Journal of Information Technology and Cyber Security 4(1) January 2026: 1-15 EISSN 2987-386X | PISSN 2987-3878
DOI:10.30996/jitcs.133060

Explainable Artificial Intelligence Analysis of Transfer Learning Models for Alzheimer's Disease MRI Classification

Dea Amanda Salsabila [1], Ghaluh Indah Permata Sari [2], Fajar Astuti Hermawati [3,*]

[1,2,3] Department of Informatics Engineering, Universitas 17 Agustus 1945 Surabaya, Indonesia
[2] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taiwan
* Corresponding author: fajarastuti@untag-aby.ac.id

Received: 14 January 2026
Accepted: 20 January 2026
Revised: 20 January 2026
Available online: 27 January 2026

To cite this article: Salsabila, D. A., Sari, G. I. P., & Hermawati, F. A. (2026). Explainable Artificial Intelligence Analysis of Transfer Learning Models for Alzheimer's Disease MRI Classification. *Journal of Information Technology and Cyber Security, 4*(1), 1-15. https://doi.org/10.30996/jitcs.133060

Abstract

Alzheimer's disease is a progressive neurodegenerative disorder that leads to cognitive decline and requires early and accurate diagnosis to slow disease progression. Magnetic resonance imaging (MRI) is widely used to detect structural brain changes associated with Alzheimer's disease; however, manual interpretation of MRI scans is time-consuming and subject to observer variability. Deep learning approaches have shown strong potential in automated MRI analysis, but their black-box nature limits clinical trust and interpretability. This study proposes a transfer learning–based deep learning framework for Alzheimer's disease classification, complemented by explainable artificial intelligence (XAI) techniques to analyze model predictions. A pretrained VGG16 model is employed to classify MRI images into four cognitive impairment categories: no impairment, very mild impairment, mild impairment, and moderate impairment. To enhance transparency, Grad-CAM, Saliency Maps, and Guided Grad-CAM are applied to visualize brain regions that contribute most to model predictions. Experimental results demonstrate that the proposed approach achieves 95.41% accuracy, indicating that a well-balanced network architecture combined with integrated explainability techniques leads to effective, interpretable classification. The visual explanations highlight clinically meaningful brain regions that align with known Alzheimer's disease–related structural changes. These findings suggest that combining deep transfer learning with explainable artificial intelligence can provide accurate and interpretable decision support for Alzheimer's disease diagnosis. This study is limited by the use of a single publicly available dataset and two-dimensional MRI slices, which may affect generalizability across clinical environments.

Keywords: Alzheimer's disease, clinical decision support, deep learning, explainable artificial intelligence, magnetic resonance imaging, transfer learning.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, characterized by gradual cognitive decline, memory impairment, and functional deterioration. The increasing aging population has led to a rapid rise in AD prevalence, creating substantial clinical, social, and economic challenges for healthcare systems. Early and accurate diagnosis is crucial to enable timely intervention and slow disease progression; however, detecting AD at early stages remains difficult due to subtle brain changes and overlapping symptoms with normal aging (Jack, et al., 2018; Livingston, et al., 2020).

Neuroimaging plays a central role in Alzheimer's disease diagnosis and progression assessment. Magnetic Resonance Imaging (MRI) is widely used to identify structural brain abnormalities associated with AD, such as hippocampal atrophy and cortical thinning. At the same time, positron emission tomography (PET) provides complementary metabolic information. Despite their diagnostic value, interpretation of neuroimaging data largely relies on expert visual assessment, which is time-consuming and subject to inter-observer variability. These limitations motivate the development of automated and objective image analysis

approaches to support clinical decision-making (Jack, et al., 2018; Rathore, Habes, Iftikhar, Shacklett, & Davatzikos, 2017).

Recent advances in artificial intelligence, particularly deep learning, have significantly improved automated diagnosis of Alzheimer's disease from neuroimaging data (Ali, et al., 2024; Basaia, et al., 2019; Bron, et al., 2021; El-Assy, Amer, Ibrahim, & Mohamed, 2024; Islam & Zhang, 2018; Komal, Dhavakumar, Rahul, Jaswanth, & Preeth, 2025; Sampath & Baskar, 2024; Sheikh, Marouf, Rokne, & Alhajj, 2025; Sorour, et al., 2024; Wen, et al., 2020). Convolutional Neural Networks (CNNs) are capable of learning hierarchical feature representations directly from MRI and PET images, enabling effective stage classification and early detection. Transfer learning using pretrained architectures such as VGG16 and VGG19 has become a widely adopted strategy to address limited and imbalanced medical datasets, demonstrating strong classification performance across multiple studies (Aderghal, Benois-Pineau, Afdel, & Gwenaëlle, 2017; El-Assy, Amer, Ibrahim, & Mohamed, 2024; Islam & Zhang, 2018; Wen, et al., 2020).

Several recent studies have explored multimodal deep learning approaches that combine MRI and PET imaging to improve diagnostic accuracy by leveraging both structural and functional brain information. Although these multimodal frameworks often achieve high performance, they introduce increased architectural complexity, higher computational costs, and reduced interpretability, which may limit their practical applicability in real clinical settings (Odusami, Maskeliūnas, Damaševičius, & Misra, 2023; Odusami, Damaševičius, Milieškaitė-Belousovienė, & Maskeliūnas, 2024). In parallel, other studies have investigated explainable convolutional neural networks and deep transfer learning paradigms for Alzheimer's diagnosis, emphasizing the importance of transparency and trust in medical artificial intelligence systems (De Santi, Pasini, Santarelli, Genovesi, & Positano, 2023; Mahmud, et al., 2024).

Despite these advances, the widespread adoption of deep learning models in clinical practice remains constrained by their black-box nature. Clinicians require not only accurate predictions but also clear explanations that justify model decisions in a manner consistent with medical knowledge. Explainable Artificial Intelligence (XAI) has emerged as a promising solution to this challenge by enabling human-understandable explanations of model behaviour. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM), Saliency Maps, and Guided Grad-CAM have been increasingly applied to visualize regions of interest that influence deep learning predictions in neuroimaging tasks (AbdelAziz, Said, AbdelHafeez, & Ali, 2024; Samek, Wiegand, & Müller, 2017; Selvaraju, et al., 2020; Sheikh, Marouf, Rokne, & Alhajj, 2025; Shuvo, Refat, Preotee, & Muhammad, 2025; Tjoa & Guan, 2021).

Although prior studies have demonstrated the potential of XAI in Alzheimer's disease diagnosis, several research gaps remain. Many existing works apply explainability techniques only as supplementary visualization tools without systematically evaluating their consistency across different disease stages. Furthermore, comparative analysis of interpretability across multiple deep learning architectures remains limited, and the balance between classification performance and explainability is often insufficiently addressed (Chattopadhyay, et al., 2024; Khosroshahi, et al., 2025).

To address these gaps, this study proposes a transfer learning–based deep learning framework complemented by explainable artificial intelligence techniques. The proposed approach employs a pretrained VGG16 model as the primary classifier and integrates multiple XAI techniques, including Grad-CAM, Saliency Maps, and Guided Grad-CAM, to provide comprehensive visual explanations of model predictions. Unlike prior studies that emphasize architectural complexity or multimodal fusion, this research focuses on achieving an effective balance between classification accuracy, computational efficiency, and interpretability using a well-established transfer learning model.

