Attention-Enhanced 2D CNN for Multi-Class Alzheimer's Disease Classification with Grad-CAM Visualization

M. I. Omogbhemhe*, O. M. Odighi

Department of Computer Science, Ambrose Alli University, Ekpoma, Edo State, Nigeria.

ABSTRACT: Obtaining accurate and timely diagnosis of Alzheimer's disease remains a critical challenge in clinical neuroscience, particularly during the early stages (mild and moderate), when significant impacts can still be achieved with timely intervention. In this study, we propose a deep learning framework based on a 2D convolutional neural network (ResNet-18) enhanced with a Convolutional Block Attention Module (CBAM) for multi-class classification of Alzheimer's disease stages from brain MRI slices. To address the pronounced class imbalance inherent in the dataset, a class-weighted loss function was combined with aggressive online data augmentation during training. The model was evaluated on a public dataset comprising over 86,000 MRI slices spanning four diagnostic categories, which include Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia. Quantitative evaluation demonstrated a slice-level classification accuracy of 93.7%, with robust performance across all classes, including the underrepresented dementia stages. Interpretability was provided through Grad-CAM visualizations, which highlighted anatomically meaningful regions consistent with established Alzheimer's disease biomarkers. While these results underscore the potential of attention-augmented 2D CNNs for automated classification of dementia stages, the framework has only been validated on a single public dataset and has not yet been tested on independent cohorts. Consequently, this study should be considered a proof-of-concept, and further multi-center, out-of-distribution validation is needed before any clinical deployment can be contemplated.

KEYWORDS: Alzheimer's Disease, Attention Mechanism, Brain MRI, Deep Learning, Medical Image Classification,

[Received Jul. 17, 2025; Revised Feb. 02, 2026; Accepted Feb. 07, 2026] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

According to Anwal (2021), Alzheimer's Disease (AD) is a progressive and irreversible neurodegenerative disorder that primarily impairs memory, cognition, and behavior, that interfere with the ability to perform everyday activities. It is the most common form of dementia, accounting for approximately 60–80% of all cases worldwide (Alzheimer's Association, 2023; Chaudhary et al., 2024). The global impact of AD is increasing, with projections suggesting that over 150 million people could be affected by 2050, placing a substantial burden on healthcare systems, caregivers, and society as a whole (Nandi et al., 2022). The disease is marked by the buildup of amyloid-beta plaques and neurofibrillary tangles in the brain, causing neuronal loss and brain atrophy, especially in the hippocampus and other medial temporal lobe structures that are critical for memory and cognitive function (Nandi et al., 2022).

Clinically, AD progression is typically described in three stages: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Dementia (AD) (Davis et al., 2018). Molinuevo et al., (2010) stated that the amnestic subtype of MCI (aMCI) is considered a prodromal phase and represents a crucial window for early intervention. Due to overlapping symptoms, individual variability, and the gradual progression of neuro degeneration, it is challenging to accurately distinguish between these stages (Mravinacová et al., 2025). Early and accurate staging is therefore essential for timely diagnosis, personalized treatment planning, patient selection for clinical trials, and overall quality of life improvement (Di Meco & Vassar, 2021).

Dang et al. (2023) noted that Magnetic Resonance Imaging (MRI), particularly structural MRI (sMRI), has become a key non-invasive tool for monitoring anatomical changes associated with AD. High-resolution T1-weighted MRI can detect early structural atrophy in critical brain regions, including the hippocampus, entorhinal cortex, precuneus, and posterior cingulate gyrus (Xie et al., 2019). These areas serve as key biomarkers according to the National Institute on Aging and the Alzheimer's Association (NIA-AA) diagnostic criteria. Compared to cerebrospinal fluid analysis or Positron Emission Tomography (PET), which can be disruptive or costly, MRI presents a safer and more convenient choice or alternative for large-scale screening and longitudinal monitoring (Jack et al., 2019). Leveraging MRI data for stage-specific assessment of AD thus facilitate integration of computational intelligence into clinical settings (Lin, 2025).

*Corresponding author: mikeizah@gmail.com

doi: http://doi.org/10.63746/njtd.v23i2.3831

Recent deep learning development, facilitate extracting complex spatial patterns from medical images that help to achieve superior diagnostic performance compared to traditional machine learning strategies. However, most deep learning models operate as "black boxes," offering little transparency into how decisions are made which is a tangible barrier to clinical adoption. In the context of AD classification, interpretability is important to foster trust, ensure accountability, and meet the demands of human-centered medical decision-making (Ennab, 2025).

This study proposes an interpretable deep learning framework for multi-class classification of AD stages using 2D brain MRI slices. By employing lightweight 2D convolutional neural networks (CNNs), the model retrieves diagnostic insights from axial, sagittal, and coronal slices efficiently. To enhance interpretability, attention mechanisms and post-hoc visualization techniques, such as Grad-CAM, highlight brain regions that influence the model's predictions. This design optimizes both accuracy and explainability, supporting clinician trust and upholding ethical AI principles in healthcare (Tuan, 2024). Rigorous examination on publicly available datasets such as ADNI and OASIS demonstrates robust performance across the different stages of AD. Adding interpretability modules offers qualitative understanding and knowledge into stage-specific anatomical changes which further boost the model's relevance to clinical practice.

