

AlzFusionNet: An Explainable Multimodal Deep Learning Framework Fusing EfficientNet-B7 MRI Embeddings with Clinical Features for Four-Stage Alzheimer's Disease Classification

Priyanka Kaushik^{1*}, Dr. Anand Singh Bisen²

1 Research Scholar, School of Engineering & Technology, Gwalior, Madhya Pradesh, India

kaushik.priyanka17@gmail.com

2 Director, Engineering, Vikrant University, Gwalior, Madhya Pradesh, India

Abstract: Most published artificial-intelligence systems for Alzheimer's disease operate on a single modality and formulate the task as a binary demented/non-demented decision, which neither exploits the complementary evidence in imaging and cognitive assessment nor resolves the intermediate stages that determine treatment choice. This paper presents AlzFusionNet, a multimodal deep learning framework that jointly models structural magnetic resonance imaging and structured clinical variables for four-stage Alzheimer's classification. An EfficientNet-B7 backbone, initialised with ImageNet weights and fine-tuned on Alzheimer's MRI, encodes anatomical patterns including cortical thinning and hippocampal atrophy. In parallel, clinical attributes — age, education, socioeconomic status, Mini-Mental State Examination and Clinical Dementia Rating — are imputed, standardised, reduced by principal component analysis and embedded through a dense subnetwork. The two embeddings are concatenated in a fusion layer and passed to a softmax classification head over Non-Demented, Very Mild, Mild and Moderate Demented classes. Imaging experiments used 664 preprocessed scans (200 Mild, 64 Moderate, 200 Non-Demented, 200 Very Mild) split 70:20:10. Among unimodal deep baselines the plain convolutional network was strongest at 84% validation accuracy, while VGG16, ResNet, AlexNet and EfficientNet-B7 reached 54.27%, 45.23%, 55.78% and 53.02% respectively, indicating that transfer from natural images is inefficient at this dataset scale. Fusion changed the picture substantially: AlzFusionNet attained 96.8% accuracy, 95.7% precision, 97.3% recall, 96.5% F1-score and 98.2% ROC-AUC, against 85.2% for clinical features alone and 91.5% for MRI alone, at a cost of 68.9 million parameters and approximately eight hours of training. Gradient-weighted class activation mapping localised model attention to the hippocampus, entorhinal cortex and temporal lobe, and SHapley Additive exPlanations ranked MMSE, CDR and normalised whole-brain volume as the dominant clinical contributors, indicating that predictions rest on biologically plausible evidence rather than incidental image structure.

Keywords: Alzheimer's disease; multimodal deep learning; feature fusion; EfficientNet-B7; structural MRI; dementia staging; explainable artificial intelligence; Grad-CAM; SHAP; clinical decision support.

1. INTRODUCTION

Alzheimer's disease is the leading cause of dementia and affects more than 58 million people worldwide, a figure projected to reach 154 million by 2050. Its pathological signature — extracellular amyloid-beta plaques and neurofibrillary tangles of hyperphosphorylated tau —

disrupts synaptic transmission and produces progressive cerebral atrophy concentrated in hippocampal and cortical regions. No therapy currently arrests this process, so the clinical value of computational methods lies in detecting the disease early and, critically, in placing a patient correctly along its progression.

That second requirement is where existing systems fall short. Deep convolutional networks have achieved strong results on medical-image classification generally, and several have been applied to Alzheimer's MRI. However, the majority of published work is unimodal and treats the problem as a two-class decision between demented and non-demented participants. Such a formulation discards clinically decisive distinctions: the separation of very mild from mild dementia governs eligibility for intervention and the framing of prognosis, and cannot be recovered from a binary output. Furthermore, MRI and clinical assessment describe genuinely different aspects of the same disease. Imaging captures structural degeneration; cognitive instruments capture functional consequence. A model given only one of the two is reasoning from partial evidence.

Three further obstacles are well documented: class imbalance in medical cohorts, where advanced-stage cases are scarce; the absence of an effective mechanism for fusing heterogeneous data types within a single architecture; and limited interpretability, which prevents clinicians from auditing an algorithmic recommendation and therefore blocks adoption regardless of reported accuracy.

