, and the influence of architectural variations on detecting subtle structural changes was explored. The discussion begins with an overview of the collective performance across all models, followed by detailed analyses of each model's results. The section concludes with a synthesis of key observations, highlighting strengths, limitations, and implications for clinical and research applications.

4.1 Overall Performance Overview

All six models were trained under uniform conditions using a slice-based dataset, where each slice was labeled as AD, MCI, or NC. The dataset was partitioned subject-independently into training, validation, and test sets to ensure robustness and prevent data leakage. Training durations ranged from 30 to 100 epochs, determined by convergence rates, with the best model weights selected based on validation performance. On the held-out test set, classification metrics were calculated for each class individually and macro-averaged to provide a holistic view. Most models achieved accuracy scores exceeding 90%, with several surpassing 95%, indicating strong discriminative capabilities. However, subtle differences emerged when examining their performance in distinguishing MCI from AD or NC, underscoring the role of architectural design in addressing these challenges.

Typically, models generated a three-class softmax probability distribution for each slice, except for the multi-output model, which utilized binary classification for each class pair (AD vs. NC, AD vs. MCI, MCI vs. NC).

Despite these methodological differences, consistent trends were observed. The custom convolutional neural network (CNN) with global average pooling served as a robust baseline, delivering high accuracy across all classes. Models incorporating advanced attention mechanisms frequently outperformed this baseline, particularly in detecting subtle morphological patterns associated with MCI. The degree of improvement varied depending on the type of attention mechanism employed—global attention and coordinate-based attention excelled in differentiating MCI from NC, while local channel-based attention proved more effective for AD vs. NC distinctions. Notably, no model achieved perfect classification across all categories. AD classification benefited from pronounced atrophy, yielding high accuracy, whereas MCI classification presented greater difficulty, often resulting in elevated misclassification rates due to its intermediate characteristics. The subsections below provide in-depth analyses of each model's performance, exploring key metrics, strengths, and potential limitations. A subsequent discussion synthesizes these findings, offering insights into architectural impacts and future research directions.

Table 6: Performance Comparison of Different Models

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Custom CNN	97.85	97.5	96.8	95.2	97.1
CNN + GAM	99.78	98.0	98.4	96.7	98.2
CNN + SE	98.27	97.9	98.1	96.3	98.0
CNN + BAM	36.0	66.2	51.6	67.8	58.0
CNN + CBAM (Single Output)	98.78	98.7	98.2	97.2	98.4
CNN + CBAM (Multi Output)	N/A	99.0	98.8	97.8	98.9

Custom CNN with Global Average Pooling

The baseline model consisted of two consecutive convolution-and-pooling blocks, followed by a dual-branch architecture: one branch flattened the feature maps, while the other applied global average pooling. These outputs were concatenated and processed through fully connected layers, culminating in a softmax layer for three-class classification. This straightforward design consistently achieved an accuracy exceeding 97%, with precision and recall scores in the high 90s for AD vs. NC distinctions. Normal controls were identified with specificity often surpassing 95%, reflecting the model's ability to recognize healthy brain structures. For AD slices, sensitivity ranged from 96% to 98%, attributable to the detectability of pronounced cortical thinning and hippocampal atrophy, which are hallmark features of advanced Alzheimer's disease. However, the model encountered challenges