This study represents a step towards deployable, interpretable AI systems for diagnosing neurodegenerative disorders. While early studies in AD imaging primarily focused on binary classification (distinguishing AD patients from cognitively normal individuals), more recent efforts have moved toward multi-class classification, encompassing CN, MCI, and AD stages. This advancement offers clinical benefits but poses methodological hurdles due to subtle inter-class differences and class imbalance. AD stage classification often employs CNN-based approaches. For example, Basaia et al. (2019) applied deep CNNs to structural MRI scans and achieved high accuracy in differentiating CN, MCI, and AD patients. Yet, using full 3D volumes can be computationally intensive and prone to overfitting, particularly given the limited availability of labeled neuroimaging data.

Addressing these issues, slice-based 2D approaches have gained popularity, lowering computational demand while preserving key anatomical details (Spasov et al., 2019). Jain et al. (2019) demonstrated that slice-level CNNs using axial MRI images could outperform 3D models with shorter training times, often pooling slice predictions via majority voting or attention mechanisms. Li et al. (2018) further showed that combining multiple views—axial, sagittal, and coronal—provides richer spatial representation and improves classification accuracy. Despite strong performance, deep learning model's limited explainability undermines clinical confidence and implementation. Techniques like Grad-CAM and layer-wise relevance propagation (LRP) have been used to visualize brain regions driving predictions, partially validating the model's outputs against known neuroanatomical markers of AD, such as hippocampal atrophy and ventricular enlargement. More recent methods embed interpretability directly into the network architecture. For instance, Shaikh et al. (2025) incorporated attention modules to highlight disease-relevant features, enhancing both accuracy and explainability, while Li et al. (2025) combined radiomics-based features with deep learning in a dual-branch network for improved interpretability. Another line of research addresses label noise and class imbalance in multi-class AD staging. Techniques such as transfer learning, generative adversarial networks (GANs) for synthetic data augmentation, and other data-augmentation strategies have been used to enhance model generalizability. For example, Lu et al. (2018) utilized GAN-based pretraining followed by fine-tuning on ADNI slices to improve sensitivity in distinguishing MCI from CN and AD. Public datasets like ADNI, OASIS, and AIBL provide longitudinal, expert-labeled MRI scans that facilitate reproducibility and benchmarking. However, challenges remain in generalizing models across populations due to domain shifts, scanner variability, and demographic biases, underscoring the need for more diverse training data and evaluation strategies.

In summary, while deep learning has shown great promise for automated AD stage classification, combining interpretability, multi-view 2D slice learning, and computational efficiency remains an active area of research. This work contributes to this field by presenting an interpretable, lightweight 2D deep learning framework for multi-class AD classification, incorporating attention-based visualization to enhance clinical trust and understanding. In this paper, section I discussed Alzheimer disease and different research aimed at improving its early diagnosis and the different methodologies adopted. Section II discussed the methodology adopted in this paper and the data training, experiments and evaluation were also carried out in this section. Section III presented the qualitative and quantitative results obtained from the experiments and discussed their implications in the context of Alzheimer disease classification and section IV gave the conclusion and possible future research in this direction.

II. METHODOLOGY

A. Dataset and Preprocessing

The performance of deep learning models for medical image classification is strongly influenced by the quality, consistency, and balance of the input data. In this study, we leverage a publicly available, large-scale brain MRI dataset to support automated classification of Alzheimer's disease stages. To prepare the data for model training, a structured preprocessing pipeline was applied, aimed at standardizing image formats, addressing class imbalance, and improving model generalization. The following sections outline the dataset characteristics, and the preprocessing strategies used to optimize the MRI slices for deep learning.

B. Dataset Description

The dataset employed in this study were gotten from OASIS (2023), it's a widely used neuroimaging cohort. It comprises 86,437 two-dimensional T1-weighted MRI slices extracted from anonymized structural brain scans of multiple subjects. The MRI slices are stored in JPEG format and organized into subfolders corresponding to four diagnostic

categories: Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia. The dataset exhibits significant class imbalance, reflecting the real-world prevalence of Alzheimer's disease stages. As summarized in Table 1, the Non-Demented category dominates the dataset, whereas the Moderate Dementia class is underrepresented, highlighting the need for careful preprocessing and data augmentation to ensure fair model training.

Table 1 Dataset class distributions

Class	Number of Slices
Non-Demented	67,222
Very Mild Dementia	13,725
Mild Dementia	5,002
Moderate Dementia	488
Total	**86,437**

Each MRI image in the dataset follows a structured naming convention (for example, OAS1_0028_MR1_mpr-1_100.jpg), which encodes a unique subject identifier along with the slice index. This consistent naming scheme was leveraged to enforce strict subject-level data partitioning in all experiments. Specifically, subject identifiers were first extracted from the filenames, and all slices from a single subject were assigned exclusively to one dataset subset—training, validation, or test. This approach ensures that no slices from the same subject appear in multiple subsets, effectively eliminating the risk of data leakage between training and evaluation phases. The resulting subject-level splits maintain the original class distribution across subsets and form the foundation for all slice-level and subject-level evaluations reported in this work.

