International Journal of DEEPTECH IN MEDICAL SCIENCE & TECHNOLOGY

An Edge-Deployed Explainable AI Framework with RAG-Augmented Large Language Models for Multi-Stage Alzheimer's Disease Staging and Conversational Decision Support

Saud Ahmad Khan[1], Saksham Srivastava[1], Himadri Srivastava[1] and Nishita Singh[1],

, Rishi Jain[2], Aakash Kuragayala[2]

1. United College of Engineering and Research, Prayagraj, India
2. Asian Institute of Technology, Bangkok, Thailand

Corresponding author email: rjain@g.ecc.u-tokyo.ac.jp

Abstract: Alzheimer's disease (AD) demands precise multi-stage classification and interpretable diagnostic workflows to support timely clinical intervention and personalised patient management. While deep learning models have shown strong predictive capabilities, conventional vision architectures frequently operate as uninterpretable "black boxes," failing to highlight neuroanatomical regions of interest or provide contextually grounded, actionable clinical guidance for clinicians, patients, and caregivers. To address these critical gaps, this paper presents a novel, edge-integrated preliminary decision-support framework that unifies deep convolutional neural networks, explainable artificial intelligence (XAI), and retrieval-augmented conversational intelligence for end-to-end Alzheimer's disease assistance. The computer vision pipeline utilises a fine-tuned VGG16 architecture trained to classify structural magnetic resonance imaging (MRI) scans across distinct stages of AD progression, ranging from non-demented to moderate cognitive impairment. To eliminate diagnostic opacity, Gradient-Weighted Class Activation Mapping (Grad-CAM) is embedded into the neural network pipeline to generate visual heatmaps, precisely isolating affected focal regions and localised neuropathological atrophy patterns within brain tissue. Beyond visual staging, the system incorporates an edge-deployed Llama 3.2 Large Language Model (LLM) integrated with a Retrieval-Augmented Generation (RAG) framework. This local RAG pipeline is domain-grounded using clinical guidelines established by the Alzheimer's Association, enabling privacy-preserving, low-latency execution directly on local edge hardware without reliance on external cloud services or compromising sensitive patient health data. Experimental evaluation demonstrates that the proposed VGG16 classifier achieves a multi-class staging accuracy of 90.0%, showing robust diagnostic feature extraction across severity levels. The Grad-CAM visual explanations effectively corroborate automated predictions with established neuroimaging biomarkers, significantly enhancing model interpretability and clinician trust. Concurrently, the integrated Llama 3.2 RAG conversational agent provides a multi-functional user interface, enabling real-time interactive query resolution and automated generation of structured, evidence-grounded diagnostic reports. By synergising deep learning classification, visual XAI, localised LLM reasoning, and domain-specific knowledge bases, this comprehensive edge AI architecture offers an interpretable, secure, and highly accessible preliminary diagnostic tool for modern neurodegenerative disease management.

Keywords: Alzheimer's Disease, Explainable AI (XAI), Large Language Models (LLM), Retrieval-Augmented Generation (RAG), Edge AI, Clinical Decision Support System (CDSS).

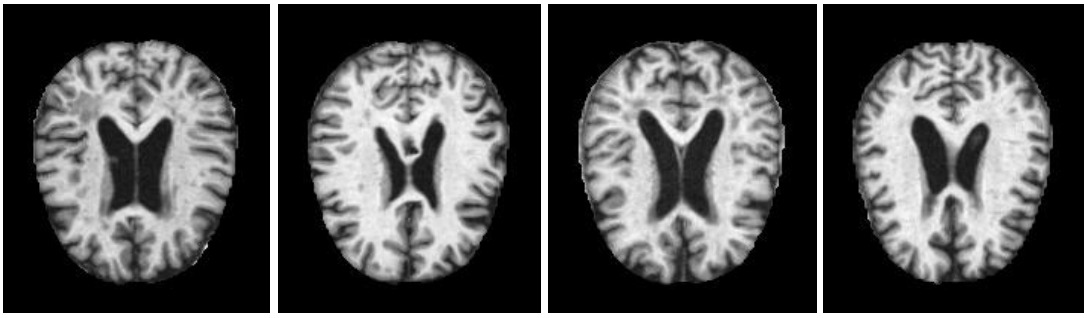
1. Introduction

Alzheimer is a progressive neurodegenerative disorder which first appeared in 1901 when a German psychiatrist, Alois Alzheimer examined a 51 years old patient, Auguste Deter at a Frankfurt Asylum, he figured out an unusual symptom which is not familiar to her age like short term memory loss, paranoia and cognitive decline. The disease Alzheimer was named in 1910 after the post-mortem of brain tissue that showed extracellular protein clumps and intracellular twisted fibres when both build up in the brain, forming sticky plaques and twisted tangles that destroy brain cells as explained in[1], [2]. But later it showed long dismissed as a rare disorder confined to younger patients, the disease was again studied in mid- century neurologist Robert Katzman's pivotal 1976 argument recognizing senile dementia and Alzheimer's as biologically the same condition. Subsequent molecular breakthroughs isolated

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

beta-amyloid in 1984, identified tau in 1986 including APP mutations and the APOE-ϵ4 risk allele. While early treatments like donepezil provided only symptomatic relief, modern PET imaging, fluid biomarkers, and monoclonal antibody therapies have shifted the field towards the actively slowing disease progression.[3], [4]



(a) Non-Demented | (b) Very Mild Demented | (c) Mild Demented | (d) Moderate Demented

Fig 1: Multi-Stage Alzheimer's Disease Staging (Alzheimer Kaggle Dataset)

Figure 1 illustrates MRI images reflecting the severity of AD a person may suffer from, which correlates with protein accumulation and depends on many factors, such as age. Early impairment in forming new memories can progress to language loss, spatial disorientation, executive dysfunction, and frequent behavioral changes, ultimately leading to the loss of basic motor functions such as walking and swallowing in its final stages.

Alzheimer's disease does not appear suddenly it first gradually shows mild changes, typically described in three stages depending on the clinical framework used. The most widely adopted model, based on the [8], divides progression into three broad phases mild (early), moderate (middle), and severe (late) each marked by a distinct pattern of cognitive, functional, and behavioral decline. Table 1 summarizes the defining features of each stage.[5], [6], [7], [8]

Table 1.0 Clinical Stages of Alzheimer's Disease

Stage	Duration (approx.)	Cognitive Symptoms	Functional Impact	Behaviour Changes	References & Primary Sources
Mild (Early)	2–4 years	Occasional word-finding difficulty, mild short-term memory lapses, difficulty planning or organizing.	Can largely live independently; minor lapses in managing finances or appointments.	Mild anxiety, occasional mood changes, awareness of decline.	[1] [10]
Moderate (Middle)	2–10 years	Increased memory loss, confusion about time/place, difficulty recognizing familiar people.	Needs assistance with daily activities (dressing, cooking); poor judgment.	Agitation, wandering, sleep disturbances, suspicion or paranoia.	[11]
Severe (Late)	1–3+ years	Severe memory loss, minimal verbal communication,	Full dependency for all daily care (feeding, bathing, mobility).	Apraxia, loss of motor control, swallowing difficulty, increased	[12]

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

		loss of awareness of surroundings.		vulnerability to infection.	

