Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00072-4

Original Article

Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease

Racheal S. Akinbo [1] , Bamidele M. Kuboye [2, ✉] , Ayanfeoluwa F. Owoseeni [2] , Joseph O. Igbogbo [2] , Tunde J. Fatoke [2]

1. *Data Science, Federal University of Technology, Akure, NGA*
2. *Information Technology, Federal University of Technology, Akure, NGA*

Received: February 22, 2026 | Review began: February 23, 2026 | Review ended: April 15, 2026 | Published: May 13, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Alzheimer's disease is a brain condition that gradually compromises memory, thinking, and learning skills. Brain changes begin years before symptoms appear, making early detection important but difficult. While previous studies use single-type imaging and small datasets, this research develops two multimodal deep learning models that combine brain MRI scans with clinical data (age and gender) to separate early stages (NonDemented, VeryMildDemented) from advanced stages (MildDemented, ModerateDemented) and to classify four severity levels. Using 6,400 MRI scans with added clinical metadata, the models address severe class imbalance through Focal Loss, class weighting, data augmentation, and 5-fold cross-validation. The architecture combines a pre-trained ResNet18 for MRI features with a two-layer multilayer perceptron for clinical features, using late fusion before task-specific output layers. On the test set, the binary model achieved strong screening performance (F1-score = 0.9612, recall = 0.9917 for early stage), showing near-perfect sensitivity for early detection. This suggests that the binary model could serve as an effective first-line screening tool to identify patients who require further clinical assessment. The four-class model yielded a weighted F1-score (0.7173) and macro-average area under the receiver operating characteristic curve (0.914), with most errors between adjacent stages due to gradual disease progression and class imbalance (particularly few ModerateDemented samples). A comparison with MRI-only models showed that the contribution of clinical features depends on task difficulty, providing marginal benefit in binary screening but substantial improvement in four-class severity staging. In subsequent studies, balanced samples of data should be taken at all stages to improve classification.

Categories: Data Analysis, Deep Learning, Medical recommender system

Keywords: alzheimer's disease, multimodal deep learning, mri analysis, disease classification, early detection

Introduction

The most prevalent type of dementia is Alzheimer's disease, a progressive brain disorder that affects millions of individuals globally [1]. The process of the disease starts many years prior to the manifestation of the symptoms. Studies indicate that alterations in the brain begin decades before individuals start experiencing problems with their memory [2]. This silent period is lengthy and significant because by detecting the disease in its early stage, a doctor can initiate treatment on time [3].

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI
https://doi.org/10.7759/s44389-026-00072-4

Traditional means of diagnosing Alzheimer's disease involve single forms of data, which can be brain scans, memory tests, or genetic tests. Nevertheless, a single type of data provides insufficient information regarding the condition of the disease [4]. Multimodal deep learning is a recent development that enables the integration of dissimilar kinds of medical data to enhance the accuracy of diagnosis. It has demonstrated the possibility to record the intricate dynamics of the progression of Alzheimer's disease [5]. Deep learning models such as Convolutional Neural Networks (CNNs) are used to discover patterns in medical images without physicians having to manually identify them [6,7].

The existing studies have significantly improved this task, yet there are some challenges. Most of the research is based on a small amount of patient data or concentrates on brain imaging only. These restrictions influence the level of usefulness of the models in actual healthcare environments [3]. This research adopts an alternative methodology by using a larger sample of participants and simpler procedures that unite brain images with basic clinical data.

Related works

A multimodal system that uses 3D CNNs for imaging data and autoencoders for clinical and genetic data was built by Venugopalan et al. [4] to predict Alzheimer's disease stages. Combining health records with either genetic or imaging data improved results compared to single sources. Nevertheless, the combined approach was not as effective because only 220 patients were available with all three types of data. Lim et al. [8] also experimented with several CNN architectures such as VGG16 and ResNet50. VGG16 was the most accurate (78.57%). The research observed that performance was influenced by the lack of training data and proposed that including clinical features could enhance results, stating the necessity of multimodal data.

Odusami et al. [9] proposed a ResNet18 framework designed with modifications that integrate positron emission tomography and MRI images for binary classification of the presence of Alzheimer's disease. They performed basic concatenation in order to combine the data and obtained an accuracy of 73.90%. The experiment involved explainable AI to distinguish the significant areas of the brain. Khan et al. [10] extended the existing approaches with transfer learning and pre-trained VGG networks. They optimized VGG16 and VGG19 models, obtaining accuracies of 97% and 98%, respectively. The transfer learning technique was effective with limited training data.

A dual-filter CNN that has parallel branches of various filter sizes to extract multi-scale features was designed by El-Assy et al. [11]. On MRI data, the model was found to be 99.43% and 99.57% accurate in three- and four-class classification, respectively. Although it was very accurate, the research had only imaging without clinical features. Gasmi et al. [12] also had the same interest in imaging by combining the models (EfficientNetV2B3 and Inception-ResNetV2) with optimization on the OASIS dataset. With the help of the optimization algorithm, the contribution of each model was adjusted according to its performance, yielding 0.9907 as the Scott's Pi agreement score.

In contrast to the imaging-only methods, Wang et al. [13] created a multimodal deep learning model to forecast the progression between mild cognitive impairment and Alzheimer's disease. They used structural MRI, clinical assessment, and genetic data, and were able to reach 92.92% accuracy and area under the receiver operating characteristic curve (AUC) of 0.962. The researchers were able to detect the established biomarkers such as hippocampal shrinkage. Nevertheless, the sample was less than 500 patients who were not well diversified in their demographic.

More recent studies have used hybrid models. Alorf [14] created a hybrid CNN-Transformer architecture that combines convolutional layers for local features with transformer blocks for global patterns. The model's accuracy was averagely 96% using functional MRI during resting state data from 204 subjects. Mmadumbu et al. [15] proposed methods consisting of long short-term memory-feedforward neural network for clinical data and ResNet50-MobileNetV2 for MRI analysis. The clinical data model achieved 99.82% accuracy, while the MRI model achieved 96.19% accuracy using transfer learning. Despite high reported accuracy, testing was done on only 26 MRI samples, and the models were developed using independent datasets. Yan et al. [16] continued the trend of hybrid models, combining pre-trained ResNet-50 with Vision Transformer (ViT) using only MRI data. ResNet-50 provided feature extraction, while ViT captured global information through attention mechanisms, achieving 97% training accuracy and 95% testing accuracy.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

S/N	Author	Year	Models	Limitations
1	Venugopalan et al. [4]	2021	3D CNN and Stacked Denoising Autoencoders	Insufficient multimodal samples (N=220)
2	Lim et al. [8]	2022	Transfer learning CNN (VGG16, ResNet50)	Limited unimodal data
3	Odusami et al. [9]	2023	Modified ResNet18	Binary classification only
4	Khan et al. [10]	2023	VGG16 and VGG19 with transfer learning	Single imaging modality; small dataset (315 MRI scans)
5	El-Assy et al. [11]	2024	Dual-Filter CNN	No clinical data integration
6	Gasmi et al. [12]	2024	EfficientNetV2B3 & Inception-ResNetV2	Imaging-focused approach
7	Wang et al. [13]	2024	CNN-based multimodal fusion (DISFC)	Small dataset (N<500)
8	Alorf [14]	2025	CNN-Transformer	Single modality evaluation
9	Mmadumbu et al. [15]	2025	LSTM-FNN and ResNet50-MobileNetV2	Very small test set (N=26)
10	Yan et al. [16]	2025	ResNet-50 and Vision Transformer	MRI-only; no multimodal fusion

TABLE 1: Summary of Reviewed Studies on Alzheimer's Disease

CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; FNN, Feedforward Neural Network

The reviewed studies in Table *1* show two main gaps. First, small datasets limit model generalization. Wang et al. [13] and Venugopalan et al. [4] used fewer than 500 patients, while Mmadumbu et al. [15] tested on only 26 samples. Second, studies like El-Assy et al. [11], Gasmi et al. [12], and Yan et al. [16] relied only on imaging data without testing if the multimodal approach of including clinical data like age and gender can improve performance. These gaps serve as the motivation of this research, which uses a ResNet18-multilayer perceptron (MLP) architecture on 6,400 MRI scans with integrated clinical data.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Materials And Methods

This research develops two multimodal deep learning models that combine brain MRI scans with clinical data for Alzheimer's disease classification. The first model separates early stage from advanced stage disease, and the second model classifies patients into four classes of disease (NonDemented, VeryMildDemented, MildDemented, and ModerateDemented). Figure *1* shows the overall development workflow from data collection to the developed models.

