ResNet-50 Transfer Learning for Early Detection of Alzheimer's Disease on OASIS-1 MRI with Explainable AI Interpretability Analysis Using Grad-CAM

1st Rangga Athari Gunawan
School of Computing
Telkom University
Bandung, Indonesia
ranggaathari@student.telkomuniversity.ac.id

2nd Jondri
School of Computing
Telkom University
Bandung, Indonesia
jondri@telkomuniversity.ac.id

3rd Indwiarti
School of Computing
Telkom University
Bandung, Indonesia
indwiarti@telkomuniversity.ac.id

***Abstract*—Early detection of Alzheimer's disease (AD) is essential for timely clinical intervention. This study proposes an interpretable deep learning framework for binary AD classification using two-dimensional coronal slices from the OASIS 1 structural MRI dataset. A pretrained ResNet-50 model was fine tuned under five training scenarios to evaluate the roles of class weighting, data augmentation, and optimization strategies. Evaluation was conducted at both slice and subject levels using multiple metrics, with the F2 score prioritized to emphasize recall-oriented performance. Multiple-run experiments using five independent random seeds demonstrated that the final configuration was stable across different initializations. The best model achieved a subject level AUC of 0.924 and an F2 score of 0.860 at the optimal threshold, with a recall of 0.987 across runs. Gradient weighted Class Activation Mapping showed that the model's attention aligned with established AD biomarkers in the medial temporal lobe, particularly the hippocampus and parahippocampal gyrus. These findings indicate that a ResNet-50-based approach can provide accurate and interpretable predictions for early AD screening using structural MRI.**

Keywords*—Alzheimer, early detection, deep learning, transfer learning, ResNet-50, MRI, Grad-CAM, explainable artificial intelligence

I. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, accounting for approximately 60–80% of all cases [1]. As a chronic and incurable condition, early detection at the prodromal or Mild Cognitive Impairment (MCI) stage is crucial for slowing disease progression and improving treatment outcomes [2]. However, early diagnosis remains challenging because clinical symptoms typically appear only after significant neuronal damage has occurred [3].

Magnetic Resonance Imaging (MRI) enables non-invasive visualization of structural brain changes such as hippocampal and cortical atrophy, which are recognized as neuroimaging biomarkers of AD [4], [5]. Nevertheless, manual MRI interpretation is subjective, time-consuming, and prone to inter-observer variability, particularly in early-stage detection [5]. To address these limitations, deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated superior performance in the automated classification of AD using structural MRI scans [2], [3].

Transfer learning approaches employing pretrained architectures such as ResNet-50 have become increasingly popular due to their ability to extract representative features from limited medical datasets and achieve high classification performance in prior studies [6]. While three-dimensional CNNs can utilize full volumetric MRI scans, they typically require substantially larger datasets and significantly higher computational resources, which often limits their applicability in clinical settings [7], [8]. Consequently, two-dimensional slice based methods remain widely adopted for their efficiency and competitive diagnostic performance [7]. Recent evidence indicates that MRI slice orientation can influence the effectiveness of deep learning models in Alzheimer's disease classification, motivating the use of anatomically informative views such as the coronal plane [9], [10].

However, CNN-based models are often criticized for their black-box nature and limited interpretability, which limits their adoption in clinical settings [11], [12]. To overcome this issue, Explainable Artificial Intelligence (XAI) techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) are employed to visualize regions that contribute to classification decisions [4], [13]. Grad-CAM has been shown to highlight anatomically relevant brain areas such as the hippocampus and temporal cortex, aligning with known neuropathological features of AD [9], [14]. This capability enhances the transparency and trustworthiness of Artificial Intelligence (AI)-based diagnostic systems [11].

Furthermore, the limited availability of data remains a key challenge in medical imaging research. Public datasets such as the Open Access Series of Imaging Studies (OASIS-1) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) are widely used but relatively small in size [15], [16]. Data augmentation techniques involving geometric and intensity transformations have proven effective in reducing overfitting and improving model generalization [8], [17]. More recent approaches employ generative models to synthesize realistic MRI data, although these methods still require careful validation [17].

Based on this background, the present study proposes an early-stage AD classification framework using a ResNet-50 architecture with transfer learning on 2D coronal MRI slices from the OASIS-1 dataset. The model is trained using conventional data augmentation techniques and evaluated based on accuracy, precision, recall, F1 score, F2 score, and Area Under the ROC Curve (AUC). To enhance interpretability, Grad-CAM is utilized to visualize the model's spatial attention during prediction. Unlike previous studies that focused primarily on accuracy, this research emphasizes recall-oriented evaluation to minimize false negatives, thereby supporting earlier and more reliable clinical decision-making.

II. Related Works

Deep learning approaches have demonstrated substantial promise in Alzheimer's disease classification using structural MRI. Multiple studies have shown that CNNs consistently outperform traditional machine-learning methods, particularly in tasks requiring the extraction of subtle anatomical abnormalities [2], [15], [18]. Among the architectures commonly employed, ResNet-50 has emerged as a reliable baseline due to its stable gradient propagation enabled by residual connections, strong generalization capability, and effectiveness when trained via transfer learning on limited medical imaging datasets [3], [6], [18].