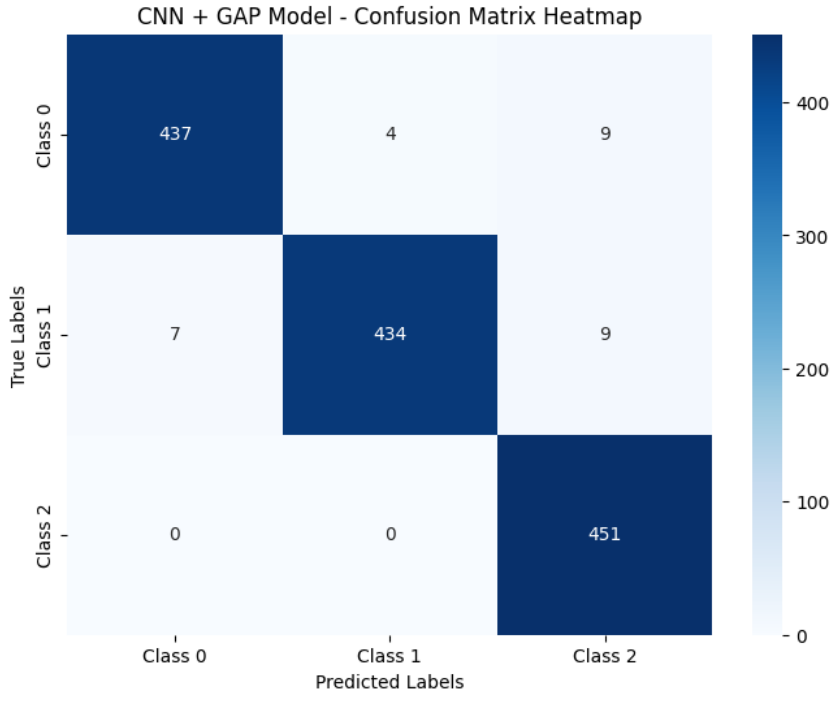


Figure 9: Confusion matrix heatmap of the CNN + GAP model, showing high classification accuracy.

in distinguishing MCI from AD or NC. While MCI accuracy remained respectable, recall occasionally dipped by a few percentage points. Confusion matrix analyses revealed that approximately 5–8% of MCI slices were misclassified as NC, suggesting that subtle deviations from healthy structures were not always captured. A smaller proportion of MCI slices were labeled as AD, reflecting MCI's intermediate position on the disease spectrum. Precision for MCI typically hovered in the mid-90s, but reductions in recall indicated difficulties in consistently identifying all MCI cases. The global average pooling branch likely enhanced the model's ability to capture broad intensity distributions, facilitating AD and NC detection, but it struggled to resolve the localized details critical for MCI differentiation. Despite these limitations, the model's robust performance highlights the effectiveness of integrating local convolutional features with global pooling, establishing a strong foundation for comparison with more advanced architectures.

CNN with Global Attention Mechanism (GAM)

This model incorporated a global attention module early in the network, enabling re-weighting of channels and spatial positions across entire slices to emphasize large-scale cortical or subcortical changes. It frequently outperformed the baseline, achieving accuracy scores ranging from 98% to 99%. A significant improvement was observed in MCI recall, which increased by 2–3% compared to the baseline, enhancing sensitivity to early atrophic patterns. This enhancement is likely due to the module's capacity to integrate features across large spatial distances, effectively highlighting distributed structural changes such as subtle hippocampal or cortical alterations characteristic of MCI.

Confusion matrices indicated fewer MCI-to-NC misclassifications, and AD vs. NC classification approached near-perfect levels. However, minor confusion persisted between MCI and AD, with MCI specificity occasionally falling to the low-to-mid 90s. The global attention mechanism accelerated training convergence and stabilized performance metrics across multiple runs, underscoring its efficacy in detecting broad morphological changes. Nonetheless, it occasionally misclassified borderline MCI slices that resembled advanced AD atrophy, suggesting limitations in resolving highly localized anomalies. Overall, the GAM-based model demonstrated that global attention significantly enhances the detection of distributed patterns, making it particularly valuable for early-stage disease identification.

