SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Advancing Neurological Disorder Diagnosis: Integrating Explainable and Generalizable Machine Learning Models

Madhuri Waman Patil[1], Dr Prakash G Burade[2]

[1]Research Scholar, Sandip University, Nashik, India, Email: madhuripatil510@gmail.com
[2]Dean, Electrical & Electronic Engineering, Sandip University, Nashik, India, Email: prakash.burade@sandipuniversity.edu.in

ABSTRACT

Alzheimer's disease (AD) and Parkinson's disease (PD) are progressive neurodegenerative disorders that place a growing burden on patients, families and health systems around the world. MRI gives a non-invasive way to look at the structural brain changes tied to these conditions, but reading scans by eye is slow, subjective and hard to scale up. This paper brings together three stages of a doctoral research programme, each built on the last, aimed at an MRI-based diagnostic framework that is accurate, efficient and explainable. The first stage ran a baseline comparison, pitting a convolutional neural network (CNN) against four classical classifiers – Support Vector Machine, Random Forest, k-Nearest Neighbours and Logistic Regression – trained on PCA-reduced features, for four-class Alzheimer's staging. The CNN came out ahead of every classical model, reaching 69% accuracy against Random Forest's best classical result of 65.7%, which suggests that features learned automatically from the images carry more diagnostic signal than handcrafted, dimensionality-reduced ones. The second stage moved away from training a CNN from scratch and toward optimized transfer learning. Pre-trained MobileNetV2 and ResNet50 backbones were adapted to the Alzheimer's Multiclass (Equal and Augmented) dataset using a two-stage freeze/fine-tune schedule, Adam optimization, data augmentation, dropout, early stopping and checkpointing. This optimization-focused approach pushed accuracy to 75% for MobileNetV2 and 67% for ResNet50, and pointed to a broader pattern: a lighter architecture, trained with a carefully designed schedule, can generalize better than a deeper one when medical-imaging data is limited. The third stage extended the framework in two directions at once – generalizing across diseases by adding an independent Parkinson's disease MRI dataset, and adding interpretability through Grad-CAM and Grad-CAM++ visualizations layered onto the MobileNetV2/ResNet50 backbones. MobileNetV2 again came out on top, reaching 86.17% accuracy on the four-class Alzheimer's dataset and 96.39% on the two-class Parkinson's dataset, while ResNet50 struggled with unstable convergence and a bias toward the majority class on both tasks. Grad-CAM++ gave sharper, more clinically localized activation maps than standard Grad-CAM, and the model's attention lined up with cortical and ventricular regions known to be involved in neurodegeneration. Together, the three stages trace a single line of argument: deep learning beats classical baselines, principled optimization of transfer learning lifts accuracy substantially further, and a lightweight, optimized backbone can end up both generalizable across diseases and explainable to clinicians. The combined results point to MobileNetV2 with Grad-CAM++ as a workable basis for computer-aided, interpretable diagnosis of neurological disease, and set up future work on more advanced architectures, additional XAI methods, and fusing MRI with PET, EEG and clinical data.

Keywords: Alzheimer's disease; Parkinson's disease; deep learning; transfer learning; MobileNetV2; ResNet50; explainable AI; Grad-CAM; Grad-CAM++; MRI classification.

INTRODUCTION

1.1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that slowly wears away memory, thinking ability and everyday functioning. The World Health Organization puts the number of people currently living with dementia worldwide at more than 55 million, with AD accounting for roughly 60–70% of these cases; as the global population ages, this burden is expected to grow substantially. Parkinson's disease (PD) is the second most common neurodegenerative disorder, and it mainly affects motor control as dopamine-producing neurons progressively die off [1]. Both diseases share a diagnostic problem: their structural signatures on MRI – hippocampal atrophy, ventricular enlargement and cortical thinning in AD, subtler basal-ganglia and midbrain changes in PD – tend to develop gradually and are easy to miss

on manual review [2].

MRI is now the standard non-invasive tool for studying these structural changes, but reading scans by hand, and even conventional machine-learning pipelines built on handcrafted features, tends to be slow, subjective and prone to missing subtle biomarkers scattered across the brain. Deep learning, and convolutional neural networks (CNNs) in particular, has changed medical image analysis by letting models learn hierarchical, task-specific features directly from pixel data instead of relying on manual feature engineering [3].

1.2 Motivation

The motivation behind this research programme is the clinical need for diagnostic tools that are accurate, computationally efficient and trustworthy all at once. Deep-learning research driven mainly by accuracy has tended to treat model architecture as the main lever for improvement, while giving comparatively little attention to (i) systematically optimizing the training process itself, (ii) whether a single framework generalizes across more than one neurological disorder, and (iii) the interpretability clinicians need before they will trust an automated prediction. This paper addresses those three gaps in sequence: first checking whether deep learning is worth pursuing at all relative to classical baselines, then optimizing the training pipeline, and finally making the resulting system both generalizable and explainable. That sequence is the narrative arc running through the three research stages consolidated here.

1.3 Objectives

The research objectives shifted somewhat across the three stages; taken together, they are:

- Develop and benchmark a CNN-based model against traditional machine-learning classifiers (SVM, Random Forest, k-NN, Logistic Regression) for multi-class Alzheimer's MRI classification.
- Design an optimized transfer-learning framework using pre-trained MobileNetV2 and ResNet50 architectures, incorporating adaptive optimization, learning-rate scheduling, fine-tuning and regularization to improve convergence and accuracy.
- Evaluate whether a lightweight architecture (MobileNetV2) can outperform a deeper architecture (ResNet50) on limited medical-imaging datasets.
- Extend the optimized framework to a second, independent disease (Parkinson's) to test cross-disease generalization.
- Integrate Explainable AI techniques (Grad-CAM and Grad-CAM++) to visualize and compare the brain regions driving model predictions.
- Comprehensively evaluate all models using accuracy, precision, recall, F1-score, class-wise classification reports and confusion-matrix analysis.
- Identify future directions, including advanced architectures, additional XAI methods and multimodal data fusion, for clinical decision-support deployment.