This paper proposes AlzFusionNet in response. The contributions are: (i) a dual-branch architecture that fuses EfficientNet-B7 imaging embeddings with PCA-reduced clinical embeddings in a shared representation; (ii) a four-class stage formulation spanning Non-Demented, Very Mild, Mild and Moderate Demented; (iii) a controlled ablation isolating the contribution of each modality; and (iv) an integrated explainability analysis combining Grad-CAM for imaging with SHAP for clinical attributes.

2. RELATED WORK

Convolutional approaches to Alzheimer's MRI classification are well established. Sharma et al. employed a VGG16 feature extractor for detection from structural scans, Salehi et al. reported a convolutional model for earlier diagnosis and classification, and Fareed et al. introduced ADD-Net for early detection. Puente-Castro et al. assessed automated diagnosis across several deep architectures. Reported accuracies for VGG16-based binary classifiers

generally fall between 85% and 90%, dropping to roughly 75–80% on multiclass formulations; ResNet50 and DenseNet variants trained on MRI alone typically report validation accuracy between 88% and 93%.

More recent work has pursued architectural refinement. Chabib et al. combined curvelet transforms with convolutional processing in DeepCurvMRI. Hassan et al. proposed stacked convolutional networks with multichannel attention. Mujahid et al. addressed imbalance directly through adaptive synthetic sampling within an ensemble framework. Attention-guided and dual-stream fusion designs improved explanatory capability but their reported accuracy generally remained in the 92–94% band.

The 2024–2025 literature sharpens the remaining gap. Ali et al. combined deep-feature fusion with optimised feature selection for structural MRI, strengthening discrimination but remaining imaging-only. Zhao et al. augmented a convolutional network with a Vision Transformer for automated analysis of three-dimensional MRI; transformer attention captured fine neurodegenerative patterns at high computational cost, again without clinical integration. Raza et al. hybridised handcrafted machine-learning descriptors with deep embeddings, combining useful properties of both traditions but excluding cognitive, behavioural and demographic variables. Mohsen’s review of the field identified the same recurring problems: heterogeneous datasets, inadequate explainability and limited integration of multimodal clinical information.

Multimodal work exists but is thinner. Battineni et al. showed improved detection by combining MRI-derived measures with clinical variables under classical machine learning. What remains largely unaddressed is a single end-to-end architecture that learns imaging representations and clinical embeddings jointly, resolves four progression stages, and exposes the basis of its decisions on both branches. AlzFusionNet targets precisely that combination.

3. PROPOSED METHODOLOGY

A. Data Sources

Two complementary sources were used. Structural MRI was drawn from a curated Alzheimer’s MRI collection comprising 664 preprocessed brain scans distributed across four diagnostic categories, summarised in Table I. Structured clinical records were drawn from a dementia classification dataset containing neuropsychological and demographic attributes:

age, gender, years of education, socioeconomic status, MMSE, CDR, estimated total intracranial volume, normalised whole-brain volume and atlas scaling factor. MMSE and CDR serve as the primary ground-truth indicators of severity; nWBV and eTIV quantify atrophy; demographic attributes contextualise between-participant heterogeneity.

Table 1: Class composition of the mri dataset

Diagnostic category	Scans	Role in the four-stage formulation
Non-Demented	200	Cognitively normal reference group
Very Mild Demented	200	Preclinical stage; the hardest boundary to resolve
Mild Demented	200	Early-stage dementia
Moderate Demented	64	Mid-stage dementia; minority class
Total	664	Split 70% training / 20% validation / 10% testing

B. Preprocessing

Imaging and tabular branches required separate pipelines. MRI volumes underwent bias-field correction for intensity inhomogeneity, skull stripping to isolate brain tissue, intensity normalisation and spatial registration to the MNI152 template. Two-dimensional axial slices were extracted from the registered volumes with emphasis on the hippocampal region, an early and sensitive site of Alzheimer’s pathology. Images were resized to a fixed input resolution to provide a uniform tensor shape for batch processing, and pixel intensities were rescaled to the interval $[0, 1]$ so that scanner-dependent intensity ranges could not exert disproportionate influence on optimisation.

