

An Efficient and Reliability-Oriented Deep Learning Framework for Alzheimer's MRI Classification

Yashwant A.*, Rajaprakash S.

Department of Computer Science and Engineering,

Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 600062, Tamil Nadu, India

Abstract—Reliable Alzheimer's disease (AD) magnetic resonance imaging (MRI) classification requires evaluation beyond accuracy alone. This study investigates a four-class image-level framework based on ImageNet-pretrained ResNet-18 using 6,400 public MRI images labelled Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. Original images were assigned to seventy percent training, fifteen percent validation, and fifteen percent test partitions before augmentation. The selected no-CLAHE pipeline used training-only transformations, 128 by 128 pseudo-RGB inputs, channel normalization with mean and standard deviation of 0.5, a 512 to 256 to 4 classifier, class-weighted cross-entropy with label smoothing, AdamW optimization, and five fixed seeds. Mean accuracy was 98.38 percent with a standard deviation of 0.38 percent, balanced accuracy was 98.86 percent with a standard deviation of 0.40 percent, macro F1-score was 90.73 percent with a standard deviation of 1.79 percent, and macro-AUC was 0.9993 with a standard deviation of 0.0009. For seed 42, validation-fitted temperature scaling reduced negative log-likelihood, Brier score, and expected calibration error without changing predicted labels. The model contains 11.31 million trainable parameters and achieved 3.08 ms batch-1 inference on a Tesla T4. Grad-CAM supported qualitative inspection. The results establish a reproducible, confidence-aware, perturbation-evaluated, and resource-efficient image-level framework within the stated public-dataset protocol.

Keywords—Alzheimer's disease; magnetic resonance imaging; ResNet-18; confidence calibration; robustness; Grad-CAM; computational efficiency

I. INTRODUCTION

Alzheimer's disease (AD) is a major cause of dementia, and structural magnetic resonance imaging (MRI) is widely used to investigate disease-related brain changes. Deep learning has been increasingly applied to medical-image representation learning and neuroimaging classification [1]–[3]. In healthcare analytics, however, a reported performance value is meaningful only when the data unit, split protocol, class prevalence, preprocessing sequence, and separation between validation and test data are clearly described. These considerations are especially important for public image collections in which subject identifiers and acquisition metadata may be unavailable [3], [4].

Residual networks use skip connections to facilitate optimization of convolutional models [5]. An ImageNet-pretrained ResNet-18 can be adapted to pseudo-RGB MRI inputs through transfer learning [6], [7]. In imbalanced medical-image collections, training-only augmentation, class-aware loss functions,

and macro-averaged evaluation may provide more useful information than accuracy alone [8], [9].

Contrast-limited adaptive histogram equalization (CLAHE) was considered as a candidate local-contrast transformation because adaptive histogram equalization improves local contrast and CLAHE limits excessive local histogram amplification [10], [11]. This rationale does not imply that CLAHE must improve an AD-MRI classifier. Contemporary MRI-based AD studies use heterogeneous image representations, architectures, and validation designs [12]–[17]; preprocessing choices should therefore be evaluated on the target collection rather than assuming that they are beneficial *a priori*.

This work evaluates a reliability-oriented and resource-efficient ResNet-18 framework for the public labels Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. In this study, “reliability-oriented” refers to the combined use of five-seed reproducibility, probability calibration, selective prediction, controlled perturbation testing, qualitative model inspection, and computational profiling as complementary technical checks within the evaluated protocol. The selected configuration preserves the resized grayscale intensity structure and uses online training augmentation. The study makes six contributions:

- It documents a fixed stratified image-level 70/15/15 partition of 6,400 original images before stochastic transformation and reports five-seed stability.
- It addresses substantial class imbalance through class-weighted optimization and class-sensitive held-out metrics, with explicit reporting of the nine-image Moderate Demented test subset.
- It reports a transparent configuration-level comparison between no CLAHE with online augmentation and CLAHE without augmentation, without claiming isolated causal effects.
- It applies Grad-CAM to a representative 0.5/0.5 normalization checkpoint for qualitative model inspection while avoiding anatomical or clinical localization claims.
- It evaluates validation-fitted temperature scaling and confidence-ranked selective prediction to characterize probability reliability and risk-coverage behaviour.
- It quantifies sensitivity to controlled image perturbations and directly measures parameters, multiply-accumulate operations, latency, throughput, checkpoint size, and peak inference memory.

*Corresponding author.

II. RELATED WORK AND MOTIVATION

A. Deep Learning for AD MRI Classification

MRI-based AD classification has been investigated using 2D and 3D convolutional networks, transfer-learning methods, multi-slice systems, attention mechanisms, and transformer-based models [12]–[17]. Direct numerical comparison across these studies is difficult because cohorts, image units, imaging protocols, label definitions, class prevalences, and validation regimes vary. A model tested on independently partitioned participants or complete volumes is not directly comparable with a model tested on individual images from a public folder structure.

Wen *et al.* emphasized the importance of carefully defined validation protocols for reproducible MRI-classification studies [3]. Ebrahimighannavieh *et al.* similarly highlighted the value of clearly specified preprocessing and evaluation procedures in neuroimaging deep learning [4]. Accordingly, this study reports reproducible image-level experimentation and retains the public dataset labels as supplied, while keeping the interpretation aligned with the evaluated protocol.

B. Online Augmentation and Class-Sensitive Evaluation

Data augmentation introduces training-time variation without altering the held-out test set. Geometric and photometric perturbations may reduce overfitting, but they should remain plausible for the target task and must not be applied before a split intended to represent unseen data [8]. Here, horizontal flips, rotations limited to $\pm 10^\circ$, and modest brightness–contrast jitter are applied only to training images.

Class imbalance affects both learning and evaluation. Moderate Demented comprises only 64 of 6,400 images. A high overall accuracy can therefore coexist with class-dependent variation in precision and F1-score. Class-weighted loss and macro-averaged reporting provide complementary information for this imbalanced setting [9]. ROC curves can remain visually favorable when prevalence is low; precision–recall curves and class-specific precision are consequently reported together with macro-AUC [18], [19].

C. Rationale for CLAHE Screening

CLAHE was evaluated as a candidate preprocessing operation, not as a presumed source of performance gain. It locally redistributes grayscale contrast while restricting the over-amplification associated with standard adaptive histogram equalization [10], [11]. In structural MRI, local intensity variation may influence anatomical representation, while enhancement may also modify intensity relationships learned by a convolutional model. CLAHE was therefore screened empirically during pipeline selection. The final selected configuration retained the resized grayscale representation without CLAHE.

The comparison documents the practical configuration-selection decision. Because CLAHE use and online augmentation changed jointly, the observed difference is reported as a pipeline-level comparison rather than attributed to either component in isolation.