1.4 Scope of Work

This consolidated research covers supervised, image-based classification of Alzheimer’s and Parkinson’s disease using publicly available MRI datasets. It spans image preprocessing, dimensionality reduction for the classical baselines, CNN and transfer-learning model development, systematic training optimization, cross-disease evaluation and post-hoc explainability analysis. More advanced architectures (e.g., Vision Transformers, EfficientNet), multimodal fusion with PET/EEG/clinical data, and prospective clinical validation are left for future work rather than tackled in the present stages.

2. Literature Review

Deep learning has increasingly been applied to MRI-based diagnosis of Alzheimer’s disease, and CNN-based frameworks have consistently outperformed traditional machine-learning pipelines built on handcrafted or dimensionality-reduced features [4]. Islam and Zhang [13] proposed an early CNN framework for staging Alzheimer’s disease from MRI and reported improved accuracy over classical baselines, while Basaia et al. [14] showed that deep neural networks can automatically classify Alzheimer’s disease and mild cognitive impairment from structural MRI. Jo et al. [15] carried out a systematic review of deep learning in neuroimaging and found that CNNs and autoencoders exceeded 90% accuracy on several multimodal datasets, underscoring the diagnostic potential of the approach.

Transfer learning has proven especially useful when labelled medical-imaging data is scarce. He et al. [6] introduced the ResNet architecture, whose residual connections make it easier to train very deep networks, and Sandler et al. [5] proposed MobileNetV2, a lightweight architecture built on depthwise-separable convolutions and inverted residual blocks that trades some representational depth for computational efficiency. AlSaeed et al. [16] applied a pre-trained

ResNet50 to Alzheimer's diagnosis and found that transfer learning cuts training time while keeping accuracy up, and Raza et al. [17] proposed a DenseNet-based transfer-learning framework centred on gray-matter segmentation.

Other work has focused on architectural refinements and attention mechanisms: Li et al. [18] proposed an enhanced residual attention network, and Khosravi et al. [19] introduced a cascade-attention CNN, both meant to focus feature extraction on diagnostically relevant brain regions. Podolszańska [20] combined CNN architectures with dimensionality-reduction techniques for MRI-based monitoring, and Shahid et al. [21] explored hybrid pipelines that pair CNN feature extraction with classical classifiers such as random forests and SVMs. Mousavi et al. [22] and Sorour et al. [23] further showed that architectures such as InceptionResNet and Xception can improve diagnostic precision over conventional machine learning.

The optimization process itself has received comparatively less attention. Kingma and Ba [12] introduced the Adam optimizer, which combines momentum with adaptive per-parameter learning rates and has become something of a default for training deep networks; Mohammed et al. [24] discussed adaptive optimizers such as Adam and RMSProp specifically for Alzheimer's detection, and Moni et al. [25] noted that telling apart adjacent disease stages (e.g., very mild versus mild dementia) remains difficult because their neuroimaging features overlap. Goodfellow et al. [28] and Ruder [32] provide the foundational treatments of deep-learning training dynamics and gradient-based optimization that underpin the fine-tuning strategies used here, while Shorten and Khoshgoftaar [29] survey the image-augmentation techniques used to fight overfitting on small medical datasets.

Beyond accuracy, the black-box nature of CNNs is increasingly seen as a barrier to clinical adoption. Gunning and Aha [7] make the case for Explainable AI (XAI) as a prerequisite for trust in automated decision systems. Selvaraju et al. [8] proposed Grad-CAM, which uses the gradients flowing into the final convolutional layer to produce a coarse localization heatmap of the regions that most influence a prediction, and Chattopadhyay et al. [9] proposed Grad-CAM++, which uses higher-order gradient information to produce sharper, more spatially precise heatmaps – particularly useful for catching subtle, early-stage pathology. More recent work has looked at transformer-based and hybrid explainable architectures (e.g., video vision transformers [36], hybrid compact convolutional transformers [37], and Vision Transformer/Bi-LSTM hybrids [38]) specifically for Alzheimer's MRI classification, part of a broader shift toward models that are both accurate and interpretable.

Overall, the literature shows steady progress on three largely separate fronts – classification accuracy, training optimization, and interpretability – but relatively few studies tackle all three at once, and fewer still test whether a single optimized, explainable framework generalizes across more than one neurodegenerative disease. That gap is what motivates the three-stage research programme brought together in this paper.

3. Research Gaps

Building on the literature review, the following gaps were identified and addressed, one stage at a time:

1. Limited comparative benchmarking. Many studies report deep-learning results in isolation, without a systematic, like-for-like comparison against classical machine-learning baselines on the same data and evaluation protocol.
2. Under-explored training optimization. Most prior work focuses on architectural novelty rather than systematically analysing the effect of optimizer choice, learning-rate scheduling, layer-wise fine-tuning and regularization on convergence and accuracy.
3. Insufficient exploration of lightweight architectures. Despite their efficiency advantages, architectures such as MobileNetV2 are less frequently benchmarked against deeper networks such as ResNet50 for medical imaging, particularly under identical optimization settings.
4. Single-disease evaluation. The large majority of studies validate a proposed framework on a single disease/dataset, leaving open the question of whether the same architecture and training recipe generalizes to a different neurodegenerative disorder.
5. Limited interpretability. The black-box nature of CNN-based diagnostic systems restricts clinical trust; where XAI is applied, it is rarely compared systematically against alternative visualization methods (e.g., Grad-CAM versus Grad-CAM++).
6. Lack of multimodal integration. Existing frameworks predominantly rely on a single imaging modality (MRI), whereas clinical diagnosis increasingly draws on PET, EEG and structured clinical data.

