An Explainable Deep Learning Model for Classification and Analysis of Alzheimer's Disease for Clinical Trust

H. C. Bharath
Bapuji Institute of Engineering and Technology, Davangere, Affiliated to Visvesvaraya Technological University, Belagavi-590018, India
bharathhonnali@gmail.com

N. Pradeep
Bapuji Institute of Engineering and Technology, Davangere, Affiliated to Visvesvaraya Technological University, Belagavi-590018, India
nmnpradeep@gmail.com

R. Shashidhar
Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysuru-570006, India
shashidhar.r@jssstuniv.in

Pavitha Nooji
Department of Artificial Intelligence, Faculty of Science and Technology, Vishwakarma University, Pune-411048, India
pavitha.nooji@vupune.ac.in

J. Meghana
Department of Computer Applications, JSS Science and Technology University, Mysuru-570006, India
meghanaj@jssstuniv.in (corresponding author)

Received: 28 June 2025 | Revised: 25 July 2025, 20 August 2025, 13 September 2025, and 16 September 2025 | Accepted: 18 September 2025

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.12997*

ABSTRACT

Alzheimer's Disease (AD), a progressive neurodegenerative disorder, presents significant challenges in early diagnosis and treatment. While deep learning models have demonstrated high accuracy in analyzing medical imaging data, their lack of interpretability limits clinical adoption. This study proposes an explainable deep learning framework that integrates a Convolutional Neural Network (CNN) with the Local Interpretable Model-agnostic Explanations (LIME) technique to enhance transparency and clinical trust in AD classification. The model highlights key biomarkers, such as hippocampal atrophy and amyloid plaque density, that contribute to its predictions. The approach addresses the need for consistent evaluation metrics and the integration of domain expertise in AI-driven diagnosis. Experiments conducted on the OASIS benchmark dataset achieved a classification accuracy of 97%, a precision of 97.2%, a recall of 96.8%, and an AUC of 0.97, with LIME providing localized, interpretable insights that align with clinical understanding. By combining predictive performance with explainability, the framework addresses ethical concerns by fostering transparency, accountability, trustworthiness, and practical AI-assisted diagnostic tools for precision healthcare.

Keywords-Alzheimer's Disease (AD); deep learning; Explainable Artificial Intelligence (XAI); Local Interpretable Model-agnostic Explanations (LIME); decision support systems

I. INTRODUCTION

AD is a long-term degenerative neurological condition and the most widespread form of dementia. It represents a significant public health issue with an increasing rate amid the aging population [1]. Early clinical intervention requires early diagnosis, which is challenging since symptoms are similar and the structural brain alterations are not obvious at the initial stages of AD.

New trends in deep learning, specifically, CNNs, have resulted in tremendous progress in the neuroimaging analysis of learning discriminative features directly on raw MRI and PET scans [2, 3]. CNNs are superior to the conventional hand-crafted feature-based Machine Learning (ML) models, demonstrating superior sensitivity and accuracy to AD classification tasks [4]. However, they are not interpretable and, hence, cannot be directly applicable in clinical practice. CNNs have a problem of being a black box, which cannot be trusted by medical professionals who need to see how decisions were made to ensure reliability in the diagnosis [5, 6].

To overcome this limitation, Explainable Artificial Intelligence (XAI) methods have become critical to serve as the middle ground between predictive accuracy and explainability. The most popular ones include LIME [7] and Shapley Additive Explanations (SHAP) [8]. Although SHAP is a global feature importance that provides results using game theory, LIME creates locally faithful approximations that are specifically helpful in visual tasks, such as determining regions of interest in brain scans. Although the XAI approaches have been successful, there has not been a consistent combination of these approaches with high-performing CNNs in AD studies, making clinical translation limited.

The present paper proposes a novel CNN-based diagnostic framework integrated with LIME to classify AD stages and provide visual explanations aligned with clinical biomarkers, such as hippocampal atrophy and cortical thinning [9]. On the OASIS dataset, the model achieved 97% accuracy, outperforming existing traditional and deep learning approaches. Its novelty lies in combining localized interpretability with biomarker-level validation, allowing clinicians to directly correlate model outputs with anatomical relevance.

