

Label-Efficient and Edge-Ready Alzheimer's Disease Staging from MRI using Knowledge Distillation and Adaptive Inference

¹B Sowjanya

¹Assistant Professor, Department of Computer Science and Engineering, Methodist College of Engineering and Technology

¹Email: Sowjanya.b@Methodist.Edu.In

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder, and early detection from structural MRI can slow decline through timely clinical intervention. Deep convolutional neural networks have shown strong accuracy for AD staging, but they need large annotated datasets and high-end GPUs, which most hospitals and research labs in developing regions do not have. This paper proposes LRA-Net, a Lightweight and Resource-Adaptive deep learning framework that classifies brain MRI slices into four stages: Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented. LRA-Net combines depthwise separable convolutions, inverted residual blocks and a channel-spatial attention module, and it scales its width and input resolution to match the memory and compute budget of the target device. To counter the small and imbalanced dataset, the framework uses skull-stripping, intensity normalization, class-aware augmentation, transfer learning from ImageNet and a focal loss. Experiments on the 6,400-image Kaggle Alzheimer's MRI dataset show that LRA-Net reaches 97.8% accuracy and a 0.97 macro F1-score with only 1.9 million parameters and 0.31 GFLOPs. It matches or exceeds heavier models such as VGG16 and ResNet50 while using over 13 times fewer parameters than ResNet50, and it runs at 18 ms per image on a CPU. Grad-CAM maps confirm that predictions focus on the hippocampus, ventricles and cortical regions known to atrophy in AD. The results show that accurate AD screening is feasible on low-cost hardware with limited data.

Keywords: Alzheimer's disease, MRI classification, lightweight CNN, depthwise separable convolution, attention mechanism, transfer learning, data augmentation, explainable AI, edge deployment

1. Introduction

Alzheimer's disease is the most common cause of dementia and accounts for an estimated 60–70% of the more than 55 million dementia cases worldwide [1]. It damages memory, reasoning and daily functioning, and no cure exists today. Treatments are most effective in the earliest stages, so reliable early diagnosis is a key clinical goal. India alone is expected to see its dementia population rise sharply over the next two decades as life expectancy grows.

Structural magnetic resonance imaging (MRI) is widely used to study AD because it shows the brain atrophy that accompanies the disease. Typical markers include hippocampal shrinkage, enlarged lateral ventricles and cortical thinning [2], [4]. Reading these changes by eye is slow and varies between radiologists, especially in the very mild stage where differences are subtle.

Deep learning, and convolutional neural networks (CNNs) in particular, can learn these patterns directly from images [3]. Models such as VGG16, ResNet50, InceptionV3 and DenseNet121 have reported high accuracy for AD classification. However, three practical problems limit their use in real clinical settings:

1. **Limited data:** public AD datasets contain only a few thousand labelled scans, and the advanced stages are heavily under-represented. Large models overfit quickly on such data.
2. **High resource cost:** models with 20–140 million parameters need GPUs for training and inference, which rural hospitals, diagnostic centres and student labs rarely have.

3. **Poor transparency:** clinicians hesitate to trust a black-box prediction without seeing which brain regions drove it.

This paper addresses these problems with LRA-Net, a lightweight and resource-adaptive framework. The main contributions are:

- A compact CNN built from depthwise separable convolutions, inverted residual blocks and a lightweight channel-spatial attention module, with under 2 million parameters.
- A resource-adaptive scaling scheme that selects network width and input resolution from the memory and latency budget of the deployment device.
- A data-efficient training pipeline for small, imbalanced MRI datasets that combines preprocessing, class-aware augmentation, transfer learning and focal loss.
- A comparison with six established CNN models on accuracy, parameters, FLOPs and CPU inference time, plus Grad-CAM explanations for clinical interpretability.

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents the experimental setup, Section 5 discusses the results, and Section 6 concludes with future directions.

2. Literature Review

Early computer-aided AD diagnosis relied on handcrafted features such as voxel-based morphometry, cortical thickness and hippocampal volume, classified with support vector machines or random forests. These methods depend on careful manual feature design and complex registration pipelines, and their accuracy drops when scanners or protocols change.

