International Journal of Advanced AI Applications
Print ISSN: 3104-932X
Online ISSN: 3104-9338

Empirical Evaluation of Artificial Intelligence Transfer Learning in Data-Constrained Dementia Screening: Analyzing Overfitting and Representation Collapse on Brain MRI

Yuling Li*, Yiti Huang, Shuangli Wu

School of Traditional Chinese Medicine, Hubei University of Chinese Medicine, Wuhan, Hubei, China

Received: Nov 1, 2026

Revised: Nov 7, 2026

Accepted: Nov 9, 2026

Published online:November 11, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 3, No. 1 (January 2027)

* Corresponding Author: Yuling Li (13659814287@163.com)

Abstract. Early screening and precise staging of dementia, particularly Alzheimer's Disease (AD), are vital for implementing timely clinical and lifestyle interventions. While deep convolutional neural networks (CNNs) have shown significant promise in structural magnetic resonance imaging (sMRI) computer-aided diagnosis, their generalization robustness remains a critical vulnerability in data-constrained clinical fine-tuning scenarios. This paper investigates the comparative learning dynamics and failure modes of two representative vision backbones—a standard residual network (ResNet18) and a lightweight architecture (MobileNetV2)—on a four-class dementia classification task (Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented) under constrained data regimes. Empirical results demonstrate severe generalization divergence across both architectures: ResNet18 exhibited unconstrained empirical memorization, achieving 100.0% training accuracy while degrading to 16.67% on unseen test instances. Conversely, MobileNetV2 suffered from complete representation collapse; despite reaching 82.50% training accuracy, its test accuracy plateaued at exactly 25.00% across all epochs, with confusion matrix profiling revealing that 100% of test predictions collapsed into the single Mild Demented category. Qualitative trajectory and error-matrix analyses confirm that generic pre-trained weights without domain-adapted regularization are highly prone to capturing non-pathological imaging artifacts or falling into trivial local optima in ultra-low data regimes. These findings highlight critical risks of decision collapse in deploying lightweight deep models for point-of-care dementia screening and underscore the imperative for domain-specific self-supervised pre-training and robust class-balancing protocols.

Keywords: *dementia screening; brain MRI; transfer learning; overfitting; mode collapse; ResNet18; MobileNetV2*

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

1. Introduction

Dementia represents one of the most critical public health challenges of the 21st century. According to clinical epidemiological surveys, Alzheimer's Disease (AD) accounts for approximately 60%–70% of all progressive neurodegenerative diagnoses globally[1]. Structural Magnetic Resonance Imaging (sMRI) serves as an indispensable in vivo neuroimaging biomarker for evaluating neurodegenerative progression[2,3]. In clinical workflows, sMRI allows neuroradiologists to assess macroscopic structural hallmarks such as bilateral hippocampal atrophy, ventricular enlargement, and accelerated neocortical thinning[1]. However, manual radiologic reading is labor-intensive, operator-dependent, and prone to inter-observer variability, especially in pre-clinical or transition phases such as Mild Cognitive Impairment (MCI) or Very Mild Dementia[3,4].

From a neuropathological perspective, the clinical continuum of dementia unfolds over an extended preclinical trajectory before overt cognitive deficits become clinically measurable. During the transition from healthy aging (Non-Demented) to Very Mild and Mild Dementia, subtle microstructural degeneration begins primarily within the medial temporal lobe structures, specifically the entorhinal cortex and hippocampal subfields, before propagating across the cerebral neocortex. On two-dimensional structural MRI slices, these transitional pathological changes manifest as millimeter-scale widening of the temporal horns of the lateral ventricles and slight sulcal enlargement. Because these subtle morphological variations closely overlap with normal age-related cerebral volume loss, visual differentiation requires exceptional radiological expertise. Furthermore, in primary care hospitals and regional memory clinics, neuroradiologists face surging imaging volumes, which exacerbates diagnostic fatigue and increases the risk of misclassifying early prodromal cases. Consequently, developing objective, quantitative, and reproducible computational screening tools capable of assisting clinicians in stratifying these four progressive stages has become an urgent clinical priority.

Deep learning (DL) has emerged as an automated solution for computer-aided diagnosis (CAD) across diverse radiological modalities[3,5]. Early convolutional architectures and custom multi-layer networks demonstrated feasibility on raw scans[3]. However, training deep networks from scratch on biomedical imaging frequently suffers from severe overfitting due to the scarcity of large, expert-annotated clinical datasets. Transfer learning—pre-training a deep network on massive natural image corpora such as ImageNet and fine-tuning it on domain-specific biomedical targets—has largely mitigated sample-efficiency constraints[4-6].

