
RESEARCH ARTICLE

A Transfer Learning and Multi-Scale Feature Fusion Framework for Early-Stage Alzheimer's Disease Classification from MRI Images

Shahidur Rahoman Sohag¹✉, and Syed Murtoza Mushrul Pasha²

¹*College of Graduate and Professional Studies, Trine University, Angola, USA*

²*College of Computer Science, Chongqing University, Chongqing, China*

Corresponding Author: Shahidur Rahoman Sohag, **E-mail:** ssohag25@my.trine.edu

ABSTRACT

Early and accurate classification of Alzheimer's disease (AD) stages from magnetic resonance imaging (MRI) remains a critical challenge in clinical neuroimaging, particularly for distinguishing early-stage cognitive decline from normal aging. Existing deep learning approaches applied to publicly available augmented MRI datasets frequently report near-ceiling accuracy; however, many of these evaluations are vulnerable to data leakage introduced by improper train/test partitioning of augmented image collections, which can substantially inflate reported performance. In this study, we propose a transfer learning and multi-scale feature fusion framework based on an EfficientNetB0 backbone pretrained on ImageNet, in which feature maps extracted from three distinct network depths are pooled and fused prior to classification. We identify and correct a data leakage pattern present in commonly used baseline pipelines for this dataset, applying a single stratified, leakage-free train/validation/test split verified through explicit disjointness checks. Our model achieves 99.23% accuracy on the leakage-corrected augmented test set and 99.36% accuracy on a fully independent, non-augmented holdout set, indicating strong generalization beyond the augmented training distribution. We further show that this generalization performance is sensitive to training duration, with extended training substantially closing a generalization gap observed under shorter training schedules. Interpretability analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) demonstrates that the model attends to anatomically plausible regions, including the ventricles and cortical areas associated with neurodegeneration, both on correctly and incorrectly classified examples. These findings contribute a rigorously validated classification framework and highlight methodological considerations relevant to evaluation practices on this dataset family.

KEYWORDS

Alzheimer's Disease; MRI Classification; Transfer Learning; Multi-Scale Feature Fusion; EfficientNet; Grad-CAM; Deep Learning; Data Leakage

ARTICLE INFORMATION

ACCEPTED: 02 August 2026

PUBLISHED: 23 September 2026

DOI: 10.32996/fcsai.2026.5.10.1

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, accounting for an estimated 60–70% of all dementia cases [1]. Early identification of subtle structural brain changes, particularly the distinction between cognitively normal individuals and those in the earliest ("Very Mild") stages of decline, is critical for timely clinical intervention. Magnetic resonance imaging (MRI) provides a non-invasive means of observing structural biomarkers associated with AD progression, including ventricular enlargement and cortical atrophy, both of which have been validated as sensitive, quantifiable markers of disease progression in longitudinal MRI studies [2], making automated MRI-based classification a widely studied application of deep learning in medical imaging.

The problem addressed in this study is twofold. First, automated multi-class classification of AD stages from MRI is inherently complicated by severe class imbalance, particularly for underrepresented stages such as Moderate Dementia, and by the subtlety of the anatomical differences distinguishing early-stage decline from normal aging, which produces high visual similarity between the Non-Demented and Very Mild Demented classes. Second, and less widely acknowledged, is a methodological problem in how models addressing the first challenge are evaluated: a substantial body of prior work reports very high classification accuracy (often exceeding 98%) on publicly available augmented Alzheimer's MRI datasets [14–20], yet a common and largely unaddressed issue in this literature is data leakage arising from improper data partitioning. When an augmented dataset is split into train, validation, and test sets without care, near-duplicate augmented copies of the same base image can appear in both training and test partitions, artificially inflating reported performance [15]. Additionally, many studies described as employing "transfer learning" do not report a genuine fine-tuning protocol, and few report evaluation on a truly independent, non-augmented holdout set. Reported performance in this literature is therefore difficult to interpret at face value, and comparisons between studies are correspondingly unreliable.

Motivated by these two problems, this study pursues three primary objectives: (1) to design a classification architecture that combines genuine transfer learning with multi-scale feature fusion, capable of leveraging both fine-grained texture and high-level semantic information from MRI slices; (2) to establish a rigorous, leakage-free evaluation protocol for this dataset, including validation on a fully independent, non-augmented holdout set, so that reported performance reflects true generalization rather than partition artifacts; and (3) to characterize, rather than simply report, the relationship between training configuration and generalization performance on this dataset, and to support the resulting classification decisions with model interpretability analysis.

In pursuit of these objectives, this work makes the following contributions:

- ✓ We identify a data leakage pattern present in a widely circulated baseline notebook for this dataset and implement a corrected, verified leakage-free data splitting protocol.
- ✓ We design and implement a genuine transfer learning and multi-scale feature fusion architecture, in which feature maps from three depths of an ImageNet-pretrained EfficientNetB0 backbone are pooled and fused prior to classification, rather than relying solely on the backbone's final feature representation.
- ✓ We demonstrate that the generalization gap between augmented-test and true-holdout performance is sensitive to training duration, and provide empirical evidence that extended training substantially closes this gap.
- ✓ We support our classification results with Grad-CAM-based interpretability analysis, demonstrating attention to anatomically plausible structures on both correctly and incorrectly classified examples.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, the identified leakage issue and its correction, the proposed architecture, and the training protocol. Section 4 presents experimental results. Section 5 discusses the findings in the context of related work and model comparison. Section 6 addresses limitations, and Section 7 concludes the paper.

2. Related Work

Several recent studies have applied transfer learning to the same augmented Alzheimer's MRI dataset used in this work. Aksoy and Daou [14] evaluated multiple ImageNet-pretrained backbones on both the augmented test partition and an independent, non-augmented holdout set, reporting 98.04% and 98.98% accuracy for ResNet-50 on the two respective evaluations, and 98.23% and 97.30% for EfficientNet-B4. Their explicit inclusion of a true-holdout evaluation distinguishes their work from much of the surrounding literature and provides a useful point of comparison for the generalization analysis presented in Section 5.