The scale of Alzheimer's disease today makes clear why early, reliable diagnosis has become such an urgent priority. According to the World Health Organisation (WHO), an estimated 57 million people were living with dementia worldwide, with more than 60% residing in low and middle income countries, regions where diagnostic and care infrastructure is often least equipped to handle the burden. Alzheimer's disease is the leading cause of this global dementia burden, generally accounting for the majority of diagnosed cases.

In the United States specifically, the Alzheimer Association facts and figures report places the number even higher within the older population: an estimated 7.4 million Americans aged 65 and older are living with Alzheimer's in 2026, with about 1 in 9 people in this age group affected, and nearly three-quarters of those diagnosed being 75 or older. The disease also affects populations unevenly. Almost two-thirds of Americans living with Alzheimer, Americans are roughly twice as likely to be diagnosed as older people, while other faces about one and a half times the risk[13]. Younger-onset cases, though far less common, are not negligible either the researchers estimate that roughly 200,000 Americans under 65 are living with younger-onset dementia.

The economic toll mirrors this human burden. Health and long-term care costs for people living with Alzheimer's and other dementias in the United States are projected to reach 409 billion dollars in 2026 alone, with total costs expected to approach 1 trillion dollars by 2050. These figures do not even account for the value of unpaid caregiving provided by family members, which represents a substantial hidden cost borne largely outside the formal healthcare system.

Alzheimer's disease quietly changes how a person thinks, remembers, and functions, often starting years before anyone clearly notices the signs. By the time family members or doctors recognize a problem, the disease may have already progressed significantly, narrowing the window for early support and care planning. Access to specialists who can read brain scans and confidently judge how advanced the condition remains limited in many places, and even where such expertise exists, getting a timely appointment can take weeks or months. This gap between when changes begin in the brain and when they are actually caught and understood is one of the most pressing challenges in caring for people affected by this disease, and it is the real-world problem this project sets out to address.

The goal of this work is to build a tool that can look at a brain scan and give an early, understandable indication of where a person may stand—whether their brain shows signs of concern, very early changes, or more noticeable ones—and to explain that judgment in plain language rather than as a mysterious verdict. Beyond just answering, the system is designed to show which part of the scan influenced its decision so the result can be checked and trusted, and to let a person ask follow-up questions to get answers grounded in real, verified information about the disease. The intention is not to replace a doctor's judgment, but to offer a fast, accessible, and transparent first step: something that could support earlier conversations, earlier care planning, and greater peace of mind for families navigating an uncertain and often frightening diagnosis.[14]

2. Literature Review

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

The literature review provides a comprehensive overview of the diverse approaches employed by various authors in predicting alzheimer disease stages, encompassing both deep learning and large language model.

The author [1] trace the origins of alzheimer's disease as a medical concept, arguing that understanding its past is essential to understand its present and future. They reviewed the etymology and history of dementia, how dementia was perceived historically, and the developments that led to the coining of the term "alzheimer's disease." A central event they cover is Alois Alzheimer's 1906 announcement that abnormal symptoms he observed originated in the cerebral cortex, based on his case study of Auguste D., a 51year old woman, presented at a psychiatry conference in Tübingen. author's [15] core argument is that alzheimer's and senile dementia are the same disease, and that recognizing this reveals alzheimer disease to be a far more prevalent and deadly killer than official statistics acknowledged a claim that reshaped how the medical community classified and prioritized the disease .

While the authors [16] and [17] noted that although no cure exists for alzheimer's disease, current drugs such as donepezil and memantine offer symptomatic relief, while CSF biomarkers combined with amyloid positron emission tomography imaging supported earlier, more accurate diagnosis. Complementing this, authors together characterize alzheimer disease as a progressive neurodegenerative disorder originating in the entorhinal cortex, emphasizing that early symptom recognition and coordinated, evidence-based clinical management remain the most effective response given the disease's irreversible course. Together, these works underscore that meaningful progress against alzheimer disease depends on earlier diagnosis through biomarkers and imaging, paired with better-coordinated care.

[18] Kim et al. found that alzheimer's patients who took their medication consistently and at the right dose had better long-term outcomes and lived longer. [2] Hajjar et al. discovered that certain blood/fluid proteins linked to blood vessel health rise even before symptoms appear, and could help predict who will get worse. [19]Feng describes a strange pattern where cancer patients are less likely to get alzheimer's, and vice versa, suggesting the two diseases might share some biological connection worth studying.

The author in [20] surveys vision transformers specifically for image classification, starting from the fundamentals of transformers and the original vision transformers, then reviews the subsequent improvements and optimizations built on vision transformers and compares their strengths and limitations against the classic convolutional neural network through experiments. while author[21] concludes that a systematic review mapping the vision transformers landscape for alzheimer disease diagnosis, concluding the field is shifting toward hybrid architectures but still lacks focus on progression prediction and reliable reproducibility .

Author [22] apply a deep neural network to classify alzheimer's from medical images, arguing that traditional feature-extraction methods have plateaued and that deep learning offers a more scalable path to accurate classification. Complementing this,[23] a review of conventional machine learning and deep learning methods in alzheimer disease neuroimaging diagnosis surveys both classical algorithms (support vector machine, random forest) and deep architectures (convolutional neural network, autoencoder, transformer), concluding that model choice must be tailored to the data since no single approach dominates and challenges like class imbalance persist. Work on [24]four-stage dementia detection extends binary classification into finer-grained severity staging, using convolutional neural network based models to distinguish

CC BY

non-demented through moderate demented cases Amin-Naji et al[25]. take a distinct architectural route, using a siamese convolutional neural network to diagnose from structural magnetic resonance imaging, achieving accuracy on the dataset and showing that similarity-learning architectures can rival standard convolutional neural network classifiers.

Finally,[14]broader machine learning and deep learning prediction studies confirms that deep learning applied directly to neuroimaging without manual feature engineering, especially when combined with additional biomarkers.[26] Fareed et al. propose alzheimer disease detection network, a lightweight custom convolutional neural network architecture built to work well even on smaller magnetic resonance imaging datasets, aiming for an efficient alternative to heavier pretrained models. Benchmarked against DenseNet169, VGG19, and InceptionResNetV2, VGG16 achieved 91.4% accuracy and 91% F1-score, outperforming all compared models across every evaluation metric. The model also generates class activation heatmaps on brain scans, adding interpretability to its early-stage alzheimer disease classification Feng et al.