Despite these advancements, an operational bottleneck persists in the translation from

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

laboratory benchmarks to real-world clinical and remote point-of-care settings. Many widely cited vision architectures possess parameter counts exceeding tens of millions, requiring dedicated server-grade GPUs[7]. Conversely, clinical endpoints, rural outpatient clinics, and mobile diagnostic units often operate with resource-constrained computational hardware. Lightweight convolutional networks, notably the MobileNet series[8], utilize depthwise separable convolutions to substantially curtail floating-point operations (FLOPs) and memory footprint.

However, deploying these architectural innovations into clinical neuroimaging workflows introduces a fundamental optimization challenge: cross-domain distribution shift. Standard transfer learning pipelines initialize convolutional kernels using weights optimized on natural trichromatic scenes, which are characterized by high-contrast object boundaries, diverse lighting conditions, and rich chromatic textures. In stark contrast, structural brain MRI scans are single-channel grayscale acquisitions defined by globally standardized anatomical topologies, subtle local tissue intensity gradients between gray matter, white matter, and cerebrospinal fluid, and high inter-subject structural self-similarity. When fine-tuning is performed on highly constrained local hospital cohorts or pilot clinical registries, this severe domain mismatch interacts dangerously with the intrinsic inductive biases of different neural network architectures.

Furthermore, in real-world clinical settings—such as regional outpatient clinics or integrative medicine departments—annotated neuroimaging cohorts are frequently restricted by patient privacy regulations, high diagnostic acquisition costs, and natural class imbalances across disease progression stages. Under such ultra-low-data regimes, the interplay between network parameter capacity and optimization stability becomes highly fragile. Over-parameterized residual networks risk memorizing patient-specific scanner noise, background intensity inhomogeneities, or skull-stripping boundaries, whereas highly compressed depthwise separable networks may experience severe gradient starvation and representation collapse during cross-domain adaptation. Understanding why and how these distinct deep learning architectures fail under data constraints is just as critical for clinical patient safety as optimizing peak accuracy on large benchmarks.

While multiple studies evaluate individual network accuracies on broad cohorts, rigorous empirical investigations into how these networks behave under low-resource or constrained fine-tuning regimes remain scarce [3-5]. To address this gap, this study benchmarks two representative architectures: ResNet18 (representing classical residual learning) and

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

MobileNetV2 (representing inverted residual and depthwise separable architectures) [7,8]. The contributions of this work are threefold:

(1) We evaluate the transfer learning adaptation of ResNet18 and MobileNetV2 on a four-class dementia classification task under controlled data constraints [2].

(2) We systematically document the divergence between training memorization and test generalization across successive epochs.

(3) We conduct an error-matrix analysis to inspect severe failure modes, explicitly identifying the mechanisms behind empirical memorization and representation mode collapse [1,4].

2. Materials and Methods

2.1. Dataset Description and Preprocessing.

This study evaluates a four-class dementia screening protocol (Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented) derived from standardized neuroimaging benchmark sets [2]. To examine model vulnerability under constrained clinical scenarios, models were trained and evaluated on stratified cohorts ensuring balanced cross-class representation. A balanced hold-out test set (N = 12, 3 independent scans per diagnostic category) was strictly isolated to measure out-of-distribution generalization and detect decision biases. All slices were resized to 224×224 pixels using bilinear interpolation. Grayscale MRI channels were duplicated across three RGB dimensions to match the input tensor specifications of ImageNet pre-trained backbones[6]. Pixel intensities were normalized using standard ImageNet channel means (μ= [0.485, 0.456, 0.406]) and standard deviations (σ = [0.229, 0.224, 0.225]). Online data augmentations during training included random horizontal flipping (p = 0.5) and slight spatial perturbations (±10° rotation) to emulate subtle variations in scanner head positioning [5].

This strictly balanced hold-out test design (N = 12, with exactly n = 3 unseen scans per diagnostic stage) was intentionally formulated as a controlled low-resource stress test. In mathematical terms, evaluating models on an equally stratified four-class cohort establishes an exact uniform prior probability distribution ($P(y = c) = 0.25$ for each class $c \in \{1, 2, 3, 4\}$). In clinical validation studies where test sets are heavily skewed toward healthy controls, a degenerate model that predicts every subject as healthy can artificially achieve an inflated overall accuracy exceeding 80%, thereby masking complete diagnostic failure on diseased cohorts. By enforcing strict four-way balance in the isolated evaluation set, any tendency of a fine-tuned network to collapse into a single majority prediction is immediately exposed as a

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

stagnant 25.00% test accuracy. Moreover, no spatial augmentations or random perturbations were applied during the inference phase on the hold-out test cohort, guaranteeing that all reported generalization metrics reflect deterministic and reproducible out-of-distribution performance.