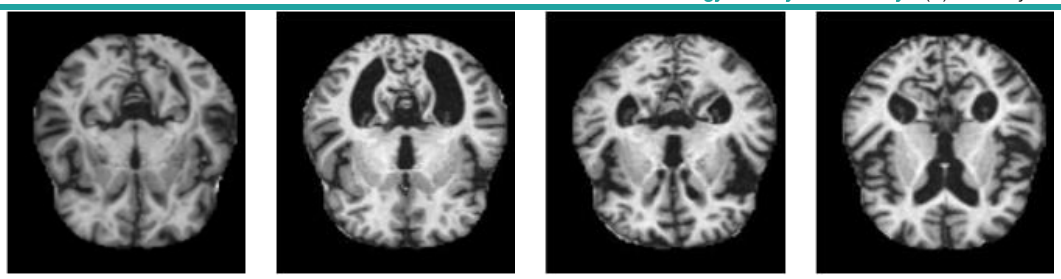
The primary objective of this study is to evaluate the interpretability of transfer learning–based deep learning models for Alzheimer's disease severity classification using MRI images. Specifically, this research aims to:

- Assess the classification performance of pretrained convolutional neural networks for distinguishing Alzheimer's disease severity levels from MRI data;
- Systematically analyse the consistency and reliability of visual explanations generated by multiple explainable artificial intelligence techniques across different disease stages;
- Investigate the relationship between model predictive performance and explainability, examining whether higher accuracy corresponds to more clinically meaningful visual explanations.

By focusing on explainability evaluation rather than architectural novelty, this study seeks to provide insights into the practical use of interpretable deep learning models in clinical decision support.

2. Methods

2.1. Dataset



(a) (b) (c) (d)

Fig. 1. Sample brain MRI images representing different severity levels of Alzheimer's disease: (a) no impairment, (b) very mild impairment, (c) mild impairment, and (d) moderate impairment.

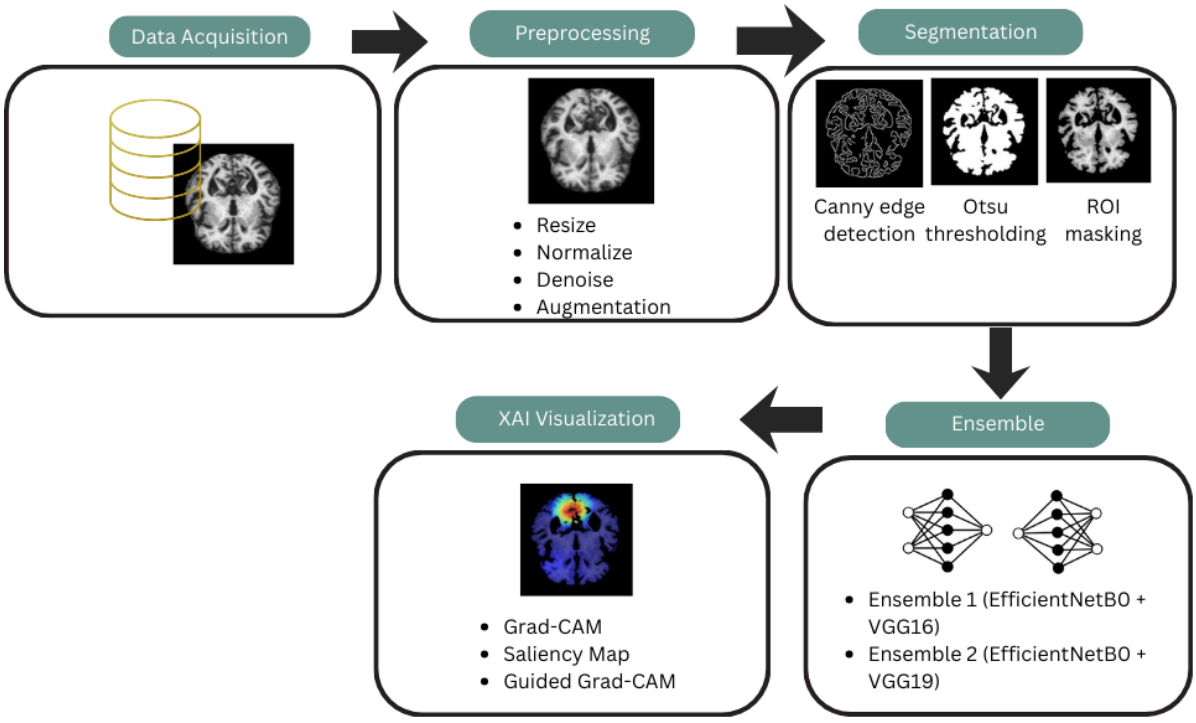


Fig. 2. Overview of the proposed deep learning framework for Alzheimer's disease classification and explainability.

This study utilizes a secondary dataset of brain Magnetic Resonance Imaging (MRI) scans obtained from a publicly available repository. The dataset consists of 5,760 MRI images categorized into four levels of cognitive impairment: no impairment, very mild impairment, mild impairment, and moderate impairment. The dataset exhibits class imbalance, with more samples in the no impairment and very mild impairment categories than in the mild and moderate impairment classes. This characteristic reflects real-world clinical conditions and presents an additional challenge for automated classification.

Using a publicly available dataset ensures reproducibility and facilitates fair comparison with previous studies. Since the dataset does not contain identifiable patient information, ethical approval was not required.

Fig. 1 presents representative MRI images from the dataset across four Alzheimer's disease severity categories, namely no impairment, very mild impairment, mild impairment, and moderate impairment. The figure illustrates the visual characteristics of each class and highlights the subtle structural differences between adjacent stages, particularly in the early phases of the disease. These similarities emphasize the challenge of automated classification and motivate the use of deep learning approaches to extract discriminative features beyond visual inspection.

2.2. Methodological Workflow

Fig. 2 illustrates the overall workflow of the proposed framework, starting from MRI data acquisition and preprocessing, followed by image segmentation and deep learning–based classification. The preprocessing stage includes image resizing, normalization, noise reduction, and data augmentation to improve model robustness. Segmentation is performed using Canny edge detection, Otsu thresholding, and

region-of-interest masking to emphasize relevant brain structures. The processed images are then fed into transfer-learning–based convolutional neural networks, including VGG16 and VGG19, with ensemble strategies used to combine model predictions. Finally, explainable artificial intelligence techniques, namely Grad-CAM, Saliency Maps, and Guided Grad-CAM, are used to generate visual explanations that highlight regions influencing the classification results.

The intended users of the proposed system are clinicians and medical researchers involved in Alzheimer's disease assessment. The system is designed as a decision-support tool, providing visual explanations to assist human interpretation of MRI-based predictions rather than performing autonomous diagnosis. The outputs of the system are intended to support clinical reasoning by highlighting image regions relevant to disease severity, while final diagnostic decisions remain under full human authority.

2.2.1. Preprocessing

Before model training, all MRI images undergo preprocessing to ensure consistency and improve data quality. Each image is resized to 224×224 pixels to match the input requirements of the pretrained convolutional neural networks. Pixel intensity values are normalized to stabilize the training process and accelerate convergence.

Noise reduction is applied to suppress irrelevant artifacts while preserving essential anatomical structures. In addition, data augmentation techniques, including rotation, horizontal flipping, zooming, and shifting, are employed to increase data diversity and mitigate the effects of class imbalance. These strategies help reduce overfitting and improve model generalization.

2.2.2. Image segmentation

To emphasize relevant brain structures and reduce background influence, an image segmentation step is applied before classification. Edge detection using the Canny algorithm is performed to highlight structural boundaries, followed by Otsu thresholding to separate foreground and background regions. A region-of-interest masking technique is then applied to focus the analysis on the brain region. This segmentation process aims to improve feature extraction by directing the model's attention to anatomically meaningful regions.

2.2.3. Model Training and Ensemble Strategy

The primary classification model employed in this study is based on the VGG16 architecture, which is widely recognized for its effectiveness in hierarchical feature extraction. Transfer learning is applied by initializing the model with pretrained weights obtained from a large-scale image dataset. The final fully connected layers are modified to accommodate the four-class Alzheimer's disease severity classification task.

During training, the convolutional base of the model is partially frozen to preserve learned low-level features, while higher-level layers are fine-tuned to adapt to MRI-specific characteristics. For comparative analysis, additional pretrained architectures are trained under the same experimental settings, and ensemble strategies are applied to combine predictions from selected models.