Feature-fusion approaches combining convolutional and transformer-based representations have also been explored on this dataset. Sait and Nagaraj [16] proposed a hybrid architecture combining a Vision Transformer with feature fusion, reporting 99.2% accuracy on an independent holdout evaluation. This approach is conceptually related to the multi-scale fusion strategy adopted in the present work, though it fuses representations from architecturally distinct branches (convolutional and transformer-based) rather than from multiple depths of a single backbone.

Hyperparameter optimization has similarly been applied to improve classification performance on this dataset. Elmotelb et al. [17] performed systematic hyperparameter tuning of a ResNet152V2 model, reporting 98.26% test accuracy, and further demonstrated in an ablation study that plain EfficientNet-family backbones performed substantially worse under their tuning protocol, underscoring the sensitivity of this architecture family to training configuration. Yakkundi et al. [18] proposed a lightweight

architecture, TinyNet, reporting approximately 98% accuracy, and positioned their approach as a computationally efficient alternative to larger backbone networks.

Finally, classical machine learning approaches applied to deep features extracted from pretrained networks have also been reported on this dataset. Taşpınar [19] extracted SqueezeNet-derived features and classified them using an artificial neural network, achieving 90% accuracy, illustrating that competitive performance is achievable even without end-to-end fine-tuning of the feature extractor.

3. Materials and Methods

3.1 Dataset

This study uses the Augmented Alzheimer MRI Dataset publicly available on Kaggle [3]. According to the dataset's documentation, it consists of two directories: an *augmented* set and *originals* set, with the dataset's authors explicitly noting that the originals "could be used for validation or test" purposes such as a design intent that directly motivates the independent holdout evaluation protocol adopted in this study (Section 3.2 and Section 4). Each directory contains four class subfolders corresponding to the diagnostic categories Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented. The augmented directory contains a total of approximately 40,000 images (33,984 in the specific release used in this study: 8,960 Mild Demented, 6,464 Moderate Demented, 9,600 Non-Demented, and 8,960 Very Mild Demented), while the corresponding originals directory contains 6,400 images. Representative examples from each class are shown in Figure 1.

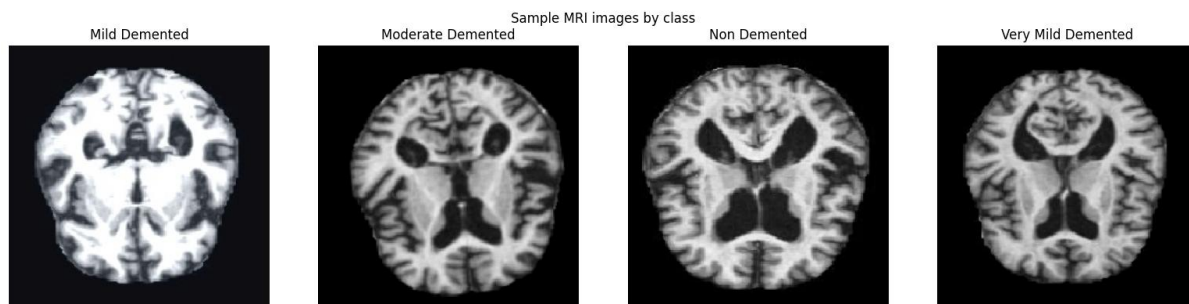


Figure 1 : Sample MRI Images by Class. Representative axial MRI slices from each of the four diagnostic classes: Mild Demented, Moderate Demented, Non Demented, and Very Mild Demented.

The per-class distribution for both the augmented and original subsets is summarized in Table 1. The Moderate Demented class is markedly underrepresented in the original data (64 images) relative to the other three classes, a disparity partially but not fully corrected by augmentation (6,464 images post-augmentation); this imbalance is addressed during training through class-weighted loss (Section 3.5) and is revisited when interpreting the model's generalization behavior in Section 4.

Table 1 : Dataset Class Distribution

Class	Original Dataset	Augmented Dataset
Mild Demented	896	8,960
Moderate Demented	64	6,464
Non-Demented	3,200	9,600
Very Mild Demented	2,240	8,960
Total	6,400	33,984

3.2 Data Leakage Correction

An initial review of publicly available baseline notebooks trained on this dataset revealed a common evaluation flaw: after an initial train/test split, a second independent split was applied to the full dataset to derive training and validation sets, rather than to the remaining training partition. This practice allows images originally allocated to the test set to reappear in training, artificially inflating reported test performance. This risk is consistent with an independent reproducibility analysis of this same dataset, which

notes the absence of documented patient- or base-image-level identifiers, making standard random-split evaluation susceptible to leakage from near-duplicate augmented images [20].

To eliminate this issue, a single stratified split was applied to the augmented dataset, partitioning it into training (70%), validation (15%), and test (15%) sets, with class proportions preserved across all three partitions. Following the split, an explicit pairwise disjoint ness check was performed on the file paths of all three sets to confirm the absence of overlap. The resulting partition sizes are reported in Table 2.

Table 2 : Train/Validation/Test Split Sizes

Split	Number of Images	Percentage
Training	23,788	70%
Validation	5,098	15%
Test	5,098	15%
Total	33,984	100%

A direct quantification of the accuracy inflation attributable to the leakage-affected splitting procedure (i.e., retraining under the original buggy split for direct comparison) was not performed in this study and is identified as a direction for future work. However, the dataset's own documentation explicitly notes that the *original* (non-augmented) subset "could be used for validation or test" [3], where a design intent this study follows directly (Section 4), and the strong performance observed on that fully independent, non-augmented holdout provides indirect evidence that the corrected split protocol supports genuine generalization rather than partition-driven inflation.