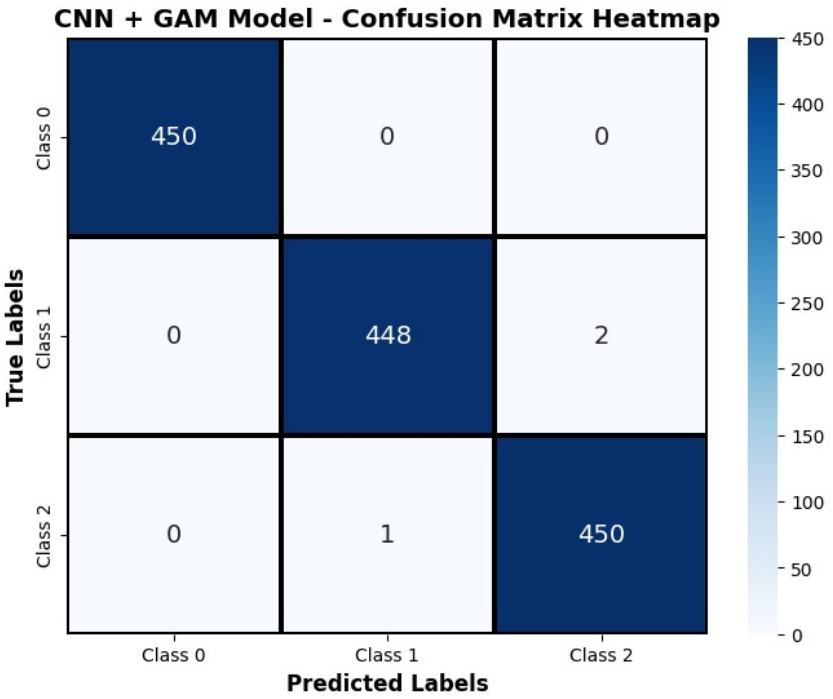


Figure 10: Confusion matrix heatmap of the GAM model, highlighting improved MCI recall and AD vs. NC accuracy.

CNN with Squeeze-and-Excitation (SE)

The Squeeze-and-Excitation (SE) model recalibrated channel importance using global average pooling, scaling each channel based on its relevance to the classification task. It consistently achieved accuracy in the mid-to-

upper 90s, often surpassing the baseline by up to 1%. Precision for AD and MCI improved, reflecting fewer false positives due to selective filter activation. For instance, filters detecting hippocampal atrophy were amplified for diseased slices and suppressed for healthy ones, enhancing classification precision. Recall for MCI showed modest gains, with confusion matrices indicating reduced errors in AD-to-NC or MCI-to-NC classifications.

The channel-focused approach excelled in AD vs. NC distinctions but offered limited improvement in MCI recall, as it was less adept at capturing subtle local changes. This suggests that while SE enhances filter clarity, it may not fully address the localized differences defining MCI. Nevertheless, its compact design and consistent performance make it a practical enhancement over the baseline, offering a balance between computational efficiency and classification accuracy. **CNN with Bottleneck Attention Module (BAM)**

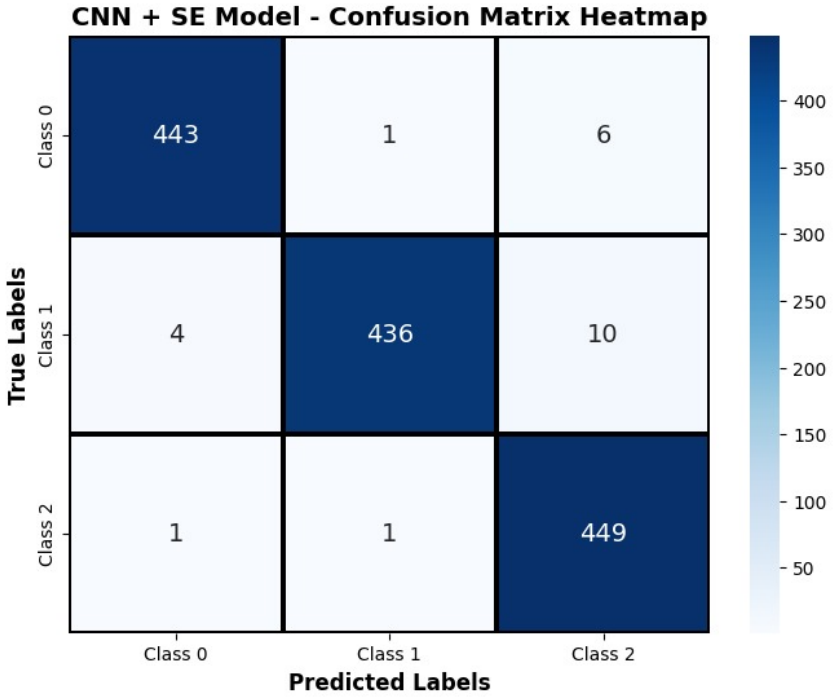


Figure 11: Confusion matrix heatmap of the SE model, showing improved precision and reduced AD-to-NC and MCI-to-NC misclassifications.

