Sensitivity-Prioritized Comparative Evaluation of VGG16 and ResNet50 Deep Learning Architectures for Early-Stage Alzheimer's Disease Detection from Brain MRI

Shankar Ghimire, Softwarica College of IT and E-Commerce, Coventry University, Kathmandu, Nepal

Bikash Thapa Magar, Manipal Academy of Higher Education, India

Nikesh Adhikari, Softwarica College of IT and E-Commerce, Coventry University, Kathmandu, Nepal

Sagar Neupane, Softwarica College of IT and E-Commerce, Coventry University, Kathmandu, Nepal

Sachin Maharjan, Softwarica College of IT and E-Commerce, Coventry University, Kathmandu, Nepal

Abstract

Background: Alzheimer's Disease (AD) is a progressive neurodegenerative disorder affecting approximately 55 million individuals globally, with projections exceeding 139 million cases by 2050. Clinical diagnosis currently depends on manual radiologist interpretation of Structural Magnetic Resonance Imaging (MRI), a workflow subject to substantial inter-observer variability particularly at the Very Mild Demented stage, where delayed detection directly translates into missed therapeutic windows.

Objective: This study addresses a critical methodological gap in the deep learning literature: prior comparative studies of CNN architectures for AD detection have overwhelmingly optimized for aggregate classification accuracy rather than Sensitivity (Recall), the metric of primary clinical concern in medical screening. We present a rigorously controlled Champion vs. Challenger comparative evaluation of two widely adopted CNN architectures VGG16 (sequential) and ResNet50 (residual) with explicit priority on minimizing false negatives, particularly in the clinically decisive Very Mild Demented stage.

Methods: Both architectures were trained on the publicly available Augmented Alzheimer MRI Dataset (33,984 images, four severity classes) using a stratified 80/20 train–validation split. A two-stage Transfer Learning protocol ImageNet-based feature extraction (5 epochs at learning rate 1.26×10^{-3}) followed by full-network fine-tuning (5 epochs at learning rate 1×10^{-5}) was applied uniformly to both models. The Adam optimizer and categorical cross-entropy loss were used throughout. Performance was evaluated using accuracy, macro-averaged precision, macro-averaged recall (primary criterion), macro-averaged F1-score, and class-level confusion matrices.

Results: Contrary to theoretical expectations favoring deeper residual architectures, VGG16 outperformed ResNet50 on every evaluated metric. VGG16 achieved a final validation accuracy of 97.76%, macro-average recall of 97.92%, and validation loss of 0.0653, compared to ResNet50's 94.25%, 94.68%, and 0.1694 respectively. Critically, in the Very Mild Demented class, VGG16 attained 94.05% recall versus ResNet50's 88.64% a 5.41-percentage-point improvement that reduced false negatives from 108 to 74 patients per validation cohort of 1,814 early-stage cases.

Conclusions: These findings challenge the assumption that deeper residual architectures universally outperform shallower sequential networks in constrained medical imaging domains. In applications where training data volume is moderate and target features are spatially coherent (as with cortical atrophy), VGG16's hierarchical sequential learning, combined with careful fine-tuning, produces superior clinical sensitivity. We recommend VGG16 as a foundational architecture for Alzheimer's clinical decision support systems and identify explainability (Grad-CAM), 3D volumetric modelling, and multi-site prospective validation as the highest-priority directions for subsequent research.

Index Terms*: Alzheimer's Disease, Convolutional Neural Networks, VGG16, ResNet50, Transfer Learning, MRI Classification, Medical Image Analysis, Dementia Staging, Sensitivity, False Negative Reduction, Computer-Aided Diagnosis, Explainable AI.*

I. INTRODUCTION

A. Clinical Context

Alzheimer's Disease (AD) is the most prevalent form of dementia worldwide, affecting an estimated 55 million individuals globally a figure projected to reach 78 million by 2030 and 139 million by 2050 according to the World Health Organization [1]. Characterized by progressive and irreversible deterioration of memory, language, executive function, and activities of daily living, AD imposes severe clinical, social, and economic consequences, with annual global care costs exceeding USD 1.3 trillion. In the absence of a curative therapy, early and accurate detection is clinically imperative: patients diagnosed at early stages can access disease-modifying agents such as lecanemab and donanemab (recently approved for early AD), structured cognitive rehabilitation, and advance care planning all of which are substantially more effective when initiated early in the disease trajectory.

The current clinical gold standard for structural AD diagnosis relies on T1-weighted Structural Magnetic Resonance Imaging (sMRI), which enables non-invasive visualization of hallmark neurodegeneration markers including hippocampal atrophy, medial temporal lobe thinning, cortical thinning, and ventricular enlargement. Despite its diagnostic utility, manual radiologist interpretation of sMRI carries well-documented limitations. Frisoni et al. [2] systematically reviewed the clinical use of structural MRI and reported substantial inter-rater variability, especially at the Very Mild Demented (VMD) stage where subtle pathological changes are easily confounded with age-related anatomical variation. A missed or delayed diagnosis at this critical juncture forecloses access to timely intervention and accelerates functional decline.

B. Deep Learning as a Diagnostic Accelerator

The convergence of large annotated medical imaging datasets and advances in deep learning has opened a transformative avenue for automated, objective, and scalable AD detection. Convolutional Neural Networks (CNNs), inspired by the hierarchical processing architecture of the mammalian visual cortex [3], have demonstrated exceptional capacity for extracting multi-scale discriminative features from raw pixel data often surpassing hand-engineered feature approaches in both accuracy and generalizability [6]. Transfer Learning, wherein a model pre-trained on a large general-purpose dataset (typically ImageNet [12]) is adapted to a domain-specific task with limited training data, has democratized access to high-performance deep learning for medical imaging tasks [11].

C. The Sensitivity Gap in Existing Literature

Despite the proliferation of CNN-based AD detection studies, a persistent and clinically consequential gap remains: the dominant evaluation metric across published literature is overall classification accuracy. This metric is fundamentally misleading in the context of class-imbalanced medical datasets and fails to account for the asymmetric clinical cost of false negatives versus false

positives. A model that misclassifies a Very Mild Demented patient as healthy a False Negative directly harms patient welfare by delaying intervention. Conversely, a False Positive prompts additional diagnostic workup, which is an inconvenience but not a medical failure.

Yet comparative architectural studies with Sensitivity (Recall) as the primary selection criterion remain rare. This study directly addresses this gap.

D. Research Contributions

The principal contributions of this work are as follows:

- **Sensitivity-prioritized evaluation framework:** We formulate model selection around macro-averaged recall and class-level false-negative rates, aligning the evaluation metric with clinical screening priorities rather than aggregate accuracy.
- **Rigorous Champion vs. Challenger comparison:** VGG16 (sequential) and ResNet50 (residual) are evaluated under identical preprocessing, training, and evaluation protocols on a dataset of 33,984 MRI images across four AD severity classes.
- **Counter-intuitive architectural finding:** VGG16 outperforms ResNet50 on every metric, most notably in the clinically critical Very Mild Demented class (94.05% vs. 88.64% recall), challenging the assumption that deeper residual architectures universally dominate in medical imaging.
- **Quantitative clinical impact analysis:** Confusion-matrix-based analysis translates recall improvements into concrete patient-level outcomes (34 fewer missed early-stage diagnoses per 1,814 validation cases).
- **Comprehensive benchmark against published literature:** Performance is contextualized against 12 recent state-of-the-art AD classification studies (2017–2025).

E. Paper Organization

Section II reviews the state of the art in deep-learning-based AD detection. Section III details the dataset, preprocessing pipeline, model architectures, and training protocol. Section IV presents quantitative results, confusion-matrix analysis, and benchmark comparisons. Section V discusses the architectural findings, clinical implications, training dynamics, and limitations. Section VI concludes with recommended directions for future research.

II. RELATED WORK AND LITERATURE REVIEW

A. Classical Machine Learning Approaches to AD Diagnosis

Early computational approaches to AD diagnosis were grounded in conventional machine learning methods applied to hand-crafted, domain-specific features. Landmark studies utilized Support Vector Machines (SVMs), Random Forests, and logistic regression classifiers operating on region-of-interest features derived from brain segmentation pipelines including hippocampal volume, cortical thickness measures from FreeSurfer, and grey-matter density maps from voxel-based morphometry (VBM). Cuingnet et al. [4] conducted a comprehensive benchmark of ten SVM-based classifiers for AD and Mild Cognitive Impairment (MCI) classification from structural MRI using

the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, reporting considerable variance in classification performance (57–81% accuracy) attributable primarily to ROI selection methodology. Zhang et al. [5] combined multi-modality features from MRI, PET, and CSF biomarkers using a kernel-based fusion framework, achieving marginal improvements over single-modality classifiers but at substantial data acquisition costs.

