International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

Ensemble Convolutional Neural Networks for Alzheimer's Disease Detection: A Systematic Review

Shrey Anand Srivastava[1], Vinayak Shukla[2], Rudra Bahadur Singh[3], Swapnil Singh[4], Vedant Bhagwani[5]
[1,4,5]Ug-Computer Science & Engineering, Bbditm, Lucknow, India
[2,3]Professor, Computer Science & Engineering, Bbditm, Lucknow, India
Emails:*shreysrivastava189@gmail.com[1],srmvinayak@gmail.com[2],rudra.rathor20@gmail.com[3]*
swa929911@gmail.com[4], vedantbhagwani777@gmail.com[5].

Abstract
Alzheimer's disease (AD) is the most prevalent form of dementia worldwide, affecting an estimated 55 million people globally and projected to triple by 2050. Timely and accurate diagnosis is critical; however, conventional clinical methods relying on neuropsychological assessments and invasive cerebrospinal fluid (CSF) biomarkers detect pathology only after significant neurodegeneration has already occurred. Magnetic Resonance Imaging (MRI) offers a non-invasive window into structural brain changes, and deep learning, particularly Convolutional Neural Networks (CNNs), has demonstrated remarkable capability in extracting discriminative spatial features from MRI volumes for automated AD classification. This systematic review investigates ensemble CNN frameworks that synergistically combine ResNet, InceptionV3, and EfficientNet architectures to overcome the inherent limitations of individual models, including sensitivity to dataset variability, architectural bias, and overfitting. Two principal ensemble strategies are critically analyzed: (1) prediction level ensembling via stacking and boosting, achieving classification accuracy up to 95%; and (2) feature-level fusion through intermediate representation concatenation, achieving accuracy as high as 99.13%. The review also examines transfer learning strategies, domain adaptation techniques, performance metrics, and the translational challenges of deploying AI based diagnostic tools in clinical settings. Findings confirm that multi-scale feature integration within ensemble CNN architectures significantly improves the robustness, generalizability, and diagnostic accuracy of non-invasive automated AD detection systems.
Keywords: *Alzheimer's Disease (AD); Convolutional Neural Networks (CNNs); Deep Learning; Dementia; EfficientNet; Ensemble Learning; Feature Fusion; Inception; MRI Classification; Neuroimaging; ResNet; Transfer Learning.*

1. Introduction
Alzheimer's disease (AD) is a chronic, progressive neurodegenerative disorder characterized by irreversible deterioration of cognitive functions, including memory, language, reasoning, and executive function. As the leading cause of dementia worldwide, AD accounts for 60–80% of all dementia cases and currently affects an estimated 55 million people globally, a figure projected to reach 139 million by 2050 as populations age [1]. Beyond its devastating personal impact on patients and caregivers, AD imposes an enormous socioeconomic burden global dementia-related costs exceeded USD 1.3 trillion in 2019 and are expected to double within the decade. At the molecular level, AD is defined by two hallmark neuropathological features: (1) extracellular accumulation of amyloid-beta (Aβ) plaques, formed by misfolded fragments of the amyloid precursor protein (APP); and (2) intracellular neurofibrillary tangles (NFTs), composed of hyperphosphorylated tau protein. These aggregates disrupt synaptic signaling, trigger neuroinflammatory cascades, and ultimately cause widespread neuronal death particularly in the hippocampus and entorhinal cortex, regions essential for memory consolidation. Notably, these pathological changes begin accumulating 15–20 years before clinical symptoms manifest, underscoring the importance of biomarker-based early detection [1].Current clinical diagnostic