D. Calibration, Selective Prediction, and Robustness

Classification accuracy and confidence calibration describe complementary properties of predictive performance. Temperature scaling provides a simple post-hoc calibration method in which a single validation-fitted scalar rescales logits without changing the model weights or predicted class ordering [26]. Selective prediction complements calibration by examining the trade-off between retained coverage and prediction risk when low-confidence cases are withheld [27]. Robustness to common image corruptions is also distinct from clean-set discrimination, and controlled perturbation tests can reveal sensitivity to noise, blur, brightness, contrast, and rotation [28]. In this study, these analyses are applied to the unchanged representative seed-42 checkpoint and are interpreted as complementary technical reliability checks within the evaluated dataset and protocol.

III. MATERIALS AND METHODS

A. Dataset and Study Scope

The MRI image collection was obtained from Mendeley Data [20]. It contained 6,400 images in four labelled folders: Non-Demented (3,200), Very Mild Demented (2,240), Mild Demented (896), and Moderate Demented (64). The dataset is a public image-classification collection, and the present study therefore reports its results using the supplied image-level organization. Thus, the experimental unit in this study is an image, and all reported performance values are interpreted at the image level.

No new human participants were recruited, no intervention was performed, and no identifiable clinical records were accessed. The directory terminology is retained exactly from the public collection; in particular, Moderate Demented is treated as a dataset class label rather than as a newly proposed clinical staging system. Table I summarizes the class composition and fixed image-level partition, while Fig. 1 shows the corresponding class-support profile.

B. Split Integrity and Evaluation Workflow

Original image records were assigned to a fixed stratified 70/15/15 split: 4,480 training images, 960 validation images, and 960 test images. The same saved split manifest was reused for every condition and seed. No stochastic transformation was applied to validation or test images. This design prevents a newly generated transformation of an image from entering a different partition during the evaluated training process.

Fig. 2 summarizes the workflow. The held-out test set remains fixed while training is repeated with seeds 42, 52, 62, 72, and 82. Variation across seeds therefore represents training stochasticity under one shared image-level partition and is reported specifically as a measure of training reproducibility.

C. Image Representation and Training Transformations

Each image was converted to grayscale, resized to 128×128 pixels using bilinear interpolation, and replicated across three channels to form a pseudo-RGB input. The selected configuration retained the resized grayscale intensities and applied channel-wise normalization with mean and standard deviation equal to 0.5:

TABLE I. DATASET COMPOSITION AND FIXED IMAGE-LEVEL PARTITION. THE SPLIT WAS CREATED BEFORE ONLINE AUGMENTATION

Class	Total	Training	Validation	Test	Test support (%)
Non-Demented	3,200	2,240	480	480	50.00
Very Mild Demented	2,240	1,568	336	336	35.00
Mild Demented	896	627	134	135	14.06
Moderate Demented	64	45	10	9	0.94
Total	6,400	4,480	960	960	100.00

Class imbalance and its implication for class-wise interpretation

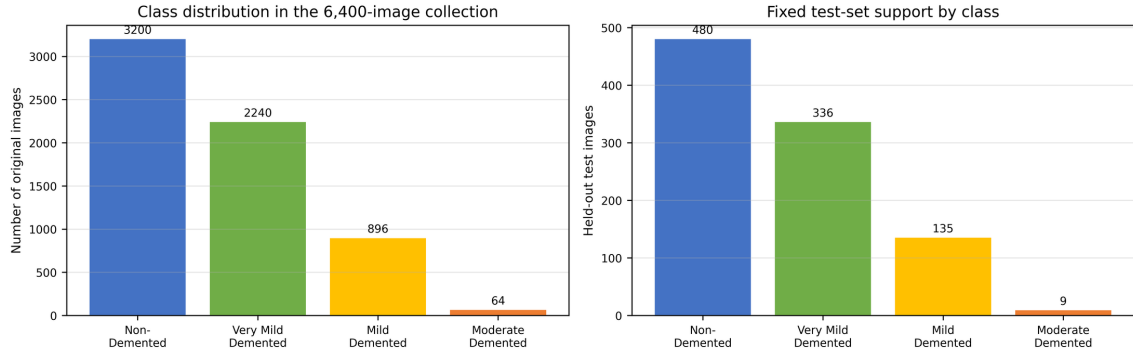


Fig. 1. Class prevalence in the public collection and fixed held-out test support. The Moderate Demented support is shown explicitly to contextualize precision, F1-score, and class-level performance differences.

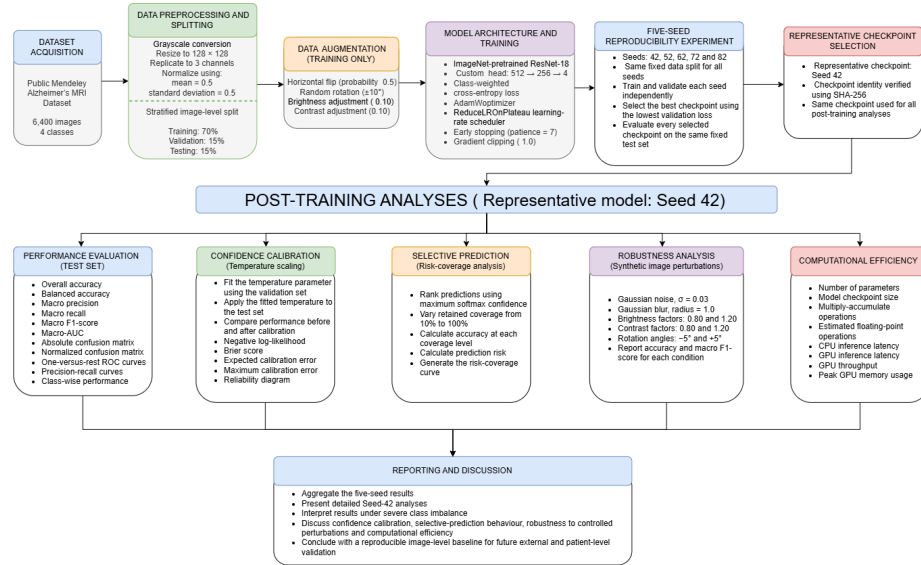


Fig. 2. Reproducible image-level evaluation workflow. Original files are split before training transformations; validation loss selects the checkpoint, and the unchanged held-out test partition is evaluated across five fixed seeds.

$$\tilde{x}_c = \frac{x_c - 0.5}{0.5}, \quad c \in \{1, 2, 3\}. \quad (1)$$

Only training images received random horizontal flipping with probability 0.5, random rotation in $[-10^\circ, 10^\circ]$, and brightness-contrast perturbation with magnitude 0.10. These transformations were re-sampled online and never stored as offline copies. Fig. 3 shows the partition composition, while Fig. 4 illustrates the candidate CLAHE operation.

D. Model Architecture, Optimization, and Selection

The backbone was ImageNet-pretrained ResNet-18 [5]. ResNet-18 was selected because its compact residual architecture provides a practical balance between representational capacity, transfer-learning support, and computational cost for a lightweight image-level framework. Its final layer was replaced by a task-specific head comprising a 512-to-256 fully connected layer, batch normalization [21], ReLU, dropout with rate 0.30 [22], and a 256-to-4 output layer. The model contains 11,309,380 trainable parameters. Table II summarizes

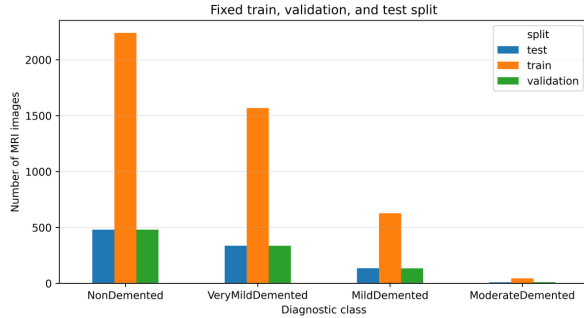


Fig. 3. Fixed train, validation, and test composition by class. The partition was created before online augmentation.

