Smart Media Journal / Vol.14, No.6 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2025.14.6.60

Toward Clinically Trustworthy Alzheimer's Diagnosis: Combining EfficientNetV2B0 and XAI Techniques for MRI Analysis

Faizaan Fazal Khan, Goo-Rak Kwon

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder where early diagnosis is essential for effective intervention. Recent advancements in neuroimaging and deep learning have led to promising automated diagnostic systems, yet their "black box" nature limits clinical adoption. In this study, we present a novel framework that combines a fine-tuned EfficientNetV2B0-based model with state-of-the-art explainable AI (XAI) techniques, including Grad-CAM and LIME, to classify brain MRI images from the OASIS-1 dataset into four distinct stages of Alzheimer's progression. The model achieves over 99% validation accuracy while generating visual explanations that reliably highlight disease-relevant brain regions, thereby enhancing interpretability and trust for clinical decision-making. These results not only validate the effectiveness of our approach but also underscore the importance of integrating transparency into AI-driven diagnostics. Future research will explore multimodal imaging data and additional XAI methods to further improve diagnostic reliability and support personalized treatment strategies.

Keywords: Alzheimer's disease | MRI | deep learning | EfficientNetV2B0 | Explainable AI | Grad-CAM | interpretability | LIME | medical imaging | OASIS dataset

I. INTRODUCTION

Alzheimer's disease (AD) remains one of the most pressing public health challenges[1], largely due to its insidious onset and progressive nature. Early diagnosis is critical for effective intervention, yet the subtle structural changes in the brain that herald AD often evade conventional diagnostic methods. With the advent of advanced neuroimaging techniques, particularly magnetic resonance imaging (MRI), clinicians now can capture high-resolution images of the brain. However, interpreting these images for early signs of Alzheimer's requires both expert knowledge and robust computational tools. Recent advances in deep learning have demonstrated impressive accuracy in classifying brain MRI scans into various stages of AD progression, yet many of these convolutional neural network (CNN) models remain opaque[2], leaving clinicians hesitant to fully trust their automated predictions.

In response to this challenge, explainable artificial intelligence (XAI) methods have been developed to provide visual insights into how deep learning models make their decisions. Techniques such as Grad-CAM, LIME, and SHAP have shown promise in highlighting the critical regions of the brain that influence the model's predictions, thereby offering a window into the "black box" of CNN-based diagnostic systems[3]. For example, Alami et al.[4] demonstrated how visual explanation methods could enhance the interpretability of CNN models in AD detection, while Vetrithangam et al.[5]

* This study was supported by research funds from Chosun University, 2024.

Manuscript : 2025.03.23
Confirmation of Publication: 2025.04.08
Corresponding Author : Goo-Rak Kwon e-mail : grkwon@chosun.ac.kr

integrated clustering approaches with CNNs to further improve model transparency. Our previous study [6] addressed the challenges of skewed data in MRI-based Alzheimer's diagnosis by comparing various CNN architectures and applying resampling techniques to improve classification fairness. These studies underscore the necessity of bridging high-performance deep learning with interpretable outputs that can be readily understood and trusted by clinicians.

Building on these advances, our work leverages the OASIS MRI dataset, which comprises over 80,000 brain images from 461 patients, to train an EfficientNetV2B0-based model for classifying AD progression into four distinct stages: non-demented, very mild demented, mild demented, and moderate demented. Our study not only focuses on achieving state-of-the-art classification accuracy but also emphasizes the integration of XAI techniques—specifically Grad-CAM and LIME—to generate visual explanations that highlight relevant brain regions. These visual outputs aim to enhance clinical decision-making by providing transparent, interpretable, and trustworthy results. Ultimately, our approach looks to empower clinicians with tools that combine the predictive prowess of deep learning with intuitive visual explanations, paving the way for more reliable early diagnosis of Alzheimer's disease.

II. METHODOLOGY

A. Proposed Model

In this study, we adopt a widely recognized deep learning architecture EfficientNetV2B0[7] as the backbone for our Alzheimer's detection framework. Figure 1 depicts our overall methodology. The EfficientNetV2B0 model, pre-trained on ImageNet[8], is chosen due to its balanced trade-off between accuracy and computational efficiency. Its established performance makes it an ideal base model for integrating explainable AI techniques such as Grad-CAM and LIME, which are critical for providing transparency in clinical applications.