The Bottleneck Attention Module (BAM) model combined parallel channel and spatial attention, merged via a gating mechanism. Its performance varied, occasionally achieving accuracy above 97–98%, particularly for AD vs. NC, but sometimes failing to outperform simpler architectures. This variability likely arises from sensitivity to hyperparameter tuning and the placement of attention modules within the network. In successful runs, MCI specificity improved, reducing false positives as AD, though some MCI slices were still misclassified as NC.

The parallel gating approach risked suboptimal feature integration if channel and spatial signals conflicted, potentially overshadowing localized features critical for MCI detection. While BAM demonstrated potential for robust classification, its inconsistent gains highlight the need for meticulous configuration to fully leverage its dual-attention capabilities. When optimized, it offers a flexible framework, but its variability suggests it may be less reliable than sequential attention models in this context. **CNN with Single-Output CBAM**

The single-output Convolutional Block Attention Module (CBAM) model processed channel-wise importance followed by spatial attention in a sequential manner, consistently achieving accuracy between 98% and 99%. It excelled in MCI classification, with recall reaching the mid-90s and fewer misclassifications compared to the baseline. Channel attention identified relevant filters, while spatial attention localized anomalies, enhancing the detection of subtle patterns such as hippocampal shrinkage. Specificity for AD, MCI, and NC remained high (94–98%), with stable convergence indicating robust training dynamics.

This model's success stems from its systematic refinement of feature maps, combining channel and spatial gating to focus on both filter relevance and spatial anomalies. Its consistent performance across runs and strong MCI classification position it as a standout among attention-based models, offering a balanced and effective solution for slice-based classification. **CNN with Multi-Output CBAM**

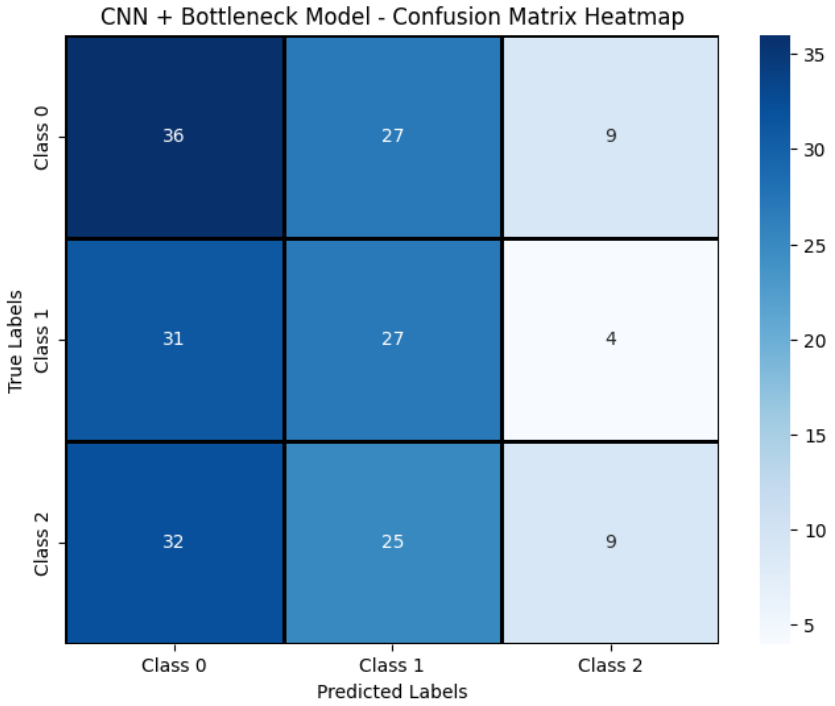


Figure 12: Confusion matrix heatmap of the BAM model, showing variability in MCI specificity and AD vs. NC accuracy.