[27] show that alzheimer's diagnosis improves when a pre-trained large language model is used to incorporate non-imaging data patient information, genetic data, medication history, cognitive tests alongside imaging, since existing models mostly relied on multi-modal image data alone and couldn't fully exploit non-image data; their approach achieved state of the art results on the dataset.

[28]Lahiri and Hu take a different angle with alzheimer retrieval augmented generation, a multimodal retrieval augmented generation application for clinical use focused on alzheimer's disease case studies from the articles, using cross-modal attention fusion to integrate textual and visual data essentially building a retrieval system to make biomedical literature more accessible, rather than a diagnostic classifier.[29] Zheng et al. instead target detection directly through language: a non-invasive, speech based method that fine tunes a pretrained language model (large language model meta artificial intelligence) using prompt tuning and improvement over a bidirectional encoder representations from transformers baseline.[30]Huang et al.'s alzheimer's disease analysis model, generation 1 goes further into system design, proposing a multi-agent large language model framework that integrates clinical, microbiome, and literature data via retrieval augmented generation, using specialized artificial intelligence agents for classification and synthesis and reporting comparable or improved F1 scores against extreme gradient boosting., with notably reduced variance.

Finally, the mass general brigham team's dementia electronic health record paper focuses on real world scalability: a retrieval augmented generation based large language model pipeline evaluated on real electronic health record data that improved dementia ascertainment compared to traditional international classification of diseases code and keyword-based methods, using both open-source models (Llama, Gemma) run locally and commercial models (GPT-4o) for extracting dementia status from unstructured clinical notes .

Authors [13], [31], [32] across these three, artificial intelligence plays three distinct roles reshaping the field broadly as a diagnostic and research accelerator, guiding discovery of biological mechanisms in models, and automated expert-level visual scoring tasks that traditionally required judgment together illustrating how artificial intelligence role in alzheimer disease research now spans prediction, mechanism discovery, and clinical workflow automation, not just image classification. [33]The case-control study found that things like loneliness, chronic illness, and family history raise alzheimer's risk, while exercise and social

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

activity seem to lower it. Lastly,[34] the treatment-patterns study showed that in real life, over half of patients switch or stop their Alzheimer's medication within six months showing a gap between what's recommended and what actually happens.

The literature indicates substantial progress in Alzheimer's disease diagnosis using machine learning, deep learning, Vision Transformers, multimodal learning, and Large Language Models. However, there remains a significant gap in developing an integrated, explainable, multimodal diagnostic framework that combines Vision Transformer-based MRI feature extraction, Retrieval-Augmented Generation for evidence retrieval, and Large Language Models for clinically interpretable diagnostic reasoning.

3. Methodology

The proposed system follows a five-stage pipeline that integrates deep learning-based classification with integration of a Large Language Model (LLM) enhanced by Retrieval Augmented Generation (RAG) to provide interpretable, actionable outputs for Alzheimer stage detection as shown in Figure 2. The system architecture includes five stages: dataset acquisition, data preprocessing, neural network classification, LLM-based diagnosis generation with RAG, and output generation.[35]

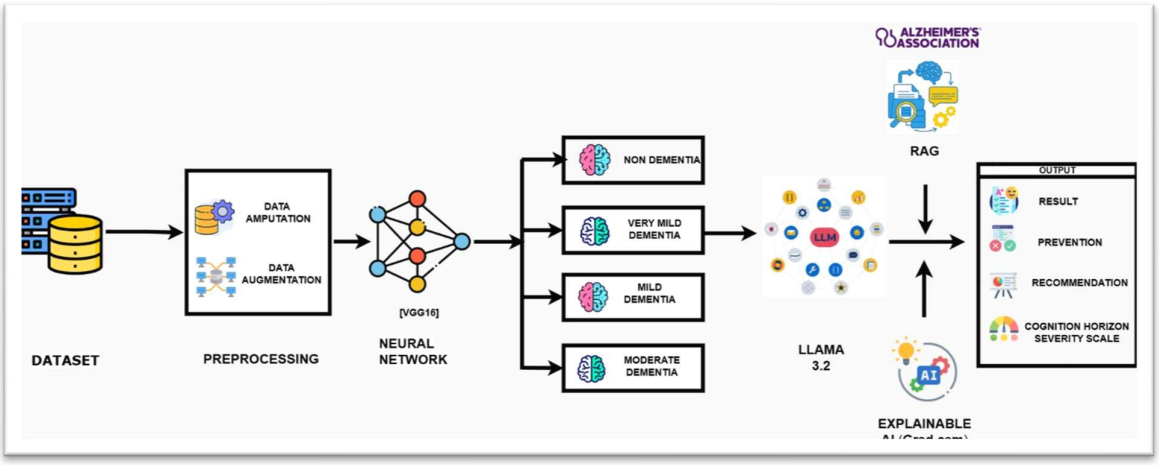


Fig. 2.0: Proposed methodology pipeline

3.1 Dataset

In this study, the dataset comprises of brain Magnetic Resonance Imaging (MRI) scans that are organized into their respective class labels, forming the input of the deep learning model. The classes represent the progressive stages of Alzheimer's type dementia[32]:

- Non-Dementia (3200 images count)
- Very Mild Dementia (8960 images count)
- Mild Dementia (869 images count)
- Moderate Dementia (64 images count)

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

Table 2.0 Dataset information

Attribute	Detail
Dataset Source	Kaggle-Alzheimer's Multiclass Dataset
Total Sample Size	379 MB
MRI Image Size	190*200
File Format	JPEG

3.2 Data preprocessing

Before model training, the raw MRI dataset underwent a preprocessing pipeline consisting of two sequential sub-processes to improve data quality and increase model generalisation.[36]

1. Data Amputation (Cleaning and Reduction): This step focused on eliminating low-quality, redundant, or noisy entries. Corrupted or unreadable images were discarded. Near-identical or exact duplicate scans were filtered out using a random-based model to prevent data leakage between training and testing splits. All retained images were then resized to a fixed input dimension of 224×224 pixels to match the VGG16 input requirements, and pixel intensity values were normalized to the range [0,1].
2. Data Augmentation**:** To address class imbalance, data augmentation artificially expanded the training set, improved model generalisation, and reduced overfitting. Minority classes were augmented at a higher rate to achieve a near-uniform class distribution, mitigating classifier bias. Yes, Test-Time Augmentation (TTA)—specifically horizontal flipping—was prepared for the evaluation phase to smooth single-view prediction noise.