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

protocols rely primarily on neuropsychological battery tests (e.g., MMSE, MoCA) and invasive CSF biomarker assays measuring Aβ42 and tau protein levels. While effective, these approaches suffer from significant limitations: neuropsychological tests lack sensitivity in the prodromal (pre-symptomatic) phase, and lumbar puncture for CSF extraction is invasive, painful, and associated with procedural risks. Positron Emission Tomography (PET) imaging with amyloid tracers offers pre-symptomatic sensitivity but remains prohibitively expensive and exposes patients to ionizing radiation. Consequently, the majority of AD diagnoses are made only after clinically significant cognitive decline has occurred, precluding the therapeutic window for disease-modifying interventions [2]. Structural MRI has emerged as the most clinically accessible and cost-effective neuroimaging modality for AD research, capturing measurable hallmarks of neurodegeneration including hippocampal atrophy, cortical thinning, and ventricular enlargement. However, visually interpreting subtle structural changes in MRI particularly in early-stage disease demands exceptional expertise and is subject to significant inter-rater variability. This limitation has motivated the development of automated, data-driven approaches grounded in machine learning and, more recently, deep learning [2]. Deep Convolutional Neural Networks (CNNs) have revolutionized medical image analysis by learning hierarchical spatial feature representations directly from raw pixel data, eliminating the need for manual feature engineering. Architectures such as ResNet, InceptionV3, and EfficientNet have consistently achieved state-of-the-art performance on large-scale image recognition benchmarks and have been successfully adapted for MRI based AD classification through transfer learning. However, individual CNN models remain susceptible to dataset-specific bias, architectural constraints, and overfitting challenges that are particularly pronounced given the relatively small scale of available annotated neuroimaging datasets. Ensemble learning addresses these limitations by strategically combining predictions or feature representations from multiple diverse models, exploiting their complementary strengths while mitigating individual weaknesses. This review systematically examines the landscape of ensemble CNN approaches for AD detection, with a focus on ResNet, InceptionV3, and EfficientNet architectures. We analyze two principal ensemble paradigms prediction- level and feature level ensembling alongside transfer learning strategies and domain adaptation techniques. The review further discusses evaluation metrics, dataset benchmarks, open research challenges, and future directions for translating AI-based AD diagnostic tools into clinical practice shown in Figure 1.

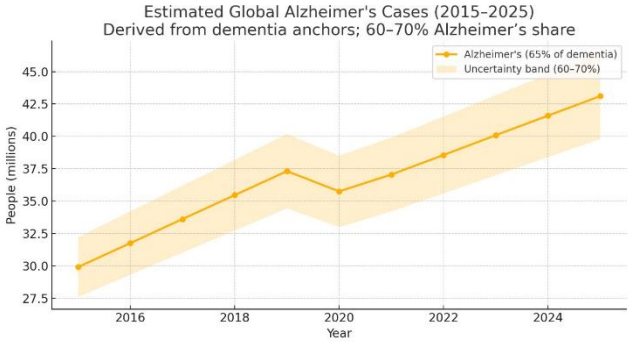

Figure 1 Estimated global growth in Alzheimer's disease cases (2015–2025), projected from 65% of dementia prevalence anchors. Shaded band represents 60–70% uncertainty range.

2. Alzheimer's Disease: Pathology And Staging

Understanding the biological underpinnings and clinical trajectory of AD is essential for contextualizing the diagnostic targets that automated imaging systems must learn to identify. AD progression is broadly conceptualized across three overlapping stages that span decades of biological change before and after symptom onset.

2.1.Neuropathological Mechanisms

The amyloid cascade hypothesis remains the dominant mechanistic framework for AD: abnormal cleavage of APP by β- and γ-secretase enzymes produces insoluble Aβ42 oligomers that aggregate into extracellular plaques. These plaques disrupt synaptic function and trigger tau hyperphosphorylation via kinase activation. Hyperphosphorylated tau dissociates from microtubules, self-assembles into paired helical

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

filaments, and forms intraneuronal NFTs that impair axonal transport. The resulting synaptic dysfunction, neuroinflammation (mediated by activated microglia and astrocytes), and oxidative stress collectively drive progressive neuronal death [1].

