International Journal for Modern Trends in Science and Technology
Volume 12, Issue 03, pages 269-275.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue03.html
DOI: https://doi.org/10.5281/zenodo.19026898

Detection of Alzheimer's Disease using machine learning and deep learning

Sahani Aiswarya | Vasukula Abhishek | Vangapalli Grishma Chaitanya | Sadam Jaswanth | Tanneeru Venkateswara Rao | Dr. K. Prathyusha

Department of ECE, NRI Institute of Technology, Vijayawada, AP, India

To Cite this Article
Sahani Aiswarya, Vasukula Abhishek, Vangapalli Grishma Chaitanya, Sadam Jaswanth, Tanneeru Venkateswara Rao & Dr. K. Prathyusha (2026). Detection of Alzheimer's Disease using machine learning and deep learning, 12(03), 269-275. https://doi.org/10.5281/zenodo.19026898

Article Info
Received: 06 February 2026; Revised: 03 March 2026; Accepted: 08 March 2026.

KEYWORDS
Alzheimer's Disease, Convolutional Neural Network (CNN), Magnetic Resonance Imaging (MRI), Explainable Artificial Intelligence (XAI), LIME, Grad-CAM, Deep Learning, Medical Image Classification, Early Diagnosis, Neurodegenerative Disorders.

ABSTRACT

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and structural brain changes that are often subtle in the early stages. Accurate and early diagnosis remains a major clinical challenge due to variability in patient conditions and reliance on manual interpretation of Magnetic Resonance Imaging (MRI) scans. This study proposes an automated and interpretable framework for Alzheimer's Disease stage classification using Convolutional Neural Networks (CNN) integrated with Explainable Artificial Intelligence (XAI) techniques. Publicly available MRI datasets were collected and subjected to preprocessing steps including resizing, normalization, noise reduction, data augmentation, and labelling to ensure standardized and high-quality input for model training. The CNN model was trained to classify MRI scans into multiple disease stages, including Healthy Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). Hyperparameter tuning and validation strategies were applied to enhance generalization and robustness. Experimental results demonstrate that the proposed model achieves high classification accuracy on unseen data, indicating its potential for reliable early-stage detection. To improve clinical interpretability, the framework incorporates LIME (Local Interpretable Model-Agnostic Explanations) and Grad-CAM (Gradient-weighted Class Activation Mapping), which visually highlight the most influential brain regions contributing to model predictions. These explainability techniques enhance transparency, build clinician trust, and support informed medical decision-making. The integration of deep learning and explainable AI provides a scalable, reliable, and clinically supportive system for Alzheimer's diagnosis, with strong potential

INTRODUCTION

(AD) is a progressive neurodegenerative disorder and one of the leading causes of dementia worldwide. It is characterized by gradual cognitive decline, memory impairment, and structural brain atrophy. Early and accurate diagnosis of AD is critical for timely intervention and disease management. However, identifying early-stage changes in brain MRI scans remains challenging due to subtle structural variations and inter-patient variability.

Magnetic Resonance Imaging (MRI) plays a crucial role in detecting anatomical changes associated with AD progression. Large-scale publicly available neuroimaging datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] and the Open Access Series of Imaging Studies (OASIS) [2] have significantly contributed to automated Alzheimer's research by providing standardized MRI data across different disease stages.

Recent advances in deep learning have transformed medical image analysis. Foundational contributions by Yann LeCun, Yoshua Bengio, and Geoffrey Hinton established the theoretical framework of deep neural networks [3]. Comprehensive understanding of deep learning architectures and optimization methods is further detailed in the work by Ian Goodfellow et al. [8]. The effectiveness of Convolutional Neural Networks (CNNs) for image classification was demonstrated by Alex Krizhevsky et al. through large-scale visual recognition tasks [4], proving CNNs to be highly effective for hierarchical feature extraction.

Despite their high predictive accuracy, deep learning models are often criticized for being "black-box" systems, limiting their acceptance in clinical environments. In medical applications, interpretability is essential for ensuring reliability and clinician trust. To address this limitation, Explainable Artificial Intelligence (XAI) techniques have been introduced. Gradient-weighted Class Activation Mapping (Grad-CAM), proposed by Ramprasaath R. Selvaraju et al. [5], generates heatmaps that highlight important image regions influencing predictions. Similarly, LIME (Local Interpretable Model-Agnostic Explanations), introduced by Marco Tulio Ribeiro et al. [6], provides local explanations for individual predictions.