The dataset is split into training and test sets at 80:20. Model training is performed using the Adam optimizer with a fixed learning rate, and categorical cross-entropy is used as the loss function. Early stopping is implemented to prevent overfitting by monitoring validation loss during training.

Model performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. Confusion matrices are also analyzed to examine misclassification patterns between different severity levels.

To further evaluate model robustness, ensemble learning is used to combine predictions from selected pretrained models. Two ensemble configurations are considered, namely Ensemble 1, which combines EfficientNetB0 and VGG16, and Ensemble 2, which combines EfficientNetB0 and VGG19. The ensemble predictions are obtained using a soft voting scheme by averaging the class probability outputs of the individual models. This approach aims to assess whether combining complementary feature representations can improve classification stability compared to single-model architectures.

2.2.4. Explainable Artificial Intelligence Analysis

To enhance interpretability and clinical transparency, explainable artificial intelligence techniques are integrated into the proposed framework. Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to generate class-specific activation maps that highlight spatial regions contributing to model predictions (Selvaraju, et al., 2020). Saliency Maps are used to visualize pixel-level sensitivity, indicating how individual pixels influence the classification outcome (Samek, Wiegand, & Müller, 2017). In addition, Guided Grad-CAM combines coarse localization with fine-grained gradient information to produce more precise, detailed visual explanations. These techniques enable qualitative assessment of whether the model

Table 1
Performance metrics.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
EfficientNetB0	75.83	80.00	76.00	75.00
VGG16	95.41	95.00	95.00	95.00
VGG19	95.02	95.00	95.00	95.00
Ensemble-1	94.78	95.00	95.00	95.00
Ensemble-2	94.34	95.00	94.00	94.00
EfficientNetB0	75.83	80.00	76.00	75.00

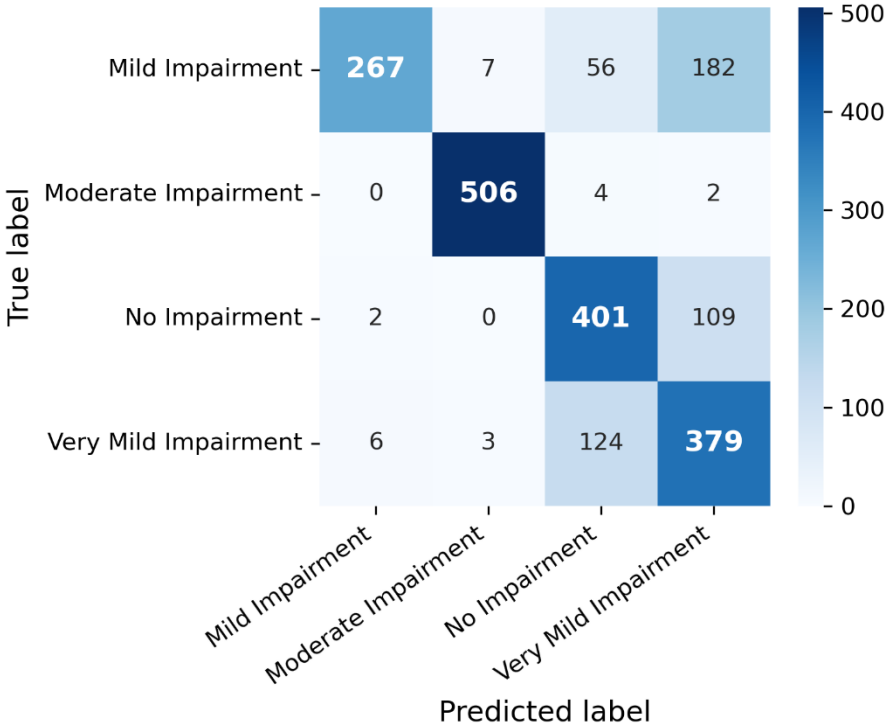


Fig. 3. Confusion matrix for the EfficientNetB0 model.

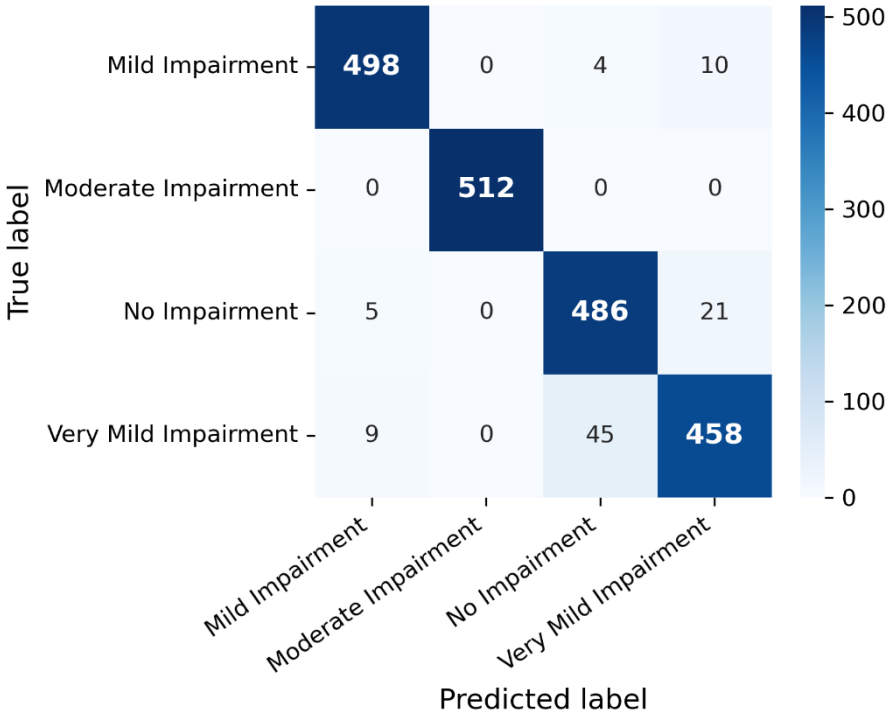


Fig. 4. Confusion matrix for the VGG16 model.

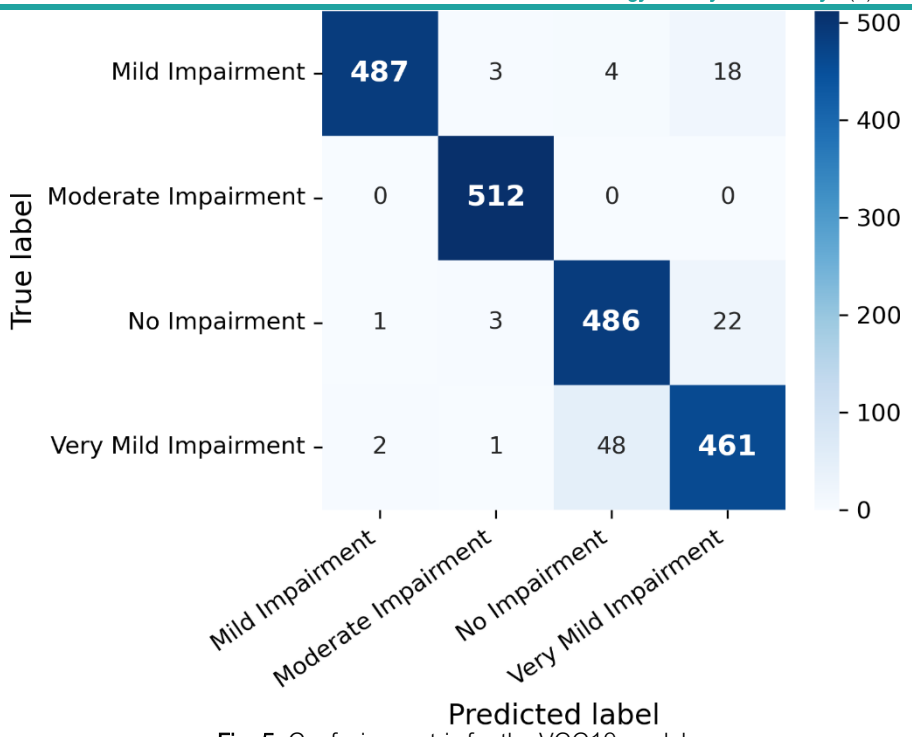


Fig. 5. Confusion matrix for the VGG19 model.