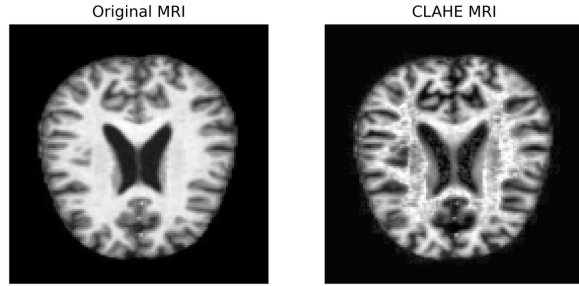


Fig. 4. Illustrative original and CLAHE-transformed grayscale MRI image. CLAHE was screened as a literature-motivated local-contrast candidate and was not retained in the final selected configuration.

the selected ResNet-18 training configuration.

TABLE II. SELECTED RESNET-18 TRAINING CONFIGURATION

Item	Setting
Input	$3 \times 128 \times 128$ pseudo-RGB image
Backbone	ImageNet-pretrained ResNet-18
Classifier head	$512 \rightarrow 256 \rightarrow 4$; batch normalization, ReLU, dropout 0.30
Trainable parameters	11,309,380
Loss	Class-weighted cross-entropy; label smoothing 0.05
Optimizer	AdamW [23]
Initial learning rate	10^{-4}
Weight decay	10^{-4}
Scheduler	ReduceLROnPlateau; factor 0.5, patience 2
Batch size	16
Epoch limit	25
Model selection	Lowest validation-loss checkpoint
Early stopping	Patience 7
Seeds	42, 52, 62, 72, 82

The training loss used class weights inversely proportional to observed training support:

$$w_k = \frac{N}{Kn_k}, \quad (2)$$

where, N is the number of training images, $K = 4$ is the number of classes, and n_k is the training support of class k .

Weighted cross-entropy was combined with label smoothing of 0.05 [24]. AdamW used initial learning rate 10^{-4} and weight decay 10^{-4} [23]. A ReduceLROnPlateau scheduler halved the learning rate after two non-improving validation-loss epochs. Gradient clipping at 1.0 and early stopping with patience 7 were used. The checkpoint with the lowest validation loss was restored before test inference.

E. Evaluation Measures and Statistical Reporting

For each seed, accuracy, balanced accuracy, macro precision, macro recall, macro F1-score, macro one-versus-rest AUC, and per-class precision, sensitivity, specificity, and F1-score were computed on the fixed test partition. In the multi-class setting, balanced accuracy is the average class recall:

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^K \frac{TP_k}{TP_k + FN_k}. \quad (3)$$

The macro F1-score is the unweighted average of class-specific F1-scores:

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K \frac{2P_k R_k}{P_k + R_k}. \quad (4)$$

Metrics are reported as mean $\pm$ standard deviation across five seeds. This summarizes variability due to initialization, data-loader order, and stochastic augmentation under the same split. Accordingly, the five-seed summary is interpreted as training reproducibility under the fixed image-level protocol rather than as a population-level estimate.

F. Configuration-Level Screening

CLAHE was evaluated as a candidate local-contrast operation, whereas online augmentation introduced limited training-time variation. The final no-CLAHE configuration with online augmentation was compared with a CLAHE-based configuration without online augmentation. Both conditions shared the fixed split, ResNet-18 architecture, 0.5/0.5 normalization, class-weighted loss, optimizer, scheduler, batch size, seed set, and checkpoint policy. Since contrast enhancement and augmentation changed jointly, this screening comparison identifies the stronger evaluated pipeline and is interpreted as a configuration-level comparison without assigning the observed difference to either component in isolation.

G. Explainability Analysis Using Grad-CAM

Gradient-weighted class activation mapping (Grad-CAM) [25] was applied to the final convolutional layer of the last ResNet-18 residual block (`layer4[-1].conv2`) to provide qualitative insight into the spatial evidence associated with model predictions. The analysis used the representative seed-42 checkpoint trained with no CLAHE, training-only online augmentation, and the 0.5/0.5 channel normalization defined in Eq. (1). For each class, one correctly classified held-out test image with confidence closest to the class-specific median among correct predictions was selected to reduce reliance on only the most confident examples. Gradients were calculated

with respect to the predicted class, and each activation map was rectified, normalized to the interval $[0, 1]$, resized to the input resolution, and overlaid on the corresponding grayscale MRI image. These maps are qualitative explanations of model attention and are not interpreted as anatomical, pathological, or clinical localization. The visualizations preserve the orientation supplied by the public image collection; no additional anatomical orientation markers or slice-index metadata are introduced.

H. Confidence Calibration and Selective Prediction

The reliability analysis used the representative seed-42 checkpoint with the 0.5/0.5 channel normalization defined in Eq. (1). A single temperature parameter $T > 0$ was fitted by minimizing cross-entropy on the fixed validation partition only [26]. Test probabilities were then calculated as:

$$p(y = k | x) = \frac{\exp(z_k/T)}{\sum_{j=1}^K \exp(z_j/T)}, \quad (5)$$

where, z_k denotes the logit for class k . Calibration was evaluated on the unchanged test partition using negative log-likelihood (NLL), multiclass Brier score, expected calibration error (ECE), and maximum calibration error (MCE). ECE and MCE used 15 equal-width confidence bins. Because scalar temperature scaling is monotonic in the logits, it changes probability sharpness but not the predicted class labels.

Selective prediction was evaluated by ranking test predictions according to maximum softmax confidence and retaining progressively larger coverage fractions from 10% to 100% [27]. Selective risk was defined as one minus accuracy among retained predictions. Macro F1-score was recorded but was not emphasized at very low coverage because some classes may be absent from the retained subset.

I. Robustness Stress-Test Protocol

The unchanged seed-42 checkpoint was evaluated under deterministic synthetic perturbations applied only at test time. The conditions were Gaussian noise with standard deviation 0.03, Gaussian blur with radius 1.0, brightness factors of 0.80 and 1.20, contrast factors of 0.80 and 1.20, and rotations of -5° and $+5^\circ$. After perturbation, images followed the same 128×128 pseudo-RGB conversion and 0.5/0.5 normalization used for the clean test set. Accuracy, balanced accuracy, macro precision, macro recall, macro F1-score, and macro-AUC were computed for every condition. These tests quantify sensitivity to the specified controlled input changes [28]; broader scanner-, site-, and acquisition-related variation is reserved for future multi-centre evaluation.