From a neuroimaging physics and data acquisition standpoint, each step of this preprocessing pipeline plays a specific role in governing downstream feature extraction. First, standardizing slice dimensions to 224 by 224 pixels via bilinear interpolation smooths out high-frequency discretization jaggedness along the cortical ribbon and ventricular margins while preserving the macroscopic volumetric ratios between cerebrospinal fluid spaces and parenchymal tissue. Second, replicating the single grayscale MRI channel across three identical RGB planes allows the network to leverage pre-trained spatial edge, corner, and texture filters in the early convolutional stem; however, it simultaneously renders the color-opponent channels of the pre-trained weights redundant, forcing the network to rely entirely on luminance gradients rather than chromatic contrast. Third, applying global ImageNet channel means and standard deviations standardizes the numerical variance of input tensors entering the first batch normalization layer, yet it does not eliminate scanner-specific intensity inhomogeneities or bias fields inherent to clinical magnetic resonance acquisitions. Finally, restricting online spatial augmentations to horizontal reflections and bounded ten-degree rotations ensures anatomical plausibility—since human brain axial sections exhibit approximate bilateral symmetry across the interhemispheric fissure—while preventing unrealistic geometric distortions that could corrupt subtle atrophic landmarks in the medial temporal lobes.

2.2. Network Architectures.

Both evaluated models utilize feature extraction backbones pre-trained on the ImageNet-1K database[6]. The original 1000-class fully connected classification heads were replaced with a customized multi-class block consisting of:

Input Features→AdaptiveAvgPool2D(1×1)→Dropout(p = 0.3)→Linear

where Din = 512 for ResNet18 [7] and Din = 1280 for MobileNetV2 [8].

ResNet18: Employs standard residual blocks where identity skip connections bypass convolutional operations[7]:

$$y = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{1}$$

This formulation addresses gradient degradation, allowing effective convergence during transfer learning[7].

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

MobileNetV2: Leverages inverted residual blocks with linear bottlenecks and depthwise separable convolutions[8], drastically reducing computational parameters to ~3.50 million while maintaining cross-channel representational capacity.

The structural divergence between ResNet18 and MobileNetV2 provides an ideal theoretical contrast for analyzing small-sample transfer learning dynamics. In ResNet18, spatial filtering and channel mixing are performed jointly through dense 3 ×3 convolutional kernels across four progressive stages with channel widths expanding from 64 to 128, 256, and 512. While the additive identity shortcut ensures unimpeded backward gradient propagation across all 18 layers, the resulting ~11.69 million trainable parameters create an over-complete hypothesis space when confronted with limited medical training instances. Conversely, MobileNetV2 decouples spatial filtering from cross-channel feature synthesis through an expansion-filtering-compression mechanism. Within each inverted residual block, a narrow input bottleneck tensor is first expanded by a factor of six via a 1× 1 pointwise convolution, processed spatially by a lightweight 3×3 depthwise convolution operating independently on each channel, and subsequently projected back into a low-dimensional linear bottleneck via another 1×1 convolution without non-linear activation. Although removing the terminal non-linearity preserves representational fidelity in low-dimensional spaces, the lack of dense cross-channel spatial coupling makes depthwise separable filters highly sensitive to initialization mismatches when adapting from natural images to grayscale medical scans.

2.3. Training Hyperparameters and Experimental Environment.

All models were implemented in PyTorch and trained on a Linux workstation equipped with an NVIDIA GPU and an Intel CPU. Training configurations were standardized across architectures: (1) Loss Function: Categorical Cross-Entropy Loss. (2) Optimizer: Adam optimizer with an initial learning rate η= 1×10-4 and weight decay λ= 1×10-5, (3) Batch Size: 32. (4) Epochs: Models were trained for 5 full epochs to record rapid empirical memorization trajectories.

To mitigate the intrinsic class imbalance present in medical registries (e.g., MoD has substantially fewer instances than ND)[3], we implemented inverse class-frequency loss weighting:

$$w_c = \frac{N_{total}}{4 \times N_c} \tag{2}$$

In this inverse class-frequency formulation, wc denotes the penalty weight assigned to diagnostic category c, Ntotal represents the total number of training instances across the entire

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

cohort, and Nc indicates the sample count belonging to class c. By integrating wc directly into the categorical cross-entropy objective, gradient updates generated by underrepresented clinical stages — most notably Moderate Demented — are proportionally amplified during backpropagation, whereas gradients from more prevalent categories are attenuated.