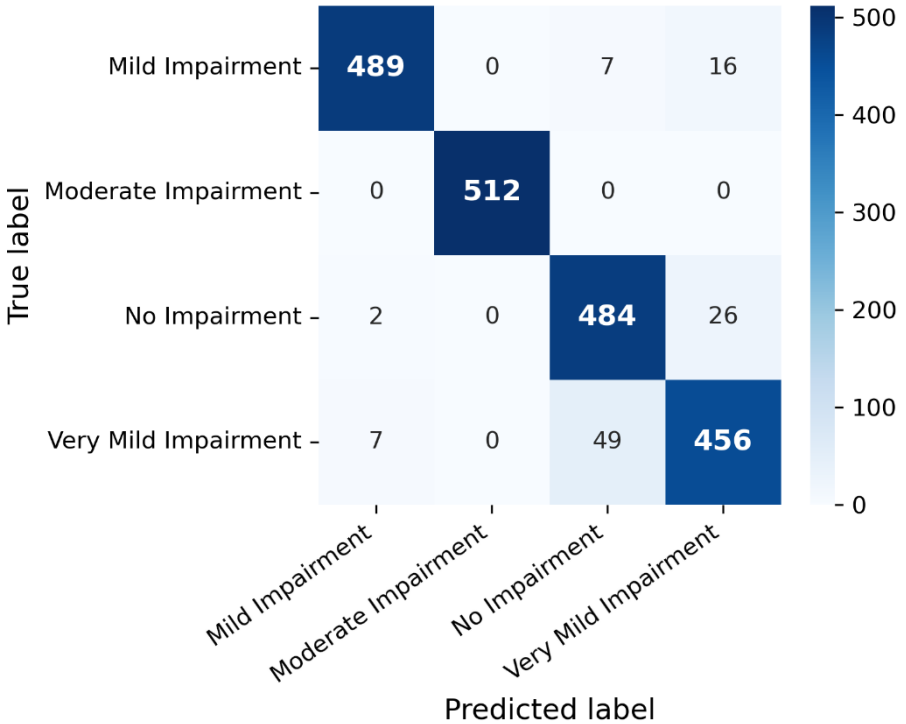


Fig. 6. Confusion matrix for the Ensemble-1 model.

focuses on anatomically and clinically relevant brain regions, thereby supporting interpretability and trust in the model's predictions (Tjoa & Guan, 2021).

3. Results and Discussion

3.1. Classification Performance

The performance of the proposed deep learning models in classifying Alzheimer's disease severity was evaluated using accuracy, precision, recall, and F1 score, as summarized in Table 1. Among the models being assessed, the VGG16 architecture achieved the highest accuracy of 95.41%, with balanced precision, recall, and F1 scores across all classes. This result indicates that VGG16 provides the most

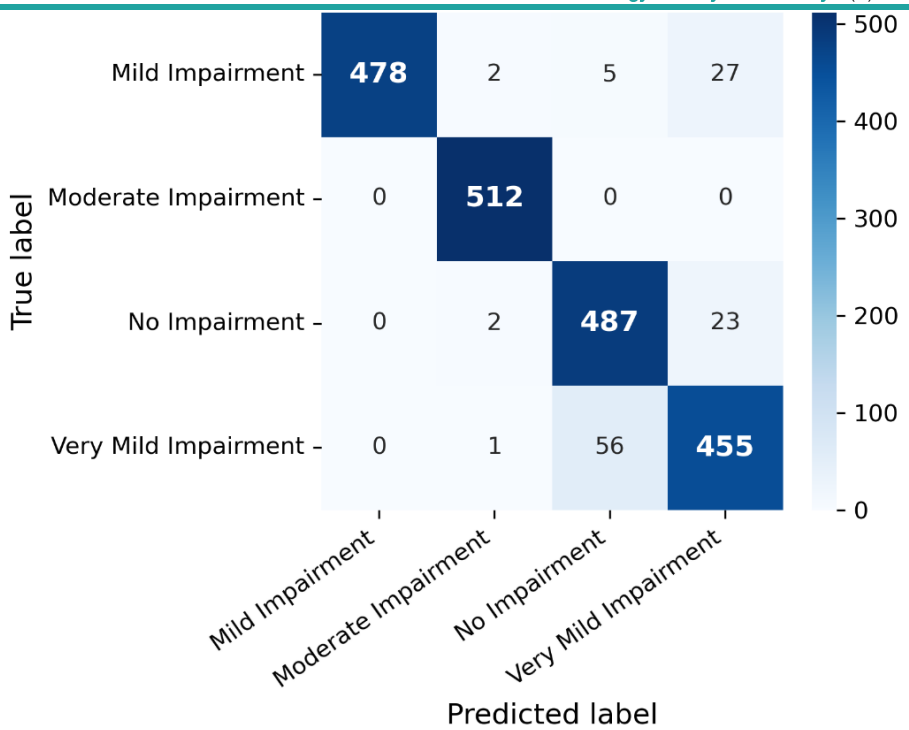


Fig. 7. Confusion matrix for the Ensemble-2 model.

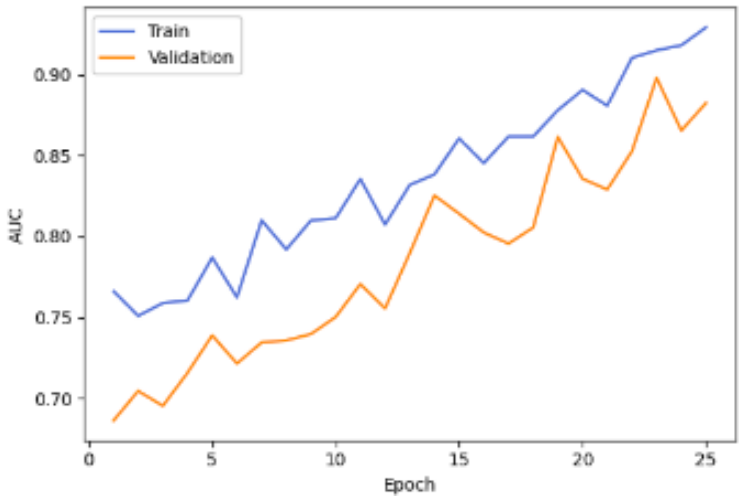


Fig. 8. AUC curve of EfficientNetB0.

reliable classification performance for MRI-based Alzheimer's disease severity assessment.

Figs. 3–5 present the confusion matrices for the individual deep learning models: EfficientNetB0, VGG16, and VGG19. As shown in Fig. 3, EfficientNetB0 yields considerable misclassifications between adjacent severity levels, particularly between no impairment and very mild impairment, as well as between very mild and mild impairment. In contrast, VGG16 (Fig. 4) demonstrates more stable performance, with most predictions concentrated along the diagonal. The VGG19 model (Fig. 5) also achieves strong performance, although minor misclassifications remain in the early-stage categories.

The confusion matrices of the ensemble models are shown in Fig. 6 and Fig. 7. Ensemble-1, combining EfficientNetB0 and VGG16, shows improved stability compared to EfficientNetB0 alone, particularly in the moderate-impairment class. Ensemble-2, combining EfficientNetB0 and VGG19, achieves similarly strong performance with minimal misclassification. However, neither ensemble model surpasses the classification accuracy achieved by the single VGG16 model, confirming that VGG16 offers the most effective balance between model complexity and performance for this dataset.