J. Computational Efficiency Evaluation

Computational demand was measured using trainable parameter count, checkpoint size, multiply-accumulate operations (MACs), estimated floating-point operations (FLOPs), inference latency, throughput, and peak allocated GPU memory. MACs were profiled for one $3 \times 128 \times 128$ input, with FLOPs approximated as twice the MAC count. Latency measurements used repeated warm-up and timed forward passes. CPU measurements used batch sizes 1 and 16; Tesla T4 GPU

measurements used batch sizes 1 and 32. These values are reported as an implementation profile for the stated hardware, batch sizes, and runtime conditions.

IV. RESULTS

A. Aggregate Configuration Comparison

The selected no-CLAHE/augmentation configuration achieved $98.38\% \pm 0.38\%$ accuracy, $98.86\% \pm 0.40\%$ balanced accuracy, $86.94\% \pm 1.79\%$ macro precision, $90.73\% \pm 1.79\%$ macro F1-score, and 0.9993 ± 0.0009 macro-AUC. The CLAHE/no-augmentation configuration achieved $97.19\% \pm 0.32\%$ accuracy, $97.95\% \pm 0.23\%$ balanced accuracy, $82.56\% \pm 0.66\%$ macro precision, $86.04\% \pm 0.76\%$ macro F1-score, and 0.9980 ± 0.0016 macro-AUC. The selected configuration showed a 1.19 percentage points higher accuracy and a 4.68 percentage points higher macro F1-score than the alternate configuration under the same five-seed image-level protocol (Fig. 5).

B. Seed-Level Stability

Table IV lists held-out metrics for all five selected-configuration runs. Accuracy ranged from 98.02% to 98.96%, while macro F1-score ranged from 87.98% to 92.66%. Fig. 6 shows seed-level accuracy and macro F1-score. Seed variation describes training stochasticity under one fixed image-level split and therefore provides a focused measure of training reproducibility.

C. Class-Wise Performance

Table V reports class-wise results for the selected configuration. Non-Demented, Very Mild Demented, and Mild Demented achieved high mean F1-scores. Moderate Demented achieved 100% mean sensitivity across seeds, with $49.89\% \pm 7.64\%$ mean precision and $66.29\% \pm 6.88\%$ mean F1-score. This class-wise pattern shows why the 98.38% overall accuracy should be interpreted together with macro-level and per-class metrics. The nine-image Moderate Demented test support is therefore reported explicitly so that the sensitivity, precision, and F1-score can be interpreted in the context of the available class support.

Fig. 7 complements Table V. The left panel compares per-class F1-scores under the two configurations. The right panel shows the selected configuration's precision-sensitivity relationship. The Moderate Demented class combines high sensitivity with lower precision, making macro-level and class-wise reporting especially informative alongside overall accuracy.

D. Representative Learning and Discrimination Analyses

Fig. 8 presents training and validation loss and macro F1-score for the representative seed-42 selected run. The lowest validation-loss checkpoint was retained for test inference. Fig. 9 presents one-versus-rest ROC and precision-recall curves for the same run. ROC and PR views are both included because class support is unequal; numerical claims remain the five-seed aggregate results in Table III.

TABLE III. PIPELINE-SCREENING COMPARISON ON THE FIXED HELD-OUT TEST SET. VALUES ARE MEAN $\pm$ STANDARD DEVIATION OVER FIVE FIXED TRAINING SEEDS. THE PIPELINES DIFFER JOINTLY IN CLAHE USE AND ONLINE AUGMENTATION

Configuration	Accuracy (%)	Balanced accuracy (%)	Macro precision (%)	Macro recall (%)	Macro F1-score (%)	Macro-AUC
No CLAHE + augmentation	98.38 $\pm$ 0.38	98.86 $\pm$ 0.40	86.94 $\pm$ 1.79	98.86 $\pm$ 0.40	90.73 $\pm$ 1.79	0.9993 $\pm$ 0.0009
CLAHE + no augmentation	97.19 $\pm$ 0.32	97.95 $\pm$ 0.23	82.56 $\pm$ 0.66	97.95 $\pm$ 0.23	86.04 $\pm$ 0.76	0.9980 $\pm$ 0.0016

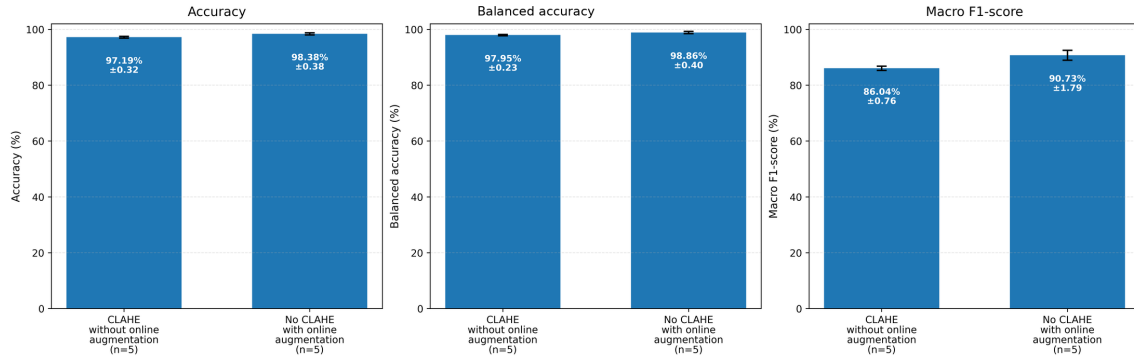


Fig. 5. Pipeline-screening comparison under the same split, model, 0.5/0.5 normalization, and five training seeds. The no-CLAHE configuration with online augmentation was the stronger of the two evaluated pipelines.

TABLE IV. SEED-LEVEL HELD-OUT RESULTS FOR THE SELECTED NO-CLAHE/AUGMENTATION CONFIGURATION. EACH ROW USES THE IDENTICAL FIXED IMAGE-LEVEL SPLIT

Seed	Best epoch	Accuracy (%)	Balanced accuracy (%)	Macro F1-score (%)	Macro-AUC
42	25	98.02	98.23	87.98	0.9994
52	25	98.96	99.32	91.91	0.9998
62	25	98.23	98.94	90.76	0.9996
72	25	98.13	98.77	92.66	0.9976
82	24	98.54	99.03	90.32	0.9999
Mean $\pm$ SD	–	98.38 $\pm$ 0.38	98.86 $\pm$ 0.40	90.73 $\pm$ 1.79	0.9993 $\pm$ 0.0009

Five-seed stability under the shared split and training protocol

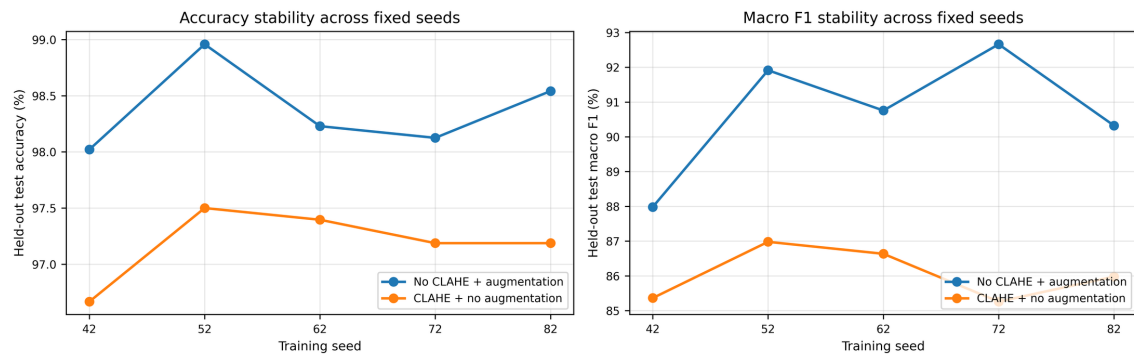


Fig. 6. Accuracy and macro F1-score across five fixed seeds. The values quantify seed-to-seed training variation under one shared image-level partition.