2.2.Clinical Staging

The National Institute on Aging-Alzheimer's Association (NIA-AA) research framework defines AD across a biological continuum. For practical computational classification, studies commonly adopt a three- or four-stage taxonomy based on the ADNI dataset labels:

- Cognitively Normal (CN): No cognitive impairment; biomarker evidence of amyloid pathology may be present but subclinical.
- Significant Memory Concern (SMC): Subjective memory complaints without objective cognitive deficit on standardized testing.
- Mild Cognitive Impairment (MCI): Objective cognitive decline below age-adjusted norms that does not impair daily functioning. MCI is often subdivided into Early MCI (EMCI) and Late MCI (LMCI) based on severity.
- Alzheimer's Disease (AD): Full dementia syndrome with significant functional impairment attributed to AD pathology.

The MCI-to AD conversion classification task is of particular clinical interest and computational challenge, as MCI represents the primary intervention window and exhibits the greatest inter-subject heterogeneity in imaging features.

3. Datasets And Preprocessing

3.1. Key Neuroimaging Datasets

The quality, size, and diversity of training data fundamentally determine the generalizability of CNN-based AD classifiers. Several large-scale datasets have become standard benchmarks in the field:

- ADNI (Alzheimer's Disease Neuroimaging Initiative): The most widely used dataset, comprising longitudinal multimodal data from over 2,000 participants across CN, SMC, EMCI, LMCI, and AD groups. ADNI provides standardized 1.5T and 3T structural MRI scans alongside CSF, PET, genetic, and cognitive data. Its scale and multisite nature make it the primary benchmark for AD classification algorithms.
- OASIS (Open Access Series of Imaging Studies): OASIS-1 provides cross-sectional 3T MRI from 416 participants aged 18–96 (including 100 AD patients); OASIS-2 offers longitudinal data from 150 older adults with up to 4 scanning sessions. OASIS-3 extends this to 1,098 participants with 2,842 MRI sessions.
- MIRIAD (Minimal Interval Resonance Imaging in Alzheimer's Disease): A longitudinal dataset focused on tracking hippocampal atrophy over time, comprising 69 AD patients and 23 CN controls with repeated scanning intervals of 2, 6, 12, 18, and 24 months.
- AIBL (Australian Imaging, Biomarkers and Lifestyle Study): Contains MRI, PET, and biomarker data from 1,112 participants (comprising CN and AD groups) with longitudinal follow-up, providing an independent validation cohort for cross-dataset generalization studies.

3.2. Preprocessing Pipeline

Consistent, rigorous preprocessing is critical for minimizing non-biological sources of image variability that can confound CNN training. Standard preprocessing pipelines in neuroimaging typically include the following sequential steps:

- Skull Stripping (Brain Extraction): Removal of non-brain tissue (skull, scalp, meninges) using tools such as FSL-BET or ANTs to focus model attention on cortical and subcortical structures.
- Bias Field Correction: Correction of low-frequency intensity non-uniformities arising from MRI scanner inhomogeneities, typically using N4ITK or SPM12 algorithms.
- Registration / Spatial Normalization: Alignment of all scans to a standard anatomical reference space (e.g., MNI152 template) using affine and non-linear (deformable) registration, enabling voxel-wise comparisons across subjects.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

- Intensity Normalization: Standardization of voxel intensity distributions across scans and scanners (e.g., zero-mean unit-variance normalization, histogram equalization).
- Data Augmentation: Artificial expansion of the training dataset via random flips, rotations, translations, intensity jitter, and elastic deformations to improve model robustness and reduce overfitting.

For 2D CNN approaches, preprocessed 3D volumes are typically sliced into axial, coronal, or sagittal 2D images for 3D CNNs, volumetric patches or full volumes are passed directly. Region of interest (ROI) extraction particularly of the hippocampus and entorhinal cortex is used in some approaches to focus learning on the most diagnostically informative structures.