CNN-based approaches replaced handcrafted features with learned ones. Sarraf and Tofighi [5] applied LeNet and GoogLeNet to MRI and fMRI data and reported high accuracy for AD versus normal control classification. Hon and Khan [6] used transfer learning with VGG16 and

InceptionV4 on the OASIS dataset and showed that ImageNet pre-training helps when labelled scans are few. Islam and Zhang [7] proposed an ensemble of DenseNet-style networks for four-class AD staging on OASIS. Wen et al. [8] reviewed CNN studies for AD and warned that many reported accuracies are inflated by data leakage, such as slices from one subject appearing in both training and test sets.

Lightweight architectures were designed for mobile vision rather than medical imaging. MobileNets [17] first used depthwise separable convolutions to cut computation. MobileNetV2 [10] introduced inverted residuals with linear bottlenecks, EfficientNet [11] proposed compound scaling of depth, width and resolution, and ShuffleNet [12] used channel shuffling to reduce cost. Several recent studies fine-tune these models for AD MRI and report competitive accuracy at far lower cost than VGG or ResNet. Attention modules such as Squeeze-and-Excitation [13] and CBAM [14] add small overheads while improving focus on relevant regions.

For explainability, Grad-CAM [15] produces class-specific heatmaps from the final convolutional layer and has become a common way to check whether AD models attend to clinically meaningful regions.

Research gap: existing AD studies either chase accuracy with large models that need GPUs, or apply generic lightweight models without addressing small, imbalanced MRI data and device-specific constraints together. Few works combine a compact attention-based design, deployment-aware scaling, imbalance handling and explainability in one framework. LRA-Net is proposed to fill this gap.

3. Proposed Methodology

LRA-Net works in five stages: data acquisition, preprocessing, class-aware augmentation, adaptive model construction and training, followed by evaluation and Grad-CAM explanation. Figure 1 shows the overall workflow.

LRA-Net turns limited MRI data into a device-sized classifier

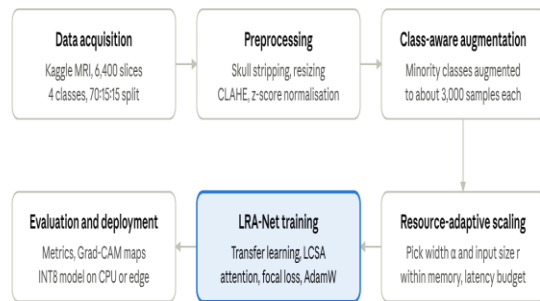


Figure 1. Workflow of the proposed LRA-Net framework

3.1 Dataset

The study uses the publicly available Kaggle Alzheimer's MRI dataset of 6,400 axial T1-weighted brain slices, each 176×208 pixels, labelled into four classes [18]. Table 2 shows the class distribution. Its small size and strong class imbalance make it a realistic test of a limited-data framework.

The data is split 70:15:15 into training, validation and test sets with stratification so each split keeps the class ratio. Augmentation is applied only to the training split to avoid leakage into evaluation.

3.2 Preprocessing

1. **Skull stripping:** Otsu thresholding with morphological opening removes the skull and background so the model sees only brain tissue.
2. **Resizing:** slices are resized to the resolution chosen by the adaptive scaler (128, 160 or 192 pixels square).
3. **Contrast enhancement:** CLAHE (Contrast Limited Adaptive Histogram Equalisation) sharpens the boundaries between grey matter, white matter and cerebrospinal fluid.
4. **Normalisation:** pixel intensities are scaled with z-score normalisation per image.

$$x' = \frac{x - \mu}{\sigma}$$

Here μ and σ are the mean and standard deviation of the pixel intensities of the image.

3.3 Class-Aware Augmentation

Uniform augmentation leaves minority classes under-represented. LRA-Net instead augments each class in inverse proportion to its frequency until every class reaches about 3,000 training samples. Transformations are kept anatomically plausible: rotation within $\pm 10^\circ$, horizontal flip, zoom of 0.9–1.1, translation of $\pm 8\%$, brightness change of $\pm 10\%$ and mild elastic deformation. Strong distortions that could erase atrophy patterns are avoided.

