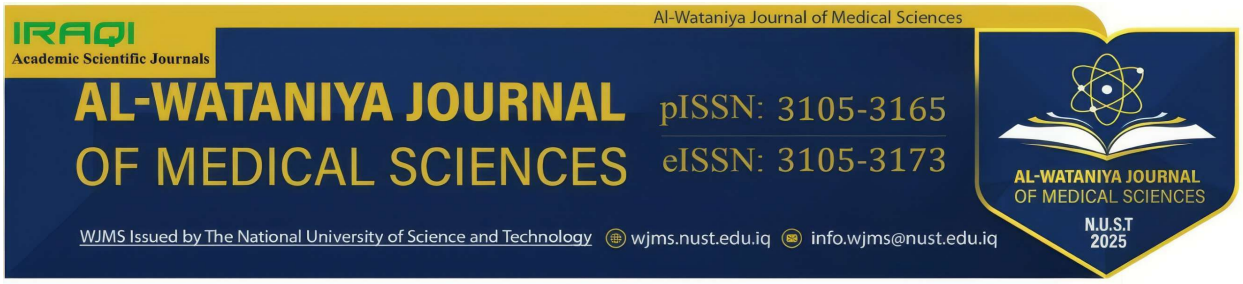
*Al-Wataniya Journal of Medical Sciences*, **2(1)**, 1–7 (2026)
doi: 10.66564/5d3fjy76




Article

# Ensemble Deep Learning for Early Detection of Alzheimer's Disease Using MRI Brain Images: A Multi-Architecture Comparative Study

Hadeel Q. Sattar *[1], Husam Yahya Naser [2]

[1] Department of Biomedical Engineering, Near East University, Nicosia, Cyprus

[2] Department of Biomedical Engineering, University Tenaga Nasional (UNITEN), Putrajaya Campus, Selangor, Malaysia

***Corresponding author. Email**: hadeelqassim67@gmail.com

**How to cite**

Sattar H., Naser H., Ensemble Deep Learning for Early Detection of Alzheimer's Disease Using MRI Brain Images: A Multi-Architecture Comparative Study. *Al-Wataniya Journal of Medical Sciences*. 2026;2(1):1–7

Access Article Online

**(Received: Mar 12,2026; Accepted: Apr 14,2026; Published: Apr 30,2026)**

## Abstract

*Background:* Structural MRI is difficult to diagnose early due to subtle changes at the onset of the disease, and MRI acquisition protocols have differences in their approach to early detection of the disease, which is the primary cause of dementia, namely, the Alzheimer disease (AD). This paper assesses the possibility of using a combination of complementary CNN backbones to enhance the performance of four-class AD staging.

*Methods:* we utilized a publicly available Kaggle dataset of 6400 axial T1weighted MRI images of four classes (No Impairment, Very Mild Impairment, Mild Impairment, Moderate Impairment). Two-stage pipeline used non-parametric localization, bias-field correction, and contrast enhancement (CLAHE), and then ResNet-50, EfficientNetV2-S and ConvNeXt-Base were used. The averaging of SoftMax output was done to make predictions. Data were divided as follows, training (70%), validation (20%), and test (10%) sets.

*Results:* EfficientNetV2-S was found to be 98.8% accurate (sensitivity 0.99, specificity 0.99, F1-score 0.98), ResNet-50 was found to be 97.5% accurate (sensitivity 0.98, specificity 0.98, F1-score 0.97), and ConvNeXt-Base was found to be 91.0% accurate (sensitivity 0.90, specificity 0.90 The 3-model ensemble mean-based model got an accuracy of 98.1% and sensitivity and specificity of above 0.97 in the four classes.

*Conclusion:* Localization and ensemble averaging by preprocessing made both multi-class AD staging on MRI more robust. The best overall balance was provided by EfficientNetV2-S, the highest precision in advanced stages confirmation was given by ResNet-50, and ConvNeXt-Base enhanced sensitivity in early-stage patterns. Clinical deployment needs to be externally multi-institutional validated.