Furthermore, because aggregate classification accuracy alone cannot distinguish between balanced multi-class competence and degenerate single-class guessing, model evaluation on the hold-out test set was expanded to incorporate granular class-specific performance indicators derived from the confusion matrix. Specifically, for each diagnostic category, we quantified True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). From these fundamental counts, class-wise Precision (TP / (TP + FP)), Recall or Sensitivity (TP / (TP + FN)), Specificity (TN / (TN + FP)), and the harmonic F1-Score (2×Precision×Recall / (Precision+ Recall)) were systematically computed alongside their unweighted macro-averages.

3. Results

3.1. Quantitative Performance Comparison.

Model testing was conducted strictly on the independent hold-out test set (N = 12). Table 1 summarizes the final training and testing accuracy for both candidate architectures.

Table 1. Empirical performance benchmarking under data-constrained conditions.

Architecture	Parameters (M)	Training Accuracy	Testing Accuracy	Diagnostic Regime
ResNet18	~11.69	1.0000 (100.0%)	0.1667 (16.67%)	Severe Overfitting
MobileNetV2	~3.50	0.8250 (82.50%)	0.2500 (25.00%)	Majority Class Collapse

As presented in Table 1, both architectures exhibited pronounced generalization gaps. ResNet18 memorized the training cohort completely, achieving 100% training accuracy within three epochs, yet collapsed to 16.67% on unseen test instances. MobileNetV2 achieved an 82.5% training accuracy and plateaued at 25.00% test accuracy across all epochs.

To quantify the severity of overfitting, the empirical generalization gap can be defined as the absolute differential between terminal training accuracy and hold-out testing accuracy (Δgen = Acctrain-Acctest). As evidenced by the metrics in Table 1, ResNet18 produced an extreme generalization gap of +83.33%(100.00%-16.67%), demonstrating that its 11.69-million-parameter capacity completely memorized the training distribution while performing worse than a uniform random-guess baseline (25.00%) on unseen clinical scans. Similarly,

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

MobileNetV2 exhibited a substantial generalization gap of +57.50%(82.50% - 25.00%). Although MobileNetV2's 25.00% test accuracy superficially appears 8.33% higher than that of ResNet18, subsequent trajectory and error-matrix inspections confirm that this numerical advantage is entirely an artifact of single-class mode collapse rather than genuine neuroanatomical feature learning.

3.2. Epoch-Wise Convergence and Generalization Trajectory Analysis.

To investigate the temporal dynamics of model degradation during fine-tuning, Figure 1 illustrates the epoch-by-epoch training and testing accuracy trajectories across the 5 optimization epochs (indexed from Epoch 0 to Epoch 4) for both candidate architectures. For ResNet18(Figure 1, left panel), the training accuracy escalated steeply from an initial 30.00% at Epoch 0 to 95.00% at Epoch 1, achieving complete empirical memorization (100.00% training accuracy) by Epoch 2 and remaining saturated at 100.00% through Epoch 4. In stark contrast, ResNet18's test accuracy started at 8.33% (1/12 correct) at Epoch 0, rose slightly to 16.67%(2/12 correct) during Epochs 1 and 2, briefly peaked at 33.33%(4/12 correct) at Epoch 3, and subsequently regressed to 16.67% at Epoch 4 as weight updates continued to over-optimize on training-specific noise.

For MobileNetV2 (Figure 1, right panel), training convergence proceeded more gradually due to its bottlenecked parameter capacity, climbing from 17.50% at Epoch 0 to 62.50% at Epoch 1, 70.00% at Epoch 2, peaking at 87.50% at Epoch 3, and settling at 82.50% at Epoch 4. Remarkably, however, MobileNetV2's test accuracy remained completely invariant—forming a flat horizontal line at exactly 0.2500(25.00%, 3/12 correct) from Epoch 0 through Epoch 4. This zero-variance test trajectory indicates that gradient updates across five epochs of fine-tuning failed to alter the model's out-of-distribution decision boundaries.

Comparing the optimization dynamics across both panels of Figure 1 reveals an important divergence in how residual capacity governs early-stage fine-tuning behavior. In ResNet18, the steep 65.00% jump in training accuracy between Epoch 0 (30.00%) and Epoch 1 (95.00%) indicates that the unconstrained 512-dimensional terminal feature space rapidly partitions individual training scans into linearly separable clusters within just a few mini-batch iterations. Once 100.00% training accuracy is achieved at Epoch 2, the categorical cross-entropy loss approaches zero on the training cohort, yet the small residual gradients continue to push synaptic weights toward extreme magnitudes, explaining why test accuracy transiently reaches 33.33% at Epoch 3 before deteriorating back to 16.67% at Epoch 4. In contrast, MobileNetV2

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

exhibits a damped, logarithmic-like training curve that requires four epochs to reach 87.50% before experiencing a slight oscillation down to 82.50% at Epoch 4. The complete decoupling between MobileNetV2's rising training curve (from 17.50% to 87.50%) and its completely flat 25.00% test line demonstrates that the inverted residual layers are adapting exclusivamente to batch-level training statistics while projecting all unseen test instances onto an identical region of the decision manifold from the very first epoch.