In the literature, two major modeling paradigms are commonly adopted for MRI-based AD classification: three-dimensional volumetric CNNs and two-dimensional slice-based CNNs. Although 3D models can exploit full volumetric information, they typically require substantially larger training cohorts and significantly higher computational resources, thereby increasing the risk of overfitting in modest-sized datasets such as OASIS-1 [7], [8]. Consequently, 2D slice-based approaches remain widely used due to their computational efficiency, compatibility with ImageNet-pretrained architectures, and competitive diagnostic performance even when training data are limited [6], [7], [11]. Several works have demonstrated that carefully selected slices, particularly those capturing the medial temporal lobe, preserve sufficient diagnostic information for early AD detection and yield robust performance across subjects [7], [11]. These findings underscore the practicality of adopting a 2D framework in resource-constrained or dataset-limited clinical research settings.

TABLE I. summarizes representative studies that apply deep learning for MRI-based AD classification. These works collectively highlight the relevance of CNN architectures, the increasing adoption of transfer learning, and the growing emphasis on model interpretability. As reported in [10], coronal slices consistently achieved superior accuracy compared with axial and sagittal orientations across multiple architectures. This aligns with broader evidence showing that slice orientation substantially influences the effectiveness of 2D deep learning models for neurodegenerative disease analysis [10]. A concise comparison of average performance across slice orientations is provided in TABLE II. further illustrating the advantage of coronal views. Coronal slices offer clearer visualization of the medial temporal lobe, including the hippocampus and parahippocampal gyrus, which are among the earliest regions affected by AD pathology [9]. Additional findings in [7] similarly show that coronal MRI frames contain richer structural information for AD stage prediction than axial or sagittal slices.

Interpretability has also emerged as a central theme in recent AD classification research [11], [12]. Explainable AI techniques, particularly Grad-CAM, have been widely employed to identify disease-relevant brain regions and verify whether network attention corresponds to established neuropathological patterns [4], [12], [13]. Prior studies have reported consistent Grad-CAM activation in the hippocampus, entorhinal cortex, and temporal lobe, which are well-established biomarkers of early Alzheimer's progression [4], [12], [13], [14], [19]. These findings highlight the need for models that are both accurate and clinically interpretable.

TABLE I. Previous Studies

Author	Dataset	Model	Results
Nithya et al. [18]	ADNI, OASIS	ResNet-50 (TL)	95.6% accuracy for AD vs. non-AD classification
Ramalho et al. [10]	ADNI	2D CNN (axial/sagittal/coronal)	Coronal slices consistently yielded the highest accuracy compared to axial and sagittal orientations
Coluzzi et al. [4]	ADNI	CNN + Grad-CAM	Identified hippocampal and temporal regions as key areas
Turrisi et al. [8]	ADNI	Various CNNs	Performance was highly dependent on architecture and training strategy

TABLE II. Average Accuracy Across Slice Orientation and Training Epochs

Epoch	Coronal	Axial	Sagittal
20	0.99	0.99	0.98
50	1.00	0.99	0.99
80	0.99	0.99	0.99
100	1.00	0.99	0.99

Values summarized from Ramalho et al. [10]

Interpretability has also emerged as a central theme in recent AD classification research [11], [12]. Explainable AI techniques, particularly Grad-CAM, have been widely employed to identify disease-relevant brain regions and verify whether network attention corresponds to established neuropathological patterns [4], [12], [13]. Prior studies have reported consistent Grad-CAM activation in the hippocampus, entorhinal cortex, and temporal lobe, which are well-established biomarkers of early Alzheimer's progression [4], [12], [13], [14], [19]. These findings highlight the need for models that are both accurate and clinically interpretable.

Despite these advances, relatively few studies have examined how training strategies such as data augmentation, class balancing, fine tuning, and slice selection interact to influence both diagnostic accuracy and interpretability. The present study addresses this gap by integrating ResNet-50 transfer learning, two dimensional coronal slice selection, comprehensive augmentation strategies, and Grad-CAM analysis into a unified framework for early AD detection using the OASIS-1 dataset. The evaluation prioritizes sensitivity and interpretability to develop a model that is diagnostically reliable and clinically meaningful.

III. Methodology

The proposed framework for early AD detection consists of six main stages: dataset preparation, data preprocessing, data augmentation, model training, interpretability analysis, and performance evaluation. The overall workflow is illustrated in Fig. 1. The dataset comprises T1-weighted structural MRI images from OASIS-1, which underwent preprocessing steps including skull stripping, intensity normalization, and spatial resampling. Following preprocessing, the 3D volumes were converted into 2D coronal slices and resized to 224×224 pixels.

To enhance model robustness and reduce overfitting, the training subset was further expanded using conventional data augmentation techniques, including rotation, flipping, translation, zooming, and brightness or contrast adjustment. After augmentation and preprocessing, the resulting slices were input into a transfer learning framework based on a ResNet-50 model pretrained on the ImageNet-1K dataset. The

original fully connected layer was replaced with a dropout layer and a final classification layer with two output nodes for binary prediction of AD versus non-AD subjects.