4. Research Contributions

This consolidated study makes the following contributions across the three research stages:

1. A systematic baseline comparison between a custom CNN and four classical machine-learning classifiers (SVM, Random Forest, k-NN, Logistic Regression) for four-class Alzheimer's MRI classification, establishing the performance ceiling of handcrafted-feature approaches on the chosen dataset.
2. An optimized transfer-learning framework built on MobileNetV2 and ResNet50, incorporating a two-stage freeze/fine-tune training schedule, Adam optimization with learning-rate reduction, dropout regularization, early stopping and model checkpointing.
3. Empirical evidence that a lightweight architecture (MobileNetV2) consistently and substantially outperforms a deeper architecture (ResNet50) on limited MRI datasets, both for Alzheimer's and Parkinson's disease.
4. Extension of the optimized framework to a second, independent neurodegenerative disease (Parkinson's disease), demonstrating cross-disease generalizability of the same architecture and training recipe.
5. Integration and comparative evaluation of Grad-CAM and Grad-CAM++ explainability techniques, showing that Grad-CAM++ produces sharper, more clinically localized activation maps than conventional Grad-CAM.
6. Comprehensive, multi-metric performance evaluation (accuracy, precision, recall, F1-score, confusion matrices and class-wise classification reports) across all three stages, enabling a coherent before/after view of the effect of optimization and architecture choice.
7. Identification of concrete future research directions – advanced architectures, additional XAI techniques, and multimodal MRI/PET/EEG/clinical data fusion – for translating the framework toward clinical decision support.

5. Problem Formulation

Alzheimer's disease and Parkinson's disease are progressive neurodegenerative disorders that cause significant cognitive and motor impairment respectively. Early, accurate diagnosis matters for appropriate clinical management, and brain MRI provides useful structural information for spotting disease-related abnormalities; but reading MRI scans by hand is time-consuming and can be unreliable, especially when the changes involved are subtle.

Conventional CNN-based diagnostic systems have two notable limitations. First, their black-box nature makes it hard to tell which image regions drove a given prediction, which limits clinical trust. Second, different transfer-learning architectures can perform quite differently on limited medical-imaging datasets, and this variability is rarely examined systematically alongside training-optimization choices.

The problem this research addresses, then, is building an accurate, computationally efficient and explainable deep-learning framework for classifying Alzheimer's and Parkinson's disease from brain MRI. The approach uses transfer learning with MobileNetV2 and ResNet50 and adds Grad-CAM and Grad-CAM++ to provide visual explanations of the model's predictions. The Alzheimer's dataset has four classes – NonDemented, VeryMildDemented, MildDemented and ModerateDemented – while the Parkinson's dataset has two, Normal and Parkinson; in both cases images are resized, normalized, augmented and split into training and validation sets.

Specifically, the research addresses the following sub-problems:

- Automated classification: developing a deep-learning-based approach for automated classification of AD and PD from MRI images.
- Model efficiency: determining whether a lightweight architecture such as MobileNetV2 can outperform a deeper architecture such as ResNet50 on limited MRI datasets.
- Model interpretability: addressing the black-box nature of CNNs by identifying the MRI regions that influence disease predictions.
- XAI comparison: comparing Grad-CAM and Grad-CAM++ in terms of localization quality and visual interpretability.
- Generalization across diseases: evaluating the transfer-learning framework on both Alzheimer's and Parkinson's disease datasets.
- Clinical decision support: developing an interpretable, computer-aided diagnostic approach that can potentially support clinicians in neurological disease assessment.

6. Datasets

6.1 Alzheimer's Disease MRI Datasets

Two related, publicly available Kaggle datasets were used across the three stages. The first stage used the Alzheimer's Classification Dataset [10], with brain MRI images pre-sorted into training, validation and testing folders across four classes – Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented. Images were resized to 64×64

pixels, normalized to [0, 1], and capped at 1,500 samples per class (6,000 images total) to avoid class imbalance, as shown in Table 1.

Table 1. Alzheimer's Classification Dataset Composition (Stage I)

Class Label	Description	Samples Used
Non-Demented	Healthy control subjects	1,500
Very Mild Demented	Early signs of Alzheimer's disease	1,500
Mild Demented	Moderate impairment due to Alzheimer's	1,500
Moderate Demented	Advanced stage of Alzheimer's disease	1,500
Total		6,000

Stages II and III moved to the larger Alzheimer's Multiclass Dataset (Equal and Augmented), which has the same four disease categories organized into class-labelled folders, split 80% training / 20% validation. Images in this dataset were resized to 224×224 pixels to match what the pre-trained MobileNetV2 and ResNet50 backbones expect.

6.2 Parkinson's Disease MRI Dataset

Stage III added a publicly available Parkinson's Brain MRI Dataset [11], with two classes – Normal and Parkinson – to check whether the optimized transfer-learning framework holds up beyond Alzheimer's disease. As with the Alzheimer's data used in Stages II and III, images were resized to 224×224 pixels, normalized, augmented and split into training and validation sets.

6.3 Data Preprocessing and Augmentation

A consistent preprocessing pipeline was used across all three stages, adjusted for the resolution each model family expects:

- Image resizing to a fixed resolution (64×64 pixels for the Stage I CNN/classical baselines; 224×224 pixels for the Stage II–III transfer-learning models).
- Pixel-intensity normalization to the [0, 1] range via rescaling (1/255) to stabilize and speed up training.
- Label encoding – one-hot encoding for CNN/transfer-learning models, integer encoding for classical ML models.
- Stratified train/validation/test splitting to preserve class proportions (80/20 in Stages II–III; 70/30 or 80/20 with stratified sampling in Stage I).
- Data augmentation – rotation, width/height shifting, zoom, shear, horizontal flipping and brightness adjustment – applied during training in Stages II and III to increase data diversity and reduce overfitting.
- For classical ML models in Stage I, images were flattened into 1D vectors, standardized, and reduced to 100 dimensions using Principal Component Analysis (PCA).