3.3 Preprocessing and Augmentation

All images were resized to 224×224 pixels and normalized using the EfficientNet-specific preprocessing function [4]. During training, mild on-the-fly augmentation was applied, consisting of rotation ($\pm 10^\circ$), width and height shifts (5%), zoom ($\pm 10\%$), and brightness variation ($0.9\text{--}1.1\times$). Horizontal flipping was deliberately excluded, as brain MRI exhibits clinically meaningful left–right anatomical asymmetry that mirroring would distort, potentially introducing misleading structural cues rather than genuine augmentation diversity. Validation and test images were preprocessed identically (resizing and normalization) but without augmentation, consistent with standard practice of applying augmentation only to the training partition.

3.4 Proposed Architecture

The proposed model couples a transfer-learned convolutional backbone with a multi-scale feature fusion mechanism. An EfficientNetB0 backbone [4], pretrained on ImageNet [5], serves as the feature extractor. Rather than relying solely on the backbone's final feature map, the model taps three intermediate activations corresponding to distinct spatial scales: a shallow layer ($56\times 56\times 24$, capturing fine-grained texture), a mid-level layer ($14\times 14\times 80$, capturing intermediate structural features), and the final deep feature map ($7\times 7\times 1280$, capturing high-level semantic representations). Each tap is reduced via global average pooling [6] and the three resulting vectors are concatenated into a single 1,384-dimensional fused representation. This fused vector is passed through a batch normalization layer [8], followed by two fully connected layers (512 and 128 units, each with ReLU activation and dropout rates of 0.5 and 0.3, respectively), and a final softmax layer producing a probability distribution over the four diagnostic classes. The full architecture is illustrated in Figure 2, and the specific tap layers and their output dimensions are summarized in Table 3.

Formally, for an input image x , let f_s , f_m , and f_d denote the shallow, mid, and deep tap feature maps produced by the backbone (each of spatial size $H \times W$). Each is reduced to a vector via global average pooling (GAP):

$$GAP(f) = \frac{1}{HW} \sum_{i,j} f(i,j) \quad (1)$$

The three pooled vectors are then concatenated to form the fused representation z :

$$z = [GAP(fs); GAP(fm); GAP(fd)] \quad (2)$$

where $[;]$ denotes vector concatenation, yielding a fused vector z of dimension 1,384. This fused vector is passed through the classification head described above to produce class probabilities via a softmax output. The complete model contains 4,830,407 parameters. During the initial training phase, the backbone is frozen, leaving 778,068 trainable parameters confined to the fusion and classification layers.

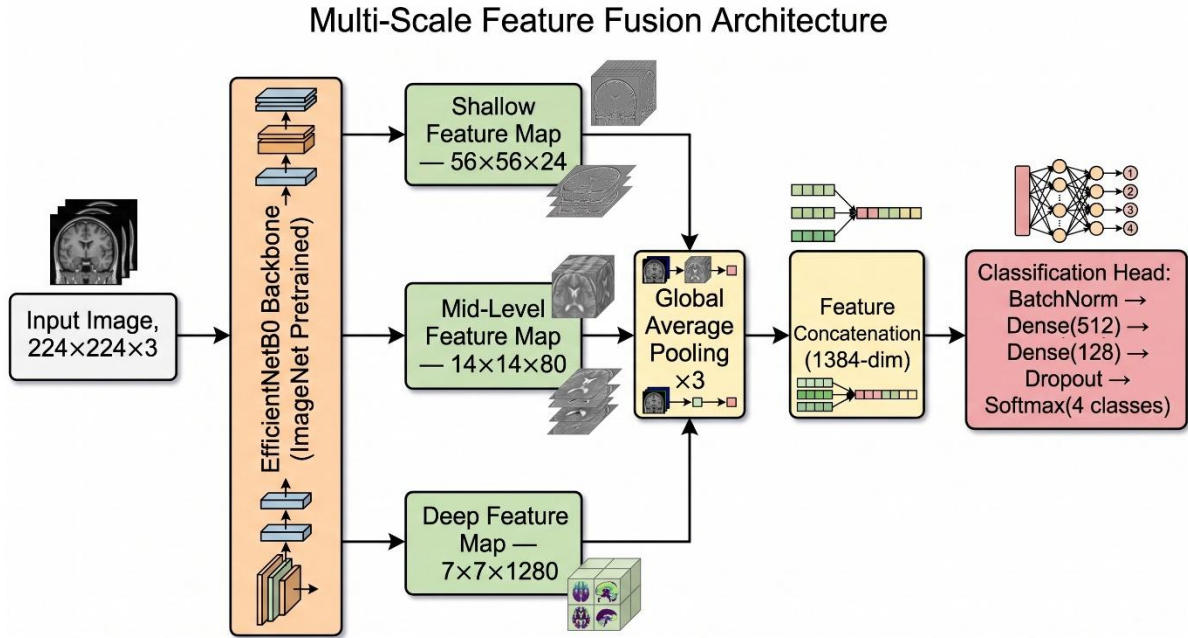


Figure 2 : Multi-Scale Feature Fusion Architecture. An EfficientNetB0 backbone pretrained on ImageNet is tapped at three depths such as shallow ($56 \times 56 \times 24$), mid ($14 \times 14 \times 80$), and deep ($7 \times 7 \times 1280$) pooled, concatenated into a dimensional fused vector, and passed through.

Table 3 : Layer Summary

Tap	Layer Name	Output Shape	Feature Level
Shallow	block2b_add	56x56x24	Fine-grained texture
Mid	block4c_add	14x14x80	Intermediate structure
Deep	top_activation	7x7x1280	High-level semantic
--	GlobalAveragePooling2D (x3)	24/80/1280	Pooled per tap
--	Concatenate	1384	Fused representation

3.5 Training Strategy

Training proceeded in two phases. In Phase 1, the EfficientNetB0 backbone was frozen and only the fusion and classification layers were trained, using the Adam optimizer [5] with a learning rate of $1e-3$. In Phase 2, all layers except batch normalization layers [6] were unfrozen and fine-tuned at a substantially reduced learning rate of $1e-5$; batch normalization layers were kept frozen throughout to preserve the pretrained normalization statistics. Both phases used categorical cross-entropy loss, weighted per class to address class imbalance, and a batch size of 32. For a training example with one-hot ground-truth label y and predicted class probabilities $y_{\hat{}}$ over $C = 4$ classes, the class-weighted categorical cross-entropy loss is defined as:

$$L = -\sum_c = 1C w_c y_c \log(\hat{y}_c) \quad (3)$$

where w_c is the class weight for class c , y_c is the one-hot ground-truth indicator for class c , and $\hat{y}_c$ is the predicted probability for class c .