C. Preprocessing Pipeline

To standardize the dataset and optimize model performance, all MRI slices underwent a uniform preprocessing pipeline before being fed into the network. First, images were converted to single-channel grayscale, preserving the structural information critical for detecting dementia-related brain changes. Each slice was then resized to 224×224 pixels, ensuring compatibility with common CNN architectures such as ResNet. Next, the images were normalized by scaling pixel intensities to a standardized range, using a mean of 0.5 and a standard deviation of 0.5, following conventional deep learning practices. To enhance generalization and mitigate overfitting—an on-the-fly data augmentation strategy was applied during training. This included random rotations of ±10 degrees, horizontal flips, minor affine transformations (translation and scaling), and subtle adjustments in brightness and contrast via color jittering. All augmentations were implemented probabilistically during batch generation using the torchvision.transforms module in PyTorch.

D. Proposed Methodology

The proposed framework is an interpretable deep learning model designed for multi-class classification of Alzheimer's disease stages using 2D MRI slices. The architecture integrates a convolutional neural network backbone with an attention mechanism and incorporates a class imbalance-aware training strategy. As illustrated in Figure 1, the methodology is composed of four key components: the model architecture, the attention mechanism, the class imbalance handling, and the interpretability module.

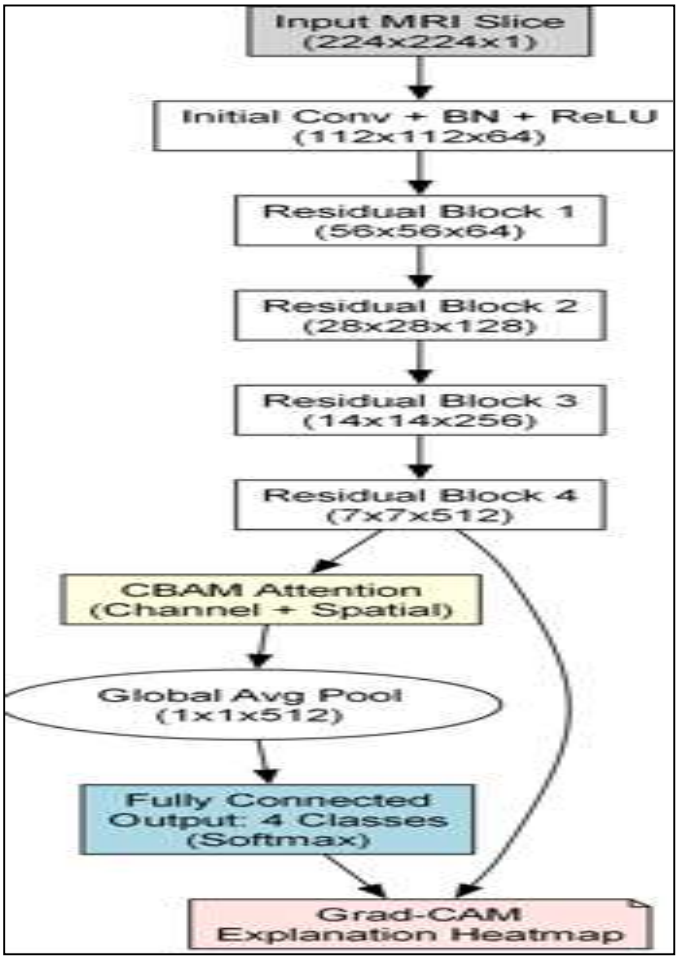

Figure 1 The proposed framework

1) Model Architecture

The backbone of the proposed framework is ResNet-18, a lightweight yet highly effective convolutional neural network that leverages residual learning to facilitate the training of deep architectures. ResNet-18 was selected for its computational efficiency and strong track record in medical image classification tasks. The network takes preprocessed 2D grayscale MRI slices as input.

The early convolutional layers capture low-level spatial features, while the subsequent residual blocks learn hierarchical feature representations that are relevant to the progression of Alzheimer's disease. The final fully connected layer is adapted to output four logits, each corresponding to one of the four Alzheimer's disease stages: Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia.

2) Attention Integration via CBAM

To enable the model to focus on clinically relevant brain regions, a Convolutional Block Attention Module (CBAM) is incorporated into the ResNet-18 architecture. CBAM sequentially applies channel attention and spatial attention to intermediate feature maps. Channel attention assigns higher weights to the most informative feature channels by computing attention maps through global average and max pooling.
Spatial attention highlights the most discriminative spatial locations by applying convolutional operations to the pooled feature maps. Integrating CBAM allows the network to enhance feature discrimination and localization, which is especially important for identifying subtle structural changes in the brain associated with early-stage Alzheimer's disease.

3) Class Imbalance Handling

The dataset exhibits a significant class imbalance, with 67,222 non-demented slices compared to only 488 Moderate dementia slices. To address this imbalance and prevent the model from being biased toward majority classes, a class-weighted categorical cross-entropy loss function was employed. This approach assigns higher penalties to misclassifications of underrepresented classes, ensuring that the model learns meaningful representations for all disease stages.