The learning behavior of each model is further analyzed using training and validation AUC and loss curves shown in Figs. 8–17. The curves indicate that all models demonstrate progressive learning with

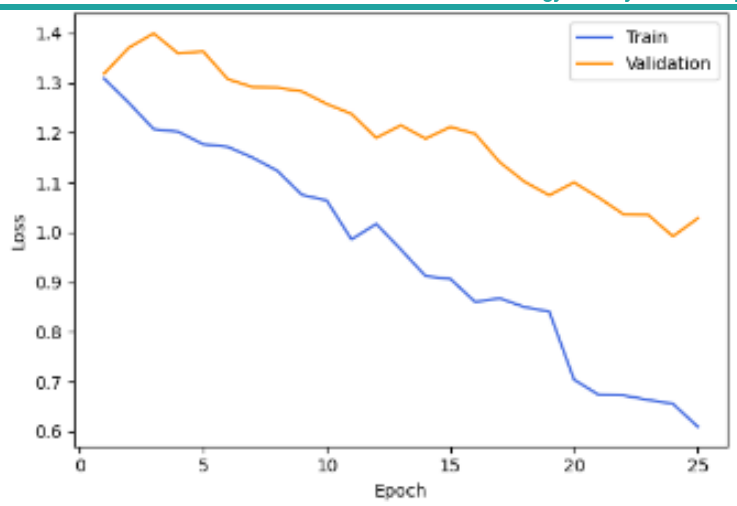


Fig. 9. Loss curve of EfficientNetB0.

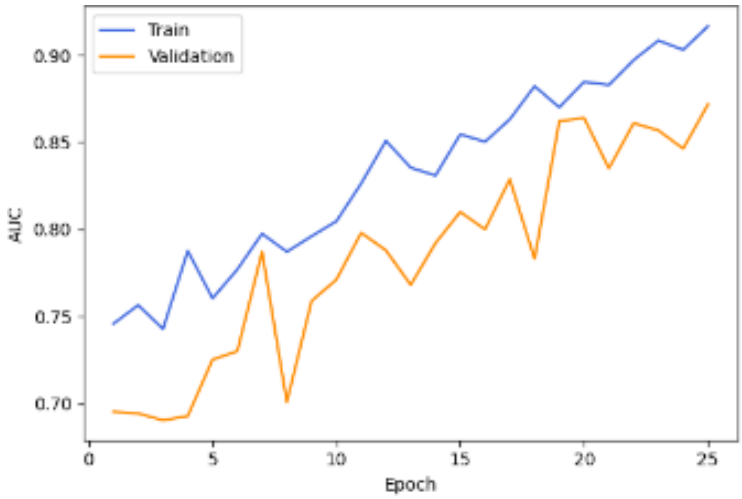


Fig. 10. AUC curve of VGG16.

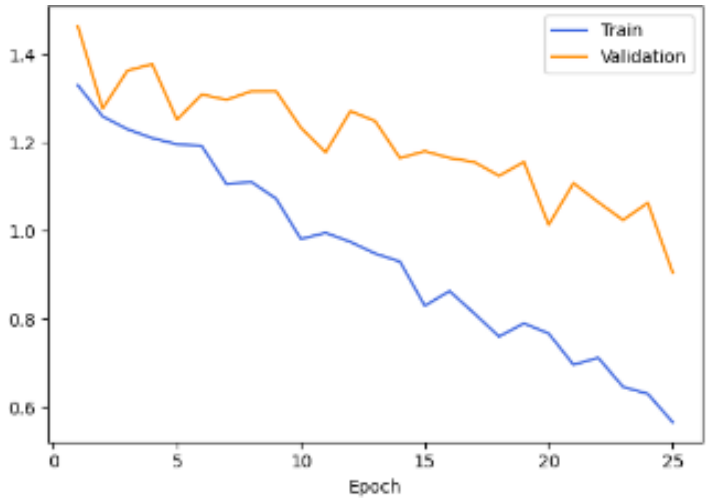


Fig. 11. Loss curve of VGG16.

decreasing loss values over epochs. Notably, the VGG16 model exhibits stable convergence with a relatively small gap between training and validation curves, suggesting good generalization capability. Although slight fluctuations are observed in the validation curves, early stopping effectively mitigates overfitting across all evaluated models.

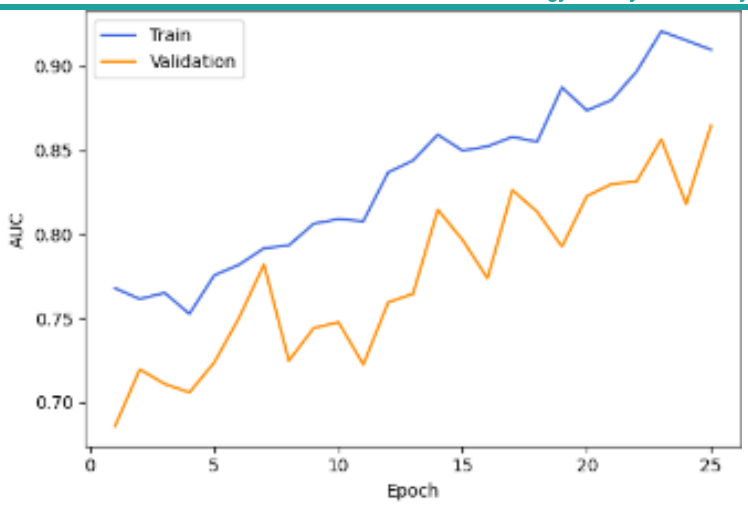


Fig. 12. AUC curve of VGG19.

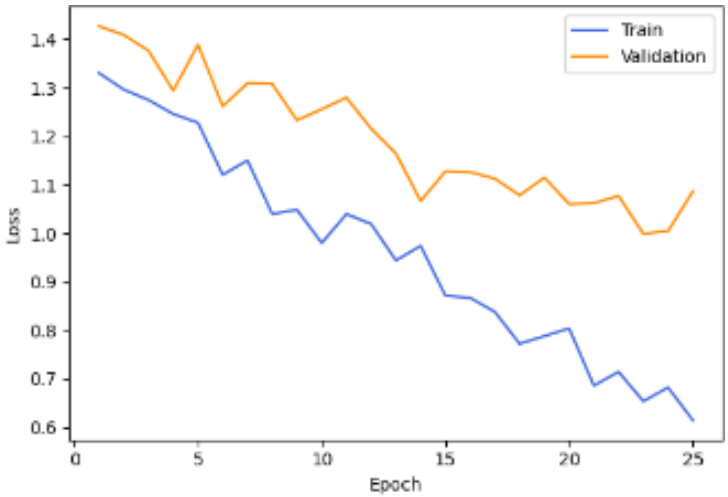


Fig. 13. Loss curve of VGG19.

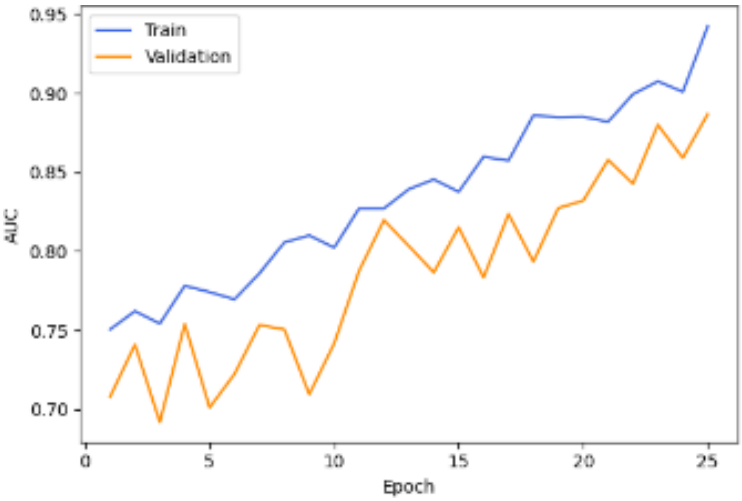


Fig. 14. AUC curve of Ensemble-1.