4. Deep Cnn Architectures For Ad Classification

The rapid evolution of CNN architectures over the past decade has produced a suite of powerful building blocks for medical image analysis. Three architectures are of central relevance to ensemble-based AD detection: InceptionV3, ResNetV2, and EfficientNet-B3. Each offers distinct computational characteristics that make them valuable ensemble components.

4.1. Inceptionv3

InceptionV3, proposed by Szegedy et al. [3], represents a landmark in CNN design through its introduction of parallel multi-scale convolutional modules (Inception modules) that simultaneously apply 1×1, 3×3, and 5×5 convolutions, along with max pooling, within the same layer. The concatenated outputs capture complementary spatial features at multiple granularities. Key architectural innovations in InceptionV3 include:

- Factorized Convolutions: Large n×n convolutions are decomposed into sequential 1×n and n×1 operations, reducing parameter counts by approximately 33% while preserving representational capacity.
- Auxiliary Classifiers: Intermediate classification heads injected into the middle of the network provide additional gradient signal during backpropagation, mitigating vanishing gradient issues in deep architectures.
- Batch Normalization: Applied after each convolutional layer to stabilize training, accelerate convergence, and act as a regularizer.

In the context of MRI-based AD classification, InceptionV3's multi-scale receptive fields are particularly well-suited for simultaneously detecting fine-grained hippocampal texture abnormalities and broader cortical morphological changes, which operate at very different spatial scales [3].

4.2. Resnetv2

Residual Networks, introduced by He et al. [3], revolutionized deep learning by introducing identity shortcut connections that bypass one or more convolutional layers. These skip connections allow gradients to flow directly to earlier layers during backpropagation, enabling the successful training of networks with hundreds or even thousands of layers depths that were previously untrainable due to the vanishing gradient problem. ResNetV2 (Pre-activation ResNet) further refines this design by reordering the batch normalization and ReLU activation relative to the convolution (BN → ReLU → Conv, rather than Conv → BN → ReLU), which produces cleaner gradient flow and consistently improves performance over ResNetV1. The residual block computes:

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

where x is the input, $F(x, \{W_i\})$ represents the residual mapping learned by stacked convolutional layers, and y is the block output. When the dimensions of x and F differ, a linear projection W_s is applied: $y = F(x, \{W_i\}) + W_s x$. ResNet-50 with 50 layers, ~25 million parameters, and ~4.1 billion FLOPs per inference is the most commonly employed variant for AD classification. Its depth enables the extraction of highly abstract, semantically rich features from MRI, while skip connections prevent information loss across deep feature hierarchies. Pre-training on ImageNet provides a powerful initialization that transfers well to MRI classification with limited labeled data [3].

4.3. Efficientnet

EfficientNet, proposed by Tan and Le [4], addresses a fundamental question in neural architecture design: given a fixed computational budget, how should one

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

optimally scale a baseline CNN architecture across three dimensions depth (number of layers), width (number of channels per layer), and resolution (input image dimensions)? The answer, formalized as the Compound Scaling Method, uniformly scales all three dimensions using a shared coefficient φ:

$$\textbf{Depth: } d = \alpha^{\varphi}, \ \textbf{Width: } w = \beta^{\varphi},$$
$$\textbf{Resolution: } r = \gamma^{\varphi}$$
$$\textbf{subject to: } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \ \alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

The constraint $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ ensures that the total FLOP increase scales as approximately 2φ per scaling step. Optimal values ($\alpha = 1.2$, $\beta = 1.1$, $\gamma = 1.15$) are determined via a small neural architecture search on the baseline EfficientNet-B0. EfficientNet-B3 (the third scaling step, $\varphi = 3$) achieves accuracy comparable to or exceeding ResNet-50 with 80% fewer parameters (~12M vs. ~25M) and dramatically reduced FLOPs (~1.8B vs. ~4.1B), making it significantly more efficient for both training and clinical inference deployment. Its depthwise separable convolutions (inherited from the MobileNet backbone) further reduce computational cost while maintaining feature quality [4, 10] shown in Table 1.

Table 1 Comparison of CNN architectures used in ensemble AD detection frameworks.