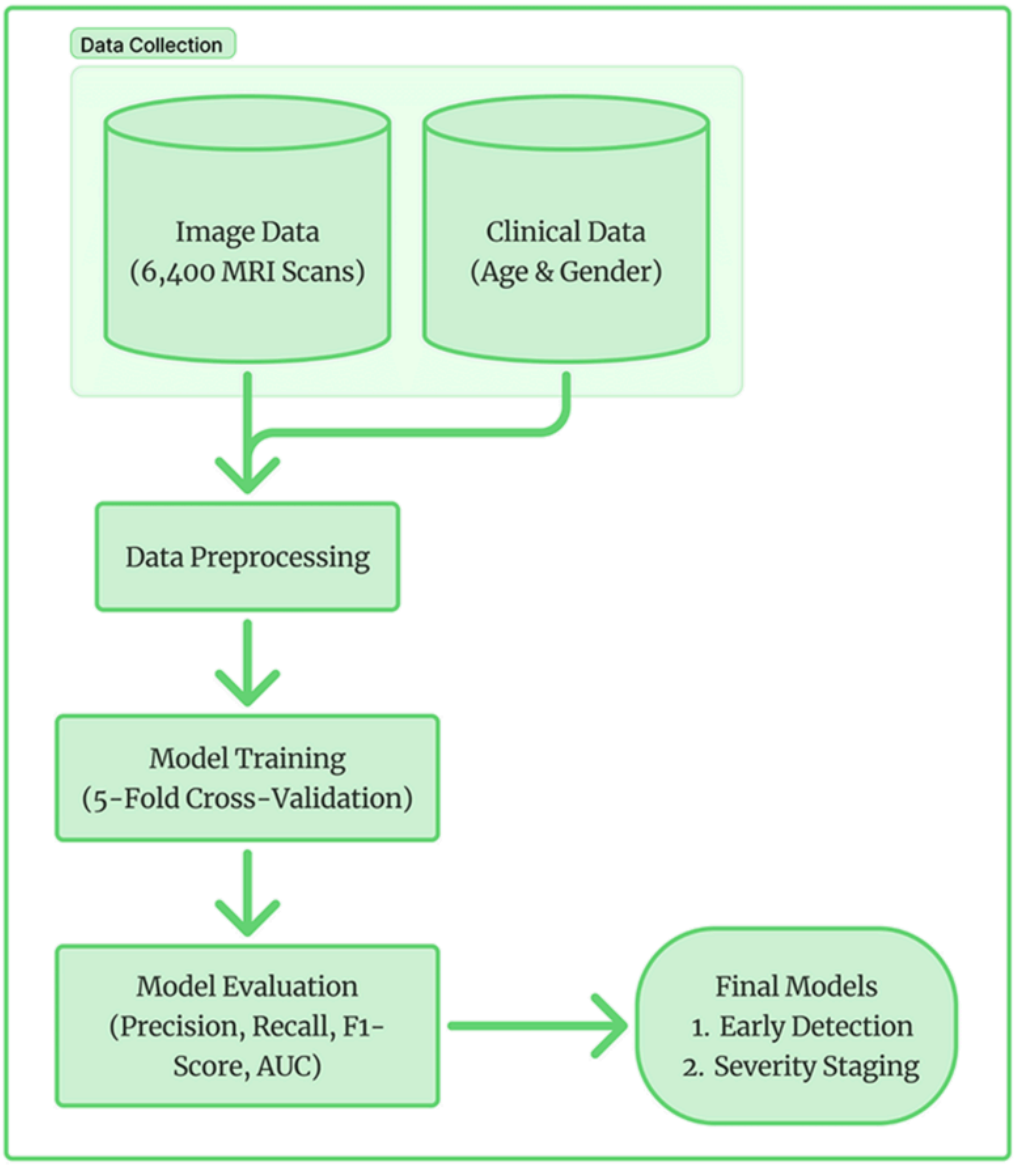


FIGURE 1: Overall Development Workflow

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Data collection

The Mendeley Alzheimer's Disease Detection repository serves as the dataset's source [17]. It contains 6,400 brain MRI scans in JPG format, already distributed into training (80%, N = 5,121) and testing (20%, N = 1,279) sets. Every scan is marked with the severity of the disease. Demographic trends (age and gender data) were included because of how common they are in Alzheimer's populations, as the severity of the disease increases with age [18]. The gender was randomly selected and allocated equally. Table *2* shows the first five entries of the clinical data that was prepared according to the stated structure in the source repository of the dataset.

Image_Path	Label	Participant_ID	Age	Gender
train/NonDemented/nonDem0.jpg	NonDemented	1	69	Male
train/NonDemented/nonDem1.jpg	NonDemented	2	66	Male
train/NonDemented/nonDem10.jpg	NonDemented	3	70	Female
train/NonDemented/nonDem100.jpg	NonDemented	4	74	Male
train/NonDemented/nonDem1000.jpg	NonDemented	5	65	Male

TABLE 2: Entries of Clinical Data

The samples are distributed in four severity classes, which are shown in Table *3*. There are high levels of class imbalance in the dataset, with NonDemented samples being 49 times more common than the ModerateDemented samples [17].

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Severity Class	Training Samples	Test Samples	Total
NonDemented	2,560	640	3,200
VeryMildDemented	1,792	448	2,240
MildDemented	717	179	896
ModerateDemented	52	12	64
Total	5,121	1,279	6,400

TABLE 3: Dataset Distribution Across Severity Classes

Figure *2* presents typical MRI scans of all the severity classes, which reveal progressive patterns of brain atrophy: (a) NonDemented: normal brain structure; (b) VeryMildDemented: minimal hippocampal atrophy; (c) MildDemented: moderate atrophy; (d) ModerateDemented: severe atrophy [17].