Although minor fluctuations are observed in the validation curves, all reported evaluation results are obtained from the held-out test set, which was not used during training, ensuring an unbiased performance assessment.

3.2. Explainable Artificial Intelligence Visualization

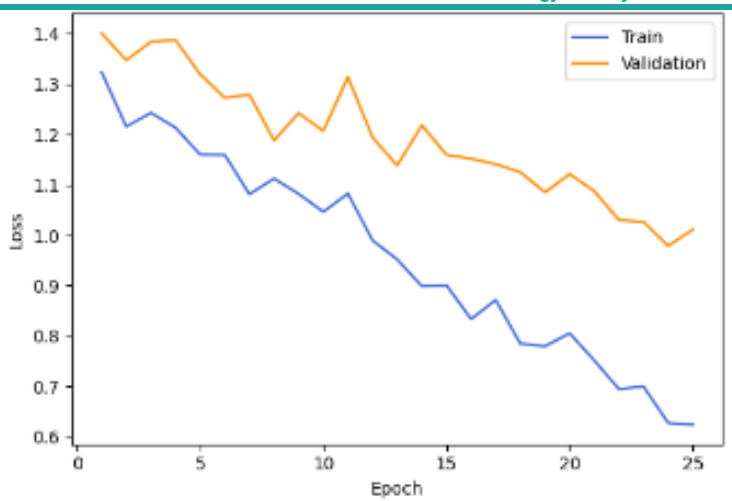


Fig. 15. Loss curve of Ensemble-1.

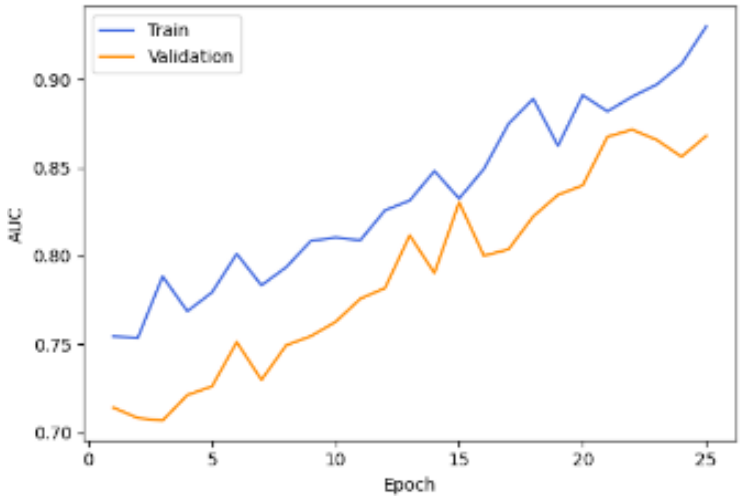
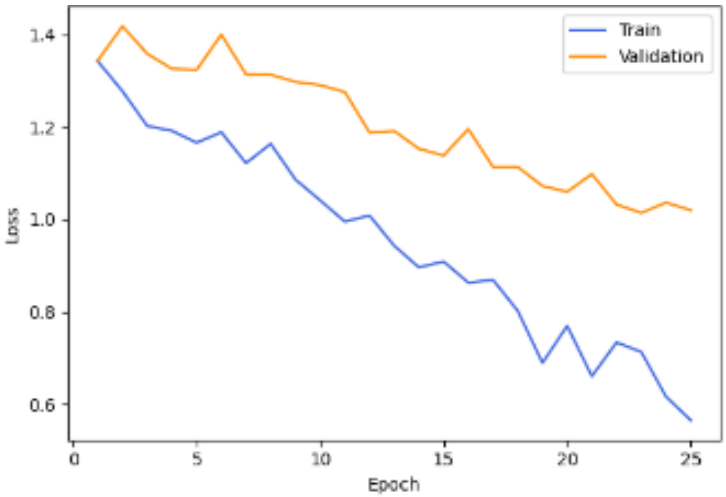


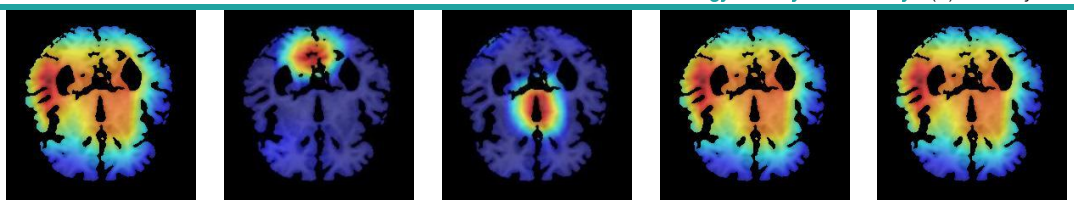
Fig. 16. AUC curve of Ensemble-2.



(b)

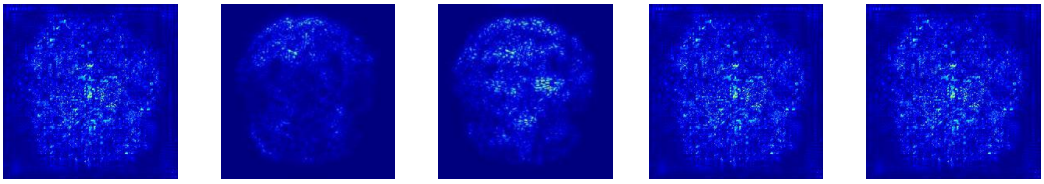
Fig. 17. Loss curve of Ensemble-2.

To enhance interpretability and assess whether the models base their predictions on clinically meaningful brain regions, several explainable artificial intelligence techniques were applied. Fig. 18 presents the Grad-CAM visualizations generated for each model. As shown in Fig. 18(b), the VGG16 model produces focused and coherent activation patterns in central brain regions commonly associated with Alzheimer's



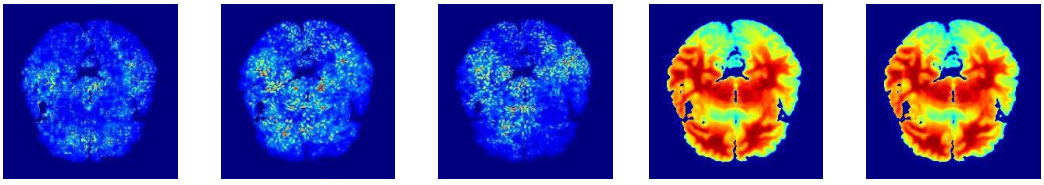
a b c d e

Fig. 18. Grad-CAM results: (a) EfficientNetB0, (b) VGG16, (c) VGG19, (d) Ensemble-1, and (e) Ensemble-2.



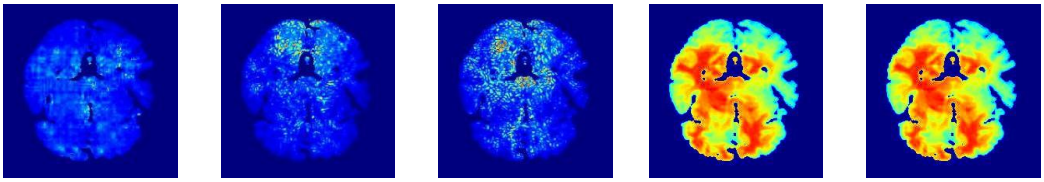
a b c d e

Fig. 19. Saliency map results: (a) EfficientNetB0, (b) VGG16, (c) VGG19, (d) Ensemble-1, and (e) Ensemble-2.