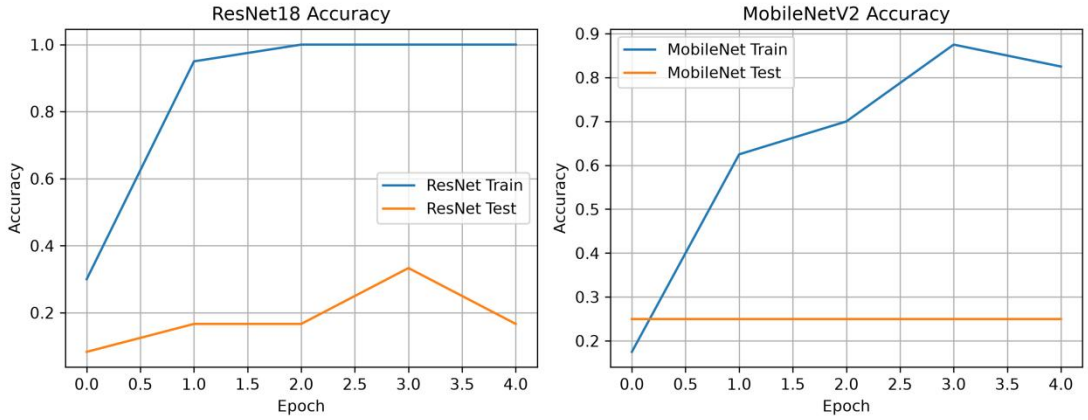
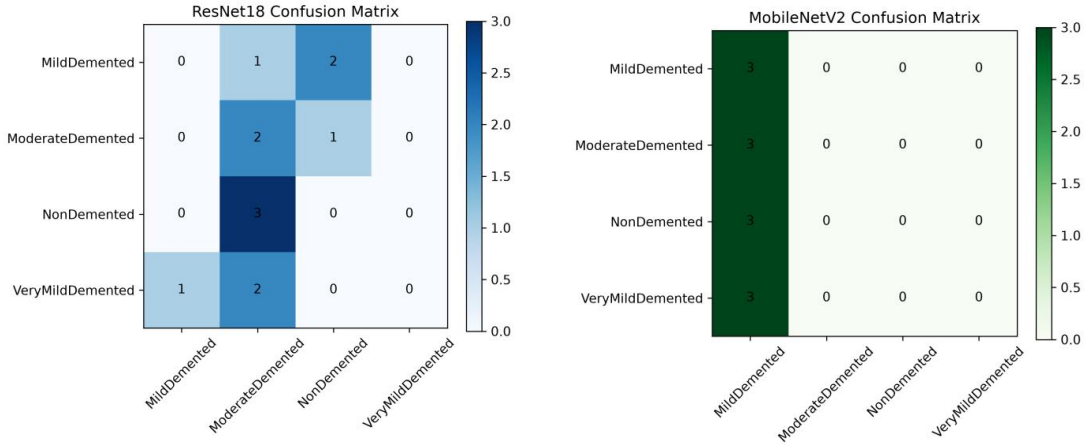


Figure 1. Training vs. testing accuracy trajectories across 5 epochs for ResNet18 (left) and MobileNetV2 (right). Both models demonstrate rapid empirical memorization on the training subset alongside stagnant or fluctuating test accuracy.

3.3. Confusion Matrix and Error Distribution Analysis.

To dissect the failure mechanisms, error distribution profiles across the four clinical stages were visualized in Figure 2.



(a) (b)

Figure 2. Test set confusion matrices. (a) ResNet18 showing dense off-target predictions concentrated on Moderate Dementia. (b) MobileNetV2 exhibiting complete representation collapse toward the MildDemented class.