3.3 Neural Network Classification This study employs the VGG16 Convolutional Neural Network (CNN) architecture for feature extraction and classification.

1. Model Architecture & Fine-Tuning: The standard VGG16 base consists of 13 convolutional layers across 5 blocks, followed by max-pooling layers. The original 1000-class output layer was replaced with a custom classification head featuring a global average pooling layer, fully connected dense layers with ReLU activation, dropout regularisation, and a final dense layer with softmax activation for the four target classes.
2. Training Configuration: The model was trained with a batch size of 32 over 50 epochs utilising weighted cross-entropy loss. A dual learning rate strategy was applied: a higher rate of 1e-3 was used for the custom classification head, while a lower rate of 1e-5 was used to fine-tune the unfrozen convolutional blocks. The system utilised hardware featuring a GPU with 2560 CUDA cores,320 Tensor cores, and 16GB VRAM.

Table 3.0 Overall performance

Input Shape	224×224×3
learning rate	1e-3, 1e-5
batch size	32

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

number of epochs	50
CUDA cores	2560 GPU
Tensor cores	320
VRAM	16GB
Loss function	Weighted cross entropy

3.4 LLM-Based Diagnosis Generation with RAG: The predicted class is passed into the LLM module responsible for generating a natural-language, clinically contextualized diagnosis[27]. The predicted class was further passed into a Large Language Model (LLM) module responsible for generation of a natural-language, clinically contextualized result that consist of knowledge base construction which is an accurate knowledge base built from the Alzheimer's Association, retrieval and generation process includes the LLM (llama 3.2). The augmented prompt is sent to Llama, which processes the context to generate the output and conversational responses for the chat interface.[30], [37]

3.5 Output Generation The final output presented to the end user is generated by the RAG-enhanced LLM and consists of the following components:

- Result & Cognition Horizon: The predicted stage of dementia, alongside a concise explanation of its clinical significance and a visual severity scale to aid user comprehension.
- Prevention: Evidence-based, stage-specific strategies for reducing risk and slowing disease progression, retrieved directly from the Alzheimer's Association knowledge base.
- Recommendation: Practical, Personalized guidance, including lifestyle modifications and advice to consult a qualified healthcare professional.
- Interactive Chatbot: A conversational module enabling users to ask follow-up questions in real time.

4. Result

To quantify the various deep neural network architectures and also their performance on each of the category, Precision, Recall, F1-score, and Support metrics were employed, as shown in Table 4. And Table 5.

Table 4.0 Comparison table of models used

CNN Models	Test Accuracy	Precision	F1-score	Recall
VGG16	91.41%	91%	91%	91%
DenseNet121	85.75%	86%	85%	86%
ResNet50	82.28%	87%	87%	88%
EfficientNetV2S	90.32%	91%	91%	91%

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

Table 5.0 Model performance is assessed on the held-out test set using classification metrics:

	Precision	Recall	F1-score	Support
NonDemented	0.81	0.87	0.84	144
VeryMildDemented	0.88	0.81	0.81	151
MildDemented	0.97	0.97	0.97	150
ModerateDemented	0.99	1.00	1.00	149
Accuracy			0.91	594

4.1 Quantitative Classification Performance

The fine-tuned VGG16 model achieved an overall test accuracy of 91.4% on a held-out set of 594 images, evaluated using four stages in which all augmented copies of a given source scan were confined to a single partition (train (85) and test (15)). This result reflects two refinements over an earlier iteration of the model (which reached 91.4% under the same evaluation protocol): unfreezing an additional convolutional block during fine-tuning to give the network more capacity to separate visually similar early-stage cases, and applying test-time augmentation (TTA) at evaluation averaging each test image's prediction with that of its horizontal flip, which recovers a small, consistent accuracy gain by smoothing out single-view prediction noise.

Figure 3 shows the training and validation loss and accuracy curves across the complete 50 epoch training run: 15 epochs with the convolutional base frozen (Phase 1), followed by 35 fine-tuning epochs with block4 and block5 unfrozen (Phase 2), separated by the dashed vertical line. Training accuracy converges to near-ceiling levels while validation accuracy plateaus around 89-92%, and validation loss stabilizes without a sustained upward trend, indicating that fine-tuning did not overfit substantially beyond this point despite the persistent gap between training and validation curves that is typical of fine-tuning on a modestly sized medical imaging dataset.



Fig 3.0 Training and validation loss and accuracy across the full 50-epoch training run (Phase1:epochs 1-15, frozen base; Phase 2: epochs 16-50, fine-tuning with block4 and block5 unfrozen).

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

Figure 4 illustrates why the final reported test accuracy (91.4%) sits visibly above the validation accuracy curve observed during training. The green marker shows the final held-out test accuracy computed with test-time augmentation, applied only at final evaluation and not during the epoch-by-epoch validation checks used for early stopping. The gap between the two is therefore expected and attributable to TTA, not an inconsistency in the reported results.



Fig 4.0 Fine-tuned phase validation accuracy curve with the final held-out test accuracy (91.4%, with TTA) marked, showing the gap between per-epoch validation accuracy and the final TTA-evaluated test result.

Figure 5 presents the confusion matrix for the four-class classification task (Non-Demented, Very Mild Demented, Mild Demented, Moderate Demented). The model achieved near-perfect classification on Moderate Demented (99.3% precision, 100% recall) and strong performance on Mild Demented (97.3% precision, 97.3% recall). The dominant source of error remains confusion between the Non-Demented and Very Mild Demented classes, which were mutually misclassified in 42 of the 594 test cases (down from 49 in the earlier iteration of the model) consistent with the well-documented clinical difficulty of distinguishing early, subtle cognitive changes from typical age-related variation, rather than indicating an arbitrary model weakness.

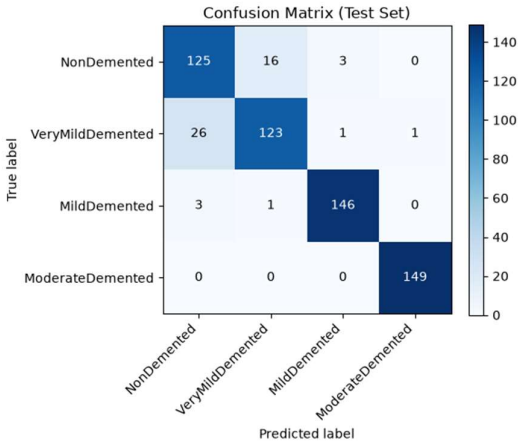


Fig 5.0 Confusion matrix for the VGG16 classifier on the held-out test set (n = 594, accuracy = 91.4%, with test-time augmentation).