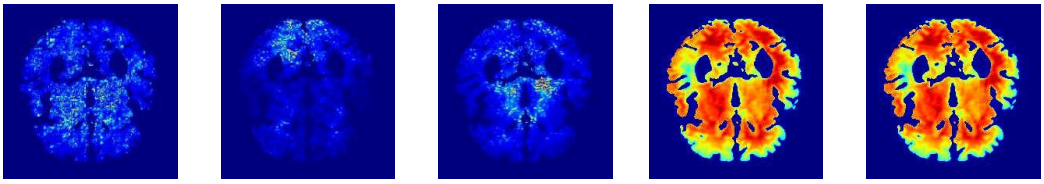
a b c d e

Fig. 20. Guided Grad-CAM results (no impairment): (a) EfficientNetB0, (b) VGG16, (c) VGG19, (d) Ensemble-1, and (e) Ensemble-2.



a b c d e

Fig. 21. Guided Grad-CAM results (very mild impairment): (a) EfficientNetB0, (b) VGG16, (c) VGG19, (d) Ensemble-1, and (e) Ensemble-2.

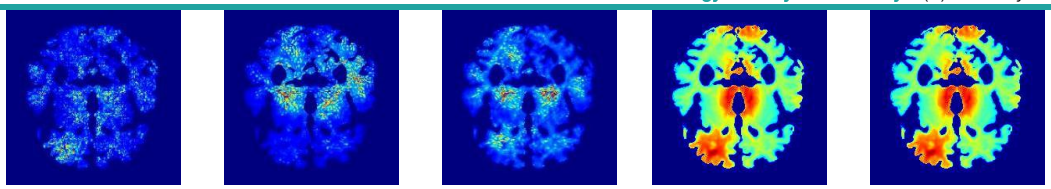


a b c d e

Fig. 22. Guided Grad-CAM results (mild impairment): (a) EfficientNetB0, (b) VGG16, (c) VGG19, (d) Ensemble-1, and (e) Ensemble-2.

disease pathology. In contrast, Fig. 18(a) shows that EfficientNetB0 exhibits more diffuse and less consistent activation, which aligns with its lower classification performance. The ensemble models in Fig. 18(d) and Fig. 18(e) demonstrate broader activation patterns that reflect the combined characteristics of their constituent models.

Pixel-level sensitivity analysis using Saliency Maps is illustrated in Fig. 19. The Saliency Map results reveal that VGG16 and VGG19 generate more concentrated and structured sensitivity patterns compared to EfficientNetB0, indicating more transparent decision-making processes. The ensemble models show



a b c d e

Fig. 23. Guided Grad-CAM results (moderate impairment): (a) EfficientNetB0, (b) VGG16, (c) VGG19, (d) Ensemble-1, and (e) Ensemble-2.

moderate improvement in sensitivity localization, although their visual explanations remain less focused than those produced by VGG16.

Further interpretability analysis is conducted using Guided Grad-CAM to examine model attention across different Alzheimer's disease severity levels. The Guided Grad-CAM visualizations for the no impairment, very mild impairment, mild impairment, and moderate impairment classes are presented in Figs. 20–23, respectively. These figures show that activation intensity and spatial coverage increase with disease severity, reflecting progressive structural brain changes captured by the model. The ensemble models generally produce more substantial and more widespread activation; however, the VGG16 model consistently highlights clinically relevant regions with clearer localization across all severity levels.

3.3. Discussion

The experimental results demonstrate that deep transfer learning is effective for MRI-based classification of Alzheimer's disease severity, particularly with the VGG16 architecture. The superior performance of VGG16, as evidenced by Table 1, Figs. 3–5, and the comparative training and validation curves shown in Figs. 8–13, suggests that its balanced network depth enables effective extraction of hierarchical features relevant to neurodegenerative changes without introducing excessive model complexity. This finding is consistent with previous studies reporting strong performance of transfer learning–based convolutional neural networks for Alzheimer's disease classification using MRI data (Basaia, et al., 2019; Islam & Zhang, 2018; Wen, et al., 2020).

Compared to existing studies that primarily emphasize improving classification accuracy through deeper architectures or multimodal fusion of MRI and PET data, the present study focuses on the systematic evaluation of explainability in transfer learning–based MRI classification of Alzheimer's disease (Mahmud, et al., 2024; Odusami, Maskeliūnas, Damaševičius, & Misra, 2023; Odusami, Damaševičius, Milieškaitė-Belousovienė, & Maskeliūnas, 2024; Sheikh, Marouf, Rokne, & Alhajj, 2025; Soladoye, Aderinto, Osho, & Olawade, 2025). Previous works such as Odusami, Maskeliūnas, Damaševičius, & Misra (2023), Odusami, Damaševičius, Milieškaitė-Belousovienė, & Maskeliūnas (2024) and Mahmud, et al. (2024) report strong performance using complex or multimodal models; however, these approaches often require increased computational resources and treat explainability as a supplementary visualization rather than a primary evaluation objective. In contrast, this study demonstrates that a single, well-optimized transfer learning model can achieve competitive performance while maintaining computational efficiency and transparent interpretability, highlighting the practical relevance of parsimonious and explainable models for clinical decision-support applications.

The explainable artificial intelligence analyses presented in Figs. 18–23 provide essential insights into model behavior. The highlighted activation regions produced by Grad-CAM, Saliency Maps, and Guided Grad-CAM align with brain areas commonly associated with Alzheimer's disease pathology, supporting the clinical plausibility of the proposed framework. These findings align with prior research highlighting the role of explainability in increasing trust and acceptance of deep learning models in medical diagnosis (Mahmud, et al., 2024; Selvaraju, et al., 2020; Khosroshahi, et al., 2025).

Unlike many existing studies that primarily apply explainability techniques as post hoc visualization tools, this work systematically evaluates interpretability across multiple disease severity levels and model architectures. The results indicate that high classification accuracy does not necessarily guarantee reliable interpretability, reinforcing the importance of jointly considering predictive performance and explanation quality. By integrating transfer learning with multiple explainable artificial intelligence methods, the proposed approach addresses a key limitation of black-box deep learning models. It strengthens their potential for real-world clinical adoption.

3.4. Limitations and Future Work

Despite the promising results, this study has several limitations that should be acknowledged. First, the experiments were conducted using a secondary, publicly available MRI dataset that may not fully reflect the heterogeneity of Alzheimer's disease across different populations, imaging protocols, and clinical settings. Consequently, the generalizability of the proposed model to data acquired from other institutions or scanners may be limited.

Second, the dataset exhibits class imbalance, particularly in the moderate impairment category, which may affect classification performance despite the use of data augmentation techniques. Although the model demonstrates strong overall accuracy, imbalanced data distribution can still affect the robustness of predictions for underrepresented classes.

Third, the proposed framework relies on two-dimensional MRI slices rather than complete three-dimensional volumetric data. While this approach reduces computational complexity and enables efficient training, it may not capture complete spatial information related to Alzheimer's disease progression. In addition, the explainable artificial intelligence analysis is primarily qualitative, and no direct clinical validation was performed to assess the alignment between highlighted regions and expert annotations quantitatively.

The impact of preprocessing and segmentation was not quantitatively evaluated through ablation experiments in this study. These steps were adopted based on established practices to improve input consistency and model convergence; however, their isolated effects remain an critical direction for future work. Future studies may incorporate systematic ablation experiments to assess the individual contribution of preprocessing and segmentation steps across different datasets.

Future work will focus on addressing these limitations by incorporating larger and more diverse multi-center datasets to improve model generalizability. The use of three-dimensional MRI volumes and longitudinal imaging data will also be explored to capture more comprehensive structural information and disease progression patterns. Furthermore, integrating quantitative evaluation of explainability, such as comparison with expert-labeled regions of interest, may strengthen the clinical relevance of the proposed framework. Finally, extending the approach to multimodal neuroimaging data, including PET or functional MRI, could further enhance diagnostic accuracy and interpretability.