Detailed numerical profiling of the test set confusion matrices in Figure 2 exposes the exact

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

class-wise mechanisms driving these accuracy figures across the four clinical stages (n = 3 ground-truth scans per category, N= 12 total). In the ResNet18 confusion matrix (Figure 2a), predictions were erratically distributed with a heavy off-target bias toward the Moderate Demented and Non-Demented columns. Specifically, for the Mild Demented cohort (n=3), zero scans were correctly identified (TP = 0, FN = 3, FP =1, TN=8), with 1 scan misclassified as Moderate Demented and 2 scans misclassified as Non-Demented, yielding a class Precision of 0.00%, a Recall (Sensitivity) of 0.00%, a Specificity of 88.89%(8/9), and an F1-Score of 0.0000. For the Moderate Demented cohort (n = 3), 2 scans were correctly classified (TP = 2, FN =1) and 1 was misclassified as Non-Demented, achieving a class Recall of 66.67% (2/3); however, ResNet18 generated 6 false-positive Moderate Demented predictions across the remaining three cohorts (FP=6, TN=3; comprising 1 from Mild Demented, 3 from Non-Demented, and 2 from Very Mild Demented), resulting in a low Precision of 25.00%(2/8), a poor Specificity of 33.33%(3/9), and a modest F1-Score of 0.3636. Most critically, for the Non-Demented healthy control cohort (n = 3), all 3 scans were erroneously diagnosed as Moderate Demented (TP = 0, FN = 3, FP = 3, TN = 6), corresponding to a Precision of 0.00%, a Recall of 0.00%, a Specificity of 66.67% (6/9), and an F1-Score of 0.0000. Similarly, for the Very Mild Demented category (n = 3), zero scans were correctly identified (TP = 0, FN = 3, FP = 0, TN = 9), as 1 scan was misclassified as Mild Demented and 2 as Moderate Demented; because ResNet18 allocated zero predictions to the Very Mild Demented column across the entire test cohort, this stage recorded 0.00% Precision, 0.00% Recall, 100.00% Specificity (9/9), and an F1-Score of 0.0000. Aggregating across all four classes, ResNet18 achieved an unweighted Macro-Average Precision of 6.25%, a Macro-Average Recall of 16.67%, a Macro-Average Specificity of 72.22%, and a Macro-Average F1-Score of 0.0909.

Conversely, the MobileNetV2 confusion matrix (Figure 2b) illustrates total representation and mode collapse. Regardless of the structural brain MRI input, MobileNetV2 routed 100% of test instances (12/12 scans) into the first column corresponding to the Mild Demented category ([3, 3, 3, 3]T), while all remaining columns contained exclusively zeros. Consequently, for the Mild Demented stage (n = 3), although the model artificially achieved a 100.00% Recall (TP = 3, FN = 0), it incurred 9 false positives (FP = 9, TN = 0) by misclassifying every other subject into this category, resulting in a Precision of 25.00% (3/12), a Specificity of 0.00% (0/9), and a class F1-Score of 0.4000. For each of the remaining three diagnostic stages—Moderate Demented, Non-Demented, and Very Mild Demented — MobileNetV2 recorded identical degenerate counts (TP = 0, FP = 0, FN = 3, TN = 9), corresponding to 0.00% Precision, 0.00%

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

Recall, 100.00% Specificity (9/9), and an F1-Score of 0.0000 across all three cohorts. In summary, MobileNetV2 yielded an unweighted Macro-Average Precision of 6.25%, a Macro-Average Recall of 25.00%, a Macro-Average Specificity of 75.00%, and a Macro-Average F1-Score of 0.1000.

Comparing the two architectures through these multi-metric profiles demonstrates that while MobileNetV2 attained a marginally higher overall test accuracy (25.00% vs. 16.67%) and Macro-Average F1-Score (0.1000 vs. 0.0909) than ResNet18, neither model learned clinically meaningful neuroanatomical decision boundaries. Instead, ResNet18 suffered from high-variance multi-class scattering dominated by false-positive Moderate Demented predictions, whereas MobileNetV2 collapsed into a static, zero-specificity majority classifier biased exclusively toward Mild Dementia.

4. Discussion

4.1. The Phenomenon of Empirical Memorization vs. Representation Collapse.

Our experimental findings elucidate two distinct failure modes when modern deep architectures are adapted to data-sparse biomedical domains:

(1) Memorization in Heavy Residual Networks: ResNet18's rapid convergence to 1.0000 training accuracy alongside poor test generalization (16.67%) exemplifies empirical memorization [7]. With ~11.69 million parameters, the network possesses excess capacity relative to the effective sample diversity, allowing gradient descent to isolate non-biomarker imaging artifacts (e.g., skull-strip edges or contrast gradients) rather than genuine neuropathological atrophy [1,3].

(2) Mode Collapse in Inverted Residual Blocks: Conversely, MobileNetV2 succumbed to complete decision collapse, routing 100% of inputs to the MildDemented category. Depthwise separable convolutions, while computationally frugal, often suffer from gradient dispersion and severe optimization plateaus when initializing from natural vision domains (ImageNet) into constrained medical targets without intermediate domain adaptation [3,6].