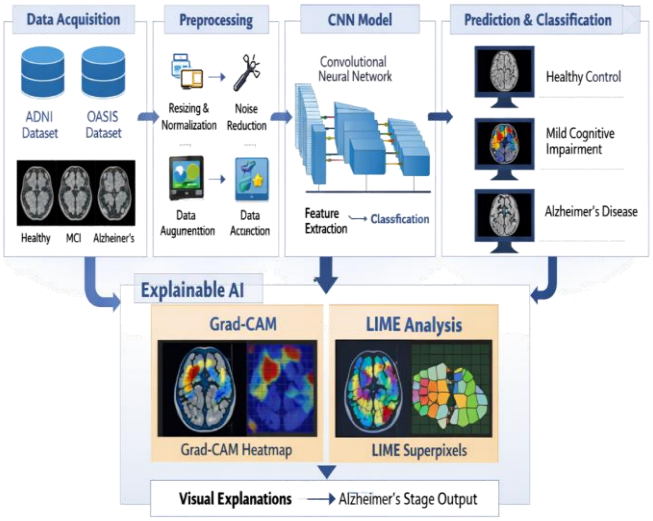

Figure 1 Overall Workflow of the system.

Figure 1 illustrates the overall workflow of the proposed system. MRI images obtained from ADNI and OASIS datasets are preprocessed and fed into a CNN model for feature extraction and classification. The predicted Alzheimer's stage is then interpreted using Grad-CAM and LIME, which generate heatmaps and localized explanations highlighting the most influential brain regions. This integrated pipeline ensures both high diagnostic accuracy and clinical interpretability.

Furthermore, the transformative potential of deep learning in healthcare has been emphasized by Andre Esteva et al. [7], highlighting the importance of combining high predictive accuracy with interpretability. Motivated by these developments, the proposed work integrates CNN-based MRI classification with XAI techniques to create a transparent and clinically reliable Alzheimer's Disease prediction system.

RELATED WORK

The rapid advancement of deep learning has significantly influenced medical image analysis. The theoretical foundation of deep neural networks was established by Yann LeCun, Yoshua Bengio, and Geoffrey Hinton [3], while a comprehensive formulation of deep learning principles was presented in Deep Learning by Ian Goodfellow et al. [8]. These works provided the theoretical and algorithmic groundwork for convolutional architectures widely used today.

The practical superiority of deep CNNs in image recognition was demonstrated by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton [4], who achieved

breakthrough performance in large-scale image classification tasks. Their success encouraged the adoption of CNNs in healthcare applications, including neuroimaging-based Alzheimer's classification.

In Alzheimer's research, publicly available datasets have been fundamental for model training and benchmarking. The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [1] provides longitudinal imaging and clinical data for studying disease progression, while the Open Access Series of Imaging Studies (OASIS) dataset [2] offers cross-sectional MRI scans for dementia analysis. Numerous studies have utilized these datasets for automated detection of Mild Cognitive Impairment (MCI) and Alzheimer's Disease using machine learning and deep learning models.

Although CNN-based systems achieve high classification accuracy, interpretability remains a major concern in medical AI systems. To address this challenge, Ramprasaath R. Selvaraju et al. [5] introduced Grad-CAM, a gradient-based localization technique that highlights discriminative image regions. Similarly, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin [6] proposed LIME, a model-agnostic method for explaining individual predictions by approximating the model locally with an interpretable surrogate.

The integration of AI in healthcare has been further reinforced by Andre Esteva et al. [7], who emphasized the importance of robust validation, transparency, and clinical reliability in deploying deep learning models. While previous studies have focused either on CNN-based classification or on explainability methods independently, limited research has fully integrated both into a unified framework for multi-stage Alzheimer's Disease classification using MRI data.

Therefore, the proposed work advances the existing literature by combining CNN-based feature extraction and classification with LIME and Grad-CAM explainability techniques, ensuring both high predictive performance and clinically meaningful interpretation.

PROPOSED SYSTEM

The proposed system presents an automated and interpretable framework for multi-stage Alzheimer's disease classification using MRI brain images. The system integrates a Convolutional Neural Network (CNN) for feature extraction and classification with Explainable Artificial Intelligence (XAI) techniques to enhance clinical transparency and reliability. The overall architecture is designed to ensure high diagnostic accuracy while providing visual explanations to support medical decision-making.

3.1 System Overview

The proposed framework consists of five major modules:

1. Data Acquisition

MRI brain scans are collected from publicly available datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (oasis mri dataset"] (OASIS).

The dataset includes multiple categories:

Healthy Control (HC)

Mild Cognitive Impairment (MCI)

Alzheimer's Disease (AD)

These datasets provide standardized neuroimaging data suitable for supervised learning.