7. Proposed Methodology

The research programme ran in three progressive stages, each building on what the last one found: (I) a baseline comparison of deep versus classical learning, (II) an optimized transfer-learning framework, and (III) an explainable, cross-disease framework. Each stage's methodology is described below.

7.1 Stage I – Baseline CNN versus Traditional Machine Learning

The first stage took a two-pronged approach to check whether deep learning offers a real advantage over classical machine learning for MRI-based Alzheimer's classification. A custom CNN was built to learn spatial features directly from raw pixel data, while four traditional classifiers were trained on PCA-reduced feature vectors as a point of comparison.

7.1.1 CNN Architecture

The CNN has two convolutional layers (32 and 64 filters respectively), each followed by max-pooling to shrink spatial dimensions while keeping the salient features, with ReLU activation adding non-linearity. The resulting feature maps are flattened and passed through a dense layer of 128 neurons, then a dropout layer (50% rate) to curb overfitting, and finally a softmax layer for four-class classification. The network was compiled with the Adam optimizer and categorical cross-entropy loss, and trained for 20 epochs with a batch size of 32, holding back 10% of the training data for validation.

7.1.2 Classical Machine-Learning Baselines

Four classical classifiers were trained on the PCA-reduced (100-component) feature representation, using an 80/20 train/test split:

- Support Vector Machine (SVM) – linear kernel, suitable for high-dimensional, multi-class problems.
- Random Forest (RF) – an ensemble of 100 decision trees trained on bootstrap samples to reduce overfitting.
- k-Nearest Neighbours (KNN) – instance-based classification using the five nearest neighbours.
- Logistic Regression (LR) – multinomial logistic regression with L2 regularization as a linear, interpretable baseline.

7.2 Stage II – Optimized Transfer-Learning Framework

Building on Stage I's finding that deep learning beats classical baselines, Stage II shifted the focus from architecture alone to systematically optimizing the training process, using pre-trained convolutional backbones instead of a CNN trained from scratch.

7.2.1 Transfer-Learning Backbones

Two ImageNet-pretrained architectures were used: MobileNetV2 [5], chosen for its lightweight, depthwise-separable-convolution design and low computational cost, and ResNet50 [6], a deeper residual-learning architecture used as a point of comparison. In both cases the original classification head was removed and replaced with a Global Average Pooling layer, a dense layer with ReLU activation, a dropout layer, and a softmax output sized to the four Alzheimer's classes.

7.2.2 Two-Stage Training and Fine-Tuning

Training followed a two-stage schedule. In the first stage, the pre-trained convolutional base was frozen and only the newly added classification layers were trained, using the Adam optimizer with an initial learning rate of 0.001, so the general-purpose ImageNet features would be preserved while the classifier head adapted to MRI data [12]. In the second stage, selected layers of the pre-trained network were unfrozen and fine-tuned at a much lower learning rate, 1e-5, letting the network adjust its deeper features to the structural patterns specific to Alzheimer's MRI without forgetting what it had already learned.

7.2.3 Regularization and Training Optimization

The following techniques were used to improve generalization and stabilize training:

- Dropout regularization within the classification head.
- Data augmentation (rotation, shifting, zoom, shear, horizontal flip) applied during training.
- Early stopping to halt training once validation performance plateaued.
- Model checkpointing to retain the best-performing weights.

Both models used categorical cross-entropy loss, which fits the four-class classification task.

7.3 Stage III – Explainable, Cross-Disease Framework

The third stage pushed the Stage II framework in two directions at once: generalizing to a second disease (Parkinson's) with the same architecture and training recipe, and adding interpretability through Grad-CAM and Grad-CAM++ explainable-AI visualizations.

7.3.1 Overall Framework

The Stage III framework consists of the following pipeline stages:

1. Dataset acquisition (Alzheimer's and Parkinson's MRI datasets).
2. MRI image preprocessing (resizing to 224×224, normalization, splitting, augmentation).
3. Transfer learning using pretrained MobileNetV2 and ResNet50 backbones.
4. Fine-tuning of unfrozen layers of the CNN architectures.
5. Disease classification (softmax output for each disease's class set).
6. Explainability visualization using Grad-CAM and Grad-CAM++.
7. Performance evaluation using accuracy, precision, recall, F1-score and confusion matrices.

7.3.2 Explainable AI: Grad-CAM and Grad-CAM++

Grad-CAM (Gradient-weighted Class Activation Mapping) [8] produces a localization heatmap by taking the gradient of the target class score with respect to the feature maps in the final convolutional layer, then combining these gradients into per-channel importance weights. The Grad-CAM formulation is given by:

$$L_{Grad\text{-}CAM}^{c} = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (1)$$

where A^k represents feature maps, and α_k^c represents importance weights.

Grad-CAM++ [9] builds on this by using higher-order gradient information to compute more precise, pixel-wise importance weights, which produces sharper and more spatially focused activation maps. The Grad-CAM++ formulation is:

$$L_{Grad\text{-}CAM++}^{c} = \sum_k w_k^c A^k \quad (2)$$

Grad-CAM++ uses higher-order gradient information to assign more precise importance weights to feature maps. In the resulting visualizations, red and yellow mark areas of high importance to the model's prediction, green marks moderate importance, and blue marks low importance. Grad-CAM tends to produce coarser localization that can spill beyond the exact pathological area, while Grad-CAM++ produces cleaner, more tightly focused maps around clinically relevant structures such as cortical and ventricular regions, with less background noise – which makes it especially useful for subtle, early-stage disease patterns.

7.3.3 Evaluation Protocol

All three stages used the same multi-metric evaluation protocol: accuracy, precision, recall and F1-score (weighted where applicable), supplemented in Stages II and III by class-wise classification reports and confusion-matrix analysis to surface per-class strengths and weaknesses that overall accuracy can hide.

8. Experimental Results

8.1 Stage I – CNN versus Classical Machine Learning

Table 2 reports how the CNN and the four classical classifiers performed on the Alzheimer's Classification Dataset.