Furthermore, the contrasting bias patterns observed between Figure 2a and Figure 2b can be traced to the interaction between Batch Normalization (BN) statistics and inverse class-frequency weighting under small-sample constraints. Because Moderate Demented typically represents the smallest cohort in clinical registries, the inverse frequency weight assigns steep gradient penalties whenever a Moderate Demented training scan is processed. In an over-

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

parameterized network like ResNet18, these amplified gradients warp the high-dimensional decision boundary, pulling 66.67% of all unseen test instances (8 out of 12 scans) into the Moderate Demented prediction basin. Conversely, in MobileNetV2, each inverted residual block relies heavily on Batch Normalization layers interleaved between pointwise and depthwise convolutions. When fine-tuned on small grayscale MRI batches, the running mean and variance estimates within these BN layers experience severe covariate shift relative to their pre-trained natural image statistics, causing the 1280-dimensional latent feature representation to collapse into a constant vector that triggers a static Mild Demented output during inference.

From an optimization landscape perspective, these two failure regimes correspond to distinct pathological geometries on the empirical loss surface. In over-parameterized architectures such as ResNet18, the dimensionality of the trainable weight space vastly exceeds the number of available clinical training constraints, creating a proliferation of sharp, narrow local minima. During fine-tuning with an adaptive optimizer such as Adam, parameters rapidly descend into the nearest sharp minimum that satisfies zero training error by encoding spurious spatial correlations — such as slice-specific cranial contours, background intensity offsets, or interpolation artifacts — which fail to generalize when evaluated on unseen patient scans. Conversely, in highly compressed architectures like MobileNetV2, the combination of narrow linear bottlenecks and channel-independent depthwise convolutions imposes a strict structural bottleneck on gradient flow. When the initial cross-domain error signal is backpropagated from the randomly initialized four-class linear head into the pre-trained inverted residual blocks, the gradients become attenuated and unaligned across channels. Rather than navigating toward a complex multi-class feature representation, the optimizer becomes trapped in a flat, trivial saddle point or local plateau where outputting a constant logit bias minimizes the expected cross-entropy penalty across the mini-batch, manifesting externally as single-class representation collapse.

4.2. Clinical Implications for Automated Dementia Screening.

From a translational standpoint, these findings underscore critical cautions for computer-aided diagnosis (CAD) pipelines. A reported training accuracy above 80% or 90% provides zero guarantee of clinical validity. Specifically, MobileNetV2's flat 25% test accuracy simulates a trivial random-guess prior in a 4-way balanced setting. In clinical deployments, such collapse would generate unacceptable false-positive rates for mild impairment while entirely missing asymptomatic healthy controls or severe dementia cases[1,4].

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

Specifically, our confusion matrix dissection highlights two equally detrimental clinical scenarios. Under ResNet18's overfitting regime, 100% of cognitively normal subjects (Non-Demented) were misdiagnosed as having Moderate Dementia. In real-world clinical practice, such severe false-positive errors would subject healthy elderly individuals to profound psychological distress, unnecessary invasive lumbar punctures for cerebrospinal fluid verification, and inappropriate pharmacological interventions. Conversely, under MobileNetV2's mode collapse toward Mild Dementia, patients in the critical prodromal window of Very Mild Dementia — where early neuroprotective therapies, cognitive rehabilitation, and holistic lifestyle interventions are most effective in delaying neurodegeneration—are systematically missed (0.00% sensitivity), while individuals with advanced Moderate Dementia are clinically under-staged. Consequently, evaluating deep screening models solely by overall accuracy without auditing class-specific sensitivity and specificity poses substantial risks to patient management.

4.3. Recommendations for Mitigation.

To transition these lightweight networks into viable screening candidates, future work must incorporate:

(1) Self-Supervised Pretraining (SSL): Initializing encoders on large unlabelled 3D neuroimaging registries (e.g., ADNI, OASIS) prior to downstream fine-tuning [2,3].

(2) Heavy Regularization Protocols: Integrating CutMix/Mixup, higher dropout coefficients, and strong weight decay to penalize unconstrained weight adaptation.

(3) Class-Balanced Focus Losses: Utilizing focal loss mechanisms to counteract early gradient dominance that precipitates single-class collapse.