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

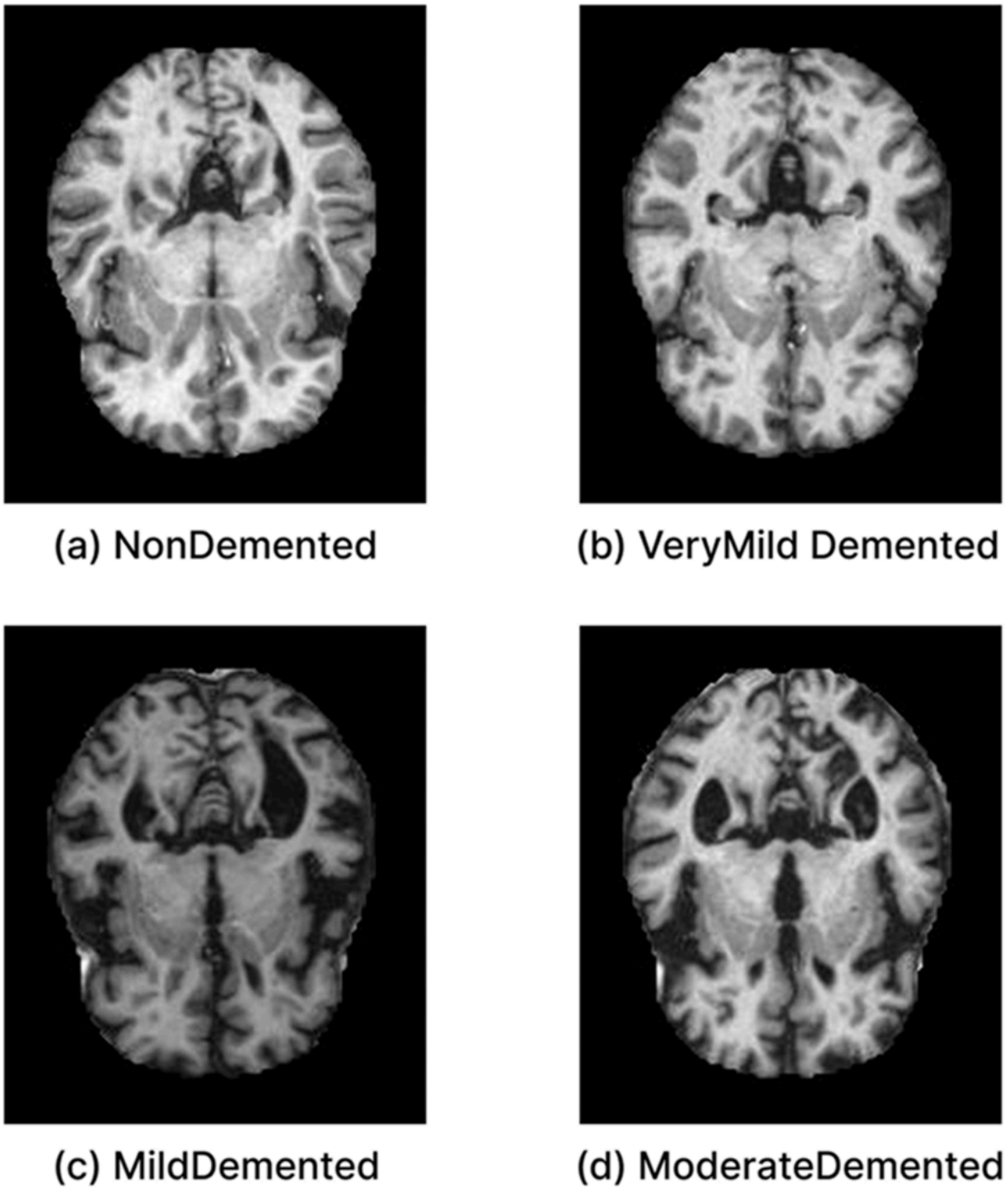


FIGURE 2: Representative MRI Scans

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Data preprocessing

The dataset was prepared using four major steps. Initially, the size of all MRI images was reduced to 224 × 224 pixels in order to fit the ResNet18 input requirement. This normalizes the size of images and retains important brain structures. Second, clinical aspects were normalized, using Z-score normalization in Equation (1) for ages of patients:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original age, μ is the mean age, and σ is the standard deviation. Only the training set was used to compute the mean and standard deviation to prevent information leakage on the test data. The gender was represented in binary values (Male = 1, Female = 0). Third, the severity labels were represented as numbers: NonDemented = 0, VeryMildDemented = 1, MildDemented = 2, and ModerateDemented = 3.

For binary classification, the early stage disease (classes 0 and 1) was set as 0, and the advanced stage disease (classes 2 and 3) was set as 1. Fourth, training was done with data augmentation to enhance generalization. Horizontal flipping (random, 50% chance), random rotation (±15 degrees), random translation (±10%), modifying the brightness and contrast (±30%), and random erasing (20% chance) were the techniques of augmentation.

Horizontal flipping was applied randomly because Alzheimer's-related brain changes do not consistently favour one side of the brain at the level captured in these scans. Random rotations were limited to ±15 degrees to simulate minor head tilts while preserving overall brain anatomy. Random translation of ±10% reflects small variations in patient positioning within the scanner frame. Brightness and contrast adjustments of ±30% account for typical variations in MRI scanner settings without creating clinically implausible images. Random erasing at 20% probability improves robustness to minor image artifacts [19]. There was no augmentation of test data so that unbiased evaluation can be carried out. Figure *3* depicts the preprocessing pipeline that shows the order of each of the stages for the raw MRI scans and clinical data.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

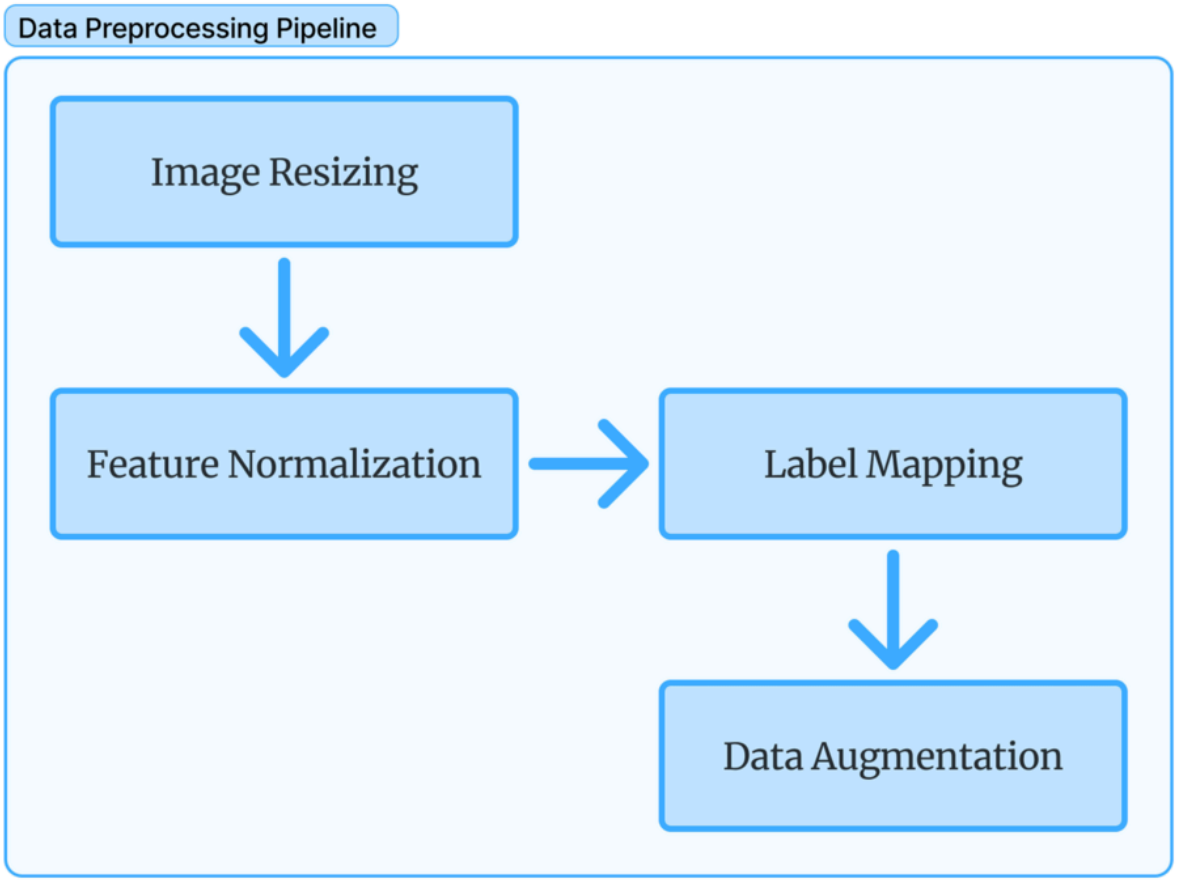


FIGURE 3: Data Preprocessing Pipeline

Model architecture

The two models have the same base architecture with different output layers. It consists of three components: an MRI feature extractor, a clinical feature processor, and a fusion classifier. In the MRI feature extractor, ResNet18 (CNN) is used. The last classification layer was eliminated, leaving behind the convolutional and pooling layers. This obtains a 512-dimensional feature vector in every MRI scan.