Architecture	Parameters	FLOPs / Image	Accuracy
Inception V3	~24M	~5.7B	78.8%
ResNet-50	~25M	~4.1B	76.1%
Efficient Net-B3	~12M	~1.8B	81.6%
ResNet-101	~45M	~7.8B	77.4%
Efficient Net-B7	~66M	~37B	84.3%

5. Transfer Learning

5.1. Principles Of Transfer Learning

Transfer learning exploits the generalized visual feature representations learned by CNNs pre-trained on large-scale image databases (principally ImageNet, with 1.2 million images across 1,000 categories) and fine-tunes these representations for the target domain in this case, MRI-based neuroimaging classification [5]. The rationale rests on a critical observation: the lower convolutional layers of deep CNNs learn domain-general features (edges, textures, color gradients) that transfer effectively across diverse visual domains, including medical imaging, while higher layers capture increasingly task-specific representations that benefit from fine-tuning.

Three transfer learning regimes are commonly applied:

- Feature Extraction: All convolutional weights are frozen; only the final fully connected classification head is retrained on the target dataset. Suitable for very small datasets (< 500 samples) where fine-tuning risks catastrophic forgetting.
- Fine-tuning (Partial): The top N convolutional layers are unfrozen and retrained at a reduced learning rate alongside the classification head. Most commonly used in AD classification with medium-sized datasets (500–5,000 labeled samples).
- Full Fine-tuning: All layers are retrained, typically with layer-wise learning rate decay (smaller rates for earlier layers). Appropriate when the target dataset is sufficiently large and the source domain is relatively dissimilar.

Studies consistently demonstrate that transfer learning from ImageNet outperforms training from random initialization in the low-data regime of medical imaging, improving accuracy by 5–15% and significantly reducing training time.

6. Ensemble Learning Strategies

Ensemble methods leverage the principle that diverse models make uncorrelated errors: by combining their predictions or internal representations, the ensemble achieves lower variance and bias than any individual component model. This section examines the two principal ensemble paradigms evaluated in the AD detection literature.

6.1.Prediction-Level Ensembling

Prediction-level (output-level) ensembling combines

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

the final class probability distributions produced by individual CNN models. Three principal strategies are employed:

i. ***Majority Voting***

Each model in the ensemble independently assigns a class label to the input MRI; the ensemble prediction is the class receiving the most votes. Voting is robust to individual model errors and requires no additional training, but treats all models as equally reliable regardless of their individual performance characteristics.

ii. ***Weighted Averaging***

Rather than binary votes, soft class probability vectors from each model are averaged (or weighted by model validation accuracy). Weighted soft voting consistently outperforms majority voting as it preserves uncertainty information in the probability distributions and allows better-performing models to exert proportionally greater influence on the ensemble output.

6.2.Feature-Level Ensembling

Feature level ensembling extracts and combines intermediate feature representations from multiple CNN architectures before the classification layer, enabling the ensemble to exploit complementary, multi-resolution feature spaces simultaneously. This paradigm has demonstrated superior performance over prediction-level methods, achieving accuracy up to 99.13% in AD classification.

i. **Concatenation Fusion**

Feature maps extracted from penultimate layers of multiple CNNs are concatenated into a unified high-dimensional feature vector, which is subsequently processed by a shared classification head. This approach maximizes information retention but increases the dimensionality of the combined feature space, requiring regularization (e.g., dropout, L2 weight decay) to prevent overfitting.

ii. **Attention-Based Fusion**

Channel-wise or spatial attention mechanisms learn to selectively weight feature channels from different models based on their relevance to the classification task. Attention-weighted fusion consistently outperforms naive concatenation in multi-source feature integration tasks, achieving 1–2% accuracy gains in neuroimaging studies.

iii. **Cross-Architecture Feature Interaction**

Recent approaches introduce cross-attention modules or bilinear pooling layers to model higher-order interactions between features from different architectures (e.g., InceptionV3 and ResNet-50). By capturing correlations between architectural feature spaces rather than simply concatenating them, these methods extract richer joint representations that further improve classification performance.