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

4.2 System Implementation: NeuroSense

The trained model[3], [38] was deployed within NeuroSense, a clinical workspace interface designed to make the classifier's output usable, explainable, and auditable. The system requires user authentication (Figure 6) and presents a dashboard summarising session activity, including total analyses performed, patients on file, the most common predicted stage, and the most recent analysis on record (Figure 7). Patient analysis records are persisted server-side in a MongoDB-backed store rather than only in browser storage, so saved records remain available across sessions and devices.

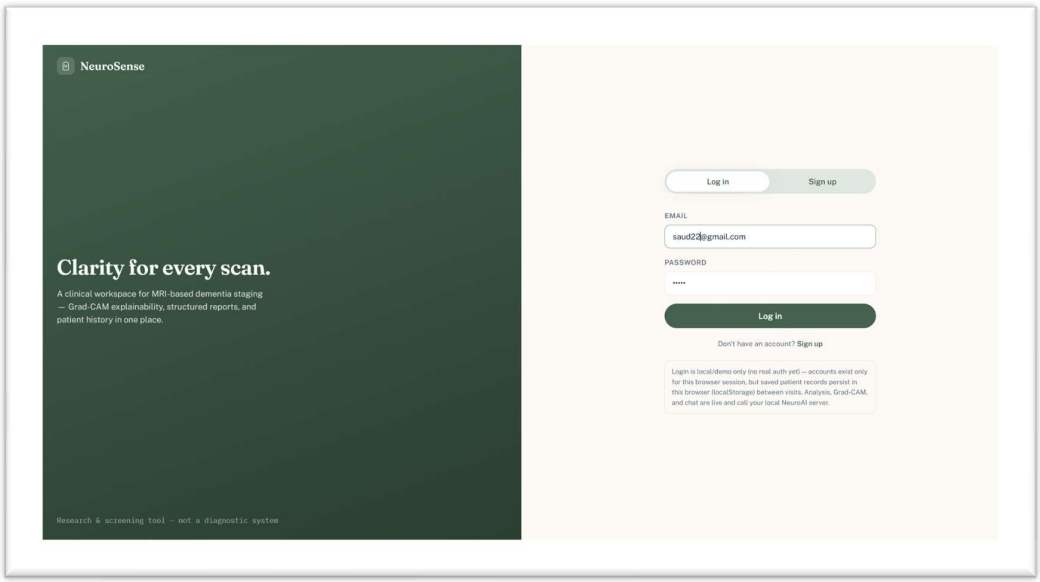


Fig 6.0 NeuroSense login interface

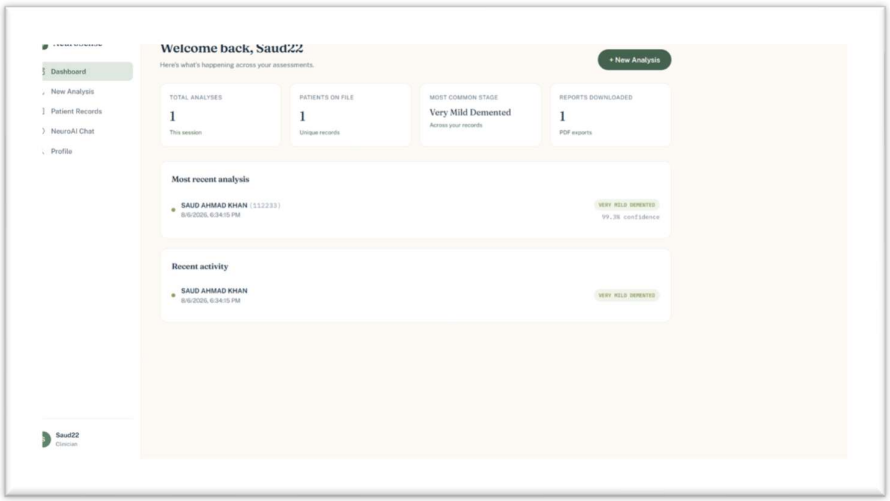


Fig 7.0 NeuroSense dashboard, summarizing session-level analysis activity and the most recent saved analysis.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

A new classification is initiated through a structured intake form (Figure 8) capturing patient identifiers, referring clinician, scan date, and optional clinical notes, alongside the MRI slice to be analyzed. Upon submission, the system returns the predicted stage, a full confidence breakdown across all four classes, a Cognition Horizon severity scale visualizing where the prediction falls along the disease progression spectrum, and a Grad-CAM heatmap overlaid on the original scan to visually indicate which regions most influenced the model's decision (Figure 9).

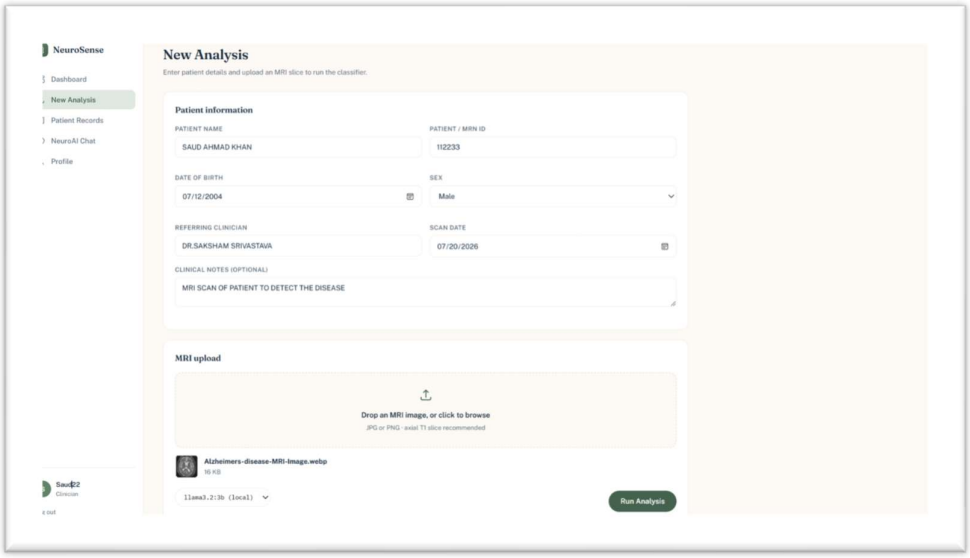


Fig 8.0 Patient intake and MRI upload interface for a new analysis.