The major interpretability problem in deep learning models is addressed using the integration of LIME to increase model transparency in the research study. The major contributions of the current work are:

- The study integrates LIME with CNN to address the interpretability gap in deep learning-based AD classification, enabling clinically relevant explanations that enhance both diagnostic confidence and potential adoption in precision healthcare.
- It correlates model-extracted features with established AD biomarkers, such as hippocampal atrophy and amyloid plaque density, to improve the clinical relevance of AI predictions, which supports domain-aware interpretability analysis.

Authors in [10] investigated boosting ensemble ML schemes, such as XGBoost, LightGBM, and Gradient Boosting (GB), for AD diagnosis and SHAP for feature selection. They proposed a scheme that achieved efficient results, with an accuracy of more than 94% and with a minimum number of features for the detection process. Authors in [11] created a deep learning model capable of providing instant explanations alongside AD predictions during clinical assessments. The specific technology is appropriate for real-world healthcare situations due to its speed and transparency. Authors in [12] presented an XAI Kernel Extreme Learning Machine Enhanced with the Crowned Porcupine Optimization Algorithm (XAIKELM-ICPOA), as a tool of XAI, and it was utilized to provide insights into how features contribute to decisions. The model was tested on the NSL-KDD dataset, obtaining a 96.82% accuracy.

One of the initial studies to emphasize explainability in AD diagnosis was [13], where the LIME framework was applied to traditional ML models, providing insights into feature importance and decision-making. Authors in [14] explored the combination of structural MRI and functional MRI data for AD classification using deep learning, demonstrating that integrating multiple imaging modalities can significantly enhance diagnostic performance. They also personalized explanation models that adapt to individual patient data, offering tailored insights for AD diagnosis and treatment planning. Authors in [15] proposed an explainable ML framework for AD classification that integrates SHAP with various ML classifiers. Their approach provided both high classification accuracy and meaningful feature attributions, allowing clinicians to interpret the model's decision-making process. The study leveraged significant features from structural MRI data and showed that XAI tools can improve trust in automated diagnostic systems. The combination of interpretability and performance is critical for deployment in real-world clinical settings.

Authors in [16] explored brain connectivity in AD using an explainable AI pipeline. By analyzing resting-state fMRI data, they built interpretable models that highlighted disrupted connectivity patterns in patients with AD. Their work emphasized the relevance of topological brain network analysis and the potential of explainable models to uncover neurobiological insights that are often overlooked by traditional black-box algorithms. The incorporation of attention mechanisms and saliency maps further enhanced the interpretability of their models. Authors in [17] developed a transfer learning approach using pre-trained CNNs for early diagnosis of AD from structural MRI. Although not fully explainable by default, their approach emphasized feature reusability from large datasets, boosting classification performance even with limited medical data. Finally, they laid the foundation for integrating XAI by highlighting how feature transfer from general-purpose image datasets can be adapted and explained using layer-wise relevance propagation or class activation maps.

Authors in [13] introduced GFE-Mamba, a Generative Feature Extraction (GFE) model based on the Mamba attention architecture, for multi-modal AD progression assessment.

Their model synthesized features from both MRI and clinical data to track the progression from Mild Cognitive Impairment (MCI) to full-blown AD. The study emphasized multi-modality and temporal modeling, offering a novel pathway to interpret disease evolution. While primarily a generative model, their architecture incorporated feature visualizations to provide a degree of transparency, aligning with XAI objectives. Authors in [18] proposed explainable deep CNN models specifically for MRI-based AD diagnosis. They integrated Grad-CAM and occlusion sensitivity methods to visualize the brain regions the model focused on while making predictions. It was demonstrated that despite CNNs being black-box in nature, combining them with visualization-based XAI methods made their decisions more understandable. This was particularly important in identifying structural changes in the hippocampus and cortical regions associated with AD [18]. While many studies focus on accuracy or basic feature importance, few offer spatially grounded, biomarker-level interpretability in brain scans. This study addresses this gap by integrating CNN classification with LIME-based super-pixel mapping, aligning AI outputs with clinical understanding.

