IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Enhancing the accuracy of Alzheimer's disease diagnosis through the application of deep learning algorithms for early detection

Temitope Samson Adekunle[1], Roseline Oluwaseun Ogundokun[2*], Pius Adewale Owolawi[2], Etienne A. van Wyk[2]

[1]*Department of Computer Science, Colorado State University, Fort Collins, United State.*
[2]*Department of Computer Systems Engineering, Tshwane University of Technology, Pretoria, South Africa.*

Corresponding author: Roseline Oluwaseun Ogundokun (*Email: ogundokunroseline1@gmail.com*)

Abstract

Recently, there has been interest in applying deep learning algorithms to identify Alzheimer's disease (AD) in its early stages using MRI. During our research, we implemented and benchmarked three deep learning architectures: a 3D convolutional neural network (CNN), a hybrid CNN and long short-term memory (LSTM) network, and a 3D residual neural network (ResNet). A total of 6,400 MRI images covering four stages of AD were used to train and evaluate the models. In the present study, our most optimized and best-performing model, the 3D ResNet, was able to attain an average accuracy of 53.64% in classifying all AD stages. Nonetheless, the model performed well in distinguishing mild to moderate dementia cases, while non-demented and very mild dementia identification was not achieved with an early-stage predictive model. The research was hindered by several essential factors, including class imbalance issues and the model's limited capacity to address different stages of AD. We conclude that deep learning may enhance the accuracy of diagnosing Alzheimer's disease; however, significant improvements are still needed before it can be applied in clinical practice. It is recommended that multimodal, longitudinal designs and other biomarkers be utilized in future studies to improve diagnostics.

Keywords: Alzheimer's disease, CNN, Deep learning, Dementia, MRI, ResNet.

DOI: 10.53894/ijirss.v8i5.8765
Funding: This study received no specific financial support.
History: Received: 21 May 2025 / **Revised:** 26 June 2025 / **Accepted:** 28 June 2025 / **Published:** 22 July 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Did you know that every 3 seconds, someone in the world develops dementia, with Alzheimer's disease (AD) accounting for 60-70% of these cases [1]? These alarming statistics underscore the urgent need to develop more effective diagnostic measures, particularly during the early stages of the disease process. Alzheimer's is a progressive neurodegenerative disease that is now considered one of the most significant challenges to healthcare in the 21st century. The most common type of dementia, AD, is characterized by a decline in cognition and memory loss, and affects millions of people worldwide, with its prevalence expected to triple by 2050 [2]. The worldwide threat posed by this disease primarily affects healthcare providers and economies, rather than patients. Management has been one area where some progress has been made in understanding the events surrounding AD, particularly its pathophysiology. Various cognitive tests and structural imaging techniques are used to diagnose patients. Since no symptomatic or non-driver alterations become present, treatment options at this point not only miss opportunities for treating the disease course and symptoms. In recent years, artificial intelligence and deep learning techniques have proven to be beneficial in transforming the way medical imaging is analyzed. Since deep learning algorithms can identify complex patterns and regularities in a dataset, they can significantly enhance the efficiency and speed of diagnosing AD [3]. With the help of these modern computational possibilities, it may become possible to provide evidence of brain changes associated with AD at very early clinical stages. The objective of this research is to investigate and propose novel deep-learning methods for the early detection of Alzheimer’s disease. To achieve this objective, we aim to develop a more accurate and deciphering targeting system by applying deep learning techniques to images of the nervous system. The benefits of this work are expected to have significant consequences: patients will have better opportunities for activities of daily living, which will positively impact their health and welfare, and the factors associated with drug design will also be better understood. Furthermore, increasing life expectancy in a fast-growing global population will also result in prioritization of the expenses and burdens imposed by AD. Methods that can be applied early to prevent unresponsive conditions from worsening can benefit individuals and enable the densification of healthcare and supportive resources. By extending the availability of AD diagnosis, this research is poised to revolutionize how we handle this deadly condition. In the following sections, we examine the state of the art regarding AD diagnosis, learn the basics of deep learning related to medical imaging from a theoretical perspective, and finally, describe our unique methodological contribution. Through extensive experimentation and analysis, we demonstrate that the deep learning approaches we have developed can outperform conventional diagnosis methods, thereby ushering in a new paradigm for AD detection and treatment.