ResNet18 was chosen because it offers a good balance between feature extraction and computational efficiency in medical imaging tasks [20]. The model was initialized with weights pre-trained on ImageNet, a common practice in medical image analysis to improve performance when training data is limited [21].

The clinical feature processor is a two-layer MLP. The initial layer transforms the 2-dimensional input (age and gender) into 64 dimensions, after which there is ReLU activation and dropout (rate = 0.5). The second layer is mapped to 128 dimensions with ReLU activation, which produces a 128-dimensional clinical feature vector. The layer sizes of 64 and 128 dimensions were chosen based on the small size of the clinical input (only two features), providing sufficient capacity to learn useful patterns without overfitting.

The fusion classifier combines both the 512-dimensional MRI features and the 128-dimensional clinical features, making a total of 640-dimensional combined representation. This fused vector is processed by three completely connected layers. The initial layer decreases the dimensions from 640 to 256, the second from 256 to 128 with ReLU activation and dropout

(rate = 0.6), and the last layer produces predictions of classes. A higher dropout rate (0.6) was used in the fusion classifier to regularize the combined representation and reduce the risk of memorizing training data [22].

For Model 1 (binary early stage detection), the output layer will have two neurons for early and advanced classification. In the case of Model 2 (multi-class severity staging), there are four neurons in the output layer, which represent the four severity classes. Figure *4* depicts the entire model architecture, which shows how the MRI and clinical paths interact and diverge for the two tasks.

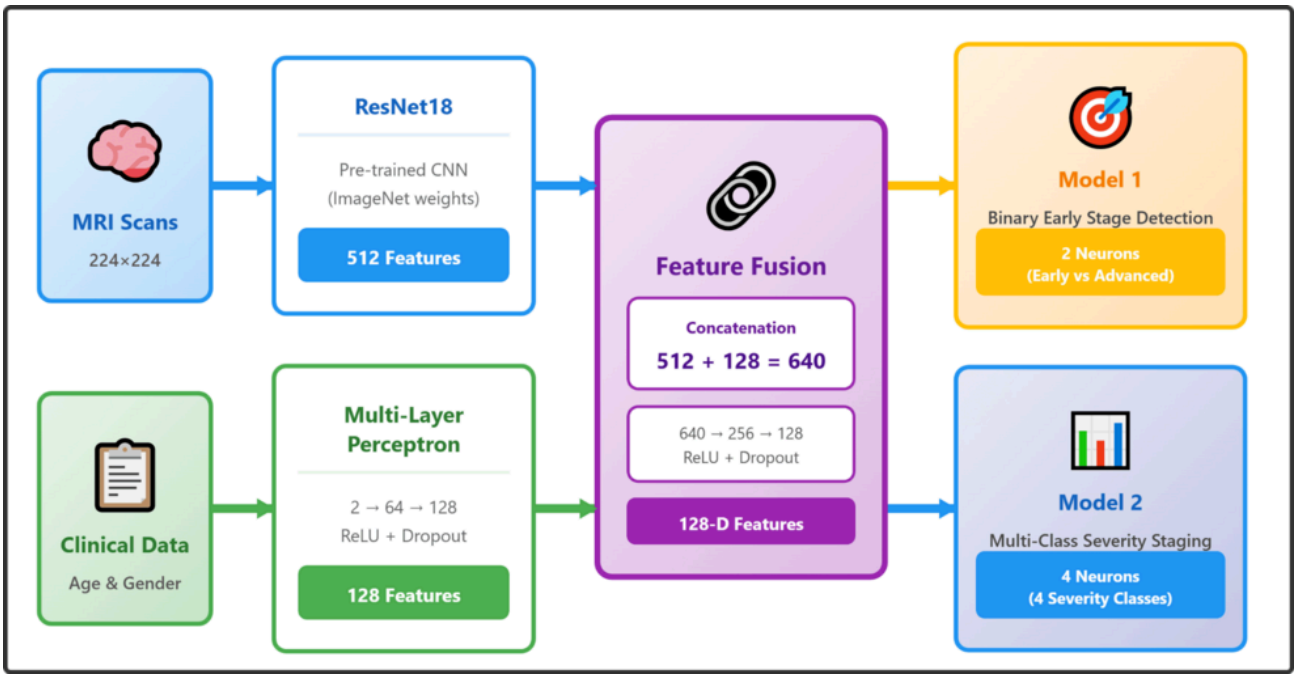


FIGURE 4: Multimodal Model Architecture

Model training

Training used 5-fold stratified cross-validation to ensure robust performance estimates. The training set was divided into five equal folds, maintaining class distribution proportions in each fold. Each fold served as the validation set once, while the other four folds were used for training. The loss function was Focal Loss, designed to handle severe class imbalance:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{2}$$

where p_t is the predicted probability for the true class, α_t is the class weight, and γ = 2.0 is the focusing parameter. Class weights were computed using balanced weighting:

$$w_c = \frac{N}{C \times N_c} \tag{3}$$

where N is the total number of samples, C is the number of classes, and N_c is the number of samples in class c. This gives higher weight to minority classes.

Adam Optimization was used with 0.0001 as its learning rate and weight decay (L2 regularization). The learning rate was reduced by 50% if validation loss did not improve for five epochs. Early stopping terminated training if validation accuracy showed no improvement for seven epochs. The maximum training was 50 epochs per fold, and the batch size was 32. Training was performed on NVIDIA Tesla T4 GPU, and all random seeds were set to 42 for reproducibility.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Model evaluation

Model performance was evaluated using these metrics: True Positives (TP) represent correctly identified positive cases, False Positives (FP) represent incorrectly classified positive cases, False Negatives (FN) represent missed positive cases, and True Negatives (TN) represent correctly identified negative cases. Precision measures the proportion of correct predictions among all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

where TP is true positives and FP is false positives.

Recall measures the proportion of actual positives correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

where FN is false negatives.

F1-score is the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Model discrimination was also assessed using the AUC. The AUC measures the model's ability to distinguish between classes across all classification thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t)) \tag{7}$$

where TPR (True Positive Rate) is Recall and FPR (False Positive Rate) is defined as:

$$\frac{FP}{FP + TN} \tag{8}$$

where TN is true negatives. AUC values range from 0 to 1, where 1.0 indicates perfect classification and 0.5 indicates random performance.

For multi-class receiver operating characteristic (ROC) analysis, the macro-average AUC provides an overall performance measure:

$$\text{Macro-Average AUC} = \frac{1}{C} \sum_{i=1}^{C} \text{AUC}_i \tag{9}$$

where AUC_i is the one-vs-rest AUC for class i, and C is the total number of classes.