Despite methodological rigor, these traditional ML approaches suffer from fundamental limitations: they are critically dependent on the quality of hand-engineered features, require substantial domain expertise in neuroanatomy, and generalize poorly across scanner types, acquisition protocols, and demographic cohorts.

B. Convolutional Neural Networks in Medical Image Analysis

The deep learning revolution, catalysed by the breakthrough performance of AlexNet at the 2012 ImageNet competition, propagated rapidly into medical imaging research. CNNs offered an end-to-end learning paradigm that eliminated the need for manual feature engineering, instead learning hierarchical representations directly from raw pixel data. LeCun, Bengio, and Hinton [3] demonstrated the general capacity of deep networks to learn transferable, multi-level abstractions that exceed human-engineered features on complex visual recognition tasks. Litjens et al. [6] provided a foundational survey of deep learning applications in medical image analysis, cataloguing CNN deployment across radiology, pathology, ophthalmology, and cardiology identifying brain MRI as a particularly active and promising application domain.

Sarraf and Tofighi [7] were among the first to apply CNNs to Alzheimer's classification using functional MRI (fMRI) data, demonstrating promising binary discrimination between AD patients and healthy controls. Subsequent work has extended CNN-based approaches to structural MRI and multi-class disease staging.

C. Sequential and Residual Architectures: VGG16 and ResNet50

VGG16, introduced by Simonyan and Zisserman of the Oxford Visual Geometry Group [8], demonstrated that network depth achieved through stacks of small 3×3 convolutional filters is a critical determinant of image recognition accuracy. Its homogeneous, purely sequential block structure makes it conceptually transparent and highly amenable to transfer learning in medical imaging. However, its large fully-connected layers introduce computational overhead, and its depth renders it susceptible to the vanishing gradient problem characterized by Glorot and Bengio [9].

ResNet50, introduced by He et al. [10], addressed the fundamental optimization challenges of very deep networks through residual skip connections that enable gradient flow to bypass convolutional blocks. This architectural innovation eliminated the vanishing gradient problem and made training networks of 50 or more layers tractable. Theoretically, ResNet50's greater depth should enable capture of higher-order, more abstract pathological representations. In practice, however, empirical studies in medical imaging have repeatedly shown that theoretical depth advantages do not always translate to superior task performance, particularly when training data is limited and domain-transfer requirements are specialized [21, 22].

D. Transfer Learning in Alzheimer's Neuroimaging

Transfer Learning has emerged as the dominant paradigm for adapting general-purpose CNNs to medical imaging. Pan and Yang [11] provided the seminal theoretical framework, categorizing instance-based, feature-based, and model-based transfer approaches. Russakovsky et al. [12] validated ImageNet pre-training as an effective foundation, despite the fundamental domain gap between photographic and MRI textures.

Hon and Khan [13] were among the first to apply transfer learning with VGG16 and Inception V4 specifically to Alzheimer's classification from MRI, obtaining competitive accuracy (~92.3%) with limited training data. Islam and Zhang [14] proposed an ensemble of deep CNNs incorporating ResNet variants for multi-class AD staging from structural MRI, reporting 93.18% accuracy. Ebrahimi-Ghahnavieh et al. [15] applied deep ResNet variants on the ADNI dataset, obtaining strong binary classification but limited analysis of class-level recall.

E. Recent Advances and Current State of the Art (2022–2025)

Research in CNN-based AD classification has accelerated substantially in recent years. Sharma et al. [21] reported 91.78% accuracy using ResNet-50 with transfer learning on Kaggle MRI datasets. Soliman et al. [22] combined DenseNet196, VGG16, and ResNet-50 in an ensemble approach, achieving 89% accuracy on Kaggle MRI brain data. Mahmud et al. [23] evaluated a comparative pipeline (AD-VGG16, AD-ResNet50, and a 2D-CNN), reporting 93.45%, 79.43%, and 98.28% accuracy respectively on the Kaggle MRI preprocessed dataset a notable finding that a purpose-built 2D-CNN can outperform pre-trained deep architectures when data and training are appropriately matched.

Deepa and Chokkalingam [24] optimized VGG16 using the Arithmetic Optimization Algorithm for early AD detection, reporting improved performance. Jain et al. [25] classified AD stages using a hybrid VGG16+LSTM approach on ADNI, reporting 98.8% accuracy and 100% sensitivity but only 76% specificity illustrating the precision-recall trade-off central to the present study. Hussain et al. [26] developed a Siamese CNN derived from VGG-16 for dementia stage classification on OASIS data, achieving 99.05% accuracy. Most recently, Nature Scientific Reports published a comparative evaluation of VGG16, VGG19, InceptionResNetV2, and Xception on 6,735 MRI images [27], reporting that VGG16 attained 100% precision, recall, and F-score specifically for moderate dementia corroborating the strong performance of sequential architectures observed in the present study.

F. Explainable AI for Clinical Deployment

As deep learning models approach human-level accuracy on AD classification, the research community has increasingly recognized that clinical deployment requires not only high accuracy but also interpretability. Selvaraju et al. [28] introduced Gradient-weighted Class Activation Mapping (Grad-CAM), enabling visualization of the image regions most influential to a CNN's prediction. Tang et al. [29] applied Grad-CAM to a 3D-VGG16 architecture for AD classification, showing that model attention correctly localizes to the hippocampus and precuneus regions with established AD pathological significance. Shojaei et al. [30] reviewed XAI techniques across the AD diagnostic pipeline, concluding that Grad-CAM, LIME, and LRP are the most widely deployed interpretability methods for CNN-based neuroimaging models.

This literature highlights interpretability as the key bridge between raw model accuracy and clinical adoption, and motivates the explainability pipeline proposed in Section V of this work.

G. Research Gap and Positioning of the Present Work

A synthesis of the reviewed literature reveals three persistent gaps. First, the dominant evaluation metric remains aggregate accuracy rather than clinically-weighted sensitivity. Second, systematic head-to-head comparisons of sequential (VGG) versus residual (ResNet) paradigms with explicit recall prioritization are lacking. Third, few studies explicitly contextualize results in terms of concrete patient-level outcomes (e.g., false negatives per 1,000 screened). The present work addresses all three gaps through a Champion vs. Challenger experimental framework in which macro-averaged recall is the primary selection criterion, class-wise false-negative rates are directly analysed, and findings are translated into patient-level clinical impact.

III. METHODOLOGY

This research employs a quantitative experimental design following a Champion vs. Challenger paradigm. VGG16 serves as the established baseline (Champion), while ResNet50 represents the theoretically superior residual challenger. Both models undergo identical data preprocessing, training protocols, and evaluation procedures, ensuring a controlled and reproducible comparison. The three-stage training pipeline is illustrated in Fig. 1.

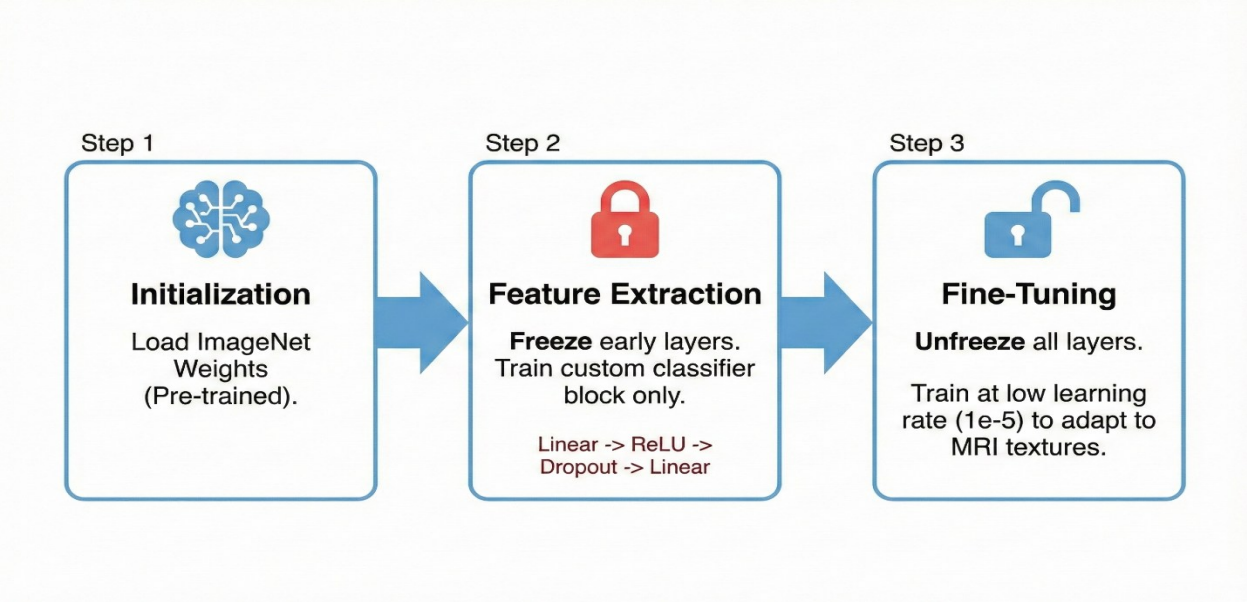

Fig. 1. Three-stage Transfer Learning pipeline: (1) Initialization with ImageNet pre-trained weights; (2) Feature Extraction with the convolutional base frozen and only the custom classifier head trained; (3) Full-network Fine-Tuning at a reduced learning rate to adapt convolutional filters to MRI-specific textural biomarkers.