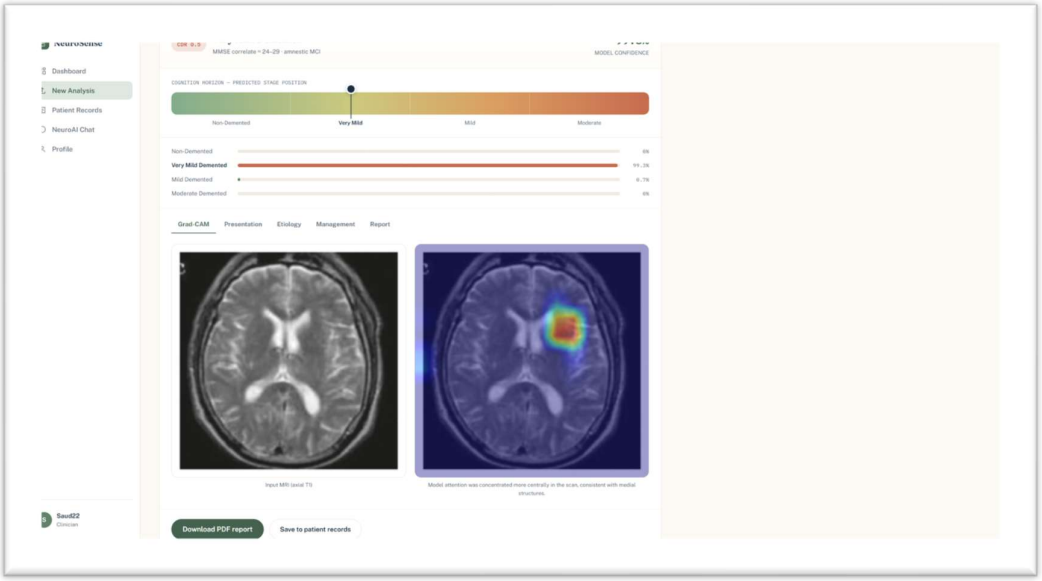


Fig 9.0 Prediction result for a Very Mild Demented case (99.3% confidence), showing the Cognition Horizon severity scale, per-class confidence breakdown, and Grad-CAM heatmap overlay alongside the original MRI slice.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

Beyond the raw prediction[39], NeuroSense generates a structured clinical report (Figure 10) combining the prediction summary with retrieval-grounded clinical notes covering typical presentation, etiology and risk factors, the Grad-CAM [40] attention region in plain language, and general management considerations. Consistent with the system's design principle that the language model explains rather than diagnoses, every generated report is explicitly framed as a research and screening aid, not a clinical diagnosis, with a formal diagnosis requiring evaluation by a qualified professional.



NeuroSense Clinical Report

Generated: 8/7/2026, 1:34:23 AM

Research / screening output — not a clinical diagnosis

Patient Information

Name:	SAUD AHMAD KHAN	Patient / MRN ID:	112233
Date of birth:	2004-07-12	Sex:	Male
Referring clinician:	DR.SAKSHAM SRIVASTAVA	Scan date:	2026-07-20

Clinical notes:
MRI SCAN OF PATIENT TO DETECT THE DISEASE

Prediction Summary

Predicted stage: Very Mild Demented (CDR 0.5)

Model confidence: 99.3%

Clinical correlate: MMSE correlate ~H 24–29 · amnestic MCI

Confidence breakdown

Non-Demented:	0.0%
Very Mild Demented:	99.3%
Mild Demented:	0.7%
Moderate Demented:	0.0%

Clinical Notes

Presentation:
Mild, consistent episodic memory deficit reported by patient/informant, often not evident on brief exam. IADLs largely preserved.

Etiology / risk factors:
Early, subtle hippocampal or entorhinal volume changes may be present; findings can be within normal limits at this stage.

Grad-CAM attention:
Subtle attention over the medial temporal lobe — early, low-magnitude activation.

Management considerations:
Structured follow-up every 6–12 months. Evaluate reversible contributors (B12, TSH, depression, sleep apnea, polypharmacy).

Fig 10.0 Generated clinical report, showing patient information, the structured prediction summary, and retrieval-grounded clinical notes.

Saved analyses are also accessible from a dedicated Patient Records view (Figure 11), which lists prior assessments with predicted stage, confidence, and date for later reference or audit.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY



Fig 11.0 Patient Records view, listing saved analyses with predicted stage, confidence, and date.

Finally, NeuroSense provides a conversational interface, NeuroAI (Figure 12), allowing the user to ask follow-up questions grounded in the most recent analysis, including requests to explain the Grad-CAM result, describe typical disease progression, or summarize documentation points for a referral.

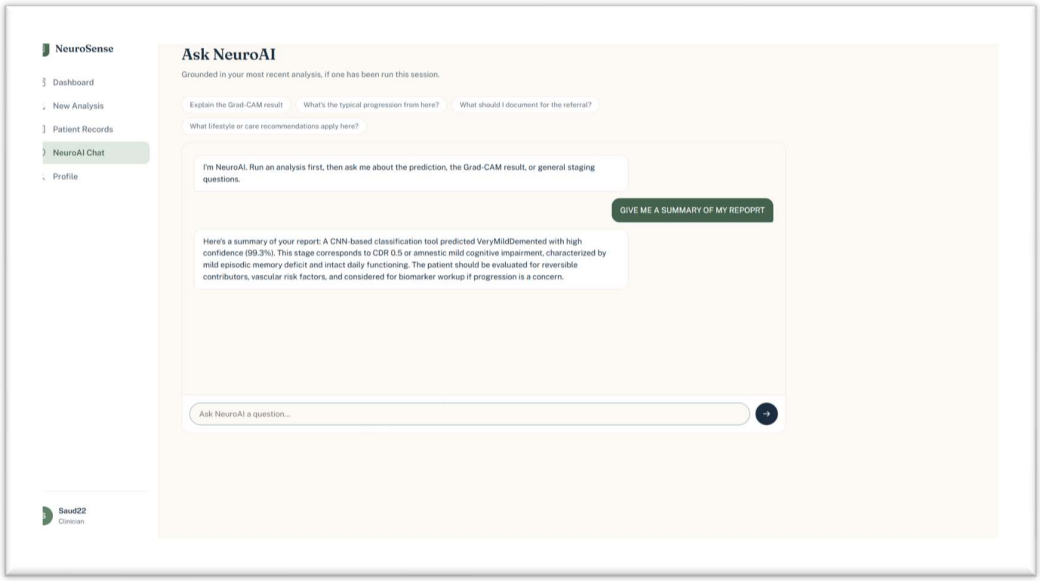


Fig 12.0 NeuroAI conversational interface, offering grounded follow-up questions about the most recent analysis.