The primary metric for Model 1 is recall because missing early stage patients in screening has severe clinical consequences. For Model 2, weighted F1-score is prioritized because it balances precision and recall while accounting for class imbalance across the four severity stages.

$$\text{Weighted F1-Score} = \frac{\sum_{i=1}^{C} w_i \times F1_i}{\sum_{i=1}^{C} w_i} \tag{10}$$

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

where w_i is the number of samples in class i (support), $F1_i$ is the F1-score for class i, and C is the total number of classes.

Results And Discussion

Experimental setup

All experiments were implemented in Python using PyTorch 1.12.0 and conducted on Google Colab with NVIDIA Tesla T4 GPU. Standard Python libraries including NumPy, Pandas, and Scikit-learn were used for data processing and evaluation. The dataset of 6,400 MRI scans was pre-split into training (5,121 samples) and testing (1,279 samples) sets. Model training used 5-fold stratified cross-validation with Focal Loss, Adam optimizer (learning rate = 0.0001), and early stopping (patience = 7 epochs). Data augmentation was applied only during training. Model performance was evaluated on the held-out test set using precision, recall, F1-score, and AUC-ROC metrics. The best-performing validation fold was selected for final test evaluation. All random processes used seed 42 for reproducibility.

Table *4* presents cross-validation results for both models across five folds. Model 1 achieved mean validation accuracy of 96.76% with standard deviation of 2.82%. Model 2 achieved mean validation accuracy of 75.90% with standard deviation of 5.15%.

Fold	Model 1: Early Detection (%)	Model 2: Severity Staging (%)
1	98.54	69.85
2	97.27	79.49
3	95.41	**82.52**
4	92.68	72.36
5	**99.90**	75.29
Mean	96.76 ± 2.82	75.90 ± 5.15

TABLE 4: Cross-Validation Results

Model 1 shows a higher mean accuracy and lower variance. The lower standard deviation indicates more consistent performance across different data splits. This demonstrates that distinguishing early stage from advanced stage disease is more reliable than fine-grained staging across four severity levels.

Test set performance: Model 1 (early detection)

Model 1 was evaluated using the best cross-validation fold (Fold 5 with a validation accuracy of 99.90%). Table *5* shows the performance metrics, and Figure *5* shows the confusion matrix, highlighting the high true positive rate for early stage detection compared to the advanced stage.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Class	Precision	Recall	F1-Score	Support
Early	0.9326	0.9917	0.9612	1088
Advanced	0.9262	0.5916	0.7220	191
Weighted Average	0.9316	0.9320	0.9255	1,279

TABLE 5: Performance of Model 1 (Early Detection)

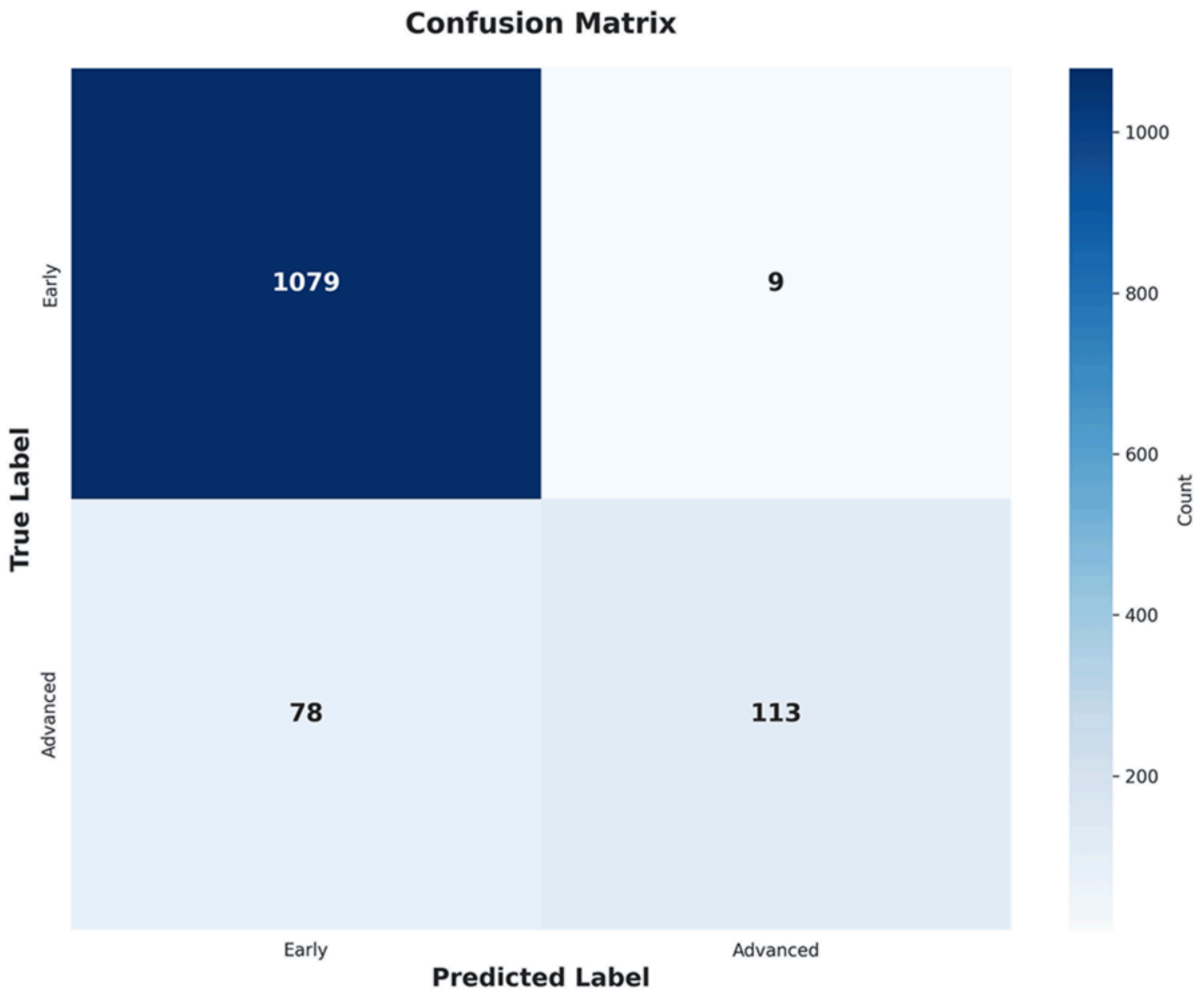


FIGURE 5: Confusion Matrix for Model 1 (Early-Stage Detection)

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

The performance of the model was excellent with an F1-score of 0.9612. Above all, the recall was 0.9917. Among the 1,088 patients for early stage, the model accurately identified 1,079 of them. This near perfect sensitivity is important for clinical screening to ensure intervention against disease progression, and it further identifies nearly all the patients requiring further assessment.

Good precision (0.9262) and lower recall (0.5916) were associated with advanced stage detection. The model accurately determined 113 out of 191 cases that were advanced, unlike 78 that were missed due to false classification as early cases. This represents appropriate caution rather than harmful misclassification because early-stage patients are monitored and provided with preventive care. The ROC curve in Figure *6* has an AUC of 0.971, which shows great discrimination between early stage and advanced stage.