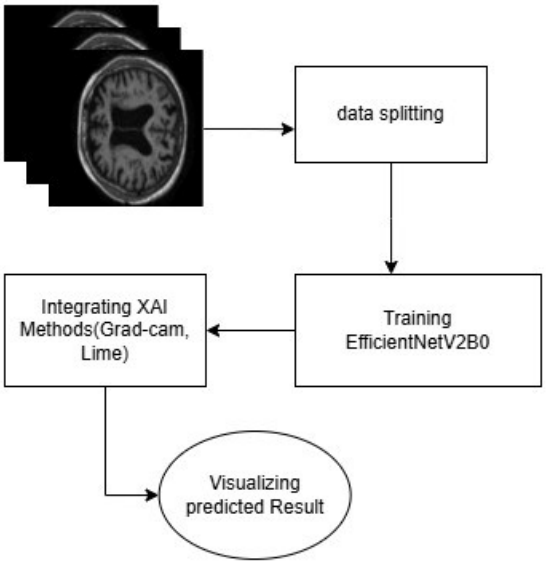

Figure 1 Methodology of our study

The pre-trained EfficientNetV2B0 is modified by removing the top classification layers and appending a custom classification head. This head comprises a Global Average Pooling layer, followed by two fully connected (dense) layers with dropout for regularization. The final dense layer outputs a softmax probability distribution over the four classes (non-demented, very mild demented, mild demented, and moderate demented).

The hyperparameters used in our model are summarized in Table 1. These hyperparameters were selected based on preliminary experiments to ensure stable performance while maintaining a robust baseline model for subsequent XAI analysis. It is important to note that the primary role of the model training in this study is to provide a reliable foundation for interpreting predictions via XAI methods rather than achieving state-of-the-art classification accuracy.

Table 1 Hyperparameters of the Proposed Model

Parameter	Value
Input Shape	128 × 128 × 3
Base Model	EfficientNetV2B0 (pre-trained on ImageNet)
Dense Layer 1 Units	256
Dropout Rate 1	30%
Dense Layer 2 Units	128
Dropout Rate 2	25%
Final Layer Activation	Softmax
Number of Classes	4

The role of the EfficientNetV2B0-based model in our study is to serve as a reliable and well-calibrated baseline. After fine-tuning the model on the OASIS MRI dataset (which consists of 80,000 images derived from 461 patients), we integrate state-of-the-art XAI techniques. These methods—such as Grad-CAM and LIME—are used to generate visual explanations that highlight the brain regions influencing the model's predictions. Such interpretability is essential for clinical adoption, as it helps bridge the gap between automated analysis and medical expertise.

B. OASIS-1 Dataset and Demographics

This study utilized the OASIS-1 (Open Access Series of Imaging Studies) dataset[9], a publicly available resource designed to support research on Alzheimer's disease progression and related cognitive impairments. The dataset comprises neuroimaging data, including MRI scans, along with detailed clinical and demographic information, making it an invaluable asset for understanding the disease through advanced imaging techniques. In our study, the OASIS-1 dataset includes data from 416 subjects. Brain MRI scans were processed by slicing each 3D volume along the z-axis into 256 slices, from which 100 to 160 slices per patient were randomly selected. This approach resulted in an original dataset of approximately 86,437 2D images, as illustrated in Figure 2 below with there labels. However, the dataset is highly skewed, with a pronounced predominance of non-demented images relative to the other classes. To address this imbalance and reduce the risk of overfitting, we applied a resampling strategy with 5000 sample size. The resulting balanced class distribution is depicted in the pie chart in Figure 3, highlighting the efforts to mitigate bias and improve the reliability of subsequent deep learning analyses.