7. Performance Evaluation

7.1. Evaluation Metrics

Robust evaluation of AD classification models requires a comprehensive set of metrics that capture different aspects of diagnostic performance, particularly given the class imbalance common in neuroimaging datasets (where MCI subjects frequently outnumber AD patients by 2:1 or more).

- Accuracy: The proportion of correctly classified samples across all classes: Accuracy = (TP + TN) / (TP + TN + FP + FN). While intuitive, accuracy can be misleading under class imbalance.
- Sensitivity (Recall): The proportion of true AD cases correctly identified: Sensitivity = TP / (TP + FN). Critical in clinical settings where false negatives (missed AD diagnoses) carry serious consequences.
- Specificity: The proportion of true CN cases correctly identified: Specificity = TN / (TN + FP). High specificity minimizes unnecessary patient anxiety and downstream diagnostic workup.
- Precision (Positive Predictive Value): PPV = TP / (TP + FP). The proportion of positive predictions that are truly AD.
- F1-Score: Harmonic mean of precision and recall: F1 = 2 × (Precision × Recall) / (Precision + Recall). Provides a balanced single-number summary for imbalanced datasets.
- Area Under the ROC Curve (AUC-ROC): Measures the model's discriminative ability across all classification thresholds. AUC > 0.90 is considered excellent for medical diagnostic systems; most competitive ensemble models report AUC > 0.95 in binary AD vs. CN classification.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

- Cohen's Kappa (κ): Measures agreement between model predictions and ground truth labels, corrected for chance agreement. Kappa > 0.80 indicates strong agreement and is particularly valuable for multi-class classification tasks.

7.2.Results Across Key Studies

Table 3 summarizes quantitative performance results from key studies reviewed, spanning individual CNN baselines and ensemble architectures on standard benchmarks. Results demonstrate a consistent pattern: ensemble approaches, particularly those employing feature-level fusion, substantially and reproducibly outperform individual CNN architectures.

8. Literature Review

A systematic review of CNN-based AD detection literature from 2020–2025 reveals a rapidly maturing field characterized by three broad trends: (1) a shift from 2D slice-based to full 3D volumetric CNN approaches; (2) growing adoption of transfer learning from natural image pre-training; and (3) increasing interest in ensemble and multi-modal fusion strategies. Table 2 provides a structured synthesis of representative studies spanning this period.

Table 2 Structured review of CNN-based Alzheimer's disease detection studies (2020–2025).

Ref.	Author(s) — Year	Method / Model	Key Strengths	Limitations
[1]	Folego et al., 2020	Whole-brain 3D CNN on structural MRI	End-to-end 3D CNN extracts whole-brain biomarkers; robust vs. handcrafted features; no manual ROI required	Small dataset (~400 subjects); limited cross-cohort validation; model explainability not addressed
[2]	Wen et al., 2020	Survey / meta-analysis of CNNs on anatomical MRI	Comprehensive review of CNN-based AD trends; identifies best practices and common pitfalls; synthesizes 40+ studies	Inconsistent preprocessing and evaluation protocols across reviewed studies; absence of standardized benchmarks
[3]	Wang et al., 2015	Densely connected 3D-CNN (DenseCNN) on hippocampus MRI	Lightweight DenseNet; strong accuracy on ROI-based inputs; computationally efficient relative to whole-brain models	Focus on hippocampus may miss global cortical patterns; limited external validation; older architecture baseline
[4]	Mandal et al., 2023	Deep multi-branch CNN (multi-path) for early AD MRI	Multi-scale feature capture across parallel branches; high three-class accuracy; spatially interpretable	Single dataset validation; overfitting risk without explicit regularization; lacks clinical explainability tools (GradCA