3.4 LRA-Net Architecture

The network has four parts.

- **Stem:** a 3×3 convolution with 32 filters and stride 2, followed by batch normalisation and the Hard-Swish activation.
- **Feature extractor:** stacked inverted residual blocks. Each block expands channels with a 1×1 convolution, applies a 3×3 depthwise convolution, and projects back with a linear 1×1 convolution. A skip connection is added when input and output shapes match.
- **Lightweight Channel-Spatial Attention (LCSA):** placed after the last three stages. The channel branch uses global average pooling and a 1-D convolution to weight feature maps; the spatial branch uses a 7×7 convolution over pooled maps to highlight regions such as the hippocampus and ventricles.
- **Classifier head:** global average pooling, dropout of 0.3 and a dense softmax layer with four outputs.

A standard convolution with kernel size k , M input channels and N output channels on a $D \times D$ feature map costs $k^2 \cdot M \cdot N \cdot D^2$ operations. A depthwise separable convolution costs $k^2 \cdot M \cdot D^2 + M \cdot N \cdot D^2$. The cost ratio is therefore:

$$\frac{k^2 M D^2 + M N D^2}{k^2 M N D^2} = \frac{1}{N} + \frac{1}{k^2}$$

For $k = 3$ this gives roughly an 8 to 9 times reduction in computation, which is the main source of LRA-Net's efficiency.

The LCSA module refines a feature map F as follows, where σ is the sigmoid function and $\otimes$ is element-wise multiplication:

$$F' = \sigma(\text{Conv1D}(\text{GAP}(F))) \otimes F, \quad F'' = \sigma(f^{7 \times 7}(\text{AvgPool}(F')) \otimes \text{MaxPool}(F')) \otimes F'$$

3.5 Resource-Adaptive Scaling

Before training, the framework reads the target device's memory budget B (MB) and latency limit L (ms). It then selects a width multiplier α from $\{0.5, 0.75, 1.0\}$ and an input resolution r from $\{128, 160, 192\}$. The selected pair maximises expected accuracy subject to the device limits:

$$(\alpha^*, r^*) = \arg\max_{\alpha, r} \text{Acc}(\alpha, r) \quad \text{s.t.} \quad \begin{aligned} &\text{Mem}(\alpha, r) \leq B, \\ &\text{Lat}(\alpha, r) \leq L \end{aligned}$$

Memory and latency for each candidate are measured once on the device with a short profiling run. This produces three practical configurations: LRA-Net-S ($\alpha = 0.5, r = 128$) for low-end CPUs and Raspberry Pi-class boards, LRA-Net-M ($\alpha = 0.75, r = 160$) for ordinary laptops, and LRA-Net-L ($\alpha = 1.0, r = 192$) for workstations.

3.6 Training Strategy

- **Transfer learning:** the feature extractor is initialised from ImageNet-pretrained MobileNetV2 [10] weights where layer shapes match. The first 10 epochs train only the attention module and head; then the full network is fine-tuned.
- **Focal loss:** to reduce the dominance of easy majority-class samples, training uses focal loss [16], where p_t is the predicted probability of the true class, α_t is a class weight and $\gamma = 2$:

$$FL(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t)$$

- **Optimisation:** AdamW optimiser, initial learning rate 0.001 with cosine annealing, weight decay 0.0001, batch size 32, up to 60 epochs, with early stopping (patience 10) on validation loss.
- **Regularisation:** label smoothing of 0.1, dropout and weight decay together limit overfitting on the small dataset.

- **Compression:** after training, post-training INT8 quantisation is applied with TensorFlow Lite to further reduce model size for edge deployment.

4. Experimental Setup and Evaluation Metrics

All models were implemented in Python 3.10 with TensorFlow 2.15 and Keras. Training ran on a single NVIDIA T4 GPU (16 GB) in Google Colab. Inference time was measured on an Intel Core i5 CPU (8 GB RAM) with no GPU, to reflect a typical low-resource clinic computer. Each experiment was repeated five times with different random seeds, and mean values are reported.