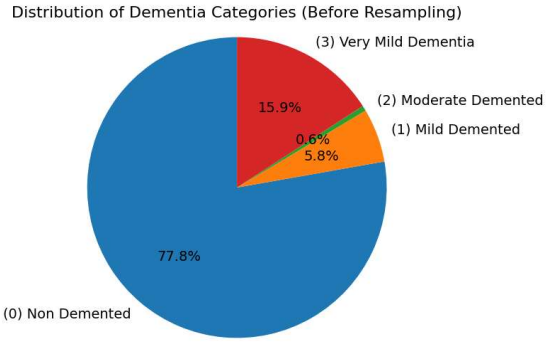

Figure 2 Class distribution before resampling

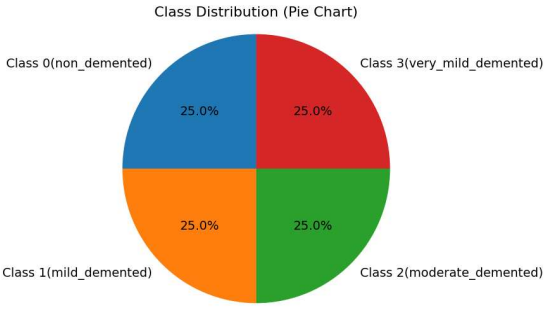

Figure 3 Class distribution after resampling

III. EXPERIMENT RESULT

A. Training and Results

This section outlines the training process of our EfficientNetV2B0 based model, detailing its convergence behavior and final performance metrics, which

serve as the foundation for our subsequent interpretability analyses. To ensure robust generalization and mitigate overfitting, we implemented an early stopping mechanism that monitored the validation loss and automatically restored the best model weights.

Training was originally scheduled for up to 50 epochs. However, the model's performance on the validation set peaked around epoch 8, where it achieved its lowest validation loss. After this point, no further improvements were observed, and the early stopping mechanism was triggered at epoch 11. The model then automatically reverted to the weights from epoch 8, which marked the optimal balance between training accuracy and generalization. This strategy helped to prevent overfitting and ensured efficient use of training resources. It's important to highlight that loss and accuracy are measured differently and do not add up to 1. Accuracy measures the percentage of correct predictions, whereas loss quantifies the prediction error — lower loss does not always directly correlate with higher accuracy and vice versa.

Table 2 Performance Metric

Training Accuracy	0.9938
Validation Accuracy	0.9958
Training Loss	0.0192
Validation Loss	0.0178

The metrics in Table 2 above show that the model reached over 99% accuracy on both the training and validation datasets, with very low loss values, suggesting effective learning and strong generalization to unseen data.

Figures 4 and 5 illustrate the evolution of the training process. In Figure 3, the training accuracy increases steeply in the initial epochs and approaches near-perfect performance, while the validation accuracy closely follows, suggesting effective learning without significant overfitting. Figure 2 shows a corresponding sharp decrease in training loss, with the validation loss decreasing in parallel during the early epochs. The slight increase in validation loss beyond epoch 8 further justifies the early stopping strategy.

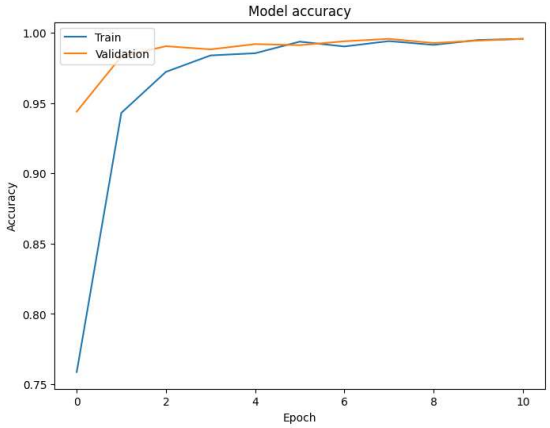

Figure 4 Training and Validation Accuracy

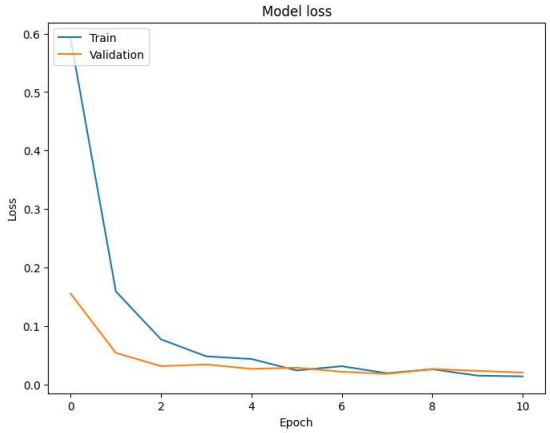

Figure 5 Training and Validation Loss

Overall, the training process demonstrates that our chosen architecture, along with the carefully tuned hyperparameters and early stopping, provides a robust and reliable model for Alzheimer's disease classification. This strong baseline is essential for the subsequent integration of explainable AI techniques, such as Grad-CAM and LIME, which will offer transparent visualizations to support clinical decision-making.