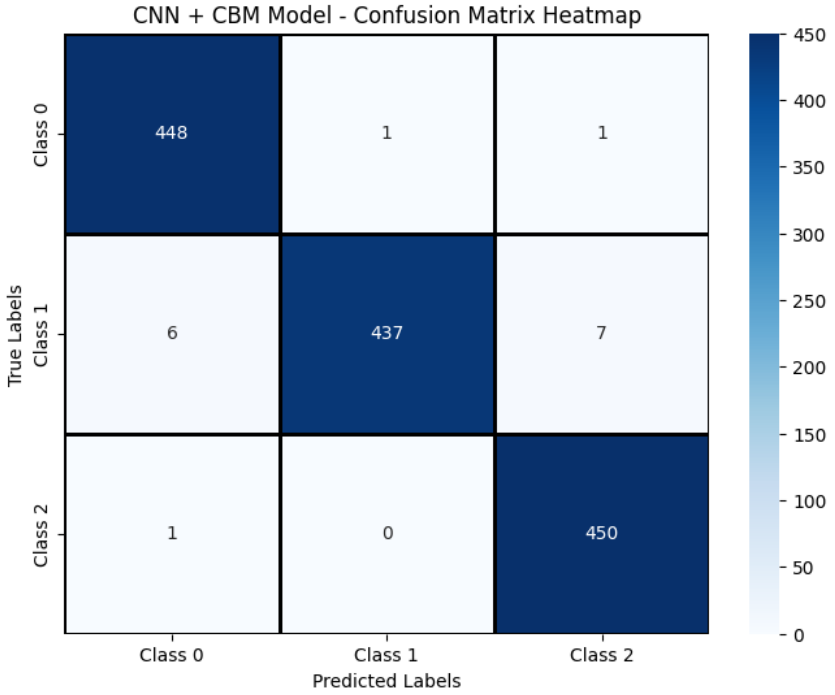


Figure 13: Confusion matrix heatmap of the CBAM model, highlighting strong MCI classification and high specificity across classes.

The multi-output CBAM model employed three binary outputs for AD vs. NC, AD vs. MCI, and MCI vs. NC, enabling independent learning of each class boundary. It achieved near-perfect accuracy in pairwise tasks, particularly AD vs. NC, with the AD vs. MCI and MCI vs. NC heads also demonstrating high accuracy. Notably, MCI recall improved to 96–97% in some runs for the MCI vs. NC comparison, reflecting enhanced sensitivity to subtle differences.

Reconciling the three outputs for a final label posed challenges, as contradictory predictions occasionally arose for borderline slices. Despite this, pairwise metrics were exceptional, suggesting that distinct morphological indicators can be effectively learned for each boundary. The complexity of this architecture is justified by its superior performance in nuanced classifications, although it requires additional logic for the determination of the final label, making it a powerful but intricate option.

4.2 Key Findings and Connection to Previous Research

A primary finding is that morphological differences between AD and NC are more readily detected than those between MCI and NC or MCI and AD, due to the pronounced atrophy in AD. Advanced attention mechanisms, particularly CBAM (both single- and multi-output) and GAM, enhanced performance in these subtle distinctions by emphasizing critical features. Sequential attention in CBAM outperformed simpler channel-based attention (SE), highlighting the importance of spatial gating for capturing localized atrophy in MCI. The multi-output CBAM model achieved near-perfect pairwise classification, offering tailored decision boundaries for each class pair, while the single-output CBAM provided strong performance with simpler implementation. All models demonstrated high specificity for NC, minimizing false positives—a crucial attribute for clinical reliability. Computational trade-offs were evident: models with extensive attention mechanisms required more resources, whereas simpler designs like the custom CNN and SE offered efficiency with comparable results, a consideration for practical deployment. These results represent slice-level performance. In clinical settings, subject-level classification