Table 2. Performance Comparison – Stage I (Alzheimer's, 4-Class)

Model	Accuracy	Precision	Recall	F1-score
CNN	0.690	0.681	0.690	0.682
Random Forest	0.657	0.652	0.657	0.651
Logistic Regression	0.595	0.588	0.595	0.590
KNN	0.543	0.536	0.543	0.537
SVM (linear)	0.364	0.363	0.364	0.354

The CNN reached the highest accuracy (69%) with a balanced precision-recall profile, which confirms that features learned automatically from raw MRI pixels carry more discriminative signal than PCA-reduced handcrafted ones. Random Forest was the strongest classical baseline (65.7%), while linear SVM performed worst (36.4%), suggesting the PCA-reduced feature space isn't linearly separable for this task. These results made the case for moving to deep learning – specifically pre-trained transfer-learning backbones – in the next stage.

8.2 Stage II – Optimized Transfer Learning (Alzheimer's)

Table 3 summarizes the overall accuracy of the optimized MobileNetV2 and ResNet50 models on the Alzheimer's Multiclass (Equal and Augmented) dataset.

Table 3. Overall Accuracy – Stage II (Alzheimer's, 4-Class)

Model	Accuracy
MobileNetV2	75%
ResNet50	67%

Table 4 gives the corresponding class-wise precision, recall and F1-score for both models.

Table 4. Class-wise Performance Comparison – Stage II

Class	MobileNetV2 P	MobileNetV2 R	MobileNetV2 F1	ResNet50 P	ResNet50 R	ResNet50 F1
MildDemented	0.82	0.63	0.71	0.54	0.69	0.60
ModerateDemented	0.99	1.00	0.99	0.95	0.99	0.97
NonDemented	0.72	0.80	0.76	0.67	0.70	0.69
VeryMildDemented	0.55	0.58	0.56	0.49	0.32	0.39

Optimization – the two-stage freeze/fine-tune schedule, Adam optimization with learning-rate reduction, dropout, augmentation, early stopping and checkpointing – lifted MobileNetV2's accuracy from the Stage I CNN baseline of 69% to 75%, while ResNet50 reached 67%, roughly on par with the Stage I CNN but behind MobileNetV2. Both models classified ModerateDemented cases almost perfectly, but both struggled with VeryMildDemented, which reflects how subtle early-stage structural changes are – a limitation that carried into Stage III and points to continued work on early-stage detection.

8.3 Stage III – Explainable, Cross-Disease Framework

8.3.1 Alzheimer's Disease Classification

Table 5. Performance Comparison – Stage III (Alzheimer's, 4-Class)

Model	Accuracy	Precision	Recall	F1-score
MobileNetV2	86.17%	86.84%	86.17%	86.25%
ResNet50	41.10%	28.33%	41.10%	31.68%

Table 6. Class-wise Classification Report – MobileNetV2 (Alzheimer's)

Class	Precision	Recall	F1-score	Support
MildDemented	0.89	0.90	0.90	2000
ModerateDemented	0.99	1.00	1.00	2000
NonDemented	0.89	0.76	0.82	2560
VeryMildDemented	0.71	0.83	0.77	2240

ResNet50 showed unstable convergence and poor class discrimination in this stage, including zero precision and recall for the MildDemented class – a sign of severe majority-class bias rather than genuine feature learning under this training configuration. MobileNetV2, by contrast, performed strongly and fairly evenly across all four classes, with near-perfect results on ModerateDemented and improved, though still comparatively weaker, performance on VeryMildDemented.

8.3.2 Parkinson's Disease Classification

Table 7. Performance Comparison – Stage III (Parkinson's, 2-Class)

Model	Accuracy	Precision	Recall	F1-score
MobileNetV2	96.39%	96.55%	96.39%	96.30%
ResNet50	73.49%	54.01%	73.49%	62.27%

Table 8. Class-wise Classification Report – MobileNetV2 (Parkinson's)

Class	Precision	Recall	F1-score	Support
Normal	0.95	1.00	0.98	122
Parkinson	1.00	0.86	0.93	44

On the Parkinson's dataset, MobileNetV2 reached 96.39% accuracy with a precision of 1.00 for the Parkinson class, meaning no healthy subjects were misclassified as Parkinsonian. ResNet50 again showed majority-class bias, with markedly lower precision and recall for the Parkinson class, correctly identifying only a minority of Parkinsonian cases. This pattern holding up across two independent diseases and class-count settings (four-class and two-class) supports the conclusion that, under the optimization regime used here, the lightweight MobileNetV2 backbone generalizes more reliably than the deeper ResNet50 backbone on limited medical-imaging datasets.

8.3.3 Explainability Analysis

Grad-CAM and Grad-CAM++ were used to visualize the brain regions driving MobileNetV2's and ResNet50's predictions for both diseases. Grad-CAM highlighted cortical and ventricular regions consistent with known Alzheimer's pathology, but its localization was relatively coarse and sometimes extended beyond the precise pathological area. Grad-CAM++ produced cleaner, more tightly focused activation maps around the same clinically relevant structures – particularly cortical thinning and ventricular enlargement – with less background noise, and was especially more informative for subtle, early-stage dementia patterns. Table 9 summarizes the qualitative comparison between the two XAI techniques.

Fig. 1–Fig. 3 show representative Grad-CAM and Grad-CAM++ visualizations for the MobileNetV2 model, and Fig. 4–Fig. 6 show the corresponding visualizations for ResNet50, on Alzheimer's MRI slices. In each figure, the leftmost panel is the original MRI slice, the centre panel is the Grad-CAM heatmap, and the rightmost panel is the Grad-CAM++ heatmap overlaid on the same slice.