B. Comparative Visual Results:

Figures 6 and 7 illustrate the difference in visual explanations generated by Grad-

CAM and LIME for MRI slices of patients at varying stages of Alzheimer's disease. Both techniques highlight high-confidence regions in the image; however, their underlying mechanisms and resulting overlays differ significantly.

1. High-Confidence Activation Thresholds

Grad-CAM: A gradient-based method that computes importance weights for each neuron in the last convolutional layer. The final heatmap is thresholded at the top 1% (or a similarly high percentile) of activation values, ensuring that only the most influential pixels are shown.

LIME: A perturbation-based, local explanation approach that identifies superpixels whose removal or alteration most affects the model's prediction. In this study, the top 4% (96th percentile) of superpixels by importance score were selected, highlighting the most critical regions for the model's decision.

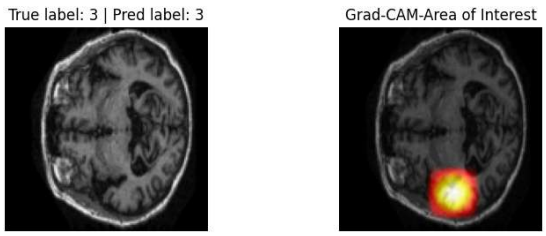

Figure 6 Grad-cam visualization

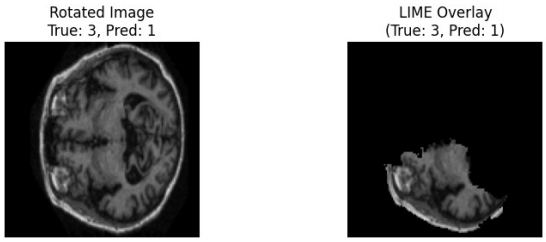

Figure 7 Lime Visualization

2. Overlay Observations

Grad-CAM Overlay (Figure 6): Focuses on a compact, gradient-driven "area of interest," typically around anatomically salient regions associated with the model's classification. This approach provides a global view of where the network's highest gradient activations lie, often localizing a tight cluster that corresponds to key features for the predicted label.

LIME Overlay (Figure 7): Produces a superpixel-based mask that zeroes in on multiple sub-regions the model relies upon. Because LIME creates explanations by locally perturbing the input, its overlay can appear more fragmented, emphasizing specific patches that have the greatest local influence on the model's output.

3. Correct vs. Misclassified Instances

In one example, LIME is shown on an instance misclassified by the network ("True: 3, Pred: 1"). Here, LIME's overlay reveals which superpixels led the model toward an incorrect conclusion where we can see the eye.

Grad-CAM, in contrast, is demonstrated on a correctly classified instance ("True label: 3, Pred label: 3"). Its heatmap pinpoints a distinct bright region that strongly influenced the correct prediction.

4. Clinical Implications

Interpretability: Grad-CAM's single heatmap can be quickly interpreted by clinicians as a focal "hotspot," while LIME's superpixel-based explanation offers finer-grained, local insights. Grad-CAM may be preferred for a fast, high-level check of whether the network focuses on disease-relevant structures, whereas LIME is beneficial for dissecting the model's behavior on difficult or misclassified cases.

Overall, these methods increase the interpretability and trust, enabling clinicians to gain a clearer understanding of why the network arrives at a given decision—particularly crucial in the high-

stakes context of Alzheimer's disease diagnosis.

C. Experiment Environment

Experiments were conducted on a Windows 10 machine running Python 3.9.13 using Visual Studio with the Python Notebook extension (IPvKernel). The system featured two NVIDIA Quadro P4000 GPUs (4GB each) and 64 GB of RAM (four 16GB Samsung modules at 2666 MHz). We primarily used TensorFlow/Keras for deep learning, along with libraries such as scikit-learn, OpenCV, and nibabel to ensure reproducibility.