5. Conclusion and Discussion

This study demonstrates a holistic paradigm shift in neurodegenerative diagnostic frameworks by unifying deep learning-based visual staging, explainable artificial intelligence (XAI), and local conversational intelligence. Addressing the longstanding limitation of "black-box" deep neural networks in clinical environments, the fine-tuned VGG16 architecture achieves a 91.4% multi-class accuracy across four progressive stages of alzheimer's disease, outperforming benchmark vision architectures including DenseNet121 (85.75%), ResNet50 (82.28%), and EfficientNetV2S (90.32%). Critical to clinician trust and diagnostic transparency, the framework seamlessly embeds Gradient-Weighted Class Activation Mapping (Grad-CAM) to

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

generate visual heatmaps that isolate localised neuropathological atrophy patterns, directly corroborating automated image classifications with established neuroimaging biomarkers.

A principal innovation of this research lies in its privacy-preserving Edge AI deployment, which operationalises a localised Llama 3.2 LLM augmented by RAG. By domain-grounding the RAG pipeline in authoritative clinical guidelines established by the Alzheimer's Association, the system provides contextually grounded, low-latency diagnostic reasoning directly on local edge hardware. This edge-centric architecture eliminates reliance on external cloud infrastructure, safeguarding sensitive patient health data against data privacy vulnerabilities.

The *NeuroSense* serves as a preliminary clinical decision support system designed to bridge the gap between initial MRI scan acquisition and preliminary neurological evaluation. The platform converts model outputs into actionable clinical deliverables, providing per-class confidence distributions, a continuous "Cognition Horizon" severity scale, automated structured report generation, and real-time interactive query resolution via a conversational agent. While quantitative performance analysis reveals that diagnostic error is predominantly concentrated in distinguishing Non-Demented from Very Mild Demented cases (reflecting the clinical challenge of isolating early subtle cognitive decline from normal ageing), the system's structured outputs provide objective preliminary screening without usurping expert medical judgment.

Ultimately, this edge-deployed, XAI-driven framework establishes a scalable blueprint for modern neurodegenerative disease management. By harmonising deep convolutional classification, visual interpretability, localised LLM reasoning, and clinical knowledge retrieval, the proposed system provides an interpretable, secure, and accessible diagnostic aid. Future work will focus on integrating longitudinal multimodal clinical data, including fluid biomarkers, cognitive scoring assessments, and electronic health records, to increase early-stage sensitivity and advance equitable, low-latency AI assistance across global clinical workflows.

References

[1] H. D. Yang, D. H. Kim, S. B. Lee, and L. D. Young, “History of Alzheimer’s Disease,” *Dement. Neurocogn. Disord.*, vol. 15, no. 4, p. 115, 2016, doi: 10.12779/DND.2016.15.4.115.

[2] I. M. Hajjar *et al.*, “Adhesion molecules provide an endothelial protein signature in preclinical and clinical Alzheimer’s disease and predict clinical progression,” *Communications Medicine 2026 6:1*, vol. 6, no. 1, pp. 283-, Mar. 2026, doi: 10.1038/s43856-026-01519-4.

[3] H. Fuse, K. Oishi, N. Maikusa, and T. Fukami, “Detection of alzheimer’s disease with shape analysis of MRI images,” *Proceedings - 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2018*, pp. 1031–1034, Jul. 2018, doi: 10.1109/SCIS-ISIS.2018.00171.

[4] B. Hemalatha, K. Venkatachalam, S. Siuly, and J. Cho, “AI-driven framework for accurate detection of Alzheimer’s disease in EEG,” *Scientific Reports 2026 16:1*, vol. 16, no. 1, pp. 5509-, Feb. 2026, doi: 10.1038/s41598-026-35184-3.

[5] M. Chethan, · P Sabarinathan, · R Arumuga Arun, · R S Rampriya, and R. Soundrapandiyan, “Building an Efficient Multi-stage Alzheimer’s Disease Classification System Using Computationally Effective Hybrid and Ensemble Models,” *International Journal of Computational Intelligence Systems 2026 19:1*, vol. 19, no. 1, pp. 268-, May 2026, doi: 10.1007/S44196-026-01292-1.

CC BY

[6] M. W. Patil, D. P. G. Burade, and Dr. M. Nikose, "Lightweight Hybrid Deep Learning and Machine Learning Framework for Multiclass Alzheimer's Detection Using PCA-Compressed CNN Features," *International Journal on Advanced Computer Theory and Engineering*, vol. 15, no. 1S, pp. 232–237, Jan. 2026, doi: 10.65521/IJACTE.V15I1S.1323.

[7] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia," *Br. J. Psychiatry*, vol. 140, no. 6, pp. 566–572, 1982, doi: 10.1192/BJP.140.6.566.

[8] B. Sheehan, "Assessment scales in dementia," *Ther. Adv. Neurol. Disord.*, vol. 5, no. 6, p. 349, 2012, doi: 10.1177/1756285612455733.

[9] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia," *Br. J. Psychiatry*, vol. 140, no. 6, pp. 566–572, 1982, doi: 10.1192/BJP.140.6.566.

[10] B. Sheehan, "Assessment scales in dementia," *Ther. Adv. Neurol. Disord.*, vol. 5, no. 6, p. 349, 2012, doi: 10.1177/1756285612455733.

[11] G. M. McKhann *et al.*, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimers Dement.*, vol. 7, no. 3, p. 263, 2011, doi: 10.1016/J.JALZ.2011.03.005.

[12] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia," *Br. J. Psychiatry*, vol. 140, no. 6, pp. 566–572, 1982, doi: 10.1192/BJP.140.6.566.

[13] S. Jati *et al.*, "AI guided discovery of a murine model of asymptomatic Alzheimer's disease," *Acta Neuropathologica Communications 2026 14:1*, vol. 14, no. 1, pp. 110-, Apr. 2026, doi: 10.1186/S40478-026-02286-Y.

[14] A. Prince Nickson, K. Ramalakshmi, S. Shirly, and R. Venkatesan, "Alzheimer's Disease Prediction and Detection using Machine Learning and Deep Learning Approaches," *2025 International Conference on Computing and Communications, COMPUTINGCON 2025*, 2025, doi: 10.1109/COMPUTINGCON64838.2025.11377300.

[15] R. Katzman, "Editorial: The prevalence and malignancy of Alzheimer disease. A major killer," *Arch. Neurol.*, vol. 33, no. 4, pp. 217–218, 1976, doi: 10.1001/ARCHNEUR.1976.00500040001001.

[16] A. Prabhu, "Alzheimer's Disease: A Comprehensive Review of its Causes, Diagnosis, and Treatment," *Journal of Student Research*, vol. 12, no. 4, Nov. 2023, doi: 10.47611/jsrhs.v12i4.5770.

[17] F. Lui and J. W. Tsao, "Alzheimer Disease," *StatPearls*, pp. 1–27, Feb. 2024, Accessed: Aug. 10, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK499922/

[18] T. Kim *et al.*, "Impact of dose and compliance of antidementia medications on long-term outcomes in Alzheimer's disease: a nationwide real-world study," *Alzheimer's Research & Therapy 2026 18:1*, vol. 18, no. 1, pp. 24-, Jan. 2026, doi: 10.1186/S13195-025-01942-0.