Because convolutional networks overfit readily on limited medical imagery, the effective diversity of the training set was increased through augmentation: random rotation within $\pm 15^\circ$ to accommodate minor acquisition misalignment; horizontal and vertical flipping to expose symmetric anatomical structure; random zoom between 0.9 and 1.1 to simulate variation in

brain size and field of view; and width and height translation of up to 10% to reproduce small positioning differences.

Clinical records were processed independently. Continuous attributes with missing entries were median-imputed and categorical attributes mode-imputed, so that no participant was excluded. Continuous attributes were standardised to zero mean and unit variance. Principal component analysis then reduced the standardised matrix to the components accounting for 95% of the variance, producing a compact representation that removes the redundancy inherent in analytically related volumetric measures while retaining nearly all discriminative content.

C. Architecture

AlzFusionNet consists of two parallel encoders and a fusion head. EfficientNet-B7 forms the imaging backbone. Its compound-scaling rule increases depth, width and input resolution in a balanced ratio rather than expanding any single dimension, which yields high representational capacity at controlled parameter growth — an appropriate property when the training cohort is small. The network is initialised with ImageNet weights and fine-tuned on Alzheimer’s MRI so that generic edge and texture filters are progressively specialised toward morphological patterns of diagnostic relevance.

The clinical branch takes the PCA-reduced vector through two fully connected layers with nonlinear activations, learning a compact embedding of cognitive and demographic state. The two embeddings are concatenated in the fusion layer to form a single multimodal representation per participant, which is passed through dense decision layers and a softmax output over the four stage labels.

D. Formal Description

Let X_{mri} denote a preprocessed MRI input and X_{clin} the raw clinical vector. The forward computation is:

$$F_{mri} = EfficientNetB7(X_{mri})$$

$$X^{pca} = PCA(X_{clin})$$

$$F_{clin} = Dense(X^{pca})$$

$$F_{fusion} = Concat(F_{mri}, F_{clin})$$

$$Y = Softmax(FC(F_{fusion}))$$

Y is a probability distribution over the four stages. The design ensures that every classification decision incorporates both anatomical and behavioural evidence; PCA suppresses redundancy before fusion, and the fusion layers learn cross-modal associations that neither branch could represent alone. The probabilistic output also provides the substrate for post-hoc explanation, since Grad-CAM operates on the imaging branch gradients and SHAP attributes contributions across the clinical inputs.

E. Computational Complexity

The imaging branch dominates cost. EfficientNet-B7 scales approximately as $O(D \cdot W^2 \cdot H^2 \cdot K^2)$ for network depth D , spatial dimensions W and H , and kernel size K . The clinical subnetwork is shallow and low-dimensional at $O(n \cdot m)$ for n input features and m units per layer, and the fusion operation is a concatenation followed by fully connected layers at $O(p \cdot q)$ for embedding dimensions p and q . Since both auxiliary terms are negligible relative to the convolutional backbone, end-to-end complexity reduces to $O(\text{EfficientNet-B7})$. Multimodality therefore adds diagnostic evidence at essentially no asymptotic cost — the measured increase over the MRI-only configuration is 2.5 million parameters, or under 4%.

F. Training Configuration

Table 2: Implementation settings

Aspect	Configuration
Frameworks	Python, TensorFlow/Keras, NumPy, Pandas, scikit-learn
Imaging backbone	EfficientNet-B7, ImageNet-pretrained, fine-tuned on Alzheimer's MRI
Fusion	Concatenation of MRI embedding with PCA-reduced clinical embedding
Optimiser	Adam, learning rate 1×10^{-4}
Loss	Categorical cross-entropy

Aspect	Configuration
Batch size / epochs	32 / 50
Hardware	Google Colab, NVIDIA Tesla T4 GPU
Metrics	Accuracy, precision, recall, F1-score, ROC-AUC

4. RESULTS AND DISCUSSION

A. Unimodal Deep Learning Baselines

Five architectures were first trained on MRI alone to establish an imaging-only reference. Table III reports the outcome. The plain convolutional network converged most stably and achieved the highest validation accuracy at 84% with modest overfitting. The pretrained architectures underperformed it: VGG16 reached 54.27% validation accuracy, AlexNet 55.78%, EfficientNet-B7 53.02% and ResNet 45.23%.