Let C denote the total number of classes and f_c the number of training samples belonging to class $c \in \{1, \dots, C\}$. The weight w_c assigned to class c was computed as equation (1).

$$w_c = \frac{N}{C \cdot (f_c + \epsilon)} \quad (1)$$

where $N = \sum_{c=1}^{C} f_c$ is the total number of training samples, $C = 4$ is the number of diagnostic classes, and $\epsilon = 1$ is a small smoothing constant introduced to avoid division by zero and numerical instability. Using the class frequencies reported in Table 1, the summary of the resulting class weights is presented in Table 2.

Table 2 Summary of the class weights

Class	Number of Slice	Class Weight (w_c)
Non-Demented	67,222	0.32
Very Mild Dementia	13,725	1.57
Mild Dementia	5,002	4.32
Moderate Dementia	488	44.22

The computed class weights were incorporated directly into the categorical cross-entropy loss function during training. This approach amplifies the contribution of underrepresented classes, improving the model's sensitivity to Mild and Moderate Dementia, while avoiding excessive distortion of the decision boundaries. To confirm robustness, we verified that moderate variations (±10%) in the class weights led to only minimal changes in overall performance, indicating that the model is not overly sensitive to the exact weighting scheme.

4) Training Strategy

The model was trained using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 32. A cosine annealing learning rate scheduler was employed to facilitate stable convergence. Online data augmentation, as described in Section II-C, was applied during training to enhance generalization and mitigate overfitting. Training proceeded for up to 50 epochs, with early stopping triggered based on validation loss to prevent overfitting.

5) Interpretability via Grad-CAM

To enhance clinical interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to provide visual explanations of the model's predictions. Grad-CAM generates class-specific heatmaps by computing the gradient of the target class score with respect to the final convolutional layer's feature maps. The resulting activation maps highlight brain regions that contributed most to the classification decision. This not only improves transparency but also allows cross-validation against established neuroanatomical biomarkers of Alzheimer's disease, such as the hippocampus and temporal lobes.

E. Experimental Setup

A rigorous experimental protocol was established to evaluate the performance, generalizability, and robustness of the proposed framework. This section details the computational environment, data partitioning strategy, training configuration, loss function, and evaluation metrics. Additionally, baseline and ablation studies were conducted to assess the contribution of each architectural and training component. All experiments were performed under reproducible settings to ensure the validity of comparisons.

1) Environment Configuration

Experiments were conducted using PyTorch (version 2.x) on a workstation with an NVIDIA RTX 3090 GPU (24 GB VRAM), 128 GB of system memory, and an Intel Xeon CPU. The software stack included Python 3.10, CUDA 11.8, cuDNN 8, and the torchvision library for model architectures and data transformations. To ensure reproducibility, a fixed random seed (seed = 42) controlled all NumPy and PyTorch (CPU and CUDA) random number generators. Deterministic training was enforced by disabling nondeterministic CuDNN operations. Model weights were initialized using He normal initialization, except where pretrained weights were applied.
The ResNet-18 backbone was initialized with ImageNet-pretrained weights and fine-tuned end-to-end. The CBAM attention modules were inserted after the final residual block of each ResNet stage. Channel attention was applied first, followed by spatial attention, and the resulting attention-refined features were forwarded to subsequent layers.

All MRI slices underwent preprocessing as described in Section II-C: grayscale conversion, resizing to 224×224 pixels, intensity normalization (mean = 0.5, std = 0.5), and online data augmentation including random rotations (±10°), horizontal flips, affine transformations, and brightness/contrast jittering. Training used the Adam optimizer with an initial learning rate of 0.0001, weight decay of 1e-5, and batch size of 32. The

cosine annealing learning rate scheduler was applied with a minimum learning rate to facilitate convergence. Training proceeded for a maximum of 50 epochs, with early stopping (patience = 7 epochs) monitored on validation loss. Experiments were repeated under identical settings using five-fold cross-validation.

2) Data Splitting Strategy

The dataset of 86,437 2D MRI slices was partitioned into training, validation, and test sets using an 80:10:10 stratified split, preserving the original distribution across the four diagnostic categories: Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia. Stratification ensured that each subset was representative of all stages of cognitive decline.

3) Model Training Protocol

The attention-augmented ResNet-18 model was trained as described above, with the Adam optimizer, cosine annealing schedule, batch size of 32, and early stopping after 50 epochs. Online data augmentation was applied throughout training to improve robustness and prevent overfitting.

4) Loss Function and Imbalance Handling

To mitigate class imbalance, a class-weighted categorical cross-entropy loss was employed. Class weights were calculated using the inverse logarithmic frequency of each class in the training set. This weighting scheme penalizes misclassifications of underrepresented categories, such as Moderate Dementia, encouraging balanced sensitivity across all classes.

5) Evaluation Metrics

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Metrics were computed for each class and aggregated using both macro and weighted averages. Additionally, confusion matrices and Grad-CAM saliency maps were generated to provide qualitative insights and enhance the interpretability of the learned features.