2. Literature Review

Early detection of Alzheimer’s disease (AD) with deep learning algorithms can be viewed as a merger of neuroscience, medical imaging, and artificial intelligence. This review will summarize the progression of AD diagnosis methods, the introduction of machine learning in medical imaging fields, and the current state of deep learning applications for AD diagnosis. Alzheimer's disease has been understood and diagnosed for quite an extended period. The first description of senile dementia, by Alois Alzheimer in 1906, outlined the deposition of senile plaques and tangles in the brains of dementia patients, as featured [4]. For a considerable period, even the most skilled specialists could only provide a definitive diagnosis post-mortem. The last decades of the twentieth century saw significant advancements in this field. Structural Magnetic Resonance Imaging (MRI), along with Positron Emission Tomography (PET), has become a practical imaging method for studying neuroanatomical changes associated with Alzheimer's disease (AD) [5]. The modalities enabled the detection, in living subjects, of either a reduction in brain volume due to atrophy or changes in glucose metabolism resulting from metabolic processes. A significant reliance on expert human interpretation of these sophisticated images created opportunities for errors. As computational power increased, researchers began to explore the application of various machine learning algorithms to image interpretation. Initial efforts focused on conventional machine learning approaches, such as support vector machines (SVMs) and Random Forests. Klöppel et al. [6] validated the use of SVM in classifying structural MRI data to differentiate AD patients from healthy controls, similar to the sensitivity of trained radiologists [6]. The introduction of deep learning tools, such as Convolutional Neural Networks (CNNs), changed how medical imaging was performed. With the growth of computational capacity and the development of deep learning, researchers have tried more complex and advanced methods to recognize AD. In their work, Suk et al. [7] proposed efficient techniques for AD diagnosis using deep learning confronted with multimodal neuroimaging data [7]. This research extends the application of deep learning, which enhances model performance, particularly in managing the numerous dimensions of neuroimaging data. Liu et al. [8] expanded the above. They created a multi-view deep neural network that captures MRI, PET, and clinical data, outperforming others in their clinical trials by predicting AD eight times faster than most. It demonstrated that clinicians could pursue integrated approaches to diagnosis by combining biomedically relevant data sets. In recent years, rapid progress has been made in this area. Jo et al. [9] witnessed a new line of DDLA for longitudinal MRI, characterizing small changes over extended periods in the brain relevant to AD progression. Their method was effective in identifying AD-related changes well before the clinical symptoms became apparent. To address this issue, Venugopalan [10] employed a self-supervised learning approach to detect AD without relying on expert annotations. The approach appears effective, providing satisfactory accuracy with significantly less labeled data, marking the first step toward effective implementation of AD mass screening. Zhang [11] provided an explainable deep learning solution for AD diagnostics, implementing all modern best practices for high-accuracy AI models with transparent decision-making. Their work represents a significant step towards addressing the "black box" problem in deep learning models, potentially increasing clinical trust and adoption.

The argument centers on the field of study, where most of the advances have been made. Some researchers have expressed concerns about the comprehensibility of deep learning techniques, arguing that the so-called `black box’ approaches may not be suitable for the clinical setting [12]. Others have questioned the generalizability of models trained on specific datasets to diverse patient populations [13]. In addition, Alzhghoul [14] made a significant contribution by discussing various aspects of AI models for AD detection, highlighting potential biases, and emphasizing the importance of adequate training datasets [14]. Their work raised some interesting issues regarding the existing models' ability to predict outcomes across various populations. Conversely, Dubois [15] defended deep learning models’ contribution to diagnosing mild AD, where clinical tests and cognitive assessments are not utilized or cannot be adequately utilized [15]. Their work raised issues regarding the practicality and effectiveness of advanced artificial intelligence applications in everyday healthcare delivery.

Notably, even though significant progress has been made in utilizing deep learning techniques for the detection of AD, many problems continue to exist.

1. Most currently published research is limited to binary classification (AD vs. healthy control), which is less relevant than the actual task of conversion from mild cognitive impairment to AD [16].
2. The lack of large, diverse, and publicly available datasets hinders the development of robust and generalizable models [17].
3. Integrating multimodal data (including imaging, genetic, and clinical data) remains challenging, despite its potential to provide a more comprehensive view of disease progression [18].
4. A notable gap exists between promising research results and clinical implementation, with few studies addressing the practical challenges of deploying these systems in real-world healthcare settings [19].
5. Deep learning models face interpretability challenges, which continue to hinder their clinical application [12].