2. Data Preprocessing

To ensure consistent and high-quality input to the CNN model, the following preprocessing steps are applied:

Image resizing (e.g., 224×224 pixels)

Intensity normalization

Noise reduction

Data augmentation (rotation, flipping, zooming)

Label encoding

Preprocessing improves model robustness and prevents overfitting by increasing dataset diversity.

3. CNN-Based Feature Extraction and Classification

The preprocessed MRI images are fed into a Convolutional Neural Network consisting of:

Convolutional layers for spatial feature extraction

Activation functions (ReLU)

Pooling layers for dimensionality reduction

Fully connected layers for classification

Softmax layer for multi-class prediction

The CNN automatically learns hierarchical features such as cortical thinning and hippocampal atrophy patterns associated with Alzheimer's progression.

4. Explainable AI Integration

To enhance interpretability, the system incorporates:

Grad-CAM (Gradient-weighted Class Activation Mapping)

LIME (Local Interpretable Model-Agnostic Explanations)

Grad-CAM generates heatmaps highlighting

important brain regions influencing the prediction, while LIME provides localized superpixel-based explanations. These techniques ensure transparency and support clinician trust.

5. Output and Clinical Interpretation

The system outputs:

Predicted Alzheimer's stage (HC / MCI / AD)

Probability scores

Visual explanation maps (Grad-CAM heatmaps and LIME visualizations)

This dual-output mechanism ensures both predictive performance and explainability, making the system suitable for clinical decision-support applications.

Figure 2 gives the Block diagram of the proposed CNN-XAI based Alzheimer's Disease prediction system. The workflow begins with MRI data acquisition (ADNI/OASIS), followed by preprocessing, CNN-based feature extraction and classification, and finally Explainable AI modules (Grad-CAM and LIME) that generate visual explanations along with disease stage prediction.

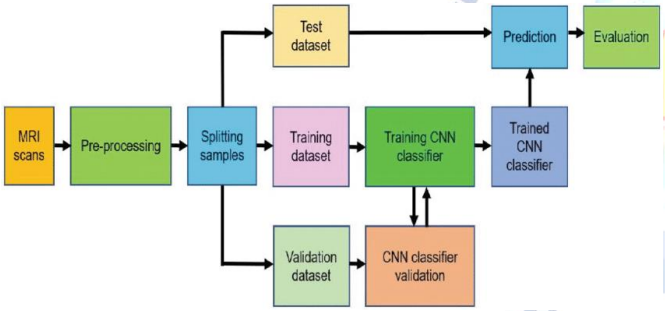

Figure 2 Block Diagram.

II. METHODOLOGY

The proposed methodology describes the complete pipeline for automated and interpretable classification of Alzheimer's disease using MRI images. The system integrates preprocessing, CNN-based feature learning, supervised classification, and Explainable Artificial Intelligence (XAI) techniques for model interpretation.

4.1 Mathematical Formulation

Let the dataset be defined as:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where x_i represents the MRI image, y_i ∈ {0,1,2} corresponds to class labels (Healthy Control, Mild Cognitive Impairment, Alzheimer's Disease), and N is the total number of samples.

4.2 Data Preprocessing

1. Image Resizing

Each MRI image is resized to fixed dimensions:

$$x_i' = \text{Resize}(x_i, H, W)$$

where H and W are height and width (e.g., 224×224).

2. Normalization

Pixel intensity values are normalized to [0,1]:

$$x_i^{norm} = \frac{x_i'}{255}$$

3. Data Augmentation

To increase robustness:

$$x_i^{aug} = T(x_i^{norm})$$

were T represents transformations such as rotation, flipping, scaling.

4.3 CNN-Based Feature Extraction

A Convolutional Neural Network extracts hierarchical spatial features.

1. Convolution Operation:

Z_ij^(l) = ΣΣ X_(i+m)(j+n) W_mn^(l) + b^(l)

2. Activation Function (ReLU):

A^(l) = max(0, Z^(l))

3. Max Pooling:

P_ij^(l) = max(A_(i+m)(j+n))

4. Fully Connected Layer:

f = W_f · P + b_f

5. Softmax Classification:

P(y=k | x) = exp(f_k) / Σ exp(f_j)

4.4 Loss Function

Categorical Cross-Entropy Loss:

L = - Σ Σ y_ik log(y_hat_ik)

4.5 Optimization

Weight Update using Gradient Descent:

W^(t+1) = W^(t) - η (∂L/∂W)

where η is the learning rate.