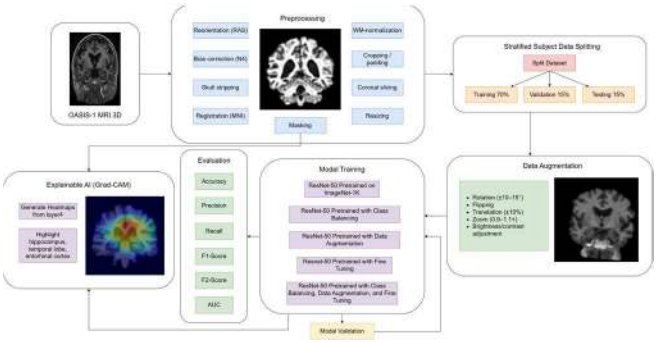

Fig. 1. Overall workflow of the proposed early AD detection framework.

To facilitate interpretability, Grad-CAM was employed to highlight brain regions that contributed most to the model's classification decisions. Model performance was evaluated using multiple metrics, including accuracy, precision, recall, F1 score, F2 score, and AUC, to provide a comprehensive assessment of both discrimination capability and recall-oriented behavior.

A. Dataset

This study utilized the OASIS-1 dataset, which provides T1-weighted MRI scans from a cross-sectional cohort of older adults. The dataset includes 416 unique participants; however, 20 non-demented individuals were scanned twice, and their follow-up acquisitions are labeled as MR2. After accounting for these repeated scans, a total of 436 MRI sessions were available, consisting of both Alzheimer's disease (AD) and non-AD cases. Clinical diagnosis was based on the Clinical Dementia Rating (CDR), where CDR = 0 denotes non-AD and CDR ≥ 0.5 denotes AD.

Each MRI volume underwent skull stripping and spatial normalization to ensure cross-subject consistency. After pre-processing, 27,904 coronal slices were obtained (AD: 6,400; non-AD: 21,504). Dataset partitioning was performed at the unique subject level to prevent leakage arising from correlated slices or repeated sessions belonging to the same individual. The distribution of subjects and slices is shown in TABLE III.

TABLE III. DATA DISTRIBUTION

Class	CDR Score[a]	Sessions	Slices
non-Alzheimer	0.0	336	21,504
Alzheimer	0.5–1.0	100	6,400

[a] Clinical Dementia Rating Score

Because demographic attributes such as age and sex were not used as model inputs, classification relied solely on structural MRI appearance and CDR-based diagnostic labels. The OASIS-1 cohort is demographically homogeneous, which has been noted as a limitation that may restrict generalizability to more diverse populations [16], [20]. In this study, each of the 436 MRI sessions was treated as an independent subject-level sample, including cases in which the same participant was scanned twice (MR1 and MR2). Here, the term subject-level refers to a single MRI session as an independent imaging observation that inherits the participant's diagnostic label. This definition is appropriate because the model operates exclusively on individual MRI volumes rather than longitudinal information, and repeated sessions can exhibit meaningful variation in head positioning, noise patterns, or acquisition conditions. Each session was therefore treated as a separate subject for data splitting and evaluation. Slice-level predictions were subsequently aggregated into one probability per session to reflect model performance on distinct imaging observations.

B. Preprocessing

All MRI volumes were processed using a standardized pipeline to ensure cross subject consistency and reproducibility. Raw DICOM files were first converted to NIfTI format using dcm2niix (version 1.0.20240202), followed by reorientation into the canonical RAS anatomical space using the FSL tool fslreorient2std (version 6.0.7.7). Intensity nonuniformities were corrected using the N4BiasFieldCorrection algorithm from ANTs (version 2.2.0). Lower head and neck regions were removed with FSL robustfov, and skull stripping was performed using FSL BET (bet2) with a fractional intensity threshold of 0.3 to isolate the intracranial region. Each T1 weighted image was then prepared for subsequent spatial normalization and tissue processing.

Each T1 weighted image was registered to the MNI152 2 mm template using FLIRT with 12 degrees of freedom and trilinear interpolation. Tissue segmentation was performed using FSL FAST to obtain gray matter, white matter, and cerebrospinal fluid partial volume maps. A white matter mask was generated by thresholding the white matter partial volume map at 0.9 and applying it to the original T1 image using fslmaths. The mean white matter intensity was computed using fslstats, and white matter based intensity normalization was applied by dividing the T1 volume by this mean value to reduce inter subject intensity variability.

After preprocessing, coronal slices were extracted from the central 20 to 80 percent of the Y axis, corresponding to the anatomical mid region of the brain. A total of 64 uniformly sampled slices were obtained for each subject. Coronal orientation was specifically selected because it provides superior visibility of medial temporal lobe structures, including the hippocampus and parahippocampal gyrus, which represent key early biomarkers of Alzheimer related neurodegeneration [4], [9]. Restricting the range to 20 to 80 percent removes boundary slices with minimal brain tissue and reduces the likelihood of obtaining near duplicate slices from adjacent positions.

For each selected slice, a tight intracranial bounding box was computed from nonzero voxel locations. The slice was cropped accordingly, padded symmetrically to preserve aspect ratio, and resized to 224 × 224 pixels using bilinear interpolation to match the ResNet-50 input requirement. A binary mask was also generated by thresholding the cropped slice. This mask was stored separately and used only during the Grad CAM interpretability stage to ensure that activation maps remained confined to brain regions. Finally, subject level splitting was applied after preprocessing to ensure that slices from the same subject did not appear across training, validation, or testing subsets, as described below.