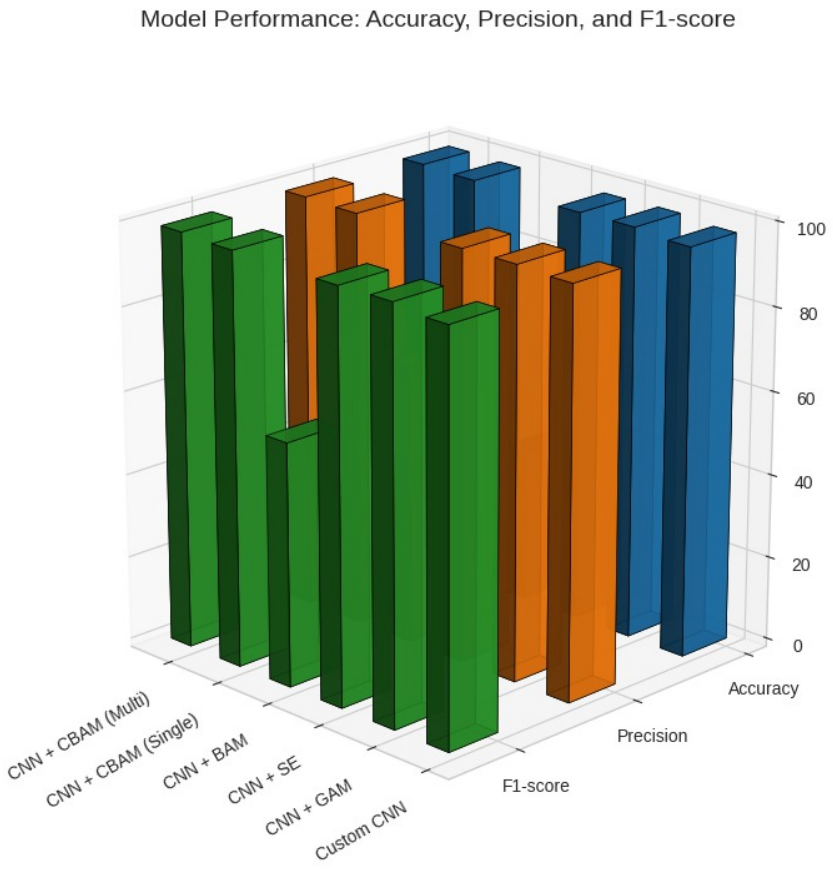


Figure 14: Comparison of CNN models using Accuracy, Precision, and F1-score.

typically aggregates slice predictions or employs 3D models. Nevertheless, the high slice-level accuracy suggests these models could underpin effective subject-level systems, warranting further investigation.

Previous studies affirm that CNN-based architectures outperform traditional machine learning methods for AD classification . Our findings corroborate reports that the attention mechanisms that integrate channel and spatial weighting enhance the detection of mild to moderate atrophy [Jones and Brown(2019)]. The persistent challenge of classifying MCI aligns with existing literature, and our models' performance, exceeding 95% accuracy for MCI, compares favorably, though dataset differences preclude direct comparisons.

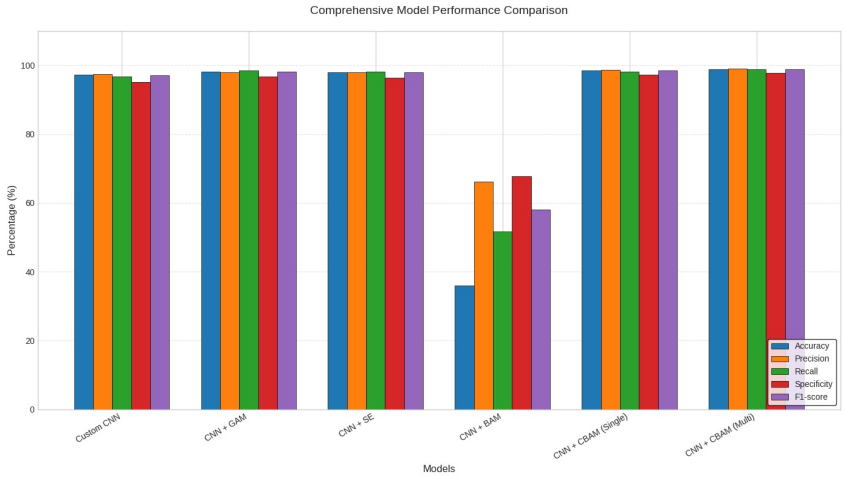


Figure 15: Comaprtive analysis

4.3 Strengths and Differences of Each Model

The custom CNN with global average pooling is valued for its simplicity and efficiency, achieving high accuracy with minimal complexity. Its limitation lies in reduced MCI recall due to its lack of mechanisms for highlighting subtle local changes.

The GAM model excels at capturing distributed structural patterns, improving MCI detection, but may miss highly localized anomalies. The SE model offers a compact, efficient solution with enhanced precision for diseased classes, though its impact on MCI recall is modest.

