

Usage of Deep Ensemble Learning in Classification of Alzheimer’s Disease: A Systematic Review

Fahad Amin¹, Saad Arsalan Wasti¹, Areesha Wasti², Maham Naveed¹, Tahzeeb Afzal¹

¹Department of Medicine, King Edward Medical University, Lahore 54000, Pakistan.

²Department of Medicine, Continental Medical College, Lahore 54760, Pakistan.

INTRODUCTION	METHODS	RESULTS	CONCLUSIONS
Alzheimer’s disease is a progressive neurodegenerative disorder marked by irreversible cognitive deterioration, where timely and accurate diagnosis becomes a key focus of current research. Ensemble deep learning methods combine the strengths of individual networks to enhance predictive performance and stability in AD diagnosis.	This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. A comprehensive literature search was conducted on PubMed, Scopus, Embase, Web of Science, and IEEE Xplore from inception to December 31, 2025, to identify publications evaluating deep ensemble learning architectures for the classification and diagnosis of Alzheimer’s disease. Data extraction focused on study characteristics, imaging modalities (predominantly structural MRI), dataset utilization (e.g., ADNI, OASIS), ensemble strategies (homogeneous vs. heterogeneous), and key performance metrics, including diagnostic accuracy, sensitivity, specificity, and area under the curve (AUC).	A total of 24 studies were included. Most included studies addressed diagnostic or classification tasks using structural MRI, with public datasets such as ADNI and OASIS most frequently utilized. Sample sizes ranged from 24 to 86,390. Both homogeneous and heterogeneous ensemble architectures were reported, typically combining multiple base models (2-8) via voting, weighting, or stacking. Reported classification accuracies ranged from approximately 86% to 99%, with area under the curve (AUC) values commonly exceeding 0.93 when reported. Sensitivity and specificity when reported typically exceeded 90% across CN, MCI, and AD classifications. In most studies, ensemble models outperformed single deep learning classifiers, although external validation was infrequently performed.	Ensemble approaches demonstrate high accuracy and reliability for AD diagnosis and classification. The improved control of false positives and false negatives enhances confidence in early detection and disease staging, facilitating prompt intervention and better patient management. Overall, ensemble approaches reduced misclassification and improved class discrimination compared with single-model methods. MRI, remains the most adopted modality owing to its cost-effectiveness, reproducibility and detailed brain imaging, enabling robust feature extraction across diverse models. Clinical translation is constrained by dataset heterogeneity, limited external validation, and risks of overfitting, underscoring the need for large-scale, standardized investigations.