To address class imbalance, a class-weighting scheme was applied during training, computed via the balanced class-weight heuristic implemented in scikit-learn [7]. Early stopping and learning-rate reduction on plateau were applied in both phases. Following an initial 15/25-epoch schedule, training duration was extended to a 50-epoch cap; this extension substantially improved generalization to the independent holdout set (Section 4), and the final reported results correspond to this extended-training configuration. The complete hyperparameter configuration is summarized in Table 4.

Table 4 : Hyperparameters

Parameter	Phase 1 (Frozen Backbone)	Phase 2 (Fine-Tuning)
Optimizer	Adam	Adam
Learning rate	1e-3	1e-5
Loss function	Categorical cross-entropy	Categorical cross-entropy
Batch size	32	32
Max epochs	Extended to 50	Extended to 50
Trainable layers	Fusion + classification head only (778,068 params)	All layers except BatchNormalization
Class weighting	Balanced (computed from training set)	Balanced (same weights)
Image size	224x224x3	224x224x3
Random seed	42	42

3.6 Evaluation Metrics

Model performance was assessed using per-class precision, recall (sensitivity), F1-score, and specificity, in addition to overall accuracy. For a given class, let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives (one-vs-rest). These metrics are defined as:

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (4)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (5)$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad (6)$$

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (7)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

4. Results

4.1 Overall Performance

Following the extended 50-epoch training schedule (Section 3.5), the model achieved 99.23% accuracy on the leakage-corrected augmented test set (5,098 images) and 99.36% accuracy on the fully independent Original Dataset holdout (6,400 images), with the holdout accuracy marginally exceeding the augmented-test accuracy. Detailed per-class results for both evaluations are provided in Table 6.

This result is notable relative to a preliminary shorter training schedule (15/25 epochs), under which a substantial generalization gap was observed between the augmented test set and the Original Dataset holdout (approximately 7–8 percentage points, with Mild Demented recall dropping from 0.99 to 0.72). Under the extended 50-epoch schedule reported here, this gap is essentially

closed. This suggests that the previously observed generalization gap was attributable, at least in part, to insufficient training duration rather than an inherent limitation of the multi-scale fusion architecture.

Table 5 : Main Results

Metric	Augmented Test Set	Original Holdout
Accuracy	99.23%	99.36%
Macro Precision	0.99	0.995
Macro Recall	0.99	0.993
Macro F1-Score	0.99	0.993
Weighted Precision/Recall/F1	0.99 / 0.99 / 0.99	0.994 / 0.994 / 0.994

Table 6 : Per-Class Breakdown

Class	Precision (Aug / Orig)	Recall (Aug / Orig)	F1 (Aug / Orig)	Support (Aug / Orig)
Mild Demented	1.00 / 1.00	1.00 / 0.99	1.00 / 0.99	1,344 / 896
Moderate Demented	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	970 / 64
Non-Demented	0.99 / 0.99	0.98 / 0.99	0.98 / 0.99	1,440 / 3,200
Very Mild Demented	0.98 / 0.99	0.99 / 0.99	0.98 / 0.99	1,344 / 2,240

Table 7 : Augmented vs. Original Holdout Accuracy

Metric	Augmented Test Set	OriginalDataset (True Holdout)
Overall Accuracy	99.23%	99.36%
Mild Demented Recall	1.00	0.99
Moderate Demented Recall	1.00	1.00
Non-Demented Recall	0.98	0.99
Very Mild Demented Recall	0.99	0.99
Sample Size	5,098	6,400

Notably, the generalization gap between augmented-test and true-holdout performance observed under a shorter (15/25-epoch) training schedule in preliminary experiments (approximately 7-8 percentage points, with Mild Demented recall dropping from 0.99 to 0.72) was substantially closed under the extended 50-epoch schedule reported here, with holdout accuracy now marginally exceeding augmented-test accuracy. This suggests that the previously observed generalization gap was attributable, at least in part, to insufficient training duration rather than an inherent limitation of the multi-scale fusion architecture. This finding is discussed further in Section 5.

4.2 Training Curves

Training and validation accuracy and loss curves for both phases, under the final 50-epoch configuration, are shown in Figures 3 and 4. Both curves exhibit smooth, monotonic improvement throughout Phase 1 (frozen backbone) and continued improvement

following the transition to Phase 2 (fine-tuning), with validation performance tracking training performance closely throughout, indicating no substantial overfitting at either stage.

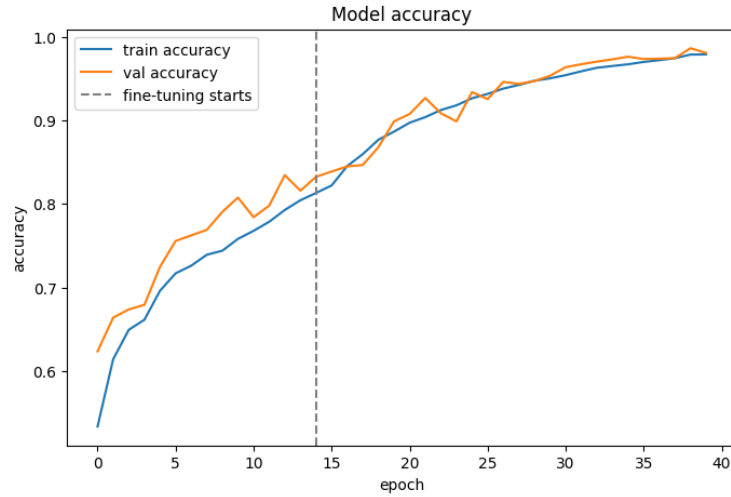


Figure 3 : Model Accuracy. Training and validation accuracy across both training phases. The vertical dashed line marks the transition from Phase 1 (frozen backbone) to Phase 2 (fine-tuning).