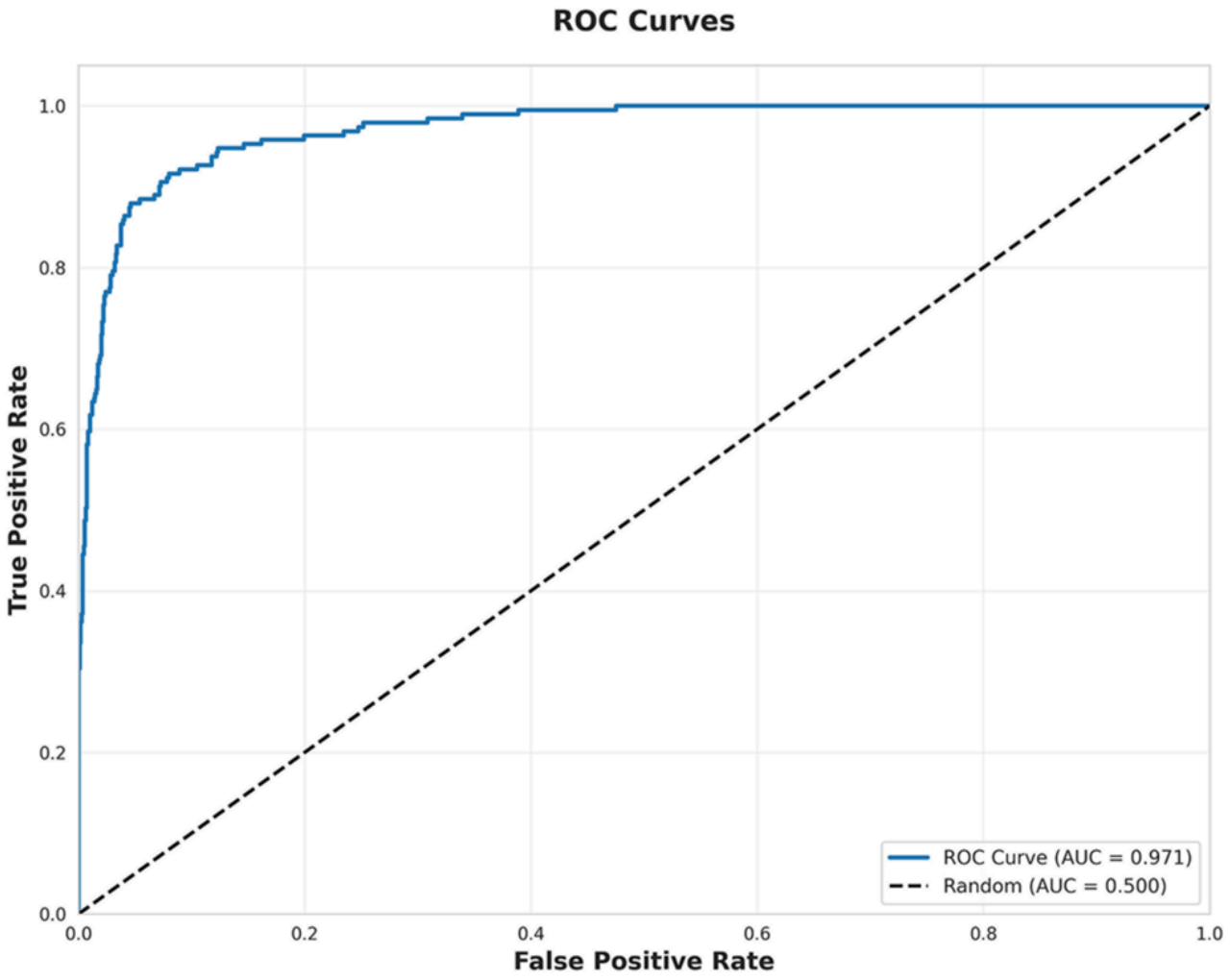


FIGURE 6: ROC Curve for Model 1 (Early-Stage Detection)

AUC, Area Under the ROC Curve; ROC, Receiver Operating Characteristic

Test set performance: Model 2 (severity staging)

The best cross-validation fold used to evaluate Model 2 was Fold 3, which has a validation accuracy of 82.52%. Table *6* indicates the performance metrics for each of the severity classes, and Figure *7* reveals the confusion matrix.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Class	Precision	Recall	F1-Score	Support
NonDemented	0.7957	0.8156	0.8056	640
VeryMildDemented	0.5860	0.6920	0.6346	448
MildDemented	0.9390	0.4302	0.5900	179
ModerateDemented	1.0000	1.0000	1.0000	12
Weighted Average	0.7442	0.7201	0.7173	1279

TABLE 6: Performance of Model 2

The model obtained a weighted F1 score of 0.7173, which shows it had a moderate performance in distinguishing the four severity stages. This is an expression of the difficulty in isolating the neighboring classes of disease, which have slight imaging distinctions. The score is lesser than that of Model 1 (0.9255), but it demonstrates that the model is capable of multi-class staging.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

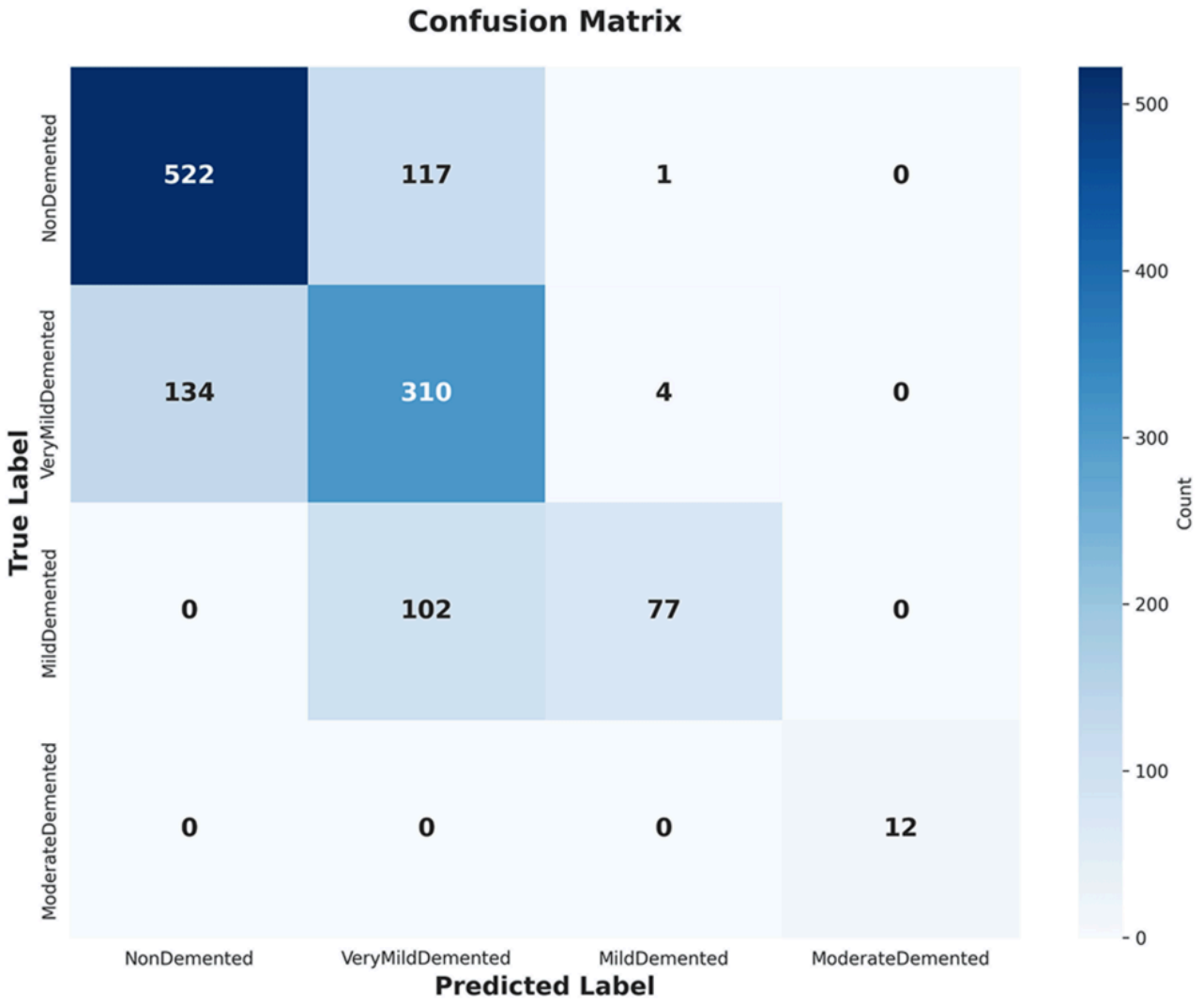


FIGURE 7: Confusion Matrix for Model 2 (Severity Staging)