C. Data Splitting

To prevent data leakage and ensure a fair evaluation, the dataset was partitioned strictly at the subject level rather than the slice level. This approach avoids the well known risk in 2D slice based pipelines, where highly similar slices from the same MRI volume may inadvertently appear across subsets and artificially inflate performance [10]. In this study, subject

assignment to training, validation, and testing sets was completed before any slice extraction or augmentation steps. All slices generated from a given subject were therefore guaranteed to remain within the same subset, eliminating cross subject contamination. The resulting distribution of subjects and slices for each subset is detailed in TABLE IV. The proportions were 70% for training, 15% for validation, and 15% for testing, using stratified sampling to preserve the distribution of AD and non AD subjects. The training set was used for parameter optimization and data augmentation, the validation set for learning rate scheduling and early stopping, and the test set was reserved exclusively for final performance evaluation.

TABLE IV. DATA SPLITTING RESULT

Split	AD	Non-AD	Total
Train	70 / 4,480	235 / 15,040	305 / 19,520
Validation	15 / 960	50 / 3,200	65 / 4,160
Test	15 / 960	51 / 3,264	66 / 4,224

Values represent subjects / slices per class.

D. Data Augmentation

Deep learning models are highly dependent on data quantity and are prone to overfitting when trained on limited medical datasets [8], [17]. To mitigate this issue, data augmentation was applied exclusively to the training subset. Each coronal slice was resized to 224×224 pixels and randomly transformed using rotation (±10°), horizontal and vertical translation (±10%), zooming (0.9–1.1×), and horizontal flipping with a probability of 0.5. The images were then converted into three-channel grayscale and normalized using the ImageNet mean and standard deviation values to ensure compatibility with the pretrained ResNet-50 weights.

This augmentation pipeline effectively increased spatial diversity in the training data while preserving the anatomical characteristics of the brain, thereby reducing the risk of overfitting under limited-data conditions.

E. Model Architecture and Training Strategy

The proposed framework employs a ResNet-50 architecture pretrained on the ImageNet-1K dataset. The original fully connected layer was replaced with a dropout layer ($p = 0.5$), followed by a linear layer with two output units for binary classification between AD and non-AD subjects. Model training was implemented using PyTorch with the AdamW optimizer (weight decay = 0.0001) and a cross-entropy loss function. Learning rate scheduling followed the ReduceLROnPlateau policy, with early stopping applied to prevent overfitting.

To complement this deterministic configuration and ensure that the reported results were not dependent on a single training trajectory, each experimental scenario was trained across five independent runs using distinct random seeds. All hyperparameters, data splits, and optimization settings were held constant for every run, with variation arising solely from random initialization and mini batch sampling. The mean and standard deviation across these runs were reported as the final performance metrics, providing a more reliable estimate of model robustness under stochastic training dynamics.

Beyond the general optimization setup, a two phase fine tuning strategy was employed to improve model adaptation to the MRI domain. In the first phase, the pretrained backbone was frozen while the classification head was trained for two epochs using a learning rate of 0.001. In the second phase, all layers were unfrozen and the model was trained end to end using differential learning rates of 0.00001 for the backbone and 0.0001 for the classification head. Mixed precision training was adopted to improve computational efficiency while preserving numerical stability.

To systematically assess the contribution of each design component, five experimental scenarios were defined as summarized in TABLE V. Scenario S1 serves as the baseline pretrained ResNet-50 model. Scenario S2 incorporates class weighted loss to address class imbalance. Scenario S3 introduces fine tuning through the freeze unfreeze strategy with differential learning rates. Scenario S4 adds data augmentation to improve generalization. Scenario S5 combines class weighting, fine tuning, and augmentation, forming the complete configuration that was subsequently used for the Grad CAM interpretability analysis.

TABLE V. EXPERIMENTAL SCENARIOS FOR MODEL TRAINING

Scenario	Description
S1	Baseline pretrained ResNet-50
S2	S1 + Class weighting
S3	S1 + Fine-tuning (freeze–unfreeze)
S4	S1 + Data augmentation
S5	S1 + Class weighting + fine-tuning + augmentation

F. Explainable Analysis Using Grad-CAM

To evaluate how the model makes its decisions, the Grad-CAM method was applied to the layer 4 block of ResNet-50, which captures high-level semantic features relevant to Alzheimer's pathology [12], [13]. The Grad-CAM activation maps highlight brain regions that contribute most strongly to the model's classification decisions.

To maintain anatomical validity, these activation maps were masked using the brain segmentation results from the preprocessing stage, ensuring that Grad-CAM attention was confined to the intracranial area. The analysis was conducted across all experimental scenarios (S1–S5) to compare spatial attention differences resulting from varying training strategies.

For each scenario (S1–S5), the resulting Grad-CAM maps were post-processed and prepared for interpretability assessment by overlaying them onto anatomical brain templates. This step facilitated the systematic comparison of attention patterns across different training settings, while enabling later evaluation of their correspondence to known neuroanatomical markers of Alzheimer's disease.

G. Evaluation Metrics

Classification performance was evaluated using several complementary metrics, including accuracy, precision, recall (sensitivity), F1 score, F2 score, and AUC. These metrics were computed at both the slice level and the subject level to comprehensively assess diagnostic capability and generalization performance, consistent with standard conventions in deep learning–based medical image analysis. Precision, recall, and F-measure are defined as follows.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Precision is calculated as the number of true positive predictions divided by the total number of positive predictions, while recall is calculated as the number of true positive predictions divided by the total number of actual positive samples.