Although several existing studies use ensemble models, such as XGBoost or LightGBM with SHAP, to get the feature attribution, they are mostly global or feature attribution methods, and they do not spatially contextualize model predictions within the brain anatomy. Others employ deep learning with built-in XAI techniques, although they do not explicitly validate the explanations against previously known clinical biomarkers, such as hippocampal atrophy or cortical thinning.

Conversely, the proposed approach incorporates a CNN-based classification model with LIME to obtain visual explanations based on super-pixels, and the localization of the regions of influence in MRI scans can be performed in spatial terms. This allows clinicians to put predictions into the context of anatomical structures, which is a very important improvement compared to the feature attribution by itself. Furthermore, the proposed approach tests these explanations against biomarkers that are medically accepted as the presence of AD, and it does so, not only with predictive accuracy, but also with clinically interpretable results.

II. DATASET DETAILS

The dataset used in this study is the publicly available Open Access Series of Imaging Studies (OASIS) Alzheimer's dataset [19]. Table I shows the four stages of the AD dataset distribution.

TABLE I. DATASET SAMPLE DISTRIBUTION

Class	Samples
Non-Demented (ND)	3200
Very Mild Demented (VMD)	2240
Mild Demented (MD)	896
Moderate Demented (MOD)	64
Total	6400

All experiments and analyses were conducted using this dataset, ensuring the transparency and reproducibility of the results.

III. METHODOLOGY

This study proposes an integrated approach that combines a customized CNN with XAI techniques to enable accurate and interpretable classification of AD stages. The model was trained using the OASIS dataset, a publicly available neuroimaging repository consisting of structural MRI scans. These scans represent various AD stages, including the ND, VMD, MD, and MOD categories. The preprocessing pipeline was critical to ensure high-quality input data. Figure 1 illustrates the proposed system's methodology.

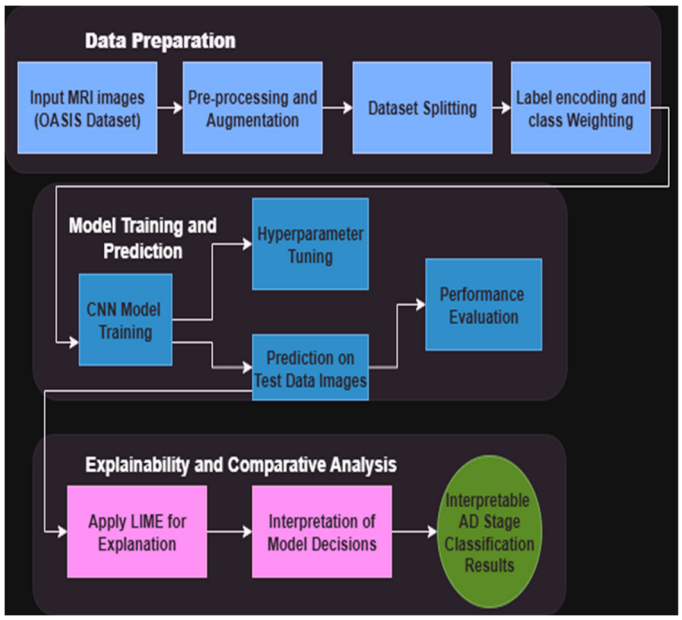

Fig. 1. Pipeline of the proposed explainable deep learning framework: integration of CNN with LIME for AD stage classification.

First, MRI images were resized to 224 × 224 pixels and normalized to (0, 1) for stable training. To improve generalization and address class imbalance, data augmentation (random horizontal flips, ±10° rotation, zoom 0.9–1.1, width/height shifts ±10%) was applied. The dataset was split into 70% training, 15% validation, and 15% testing, with labels encoded into four AD stages. Class weights were incorporated during training to compensate for underrepresented classes such as MOD.

For interpretability, the CNN was integrated with LIME, a model-agnostic XAI technique that generates localized super-pixel explanations. LIME highlights influential brain regions (e.g., hippocampal atrophy, cortical thinning) that contribute to classification, allowing the clinical validation of the model outputs. The CNN was trained and evaluated on the OASIS dataset, after which LIME explanations were generated for each prediction. Finally, a comparative analysis was performed between the chosen XAI technique and existing approaches. This combined methodology leverages the CNN's ability to extract intricate features with LIME's interpretability, improving both accuracy and transparency in AD classification, ultimately supporting reliable and informed medical decision-making.