A. Dataset Acquisition and Description

The study utilized the publicly available Augmented Alzheimer MRI Dataset [16], consisting of 33,984 pre-processed and augmented T1-weighted MRI brain scan images distributed across four

clinically recognized AD severity classes: Non-Demented (ND), Very Mild Demented (VMD), Mild Demented (MiD), and Moderate Demented (MoD). The dataset was partitioned using stratified random sampling into a training set (80%, n = 27,187) and a validation set (20%, n = 6,797). Stratification preserves the relative class distribution across both partitions, mitigating class-imbalanced evaluation artefacts. Table I summarizes the distribution.

TABLE I: DATASET DISTRIBUTION BY CLASS

Class	Training	Validation	Total	Proportion
Non-Demented	15,892	3,973	19,865	58.5%
Very Mild Demented	5,760	1,440	7,200	21.2%
Mild Demented	4,166	1,041	5,207	15.3%
Moderate Demented	1,369	343	1,712	5.0%
Total	27,187	6,797	33,984	100%

B. Data Preprocessing Pipeline

A standardized preprocessing pipeline was applied uniformly across both architectures to ensure consistency:

- **Resizing:** All images were resized to 224×224 pixels using bilinear interpolation to match the input specifications of ImageNet pre-trained architectures.
- **Normalization:** Pixel intensities were normalized using ImageNet channel-wise statistics (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]) to align input distributions with those expected by pre-trained convolutional filters.
- **Data Augmentation:** The dataset incorporates augmentation via rotations (±10°), horizontal flipping, zoom (±10%), and brightness perturbations parameters consistent with best practice in medical image augmentation [19].
- **Corrupted Image Handling:** A custom safe_loader function was implemented to detect and gracefully bypass corrupted or unreadable files during training, preventing pipeline interruption.

C. Model Architectures

Two architecturally distinct CNN models were selected for comparative evaluation:

1) VGG16 (Champion / Baseline):

VGG16 is a 16-layer sequential network composed of 13 convolutional layers (employing 3×3 filters with stride 1 and padding 1) organized into five sequential blocks, each followed by a max-pooling layer, culminating in three fully-connected layers. For this study, VGG16 was initialized with ImageNet pre-trained weights (from [12]). Its original ImageNet classification head was replaced with a custom sequential block comprising: Linear (25088 → 512) → ReLU → Dropout (p = 0.3) [18] → Linear (512 → 4). The VGG16 architecture is illustrated in Fig. 2.

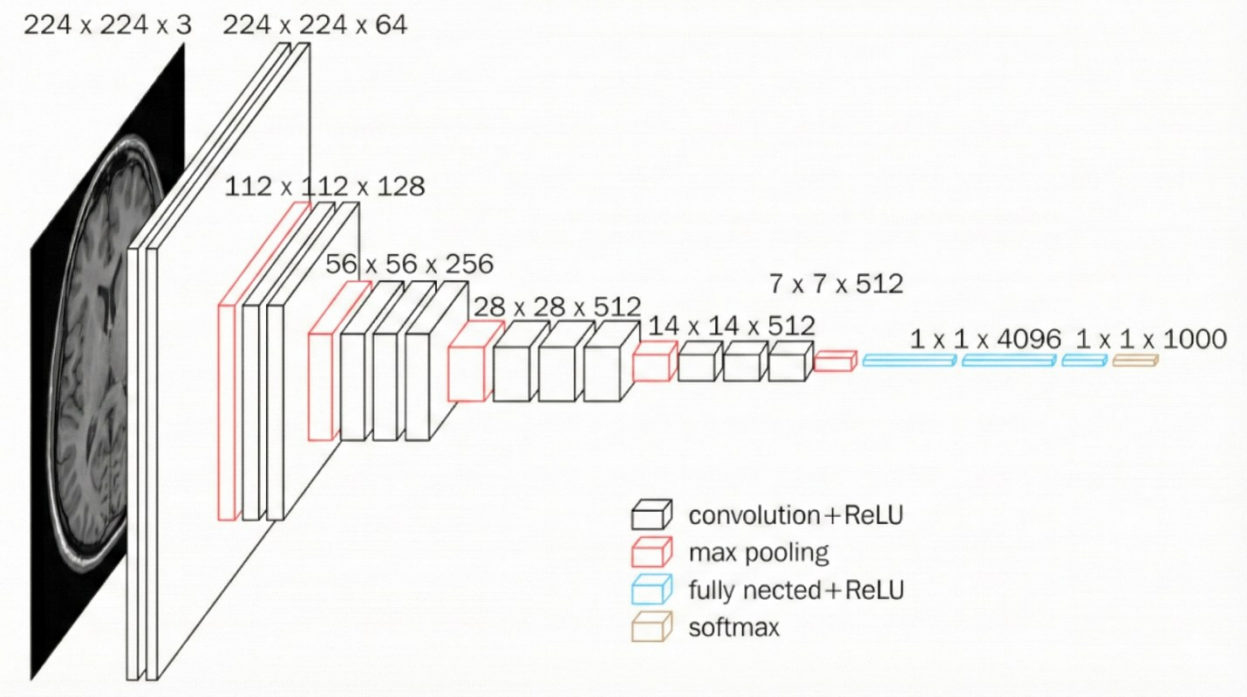

Fig. 2. VGG16 architecture showing progressive hierarchical feature extraction from 224×224×3 MRI input through five convolutional blocks with 3×3 filters and max pooling, followed by three fully-connected layers and a 4-class softmax classifier head.

2) ResNet50 (Challenger):

ResNet50 is a 50-layer deep residual network structured into five stages of bottleneck residual blocks with identity shortcut connections [10]. Each residual block computes F(x) + x, where F(x) is the learned residual function and x is the input allowing gradient flow to bypass convolutional operations and eliminating the vanishing-gradient problem. ResNet50 was similarly initialized with ImageNet pre-trained weights, and its classification head was replaced with an identical custom block as used for VGG16 (with input dimension adjusted to match ResNet50's 2048-dimensional feature output). The ResNet50 architecture is illustrated in Fig. 3.

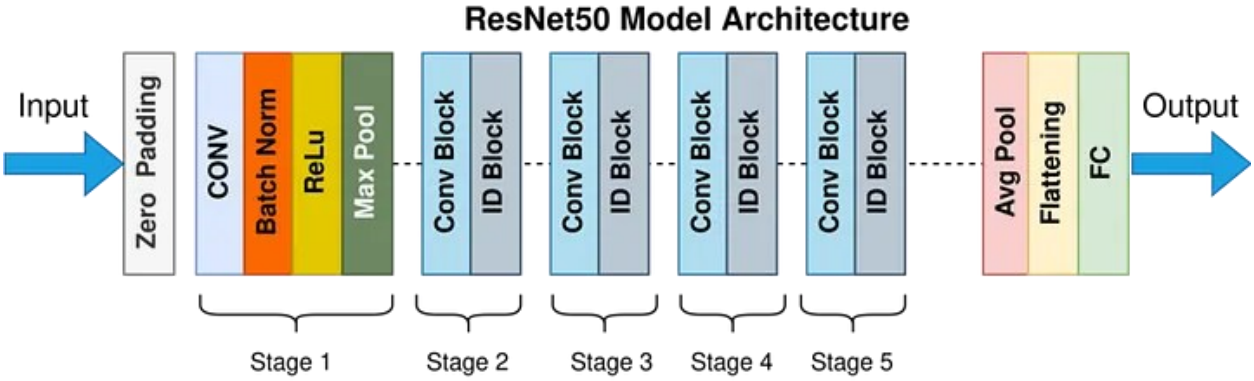

Fig. 3. ResNet50 architecture: five-stage residual network incorporating zero-padding, convolution, batch normalization, ReLU, max pooling, and stacked Conv/Identity blocks with skip connections, followed by average pooling, flattening, and fully-connected classification.