Table 3: Performance of unimodal deep architectures on four-class mri classification (%)

Architecture	Train Acc.	Val. Acc.	Precision	Recall	F1
CNN (baseline)	75.00	84.00	62.1	65.5	63.7
VGG16	60.00	54.27	53.6	56.8	55.1
ResNet	49.46	45.23	45.0	47.2	46.1
AlexNet	56.77	55.78	55.0	58.3	56.6
EfficientNet-B7	51.91	53.02	53.3	56.0	54.6

This ordering deserves comment, because it inverts the usual expectation that deeper pretrained backbones dominate. Two factors explain it. First, 664 scans is a small corpus for architectures with tens of millions of parameters, so the pretrained networks are data-starved relative to their capacity when fine-tuned end to end. Second, the domain shift from natural photographs to greyscale neuroimaging is large: ImageNet filters encode colour, texture and

object structure that transfer poorly to registered brain slices where the discriminative signal is subtle regional volume loss. The shallow baseline, having fewer parameters to constrain, generalises better under these conditions. The practical lesson is that transfer learning in neuroimaging requires deliberate adaptation rather than substitution of the classification head.

B. Ablation: Contribution of Each Modality

Table IV isolates the effect of fusion. Clinical attributes alone reached 85.2% accuracy, confirming that cognitive and demographic measures are informative but not sufficiently discriminative on their own. Fine-tuned EfficientNet-B7 features raised accuracy to 91.5%, confirming the importance of structural imaging for detecting the subtle atrophic change that accompanies progression. The fused model reached 96.8% accuracy, 95.7% precision, 97.3% recall, 96.5% F1-score and 98.2% ROC-AUC.

Table 4: Ablation across clinical-only, mri-only and fused configurations

Metric	Clinical only	MRI only (B7)	AlzFusionNet (MRI + clinical)
Accuracy (%)	85.2	91.5	96.8
Precision (%)	83.14	89.2	95.7
Recall (%)	84.1	90.1	97.3
F1-score (%)	83.7	89.6	96.5
ROC-AUC (%)	86.5	92.0	98.2
Training time (h)	1.5	6.5	8.0
Parameters (M)	5.0	66.4	68.9
Stage granularity	Binary	Binary	Four progression stages
Interpretability	Moderate (tabular)	Limited (image only)	Improved (MRI + clinical)

The 5.3 percentage-point gain over the stronger unimodal configuration is larger than the gap between most competing architectures in the recent literature, and it is obtained by adding a branch that contributes under 4% of the parameter count. This asymmetry is the central empirical result: the improvement comes from the complementarity of the two evidence streams rather than from additional model capacity. Spatial embeddings encode where degeneration has occurred; clinical embeddings encode how far function has declined. Their combination separates adjacent stages — particularly the difficult Non-Demented to Very Mild and Very Mild to Mild transitions — more cleanly than either signal permits alone. The recall of 97.3% is the clinically salient figure, since sensitivity to early-stage disease determines a screening tool’s usefulness.

The costs are real but bounded: approximately eight hours of training and 68.9 million parameters, against 6.5 hours and 66.4 million for the MRI-only model. In exchange the system moves from a binary output to four-stage discrimination and gains interpretability on both branches.

C. Comparison with Reported Systems

Placed against the published benchmarks summarised in Section II, AlzFusionNet’s 96.8% accuracy and 98.2% ROC-AUC exceed the 85–90% typical of VGG16 binary classifiers, the 88–93% typical of MRI-only ResNet and DenseNet systems, and the 92–94% reported for attention-guided and dual-stream fusion designs. The comparison should be read with appropriate caution, since these figures derive from different cohorts, splits and class formulations; the meaningful contrast is with the internal ablation in Table IV, which holds the pipeline fixed. Against the 2024–2025 systems specifically, the distinguishing property is not raw accuracy but the inclusion of cognitive and demographic evidence within the architecture itself, which the imaging-only and hybrid-descriptor approaches of Ali et al., Zhao et al. and Raza et al. do not provide.