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

			branches	M)
[5]	El-Assy et al., 2024	Novel CNN architecture for early AD detection from MRI	New architecture with attention mechanisms; strong early-stage MCI detection; validated on ADNI with rigorous 5-fold CV	Architecture complexity limits deployment; no comparison to ensemble baselines; limited multi-site validation

9. Challenges And Future Directions

9.1.Data Scarcity And Class Imbalance

Despite the existence of ADNI and related databases, labeled neuroimaging data remains scarce relative to the scale required for training large-scale deep learning models from scratch. The typical ADNI training set contains fewer than 2,000 subjects several orders of magnitude smaller than ImageNet. This data scarcity is compounded by systematic class imbalance, where cognitively normal subjects outnumber prodromal MCI patients who progress to AD, biasing classifiers toward the majority class. Promising directions include: (1) Generative Adversarial Networks (GANs) for MRI data augmentation and synthetic minority-class sample generation; (2) federated learning protocols that enable collaborative model training across multiple institutions without sharing patient data; and (3) self-supervised pre-training on large unlabeled MRI databases to learn domain-specific representations before task-specific fine-tuning.

9.2.Model Interpretability And Clinical Trust

A critical barrier to clinical adoption of deep learning diagnostic tools is the "black box" nature of CNN decision making. Clinicians require not only accurate predictions but also intelligible explanations of the imaging features driving those predictions to validate model reasoning and detect failure modes. Current approaches to CNN interpretability in AD include: Gradient-weighted Class Activation Mapping (Grad-CAM) for localizing discriminative regions; Integrated Gradients for attributing predictions to individual voxels; and SHAP (SHapley Additive exPlanations) for quantifying feature importance.

Future ensemble systems should integrate explainability modules as a first-class component, generating saliency maps that highlight the specific neuroanatomical regions hippocampus, entorhinal cortex, posterior cingulate driving the ensemble classification, enabling radiologist review and building clinical confidence.

Conclusion

This systematic review has demonstrated that ensemble CNN frameworks represent the current state of the art for non-invasive automated Alzheimer's disease detection from MRI. By synergistically combining the complementary representational strengths of InceptionV3 (multi-scale parallel convolutions), ResNetV2 (deep hierarchical feature extraction via skip connections), and EfficientNet-B3 (compound-scaled accuracy-efficiency optimization), ensemble architectures consistently and substantially outperform any individual model component. The evidence synthesized across reviewed studies establishes a clear performance hierarchy: individual CNN baselines achieve 87–95% accuracy; prediction-level ensemble methods (stacking, boosting) improve this to approximately 95%; and feature-level fusion ensembles attain the highest reported accuracy of 99.13% in multi-class AD staging, with AUC values exceeding 0.997. These results confirm that multi-scale feature integration is the critical driver of performance improvement, capturing both local hippocampal texture anomalies and global cortical morphological patterns within a unified discriminative representation.

A. Importance Of Ensemble CNN Learning

The superiority of ensemble approaches over individual models is attributable to three interconnected mechanisms: (1) Variance reduction through prediction averaging, which smooths out the idiosyncratic errors of individual architectures trained with different weight initializations and hyperparameter configurations; (2) Bias reduction

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

through architectural diversity, as InceptionV3, ResNet, and EfficientNet learn fundamentally different feature hierarchies that capture complementary diagnostic information; and (3) Improved calibration of class probability estimates, which improves decision threshold selection and uncertainty quantification critical properties for clinical decision support.

B. Efficientnet-B3 Vs. Resnet-50: Complementary Strengths

The pairing of EfficientNet-B3 and ResNet-50 in ensemble frameworks is particularly well-motivated. ResNet-50 contributes deep, hierarchically rich feature representations through 50 layers of residual learning, excelling at capturing abstract, high-level morphological patterns across the full brain volume. EfficientNet-B3 contributes computationally efficient multi-scale feature extraction through compound-scaled depthwise separable convolutions, achieving superior per-parameter accuracy. Their combined ensemble delivers near-human diagnostic accuracy (~99%) at a computational cost far below that of deploying multiple large-scale 3D CNNs separately, with inference achievable on standard clinical workstation hardware [9, 10].