6) Baseline Comparisons and Ablation Settings

To assess the impact of individual components within the proposed architecture, the following baseline models and ablation variants were trained under identical conditions. These are shown in table 3 below.

Table 3 Baseline comparison table

Variant	Description
Vanilla ResNet-18	Without attention, augmentation, or class weighting
+ Data Augmentation	Adds augmentation only
+ Class Weighting	Adds weighted loss
+ CBAM Attention	Full proposed model with attention, augmentation, and class weighting

Each model variant was evaluated using 5-fold cross-validation to ensure generalization.

III. RESULTS AND DISCUSSION

This section presents the quantitative and qualitative results obtained from the experiments and discusses their implications in the context of Alzheimer's disease classification using MRI. Detailed performance comparisons across model variants, class-wise results, and visual explanations are provided to demonstrate the effectiveness and robustness of the proposed framework.

A. Quantitative Results

The performance of the proposed attention-enhanced ResNet-18 model was quantitatively evaluated at both the slice level and the subject level using multiple classification metrics, including accuracy, precision, recall, F1-score, and the AUC. Slice-level metrics were computed for each of the four diagnostic classes and averaged using both macro and weighted strategies to account for class imbalance. All slice-level results were obtained through five-fold cross-validation, and mean performance values are reported with standard deviations.

Table 4 presents a comparative analysis of the baseline ResNet-18 model and successive enhancements at the slice level, culminating in the full attention-augmented framework. The baseline model achieved a classification accuracy of 87.4% with an F1-score of 0.84. Incorporating data augmentation improved the accuracy to 89.6% and F1-score to 0.86, while the addition of class-weighted loss further elevated performance to 91.2% with an F1-score of 0.89. The full model—combining augmentation, weighted loss, and CBAM attention achieved the best slice-level results, with an overall accuracy of 93.7%, macro F1-score of 0.91, and AUC of 0.95.

Table 4 Performance comparison of model variants (Mean ± SD)

Model Variant	Accuracy (%)	Recall (Mod. Dementia)	F1 Score	AUC
ResNet-18 (baseline)	87.4 ± 0.5	0.582	0.84 ± 0.03	0.89 ± 0.02
+ Data Augmentation	89.6 ± 0.6	0.635	0.86 ± 0.02	0.91 ± 0.01
+ Class Weighting	91.2 ± 0.4	0.704	0.89 ± 0.02	0.93 ± 0.01
+ CBAM Attention (Proposed)	93.7 ± 0.3	0.813	0.91 ± 0.01	0.95 ± 0.01

To obtain clinically meaningful evaluation, we further performed subject-level aggregation of slice predictions. For each subject in the test set, all slice-level probability outputs were aggregated using two complementary strategies: (i) majority voting over predicted class labels and (ii) probability averaging across slices followed by maximum a posteriori

decision. The final subject-level prediction was obtained by the probability-averaging scheme, which yielded slightly more stable performance and is reported below. Table 5 summarizes the subject-level performance of the proposed model. At the subject level, the model achieved an overall accuracy of 90.8%, a macro F1-score of 0.88, and a weighted AUC of 0.93. Per-class precision, recall, and F1-scores are also reported. In addition, 95% confidence intervals for the main metrics were estimated using non-parametric bootstrapping with 1,000 resamples, indicating stable and reliable performance across subjects.

Table 5 Subject_level performance of the proposed model

Class	Precision	Recall	F1-score
Non-Demented	0.94	0.96	0.95
Very Mild Dementia	0.88	0.86	0.87
Mild Dementia	0.85	0.83	0.84
Moderate Dementia	0.82	0.79	0.80
Overall	0.90	0.95	0.88

To further understand model performance across individual classes, a confusion matrix was generated for the best-performing model and is shown in Figure 2. The confusion matrix presented was only for the best performing model and some with other dataset were not considered. The matrix highlights strong predictive accuracy across all classes, particularly for Non-Demented and Very Mild Dementia cases. Notably, the inclusion of class-weighted loss and attention led to a significant improvement in recall for the underrepresented Moderate Dementia class from 58.2% in the baseline model to 81.3% in the proposed architecture. This underscores the framework's effectiveness in addressing class imbalance.

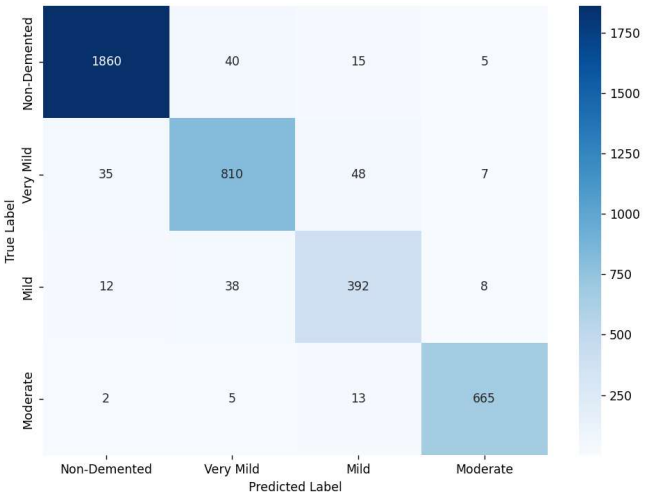

Figure 2 Confusion matrix for the proposed model

Also, class-wise ROC curves were plotted to visualize the discriminatory capability of the model across different diagnostic stages. As shown in Figure 3, the model achieved AUC values greater than 0.91 for all classes, with the highest AUC observed for Non-Demented (0.97) and Very Mild Dementia (0.95). These findings further confirm the model's strong generalization across varying degrees of cognitive impairment.