The general F-measure is calculated using the following formula:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \times Recall}{\beta^2 \cdot Precision + Recall} \quad (3)$$

with β = 1 for the F1 score and β = 2 for the F2 score. Here, TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. The AUC value was computed based on the ranking of predicted probabilities for the AD class, reflecting the model's ability to distinguish between classes across different decision thresholds.

At the slice level, metrics were calculated directly from individual 2D prediction results. At the subject level, the probabilities of all slices belonging to the same subject were averaged to obtain a single prediction score, followed by binary classification using a threshold. An initial threshold of 0.50 was used for binary labeling, and further threshold tuning was performed on the validation data to determine the optimal operating point.

Since the primary objective of this study is the early detection of AD, the F2 score was chosen as the main evaluation criterion to emphasize recall and reduce false negatives, aligning with priorities in clinical screening [4], [11]. This recall-oriented evaluation approach ensures high sensitivity to early pathological features while maintaining an acceptable level of precision. All results reported below are based on these metrics, supplemented by confusion matrix analyzes to provide further insights into the model's reliability across all experimental scenarios (S1–S5).

IV. Result and Discussion

A. Training and Convergence Analysis

All training scenarios (S1–S5) demonstrated stable convergence on the validation data. The use of a learning rate scheduler and early stopping proved effective in preventing overfitting, while mixed-precision training improved computational efficiency without degrading model performance, consistent with best practices in deep learning research within the medical domain [11], [14], [19], [21]. The baseline model (S1) showed rapid convergence but tended to overfit the training data. The application of class weighting (S2) improved recall stability by emphasizing the minority class (AD). The two-stage fine-tuning strategy (freeze–unfreeze) in scenario S3 accelerated feature adaptation to the medical domain but required careful tuning to avoid over-specialization. Data augmentation (S4) effectively delayed overfitting but was insufficient to address class imbalance on its own. The combination of all strategies (S5) produced the most stable and recall-oriented training process, demonstrating strong synergy between class balancing, fine-tuning, and data augmentation.

B. Classification Performance

1) Slice-Level Results: At the slice level, TABLE VI. summarizes the performance across the five experimental scenarios, reported as the mean and standard deviation over multiple independent runs. The baseline configuration (S1) achieved a relatively high AUC (0.917 ± 0.005) but exhibited low recall (0.461 ± 0.108), indicating that the pretrained ResNet-50 tended to miss a substantial portion of AD slices.

TABLE VI. Slice-level Performance (Mean ± STD) for Five Model Scenarios on Alzheimer's Detection

Scenario	ACC	F1	F2	AUC	Precision	Recall
S1	0.837 ± 0.011	0.554 ± 0.077	0.493 ± 0.092	0.917 ± 0.005	0.728 ± 0.036	0.461 ± 0.108
S2	0.850 ± 0.010	0.671 ± 0.061	0.687 ± 0.146	0.924 ± 0.006	0.675 ± 0.044	0.706 ± 0.175
S3	0.831 ± 0.010	0.593 ± 0.045	0.563 ± 0.062	0.910 ± 0.004	0.656 ± 0.016	0.544 ± 0.069
S4	0.832 ± 0.014	0.539 ± 0.081	0.480 ± 0.110	0.919 ± 0.005	0.713 ± 0.035	0.450 ± 0.106
S5	0.835 ± 0.005	0.703 ± 0.015	0.790 ± 0.046	0.909 ± 0.004	0.596 ± 0.015	0.862 ± 0.068

TABLE VII. Subject-level Performance (Mean ± STD) for Five Model scenarios on Alzheimer detection

Scenario	ACC	F1	F2	AUC	Precision	Recall
S1	0.854 ± 0.032	0.573 ± 0.176	0.510 ± 0.217	0.947 ± 0.008	0.832 ± 0.087	0.480 ± 0.230
S2	0.866 ± 0.026	0.708 ± 0.088	0.740 ± 0.162	0.947 ± 0.010	0.734 ± 0.099	0.654 ± 0.117
S3	0.842 ± 0.023	0.587 ± 0.106	0.542 ± 0.130	0.932 ± 0.012	0.725 ± 0.083	0.520 ± 0.143
S4	0.842 ± 0.028	0.521 ± 0.185	0.455 ± 0.223	0.946 ± 0.012	0.818 ± 0.113	0.427 ± 0.224
S5	0.842 ± 0.025	0.731 ± 0.032	0.840 ± 0.016	0.924 ± 0.004	0.601 ± 0.043	0.933 ± 0.000

TABLE VIII. Subject-level performance at the optimal F2 threshold

Scenario	Threshold	F1	F2	AUC	Precision	Recall
S1	0.116 ± 0.014	0.738 ± 0.027	0.843 ± 0.026	0.947 ± 0.008	0.613 ± 0.046	0.934 ± 0.047
S2	0.208 ± 0.117	0.746 ± 0.023	0.867 ± 0.020	0.947 ± 0.012	0.605 ± 0.033	0.973 ± 0.030
S3	0.160 ± 0.041	0.732 ± 0.016	0.872 ± 0.014	0.932 ± 0.012	0.578 ± 0.020	1.000 ± 0.000
S4	0.121 ± 0.035	0.736 ± 0.016	0.849 ± 0.020	0.945 ± 0.011	0.602 ± 0.020	0.947 ± 0.030
S5	0.302 ± 0.062	0.722 ± 0.008	0.860 ± 0.014	0.924 ± 0.004	0.570 ± 0.011	0.987 ± 0.027