[19] L. R. Feng, "The cancer Alzheimer's disease paradox," *npj Aging 2026 12:1*, vol. 12, no. 1, pp. 93-, Jul. 2026, doi: 10.1038/s41514-026-00442-1.

[20] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies (Basel).*, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/technologies13010032.

[21] I. Afifi, M. Elgendy, M. Abdelfatah, and S. El-Sappagh, "Vision and convolutional transformers for Alzheimer's disease diagnosis: a systematic review of architectures, multimodal fusion and critical gaps," *Brain Informatics 2025 13:1*, vol. 13, no. 1, pp. 1-, Dec. 2025, doi: 10.1186/S40708-025-00286-7.

[22] S. S. Bamber and T. Vishvakarma, "Medical image classification for Alzheimer's using a deep learning approach," *Journal of Engineering and Applied Science 2023 70:1*, vol. 70, no. 1, pp. 54-, May 2023, doi: 10.1186/S44147-023-00211-X.

CC BY

[23] Z. Zhao *et al.*, "Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: A review," *Front. Comput. Neurosci.*, vol. 17, Feb. 2023, doi: 10.3389/fncom.2023.1038636.

[24] K. S. Manojee, S. Ananthi, V. C. Bharathi, S. Lavanya, R. Aruna, and M. Mythili, "AI-Driven Detection of Alzheimer's Disease: Deep Learning for Identifying Four Stages of Dementia," *International Conference Intelligent Computing and Control Systems*, pp. 1468–1474, 2026, doi: 10.1109/ICICCS67901.2026.11502853.

[25] M. Amin-Naji, H. Mahdavinataj, and A. Aghagolzadeh, "Alzheimer's disease diagnosis from structural MRI using Siamese convolutional neural network," *4th International Conference on Pattern Recognition and Image Analysis, IPRIA 2019*, pp. 75–79, Mar. 2019, doi: 10.1109/PRIA.2019.8786031.

[26] M. M. S. Fareed *et al.*, "ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans," *IEEE Access*, vol. 10, pp. 96930–96951, 2022, doi: 10.1109/ACCESS.2022.3204395.

[27] Y. Feng, X. Xu, Y. Zhuang, and M. Zhang, "Large language models improve Alzheimer's disease diagnosis using multi-modality data," *Proceedings - 2023 1st IEEE International Conference on Medical Artificial Intelligence, MedAI 2023*, pp. 61–66, May 2023, doi: 10.1109/MedAI59581.2023.00016.

[28] A. K. Lahiri and Q. V. Hu, "AlzheimerRAG: Multimodal Retrieval-Augmented Generation for Clinical Use Cases," *Machine Learning and Knowledge Extraction 2025, Vol. 7, Page 89*, vol. 7, no. 3, p. 89, Aug. 2025, doi: 10.3390/MAKE7030089.

[29] T. Zheng, X. Xie, X. Peng, H. Chen, and F. Tian, "Alzheimer's disease detection based on large language model prompt engineering," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* , vol. 15170 LNAI, pp. 207–216, Jan. 2025, doi: 10.1007/978-981-96-1151-5_21.

[30] Z. Huang *et al.*, "ADAM-1: A Novel AI driven approach to Alzheimer's disease Research and Multi-modal Data Analytics," *Alzheimer's & Dementia*, vol. 21, no. Suppl 1, p. e100530, Dec. 2025, doi: 10.1002/ALZ70855_100530.

[31] S. Singh *et al.*, "Next generation preventive neurology: how artificial intelligence and machine learning are reshaping Alzheimer's disease research," *Behavioral and Brain Functions 2026 22:1*, vol. 22, no. 1, pp. 21-, Mar. 2026, doi: 10.1186/S12993-026-00329-X.

[32] L. T. Naing *et al.*, "AI-powered evaluation of dementia severity based on clinical data and visual scoring systems (MTA, ERICA, GCA) from MRI," *Scientific Reports 2026 16:1*, vol. 16, no. 1, pp. 21545-, May 2026, doi: 10.1038/s41598-026-51725-2.

[33] A. A. Sarhan, K. A. El-Sharkawy, T. S. Elserafy, and E. M. A. Alrasheed, "Novel risk factors and comprehensive assessment of Alzheimer's disease patients: case control hospital-based study," *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery 2026 62:1*, vol. 62, no. 1, pp. 33-, Feb. 2026, doi: 10.1186/S41983-025-01061-5.

[34] H. Deng *et al.*, "Treatment patterns of symptomatic treatments for Alzheimer's disease and related dementias," *BMC Geriatr.*, vol. 26, no. 1, p. 39, Dec. 2025, doi: 10.1186/S12877-025-06745-4.

[35] S. Muthuchetty, "A proposed framework for multimodal diagnosis in Alzheimer's disease," *Discover Neuroscience 2025 21:1*, vol. 21, no. 1, pp. 1-, Jan. 2026, doi: 10.1186/S13064-025-00227-4.

[36] K. Sekar, R. Meena, S. Monisha, T. Dheepa, A. K. Niraimathi, and M. D. Borah, "Early Detection Of Alzheimer's Disease using Transfer Learning on MRI Data," *3rd International Conference on Integrated Circuits and Communication Systems, ICICACS 2025*, 2025, doi: 10.1109/ICICACS65178.2025.10968359.

[37] L. Wang *et al.*, "A retrieval-augmented generation large language model framework for accurate dementia identification from electronic health records," *medRxiv*, p. 2026.01.24.26344477, Jan. 2026, doi: 10.64898/2026.01.24.26344477.

CC BY

[38] S. Darsana and L. Abraham, “Multistructure brain registration using multimodal neuroimaging for the detection of Alzheimer’s disease,” *2016 International Conference on Communication Systems and Networks, ComNet 2016*, pp. 221–226, Jan. 2017, doi: 10.1109/CSN.2016.7824018.

[39] S. Mostafaei, D. W. Shemer, J. K. L. Mak, I. K. Karlsson, and S. Hägg, “Improved polygenic risk prediction for alzheimer’s disease and related dementias using deep learning: age and APOE-stratified analysis,” *Alzheimers Res. Ther.*, vol. 18, no. 1, p. 76, Dec. 2026, doi: 10.1186/S13195-026-02011-W.

[40] J. Hariharan and R. Jothi, “Alzheimer’s disease prediction using deep learning and XAI based interpretable feature selection from blood gene expression data,” *Scientific Reports 2026 16:1*, vol. 16, no. 1, pp. 8022-, Feb. 2026, doi: 10.1038/s41598-026-35260-8.