A. Convolutional Neural Network

The CNN architecture used in this work is lightweight yet effective, as depicted in Figure 2. It began with an input layer of 3 × 224 × 224 size, followed by two convolutional layers with 32 and 64 filters, each using 3 × 3 kernels and ReLU activation. A MaxPooling layer with a 2 × 2 window was added to reduce the spatial dimensions. The feature maps were then flattened and passed to a fully connected (Dense) layer with 128 neurons using ReLU, followed by a Dropout layer with a rate of 0.5 to prevent overfitting. Finally, the output layer deployed Softmax activation to classify the input into one of the four AD stages.

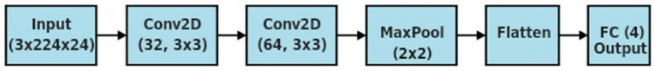

Fig. 2. Architecture of a CNN showing convolution, pooling, and fully connected (dense) layers.

For training, the study employed the Adam optimizer with a learning rate of 0.001, batch size of 32, and trained the model for 15 epochs. Categorical Cross-Entropy was utilized as the loss function due to the multi-class nature of the problem. To prevent overfitting and reduce unnecessary computation, Early Stopping was employed with a patience parameter of 5 epochs. Hyperparameters, such as learning rate, batch size, and number of epochs, were optimized using a grid search strategy. The best performance was obtained with a learning rate of 0.001, batch size of 32, and 15 training epochs.

All experiments were conducted on GPU-enabled environments to ensure efficient training and evaluation. Model performance was assessed using multiple evaluation metrics to provide a comprehensive analysis. These included Accuracy, Precision, Recall, F1 Score, AUC, and ROC curve. A confusion matrix was used to visualize the classification performance across classes. To statistically validate the performance of the proposed customized CNN model, a 5-fold cross-validation was conducted and repeated three times (for a total of 15 runs). The model achieved an average accuracy of 96.98% with a standard deviation of ±0.47%, demonstrating consistent performance across different folds. A one-sample t-test examined whether the observed mean accuracy significantly differs from a baseline of 95%. The test yielded a t-statistic of 7.21 and a p-value of 0.00003 ($p < 0.01$), confirming that the observed improvement is statistically significant and not due to chance.

B. Local Interpretable Model-Agnostic Explanations

The XAI movement attempts to create techniques for explaining the inner workings of complex models. Local LIME is a well-known XAI method that provides localized explanations for individual predictions from any classifier models [20, 21]. In the proposed framework, LIME was applied to MRI images classified by the CNN model, and the process consists of the following five key steps:

- Perturbation of Input Data: Given an input MRI image, LIME generates a set of perturbed samples by modifying regions of the image (e.g., removing super-pixels or adding noise). Each perturbed image is passed through the deep learning model to obtain predictions.
- Prediction Generation: Each perturbed image is passed through the pre-trained CNN, and the resulting predictions are recorded. This yields a set of prediction outputs corresponding to the varied inputs.
- Similarity Weighting: According to LIME, it assigns weights to perturbed samples according to their similarity to the original image. A kernel function is often used to measure similarity. A common approach is to use an exponential function:

$$w(x, x') = exp(-\frac{D(x,x\prime)^2}{\sigma^2}) \quad (1)$$

where $D(x, x')$ represents the distance between the original and perturbed image, and sigma controls the sensitivity.

- Training an Interpretable Model: A basic model that someone can easily understand (such as a linear model) undergoes training using the modified samples. The goal is to estimate local deep learning predictions by minimizing weighted prediction errors:

$$arg\,\frac{min}{y}\,\frac{\sum}{x\prime} w(x, x') \cdot (f(x') - g(x'))^2 + \Omega(g) \quad (2)$$

where $\Omega(g)$ is a complexity term that ensures interpretability.