As shown in Figure *7*, misclassification is concentrated between adjacent severity classes: 117 NonDemented cases were misclassified as VeryMildDemented, 134 VeryMildDemented cases as NonDemented, and 102 VeryMildDemented cases as MildDemented. This pattern reflects the gradual nature of Alzheimer's progression, where neighbouring stages share overlapping brain changes that are difficult to separate even on MRI scans.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

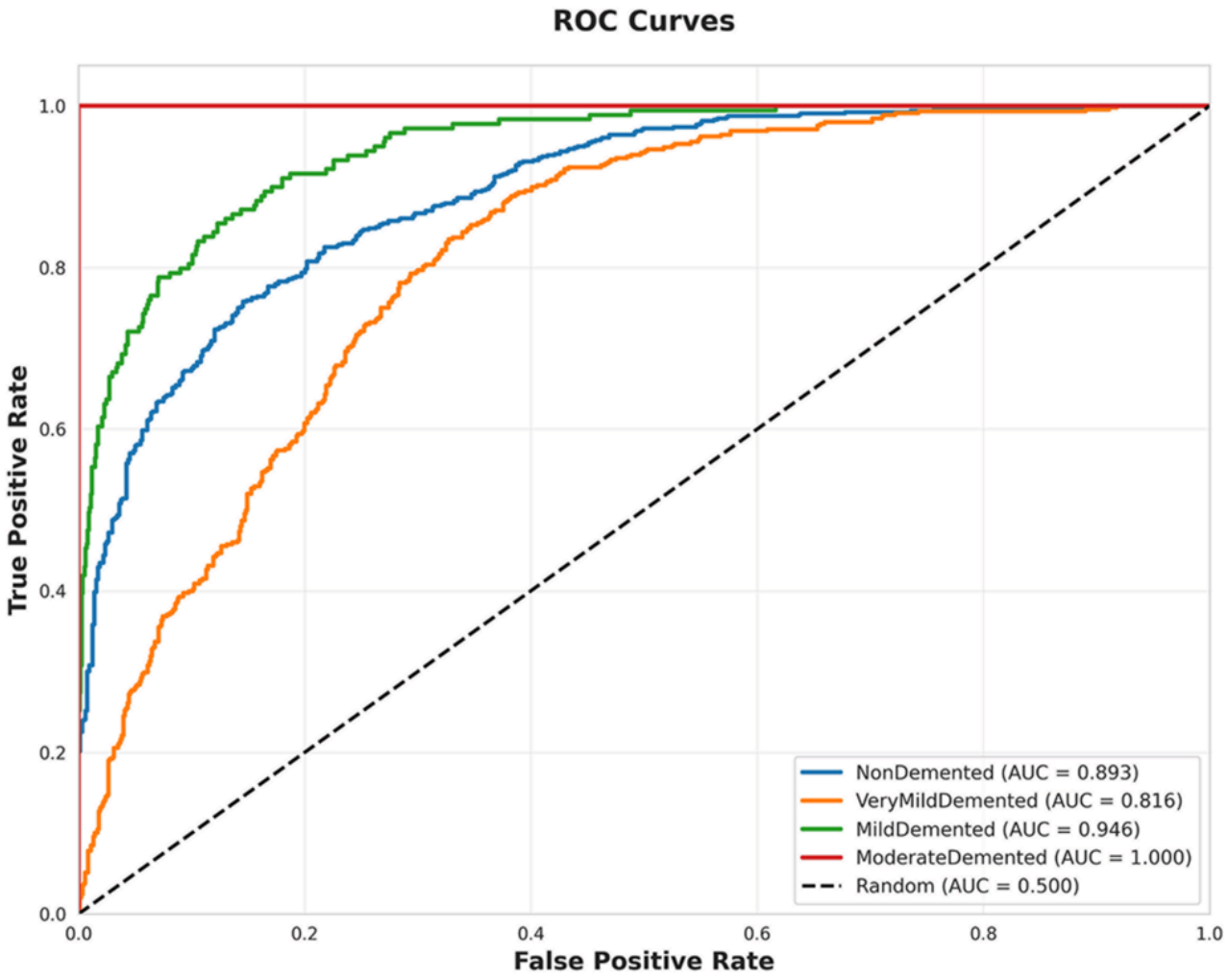


FIGURE 8: ROC Curves for Model 2 (Severity Staging)

AUC, Area Under the ROC Curve; ROC, Receiver Operating Characteristic

Figure *8* presents the ROC curves of each severity class where the macro-average AUC is 0.914, which reveals good overall discrimination ability. VeryMildDemented had the smallest AUC (0.816), indicating that it was difficult to distinguish between very early disease from normal aging, and ModerateDemented had a perfect discrimination (AUC = 1.000) because of a limited number of samples.

Comparison with MRI-only models

To verify that the clinical features contribute beyond what MRI alone provides, an MRI-only variant of each model was developed. The clinical MLP branch and the fusion step were removed from the architecture, while the classifier head was kept structurally identical (same depth, dropout of 0.6, and activations) with only the input dimensionality changed from 640 to 512. All other training settings such as the random seed, 5-fold cross-validation, Focal Loss, optimizer, scheduler, early stopping, augmentation, batch size, and epoch budget were held constant. Table *7* and Table *8* present the comparisons.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

Model	Configuration	Precision	Recall	F1-Score	AUC
Model 1	MRI-only	0.9221	0.9257	0.9228	0.9449
Model 1	Multimodal	0.9316	0.9320	0.9255	0.9710
Model 2	MRI-only	0.6636	0.6091	0.5911	0.8371
Model 2	Multimodal	0.7442	0.7201	0.7173	0.9140

TABLE 7: Test Performance of MRI-only Models vs. Multimodal Models

Precision, recall, and F1-score in Table *7* are weighted averages across classes, while AUC is reported as binary AUC for Model 1 and macro-average AUC for Model 2. For Model 1, the multimodal model recorded only a slight improvement in weighted F1-score (0.9255 against 0.9228), but a more noticeable improvement in AUC (0.9710 against 0.9449). For Model 2, the multimodal advantage was substantial. Weighted F1-score increased from 0.5911 to 0.7173, which is a 21.4% relative improvement. This indicates that the imaging features were insufficient for fine-grained severity staging on their own, and the inclusion of demographic features restored the discrimination between adjacent classes.