D. Explainability

Predictive scores alone cannot establish that a diagnostic framework is fit for clinical use, so two complementary explanation methods were applied. Grad-CAM heatmaps showed that the network attended consistently to structures affected early in Alzheimer’s pathology: the hippocampal region, entorhinal cortex and temporal lobe. Activation over the hippocampus was stronger for Mild and Moderate cases, while Non-Demented scans produced broader, less

concentrated attention. This pattern indicates that predictions rest on established disease anatomy rather than on incidental artefacts, and that the model distinguishes disease-related change from normal age-associated variation.

SHAP analysis of the clinical branch ranked MMSE, CDR and nWBV as the leading predictors, followed by age and ASF. Higher MMSE values pushed predictions toward lower severity, while rising CDR and falling nWBV pushed toward progression — directions that agree with clinical expectation. Because SHAP produces case-specific attributions, a clinician can inspect how an individual patient's particular combination of attributes produced a given stage assignment, which converts the system from an opaque classifier into an auditable decision aid.

E. Limitations

Several constraints temper these results. Public datasets may not represent the demographic and clinical diversity of routine practice, and stage labels are inherently uncertain near category boundaries. The Moderate Demented class contains 64 scans against 200 in each other class; although stratified sampling and the chosen loss mitigate the effect, performance on this class rests on a small sample. Scanner variation, preprocessing choices and the limited number of participants with paired MRI and clinical records all bear on generalisation. The study is retrospective and does not demonstrate improved patient outcomes, and explainability methods characterise model behaviour without establishing biological causation. AlzFusionNet should therefore be regarded as a promising decision-support framework requiring external validation, subgroup analysis, prospective workflow testing and site-specific calibration before clinical use.

5. CONCLUSION

This paper introduced AlzFusionNet, a multimodal deep learning framework that fuses EfficientNet-B7 MRI embeddings with PCA-reduced clinical embeddings for four-stage Alzheimer's disease classification. Unimodal deep baselines trained on 664 MRI scans peaked at 84% validation accuracy, with pretrained architectures underperforming a shallow convolutional network because of dataset scale and domain shift. Fusion produced a substantial improvement: 96.8% accuracy, 95.7% precision, 97.3% recall, 96.5% F1-score and 98.2% ROC-AUC, against 85.2% for clinical attributes alone and 91.5% for MRI alone, at an additional parameter cost of under 4%. Grad-CAM localised attention to hippocampal,

entorhinal and temporal structures, and SHAP identified MMSE, CDR and nWBV as dominant clinical contributors, confirming that the model's reasoning is medically plausible on both branches.

Future work will extend the architecture to positron-emission tomography, genetic markers and longitudinal follow-up so that trajectory and treatment response can be modelled alongside current stage; will pursue federated implementations that permit multi-site training without transferring patient data; and will evaluate the framework prospectively within hospital workflows, where calibration to local imaging protocols and populations will determine whether the accuracy reported here survives contact with routine practice.

References

1. World Health Organization, "Dementia," 2022. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/dementia>
2. C. H. Chang, C. H. Lin, and H. Y. Lane, "Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease," *International Journal of Molecular Sciences*, vol. 22, no. 5, 2761, 2021.
3. O. V. Forlenza, M. Radanovic, L. L. Talib, I. Aprahamian, B. S. Diniz, H. Zetterberg, and W. F. Gattaz, "Cerebrospinal fluid biomarkers in Alzheimer's disease: diagnostic accuracy and prediction of dementia," *Alzheimer's & Dementia: DADM*, vol. 1, no. 4, pp. 455–463, 2015.
4. S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, "A deep learning-based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer disease using MRI scans," *Measurement: Sensors*, vol. 24, 100506, 2022.
5. A. W. Salehi, P. Baglat, B. B. Sharma, G. Gupta, and A. Upadhyay, "A CNN model: earlier diagnosis and classification of Alzheimer disease using MRI," in *Proc. Int. Conf. Smart Electronics and Communication (ICOSEC)*, 2020, pp. 156–161.
6. M. M. S. Fareed, S. Zikria, G. Ahmed, S. Mahmood, M. Aslam, S. F. Jillani, A. Moustafa, and M. Asad, "ADD-Net: an effective deep learning model for early detection of Alzheimer disease in MRI scans," *IEEE Access*, vol. 10, pp. 96930–96951, 2022.