TABLE V. PER-CLASS RESULTS FOR THE SELECTED NO-CLAHE/AUGMENTATION CONFIGURATION. VALUES ARE MEAN $\pm$ STANDARD DEVIATION OVER FIVE SEEDS; SUPPORT IS THE FIXED TEST-SET COUNT

Class	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	Support
Non-Demented	99.46 $\pm$ 0.68	98.58 $\pm$ 0.56	99.46 $\pm$ 0.68	99.02 $\pm$ 0.33	480
Very Mild Demented	99.28 $\pm$ 0.54	97.74 $\pm$ 0.98	99.62 $\pm$ 0.29	98.50 $\pm$ 0.51	336
Mild Demented	99.12 $\pm$ 0.80	99.11 $\pm$ 1.61	99.85 $\pm$ 0.13	99.11 $\pm$ 0.69	135
Moderate Demented	49.89 $\pm$ 7.64	100.00 $\pm$ 0.00	99.01 $\pm$ 0.31	66.29 $\pm$ 6.88	9

Class-wise analysis on the fixed held-out test set

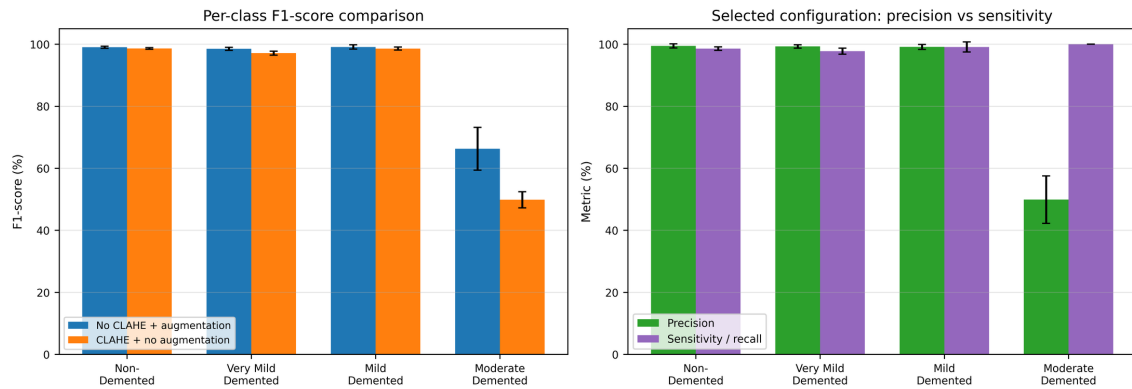


Fig. 7. Class-wise visualization based on five-seed aggregate results. Left: mean F1-score with standard deviation for both configurations. Right: selected configuration precision and sensitivity by class. The 100% Moderate Demented sensitivity is shown together with the nine-image fixed test support to provide the appropriate class-level context.

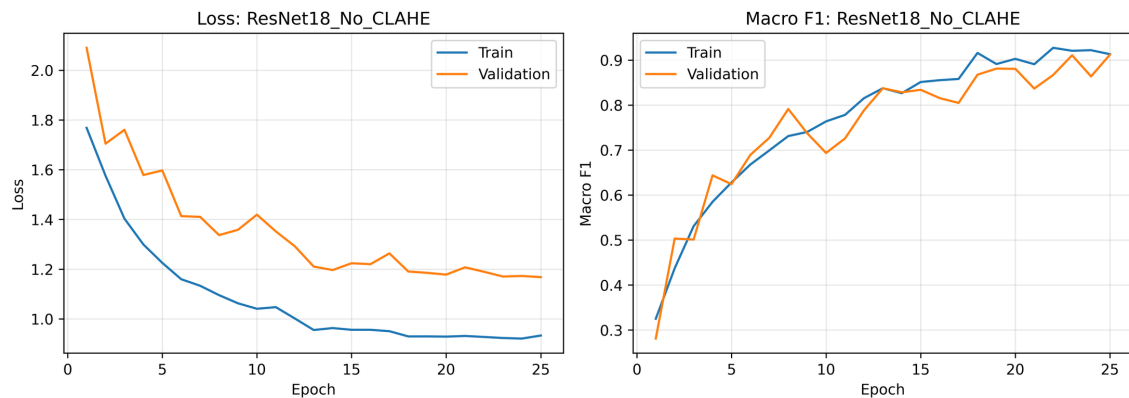


Fig. 8. Training and validation dynamics for the representative seed-42 selected run. The held-out test result was calculated only after restoring the lowest-validation-loss checkpoint.

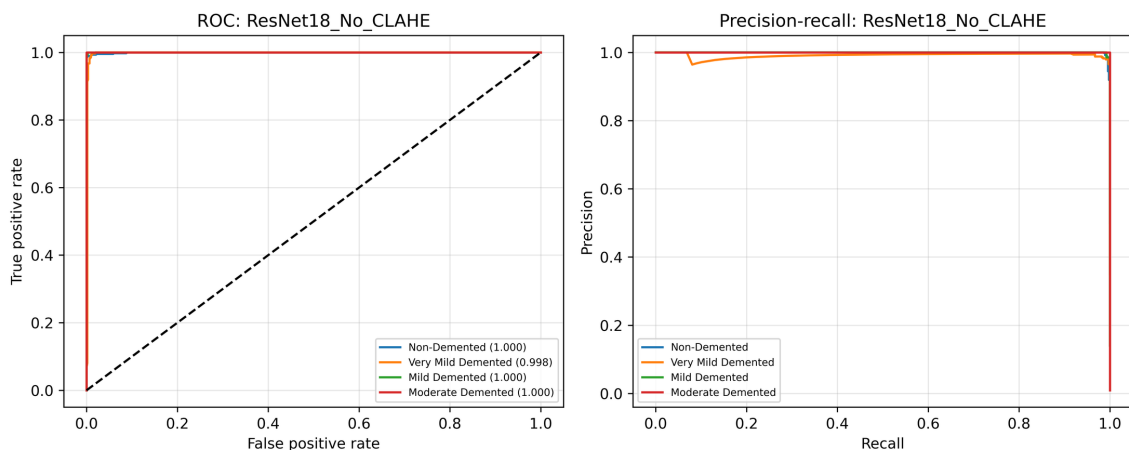


Fig. 9. One-versus-rest ROC and precision-recall curves for the representative seed-42 selected configuration on the held-out test set.