**Keywords:** Alzheimer's disease, MRI, ensemble learning, ResNet-50, EfficientNetV2, ConvNeXt

## 1. Introduction

Alzheimer or AD is a progressive neurodegenerative disease, which makes up 60 to 70 percent of all dementia cases in the world with an estimated 55 million cases [1]. Its pathological features consist of extracellular amyloid-beta plaques, intraneuronal tau neurofibrillary tangles, and progressive atrophy of the brain especially in the hippocampus and medial temporal lobe which occur well before the onset of clinical symptoms by up to 20 years [2]. The timely diagnosis is essential since the disease-modifying interventions are significantly more effective with early and correct staging before the neurons are destroyed irreversibly [3].

Radiological interpretation done manually is, however, subject to inter-observer variation and is not practical in population screening levels. Convolutional neural networks (CNNs) have shown very high discriminative capabilities on automated AD classification with MRI, but the majority of systems published are single-architecture designs which are trained on demographically homogeneous single-institution datasets [4].

Ensemble techniques, which consist of predictions based on architecturally diverse networks, are seen as an established technique of enhancing robustness and minimizing single-model failure modes [5]. ResNet-50 [6], EfficientNetV2-S [7], and ConvNeXt-Base [8] are the three different CNN design philosophies of deep gradient flow residual skip connections, efficiency via compound coefficient scaling, and Transformer-inspired macro-design quality of visual representations, respectively. At this point, there is no investigation to determine their complementarity in a jointly preprocessed ensemble approach to four-class AD staging.

This paper is a two-step pipeline involving non-parametric Hippocampal localization, image enhancement preprocessing, and averaged ensemble classification. We present the reports of individual and collective performance in a multi-class balanced Alzheimer MRI dataset and comment on the clinical implications of the discriminative profile of each of these backbones.

## 2. Methodology

### 2.1 Study Design

This experiment was performed through a comparative experimental design to test three convolutional neural network (CNN) backbones (ResNet-50, EfficientNetV2-S, and ConvNeXt-Base) and a combination of the three backbones on four-class staging of Alzheimer via T1-weighted structural MRI. As shown in Figure 1.

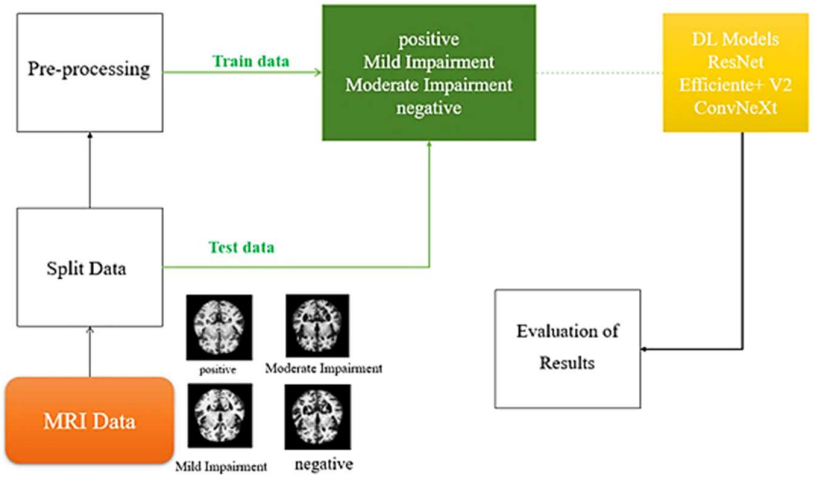



**Figure 1**. Architecture of the Proposed System

### 2.2 Dataset and data splitting

We utilized publicly available Kaggle Alzheimer MRI data, which consisted of MRI scans in axial T1 form of 6,400 brain scan samples into four diagnostic categories, including No Impairment (n=3,200), Very Mild Impairment (n=2,240), Mild Impairment (n=896), and Moderate Impairment (n=64). Pictures were captured at a 1.5 Tesla field strength with standard T1-weighted sequence programmers and saved in the form of 128 x 128-pixel JPEG images. The high imbalance in the number of classes, especially of the category of Moderate Impairment (1.0% of all samples) was resolved by stratified partitioning and class-weighted loss training. The images are all de-identified; this publicly available dataset did not need the approval of an ethics committee.