- Explanation Generation: Finally, LIME identifies the most influential super-pixels, i.e., brain regions that contributed to the CNN's decision. These regions are highlighted as salient visual areas (e.g., hippocampus, cortical thinning zones), enabling clinicians to visually validate the rationale behind the model's classification.

LIME has been chosen because it offers localized, model-agnostic, and human-interpretable super-pixel level explanation, which is more clinically understandable. In contrast to SHAP, where the importance of features is measured across the board, or Grad-CAM, only applicable to CNN and only on the deeper layers. LIME allows segmentations of images in the image space, which seem to fit the anatomy of the brain. This property enables a clinician to physically evaluate the model by determining whether the model's attention is aligned with a medically relevant area, e.g., the hippocampus or the temporal lobe.

IV. RESULTS

A systematic hyperparameter tuning was performed using grid search to evaluate the effectiveness and feasibility of the proposed CNN with the LIME model for AD classification. Optimal results were achieved with the Adam optimizer (learning rate = 0.001, batch size = 32, 15 epochs). This configuration significantly improved performance, especially for the challenging MOD class. The model comprised approximately 3.5 million trainable parameters, making it lighter than ResNet or VGG, and was trained on an NVIDIA RTX 3060 GPU (12GB) to converge within 32 min. The average inference time, including LIME explanation, was 0.18 s per image. This demonstrates a practical balance between

accuracy, interpretability, and computational efficiency for real-time clinical use. The comprehensive evaluation confirmed a balanced performance, with accuracy of 97%, precision of 97.2%, recall of 96.8%, and F1 Score of 96.9%, as displayed in Table II.

TABLE II. PERFORMANCE EVALUATION METRICS

Metric	Value
Accuracy	97.0%
Precision	97.2%
Recall (Sensitivity)	96.8%
F1 Score	96.9%
AUC	0.97

Figure 3 presents the multi-class ROC curve for classifying four AD stages using the proposed CNN model on the OASIS dataset. The curves represent ND (AUC: 0.97), VMD (AUC: 0.96), MD (AUC: 0.96), and MOD (AUC: 0.95), with an overall AUC of 0.97. The slightly lower AUC for MOD reflects the challenges in distinguishing advanced AD stages due to overlapping structural MRI features (e.g., hippocampal atrophy, cortical thinning), despite the class weights applied to address the dataset's imbalance (MOD: 64 samples versus ND: 3200). This indicates the intrinsic difficulty in differentiating the Mild and Moderate stages using structural MRI alone.

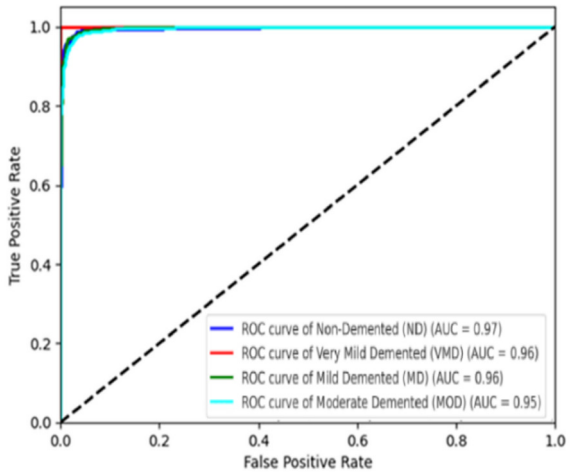

Fig. 3. ROC curves for four classes of AD.

Figure 4 depicts the training and validation accuracy over 15 epochs. Both increase rapidly, with the training accuracy nearing 99% and the validation accuracy stabilizing around 97%, indicating effective learning and minimal overfitting.

Figure 5 illustrates the loss trends. The training and validation loss drop quickly, with the training loss nearing zero and the validation loss stabilizing at a low value, suggesting strong optimization and minimal overfitting. Figure 6 presents the confusion matrix for AD classification. It shows that the CNN model effectively distinguishes between the four AD stages, achieving 97% overall accuracy. Minor misclassifications occur between adjacent stages, indicating areas for further improvement in clinical reliability.