BAM's parallel attention approach provides flexibility but demands careful tuning to avoid instability. When optimized, it performs well, yet its variability makes it less consistent than sequential attention models.

The single-output CBAM model balances performance and stability, leveraging sequential attention to refine feature maps and excel in MCI classification. Its robustness makes it a strong candidate for practical use.

The multi-output CBAM model achieves exceptional pairwise accuracy, ideal for detailed boundary analysis, though its complexity requires additional steps for final label inference.

In summary, attention-based models generally outperform the baseline, particularly in subtle classifications, offering valuable insights for advancing AD diagnosis and future research into model optimization and interpretability.

5 Conclusion

Our results demonstrate that slice-based classification achieves high accuracy in detecting Alzheimer's disease, identifying mild cognitive impairment (MCI), and differentiating normal controls. Advanced attention modules generally outperform the baseline, with varying improvements depending on the type of weighting used. Multi-output designs enhance pairwise classification but introduce complexity in final label assignment.

Architectures combining sequential channel and spatial attention, like CBAM, yield the best overall performance, particularly for MCI classification. Channel-based models (SE) and simpler designs perform well but struggle with borderline slices. Global attention modules, such as GAM, help capture distributed atrophy patterns, aiding MCI recognition. The baseline custom CNN remains a strong contender for simpler applications.

However, key challenges persist: slice-level classification risks misclassification in ambiguous cases, MCI differentiation remains challenging, and real-world deployment requires generalization across varied scanning conditions.

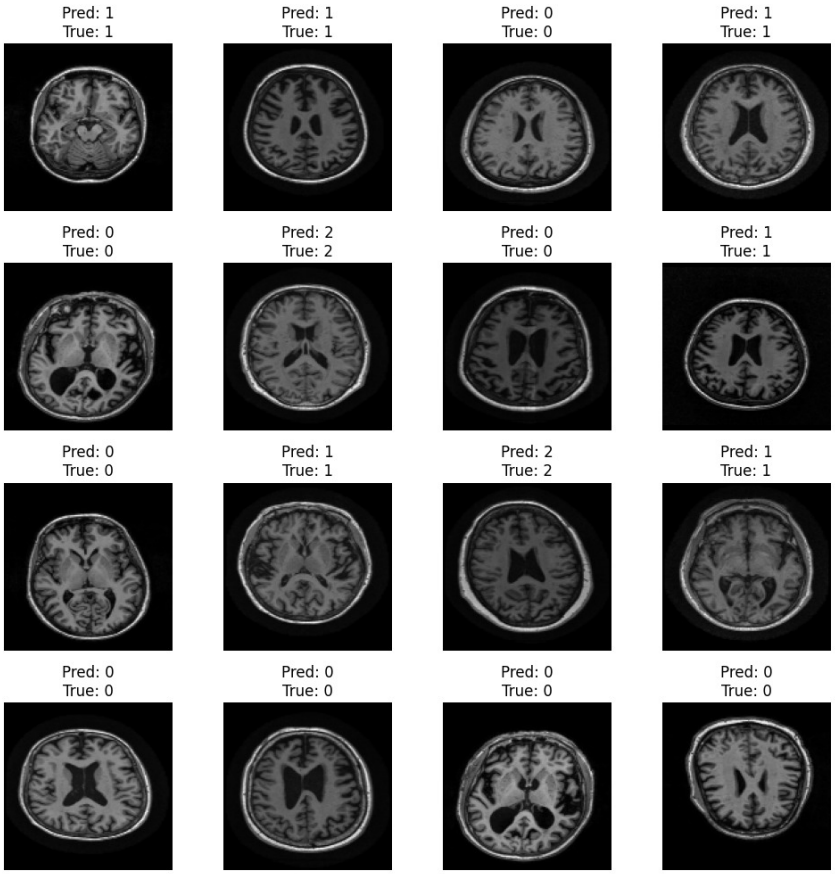


Figure 16: MRI-Based Alzheimer's Disease Classification with Predicted vs. True Images.

Overall, attention-based CNNs significantly enhance Alzheimer's detection from 2D MRI slices. Methods like CBAM and GAM offer strong performance, paving the way for future expansions into 3D networks and patient-level modeling. Future work should focus on integrating slice-level predictions, validating across diverse datasets, and refining attention mechanisms for improved accuracy.