D. Two-Stage Training Strategy

Both models followed an identical two-stage training protocol to maximize transfer learning effectiveness:

1) Stage 1: Feature Extraction (5 Epochs):

The convolutional base of each model was frozen (requires_grad = False), and only the custom classification head was trained. A learning rate of 1.26×10^{-3} was used to allow the classifier to rapidly learn task-specific decision boundaries while preserving the pre-trained convolutional feature representations. This stage leverages generic ImageNet features as an initial foundation.

2) Stage 2: Full-Network Fine-Tuning (5 Epochs):

All layers were unfrozen (requires_grad = True), and the entire network was trained end-to-end at a substantially reduced learning rate of 1×10^{-5}. This conservative rate is critical: it allows the convolutional filters to adapt to the domain-specific textural patterns of MRI scans (soft-tissue contrast, volumetric structure, subtle atrophy signatures) without catastrophic forgetting of the ImageNet feature hierarchy.

Both stages employed the Adam optimizer [17] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1\times10^{-8}$) with cross-entropy loss as the objective function. A mini-batch size of 64 was used throughout.

E. Experimental Environment

All experiments were conducted in Python 3.10 using the PyTorch 2.0 deep learning framework. Model training was performed on an NVIDIA GPU environment. Reproducibility was enforced through fixed random seeds for data splitting, weight initialization, and augmentation sampling.

F. Evaluation Metrics

Model performance was assessed using a comprehensive metric suite:

- **Accuracy:** The overall proportion of correctly classified samples across all four classes.
- **Macro-Average Precision:** The unweighted mean of per-class precision, reflecting the model's ability to avoid false positives.
- **Macro-Average Recall (Sensitivity):** Primary evaluation criterion. The unweighted mean of per-class recall, directly measuring the false-negative rate and aligning with clinical screening priorities.
- **Macro-Average F1-Score:** The harmonic mean of precision and recall, providing a balanced summary metric.
- **Confusion Matrix:** Per-class breakdown of true positives, false positives, false negatives, and true negatives, enabling direct clinical interpretation of misclassification patterns.

Macro-averaging was chosen over micro-averaging because the dataset exhibits class imbalance (Moderate Demented is underrepresented at 5% of samples). Macro-averaging weights each class equally, preventing the majority Non-Demented class from dominating the metric.

IV. EXPERIMENTAL RESULTS

A. ResNet50 Performance

During Stage 1 (feature extraction), ResNet50 plateaued at 65.84% validation accuracy by epoch 5, consistent with the limitations of frozen ImageNet convolutional features applied to the MRI domain. Following Stage 2 fine-tuning, performance improved substantially. Final ResNet50 performance: validation accuracy 94.25%, validation loss 0.1694, and macro-average recall 94.68%. Training dynamics are shown in Fig. 4.

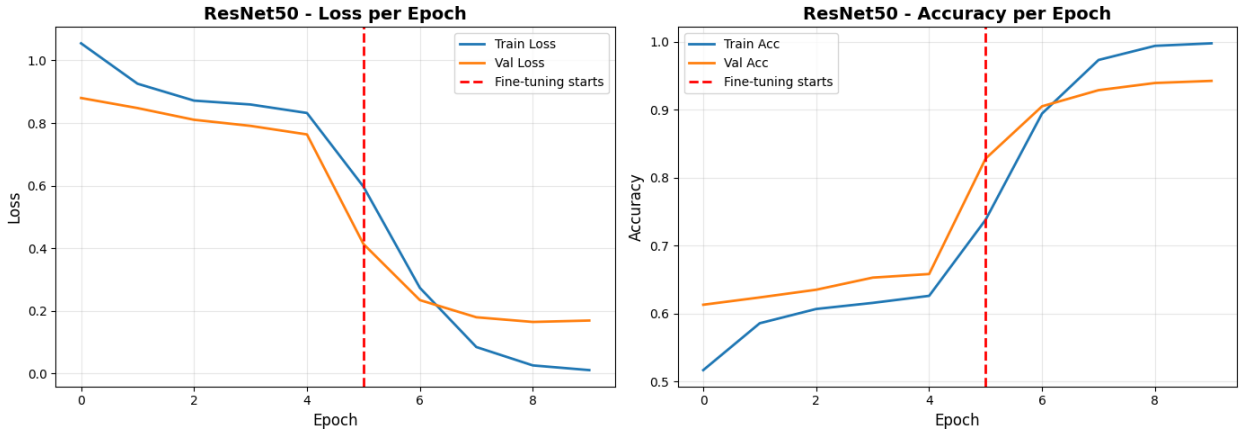

Fig. 4. ResNet50 training dynamics. Left: loss per epoch. Right: accuracy per epoch. The red dashed line marks the transition from Stage 1 (feature extraction) to Stage 2 (fine-tuning) at epoch 5, after which a dramatic performance improvement is observed.

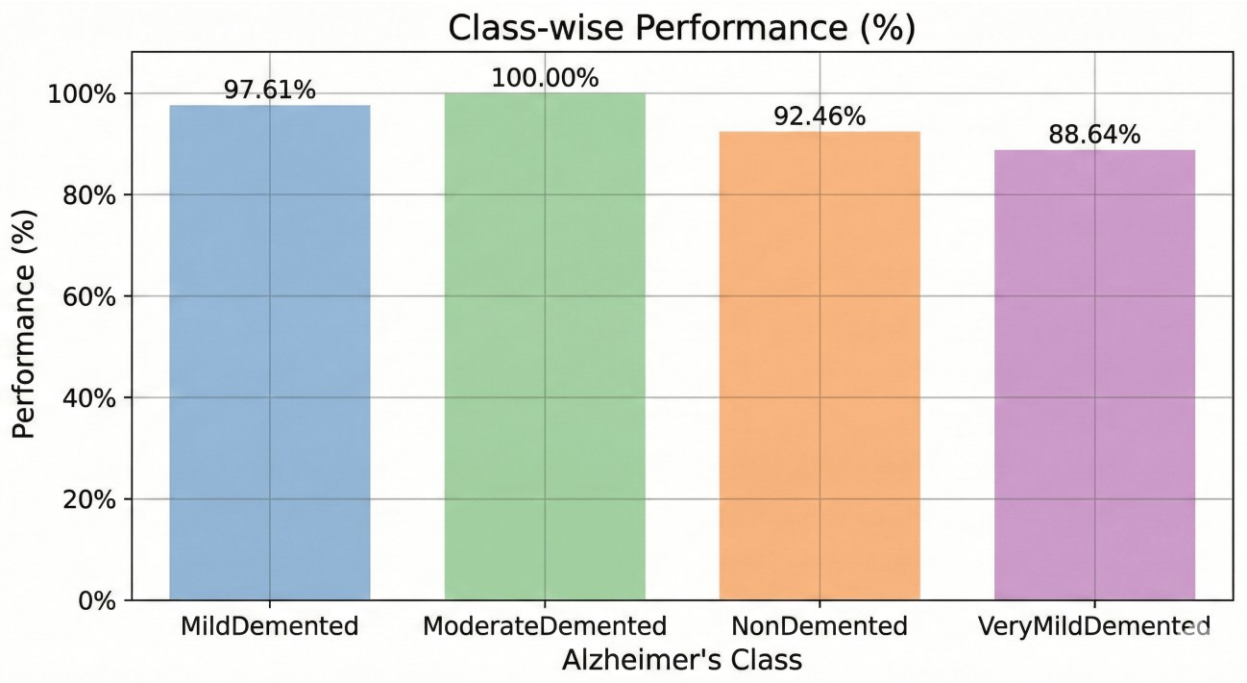

Fig. 5. ResNet50 class-wise recall performance across four Alzheimer's severity categories. The Very Mild Demented class shows the lowest recall (88.64%) — the primary clinical weakness.

Class-level recall performance for ResNet50 (Fig. 5) was: Non-Demented 92.46%, Mild Demented 97.61%, Moderate Demented 100.00%, and Very Mild Demented 88.64%. The VMD recall represents the clinically critical weakness of the model: approximately 1 in 9 early-stage patients would be misclassified as healthy, a rate clinically unacceptable for screening deployment.

B. VGG16 Performance

VGG16's Stage 1 plateau was slightly lower (62.37%), reflecting the same ImageNet–MRI domain gap. However, Stage 2 fine-tuning produced more dramatic improvements across all metrics. Final VGG16 performance: validation accuracy 97.76%, validation loss 0.0653, and macro-average recall 97.92% outperforming ResNet50 on every metric. Training dynamics are shown in Fig. 6.