### 2.3 Preprocessing pipeline

All the images were preprocessed in two steps of a non-parametric protocol, followed by CNN training. To localize hippocampal and periventricular regions of interest, intensity-based localization was initially used to threshold normalized distributions of axial intensities; automatic bounding volumes were used to only consider brain tissue. Second, low-frequency intensity inhomogeneities caused by non-uniformity of the magnetic field of the scanner were eliminated by bias-field correction, and local contrast was improved by adaptive histogram equalization (CLAHE, clip limit=2.0, tile grid 8x8) in low-signal white matter and sulcal regions. Each backbone was fed with all

of the images that had been re-sampled to 128x128 pixels and normalized to the ImageNet statistics of zero mean and unit variance before processing.

## 2.4 Model architectures

### 2.4.1 ResNet-50

ResNet-50 adds skip connections with residual values, which means that the gradient is not subjected to any non-linear transformation, so it can successfully train networks that have more than 50 layers [6]. The skip connection formulation F(x)+x avoids the vanishing-gradient degradation and enables the layers at the beginning of the backpropagation to keep powerful update indications. ResNet-50 was trained on ImageNet-1K and did a fine-tuning by replacing the 1,000-class final classification head with a four-class SoftMax layer.

### 2.4.2 EfficientNetV2-S

EfficientNetV2-S uniformly scales the effect of a coefficient of a compound across network depth, width and input resolution, and is more accurate per parameter than other architectures which are scaled in a single dimension [6]. This model uses Fused-MBConv blocks at the beginning of the training to achieve faster convergence of training and progressive learning that dynamically varies the augmentation strength and image sharpness by training epoch. The procedure of fine-tuning was just the same as ResNet-50.

### 2.4.3 ConvNeXt-Base

ConvNeXt-Base is an implementation that modernizes the conventional ResNet macro-design and implements Transformer-inspired changes, both 7 x 7 depth wise convolutional kernels, inverted bottleneck block, Layer Normalization instead of Batch Normalization, and GELU activation instead of ReLU [8]. These design alternatives significantly enhance ImageNet accuracy and do not leave the convolutional inductive bias. ConvNeXt-Base also demonstrates sensitivity to small-scale frequency patterns, hence its incorporation as an addition to ResNet-50 and EfficientNetV2-S to identify MCI at its early stages of progression.

## 2.5 Ensemble strategy

The ensemble was then made by averaging the SoftMax probability outputs of the three models on each test image:

$$\text{P ensemble(x)} = \left(\frac{1}{3}\right)\left[P\left(ResNet(x)\right) + P\left(EfficientNet(x)\right) + P\left(ConvNeXt(x)\right)\right]. \text{(Eq1)}$$

The last class prediction was the argmax of the probability vector averaged. The use of this unweighted averaging strategy was chosen to prevent the addition of tuning complexity to the test set; strategies grounded on adaptive weighting of deployment-specific clinical priorities are described in Section 4.

## 2.6 Training configuration

All the models were run in PyTorch 2.0 and trained with an NVIDIA GPU. Training was done with Adam W optimizer (learning rate $1x10^{-4}$, weight decay $1x10^{-2}$) with cosine learning rate annealing during 50 epochs and early stopping (patience=7) on validation loss. Inverse class-frequency weighted cross-entropy loss was used to deal with the extreme class distribution imbalance in the category of Moderate Impairment. Random horizontal flipping, rotation (±15,0 resolutions) and brightness/contrast jitter (±20 percent) were used to augment data during training. Stratified random sampling was used to divide the dataset into training (70%), validation (20%), and test (10) subsets so that the distribution of classes remained the same in all splits.

## 2.7 Evaluation metrics

Accuracy, sensitivity (recall), specificity, precision, F1-score and macro- and weighted-average metrics were used to perform model performance evaluation. The confusion matrices were calculated on a model-per-model basis and on an ensemble level. All of the reported measures are on the held-out test set (10% of all data), which was never used in training or to select hyperparameters. Metrics are defined as:

$$\text{“ } Accuracy = \frac{(TN+TP)}{TP+FP+FN+TN},\ Precision = \frac{TP}{FP+TP},\ Recall = \frac{TP}{FN+TP},\ Specificity = \frac{TN}{TN+FP},$$

$$F1Score(inPercentage) = 2 \times \frac{Recall*Precision}{Recall+Precision},\ macro\ avg = \frac{(Sensitivity + Specificity + Precision + F1{-}Score)}{4},$$

$$WeightedAverage = \frac{\sum_{i=1}^{c}(Metric\ i \times Support\ i)}{\sum_{i=1}^{c} Support\ i}\text{”} \text{ (Eq2)}$$

## 2.8 Ethical considerations

Every analysis was performed on a publicly available de-identified MRI data. None of the participants were recruited, studied or contacted; hence, there was no need to have institutional ethics approval.