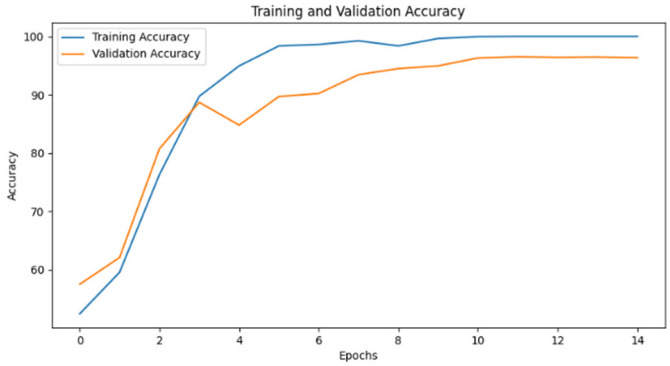

Fig. 4. Training and validation accuracy.

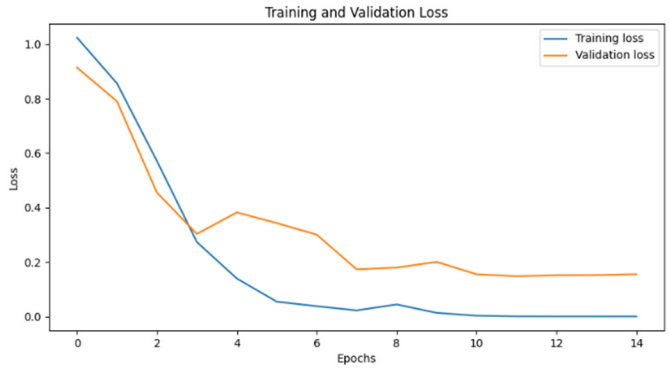

Fig. 5. Training and validation loss.

Figure 7 exhibits how LIME generates localized super-pixel visualizations that highlight the most influential brain regions contributing to CNN predictions. Bright regions correspond to areas with high impact on classification, such as hippocampal atrophy in MD cases and cortical thinning in MOD cases. These biomarker-aligned explanations link the model outputs with clinically validated markers, enhancing interpretability and trust in AI-assisted AD diagnosis.

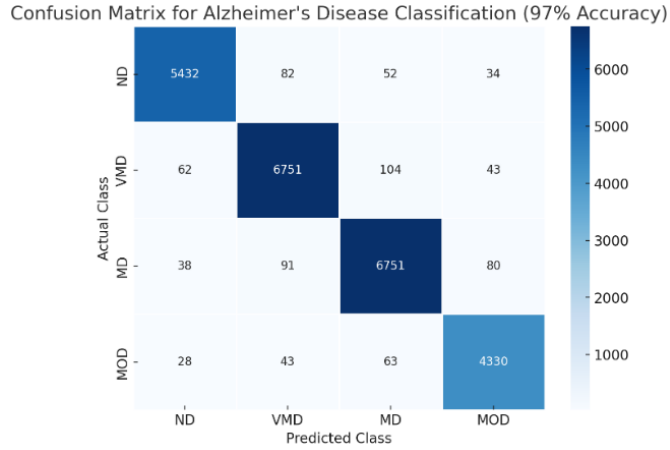

Fig. 6. Confusion matrix of the distribution of 4 classes of Dementia.

Table III presents a comparative analysis between the proposed method and existing approaches using the ADNI and

OASIS datasets. The proposed method outperforms others by achieving the highest accuracy of 97.02%.

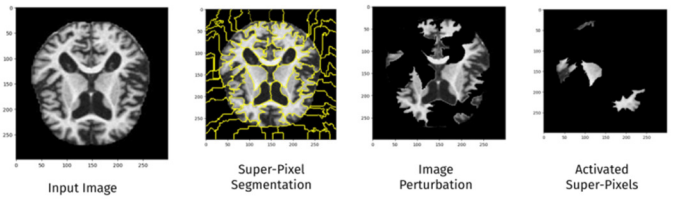

Fig. 7. LIME-based super-pixel explanations for CNN predictions of AD stages.