IV. Discussion and Conclusion

Our study demonstrates that integrating explainable AI methods—Grad-CAM and LIME—with a fine-tuned EfficientNetV2B0-based model on the OASIS-1 dataset produces both high diagnostic accuracy and clinically meaningful visual explanations. The model quickly achieved near-optimal performance, and the generated heatmaps reliably highlight brain regions known to be affected by Alzheimer's, thereby enhancing the interpretability of the predictions. These findings are consistent with recent literature emphasizing the need for transparency in AI-driven medical diagnostics[10], [11], [12]. Overall, our framework provides a robust foundation for early Alzheimer's detection and that can offer valuable insights for clinical decision-making. Future work will explore multimodal imaging and additional XAI techniques to further improve diagnostic reliability.

ACKNOWLEDGMENT

This study was supported by research funds from Chosun University, 2024. Data collection and sharing for this project was funded by the OASIS (Open Access Series of Imaging Studies)[9]. Available at: https://sites.wustl.edu/oasisbrains/.

REFERENCES

[1] Uzer Altaf Shaikh and Sanket P. Shinde, "Alzheimer Disease," *World J. Biol. Pharm. Health Sci.*, vol. 18, no. 2, pp. 049–057, May 2024, doi: 10.30574/wjbphs.2024.18.2.0239.

[2] S. Ahmed *et al.*, "Ensembles of Patch-Based Classifiers for Diagnosis of Alzheimer Diseases," *IEEE Access*, vol. 7, pp. 73373–73383, 2019, doi: 10.1109/ACCESS.2019.2920011.

[3] U. Khatri, J.-H. Kim, and G.-R. Kwon, "Alzheimer's Disease and Mild Cognitive Impairment Detection Using sMRI With Efficient Receptive Field and Enhanced Multi-Axis Attention Fusion," *IEEE Access*, vol. 12, pp. 100848–100861, 2024, doi: 10.1109/ACCESS.2024.3430325.

[4] A. Alami, J. Boumhidi, and L. Chakir, "Explainability in CNN based Deep Learning models for medical image classification," in *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, May 2024, pp. 1–6. doi: 10.1109/ISCV60512.2024.10620149.

[5] D. Vetrithangam *et al.*, "Towards Explainable Detection of Alzheimer's Disease: A Fusion of Deep Convolutional Neural Network and Enhanced Weighted Fuzzy C-Mean," *Curr. Med. Imaging*, vol. 20, p. e15734056317205, 2024, doi: 10.2174/0115734056317205241014060633.

[6] Korean Institute of Smart Media, F. F. K. Faizaan Fazal Khan, and G.-R. Kwon, "Comparison and analysis of CNN models to Address Skewed Data Issues in Alzheimer's Diagnosis,"

Korean Inst. Smart Media, vol. 13, no. 10, pp. 28–34, Oct. 2024, doi: 10.30693/SMJ.2024.13.10.28.

[7] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Jun. 23, 2021, *arXiv*: arXiv:2104.00298. doi: 10.48550/arXiv.2104.00298.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[9] 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," *J. Cogn. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007, doi: 10.1162/jocn.2007.19.9.1498.

[10] N. N. Khanna *et al.*, "Polygenic Risk Score for Cardiovascular Diseases in Artificial Intelligence Paradigm: A Review," *J. Korean Med. Sci.*, vol. 38, no. 46, p. e395, Nov. 2023, doi: 10.3346/jkms.2023.38.e395.

[11] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An explainable machine learning approach for Alzheimer's disease classification," *Sci. Rep.*, vol. 14, no. 1, p. 2637, Feb. 2024, doi: 10.1038/s41598-024-51985-w.

[12] N. N. Khanna *et al.*, "Polygenic Risk Score for Cardiovascular Diseases in Artificial Intelligence Paradigm: A Review," *J. Korean Med. Sci.*, vol. 38, no. 46, Nov. 2023, doi: 10.3346/jkms.2023.38.e395.

Authors

Faizaan Fazal Khan

He has completed his bachelor's degree at the department of CS & IT from Air University, Islamabad, Pakistan, in 2023. Currently, he is pursuing his Master studies in the department of Information and Communication Engineering at Chosun University, Gwangju, Republic of Korea. His research interests involve digital image processing, Machine learning utilizing Medical Imaging.

Goo-Rak Kwon

He got a Ph.D. from the Department of Mechatronic Engineering, Korea University, in 2007. He has been a professor at Chosun University, since 2017. His research focus includes medical image analysis, A/V signal processing, video communication, and applications. He is a senior member of the IEEE.