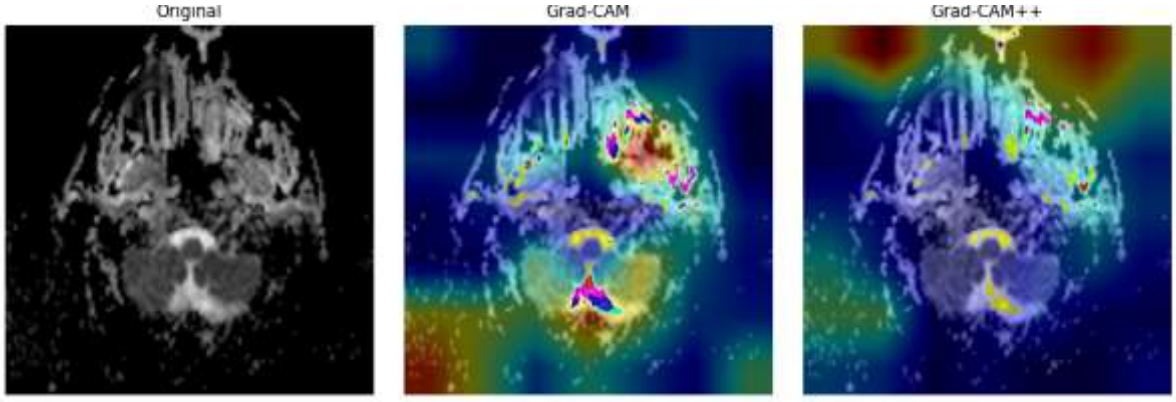

Fig. 1. Grad-CAM and Grad-CAM++ visualization for a representative Alzheimer's MRI slice using MobileNetV2 (original image, Grad-CAM heatmap, and Grad-CAM++ heatmap, left to right).

Fig. 2. Grad-CAM and Grad-CAM++ visualization for a second representative Alzheimer's MRI slice using MobileNetV2.

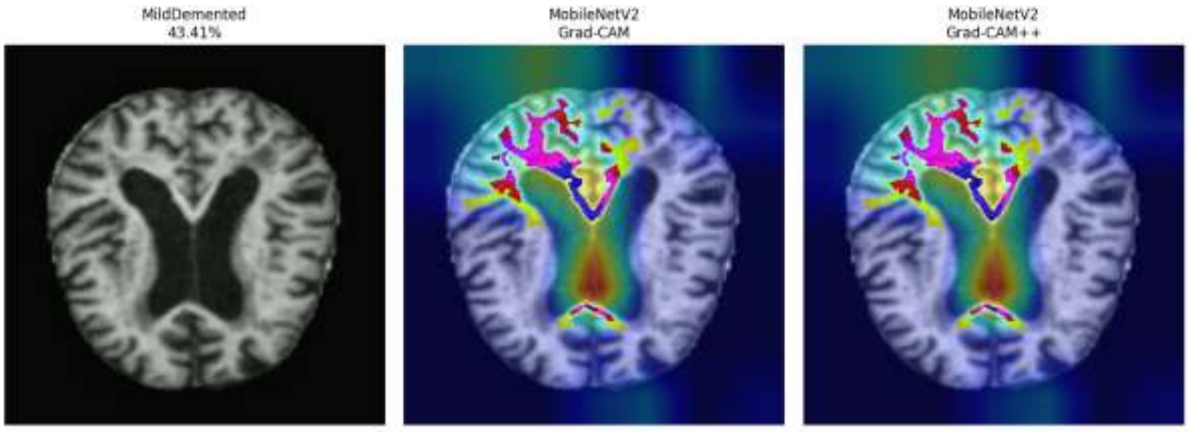

Fig. 3. Grad-CAM and Grad-CAM++ visualization for an MRI slice classified as Mild Demented (43.41% confidence) using MobileNetV2.

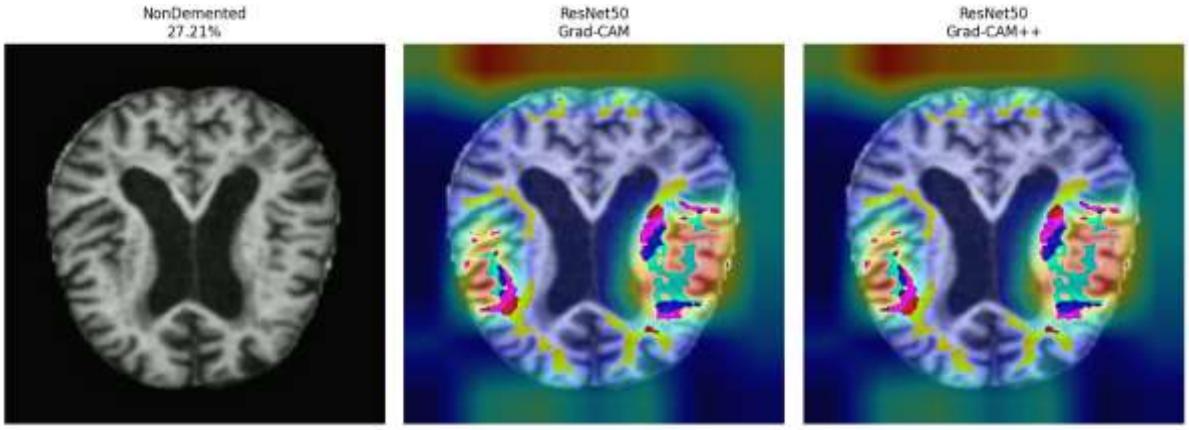

Fig. 4. Grad-CAM and Grad-CAM++ visualization for an MRI slice classified as Non-Demented (27.21% confidence) using ResNet50.