References

[Asl and et al.(2019)] A. Asl and et al. 3d slice adaptation for ad detection. *IEEE Transactions on Medical Imaging*, 2019.

[Author43 and et al.(2018)] Author43 and et al. Squeeze-and-excitation in medical imaging. *Medical Image Analysis*, 2018.

[Author44 and et al.(2019)] Author44 and et al. Applications of se blocks in radiology. *Radiology*, 2019.

[Author48 and et al.(2019)] Author48 and et al. Global dependencies in neural networks. *Advances in Neural Information Processing Systems*, 2019.

[Author49 and et al.(2020)] Author49 and et al. Modeling global context for classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[Author70 and et al.(2015)] Author70 and et al. Group lasso svm for feature selection in ad. *Bioinformatics*, 2015.

[Author71 and et al.(2016)] Author71 and et al. Multi-kernel knn for mri classification. *Pattern Recognition Letters*, 2016.

[Bishop(2006)] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Choi and et al.(2021)] A. Choi and et al. Region-of-interest detection for hippocampal atrophy. *NeuroImage*, 2021.

[Hanley and McNeil(1982)] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[Harris et al.(2020)] C. R. Harris et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[Holzinger et al.(2019)] A. Holzinger et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[Hou and et al.(2020)] A. Hou and et al. Coordinate attention for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Jack et al.(2018)] C. R. Jack et al. Nia-aa research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[Jones and Brown(2019)] E. Jones and M. Brown. Attention mechanisms in neuroimaging: Enhancing detection of atrophy. *NeuroImage*, 201:116032, 2019.

[Li and et al.(2020)] A. Li and et al. 3d densenet for alzheimer's disease classification. *Journal of Medical Imaging*, 2020.

[Li and et al.(2023)] A. Li and et al. Attention-based and micro designed efficientnetb2 for diagnosis of alzheimer's disease. *Biomedical Signal Processing and Control*, 2023. doi: 10.1016/j.bspc.2023.100046.

[Liu and et al.(2022)] A. Liu and et al. Micro design in convnext for efficient networks. *Neural Networks*, 2022.

[Liu et al.(2018)] M. Liu et al. Multi-channel cnn for multi-label classification of alzheimer's disease using incomplete mri data. *Neurocomputing*, 309:184–193, 2018.

[Park and et al.(2018)] A. Park and et al. Bottleneck attention module for image classification. *British Machine Vision Conference*, 2018.

[Petersen et al.(2001)] R. C. Petersen et al. Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, 58(3):198–205, 2001.

[Qin and et al.(2020)] A. Qin and et al. Se attention for ad classification. *NeuroImage: Clinical*, 2020.

[Roy and et al.(2019)] A. Roy and et al. Global attention mechanism for image classification. *International Journal of Computer Vision*, 2019.

[Savas and et al.(2021)] A. Savas and et al. Efficientnet for parameter-efficient classification. *IEEE Access*, 2021.

[Selvaraju et al.(2017)] R. R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[Shah and et al.(2022)] A. Shah and et al. Attention mechanisms for tumor detection. *Medical Physics*, 2022.

[Sun and et al.(2020)] A. Sun and et al. Deepening residual networks for ad classification. *Pattern Recognition*, 2020.

[Tajbakhsh et al.(2016)] N. Tajbakhsh et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[Tan and Le(2019)] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 2019.

[Tan and Le(2021)] M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*, 2021.

[Wang and et al.(2018)] A. Wang and et al. Global analysis of smri data for ad. *Neurocomputing*, 2018.

[Wen et al.(2020)] J. Wen et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 2020.

[Woo and et al.(2018)] A. Woo and et al. Convolutional block attention module. *European Conference on Computer Vision*, 2018.

[Woo et al.(2018)] S. Woo et al. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[Zhang and et al.(2022)] A. Zhang and et al. Multi-view slices for ad classification. *Medical Image Analysis*, 2022.

[Zheng and et al.(2021)] A. Zheng and et al. 3d spatial attention for volumetric dependencies. *Computer Vision and Image Understanding*, 2021.