TABLE III. COMPARISON OF THE PROPOSED METHOD WITH EXISTING METHODS

Model/references	Dataset	Accuracy (%)	Loss (%)
VGG architecture [17]	ADNI	83.72	33.2
3D-CNN-SVM [13]	ADNI	93.71	31.09
GFE [18]	ADNI	94.31	28.45
Deep Transfer Learning with XAI [22]	OASIS	96	29.01
Deep CNN with XAI [19]	OASIS	92	26.23
Proposed method	OASIS	97.02	18.76

V. CONCLUSION

In this research, a new explainable deep learning model, which combines a customized Convolutional Neural Network (CNN) with Local Interpretable Model-agnostic Explanations (LIME), is proposed to achieve high diagnostic performance and interpretability of Alzheimer's Disease (AD) classification based on MRI data. The proposed framework has a high classification accuracy of 97%. Additionally, the LIME visualizations offer spatially-constrained interpretations that can be correlated to clinically-validated biomarkers, including hippocampal atrophy and cortical thinning across the most significant division between algorithmic prediction and clinical knowledge.

In spite of these findings, the current study has significant limitations. The present application is only restricted to structural MRI data without considering multi-modal imaging (e.g., PET, fMRI) and patient-specific demographic factors (e.g., age and gender), which may affect disease manifestation and model generalization. Additionally, while LIME provides local interpretability, its comparative performance against other explainability techniques, such as SHAP, Grad-CAM, or Integrated Gradients, has not been exhaustively evaluated. SHAP mainly offers global or feature-level importance without spatial alignment to brain anatomy, and Grad-CAM highlights broader regions restricted to CNNs that may not consistently overlap with clinically recognized biomarkers. In contrast, the proposed method produces super-pixel–based visual explanations that can be directly mapped to biomarkers, such as hippocampal atrophy and cortical thinning, offering a finer degree of anatomical relevance and clinical trust. The computational overhead associated with explanation generation, however, may impact real-time deployment in clinical workflows.

This work focused on the ethical aspects of explainability, since the latter is not just a technical necessity, but also an ethical one in medical AI systems. The proposed method will enhance the transparency of model decisions, which in turn will foster accountability, decrease diagnostic opacity, as well as establish trust among clinicians as the key factors to ethical adoption of AI in healthcare.

The present study may also help precision medicine. The proposed framework enables more individualized, transparent, and evidence-based diagnostics by matching predictions of deep learning models with more clinically interpretable explanations. The future research directions will entail the combination of heterogeneous clinical data, demographic bias mitigation practices, multi-center validation, and real-time performance optimization, thus enhancing the clinical preparedness and ethical resilience of explainable AI in the diagnosis of neurodegenerative diseases.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the institutions and research facilities that supported this study. Additionally, they acknowledge the contributions of the Explainable Artificial Intelligence (XAI) community, whose methodologies, such as LIME, were instrumental in enhancing model interpretability.

DATASET AVAILABILITY

The dataset used in this study is publicly available at: https://sites.wustl.edu/oasisbrains/

REFERENCES

[1] E. M. McDade, "Alzheimer Disease," *Continuum*, vol. 28, no. 3, pp. 648–675, Jun. 2022, https://doi.org/10.1212/CON.0000000000001131.

[2] S. Gauthier, C. Webster, S. Sarvaes, J. Morais, and P. Rosa-Neto, *World Alzheimer Report 2022: Life After Diagnosis: Navigating Treatment, Care and Support*. Lincolnshire, IL, USA: Alzheimer's Disease International., 2022.

[3] N. Shaffi, F. Hajamohideen, M. Mahmud, A. Abdesselam, K. Subramanian, and A. A. Sariri, "Triplet-Loss Based Siamese Convolutional Neural Network for 4-Way Classification of Alzheimer's Disease," in *Brain Informatics*, vol. 13406, M. Mahmud, J. He, S. Vassanelli, A. Van Zundert, and N. Zhong, Eds. Cham: Springer International Publishing, 2022, pp. 277–287.

[4] S. A. Tatulian, "Challenges and Hopes for Alzheimer's disease," *Drug Discovery Today*, vol. 27, no. 4, pp. 1027–1043, Apr. 2022, https://doi.org/10.1016/j.drudis.2022.01.016.