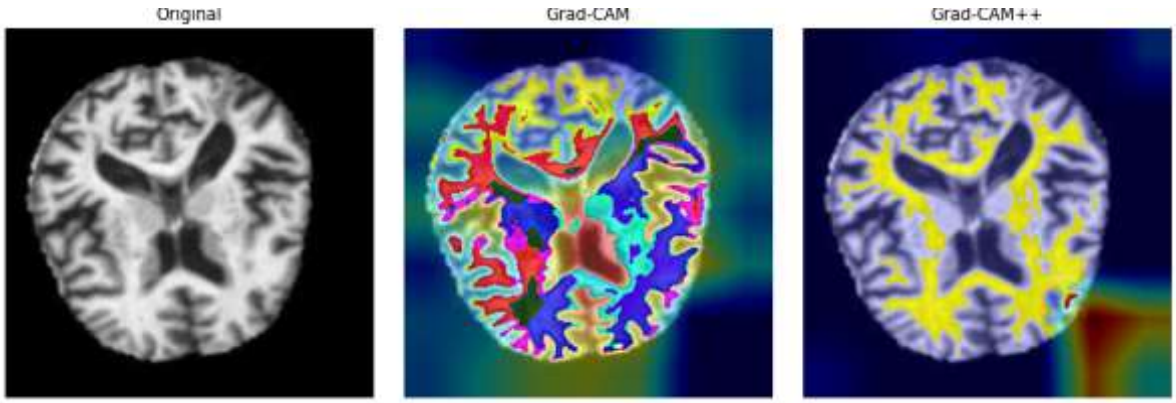

Fig. 5. Grad-CAM and Grad-CAM++ visualization for a representative Alzheimer's MRI slice using ResNet50.

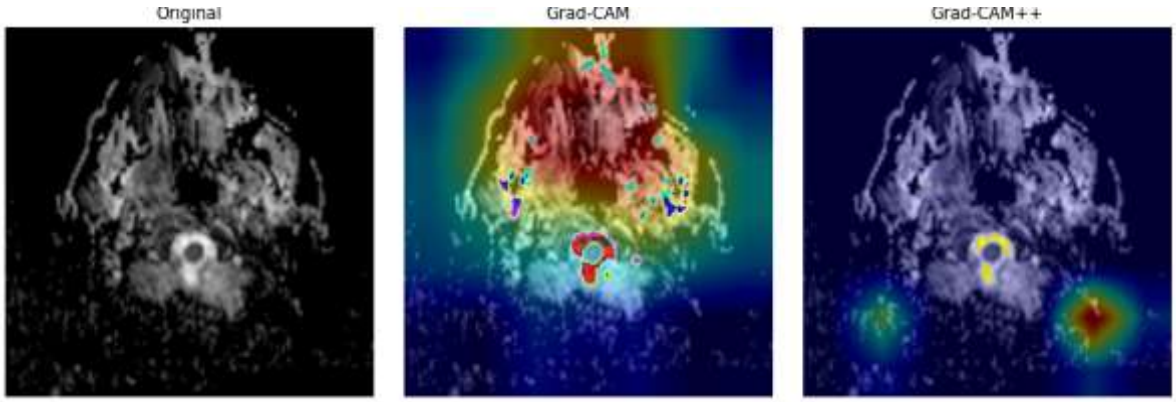

Fig. 6. Grad-CAM and Grad-CAM++ visualization for a further representative Alzheimer's MRI slice using ResNet50.

Table 9. Comparison Between Grad-CAM and Grad-CAM++

Feature	Grad-CAM	Grad-CAM++
Localization quality	Coarse	More precise
Heatmap sharpness	Moderate	High
Noise reduction	Lower	Better
Interpretability	Good	Excellent
Medical-imaging suitability	Moderate	High

9. Discussion

Taken together, the three stages tell one continuous story. Stage I showed that a modestly sized CNN, trained end-to-end on raw MRI pixels, already beats four classical machine-learning classifiers working on PCA-reduced features (69% versus a best classical result of 65.7%), which justified sticking with deep-learning approaches. Stage II showed that accuracy from a comparable architecture family can be pushed substantially further – from 69% to 75% for MobileNetV2 – not through architectural novelty but through systematic training optimization: transfer learning from ImageNet, a two-stage freeze/fine-tune schedule, adaptive optimization with learning-rate reduction, and a set of regularization techniques (dropout, augmentation, early stopping, checkpointing).

Stage III then pushed the framework further along two independent dimensions at once. Cross-disease evaluation on an independent Parkinson's dataset showed that the same architecture and training recipe generalizes well when paired with MobileNetV2 (96.39% accuracy) but drops off sharply with ResNet50 (73.49%, with notably weaker performance on the minority Parkinson class). The same pattern shows up on the Alzheimer's dataset in this stage (86.17% for MobileNetV2 versus 41.10% for ResNet50, with ResNet50 collapsing entirely on the MildDemented class), and it is a striking, consistent finding: under the optimization settings used here, model depth on its own does not translate into better generalization on limited medical-imaging datasets, and can actually work against it – likely some combination of overfitting risk, a harder optimization landscape for very deep residual networks [6], and fine-tuning datasets that are small relative to ResNet50's parameter count.

Adding Grad-CAM and Grad-CAM++ addresses the interpretability gap raised in the literature review. The consistent finding that Grad-CAM++ gives sharper, more anatomically focused heatmaps than Grad-CAM – without any change to the underlying classifier – shows that interpretability gains can, to a meaningful degree, be pursued separately from accuracy gains, as their own axis of framework design. One challenge that persisted across all three stages, though, is reliably identifying early-stage disease (VeryMildDemented in the Alzheimer's task): even the best-performing Stage III MobileNetV2 model scores lowest on this category, which reflects how genuinely subtle and overlapping early neurodegeneration's structural signatures are, rather than a shortcoming specific to any one architecture or optimization choice.

10. Conclusion and Future Work

This paper has brought together three successive stages of doctoral research into one coherent account of progress toward an accurate, efficient and explainable framework for MRI-based neurological disease diagnosis. Starting from a baseline comparison that established deep learning's advantage over classical machine learning (CNN: 69% accuracy versus a best classical result of 65.7% for Random Forest), the research moved to an optimized transfer-learning framework that raised accuracy to 75% through principled training-schedule design rather than architectural change, and ended with an explainable, cross-disease framework that reached 86.17% accuracy on Alzheimer's disease and 96.39% on Parkinson's disease using an optimized MobileNetV2 backbone – while consistently showing that the deeper ResNet50 backbone is more prone to unstable convergence and majority-class bias at this scale of data.