E. Confusion-Matrix Analysis

Fig. 10 provides absolute and row-normalized matrices for the representative seed-42 run. It complements the five-

seed class-wise summary and makes the relationship between recall, false positives, and the nine-image Moderate Demented support directly visible. For this representative run, all 9 Moderate Demented test images were correctly identified,

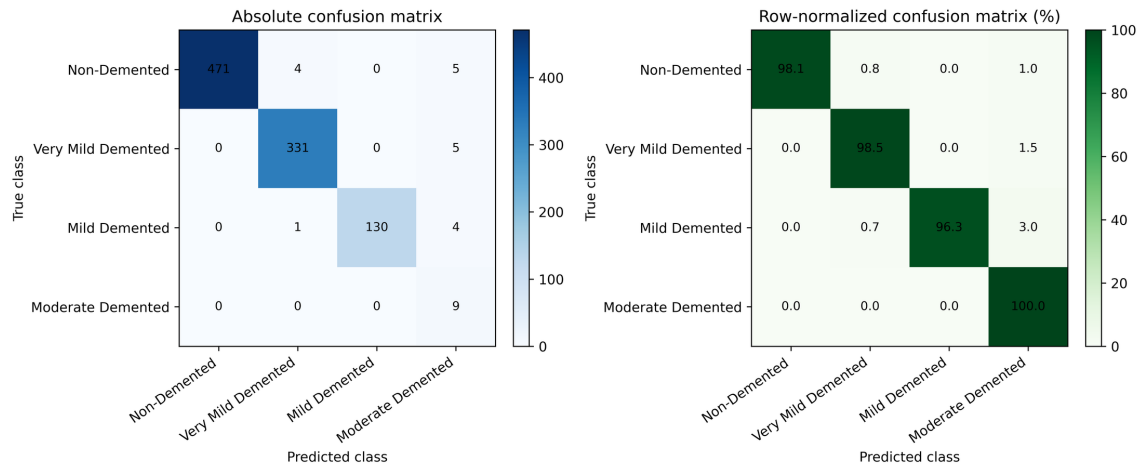


Fig. 10. Absolute and row-normalized confusion matrices for the representative seed-42 selected run. Interpret the figure together with the five-seed per-class summary in Table V.

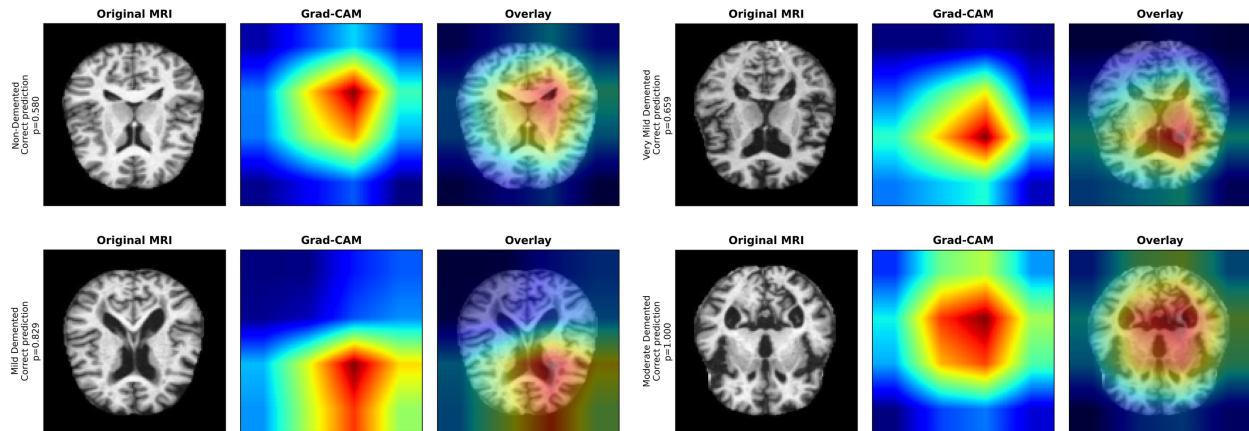


Fig. 11. Grad-CAM visualizations from the representative seed-42 checkpoint for correctly classified held-out test images from the four AD MRI classes. Each class panel shows the original grayscale MRI image, the Grad-CAM map, and the corresponding overlay. The displayed example for each class is the correctly classified image with confidence closest to the median confidence of correct predictions for that class. Warmer map values indicate locations with greater contribution to the predicted class. The maps are presented as qualitative model-inspection evidence, with quantitative anatomical validation reserved for future expert-annotated studies.

TABLE VI. CONFIDENCE CALIBRATION ON THE FIXED TEST PARTITION. TEMPERATURE WAS FITTED USING ONLY THE VALIDATION PARTITION

Output	NLL	Brier	ECE	MCE
Uncalibrated	0.4766	0.2613	0.3472	0.4188
Temperature-scaled	0.1756	0.0449	0.0722	0.5301

TABLE VII. SYNTHETIC ROBUSTNESS EVALUATION USING THE UNCHANGED SEED-42 CHECKPOINT. CHANGE IS RELATIVE TO CLEAN-TEST MACRO F1-SCORE

Condition	Accuracy (%)	Balanced accuracy (%)	Macro F1-score (%)	Δ Macro F1-score (pp)
Clean	98.02	98.23	87.98	0.00
Gaussian noise ($\sigma = 0.03$)	97.19	98.24	86.22	-1.76
Gaussian blur (radius 1.0)	93.65	96.15	79.42	-8.57
Brightness 0.80	98.02	98.65	88.62	+0.63
Brightness 1.20	95.10	95.71	82.24	-5.74
Contrast 0.80	98.02	98.52	89.48	+1.49
Contrast 1.20	96.56	97.68	84.66	-3.32
Rotation -5°	98.33	99.12	90.97	+2.99
Rotation $+5^\circ$	98.33	98.92	90.88	+2.89

TABLE VIII. MEASURED COMPUTATIONAL EFFICIENCY PROFILE

Property	Value
Trainable parameters	11,309,380
Checkpoint size	43.22 MB
GMACs per image	0.596
Estimated GFLOPs per image	1.191
CPU latency, batch 1	76.70 ms/image
GPU latency, batch 1	3.08 ms/image
GPU throughput, batch 32	3341 images/s
Peak GPU memory	126.88 MB

while 14 images from the other three classes were assigned to the Moderate Demented label, giving 9 true positives, 0 false negatives, and 14 false positives for that class.

F. Grad-CAM Explainability Results

Fig. 11 presents Grad-CAM visualizations for representative correctly classified images from the four classes. The seed-42 checkpoint used for this analysis obtained 98.02% accuracy, 98.23% balanced accuracy, and 87.98% macro F1-score on the fixed held-out test partition, consistent with the corresponding row in Table IV. The maps show the spatial distribution of image regions contributing to each predicted class. They are presented as qualitative model-inspection evidence; no anatomical localization claim is made because no expert-defined region-of-interest reference is used in this analysis.