[5] A. Rai, "Explainable AI: from Black Box to Glass Box," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, Jan. 2020, https://doi.org/10.1007/s11747-019-00710-5.

[6] G. Yang, Q. Ye, and J. Xia, "Unbox the Black-box for the Medical Explainable AI via Multi-modal and Multi-centre Data Fusion: A Mini-review, Two Showcases and Beyond," *Information Fusion*, vol. 77, pp. 29–52, Jan. 2022, https://doi.org/10.1016/j.inffus.2021.07.016.

[7] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection," *Brain Informatics*, vol. 11, no. 1, Dec. 2024, Art. no. 10, https://doi.org/10.1186/s40708-024-00222-1.

[8] B. Dubois, G. Picard, and M. Sarazin, "Early Detection of Alzheimer's Disease: New Diagnostic Criteria," *Dialogues in Clinical Neuroscience*, vol. 11, no. 2, pp. 135–139, Jun. 2009, https://doi.org/10.31887/DCNS.2009.11.2/bdubois.

[9] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "Stop Ordering Machine Learning Algorithms by Their Explainability! An Empirical Investigation of the Tradeoff Between Performance and Explainability," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, vol. 12896, D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, and M. Mäntymäki, Eds. Cham: Springer International Publishing, 2021, pp. 245–258.

[10] B. K. Raghupathy, M. R. Reddy, Prasad Theeda, E. Balasubramanian, R. K. Namachivayam, and M. Ganesan, "Harnessing Explainable Artificial Intelligence (XAI) based SHAPLEY Values and Ensemble Techniques for Accurate Alzheimer's Disease Diagnosis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20743–20747, Apr. 2025, https://doi.org/10.48084/etasr.9619.

[11] W. Feng *et al.*, "Automated MRI-Based Deep Learning Model for Detection of Alzheimer's Disease Process," *International Journal of Neural Systems*, vol. 30, no. 06, Jun. 2020, Art. no. 2050032, https://doi.org/10.1142/S012906572050032X.

[12] S. C. Emerald and T. Vengattaraman, "Explainable Artificial Intelligence with Single Layer Feedforward Neural Network and Improved Crowned Porcupine Optimization Algorithm for Classification Problems," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21593–21598, Apr. 2025, https://doi.org/10.48084/etasr.10070.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.

[14] M. Liu, D. Cheng, K. Wang, and Y. Wang, "Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis," *Neuroinformatics*, vol. 16, no. 3–4, pp. 295–308, Oct. 2018, https://doi.org/10.1007/s12021-018-9370-4.

[15] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An Explainable Machine Learning Approach for Alzheimer's Disease Classification," *Scientific Reports*, vol. 14, no. 1, Feb. 2024, Art. no. 2637, https://doi.org/10.1038/s41598-024-51985-w.

[16] N. Amoroso, S. Quarto, M. La Rocca, S. Tangaro, A. Monaco, and R. Bellotti, "An eXplainability Artificial Intelligence Approach to Brain Connectivity in Alzheimer's Disease," *Frontiers in Aging Neuroscience*, vol. 15, Aug. 2023, Art. no. 1238065, https://doi.org/10.3389/fnagi.2023.1238065.

[17] A. Mehmood *et al.*, "A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images," *Neuroscience*, vol. 460, pp. 43–52, Apr. 2021, https://doi.org/10.1016/j.neuroscience.2021.01.002.

[18] Z. Fang *et al.*, "GFE-Mamba: Mamba-based AD Multi-modal Progression Assessment via Generative Feature Extraction from MCI." arXiv, Jan. 29, 2025, https://doi.org/10.48550/arXiv.2407.15719.

[19] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, "Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom, Jul. 2020, pp. 1–8, https://doi.org/10.1109/IJCNN48605.2020.9206837.

[20] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007, https://doi.org/10.1162/jocn.2007.19.9.1498.

[21] "Alzheimer's Disease Association," *Understanding Alzheimer's Disease and Dementia*, 2024. https://www.alz.org/.

[22] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, "An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning," *Diagnostics*, vol. 14, no. 3, Feb. 2024, Art. no. 345, https://doi.org/10.3390/diagnostics14030345.