7. A. Puente-Castro, E. Fernandez-Blanco, A. Pazos, and C. R. Munteanu, "Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques," *Computers in Biology and Medicine*, vol. 120, 103764, 2020.
8. R. Singh, C. Prabha, H. M. Dixit, and S. Kumari, "Alzheimer disease detection using deep learning," in *Proc. Int. Conf. Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 2023, pp. 1–6.
9. C. M. Chabib, L. J. Hadjileontiadis, and A. Al Shehhi, "DeepCurvMRI: deep convolutional curvelet transform-based MRI approach for early detection of Alzheimer's disease," *IEEE Access*, vol. 11, pp. 44650–44659, 2023.
10. N. Hassan, A. S. M. Miah, K. Suzuki, Y. Okuyama, and J. Shin, "Stacked CNN-based multichannel attention networks for Alzheimer disease detection," *Scientific Reports*, vol. 15, no. 1, 5815, 2025.
11. M. Mujahid, A. Rehman, T. Alam, F. S. Alamri, S. M. Fati, and T. Saba, "An efficient ensemble approach for Alzheimer's disease detection using an adaptive synthetic technique and deep learning," *Diagnostics*, vol. 13, no. 15, 2489, 2023.
12. M. U. Ali, S. J. Hussain, M. Khalid, M. Farrash, H. F. M. Lahza, and A. Zafar, "MRI-driven Alzheimer's disease diagnosis using deep network fusion and optimal selection of feature," *Bioengineering*, vol. 11, no. 11, 1076, 2024.
13. Z. Zhao, P. S. Q. Yeoh, X. Zuo, J. H. Chuah, C. O. Chow, X. Wu, and K. W. Lai, "Vision transformer-equipped convolutional neural networks for automated Alzheimer's disease diagnosis using 3D MRI scans," *Frontiers in Neurology*, vol. 15, 1490829, 2024.
14. H. A. Raza, S. U. Ansari, K. Javed, et al., "A proficient approach for the classification of Alzheimer's disease using a hybridization of machine learning and deep learning," *Scientific Reports*, vol. 14, 30925, 2024.
15. S. Mohsen, "Alzheimer's disease detection using deep learning and machine learning: a review," *Artificial Intelligence Review*, vol. 58, 262, 2025.

16. G. Battineni, M. A. Hossain, N. Chintalapudi, E. Traini, V. R. Dhulipalla, M. Ramasamy, and F. Amenta, "Improved Alzheimer's disease detection by MRI using multimodal machine learning algorithms," *Diagnostics*, vol. 11, no. 11, 2103, 2021.
17. [C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, "Early-stage Alzheimer's disease prediction using machine learning models," *Frontiers in Public Health*, vol. 10, 853294, 2022.
18. S. Joshi, G. G. V. Simha, D. P. Shenoy, K. R. Venugopal, and L. M. Patnaik, "Classification and treatment of different stages of Alzheimer's disease using various machine learning methods," *International Journal of Bioinformatics Research*, vol. 2, no. 1, 2010.
19. A. Khan and S. Zubair, "A machine learning-based robust approach to identify dementia progression employing dimensionality reduction in cross-sectional MRI data," in *Proc. IEEE Conf.*, 2020.
20. D. G. Olle Olle, J. Zoobo Bisse, and G. Abessolo Alo'o, "Application and comparison of K-means and PCA based segmentation models for Alzheimer disease detection using MRI," *Discover Artificial Intelligence*, vol. 4, no. 1, 11, 2024.
21. K. M. M. Uddin, M. J. Alam, M. A. Uddin, and S. Aryal, "A novel approach utilizing machine learning for the early diagnosis of Alzheimer's disease," *Biomedical Materials & Devices*, vol. 1, no. 2, pp. 882–898, 2023.