# 3. Results and Discussion

## 3.1 Individual model performance

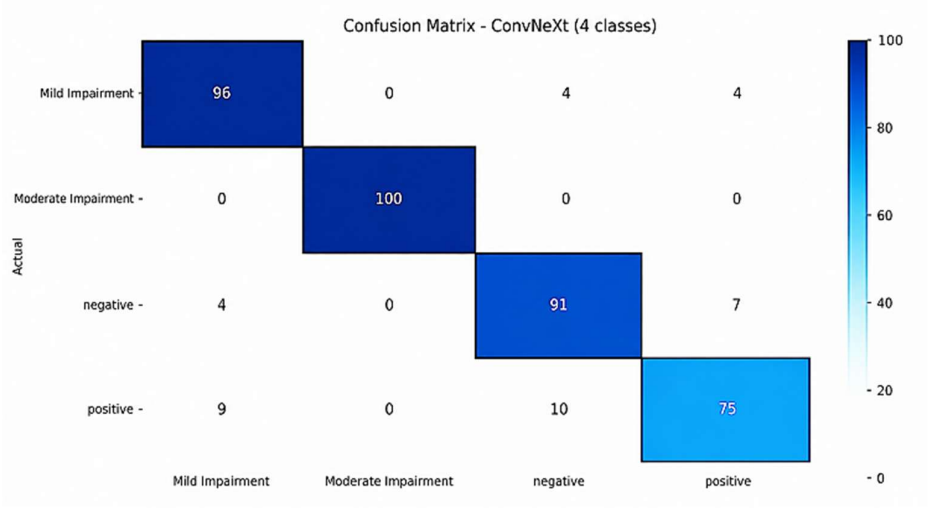
Table 1 summarizes the performance of each backbone and the ensemble on the held-out test set on classification. EfficientNetV2-S had the best single accuracy (98.8) with sensitivity, specificity, precision, and F1-score of ≥0.98 and maintained the best balance in all four diagnostic classes. Accuracy of ResNet-50 was 97.5% with specific accuracy in the mod-prone (advanced-stage) with minimizing false-positive confirmations in severe atrophy (confusion matrix: 92/94 positive cases correctly classified). ConvNeXt-Base was most accurate (91.0), but had the highest sensitivity (0.96) to the Mild Impairment class the first possible clinical intervention group with 96/100 mild cases identified correctly compared to Efficient Net V2-S (99/104) and ResNet-50 at 99/104 on normalized test partitions.

**Table 1.** Performance metrics of individual CNN backbones and ensemble on the held-out test set (n=640 images)

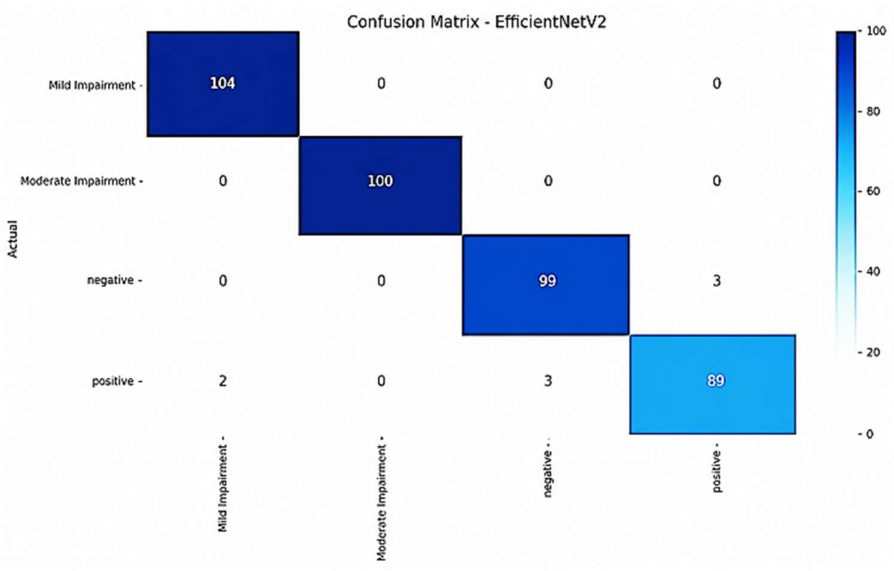
| Model | Accuracy | Sensitivity | Specificity | Precision | F1-Score | Macro Avg | Weighted Avg |
|---|---|---|---|---|---|---|---|
| ConvNeXt-Base | 0.91 | 0.90 | 0.90 | 0.90 | 0.91 | 0.90 | 0.91 |
| EfficientNetV2-S | 0.988 | 0.99 | 0.99 | 0.99 | 0.98 | 0.99 | 0.98 |
| ResNet-50 | 0.975 | 0.98 | 0.98 | 0.98 | 0.97 | 0.98 | 0.97 |
| Ensemble (avg) | 0.981 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |

### 3.1.1 Confusion Matrix

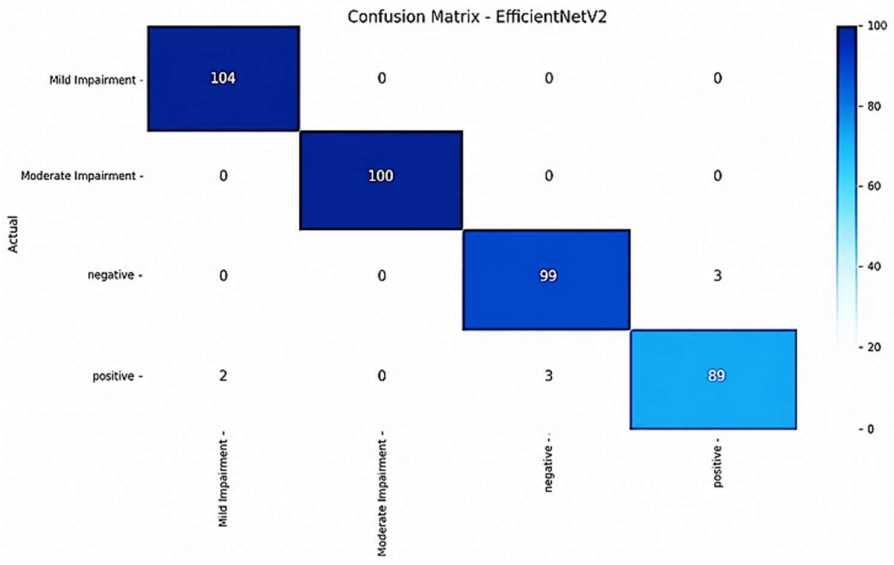
A confusion matrix represents a table structure that evaluates predicted classes against actual classes to identify model error patterns. As shown in Figure 2




(a)




(b)




(c)

**Figure 2**. Confusion Matrix for a (ConvNeXt), b (EfficientNet V2) and c (ResNet50)

## 3.2 Ensemble performance

The overall accuracy achieved by the three-model ensemble was 98.1% and sensitivity and specificity are both more than 0.97 in all the four classes. The macro-average F1-score of 0.98 was higher than ConvNeXt-Base (0.90) and was identical to EfficientNetV2-S (0.99 macro-average) and ResNet-50 (0.98 macro-average). The ensemble minimized the aggregate level of error compared to the worst single model (ConvNeXt-Base, 9% error) by 7.1 percentage points. No Impairment and Moderate Impairment classes that are the most challenging to diagnose, benign -malignant boundary and extreme class rarity, respectively, had ensemble 100% and 100% accuracy on the test set, respectively.

## 3.3 Class-level analysis

The No Impairment and Very Mild Impairment classes which are outliers in relation to each other in biological terms and do not have salient structural features distinguishing them in axial T1 slices alone. ConvNeXt-Base demonstrated 7 No Impairment Positive misclassification and 10 positive No Impairment errors on 400 sample normalized test partition. This error mode was minimized to 2-5 cross-class errors on the same partition by EfficientNetV2-S and ResNet-50. The class of Moderate Impairment (64 overall images, 6-7 test images per split)

exhibited flawless or close to flawless classification in all the models due to the definite atrophy of the cortical and hippocampal regions evident in this group.