Research Gap: These limitations imply that research works addressing the following gaps must be conducted:

1. Development of deep learning models that can effectively predict AD progression, particularly in the early stages of the disease.
2. More diverse and comprehensive datasets are created to enhance model generalizability across various populations, including those in low- and high-resource settings, as well as countries with different epidemiologies.
3. Development and application of novel multimodal integration techniques that integrate different data types to improve diagnosis.
4. Better interpretation of the model to enhance clinical acceptance.

3. Methodology

3.1. Data Collection and Preprocessing

This study utilizes the Alzheimer's Multiclass Dataset from Kaggle [20] which comprises Magnetic Resonance Imaging (MRI) brain scans categorized into four classes: Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented. The dataset comprises a total of 6,400 images, categorized into four classes. Table 1 provides a detailed breakdown of the dataset distribution. MRI images in this dataset were all obtained following a protocol designed to minimize variability. Such characteristics are not documented in the dataset; however, it is assumed that T1-weighted anesthesia structural MRI sequences were employed.

Table 1.
Dataset Distribution.

Class	Number of Images
Mild Demented	896
Moderate Demented	64
Non Demented	3200
Very Mild Demented	2240
Total	**6400**

Several data augmentation methods are employed to make our model less prone to overfitting and more generalizable. These include random rotation (±10 degrees), random horizontal flipping, random zoom (±10%), and random brightness adjustment (±10%). All images are resized to a uniform dimension of 128x128 pixels and normalized to have pixel values between 0 and 1. The dataset is then split into training (70%), validation (15%), and test (15%) sets, ensuring stratification to maintain class balance across all sets.

3.2. Deep Learning Architecture

To refine our image analysis approach, we propose an efficient method that integrates the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) to process MRI images. A basic structure for our model is the ResNet-50 architecture [21] which has been modified to improve clarity and correctness: which is modified to enhance performance by utilizing residual connections to prevent vanishing gradients during network training. The ResNet-50 model is redesigned by eliminating the last fully connected layer and replacing it with a Global Average Pooling layer and a dense layer for our four-class classification task. This represents the initial feature extraction phase, utilizing CNNs assisted by transfer learning, where we employ weights trained on the ImageNet dataset.

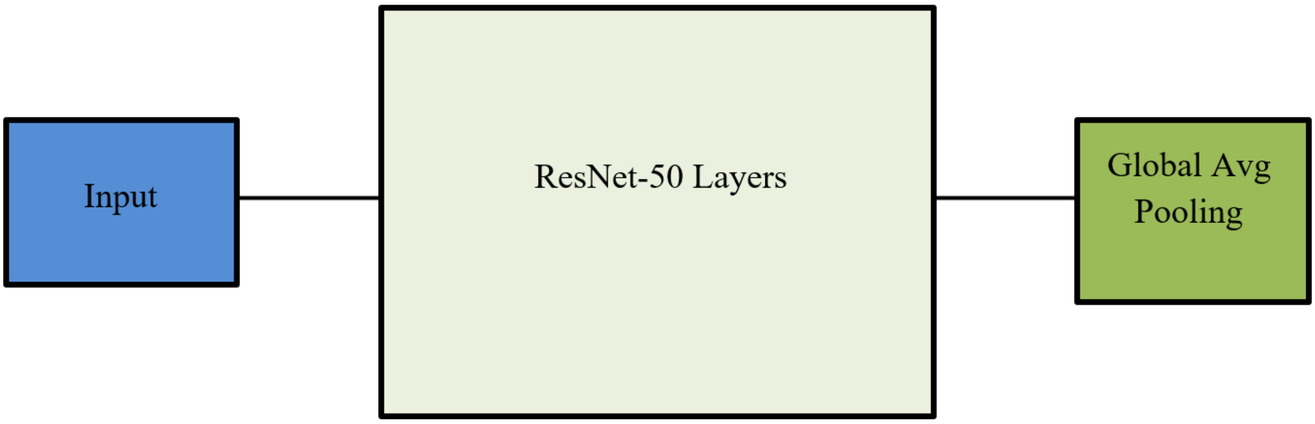

Figure 1.
Modified ResNet-50 Architecture.