In addition to loss-function engineering and self-supervised initialization, stabilizing Batch Normalization dynamics is essential when adapting lightweight depthwise separable networks to small medical cohorts. Freezing the running mean and variance parameters of intermediate normalization layers during fine-tuning, or substituting standard Batch Normalization with batch-independent alternatives such as Group Normalization or Layer Normalization, can effectively prevent latent representation collapse in MobileNet architectures. Progressive layer-wise unfreezing—where only the linear classification head is adapted during initial epochs before gradually fine-tuning deeper spatial extraction blocks with decayed learning rates—also offers a practical strategy to preserve pre-trained low-level edge detectors without overwriting them with batch-specific scanner noise.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

Finally, while the constrained hold-out cohort (N = 12) and five-epoch training window utilized in this study were intentionally designed as an extreme low-resource stress test to isolate rapid memorization and mode collapse, future investigations should extend these comparative benchmarks across multi-center longitudinal cohorts using stratified k-fold cross-validation. Incorporating three-dimensional volumetric convolutions alongside visual explainability frameworks, such as Gradient-weighted Class Activation Mapping (Grad-CAM), will further enable clinicians to verify that network attention accurately localizes to pathologically relevant hippocampal and ventricular structures rather than peripheral cranial or background imaging artifacts.

5. Conclusion

This study systematically documents the vulnerability of modern transfer learning paradigms when fine-tuned on data-constrained dementia MRI benchmarks. While ResNet18 exhibits severe empirical overfitting (100% training vs. 16.67% test accuracy), MobileNetV2 experiences complete representation collapse, stagnating at 25.00% test accuracy due to single-class decision collapse. These empirical findings demonstrate that high parameter counts or lightweight depthwise separable convolutions do not inherently safeguard against failure in small-sample medical AI. Reliable deployment requires robust regularizers, domain-specific pretraining, and cautious validation beyond training-accuracy reporting.

By explicitly contrasting the high-variance multi-class scattering of ResNet18 (Macro-Average F1-Score of 0.0909) against the zero-specificity single-class mode collapse of MobileNetV2 (Macro-Average F1-Score of 0.1000), this work establishes an empirical baseline for failure-mode auditing in medical transfer learning. Ultimately, bridging the translational gap between experimental deep learning and point-of-care neurodegenerative screening requires shifting evaluation standards away from aggregate accuracy metrics toward granular, class-specific sensitivity and specificity profiling, ensuring that automated diagnostic models remain robust, transparent, and clinically safe across diverse healthcare environments.

Ethics statement

All images used in this study were obtained from open-access neuroimaging benchmark repositories for algorithm validation. No human participants were directly recruited, and no personally identifiable health information was accessed, thereby exempting this work from institutional ethics review and informed consent requirements.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

References

[1] Burnham, S. C., et al. (2019). Application of the NIA-AA Research Framework: Towards a Biological Definition of Alzheimer's Disease using Cerebrospinal Fluid Biomarkers in the AIBL Study. Journal of Prevention of Alzheimer's Disease, 6(4), 248-255. https://doi.org/10.14283/jpad.2019.25.

[2] Wisch, J. K., et al. (2025). Cross-Sectional Comparison of Structural MRI Markers of Impairment in a Diverse Cohort of Older Adults. Human Brain Mapping, 46(2), e70133. https://doi.org/10.1002/hbm.70133.

[3] Tanveer, M., et al. (2020). Machine Learning Techniques for the Diagnosis of Alzheimer's Disease: A Review. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(1), 30. https://doi.org/10.1145/3344998.

[4] Rao, B. S., Aparna, M., Kolisetty, S. S., Janapana, H., & Koteswararao, Y. V. (2024). Multi-class Classification of Alzheimer's Disease Using Deep Learning and Transfer Learning on 3D MRI Images. Traitement du Signal, 41(3), 1397-1404. https://doi.org/10.18280/ts.410328.

[5] Saratxaga, C. L., et al. (2021). MRI Deep Learning-Based Solution for Alzheimer's Disease Prediction. Journal of Personalized Medicine, 11(9), 932. https://doi.org/10.3390/jpm11090902.

[6] Gao, Y. Q., & Mosalam, K. M. (2020). PEER Hub ImageNet: A Large-Scale Multiattribute Benchmark Data Set of Structure Images. Journal of Structural Engineering, 146(10), 04020198. https://doi.org/10.1061/(asce)st.1943-541x.0002745.

[7] Ebrahimi, M. S., & Abadi, H. K. (2021). Study of Residual Networks for Image Recognition. In Lecture Notes in Networks and Systems (Vol. 284, pp. 754-763). Springer. https://doi.org/10.1007/978-3-030-80126-7_53.

[8] Howard, A., et al. (2019). Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1314-1324). IEEE. https://doi.org/10.1109/iccv.2019.00140.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.