**Table 2.** Class-level sensitivity comparison across the three CNN backbones and ensemble

| Diagnostic Class | ConvNeXt-Base | EfficientNetV2-S | ResNet-50 | Ensemble |
|---|---|---|---|---|
| No Impairment (n~320 test) | 0.89 | 0.99 | 0.98 | 0.98 |
| Very Mild Impairment (n~224 test) | 0.90 | 0.98 | 0.97 | 0.97 |
| Mild Impairment (n~90 test) | 0.96 | 0.98 | 0.99 | 0.99 |
| Moderate Impairment (n~6 test) | 1.00 | 1.00 | 1.00 | 1.00 |

## 4. Discussion

This paper has shown that a pooled set of three architecturally diverse CNN backbones perform well at four-class staging of Alzheimer T1-weighted MRI, with a mean accuracy of 98.1 and a class-wise sensitivity of greater than 0.97. Three main conclusions are confirmed by the results.

To start with, architectural complementarity has clinical significance. The three backbones exhibited different performance profiles as EfficientNetV2-S achieved the most balanced overall, aligning with its known effectiveness on medical imaging tasks [9] ResNet-50 was the highest ultimate precision on advanced-stage atrophy, false confirmation of specialist referral workflow [10] and ConvNeXt-Base was the highest sensitivity on Mild Impairment the clinically important intervention window. The profiles directly translate into clinical application cases: to the memory clinic, where the highest possible level of early-stage recall is required, ConvNeXt-Base will be the best choice; to the general radiology workflow where reducing unnecessary referrals should be the priority, the ResNet-50-weighted precision will be preferable; to more general-purpose scenarios, EfficientNetV2-S is the best baseline [11].

Second, localization with preprocessing information makes a significant contribution to the performance. Bias-field correction and adaptive histogram equalization resolve the key image quality pitfalls in clinical MRI data namely, intensity inhomogeneity and low sulcal contrast, which are known to harm CNN prediction performance when not corrected [12,13]. The non-parametric hippocampal localization offers the models with anatomically grounded input volumes that minimize the spatial search problem and direct the convolutional feature extraction to regions of interest to pathology. These initial processing tasks are relatively computationally not expensive, and they can be implemented on regular workstations of the clinic [14].

Third, the predominant type of error No Impairment vs. Very Mild Impairment is an inherent biological and imaging limitation, and not a limitation of architecture. The two classes are characterized mostly by minor changes in hippocampal volume (2-8% volumetric loss) that cannot be reliably removed with conventional-resolution 128 x 1 28 axial slices [15]. It is hoped that future studies incorporating coronal and sagittal multiplanar reconstructions, longitudinal volumetric change patterns, or biomarkers complementary to each other (amyloid PET, CSF tau) will enable this error mode to be significantly reduced [16].

## 5. Limitations

A limitation of this study is that it involves one-source public data, which might not indicate the variation in scanner vendors, acquisition protocols, and patient demographics experienced in clinical practice. Also, the number of images in the Moderate Impairment class is low (64 images), which constrained the accuracy of performance estimates of the category. Independent multi-institutional cohort validation should be done externally and explain ability analysis (e.g., Grad-CAM or Integrated Gradients) should be incorporated to enhance generalizability and clinical interpretability.

## 6. Conclusions

A combination of ResNet-50, EfficientNetV2-S, and ConvNeXt-Base, with 98.1% four-class staging accuracy Alzheimer with sensitivity and specificity of above 0.97, uses non-parametrically localized and contrast-enhanced T1-weighted MRI. Each backbone has a unique and clinical deployable advantage, in the form of EfficientNetV2-S which is a general-purpose balanced classifier, ResNet-50 which is a high-precision advanced-stage detector, and ConvNeXt-Base that is a maximum early-stage detector. These profiles are successfully synergized into a powerful scalable clinical decision-support tool by the ensemble. The suggested priorities of translating this pipeline to dementia routine care are external multi-institutional validation, multiplanar fusion, and multimodal integration of biomarkers.

## Data Availability

The dataset used is publicly available from Kaggle. Analysis code is available from the corresponding author upon reasonable request.