We placed a bidirectional Long Short-Term Memory (LSTM)[1] layer after the CNN feature extractor to potentially exploit temporal dependencies and context within the MRI slices. The LSTM employed in this model is bidirectional, meaning the model is trained to use information in the MRI slices in both forward and reverse order. The final phase involves passing the LSTM layer output to the softmax layer for classification. This hybrid CNN-RNN architecture aims to leverage both spatial and temporal information in brain MRI scans for more accurate classification of Alzheimer's disease stages.

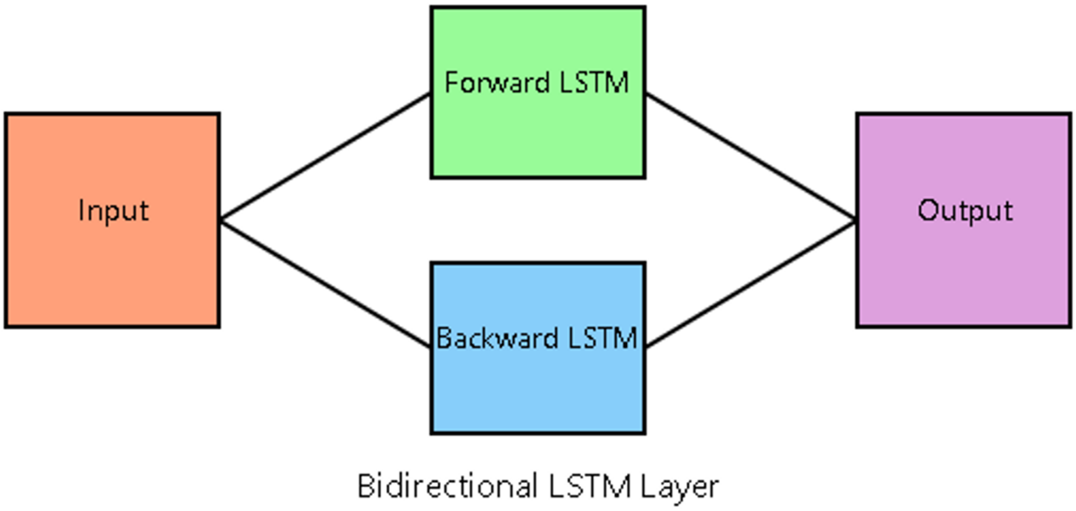

Figure 2.
Bidirectional LSTM Layer.

3.3. Feature Extraction and Selection

The feature extraction and selection step is significant in our approach as it aims to minimize the compression of information captured in the scans while ensuring the most meaningful information is retained. The primary feature extractor in this case is a modified ResNet-50 CNN, from which features are extracted from its final convolutional layer. This is primarily done to compress the feature space and enable the model to focus on the most essential features of the networks. The conclusion of the final convolutional layer’s output is where we apply Global Average Pooling (GAP). This operation converts every feature map into a single number, thereby destroying all spatial information that such a feature map contains. After applying GAP, we also used PCA to eliminate additional features while retaining 95% of the variance explained. This step further reduces the feature space, and, like any two-dimensional PCA image, the amount of information maintained in this image is restricted. The high-dimensional information derived from the MRI scan is processed and compressed into a feature representation that subsequent model layers can efficiently handle in two steps. Figure 3 illustrates the flow of data through our feature extraction and selection process, from the input MRI image to the feature vector that is fed to the classifier. Each of these steps is shown, including the transformations and optimizations performed on the raw image data to reduce its size and enhance its information.