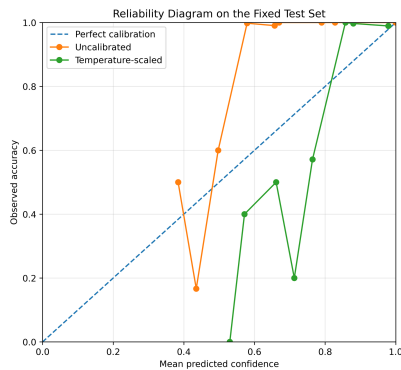


Fig. 12. Reliability diagram for the representative seed-42 checkpoint before and after validation-fitted temperature scaling.

G. Confidence Calibration and Selective Prediction

The validation-fitted temperature was $T = 0.2135$. As shown in Table VI, temperature scaling reduced NLL from 0.4766 to 0.1756, Brier score from 0.2613 to 0.0449, and ECE from 0.3472 to 0.0722. Accuracy and macro F1-score remained 98.02% and 87.98%, respectively, because scalar temperature scaling did not change class predictions. MCE increased from 0.4188 to 0.5301, indicating that a sparsely populated confidence bin retained a larger local calibration gap even though NLL, Brier score, and ECE improved. Fig. 12 therefore shows that the aggregate and maximum-bin calibration measures provide complementary information.

The confidence-ranked risk-coverage analysis is shown in Fig. 13. At 95% coverage, retained accuracy was 99.56% compared with 98.02% at full coverage. The temperature-scaled and uncalibrated curves were nearly identical because scalar temperature scaling preserves the score ordering apart from numerical ties. This experiment characterizes the confidence-coverage trade-off within the fixed test partition and provides a complementary view of confidence-ranked prediction behaviour.

H. Robustness Under Controlled Image Perturbations

Table VII and Fig. 14 summarize the synthetic stress tests. Gaussian noise produced a 1.76-percentage-point reduction in macro F1-score, whereas Gaussian blur produced the largest degradation, reducing macro F1-score from 87.98% to 79.42%.

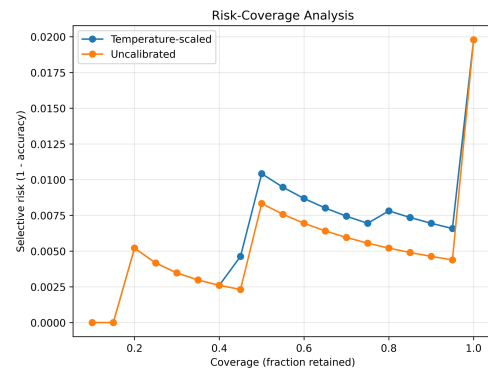


Fig. 13. Selective-prediction risk for the representative seed-42 checkpoint as a function of retained coverage.

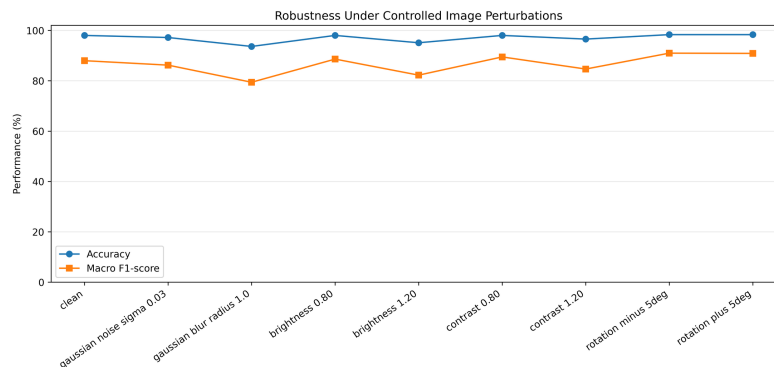


Fig. 14. Accuracy and macro F1-score for the representative seed-42 checkpoint under the specified synthetic test-time perturbations.

Increased brightness and contrast also reduced macro F1-score by 5.74 and 3.32 percentage points, respectively. Mild brightness reduction, contrast reduction, and $\pm 5^\circ$ rotations produced small apparent improvements on the fixed test partition. Because the Moderate Demented test support contains nine images, these increases are reported as fixed-sample variation and are not interpreted as a beneficial effect of perturbation.

I. Computational Efficiency Profile

The model contained 11,309,380 trainable parameters, occupied 43.22 MB as a checkpoint, and required 0.596 GMACs (approximately 1.191 GFLOPs) per 128×128 image. On the measured environment, CPU batch-1 inference averaged 76.70 ms per image. Tesla T4 batch-1 inference averaged 3.08 ms per image, while batch-32 throughput reached 3341 images/s with 126.88 MB peak allocated GPU memory. Table VIII provides the complete profile. These measurements support the resource-efficient characterization of the framework for the stated hardware, batch sizes, and runtime conditions.

V. DISCUSSION

A. Main Finding

Under the evaluated protocol, the no-CLAHE/augmentation configuration provided the stronger aggregate performance of the two screened pipelines. It obtained a macro F1-score above 90% while remaining compact relative to many 3D and attention-heavy neuroimaging systems..

The comparison supports empirical evaluation of preprocessing choices within the target collection. CLAHE was included as a local-contrast candidate based on established adaptive-histogram principles [10], [11]. Under the evaluated protocol, the CLAHE-based pipeline without online augmentation produced lower aggregate accuracy, macro precision, and macro F1-score. Because the pipelines differ in both contrast enhancement and augmentation, the difference may reflect preservation of intensity structure, online augmentation, their interaction, or another configuration-level effect. A complete 2×2 factorial design remains a useful extension for estimating independent main effects and interactions.

B. Interpretation Under Imbalance

The difference between the high overall accuracy and the macro-level metrics is mainly associated with the Moderate Demented class. The selected configuration achieved 100% mean sensitivity for that class but lower precision, implying that some images from other classes were assigned to the rare label. This is why aggregate accuracy must not be used alone in imbalanced settings [9], [18]. Macro F1-score, macro precision, per-class support, and confusion patterns provide a more complete assessment.

The Moderate Demented class contains nine test images. Consequently, individual prediction changes have a visible effect on sensitivity and precision; for example, one additional false negative would change sensitivity by more than 11 percentage points. Reporting support together with precision, sensitivity, and F1-score therefore provides the most informative interpretation of this class.

C. Qualitative Explainability

The Grad-CAM maps provide a visual check that the representative seed-42 classifier drew evidence from spatially distributed regions within the MRI field rather than exclusively from image borders or uniform background. This observation is useful for model inspection. The heatmaps are intentionally presented as qualitative visual evidence because the analysis does not use expert segmentation, an anatomical atlas, or pathology-grounded region annotations. Consequently, Grad-CAM complements the predictive metrics without making an anatomical localization claim.

D. Reliability, Robustness, and Resource Efficiency

The calibration results show that high discrimination and reliable confidence are separate properties. For the seed-42 checkpoint, validation-fitted temperature scaling improved NLL, Brier score, and ECE without altering the predicted labels. MCE moved in the opposite direction, showing that aggregate calibration measures and the largest individual-bin deviation capture different aspects of probability behaviour. Selective prediction further showed that higher-confidence subsets had lower error; the full multiclass metrics remain the primary performance summary.