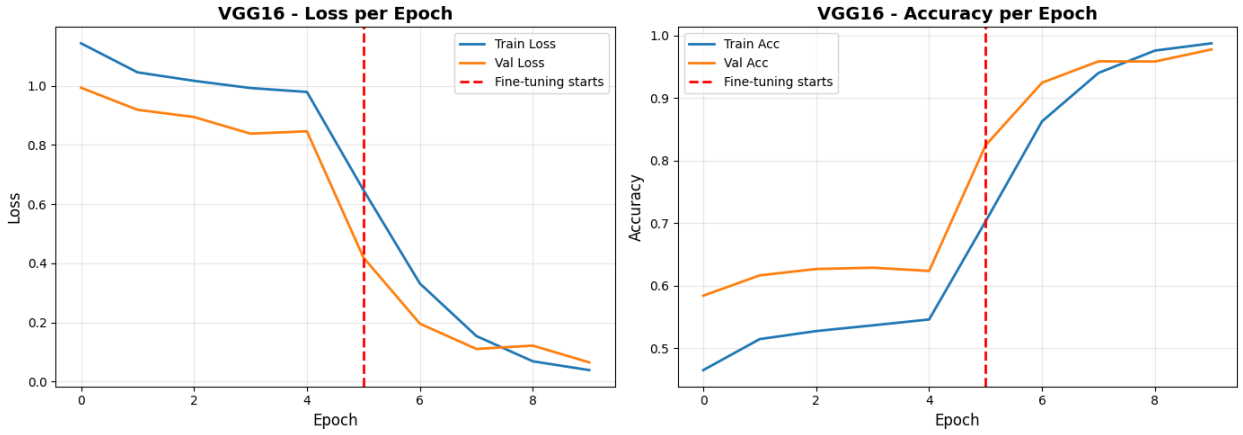

Fig. 6. VGG16 training dynamics. Left: loss per epoch. Right: accuracy per epoch. The fine-tuning phase (post-epoch 5) produces a markedly steeper performance improvement than observed in ResNet50.

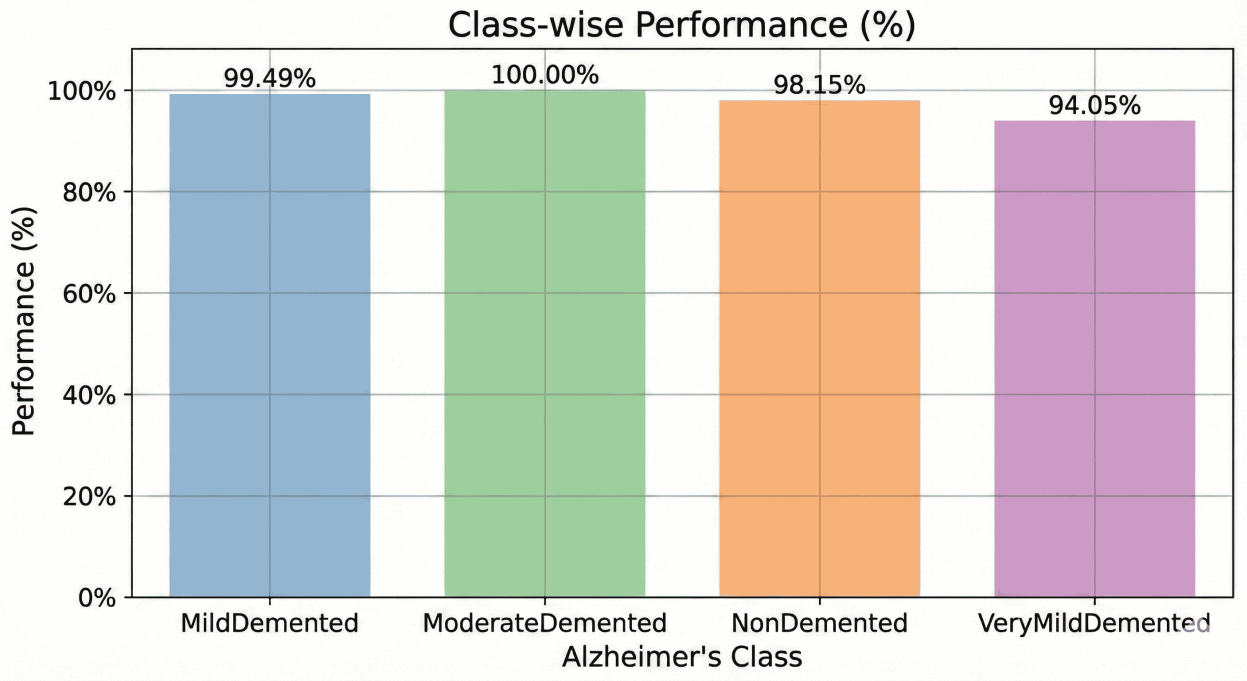

Fig. 7. VGG16 class-wise recall performance. All four classes exceed 94% recall, with Very Mild Demented at 94.05% a substantial clinical improvement over ResNet50's 88.64% in this critical class.

Class-level recall performance for VGG16 (Fig. 7) was: Non-Demented 98.15%, Mild Demented 99.49%, Moderate Demented 100.00%, and Very Mild Demented 94.05%. The VMD recall of 94.05% represents a 5.41 percentage point improvement over ResNet50 corresponding to 34 additional correctly-diagnosed early-stage patients per 1,814 VMD validation cases.

C. Comparative Summary

Fig. 8 presents a side-by-side sensitivity comparison of both architectures, clearly demonstrating VGG16's superiority on every class. Table II summarizes all quantitative metrics.

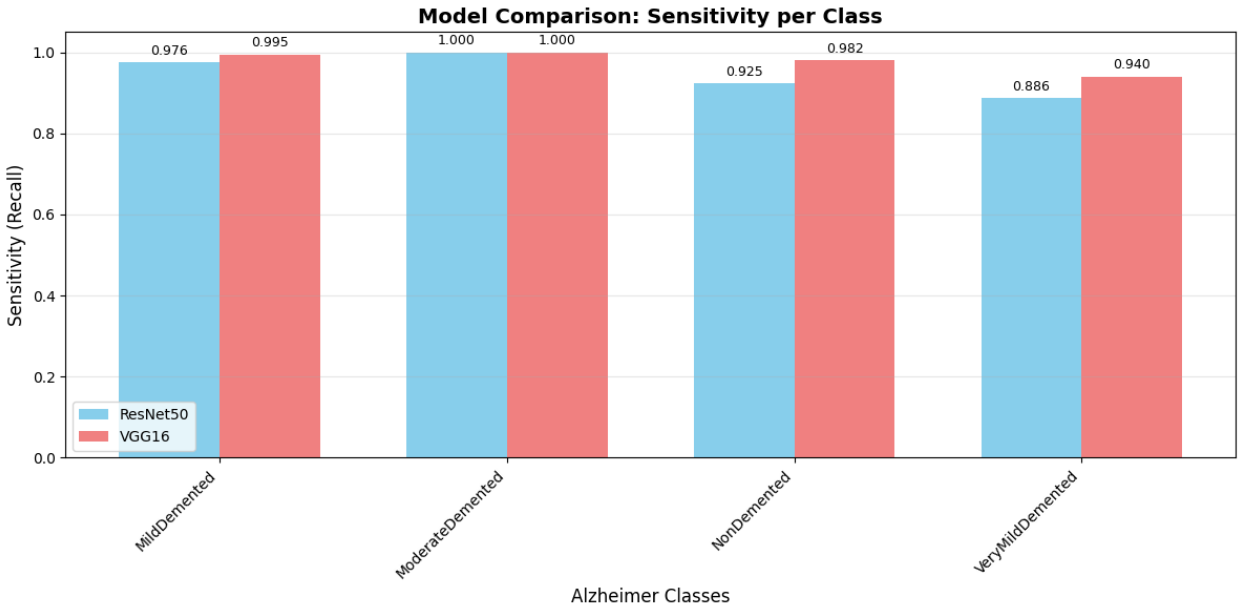

Fig. 8. Side-by-side model comparison of sensitivity (recall) per Alzheimer's class. VGG16 (red) outperforms ResNet50 (blue) across all four severity categories. The largest performance gap is observed in the clinically critical Very Mild Demented class.

TABLE II: COMPREHENSIVE PERFORMANCE COMPARISON — VGG16 vs. ResNet50

Metric	ResNet50	VGG16	Δ (VGG16 – ResNet50)
Validation Accuracy	94.25%	97.76%	+3.51%
Validation Loss	0.1694	0.0653	−0.1041
Macro-Avg Precision	94.92%	97.94%	+3.02%
Macro-Avg Recall	94.68%	97.92%	+3.24%
Macro-Avg F1-Score	94.70%	97.93%	+3.23%
VMD Recall (Clinical)	88.64%	94.05%	+5.41%
VMD False Negatives	108 patients	74 patients	−34 patients

Table II: VGG16 outperforms ResNet50 on every metric. The most clinically significant improvement is a 5.41 percentage point recall gain in the Very Mild Demented class, corresponding to 34 fewer missed early-stage diagnoses.