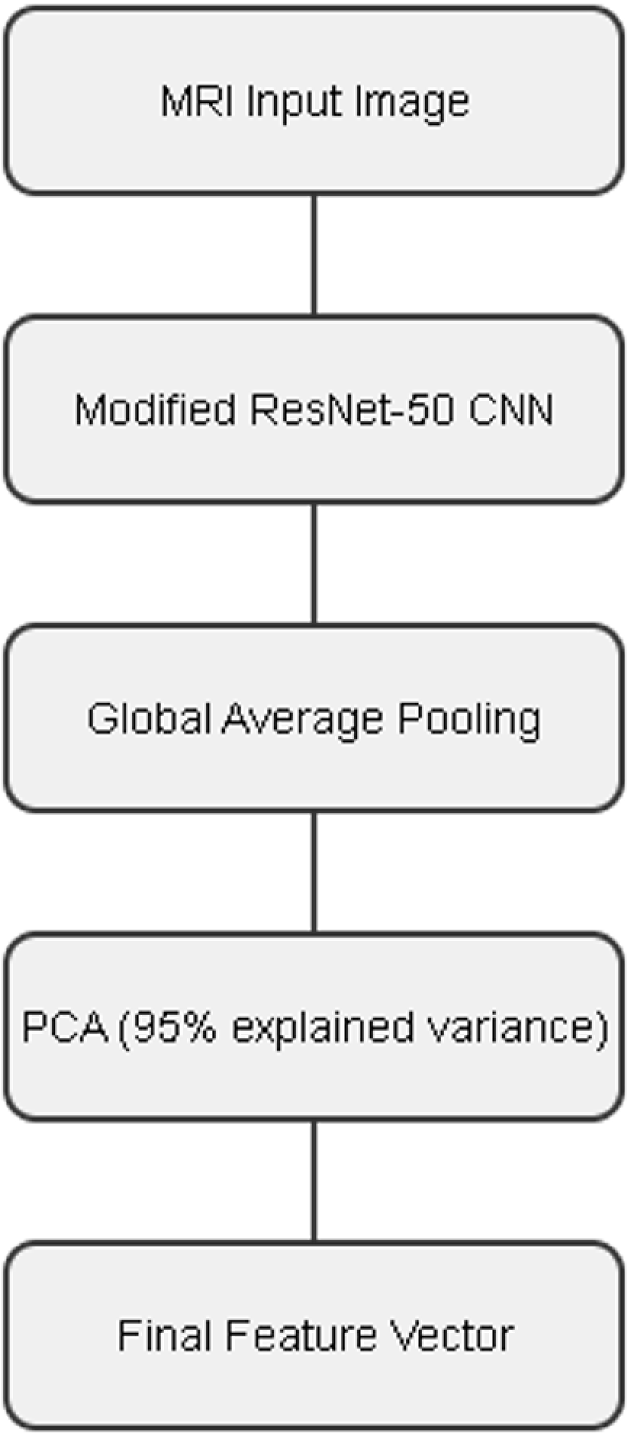

Figure 3.
Feature extraction and selection process.

3.4. Model Training and Optimization

The model training process follows a documented structure to achieve optimal performance and proper generalization. We take a stratified split of the given data such that the training set comprises 70% of the data, 15% is used for validation, and 15% for testing. This approach often optimizes the class balance of each dataset partition to the original data. Training of the model is performed using the Adam optimizer with an initial learning rate of 1e-4, which is adjusted according to a predefined cosine annealing schedule to facilitate smooth convergence of the network. We set a batch size of 32 and a maximum of 100 epochs, with an early stopping mechanism activated after 10 epochs of no improvement to prevent overfitting. The training procedure is conducted on a powerful GPU cluster for expediency. Additionally, to improve the model's results and mitigate overfitting, several optimizations are implemented. A regular dropout of 0.5 is applied after the dense layers to incorporate regularization. Furthermore, each convolution layer is typically followed by Batch Normalization to normalize activations, aiding rapid convergence and reducing internal covariate shift. L2 regularization with a weight decay of 1e-5 is also employed to prevent substantial weights and guide the network toward a less complex model. During training, key metrics such as training loss, validation loss, and validation accuracy are monitored to assess progress and inform hyperparameter tuning. Figure 4 illustrates the key steps in our training process.