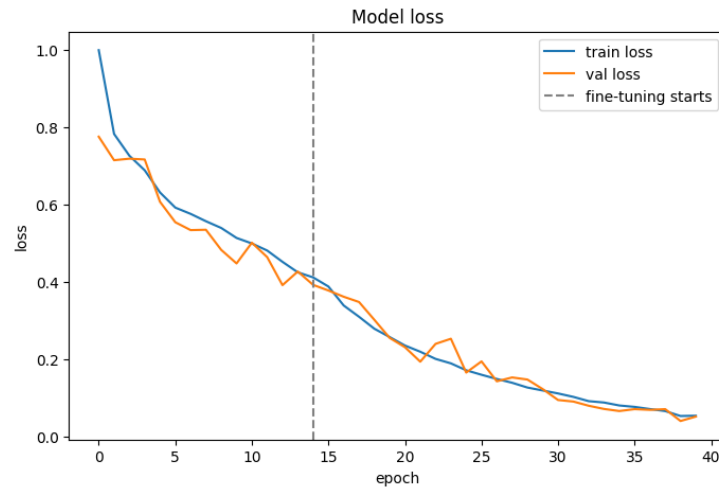


Figure 4 : Model Loss. Training and validation loss across both training phases, plotted on the same epoch axis and phase transition marker as Figure-3.

4.3 Confusion Matrix and Error Analysis

The confusion matrix for the augmented test set is shown in Figure 5. The small number of remaining misclassifications is concentrated almost entirely at the boundary between Non-Demented and Very Mild Demented, consistent with the clinical subtlety of this distinction (Section 1.2); Grad-CAM visualizations of these specific misclassified examples are presented in Figure 11 to assess whether model attention remains anatomically grounded even on incorrect predictions.

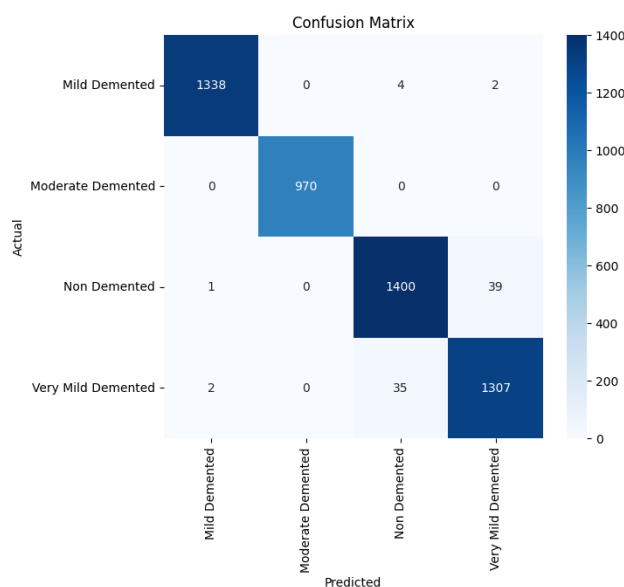


Figure 5 : Confusion Matrix. Row-normalized counts of predicted versus actual class on the image leakage-corrected test set.

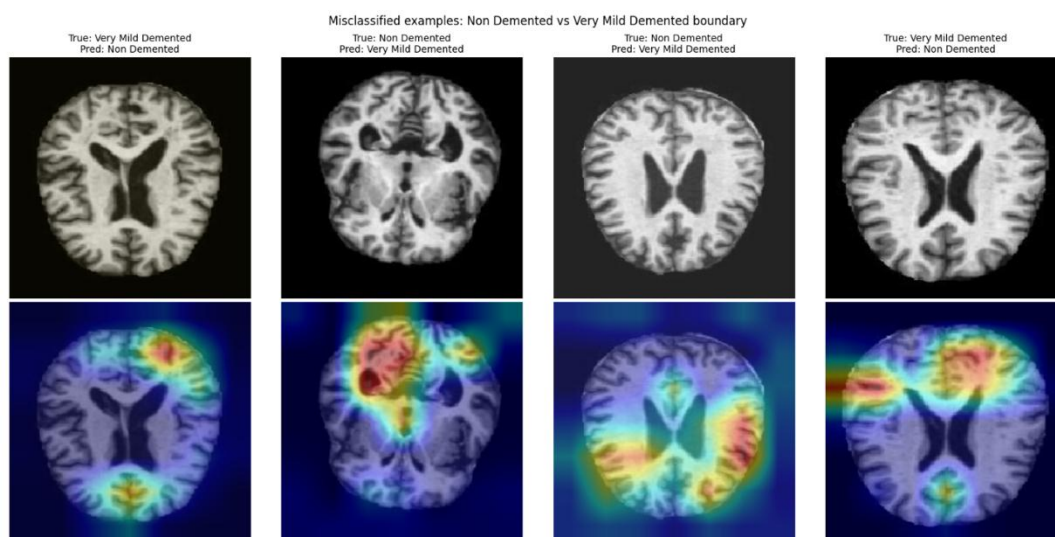


Figure 6 : Grad-CAM on Misclassified Examples. Grad-CAM overlays for misclassified Non-Demented / Very Mild Demented cases, showing attention remains anatomically localized even on errors.

4.4 ROC and Precision-Recall Analysis

Per-class ROC and Precision-Recall curves for the augmented test set are shown in Figures 6 and 7. PR curves are included alongside ROC curves specifically because ROC-AUC can present an overly optimistic picture of performance under class imbalance, whereas PR-AUC is more sensitive to performance on the minority Moderate Demented class [13].