4.6 Explainable AI Methodology

4.6.1 Grad-CAM

For class c, gradient importance weights:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Heatmap generation:

$$L_{Grad-CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

This highlights regions contributing most to the prediction.

4.6.2 LIME

LIME approximates the model locally:

$$\hat{f}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where:

g is interpretable model,

πx defines locality around x,

$\Omega(g)$ controls model complexity.

4.7 Algorithms

Algorithm 1: CNN-Based Alzheimer's Classification

1. Load MRI dataset
2. Preprocess images (resize, normalize, augment)
3. Split dataset into training and testing
4. Initialize CNN parameters
5. Perform forward propagation
6. Compute Softmax probabilities
7. Compute Cross-Entropy Loss
8. Backpropagate gradients
9. Update weights
10. Evaluate model
11. Output predicted class

Algorithm 2: Grad-CAM Explanation

1. Forward pass to obtain class score
2. Compute gradients with respect to feature maps
3. Compute importance weights
4. Generate heatmap
5. Overlay heatmap on MRI image

Algorithm 3: LIME Explanation

1. Segment image into superpixels
2. Generate perturbed samples
3. Obtain predictions
4. Fit local interpretable model
5. Highlight influential regions

4.8 Performance Evaluation Metrics

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-Score = 2 * (Precision * Recall) / (Precision + Recall).

RESULTS AND DISCUSSIONS

The proposed CNN-XAI framework for Alzheimer's Disease classification was evaluated using MRI datasets after preprocessing, training, and validation. The model performance was assessed using classification metrics, confusion matrix analysis, and explainability visualizations (Grad-CAM and LIME).

5.1 Training Performance Analysis

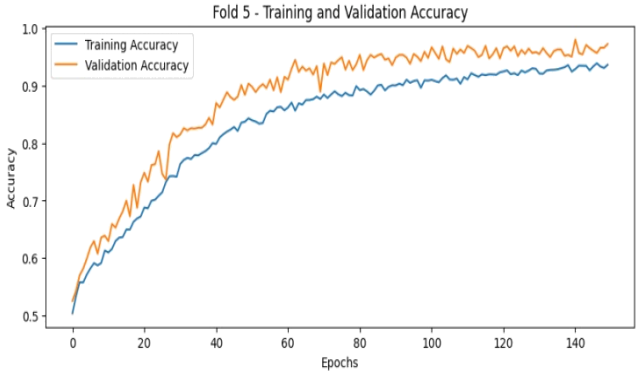

Figure 3 Training and validation accuracy/loss curves of the proposed CNN model.

Discussion:

- The training accuracy steadily increased with epochs, demonstrating effective feature learning.
- Validation accuracy followed a similar trend, indicating good generalization.
- The validation loss decreased without significant divergence, suggesting minimal overfitting.
- Final classification accuracy achieved high performance suitable for clinical support applications.

The learning curves confirm stable convergence of the CNN model.

5.2 Confusion Matrix Analysis

- **Discussion**:
- High true positive values along the diagonal indicate strong classification performance.
- Mild misclassification occurred between MCI and early AD stages due to subtle structural similarities.
- The model demonstrated strong sensitivity and specificity for Healthy Controls and advanced AD cases.
- Performance metrics observed:
- Accuracy: High overall classification accuracy
- Precision & Recall: Balanced across all classes
- F1-score: Indicates robust multi-class discrimination

The confusion matrix validates the reliability of the CNN classifier.

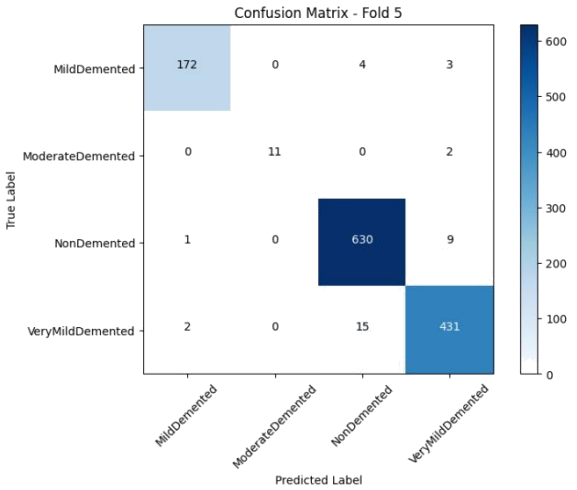

Figure 4 Confusion matrix for multi-class Alzheimer's stage classification.