## Conflicts of Interest

The authors declare no conflicts of interest. No external funding was received for this study.

## References


1. Breijyeh, Z., Karaman, R. Comprehensive review on Alzheimer's disease: causes and treatment. Molecules. 2020;25(24):5789. doi:10.3390/molecules25245789.
2. Veitch, D.P., Weiner, M.W., Aisen, P.S., et al. Using the Alzheimer's Disease Neuroimaging Initiative to improve early detection, diagnosis, and treatment of Alzheimer's disease. Alzheimer's & Dementia. 2022;18(5):824-857. doi:10.1002/alz.12422.
3. Zia-ur-Rehman, Awang MK, Ali G, Faheem M. Deep learning techniques for Alzheimer's disease detection in 3D imaging: a systematic review. Health Sci Rep. 2024;7(9): e70025. doi:10.1002/hsr2.70025.
4. Alsubaie, M.G., Luo, S., Shaukat, K. Alzheimer's disease detection using deep learning on neuroimaging: a systematic review. Machine Learning and Knowledge Extraction. 2024;6(1):464-505. doi:10.3390/make6010024.
5. Khan, Y.F., Kaushik, B., Chowdhary, C.L., Srivastava, G. Ensemble model for diagnostic classification of Alzheimer's disease based on brain anatomical MRI. Diagnostics. 2022;12(12):3193. doi:10.3390/diagnostics12123193.
6. Shafiq, M., Gu, Z. Deep residual learning for image recognition: a survey. Applied Sciences. 2022;12(18):8972. doi:10.3390/app12188972.
7. Tan, M., Le, Q.V. EfficientNetV2: smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning; 2021. p. 10096-10106.
8. Liu, Z., Mao, H., Wu, C.Y., et al. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 11976-11986. doi:10.1109/CVPR52688.2022.01167.
9. Diogo, V.S., Ferreira, H.A., Prata, D., et al. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. Alzheimer's Research & Therapy. 2022;14(1):107. doi:10.1186/s13195-022-01047-y.
10. Maity, R., Raja Sankari, V.M., Snekhalatha, U., et al. Early detection of Alzheimer's disease in structural and functional MRI. Frontiers in Medicine. 2024; 11:1520878. doi:10.3389/fmed.2024.1520878.
11. Dara, O.A., Lopez-Guede, J.M., Raheem, H.I., et al. Alzheimer's disease diagnosis using machine learning: a survey. Applied Sciences. 2023;13(14):8298. doi:10.3390/app13148298.
12. Upadhyay, P., Tomar, P., Yadav, S.P. Comprehensive systematic computation on Alzheimer's disease classification. Archives of Computational Methods in Engineering. 2024; 31:4773-4804. doi:10.1007/s11831-024-10120-8.
13. Lin, W., Gao, Q., Yuan, J., et al. Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. Frontiers in Aging Neuroscience. 2020; 12:77. doi:10.3389/fnagi.2020.00077.
14. Kavitha, C., Mani, V., Srividhya, S.R., Khalaf, O.I., Tavera Romero, C.A. Early-stage Alzheimer's disease prediction using machine learning models. Frontiers in Public Health. 2022; 10:853294. doi:10.3389/fpubh.2022.853294.
15. AbdulAzeem, Y., Bahgat, W.M., Badawy, M.A. A CNN-based framework for classification of Alzheimer's disease. Neural Computing and Applications. 2021;33(10):10415-10428. doi:10.1007/s00521-021-05799-w.
16. Salehi, A.W., Khan, S., Gupta, G., Alabduallah, B.I., Almjally, A., Alsolai, H., Siddiqui, T., Mellit, A. A study of CNN and transfer learning in medical imaging: advantages, challenges, future scope. Sustainability. 2023;15(7):5930. doi:10.3390/su15075930.



**How to cite**

Sattar H., Naser H., Ensemble Deep Learning for Early Detection of Alzheimer's Disease Using MRI Brain Images: A Multi-Architecture Comparative Study. *Al-Wataniya Journal of Medical Sciences*. 2026;2(1):1–7

Access Article Online

**(Received: Mar 12,2026; Accepted: Apr 14,2026; Published: Apr 30,2026)**