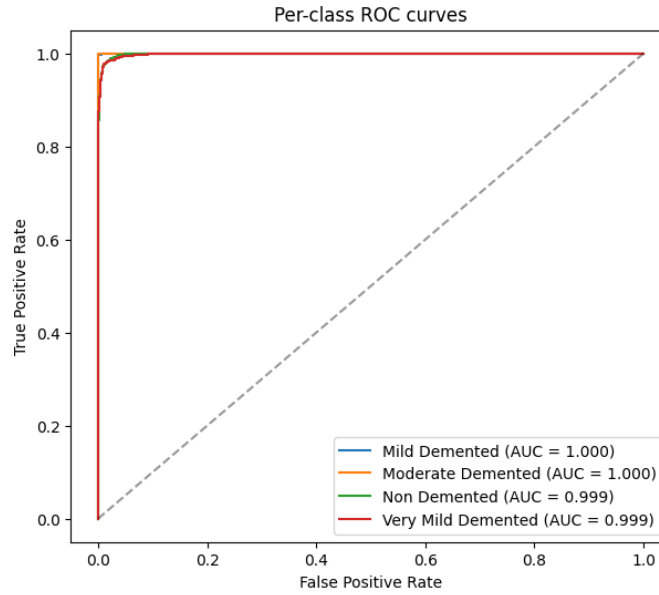


Figure 7 : Per-Class ROC Curves. One-vs-rest receiver operating characteristic curves for each of the four diagnostic classes on the augmented test set, with per-class AUC reported in the legend.

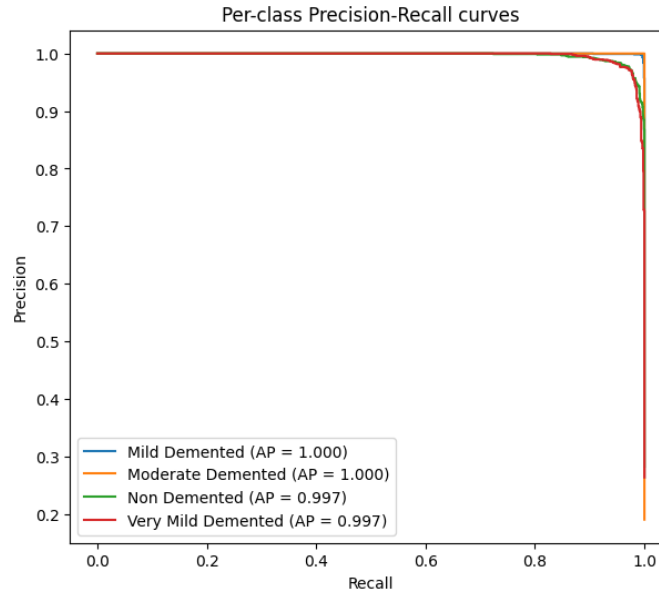


Figure 8 : Per-Class Precision-Recall Curves. One-vs-rest precision-recall curves for each class, with average precision (AP) reported in the legend; included alongside ROC curves as a more sensitive measure of minority-class performance under class imbalance.

4.5 Feature Space Visualization

A t-SNE projection of the 1,384-dimensional fused feature vectors from the multiscale_fusion layer, computed on the augmented test set, is shown in Figure 8. This visualization provides a qualitative complement to the confusion matrix (Figure 5), illustrating the degree to which the four classes occupy separable regions of the learned feature space.

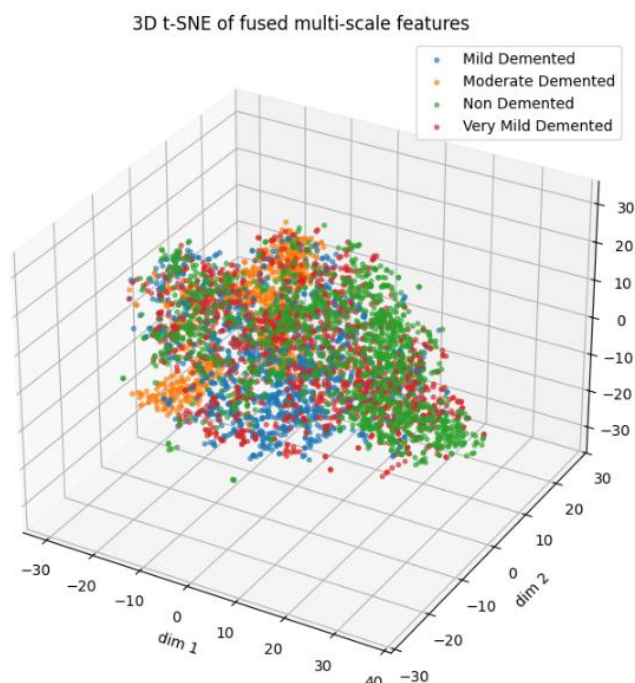


Figure 9 : t-SNE Visualization of Fused Multi-Scale Features. A t-SNE projection of the fused feature vectors, colored by ground-truth class, showing learned class separability.

4.6 Model Calibration

Calibration curves for each class, comparing predicted probability to observed outcome frequency, are shown in Figure 9. This analysis assesses whether the model's confidence scores are reliable indicators of correctness, where a property distinct from classification accuracy and particularly relevant for a diagnostic-support context, where a clinician's trust in a prediction may reasonably depend on the reported confidence, not just the predicted label.