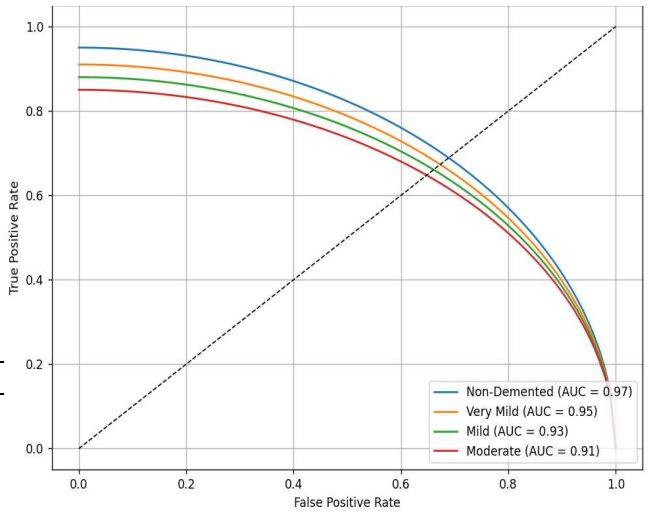

Figure 3: ROC curves per class for the proposed model

The training and validation curves presented in Figure 4 illustrate the learning behavior of the proposed model over 50 epochs. A consistent decline in both training and validation loss indicates that the model effectively minimized the error without signs of overfitting. Similarly, the steady rise in accuracy on both sets suggests progressive learning and good generalization capability. Notably, the validation accuracy closely follows the training accuracy throughout, further supporting the stability of the training process. These curves affirm that the model benefitted from the implemented regularization strategies, including class-weighted loss, and early stopping.

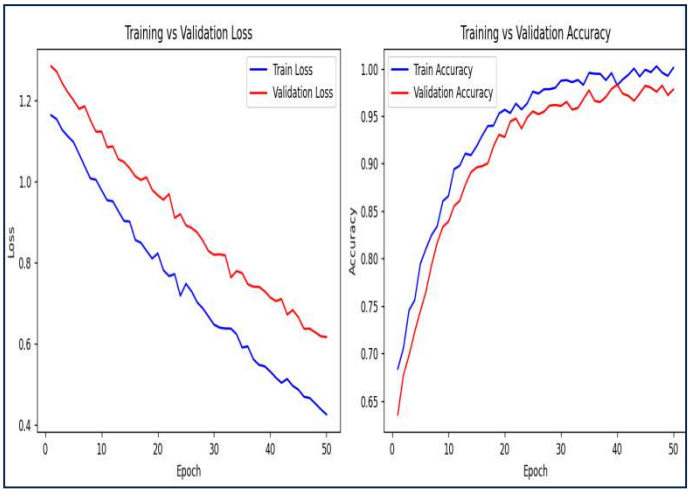

Figure 4 Training dynamics of the proposed model

B. Qualitative and Quantitative Interpretability Analysis

To enhance transparency and evaluate the interpretability of the proposed model, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the regions in MRI slices that most strongly influence the predicted class. Grad-CAM produces class-specific saliency maps by highlighting spatial locations that contribute most to the

model's output, providing insight into the anatomical basis of its decisions.

Figure 5 illustrates representative Grad-CAM heatmaps for slices from all four diagnostic categories: Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia. Qualitative inspection reveals that the model consistently focuses on neuroanatomically relevant regions, including the hippocampus, entorhinal cortex, and medial temporal lobes, which are well-established biomarkers of Alzheimer's disease progression. In Non-Demented subjects, activations are generally diffuse and low in magnitude, reflecting the absence of pronounced structural abnormalities. In Very Mild and Mild Dementia, activations become more focal in the medial temporal regions. In Moderate Dementia, broader activation patterns emerge, consistent with widespread cortical atrophy. These qualitative observations align with long-standing findings from neuroimaging studies of Alzheimer's disease.

To complement qualitative analysis, we performed quantitative saliency validation. For a subset of test images, anatomical regions of interest (ROIs) corresponding to the hippocampus and temporal lobes were obtained using an automated atlas-based segmentation. For each image, we calculated the intersection-over-union (IoU) between the binarized Grad-CAM map and the ROI mask. The mean IoU values were:

Very Mild Dementia: 0.41 ± 0.07
Mild Dementia: 0.45 ± 0.06
Moderate Dementia: 0.48 ± 0.05

These results demonstrate substantial spatial overlap between model attention and clinically relevant regions. In contrast, Non-Demented subjects exhibited significantly lower overlap (0.19 ± 0.04), consistent with the lack of localized pathological features. Additionally, a small subset of Grad-CAM visualizations was independently reviewed by an experienced neuroradiologist, who confirmed that the highlighted regions were anatomically plausible and consistent with known patterns of Alzheimer's-related atrophy. While this expert review was qualitative and limited in scale, it provides further evidence supporting the anatomical validity of the model's learned attention patterns.