The controlled perturbation results provide a differentiated sensitivity profile. The framework remained comparatively stable under mild Gaussian noise and small rotations, while blur and intensity increases produced larger metric changes. These results complement clean-test performance, while scanner differences, reconstruction pipelines, motion effects, and institutional variation remain important directions for future multi-centre evaluation.

The measured 0.596 GMACs per image, 43.22 MB checkpoint, and low Tesla T4 inference latency provide direct evidence for resource efficiency. Future translational evaluation can extend this evidence through external validation, use-case-specific calibrated thresholds, software-system testing, and prospective workflow assessment.

E. Study Scope and Future Work

The scope of the present study is a reproducible image-level evaluation on one public MRI collection. The available data support image-level rather than verified subject-level partitioning, and the reported five-seed standard deviations therefore summarize training reproducibility under one fixed split. The Moderate Demented class is reported with its nine-image test support so that class-specific metrics are interpreted with their sample context. The framework uses 2D images, while volumetric and multi-slice representations offer a natural extension when complete acquisition context is available. Grad-CAM is used for qualitative model inspection, calibration is assessed within the supplied dataset partitions, and the perturbation study is restricted to the specified synthetic transformations. Computational measurements are reported for the stated hardware and runtime configuration.

Future work can extend the present framework through datasets with subject-level and site-level identifiers, external multi-centre evaluation, volumetric or multi-slice representations with preserved acquisition context. Additional directions

include calibration assessment under dataset shift, use-case-specific selective-decision thresholds, quantitative explainability assessment against expert-defined anatomical references, and a complete 2×2 factorial study of CLAHE and augmentation.

VI. CONCLUSION

This study developed a reliability-oriented and resource-efficient evaluation of a lightweight ResNet-18 framework for four-class AD MRI image classification. Under a fixed image-level split and five training seeds, the selected no-CLAHE configuration with online augmentation achieved $98.38\% \pm 0.38\%$ accuracy, $98.86\% \pm 0.40\%$ balanced accuracy, $90.73\% \pm 1.79\%$ macro F1-score, and 0.9993 ± 0.0009 macro-AUC. For the representative seed-42 checkpoint, temperature scaling improved NLL, Brier score, and ECE, while selective prediction characterized the confidence-coverage trade-off. Controlled stress tests showed larger performance changes under Gaussian blur and intensity increases than under mild noise or small rotations. The model required 0.596 GMACs per image and achieved 3.08 ms batch-1 inference on a Tesla T4, providing measured evidence of resource efficiency. Grad-CAM provided qualitative model-inspection evidence without making an anatomical localization claim. Collectively, the findings support a reproducible, confidence-aware, perturbation-tested, and computationally profiled image-level framework within the evaluated public dataset and protocol.

REFERENCES

- [1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, *Med. Image Anal.* 42 (2017) 60–88. doi:10.1016/j.media.2017.07.005.
- [2] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, *Annu. Rev. Biomed. Eng.* 19 (2017) 221–248. doi:10.1146/annurev-bioeng-071516-044442.
- [3] J. Wen, E. Thibaut-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation, *Med. Image Anal.* 63 (2020) 101694. doi:10.1016/j.media.2020.101694.
- [4] M. A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review, *Comput. Methods Programs Biomed.* 187 (2020) 105242. doi:10.1016/j.cmpb.2019.105242.
- [5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
- [6] S. J. Pan, Q. Yang, A survey on transfer learning, *IEEE Trans. Knowl. Data Eng.* 22 (10) (2010) 1345–1359. doi:10.1109/TKDE.2009.191.
- [7] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning?, *IEEE Trans. Med. Imaging* 35 (5) (2016) 1299–1312. doi:10.1109/TMI.2016.2535302.
- [8] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data* 6 (2019) 60. doi:10.1186/s40537-019-0197-0.
- [9] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, *Neural Netw.* 106 (2018) 249–259. doi:10.1016/j.neunet.2018.07.011.
- [10] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. M. ter Haar Romeny, J. B. Zimmerman, K. J. Zuiderveld, Adaptive histogram equalization and its variations, *Comput. Vision Graph. Image Process.* 39 (3) (1987) 355–368. doi:10.1016/S0734-189X(87)80186-X.
- [11] K. J. Zuiderveld, Contrast limited adaptive histogram equalization, in: P. S. Heckbert (Ed.), *Graphics Gems IV*, Academic Press Professional, San Diego, CA, USA, 1994, pp. 474–485. doi:10.1016/B978-0-12-336156-1.50061-6.
- [12] J. S. Kim, J. W. Han, J. B. Bae, et al., Deep learning-based diagnosis of Alzheimer's disease using brain magnetic resonance images: An empirical study, *Sci. Rep.* 12 (2022) 18007. doi:10.1038/s41598-022-22917-3.
- [13] L. Chen, H. Qiao, F. Zhu, Alzheimer's disease diagnosis with brain structural MRI using multiview-slice attention and 3D convolution neural network, *Front. Aging Neurosci.* 14 (2022) 871706. doi:10.3389/fnagi.2022.871706.
- [14] F. Hajamohideen, et al., Four-way classification of Alzheimer's disease using deep Siamese convolutional neural network with triplet-loss function, *Brain Inform.* 10 (2023) 5. doi:10.1186/s40708-023-00184-w.
- [15] P. Carcagni, M. Leo, M. Del Coco, C. Distanto, A. De Salve, Convolution neural networks and self-attention learners for Alzheimer dementia diagnosis from brain MRI, *Sensors* 23 (3) (2023) 1694. doi:10.3390/s23031694.
- [16] J. Xin, A. Wang, R. Guo, W. Liu, X. Tang, CNN and Swin-Transformer based efficient model for Alzheimer's disease diagnosis with sMRI, *Biomed. Signal Process. Control* 86 (2023) 105189. doi:10.1016/j.bspc.2023.105189.
- [17] B. S. Rao, M. Aparna, S. S. Kolisetty, H. Janapana, Y. V. Koteswararao, Multi-class classification of Alzheimer's disease using deep learning and transfer learning on 3D MRI images, *Traitement du Signal* 41 (3) (2024) 1397–1404. doi:10.18280/ts.410328.
- [18] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLOS ONE* 10 (3) (2015) e0118432. doi:10.1371/journal.pone.0118432.
- [19] T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.* 27 (8) (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.
- [20] A. Yakkundi, Alzheimer's disease dataset, Mendeley Data (2023). doi:10.17632/ch87yswbz4.1.
- [21] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.
- [22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* 15 (2014) 1929–1958.
- [23] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: *International Conference on Learning Representations*, 2019.
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, D. Wojna, Rethinking the Inception architecture for computer vision, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308.
- [25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.
- [26] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70, 2017, pp. 1321–1330.
- [27] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, in: *Advances in Neural Information Processing Systems* 30, 2017.
- [28] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in: *International Conference on Learning Representations*, 2019.