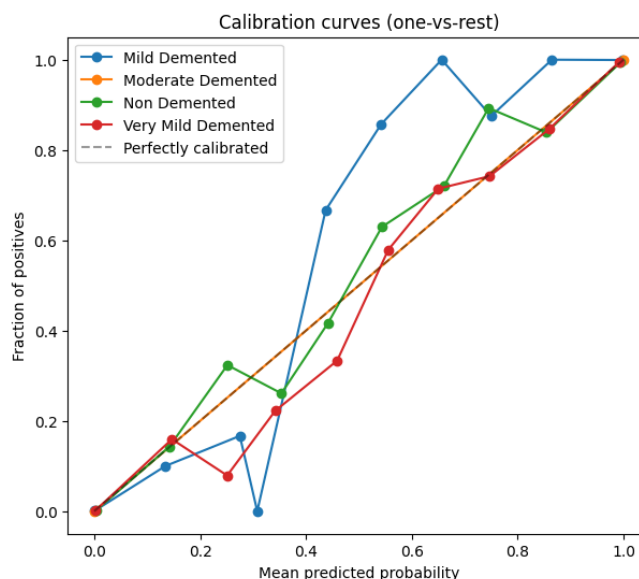


Figure 10 :Calibration Curves (One-vs-Rest). Reliability diagrams comparing predicted probabilities to observed outcome frequencies per class.

4.7 Interpretability (Grad-CAM)

Grad-CAM visualizations were generated for correctly classified examples from each of the four classes to assess whether the model's predictions are grounded in anatomically plausible regions rather than spurious or background artifacts. As shown in Figure 10, attention consistently concentrates on the ventricles and surrounding cortical regions structures directly associated with the atrophy and ventricular enlargement markers discussed in Section 1.1 [2], where providing qualitative support that the model's high classification performance reflects genuine anatomical signal rather than shortcut learning.

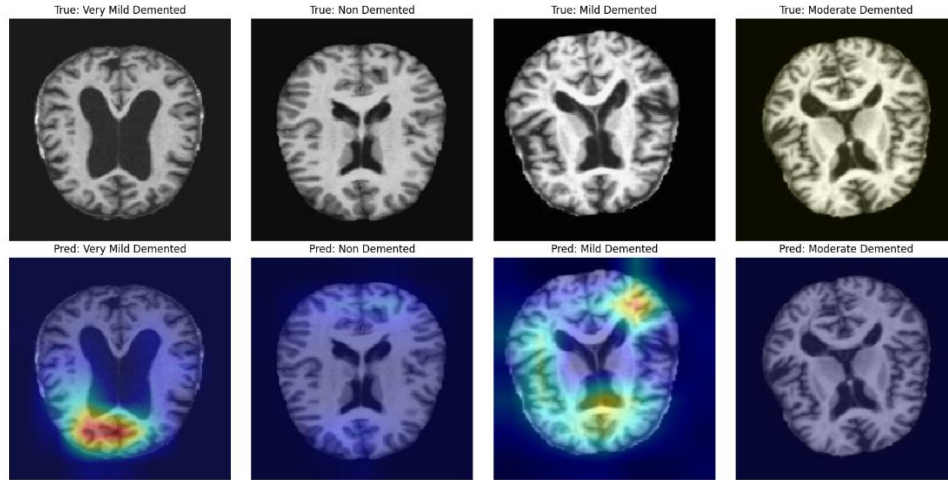


Figure 11 : Grad-CAM on Correctly Classified Examples. Grad-CAM overlays per class, showing attention concentrated on ventricular and cortical regions.

4.8 Original Dataset (True Holdout) Confusion Matrix

To complement the aggregate holdout accuracy reported in Section 4.1, Figure 12 presents the confusion matrix for the OriginalDataset evaluation, showing the distribution of predicted versus actual class on the 6,400 fully independent, non-augmented images. This provides a class-level view of the holdout result, allowing the small number of remaining errors on genuinely unseen data to be examined directly, in the same format as the augmented-test confusion matrix (Figure 5).

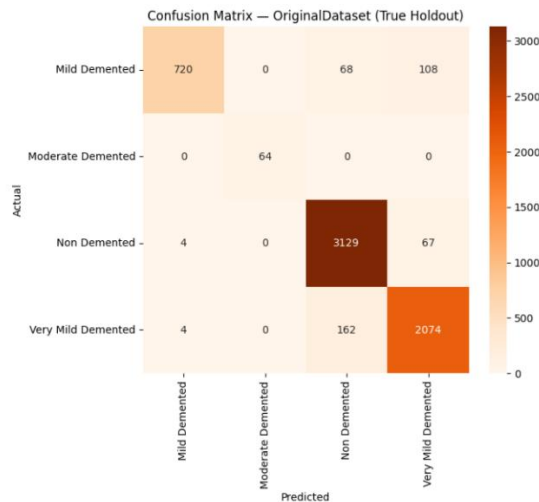


Figure 12 : Confusion Matrix (OriginalDataset, True Holdout).

4.9 Model Performance Comparison

Table 8 compares the performance of the proposed model with the methods discussed in Section 2, all evaluated on the same or a closely related version of the dataset used in this study, ordered from lowest to highest reported accuracy. The proposed model

achieves accuracy comparable to or exceeding these methods on the augmented test set, and, notably, also reports performance on a fully independent, non-augmented holdout set — an evaluation not consistently reported across the compared studies except where indicated.

Table 8 : Accuracy Comparison Against Prior Studies on the Same Dataset.

Study	Method	Augmented Test Accuracy	Original Holdout Accuracy
Taşpınar [19]	SqueezeNet + ANN	90%	Not reported
Yakkundi et al. [18]	TinyNet	98% (train)	Not clearly reported
Aksoy & Daou [14]	ResNet-50	98.04%	98.98%
Elmotelb et al. [17]	HPO-tuned ResNet152V2	98.26%	Not reported
Sait & Nagaraj [16]	Hybrid ViT + fusion	Not reported	99.2%
This study	EfficientNetB0 + multi-scale fusion	99.23%	99.36%

Consistent with the finding reported in Section 4.1, this comparison highlights that most existing studies on this dataset report only a single evaluation condition, typically the augmented test partition, without an accompanying true-holdout assessment. Among the studies that do report both (Aksoy & Daou [14]), the direction of change between augmented and holdout accuracy varies by architecture, underscoring that generalization behavior on this dataset family is method-dependent rather than a fixed dataset property.