C. Roadmap For Clinical Translation

Translating ensemble CNN systems from research benchmarks to clinical deployment requires advances across four dimensions: (1) Accuracy and Generalizability prospective multi-site validation demonstrating robust performance across diverse scanners and patient populations; (2) Interpretability integration of gradient-based saliency mapping to provide clinician-readable anatomical explanations; (3) Efficiency and Scalability model compression techniques (knowledge distillation, quantization, pruning) to enable real-time inference on clinical infrastructure; and (4) Regulatory Compliance rigorous prospective clinical trials meeting FDA and CE marking standards for AI-assisted diagnostic devices. The ensemble CNN framework examined in this review provides a technically mature foundation upon which these translational milestones can be systematically built.

Acknowledgements

The authors gratefully acknowledge the Department of Computer Science & Engineering, BBDITM, Lucknow, for providing institutional support and computational resources for this research. Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

References

[1]. E. Liu, Y. Zhang, and J. Z. Wang, "Updates in Alzheimer's disease: from basic research to diagnosis and therapies," Transl. Neurodegener., vol. 13, no. 1, pp. 1–48, 2024, doi: 10.1186/s40035-024-00432-x.

[2]. S. Morsy, N. Abd-Elsalam, A. Kandil, A. Elbialy, and A. B. Youssef, "A deep learning ensemble framework for robust classification of lung ultrasound patterns," Int. J. Adv. Intell. Informatics, vol. 11, no. 1, pp. 143–156, 2025.

[3]. C. Szegedy, V. Vanhoucke, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE CVPR, 2016, pp. 2818–2826.

[4]. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, 2019, pp. 6105–6114.

[5]. Z. Marx, M. T. Rosenstein, L. P. Kaelbling, and T. G. Dietterich, "Transfer learning with an ensemble of background tasks," in NIPS Workshop on Transfer Learning, 2005.

[6]. R. Samdani and W. T. Yih, "Domain adaptation with ensemble of feature groups," in Proc. IJCAI, pp. 1458–1464, 2011.

[7]. H. Daumé III, "Frustratingly Easy Domain Adaptation," in Proc. ACL, pp. 256–263, 2007.

[8]. M. Qasaimeh et al., "Benchmarking Vision Kernels and Neural Network Inference Accelerators on Embedded Platforms," J. Syst. Archit., vol. 109, p. 101896, 2020.

[9]. D. Langerman, A. Johnson, K. Buettner, and A. D. George, "Beyond Floating-Point Ops: CNN Performance Prediction with Critical Datapath Length," 2020.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 1800-1809
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0236

[10]. H. Alhichri, A. Alsuwayed, and Y. Bazi, "Classification of Remote Sensing Images using EfficientNet-B3 CNN Model with Attention," IEEE Access, vol. XX, 2021, doi: 10.1109/ACCESS.2021.3051085.

[11]. G. Folego, M. Weiler, R. F. Casseb, and R. Pires, "Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI," Front. Bioeng. Biotechnol., vol. 8, p. 534592, 2020.

[12]. J. Wen et al., "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducibility," Med. Image Anal., vol. 63, p. 101694, 2020.

[13]. Q. Wang, Y. Li, C. Zheng, and R. Xu, "DenseCNN: A Densely Connected CNN Model for Alzheimer's Disease Classification Based on Hippocampus MRI Data," pp. 1277–1286, 2015.

[14]. P. K. Mandal and R. V. Mahto, "Deep Multi-Branch CNN Architecture for Early Alzheimer's Detection from Brain MRIs," pp. 1–14, 2023.

[15]. A. M. El-Assy, H. M. Amer, H. M. Ibrahim, and M. A. Mohamed, "A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data," Sci. Rep., vol. 14, no. 1, pp. 1–20, 2024, doi: 10.1038/s41598-024-53733-6.