Fold	Model 1 MRI-only (%)	Model 1 Multimodal (%)	Model 2 MRI-only (%)	Model 2 Multimodal (%)
1	98.93	98.54	63.22	69.85
2	96.19	97.27	71.19	79.49
3	96.39	95.41	57.62	82.52
4	99.61	92.68	68.16	72.36
5	99.80	99.90	70.51	75.29
Mean	98.18 ± 1.58	96.76 ± 2.82	66.14 ± 4.93	75.90 ± 5.15

TABLE 8: Cross-Validation Results of MRI-only Models vs. Multimodal Models

Table *8* shows that both configurations in Model 1 had comparable mean validation accuracies (98.18% for MRI-only and 96.76% for multimodal). This reveals that the binary screening task was largely saturated by the imaging signal alone, and the clinical features contributed marginally. For Model 2, the cross-validation mean improved from 66.14% to 75.90%,

with the multimodal model outperforming MRI-only on every fold. Overall, the contribution of clinical features depends on the difficulty of the task. It was marginal in binary screening, where MRI alone proved sufficient, and substantial in fine-grained severity staging.

Conclusions

This research investigated two gaps in the classification of Alzheimer's disease: insufficient data and the use of single imaging modality. The binary early detection model (recall = 0.9917, F1-score = 0.9612) accurately identified almost all patients diagnosed with early stage of the disease. This level of performance renders it appropriate for screening, where detection of missing cases has severe impacts. The four-class severity staging model had a lower performance with a weighted F1-score of 0.7173. It is effective in early stages (NonDemented F1-Score = 0.8056, VeryMildDemented F1-Score = 0.6346) but has difficulty with later stages because of class imbalance.

The difference in the results of both models demonstrates that easier tasks provide more consistent results. The binary classification is suitable in screening, and four-class staging could be enhanced with more data. Subsequent studies should involve bigger and balanced datasets at all disease stages, particularly at the stage of ModerateDemented, which used a limited sample of 52 subjects to train. Both models might be enhanced by the addition of richer inputs such as cognitive test scores and genetic information. Additionally, since the data originates from a single repository, testing on independent datasets from diverse populations is needed to confirm generalizability. Our main finding is that the contribution of demographic clinical features depends on the difficulty of the task. MRI alone was sufficient for binary screening in this study, while the addition of clinical features provided substantial improvement in fine-grained severity staging.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ayanfeoluwa F. Owoseeni, Racheal S. Akinbo, Bamidele M. Kuboye, Joseph O. Igbogbo

Acquisition, analysis, or interpretation of data: Ayanfeoluwa F. Owoseeni, Racheal S. Akinbo, Bamidele M. Kuboye, Joseph O. Igbogbo, Tunde J. Fatoke

Drafting of the manuscript: Ayanfeoluwa F. Owoseeni, Racheal S. Akinbo, Joseph O. Igbogbo

Critical review of the manuscript for important intellectual content: Ayanfeoluwa F. Owoseeni, Racheal S. Akinbo, Bamidele M. Kuboye, Joseph O. Igbogbo, Tunde J. Fatoke

Supervision: Racheal S. Akinbo, Bamidele M. Kuboye, Tunde J. Fatoke

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request. The MRI dataset supporting this study is openly available on Mendeley Data at https://doi.org/10.17632/xx9zzz6t54.1

References

1. Dhapola R, Sarma P, Medhi B, Prakash A, Reddy DH: Recent advances in molecular pathways and therapeutic implications targeting mitochondrial dysfunction for Alzheimer's disease. Molecular Neurobiology. 2022, 59:535-555. 10.1007/s12035-021-02612-6
2. Rafii MS, Aisen PS: Detection and treatment of Alzheimer's disease in its preclinical stage. Nature Aging. 2023, 3:520-531. 10.1038/s43587-023-00410-4
3. Mohsen S: Alzheimer's disease detection using deep learning and machine learning: a review. Artificial Intelligence Review. 2025, 58:262. 10.1007/s10462-025-11258-y
4. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD: Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific Reports. 2021, 11:3254. 10.1038/s41598-020-74399-w
5. Singh SG, Das D, Barman U, Saikia MJ: Early Alzheimer's disease detection: A review of machine learning techniques for forecasting transition from mild cognitive impairment. Diagnostics. 2024, 14:1759. 10.3390/diagnostics14161759
6. Sarvamangala DR, Kulkarni RV: Convolutional neural networks in medical image understanding: a survey. Evolutionary Intelligence. 2022, 15:1-22. 10.1007/s12065-020-00540-3
7. Sistaninejhad B, Rasi H, Nayeri P: A review paper about deep learning for medical image analysis. Computational and Mathematical Methods in Medicine. 2023, 2023:7091301. 10.1155/2023/7091301
8. Lim BY, Lai KW, Haiskin K, et al.: Deep learning model for prediction of progressive mild cognitive impairment to Alzheimer's disease using structural MRI. Frontiers in Aging Neuroscience. 2022, 14:876202-876202. 10.3389/fnagi.2022.876202
9. Odusami M, Maskeliūnas R, Damaševičius R, Misra S: Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images. Journal of Medical and Biological Engineering. 2023, 43:291-302. 10.1007/s40846-023-00801-3
10. Khan R, Akbar S, Mehmood A, et al.: A transfer learning approach for multiclass classification of Alzheimer's disease using MRI images. Frontiers in Neuroscience. 2023, 16:1050777. 10.3389/fnins.2022.1050777
11. El-Assy AM, Amer HM, Ibrahim HM, Mohamed MA: A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data. Scientific Reports. 2024, 14:3463. 10.1038/s41598-024-53733-6
12. Gasmi K, Alyami A, Hamid O, et al.: Optimized hybrid deep learning framework for early detection of Alzheimer's disease using adaptive weight selection. Diagnostics. 2024, 14:2779. 10.3390/diagnostics14242779
13. Wang Y, Gao R, Wei T, et al.: Predicting long-term progression of Alzheimer's disease using a multimodal deep learning model incorporating interaction effects. Journal of Translational Medicine. 2024, 22:265. 10.1186/s12967-024-05025-w
14. Alorf A: Transformer and convolutional neural network: A hybrid model for multimodal data in multiclass classification of Alzheimer's disease. Mathematics. 2025, 13:1548. 10.3390/math13101548
15. Mmadumbu AC, Saeed F, Ghaleb F, Qasem SN: Early detection of Alzheimer's disease using deep learning methods. Alzheimer's & Dementia. 2025, 21:e70175. 10.1002/alz.70175
16. Yan H, Mubonanyikuzo V, Komolafe TE, Zhou L, Wu T, Wang N: Hybrid-RViT: Hybridizing ResNet-50 and vision transformer for enhanced Alzheimer's disease detection. PLoS ONE. 2025, 20:e0318998. 10.1371/journal.pone.0318998
17. Sufian A: Advancing Alzheimer's disease detection in clinical settings: MRI image data [Dataset]. Mendeley Data. (2025). https://doi.org/10.17632/xx9zzz6t54.1.
18. Livingston G, Huntley J, Sommerlad A, et al.: Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet. 2020, 396:413-446. 10.1016/s0140-6736(20)30367-6
19. Shorten C, Khoshgoftaar TM: A survey on image data augmentation for deep learning. Journal of Big Data. 2019, 6:60. 10.1186/s40537-019-0197-0

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4

20. Liu Z: A look at the use of the ResNet18 model in medical and industrial visual checking. Applied and Computational Engineering. 2025, 196:38-45. 10.54254/2755-2721/2025.ld28277

21. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J: Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE Transactions on Medical Imaging. 2016, 35:1299-1312. 10.1109/tmi.2016.2535302

22. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research. 2014, 15:1929-1958.

How to cite this article:
Akinbo R S, Kuboye B M, Owoseeni A F, et al. (May 13, 2026) Development of Multimodal Deep Learning Models for Early Detection and Severity Staging of Alzheimer's Disease. Cureus J Comput Sci 3 : es44389-026-00072-4. DOI https://doi.org/10.7759/s44389-026-00072-4