4. Conclusions

This study presents a transfer learning–based deep learning framework complemented by explainable artificial intelligence (XAI) techniques for Alzheimer's disease classification using brain MRI images. By leveraging a pretrained VGG16 model, the proposed approach achieves high classification accuracy while maintaining stable performance across different disease severity levels. The integration of Grad-CAM, Saliency Maps, and Guided Grad-CAM provides transparent visual explanations that highlight clinically relevant brain regions associated with Alzheimer's disease pathology.

The results demonstrate that a single, well-optimized convolutional neural network, when combined with explainability methods, can offer a favorable balance between predictive performance and interpretability. These findings indicate that the proposed framework has potential as a reliable, interpretable clinical decision-support tool for assisting in Alzheimer's disease diagnosis. Future work will focus on validation using larger and multi-center datasets, three-dimensional MRI representations, and quantitative clinical evaluation to assess generalizability and clinical utility further.

5. Declaration of AI and AI assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to assist in language refinement and clarity of presentation. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

6. CRediT Authorship Contribution Statement

Dea Amanda Salsabila: Conceptualization, Data curation, Formal analysis, Investigation, Software development, Validation, Visualization, and Writing – original draft. **Ghaluh Indah Permata Sari**: Conceptualization, Methodological guidance, Supervision, Validation, and Writing – review & editing. **Fajar Astuti Hermawati**: Formal analysis, Methodological guidance, Resources, Project administration, and Writing – review & editing.

7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. Acknowledgments

The authors would like to thank the institutional support and facilities that made this research possible.

9. Data Availability

The dataset used in this study is publicly available and can be accessed from the original source at: https://www.kaggle.com/datasets/lukechugh/best-alzheimer-mri-dataset-99-accuracy.

10. Funding

The authors declare that this research received no external funding.

11. Ethical Approval

This study used publicly available secondary data and did not involve patient-identifiable information. Therefore, ethical approval was not required.

12. References

AbdelAziz, N. M., Said, W., AbdelHafeez, M. M., & Ali, A. H. (2024). Advanced interpretable diagnosis of Alzheimer's disease using SECNN-RF framework with explainable AI. *Frontiers in Artificial Intelligence, 7*. doi:https://doi.org/10.3389/frai.2024.1456069

Aderghal, K., Benois-Pineau, J., Afdel, K., & Gwenaëlle, C. (2017). FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+ε projections. *CBMI '17: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing.* Florence, Italy: ACM. doi:https://doi.org/10.1145/3095713.3095749

Ali, M. U., Hussain, S. J., Khalid, M., Farrash, M., Lahza, H. F., & Zafar, A. (2024). MRI-Driven Alzheimer's Disease Diagnosis Using Deep Network Fusion and Optimal Selection of Feature. *Bioengineering, 11*(11). doi:https://doi.org/10.3390/bioengineering11111076

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical, 21*. doi:https://doi.org/10.1016/j.nicl.2018.101645

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., . . . W. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage: Clinical, 31*. doi:https://doi.org/10.1016/j.nicl.2021.102712

Chattopadhyay, T., Joshy, N. A., Jagad, C., Gleave, E. J., Thomopoulos, S. I., Feng, Y., . . . Thompson, P. M. (2024, September 17). *Comparison of Explainable AI Models for MRI-based Alzheimer's Disease Classification.* doi:https://doi.org/10.1101/2024.09.17.613560

De Santi, L. A., Pasini, E., Santarelli, M. F., Genovesi, D., & Positano, V. (2023). An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F-FDG PET. *Journal of Digital Imaging, 36*, 189-203. doi:https://doi.org/10.1007/s10278-022-00719-3

El-Assy, A. M., Amer, H. M., Ibrahim, H. M., & Mohamed, M. A. (2024). A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data. *Scientific Reports, 14*. doi:https://doi.org/10.1038/s41598-024-53733-6

Islam, J., & Zhang, Y. (2018). Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks Original research Open access Published: 31 May 2018. *Brain Informatics, 5*. doi:https://doi.org/10.1186/s40708-018-0080-3

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., . . . Ran, K. P. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's and Dementia, 14*(4), 535-562. doi:https://doi.org/10.1016/j.jalz.2018.02.018

Khosroshahi, M. T., Morsali, S., Gharakhanlou, S., Motamedi, A., Hassanbaghlou, S., Vahedi, H., . . . Jafarizadeh, A. (2025). Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. *Diagnostics, 15*(5). doi:https://doi.org/10.3390/diagnostics15050612

Komal, R., Dhavakumar, P., Rahul, K., Jaswanth, B., & Preeth, R. (2025). Hybrid deep learning framework for magnetic resonance imaging-based classification of Alzheimer's disease. *Brain Network Disorders, 1*(4), 239-249. doi:https://doi.org/10.1016/j.bnd.2025.06.002

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., . . . Mika. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet, 396*, 413–46. doi:https://doi.org/10.1016/S0140-6736(20)30367-6

Mahmud, T., Barua, K., Habiba, S. U., Sharmen, N., Hossain, M. S., & Andersson, K. (2024). An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning. *Diagnostics, 14*(3). doi:https://doi.org/10.3390/diagnostics14030345

Odusami, M., Damaševičius, R., Milieškaitė-Belousovienė, E., & Maskeliūnas, R. (2024). Alzheimer's disease stage recognition from MRI and PET imaging data using Pareto-optimal quantum dynamic optimization. *Heliyon, 10*. doi:https://doi.org/10.1016/j.heliyon.2024.e34402

Odusami, M., Maskeliūnas, R., Damaševičius, R., & Misra, S. (2023). Explainable Deep-Learning-Based Diagnosis of Alzheimer's Disease Using Multimodal Input Fusion of PET and MRI Images. *Journal of Medical and Biological Engineering, 43*, 291-302. doi:https://doi.org/10.1007/s40846-023-00801-3

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage, 155*, 530-548. doi:https://doi.org/10.1016/j.neuroimage.2017.03.057

Samek, W., Wiegand, T., & Müller, K.-R. (2017, Aug 28). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.* doi:https://doi.org/10.48550/arXiv.1708.08296

Sampath, R., & Baskar, M. (2024). Alzheimer's Disease Prediction Using Fly-Optimized Densely Connected Convolution Neural Networks Based on MRI Images. *The Journal of Prevention of Alzheimer's Disease, 11*(4), 1106-1121. doi:https://doi.org/10.14283/jpad.2024.66

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision, 128*, 336-359. doi:https://doi.org/10.1007/s11263-019-01228-7

Sheikh, F., Marouf, A. A., Rokne, J. G., & Alhajj, R. (2025). Lightweight Deep Learning Models with Explainable AI for Early Alzheimer's Detection from Standard MRI Scans. *Diagnostics, 15*(21). doi:https://doi.org/10.3390/diagnostics15212709

Shuvo, S. S., Refat, S. R., Preotee, F. F., & Muhammad, T. (2025). Advanced CNN and Explainable AI Based Architecture for Interpretable Brain MRI Analysis. *ICCA '24: Proceedings of the 3rd International Conference on Computing Advancements* (pp. 319-326). Dhaka, Bangladesh: ACM. doi:https://doi.org/10.1145/3723178.3723220

Soladoye, A. A., Aderinto, N., Osho, D., & Olawade, D. B. (2025). Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data. *International Journal of Medical Informatics, 204*. doi:https://doi.org/10.1016/j.ijmedinf.2025.106093

Sorour, S. E., El-Mageed, A. A., Albarrak, K. M., Alnaim, A. K., Wafa, A. A., & El-Shafeiy, E. (2024). Classification of Alzheimer's disease using MRI data based on Deep Learning Techniques. *Journal of King Saud University - Computer and Information Sciences, 36*(2). doi:https://doi.org/10.1016/j.jksuci.2024.101940

Tjoa, E., & Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793-4813. doi:https://doi.org/10.1109/TNNLS.2020.3027314

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., . . . Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis, 63*. doi:https://doi.org/10.1016/j.media.2020.101694