Adding Grad-CAM and Grad-CAM++ showed that these models can be made interpretable without giving up accuracy, with Grad-CAM++ providing sharper, more clinically meaningful localization of the brain regions behind each prediction. Taken as a whole, the findings support a lightweight, optimized and explainable transfer-learning framework – MobileNetV2 paired with Grad-CAM++ – as a practical, generalizable foundation for computer-aided neurological disease diagnosis, and as a promising building block for future clinical decision-support tools.

Building on these three stages, future work will pursue the following directions:

- Advanced architectures: evaluating EfficientNet, DenseNet, Vision Transformers and attention-based hybrid models against the MobileNetV2 baseline established here.
- Additional explainability techniques: incorporating SHAP and LIME alongside Grad-CAM/Grad-CAM++ for a broader, complementary view of model behaviour.
- Early-stage detection: targeted strategies (e.g., class-balanced losses, attention mechanisms, additional augmentation) to improve the persistently weaker performance on early-stage disease categories such as VeryMildDemented.
- Multimodal data fusion: integrating MRI with PET, EEG and structured clinical data through early, late or hybrid fusion strategies to improve diagnostic robustness.
- Broader cross-disease and cross-dataset validation: testing the framework on additional neurodegenerative disorders and on larger, clinically verified, multi-site datasets to assess real-world generalizability.

Clinical translation: prospective evaluation with radiologists and clinicians to assess the practical utility of Grad-CAM++ visualizations in real diagnostic workflows.

References

1. J. Jankovic, "Parkinson's disease: clinical features and diagnosis," Journal of Neurology, Neurosurgery & Psychiatry, vol. 79, no. 4, pp. 368–376, 2008.
2. C. R. Jack Jr. et al., "The role of MRI as a biomarker in Alzheimer's disease," NeuroRx, vol. 2, no. 2, pp. 226–236, 2005.
3. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
4. A. Sarraf and G. Tofighi, "DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI," bioRxiv, 2016.
5. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF CVPR, 2018, pp. 4510–4520.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.
7. D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.
8. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626.
9. A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved visual explanations for deep convolutional networks," in Proc. IEEE WACV, 2018, pp. 839–847.
10. A. Singhal, "Alzheimer MRI Dataset (4 Class of Images, Equal and Augmented)," Kaggle, 2023.
11. I. Sheriff, "Parkinson's Brain MRI Dataset," Kaggle, 2023.
12. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.
13. J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using deep learning," 2018.
14. S. Basaia et al., "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," NeuroImage: Clinical, 2019.
15. T. Jo, K. Nho, and A. J. Saykin, "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," Frontiers in Aging Neuroscience, 2019.
16. D. AlSaeed et al., "Brain MRI analysis for Alzheimer's disease diagnosis using a pre-trained ResNet50," 2022.
17. N. Raza et al., "Alzheimer disease classification through transfer learning approach of DenseNet with gray matter segmentation," 2023.
18. X. Li et al., "Enhanced residual attention network for Alzheimer's disease detection using MRI," 2025.
19. M. Khosravi et al., "Cascade attention CNN for Alzheimer's disease classification using MRI images," 2024.
20. J. Podolszańska, "Development and optimization of deep learning systems for MRI analysis in Alzheimer's disease monitoring," 2024.
21. S. B. Shahid et al., "Novel deep learning approaches for multiclass Alzheimer's disease classification," 2025.
22. S. M. Mousavi et al., "CNN-based classification of Alzheimer's disease using MRI images," 2025.
23. S. E. Sorour et al., "Deep learning framework for early Alzheimer's detection using MRI data," 2024.
24. E. M. Mohammed et al., "Detection of Alzheimer's disease using deep learning models and adaptive optimization," 2024.
25. M. A. Moni et al., "Deep learning-based multiclass classification of Alzheimer's disease," 2025.
26. M. Liu, D. Zhang, and D. Shen, "Deep learning based Alzheimer's disease diagnosis using MRI imaging data," 2020.
27. J. Wen, E. Thibeau-Sutre, M. Diaz-Melo et al., "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation," Medical Image Analysis, 2020.
28. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
29. C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, 2019.
30. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019.
31. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE CVPR, 2017.
32. S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747, 2016.
33. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, 2009.
34. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, 2006.
35. A. M. El-Assy, H. M. Amer, H. M. Ibrahim et al., "A novel CNN architecture for accurate early

detection and classification of Alzheimer's disease using MRI data," Scientific Reports, vol. 14, art. 3463, Feb. 2024.
36. T. Akan, S. A. Alp, M. A. N. Bhuiyan et al., "Leveraging Video Vision Transformer for Alzheimer's Disease Diagnosis from 3D Brain MRI," arXiv, Jan. 2025.
37. A. Majee, A. Gupta, S. Raha, and S. Das, "Enhancing MRI-based classification of Alzheimer's disease with explainable 3D hybrid compact convolutional transformers," arXiv, Mar. 2024.
38. T. Akan, S. A. Alp, and M. A. N. Bhuiyan, "Vision Transformers and Bi-LSTM for Alzheimer's Disease Diagnosis from 3D MRI," arXiv, Jan. 2024.
39. "A multistage convolutional neural network (CNN)-based framework for AD detection and sub-classification using brain MRI," Frontiers in Psychiatry, 2024.
40. "Early detection of Alzheimer's disease in structural and functional MRI," Frontiers in Medicine, 2024.
41. "A novel hybrid filtering and deep learning approach for early Alzheimer's detection," Scientific Reports, June 2025.
42. "Improving early detection of Alzheimer's disease through MRI slice selection with Vision Transformers and EfficientNet," Scientific Reports, Aug. 2025.
43. "Alzheimer's Disease Neuroimaging Initiative (ADNI)," reference dataset repository.