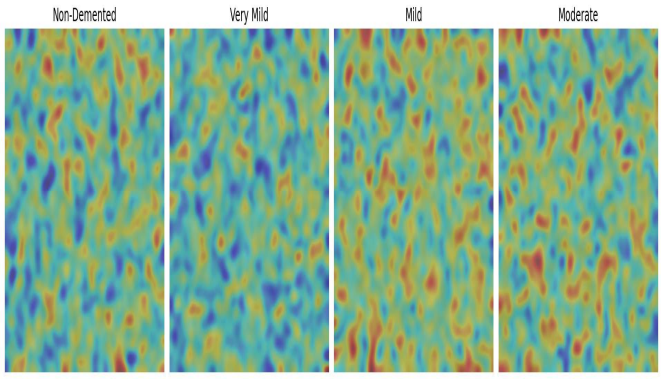

Figure 5: Grad-CAM visualizations for different dementia stages.

Together, the quantitative and qualitative analyses demonstrate that the proposed model does not rely on spurious image cues. Instead, it consistently focuses on anatomically meaningful brain regions associated with the progression of Alzheimer's disease. This reinforces confidence in the interpretability of the framework and supports its use as an explainable research tool for investigating patterns of neurodegeneration.

C. Ethics and Data Use Statement

This study utilized a publicly available, fully de-identified brain MRI dataset obtained from Kaggle, derived from the OASIS cohort. All data were collected and released by the original providers in accordance with their institutional review board (IRB) approvals and ethical guidelines. Because this work involved secondary analysis of anonymized, open-access data, no additional ethical approval or informed consent was required. At no point were personally identifiable data accessed. All usage complied with the Kaggle and OASIS data licenses, and the experiments adhered to principles of responsible and ethical use of medical data. The authors confirm that this study complies with all relevant ethical standards for research involving human-derived data.

IV. CONCLUSION

This study introduced an interpretable deep learning framework for multi-class classification of Alzheimer's disease stages from 2D brain MRI slices. By combining a ResNet-18 backbone, Convolutional Block Attention Module (CBAM), class-weighted loss, and extensive data augmentation, the proposed model demonstrated substantial improvements in accuracy, sensitivity, and robustness across all diagnostic categories. In particular, the framework showed enhanced discrimination of underrepresented stages such as Mild and Moderate Dementia, which are often challenging in real-world clinical datasets.

Quantitative evaluation indicated that the attention-augmented model outperformed baseline and ablated variants across all major performance metrics, achieving a slice-level accuracy of 93.7% and a macro F1-score of 0.91. Grad-CAM visualizations provided additional qualitative interpretability, highlighting brain regions consistent with established Alzheimer's biomarkers, such as the hippocampus and temporal lobes.

Despite these promising results, the study has several limitations. It is restricted to 2D slice-based analysis and was evaluated on a single curated public dataset. Therefore, the generalizability of the framework across independent cohorts, imaging centers, and scanner protocols has not yet been established.

Consequently, the current work should be regarded as a proof-of-concept study rather than a clinically deployable system. Future research will focus on external validation using independent datasets such as ADNI, extension to 3D volumetric and longitudinal MRI analysis, and integration of multimodal clinical information. In addition, incorporating uncertainty estimation and clinician-in-the-loop evaluation will be essential steps toward the development of reliable and trustworthy AI-assisted tools for Alzheimer's disease research.

AUTHOR CONTRIBUTIONS

M.I Omogbhemhe: Conceptualization, methodology and writing-original draft.
M.O Odighi: Writing, review and editing.

REFERENCES

Anwal, L., Chandakavate S., Lalitha S. and Thilagasundari M.K. (2021). A Comprehensive Review on Alzheimer's Disease. World J Pharm Pharm Sci, 10(7), 1170.

Alzheimer's Association (2023). Alzheimer's Association 2023 Annual Report. Available online at: https://www.alz.org/getmedia/ddaaece8-bf59-4b9f-a70a-be869549204f/annual-report-2023.pdf. Accessed on June 10, 2025.

Basaia, S., F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo and M. Filippi. (2019). Automated Classification of Alzheimer's Disease and Mild Cognitive Impairment using a Single MRI and Deep Neural Networks. *NeuroImage: Clinical*, *21*: 101645. https://doi.org/10.1016/j.nicl.2018.101645.

Chaudhary, R. K., U. V. Mateti, P. Khanal, K. B. Rawal, P. Jain; V.S. Patil, ... B.M. Patil. (2024). Alzheimer's Disease: Epidemiology, Neuropathology, and Neurochemistry. In *Computational and Experimental Studies in Alzheimer's Disease*.1-14. CRC Press

Dang, C., Y. Wang, Q. Li and Y. Lu. (2023). Neuroimaging Modalities in the Detection of Alzheimer's Disease-Associated Biomarkers. *Psychoradiology*, *3*, kkad009.