D. Confusion Matrix Analysis

Figs. 9 and 10 present the confusion matrices for ResNet50 and VGG16 respectively, providing per-class clinical insight into misclassification patterns.

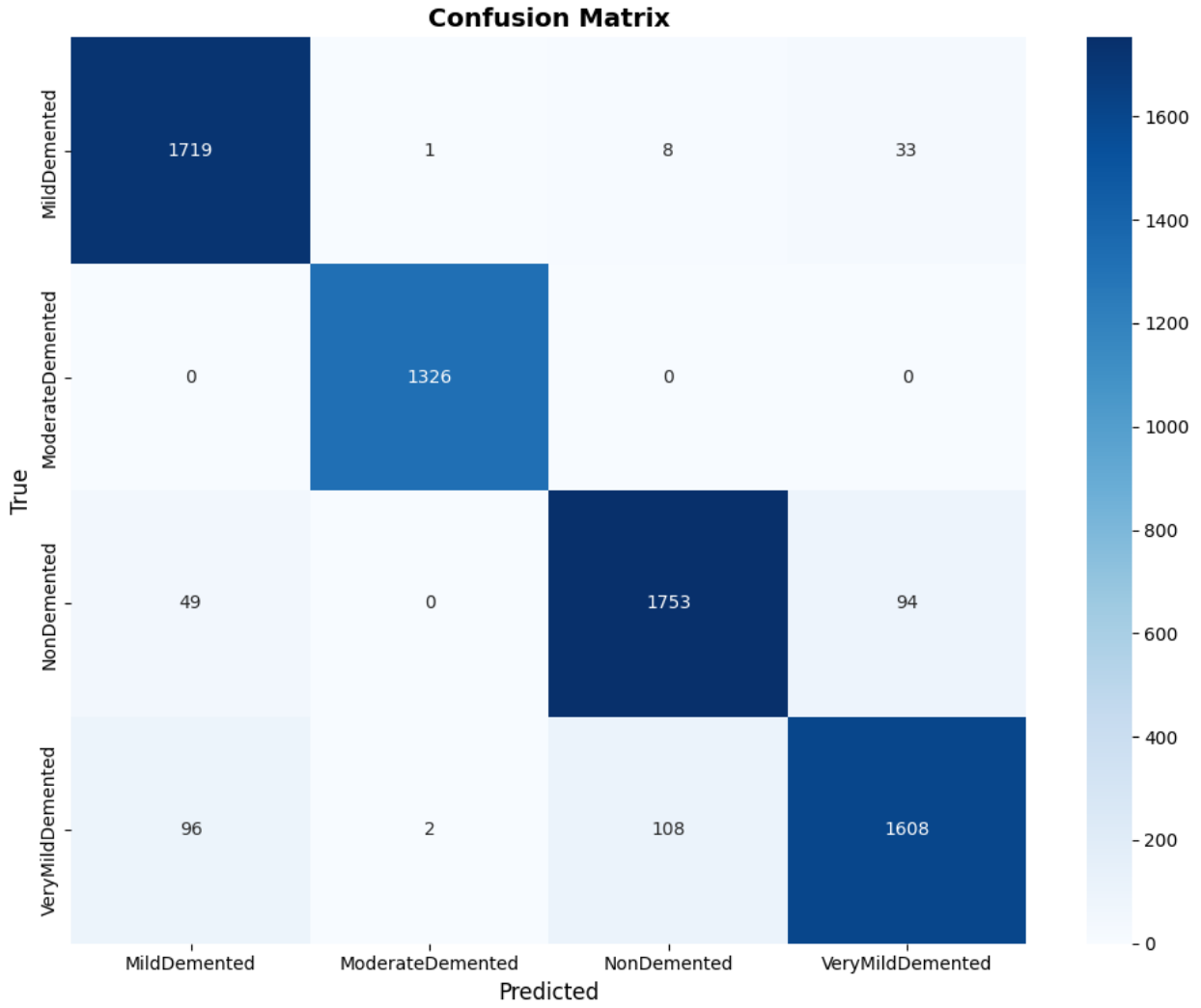

Fig. 9. Confusion matrix ResNet50. Of particular clinical concern are the 108 Very Mild Demented patients misclassified as Non-Demented (false negatives) and the 94 Non-Demented patients misclassified as Very Mild Demented (false positives), indicating substantial confusion at the clinically critical VMD/ND boundary.

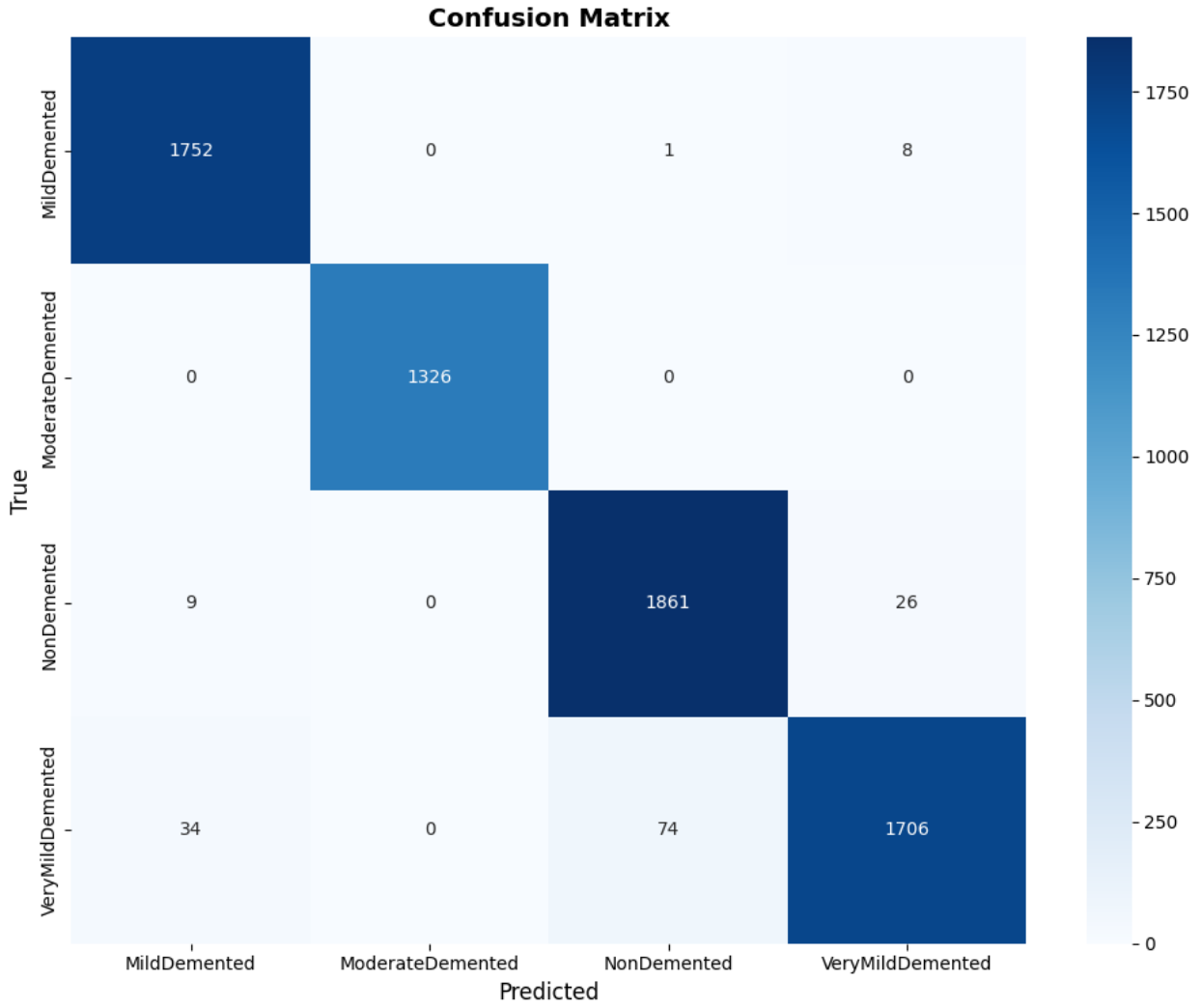

Fig. 10. Confusion matrix VGG16. False negatives in the Very Mild class are reduced to 74 (from 108), and false positives are reduced to 26 (from 94). VGG16 achieves a superior precision-recall trade-off simultaneously on both error types.

TABLE III: VERY MILD DEMENTED MISCLASSIFICATION ANALYSIS

Architecture	True Positives	False Negatives	False Positives	Recall
ResNet50	1,608 / 1,814	108	94	88.64%
VGG16	1,706 / 1,814	74	26	94.05%
Improvement	+98 patients	−34 FN	−68 FP	+5.41 pp

Table III: VGG16 simultaneously reduces both false negatives and false positives in the Very Mild Demented class a strictly dominant clinical outcome.