Building on these baseline outcomes, the introduction of class weighting in S2 markedly improved sensitivity, increasing recall to 0.706 ± 0.175 and producing corresponding gains in both F1 and F2 scores. As further shown in TABLE VI. this improvement was accompanied by a slight reduction in precision due to the stronger emphasis on identifying minority-class samples. Scenarios S3 and S4, which applied fine tuning or augmentation in isolation, yielded less consistent improvements and in some cases reduced recall, suggesting that these components alone were insufficient to stabilize slice-level predictions.

As indicated in TABLE VI. the full configuration (S5) demonstrated the most recall-oriented behavior, achieving the highest F1 (0.703 ± 0.015) and F2 (0.790 ± 0.046) scores along with a substantial increase in sensitivity (0.862 ± 0.068). These results highlight that the combined use of class weighting, fine tuning, and augmentation is essential for obtaining robust and generalizable slice-level predictions.

2) Subject-Level Results: At the subject level, TABLE VII. reports the performance averaged across multiple independent runs. The baseline configuration (S1) achieved a high AUC (0.947 ± 0.008) but yielded a low and highly variable recall (0.480 ± 0.230), indicating substantial under detection of AD subjects. Building on these baseline outcomes, the introduction of class weighting in S2 produced marked improvements in both F1 and F2 scores, with recall increasing to 0.654 ± 0.117. This result demonstrates the effectiveness of class weighting in mitigating subject level class imbalance. As further shown in TABLE VII. scenarios S3 and S4, which applied fine tuning or augmentation in isolation, offered only limited benefits and in some cases reduced sensitivity, suggesting that these components alone were insufficient to stabilize performance across subjects.

As shown in TABLE VII. the complete configuration in S5 achieved the strongest and most consistent performance, yielding the highest F1 score (0.731 ± 0.032), the highest F2 score (0.840 ± 0.016), and a consistently high recall across runs (0.933 ± 0.000). These results indicate that the combined use of class weighting, fine tuning, and augmentation is essential for obtaining robust and clinically meaningful subject level predictions.In addition to these improvements, AUC values remained consistently high across all scenarios (0.932–0.947), confirming that the model preserved strong discriminative capability regardless of the specific training configuration.

TABLE IX. CONFUSION MATRIX ON S5 MODEL WITH F2 THRESHOLD OPTIMIZATION (BEST RECALL)

Class	Predicted non-AD	Predicted AD	Total
non-AD	39	12	51
AD	0	15	15
Total	41	25	66

As shown in TABLE IX. the confusion matrix for the combined model (S5), evaluated at an optimal threshold of 0.41, shows that all AD subjects were correctly classified (FN = 0), resulting in perfect recall. Although 10 non AD subjects were misclassified as AD (FP = 10), this trade off is acceptable in early screening contexts, where maximizing true positive detection is prioritized over minimizing false negatives [11], [14], [19].

While the false positives do not reduce sensitivity, they may influence follow-up clinical decision making. Misclassifying non-AD individuals can lead to unnecessary evaluations, additional imaging, or increased psychological burden for patients receiving an incorrect preliminary indication of cognitive impairment. These considerations underscore that the optimized threshold should be interpreted as a screening-oriented operating point rather than a diagnostic cutoff. Although a recall-maximizing strategy is suitable for early detection, future work should consider cost-sensitive evaluation, calibration methods, or multi-stage screening pipelines to better balance sensitivity and specificity in real-world clinical settings.

3) Comparison with Previous Works: In comparison with prior MRI based studies, the proposed framework demonstrates competitive or superior sensitivity. Nithya et al. [18] reported an F1 score of 0.71 using a ResNet 50 architecture on the OASIS 1 dataset, while Arya [6] achieved an AUC of 0.918 using a CNN ensemble. As shown by the subject level results obtained in this study, the proposed model achieved an AUC of 0.924 and an F2 score of 0.860 at the optimized threshold, indicating improved recall-oriented performance. This improvement is particularly relevant for early stage AD detection, where reducing false negatives is essential for clinical decision support [9], [12], [18].

C. Threshold Optimization

To assess the effect of threshold refinement on subject level classification, TABLE VIII. reports the performance obtained after optimizing the decision threshold to maximize the F2 score. Across all scenarios, the optimized thresholds substantially increased sensitivity, reflecting the recall oriented nature of the F2 objective. The baseline model (S1), which originally exhibited low recall, improved to 0.934 ± 0.047, with corresponding gains in both F1 and F2. Class weighting in S2 further enhanced sensitivity, producing a recall of 0.973 ± 0.030 and an F2 score of 0.867 ± 0.020. As shown in TABLE VIII. fine tuning in S3 yielded the highest F2 score (0.872 ± 0.014) and achieved perfect recall across all runs, indicating strong class separation at the subject level. Scenario S4 produced similar improvements but with slightly lower sensitivity. The complete configuration (S5) maintained high recall (0.987 ± 0.027) with a stable threshold distribution, demonstrating that the combined use of class weighting, fine tuning, and augmentation consistently promotes recall oriented behavior under threshold optimization. These findings highlight the importance of threshold selection in maximizing early detection performance in subject level classification.