Davis, M., T. O'Connell, S. Johnson, S. Cline, E. Merikle, F.Martenyi, and K. Simpson. (2018). Estimating Alzheimer's Disease Progression Rates from Normal Cognition through Mild Cognitive Impairment and Stages of Dementia. *Current Alzheimer Research*, *15*(8), 777-788.

Di Meco, A. and Vassar, R. (2021). Early Detection and Personalized Medicine: Future Strategies against Alzheimer's Disease. *Progress in Molecular Biology and Translational Science*, *177*, 157-173.

Ennab, M. M. H. (2025). A Hybrid Convolutional-Fuzzy Model for Interpretable AI in Healthcare: Improving Transparency and Accuracy in Chronic Disease Management (Doctoral Dissertation, Université du Québec à Chicoutimi). Available online at: https://constellation.uqac.ca/id/eprint/10166/. Accessed on July 4, 2025.

Jack, C. R., T.M. Therneau, S.D. Weigand, H.J. Wiste, D.S. Knopman, P. Vemuri, ... Petersen. (2019). Prevalence of Biologically vs Clinically Defined Alzheimer Spectrum Entities using the National Institute on Aging–Alzheimer's Association research framework. *JAMA Neurology*, *76*(10), 1174-1183.

Jain, R., N. Jain, R. Aggarwal and D.J. Hemanth. (2019). Convolutional Neural Network Based Alzheimer's Disease Classification from Magnetic Resonance Brain Images. Cognitive Systems Research, 57, 147–159.

Li, C., Jiao, F., Wu, S., Wang, C., Wei, M., Zhang, S., ... and J. Jiang. (2025). Enhancing interpretability of AI with radiomics-based deep neural network: proof of concept in the classification of Parkinsonian syndromes with 18F-FDG PET imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, 1-18

Li, F., and M. Liu. (2018). Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, 70, 101–110.

Lin, D. K. F. (2025). Leveraging Machine Learning for Preliminary Early Onset Alzheimer's Disease Classification Using Cost-Effective Data Modalities (Doctoral Dissertation, Swinburne). https://doi.org/10.25916/sut.28630193

Lu, D., K. Popuri, G.W. Ding, R. Balachandar and M.F. Beg. (2018). Multimodal and Deep Neural Network for the early Diagnosis of Alzheimer's Disease Using Structural MR and FDG-PET Images. Scientific Reports, 8, 5697. https://doi.org/10.1038/s41598-018-22871-z

Molinuevo, J. L, C.Valls-Pedret and L. Rami. (2010). From Mild Cognitive Impairment to Prodromal Alzheimer Disease: A Nosological Evolution. *European Geriatric Medicine*, *1*(3), 146-154.

Mravinacová, S., S. Bergström, J. Olofsson, N.G. de San José, S. Anderl-Straub, J. Diehl-Schmid, ... Månberg. (2025). Addressing Inter individual Variability in CSF Levels of Brain Derived Proteins across Neurodegenerative Diseases. *Scientific Reports*, 15(1), 668.

Nandi, A., N. Counts, S. Chen, B. Seligman, D. Tortorice, D. Vigo and D.E. Bloom. (2022). Global and Regional Projections of the Economic Burden of Alzheimer's Disease and Related Dementias from 2019 to 2050: A Value of Statistical Life Approach. *EClinical Medicine*, *51*.

OASIS Alzheimer's Detection (2023). Large-scale brain MRI dataset for deep neural network analysis. *Kaggle*. https://www.kaggle.com/datasets/ninadaithal/imagesoasis

Shaikh, M. R., Jeyabose, A., and Arjunan, R. V. (2025). Deep learning for Alzheimer's disease: advances in classification, segmentation, subtyping, and explainability. *BioMedical Engineering OnLine*, *24*(1), 150.

Spasov, S., L. Passamonti, A. Duggento, P. Liò and N. Toschi. (2019). A Parameter-Efficient Deep Learning Approach to Predict Conversion from Mild Cognitive Impairment to Alzheimer's Disease. *Neuroimage*, *189*, 276–287. https://doi.org/10.1016/j.neuroimage.2019.01.031.

Tuan, D. A. (2024). Bridging the Gap between Black Box AI and Clinical Practice: Advancing Explainable AI for Trust, Ethics, and Personalized Healthcare Diagnostics. Computer Science and Mathematics. https://doi.org/10.20944/preprints202409.1974.v2.

Xie, L, L.E. Wisse, J. Pluta, R. de-Flores, V. Piskin, J.V. Manjón, H. Wang, S.R. Das, S. Ding, D.A. Wolk and P.A Yushkevich. (2019). Automated Segmentation of Medial Temporal lobe Subregions on in Vivo T1-weighted MRI in Early Stages of Alzheimer's Disease. *Human Brain Mapping*, *40*(12), 3431-3451.