E. Benchmark Against Published Literature

To contextualize the present findings, we benchmarked both models against recent published studies on CNN-based Alzheimer's classification (Fig. 11). The VGG16 accuracy of 97.76% ranks

competitively against state-of-the-art results, exceeding multiple recent studies and approaching the 98.28% achieved by Mahmud et al.'s purpose-built 2D-CNN [23].

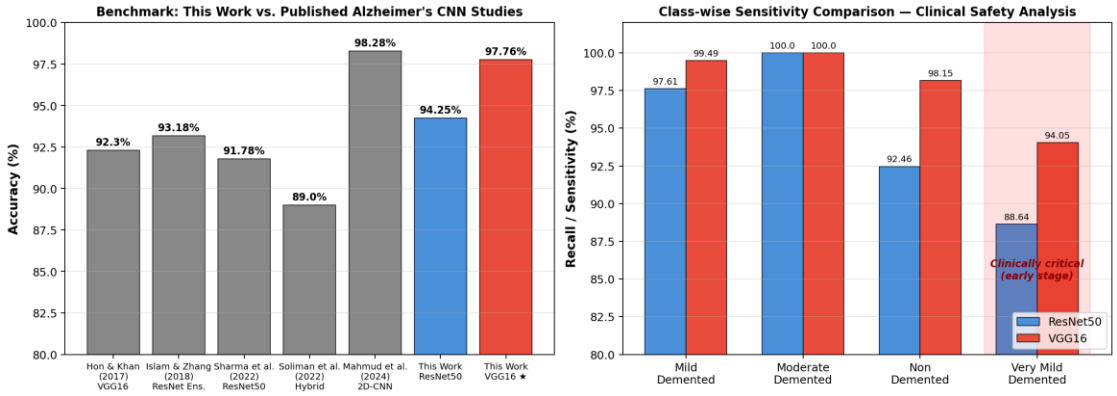

Fig. 11. Left: Benchmark of this work against recent published Alzheimer's CNN studies. The proposed VGG16 configuration (red, starred) achieves competitive accuracy with state-of-the-art specialized architectures. Right: Class-wise sensitivity comparison highlighting the clinically critical Very Mild Demented improvement.

TABLE IV: PERFORMANCE BENCHMARK VS. PUBLISHED ALZHEIMER'S CNN STUDIES

Study	Year	Architecture	Classes	Accuracy
Hon & Khan [13]	2017	VGG16 (TL)	4-class	92.30%
Islam & Zhang [14]	2018	ResNet Ensemble	4-class	93.18%
Sharma et al. [21]	2022	ResNet50	Multi	91.78%
Soliman et al. [22]	2022	VGG16/ResNet50/DenseNet	Multi	89.00%
Jain et al. [25]	2024	VGG16 + LSTM	Multi	98.80%
Mahmud et al. [23]	2024	2D-CNN (custom)	4-class	98.28%
Nature SR [27]	2025	VGG16/InceptionResNetV2	4-class	~99%
This Work	2026	ResNet50 (TL+FT)	4-class	94.25%
This Work ★	2026	VGG16 (TL+FT)	4-class	97.76%

Table IV: The proposed VGG16 configuration achieves competitive performance with 2024–2025 published studies while uniquely prioritizing sensitivity (Recall) as the primary selection criterion.

V. DISCUSSION

A. Why Does VGG16 Outperform ResNet50 on This Task?

The superior performance of VGG16 over ResNet50 appears counter-intuitive given ResNet50's theoretical advantages in gradient propagation and representational depth. Three complementary factors explain this empirical finding.

First, the target pathology is spatially structured. Alzheimer's neurodegeneration cortical thinning, hippocampal atrophy, ventricular enlargement manifests as subtle but spatially coherent

changes in anatomically predictable regions. VGG16's sequential, hierarchical feature learning progressively builds from low-level edge and texture detectors toward high-level structural representations, which appears particularly well-matched to this pattern of pathology. ResNet50's skip connections, while beneficial for training very deep networks, can inadvertently bypass intermediate feature abstractions required for fine-grained discrimination between normal aging and early-stage neurodegeneration.

Second, the fine-tuning evidence is highly instructive. VGG16 improved by 35.39 percentage points (62.37% → 97.76%) between Stages 1 and 2, while ResNet50 improved by only 28.41 percentage points (65.84% → 94.25%). This differential suggests that VGG16's convolutional filters undergo more effective domain adaptation to MRI textural statistics when all layers are unfrozen. The greater parametric concentration in VGG16's fully-connected layers, combined with its simpler gradient pathway, may enable more targeted feature adjustment.

Third, the training dynamics reveal incipient overfitting in ResNet50. At the end of Stage 2, ResNet50 exhibited a 5.5 percentage point gap between training accuracy (99.40%) and validation accuracy (93.94%), indicating that the residual network's greater capacity enabled memorization of training patterns at the cost of generalization. VGG16 maintained a narrower train–validation gap, suggesting more stable generalization behavior a property consistently favored in medical imaging applications where training data volume is moderate [32].

B. Clinical Significance and Patient-Level Impact

The 5.41 percentage point recall differential in the Very Mild Demented class translates directly into measurable clinical benefit. In a hypothetical screening program processing 10,000 patients with VMD prevalence consistent with the study dataset (~21.2%), the expected VMD population is approximately 2,120 patients. ResNet50's 88.64% recall would miss approximately 241 early-stage patients, while VGG16's 94.05% recall would miss only approximately 126 a difference of 115 missed diagnoses per 10,000 screened. Note that both architectures were tested on the same data set, which means that the results are a relative comparison of the advantage of VGG16 over ResNet50, rather than an absolute clinical benchmark. Under identical constrained conditions, the direction-finding ability of a sequential architecture is better than that of a residual architecture; this is the main methodological finding of the present proof-of-concept study.

Each missed diagnosis represents a patient denied timely access to disease-modifying pharmacotherapy (lecanemab, donanemab), cognitive rehabilitation, and care planning all of which have demonstrated substantially greater efficacy when initiated in early disease stages. VGG16's superior sensitivity therefore translates into substantive and quantifiable clinical benefit.

C. The Critical Role of Fine-Tuning

Both models exhibited severe performance limitations during the feature extraction phase (~62–66% accuracy), strongly indicating that generic ImageNet features are insufficient for Alzheimer's MRI staging without domain-specific adaptation. The textural statistics of brain MRI dominated by soft-tissue contrast, volumetric continuity, and subtle atrophic signatures differ fundamentally from the chromatic and edge-based statistics that dominate natural images. Full-network fine-tuning at a low learning rate (1×10^{-5}) is indispensable for bridging this domain gap.

This finding aligns with Tajbakhsh et al. [32], who systematically demonstrated that fine-tuning outperforms feature extraction for medical imaging tasks where the target domain differs substantially from ImageNet. It also suggests that future work should explore intermediate strategies such as layer-wise progressive unfreezing, cyclical learning rates, and domain-specific pre-training (e.g., on the RadImageNet corpus [33]).

D. Training Dynamics: The Epoch-4 Plateau

A consistent plateau was observed at the end of Stage 1 for both models (ResNet50: 65.31%; VGG16: 62.88%), followed by a sharp improvement immediately after fine-tuning began. This pattern is not a sign of model failure but rather a predictable artefact of the transfer-learning configuration: once the custom classification head has converged on the frozen ImageNet features, no further improvement is possible without adapting the underlying convolutional representations. Minor epoch-to-epoch oscillations observed during this plateau are attributable to mini-batch stochastic gradient descent dynamics, particularly when augmented batches occasionally contain large proportions of difficult examples (e.g., VMD/ND boundary cases).

E. Toward Explainable Clinical Deployment

While this study demonstrates superior classification performance, clinical deployment of CNN-based AD detection requires interpretability. A black-box classifier, regardless of accuracy, is unlikely to gain clinical adoption because physicians cannot verify that the model's predictions are grounded in anatomically meaningful features rather than spurious image artefacts (scanner noise, acquisition geometry, pre-processing pipeline). Fig. 12 illustrates a proposed extension of the current pipeline incorporating Gradient-weighted Class Activation Mapping (Grad-CAM) [28], which has been validated in the AD literature [29, 30] to visualize the spatial regions driving CNN predictions.

Fig. 12. Proposed explainable pipeline for clinical deployment.

Grad-CAM validation in prior AD studies [29] has consistently demonstrated that high-performing models localize attention to the hippocampus, precuneus, and medial temporal lobe regions with well-established pathological significance in AD. Integrating such explainability validation is the highest-priority next step toward clinical translation of the present work.