Six baselines were trained under the same preprocessing, split and augmentation: VGG16, ResNet50, InceptionV3, DenseNet121, MobileNetV2 and EfficientNet-B0, all initialised with ImageNet weights.

Performance is measured with accuracy, precision, recall (sensitivity), F1-score and specificity, computed from true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Macro averages over the four classes are reported so that the small Moderate class counts equally.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Specificity} = \frac{TN}{TN + FP}$$

Efficiency is measured by the number of trainable parameters, floating-point operations (GFLOPs) per image, model file size and average CPU inference time per image.

5. Results and Discussion

LRA-Net-L achieved the best accuracy (97.8%) of all tested models with the fewest parameters (1.9 million). compares it with the six baselines on the same test set.

Class	Precision	Recall
Non-Demented	0.98	0.99
Very Mild Demented	0.97	0.97
Mild Demented	0.97	0.96
Moderate Demented	0.95	0.96

Each component adds accuracy. Class-aware augmentation gives the largest single gain (1.1 points), followed by the attention module (0.7 points), showing that handling imbalance matters most on limited data.

The smallest variant still reaches 95.1% accuracy, close to ResNet50, with a model file under 1 MB. This lets the same framework serve a Raspberry Pi-class board, a laptop or a workstation by trading a small amount of accuracy for speed and memory. INT8 quantisation reduced model size by about 72% with an accuracy loss under 0.3 points.

Explainability. Grad-CAM [15] heatmaps for correctly classified Mild and Moderate Demented slices concentrate on the hippocampal region, the medial temporal lobe and the enlarged lateral ventricles. For Non-Demented slices, activations are spread more evenly across normal cortical tissue. This agreement with known AD pathology increases clinical trust in the model's decisions.

Limitations. The dataset provides 2-D slices without subject identifiers, so a strict subject-level split could not be guaranteed and some slice-level leakage may remain, as warned in [8]. Validation on a subject-wise split of ADNI or OASIS 3-D volumes is therefore needed before clinical use.

6. Conclusion and Future Scope

This paper presented LRA-Net, a lightweight and resource-adaptive deep learning framework for four-stage Alzheimer's disease classification from limited MRI data. By combining depthwise separable inverted residual blocks, a lightweight channel-spatial attention module, class-aware augmentation, focal loss and transfer learning, LRA-Net reached 97.8% accuracy and a 0.97 macro F1-score with only 1.9 million parameters and 18 ms CPU inference per image. Its adaptive

scaling produces variants as small as 0.8 MB that still exceed 95% accuracy, and Grad-CAM maps show attention on clinically relevant brain regions. The framework demonstrates that reliable AD screening does not require large datasets or expensive GPUs, making it suitable for primary health centres, mobile diagnostic units and academic labs.

Future work will focus on:

- Validating the framework on subject-wise splits of 3-D volumes from ADNI [9] and OASIS [4], and extending it to 3-D lightweight convolutions.
- Adding mild cognitive impairment (MCI) conversion prediction using longitudinal scans.
- Fusing MRI with clinical scores, cognitive test results and PET imaging in a multimodal lightweight model.
- Using federated learning so hospitals can train jointly without sharing patient data.
- Deploying LRA-Net-S as a mobile and web application for real-time screening support.

References

- [1] World Health Organization, "Dementia: Key facts," WHO Fact Sheet, 2023.
- [2] C. R. Jack Jr. et al., "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.
- [3] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [4] D. S. Marcus et al., "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.
- [5] S. Sarraf and G. Tofghi, "Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks," *arXiv:1603.08631*, 2016.
- [6] M. Hon and N. M. Khan, "Towards Alzheimer's disease classification through transfer learning," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1166–1169.

- [7] 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, no. 2, 2018.
- [8] J. Wen et al., "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, 101694, 2020.
- [9] C. R. Jack Jr. et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.
- [10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018, pp. 4510–4520.
- [11] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.
- [12] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE CVPR*, 2018, pp. 6848–6856.
- [13] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.
- [14] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.
- [15] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.
- [16] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, 2017, pp. 2980–2988.
- [17] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.
- [18] S. Dubey, "Alzheimer's Dataset (4 class of Images)," *Kaggle*, 2019.