D. Grad-CAM Visualization and Interpretability

To examine the spatial decision patterns learned by the model, Grad-CAM visualizations were generated from the best-performing configuration (S5). Example results are shown in Fig. 2. In AD cases, the model's activation areas were concentrated within the medial temporal lobe, particularly the hippocampus and parahippocampal gyrus, which are among the earliest regions affected by Alzheimer's-related neurodegeneration [4], [14], [20]. In contrast, non-AD subjects exhibited more diffuse and anatomically nonspecific attention patterns without a prominent focus on neurodegenerative structures.

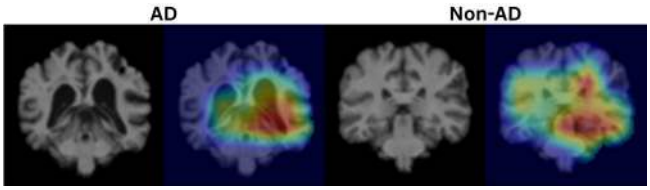

Fig. 2. Example Grad-CAM visualizations from the best-performing configuration (S5).

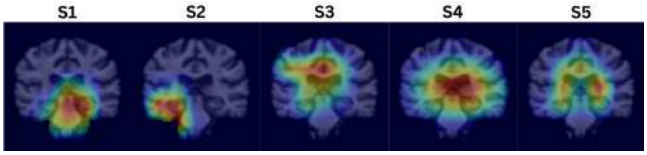

Fig. 3. Comparison of Grad-CAM activations across training scenarios (S1–S5) for the same AD subject.

To further evaluate how the training strategies affected model interpretability, Fig. 3 presents Grad-CAM activations across scenarios (S1–S5) for the same AD subject. In the early configurations (S1–S2), the attention maps already highlighted regions within the medial temporal lobe, although the activations were relatively broad and showed varying lateral emphasis between hemispheres. With the introduction of fine-tuning in S3, the model began to capture additional deep structures, including the thalamus and amygdala, while maintaining focus on the hippocampal region. In S4, the inclusion of data augmentation produced a more spatially coherent pattern centered on midline structures and extending toward both hippocampi. The final configuration (S5) yielded the most anatomically consistent activation, with balanced focus across the left and right hippocampus and parahippocampal gyrus. This spatial progression aligns with known neurodegenerative trajectories in AD, where hippocampal and parahippocampal atrophy are among the earliest and most prominent imaging biomarkers [4], [9], [14]. Moreover, the localization of attention to the thalamus and amygdala is also supported by recent XAI-based analyses, which demonstrate strong correspondence between Grad-CAM saliency and neuroanatomical substrates of AD [4]. Overall, the progression from S1 to S5 reflects increasing alignment between model attention and clinically validated AD biomarkers as the training pipeline became more refined.

While these results provide encouraging evidence of biologically meaningful attention patterns, it is important to recognize the inherent limitations of Grad-CAM. As a post hoc saliency-based approach, Grad-CAM highlights regions associated with the model's prediction but does not offer causal explanations of feature relevance. Its reliance on gradients also makes the heatmaps sensitive to architectural choices, preprocessing steps, and fine tuning dynamics, which can lead to diffuse or partially misplaced activations [11], [12]. Furthermore, the coarse spatial resolution of Grad-CAM, constrained by the granularity of the final convolutional layer, limits its ability to delineate precise anatomical boundaries and therefore restricts its utility for detailed neuroanatomical interpretation.

Given these considerations, Grad-CAM should be interpreted as supportive rather than definitive evidence of biomarker localization. Future work may incorporate complementary explainability methods or attention mechanisms that provide more anatomically detailed and spatially reliable insights into model behavior.

E. Limitations

Several limitations of this study should be acknowledged. The use of two-dimensional slices, while computationally efficient, reduces volumetric context and introduces high correlation among adjacent slices from the same subject. The OASIS-1 cohort also exhibits limited demographic diversity, and all evaluations were conducted on a single dataset, which restricts generalizability to broader clinical populations [16], [20]. Although multiple-run training improved robustness with respect to random initialization, external validation on independent cohorts such as ADNI or OASIS-3 were not performed. Furthermore, Grad-CAM provides post hoc saliency maps with limited spatial resolution [11], [12], which constrains its ability to capture fine-grained neuroanatomical patterns. These limitations indicate important opportunities for future work.

V. CONCLUSION

This study proposed a recall-oriented and interpretable deep learning framework for early detection of Alzheimer's disease using T1-weighted MRI from the OASIS-1 dataset. By integrating a pretrained ResNet-50 model with class weighting, fine-tuning, and conventional augmentation, the combined configuration achieved robust subject-level performance across multiple runs. The model attained high sensitivity after threshold optimization, highlighting its suitability for early-stage screening applications.

Grad-CAM analysis showed that network attention was concentrated in medial temporal lobe structures, including the hippocampus and parahippocampal gyrus, indicating alignment with established Alzheimer's biomarkers. These findings support the potential of the framework to provide biomarker-aligned, transparent predictions in structural MRI classification.