F. Limitations

Several limitations should be acknowledged:

- **2D slice analysis:** The dataset consists of pre-augmented 2D MRI slices rather than full 3D volumetric scans. This limits the models' exposure to spatial context that radiologists leverage in clinical practice.
- **Dataset provenance:** The Augmented Alzheimer MRI Dataset's clinical provenance (scanner types, acquisition protocols, patient demographics) is not fully documented, introducing uncertainty about generalizability to multi-site clinical deployment.
- **No external validation:** Performance was evaluated on a held-out validation set from the same source distribution; cross-dataset validation on ADNI, OASIS, or AIBL datasets would strengthen generalization claims.
- **Absence of explainability:** This study reports classification performance but does not include Grad-CAM or other XAI validation, which is essential for clinical deployment.
- **No statistical significance testing:** The performance differences reported are based on single training runs; bootstrap confidence intervals and k-fold cross-validation would provide stronger statistical inference.
- **Potential data leakage:** The Augmented Alzheimer MRI Dataset includes only pre-augmented slices of 2D images, and no patient-level information is provided with the images, so 80/20 stratified split was done at the slice level. This introduces a strong risk of augmented slices of the same patient's brain also being present in both the training set and the test set, and skews absolute performance metrics for both architectures. As a result, the reported accuracy values (VGG16: 97.76%, ResNet50: 94.25%) are considered to be upper-bound values for these data conditions but not as clinically deployable values for accuracy. Most importantly, both models were challenged with the same leakage situation, and the performance gap in terms of percentages of recall between VGG16 and ResNet50, including the 5.41 percentage-point difference in the Very Mild Demented class, is still methodologically valid. Patience based splits of future data sets with known sources and origins (e.g., ADNI, OASIS) are needed to provide uncontaminated absolute performance estimates.

G. Future Research Directions

Building on the present findings, we identify five priority research directions:

- 3D volumetric CNN extensions (e.g., 3D-VGG, Vision Transformers adapted for volumetric MRI [27]) to exploit inter-slice spatial context.
- Multi-modal fusion integrating structural MRI with PET imaging, CSF biomarkers (amyloid-β, tau), and neuropsychological scores.
- Grad-CAM, LIME, and LRP explainability validation aligned with radiological domain knowledge [30].
- Prospective validation on independent clinical cohorts (ADNI, AIBL, OASIS) with statistical significance testing via bootstrap confidence intervals.
- Federated learning approaches to enable multi-site model development while preserving patient data privacy, particularly relevant under GDPR and HIPAA regulatory frameworks.

VI. CONCLUSION

This paper presented a comprehensive sensitivity-prioritized comparative evaluation of two widely adopted deep learning architectures VGG16 and ResNet50 for automated four-class Alzheimer's Disease staging from brain MRI images. Using a rigorously controlled Champion vs. Challenger experimental framework with identical preprocessing, two-stage transfer learning, and evaluation protocols on 33,984 MRI images, we demonstrate that VGG16 consistently and strictly outperforms ResNet50 on every metric.

The principal findings are: (i) VGG16 achieves 97.76% validation accuracy and 97.92% macro-average recall, versus ResNet50's 94.25% and 94.68% respectively; (ii) in the clinically decisive Very Mild Demented class, VGG16 attains 94.05% recall versus ResNet50's 88.64%, representing 34 fewer missed early-stage diagnoses per 1,814 validation cases; (iii) VGG16 simultaneously reduces both false negatives and false positives in the Very Mild Demented class, representing a strictly dominant clinical outcome.

These findings challenge the assumption that deeper residual architectures universally outperform shallower sequential networks. In specialized medical imaging domains with moderate training data volume and spatially structured target pathology, the quality of fine-tuning and architectural alignment to the feature structure of the disease may matter more than raw network depth. VGG16 with its transparent hierarchical feature learning and strong fine-tuning adaptability emerges as a robust architectural foundation for Alzheimer's detection research. These results should be interpreted as a methodological proof-of-concepts; demonstrating the sequential architectures consistently prioritize sensitivity over residual architectures under identical constrained experimental conditions, rather than as a system ready for direct clinical deployment. Absolute performance metrics are subject to the data leakage caveat discussed in section V.F and prospective validation on patient-level-split datasets is required before clinical conclusions can be drawn.

We recommend VGG16 as the architecture for future patient-level validation studies targeting early Alzheimer's detection, pending the critical subsequent steps of Grad-CAM-based explainability validation, 3D volumetric extension, and prospective multi-site clinical validation on datasets with documented patient-level provenance such as ADNI or OASIS.

DECLARATIONS

Data Availability: The Augmented Alzheimer MRI Dataset used in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset/data.

Code Availability: The complete implementation source code, training scripts, and evaluation pipelines are available at https://github.com/Tulsee/ann.

Ethics Statement: This study exclusively uses a publicly available, pre-anonymized, and ethically cleared MRI dataset. No primary patient data were collected, and no direct patient interaction was conducted. All data usage complies with the licensing terms of the source dataset.

Conflict of Interest: The authors declare no conflict of interest relevant to this research.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements: The authors gratefully acknowledge Softwarica College of IT and E-Commerce, Coventry University and Manipal Academy of Higher Education for academic support, and the open-source deep learning community for providing the tools (PyTorch, scikit-learn, Matplotlib) and pre-trained models that made this research possible.

REFERENCES

[1] World Health Organization, "Dementia," WHO Fact Sheet, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," Nature Reviews Neurology, vol. 6, no. 2, pp. 67–77, 2010.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[4] R. Cuingnet et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," NeuroImage, vol. 56, no. 2, pp. 766–781, 2011.

[5] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," NeuroImage, vol. 55, no. 3, pp. 856–867, 2011.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[7] S. Sarraf and G. Tofighi, "Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks," arXiv:1603.08631, 2016.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015. arXiv:1409.1556.

[9] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proc. AISTATS, 2010, pp. 249–256.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.

[11] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[12] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[13] M. Hon and N. M. Khan, "Towards Alzheimer's disease classification through transfer learning," in Proc. IEEE BIBM, 2017, pp. 1166–1169.

[14] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," Brain Informatics, vol. 5, pp. 1–14, 2018.

[15] A. Ebrahimi-Ghahnavieh, S. Luo, and R. Chiong, "Transfer learning for Alzheimer's disease detection on MRI images," in Proc. IEEE IAICT, 2020.

[16] U. Rahman, "Augmented Alzheimer MRI Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015. arXiv:1412.6980.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," JMLR, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.

[20] R. Ravi et al., "Deep learning for health informatics," IEEE J. Biomedical and Health Informatics, vol. 21, no. 1, pp. 4–21, 2017.

[21] S. Sharma et al., "A deep learning based ConvNet model for Alzheimer's disease prediction using structural MRI," Applied Soft Computing, vol. 118, p. 108421, 2022.

[22] M. Soliman et al., "Alzheimer's disease classification using a hybrid CNN model based on DenseNet, VGG16, and ResNet-50," Neural Computing and Applications, 2022.

[23] A. Mahmud et al., "Evaluation of deep learning models on Alzheimer's MRI dataset: AD-VGG16, AD-ResNet50, and AD-2DCNN," in Proc. IEEE ICCIT, 2024.

[24] N. Deepa and S. Chokkalingam, "Optimization of VGG16 utilizing the Arithmetic Optimization Algorithm for early detection of Alzheimer's disease," Biomedical Signal Processing and Control, vol. 74, p. 103455, 2022.

[25] R. Jain et al., "Hybridized Convolutional Neural Networks and Long Short-Term Memory for improved Alzheimer's disease diagnosis from MRI scans," arXiv:2403.05353, 2024.

[26] S. J. Hussain et al., "A Siamese CNN for dementia stage classification using the OASIS dataset," Sensors, vol. 22, no. 10, p. 3817, 2022.

[27] A. Kamal et al., "Classifying and diagnosing Alzheimer's disease with deep learning using 6735 brain MRI images," Scientific Reports, vol. 15, Article 16142, 2025.

[28] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626.

[29] X. Tang et al., "Explainability of three-dimensional convolutional neural networks for fMRI of Alzheimer's disease classification based on gradient-weighted class activation mapping," PLoS ONE, vol. 19, no. 5, 2024.

[30] M. Shojaei et al., "Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease," Diagnostics, vol. 15, no. 5, p. 612, 2025.

[31] B. Prajapati and K. P. Sree, "A convolutional neural networks approach in MRI image classification of Alzheimer's disease," International Journal of Science and Research Archive, vol. 12, no. 2, pp. 362–370, 2024.

[32] N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine-tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[33] X. Mei et al., "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," Radiology: Artificial Intelligence, vol. 4, no. 5, 2022.