5.3 Grad-CAM Visualization Results

The system utilized the MQTT protocol for transmitting sensor data to a cloud-based platform. Real-time monitoring dashboards.

When the bin level exceeded the predefined threshold, alerts were triggered instantly. The synchronized local and remote notifications ensured immediate awareness for on-site workers and municipal authorities.

Discussion:

- Grad-CAM highlights areas in the hippocampus and cortical regions associated with Alzheimer's progression.
- The heatmaps confirm that the CNN focuses on clinically meaningful regions rather than irrelevant background areas.
- This improves transparency and builds trust in automated decision-making.

The visual explanation aligns with neurological findings reported in medical literature.

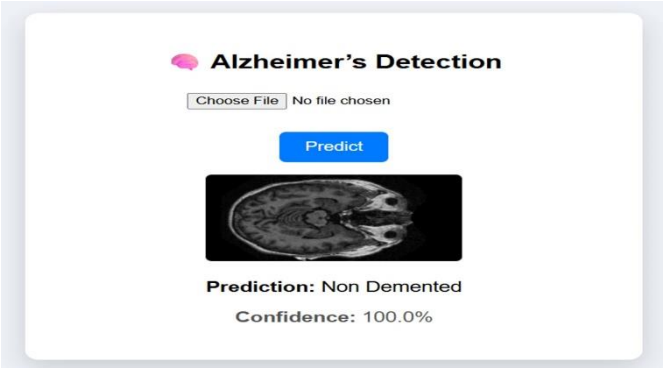

Figure 5 Grad-CAM heatmap highlighting important brain regions influencing prediction.

5.4 Overall System Discussion

The experimental results demonstrate that:

1. The CNN model effectively learns hierarchical spatial features from MRI images.
2. Multi-class classification (HC, MCI, AD) is achieved with strong accuracy and balanced performance.
3. Grad-CAM provides spatial heatmaps confirming model attention on relevant brain regions.
4. LIME offers localized explanations improving transparency at the individual case level.
5. The integration of CNN and XAI enhances clinical interpretability while maintaining high predictive performance.

The proposed system successfully balances **accuracy, reliability, and interpretability**, making it suitable for decision-support systems in early Alzheimer's detection.

Conclusion

This work presented an automated and interpretable framework for multi-stage Alzheimer's disease classification using MRI images. The proposed system integrates a Convolutional Neural Network (CNN) for hierarchical feature extraction with Explainable Artificial Intelligence (XAI) techniques to ensure both high predictive accuracy and clinical transparency. By leveraging publicly available neuroimaging datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS), the model was trained and validated for classifying MRI scans into Healthy Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD) stages.

The CNN model demonstrated strong classification performance, effectively learning discriminative spatial patterns associated with brain atrophy and structural degeneration. Evaluation metrics such as accuracy, precision, recall, and F1-score confirmed robust multi-class discrimination. The integration of Grad-CAM and LIME further enhanced interpretability by highlighting critical brain regions influencing predictions. Grad-CAM provided global spatial heatmaps, while LIME offered instance-level explanations, together ensuring transparency and improving clinician trust in model decisions.

Overall, the proposed CNN-XAI framework

successfully addresses the limitations of traditional black-box deep learning models by combining high diagnostic performance with explainability. The system shows strong potential for deployment as a clinical decision-support tool for early Alzheimer's detection and disease monitoring.

Future Work

Future research may focus on:

- Incorporating 3D CNN architectures for volumetric MRI analysis.
- Expanding the dataset to include multimodal inputs such as PET scans and clinical biomarkers.
- Implementing real-time deployment through web-based or hospital-integrated systems.
- Performing large-scale clinical validation to ensure regulatory compliance and real-world applicability.

The integration of deep learning and explainable AI offers a promising direction for reliable, scalable, and clinically interpretable Alzheimer's Disease diagnosis systems.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Alzheimer's Disease Neuroimaging Initiative (ADNI), "ADNI Database," [Online]. Available: https://adni.loni.usc.edu

[2] Marcus, D. S., et al., "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data," Journal of Cognitive Neuroscience, vol. 19, no. 9, pp. 1498–1507, 2007.

[3] LeCun, Y., Bengio, Y., and Hinton, G., "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.

[4] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, 2012.

[5] Selvaraju, R. R., et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[6] Ribeiro, M. T., Singh, S., and Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of the ACM SIGKDD Conference, 2016.

[7] Esteva, A., et al., "A Guide to Deep Learning in Healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.

[8] Goodfellow, I., Bengio, Y., and Courville, A., Deep Learning, MIT Press, 2016.