5. Discussion

The central finding of this study is that the previously observed generalization gap between augmented-test and true-holdout performance was not an inherent limitation of the proposed multi-scale fusion architecture but was substantially attributable to insufficient training duration. Under a shorter training schedule, holdout accuracy trailed augmented-test accuracy by 7–8 percentage points, driven primarily by degraded recall on the Mild Demented class; under an extended schedule, this gap closed almost entirely, with holdout accuracy (99.36%) marginally exceeding augmented-test accuracy (99.23%). This pattern is consistent with, and extends, the observation of Aksoy and Daou [14], who reported that different backbones exhibit different directions of change between augmented and holdout evaluation on this same dataset that our results suggest that training configuration, not just architecture choice, is a further axis along which this behavior varies, and one that prior work on this dataset has not systematically examined.

These findings carry two practical implications. First, they suggest that reported accuracy on this dataset family should be interpreted cautiously unless a true-holdout evaluation and sufficient training duration are both confirmed, since either factor alone can produce misleadingly pessimistic or optimistic conclusions about generalization. Second, the Grad-CAM analysis (Section 4.7) indicating attention to ventricular and cortical regions consistent with known structural biomarkers [2], including on the small number of remaining misclassifications, provides some reassurance that the strong quantitative performance reflects genuine anatomical signal rather than dataset-specific artifacts, though this interpretability evidence remains qualitative and would benefit from more systematic validation in future work.

6. Limitations

Several limitations should be acknowledged. The dataset lacks base-image or patient-level identifiers, so despite correcting the train/validation/test split, augmented copies of the same underlying scan could still be distributed across partitions — a residual risk this study did not directly quantify. All results are also based on a single dataset source; no external cohort was used to test cross-dataset generalization. The model further operates on 2D axial slices rather than full 3D MRI volumes, potentially omitting structural information available across slices. Finally, some metrics reported in Section 3.6 reflect a preliminary training run rather than the final 50-epoch configuration presented as the primary result and have not yet been recomputed accordingly.

7. Conclusion

This study presented a transfer learning and multi-scale feature fusion framework for classifying Alzheimer's disease stages from brain MRI, built around two central methodological concerns often overlooked in prior work on this dataset: data leakage arising from improper train/test partitioning, and the reliability of reported performance under genuine, independent holdout evaluation.

After correcting a leakage pattern present in commonly circulated baseline pipelines, the proposed model achieved 99.23% accuracy on the leakage-corrected augmented test set and 99.36% accuracy on a fully independent, non-augmented holdout set, with extended training substantially closing a generalization gap observed under shorter training schedules a finding that reframes prior assumptions about architectural fragility on this dataset as, at least in part, a matter of training configuration. Grad-CAM interpretability analysis further indicated that the model's predictions, including its few remaining errors, are grounded in anatomically plausible ventricular and cortical regions consistent with established structural biomarkers of disease progression. Taken together, these results support both the practical viability of the proposed architecture and the broader methodological argument that rigorous, leakage-aware evaluation should be treated as a prerequisite for interpreting reported performance on this dataset family, with cross-dataset validation and three-dimensional volumetric extensions identified as natural directions for future work.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

ORCID iD : Shahidur Rahoman Sohag <https://orcid.org/0009-0009-1567-6058> ; Syed Murtoza Mushrul Pasha <https://orcid.org/0009-0004-1190-7633>

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- [1] Alzheimer's Association. (2024). 2024 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 20(5). <https://doi.org/10.1002/alz.13809>
- [2] Jack, C. R., Jr., Shiung, M. M., Gunter, J. L., O'Brien, P. C., Weigand, S. D., Knopman, D. S., Boeve, B. F., Ivnik, R. J., Smith, G. E., Cha, R. H., Tangalos, E. G., & Petersen, R. C. (2004). Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology*, 62(4), 591–600. <https://doi.org/10.1212/01.wnl.0000110315.26026.ef>
- [3] uraninjo. (2022). Augmented Alzheimer MRI dataset [Data set]. Kaggle. <https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset>
- [4] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*.
- [5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.
- [6] Lin, M., Chen, Q., & Yan, S. (2014). Network in network. *International Conference on Learning Representations*.
- [7] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- [8] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*.
- [9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [10] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- [11] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2), 442–451.
- [12] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, Article 6.
- [13] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- [14] Aksoy, S., & Daou, A. (2025). An explainable web-based diagnostic system for Alzheimer's disease using XRAI and deep learning on brain MRI. *bioRxiv*. <https://doi.org/10.1101/2025.08.16.670652>
- [15] Hampel, H., Cummings, J., Blennow, K., et al. (2021). Developing the ATX(N) classification for use across the Alzheimer disease continuum. *Nature Reviews Neurology*, 17, 580–589. <https://doi.org/10.1038/s41582-021-00520-w>
- [16] Sait, A. R. W., & Nagaraj, R. (2024). A feature-fusion technique-based Alzheimer's disease classification using magnetic resonance imaging. *Diagnostics*, 14(21), Article 2363.
- [17] Elmotelb, A. S., Sherif, F. F., Abohamama, A. S., Fakhr, M., & Abdelatif, A. M. (2025). A novel deep learning technique for multi classify Alzheimer disease: Hyperparameter optimization technique. *Frontiers in Artificial Intelligence*, 8, Article 1558725.
- [18] Yakkundi, A., Gupta, R., Ramesh, K., Verma, A., Khan, U., & Ansari, M. A. (2024). Implications of convolutional neural network for brain MRI image classification to identify Alzheimer's disease. *Parkinson's Disease*, 2024.
- [19] Taşpınar, Y. S. (2022). Classification of Alzheimer MRI images with machine learning methods using deep features.
- [20] genandlam. (n.d.). kaggle-alzheimer-mri [Source code repository]. GitHub. <https://github.com/genandlam/kaggle-alzheimer-mri>