Future work may explore three-dimensional or hybrid architectures to incorporate volumetric context, cross-dataset validation to improve generalization, and complementary explainability techniques with higher spatial specificity. Generative augmentation and domain adaptation methods may also further address class imbalance and enhance robustness in diverse clinical settings.

ACKNOWLEDGMENT

Data were provided in part by OASIS-1: Cross-Sectional. Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

REFERENCES

[1] S. Safiri, A. Gha, and A. Fazlollahi, "Alzheimer ' s disease : a comprehensive review of epidemiology , risk factors , symptoms diagnosis , management , caregiving , advanced treatments and associated challenges".

[2] S. Dardouri, "An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks," *Front. Artif. Intell.*, vol. 8, no. April, pp. 1–11, 2025, doi: 10.3389/frai.2025.1563016.

[3] F. Momeni, D. Shahbazi-Gahrouei, T. Mahmoudi, and A. Mehdizadeh, "Transfer Learning and Neural Network-Based Approach on Structural MRI Data for Prediction and Classification of Alzheimer's Disease," *Diagnostics*, vol. 15, no. 3, 2025, doi: 10.3390/diagnostics15030360.

[4] D. Coluzzi *et al.*, "Biomarker Investigation Using Multiple Brain Measures from MRI Through Explainable Artificial Intelligence in

Alzheimer's Disease Classification," *Bioengineering*, vol. 12, no. 1, pp. 1–28, 2025, doi: 10.3390/bioengineering12010082.

[5] Z. Knapińska and J. Mulawka, "Patient-Tailored Dementia Diagnosis with CNN-Based Brain MRI Classification," *Appl. Sci.*, vol. 15, no. 9, 2025, doi: 10.3390/app15094652.

[6] L. Arya, "Comparative Evaluation of Deep Learning Models in Alzheimer 's," *Procedia Comput. Sci.*, vol. 258, pp. 2352–2361, 2025, doi: 10.1016/j.procs.2025.04.498.

[7] J. ur Rahman, M. Hanif, O. ur Rehman, U. Haider, S. Mian Qaisar, and P. Pławiak, "Stages prediction of Alzheimer's disease with shallow 2D and 3D CNNs from intelligently selected neuroimaging data," *Sci. Rep.*, vol. 15, no. 1, pp. 1–17, 2025, doi: 10.1038/s41598-025-93560-x.

[8] R. Turrisi, A. Verri, and A. Barla, "Deep learning-based Alzheimer's disease detection: reproducibility and the effect of modeling choices," *Front. Comput. Neurosci.*, vol. 18, 2024, doi: 10.3389/fncom.2024.1360095.

[9] M. Taiyeb Khosroshahi *et al.*, "Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease," *Diagnostics*, vol. 15, no. 5, 2025, doi: 10.3390/diagnostics15050612.

[10] B. A. C. Ramalho *et al.*, "The impact of the orientation of MRI slices on the accuracy of Alzheimer's disease classification using convolutional neural networks (CNNs)," *J. Med. Artif. Intell.*, vol. 7, pp. 0–3, 2024, doi: 10.21037/jmai-24-51.

[11] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, *Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review*, vol. 16, no. 1. Springer US, 2024. doi: 10.1007/s12559-023-10192-x.

[12] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, p. 102470, 2022, doi: 10.1016/j.media.2022.102470.

[13] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, "An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning," *Diagnostics*, vol. 14, no. 3, pp. 1–24, 2024, doi: 10.3390/diagnostics14030345.

[14] D. Wang *et al.*, "Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies," *Neuroimage*, vol. 269, no. January, 2023, doi: 10.1016/j.neuroimage.2023.119929.

[15] S. S. Vinayak E, S. A, and N. Khan A, "Dementia Prediction on OASIS Dataset using Supervised and Ensemble Learning Techniques," *Int. J. Eng. Adv. Technol.*, vol. 10, no. 1, pp. 244–254, 2020, doi: 10.35940/ijeat.a1827.1010120.

[16] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *J. Cogn. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. , doi: 10.1162/jocn.2007.19.9.1498.

[17] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review," *J. Imaging*, vol. 9, no. 4, 2023, doi: 10.3390/jimaging9040081.

[18] V. P. Nithya, N. Mohanasundaram, and R. Santhosh, "An Early Detection and Classification of Alzheimer's Disease Framework Based on ResNet-50," *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.*, vol. 20, pp. 1–20, 2023, doi: 10.2174/1573405620666230825113344.

[19] F. Hasan Saif, M. N. Al-Andoli, and W. M. Y. W. Bejuri, "Explainable AI for Alzheimer Detection: A Review of Current Methods and Applications," *Appl. Sci.*, vol. 14, no. 22, 2024, doi: 10.3390/app142210121.

[20] D. P. Veitch *et al.*, "The Alzheimer's Disease Neuroimaging Initiative in the era of Alzheimer's disease treatment: A review of ADNI studies from 2021 to 2022," *Alzheimer's Dement.*, vol. 20, no. 1, pp. 652–694, 2024, doi: 10.1002/alz.13449.

[21] J. Zhou *et al.*, "A deep learning model for early diagnosis of alzheimer's disease combined with 3D CNN and video Swin transformer," *Sci. Rep.*, vol. 15, no. 1, pp. 1–17, 2025, doi: 10.1038/s41598-025-05568-y.