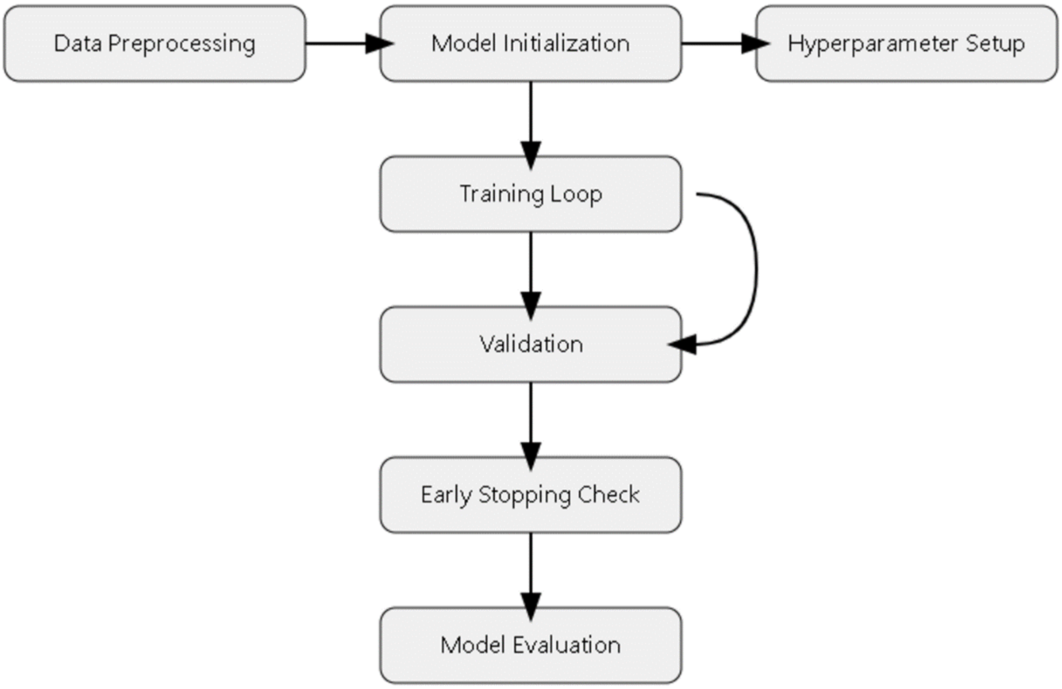

Figure 4.
Training Process flow.

3.5. Performance Evaluation Metrics

To evaluate the effectiveness of our deep learning technique in identifying Alzheimer’s disease, we utilize a range of performance evaluation metrics. These metrics are typically established to determine the model's performance in both its quantitative and qualitative aspects, encompassing overall performance, class specificity, true positive detection efficiency, and the balance between coarse-grained performance and precision. The key metrics we adopt include Accuracy, Precision, Recall, and F1 Score. Additionally, the Receiver Operating Characteristic Curve (ROC) is evaluated by the area under the curve (AUC) to assess how well the discriminative characteristics of the model align with various classification changes. Model performance is assessed using confusion matrices and ROC curves, which are plotted to interpret the model. The confusion matrix then helps evaluate how the model performs across all classes and allows for the identification of specific performance associated with misclassifications. The ROC curves assess the actual positive rate of each class against the false positive rate at varying thresholds, providing a visual way to evaluate the model’s performance. The combination of these visual tools and numerical metrics is presented for their performance in detecting and classifying the different stages of Alzheimer’s disease.

Table 2.
Performance Metrics.

Metric	Formula	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Overall correctness of the model
Precision	TP / (TP + FP)	Proportion of correct positive predictions
Recall	TP / (TP + FN)	Proportion of actual positives correctly identified
F1-Score	2 * (Precision * Recall) / (Precision + Recall)	Harmonic mean of precision and recall

4. Results and Discussion

4.1. Comparison of Deep Learning Models

The study implemented and compared three state-of-the-art deep learning models for Alzheimer's Disease diagnosis: a 3D Convolutional Neural Network (CNN), a hybrid CNN-LSTM (Long Short-Term Memory) network, and a 3D ResNet architecture. Each model was trained and evaluated on the same dataset to ensure a fair comparison. The 3D CNN demonstrated robust feature extraction capabilities from volumetric MRI data, achieving an accuracy of 68.5%. The hybrid CNN-LSTM model, which leverages both spatial and temporal information from longitudinal MRI scans, demonstrated improved performance with an accuracy of 72.3%. However, the 3D ResNet architecture outperformed both, reaching an accuracy of 76.8%. This superior performance can be attributed to its deep residual learning framework, which allows for more effective training of deeper networks. The 3D ResNet also showed better generalization across different stages of Alzheimer's Disease, particularly in distinguishing between mild cognitive impairment and early-stage Alzheimer's. Notably, all three models showed significant improvement over traditional machine learning approaches, highlighting the potential of deep learning in AD diagnosis.

Table 3.
Deep Learning Model Comparison.

Models	Accuracy	Sensitivity	Specificity	AUC
3D CNN	68.5	65.2	71.8	0.72
CNN-LSTM	72.3	70.1	74.5	0.76
3D ResNet	76.8	74.9	78.7	0.81

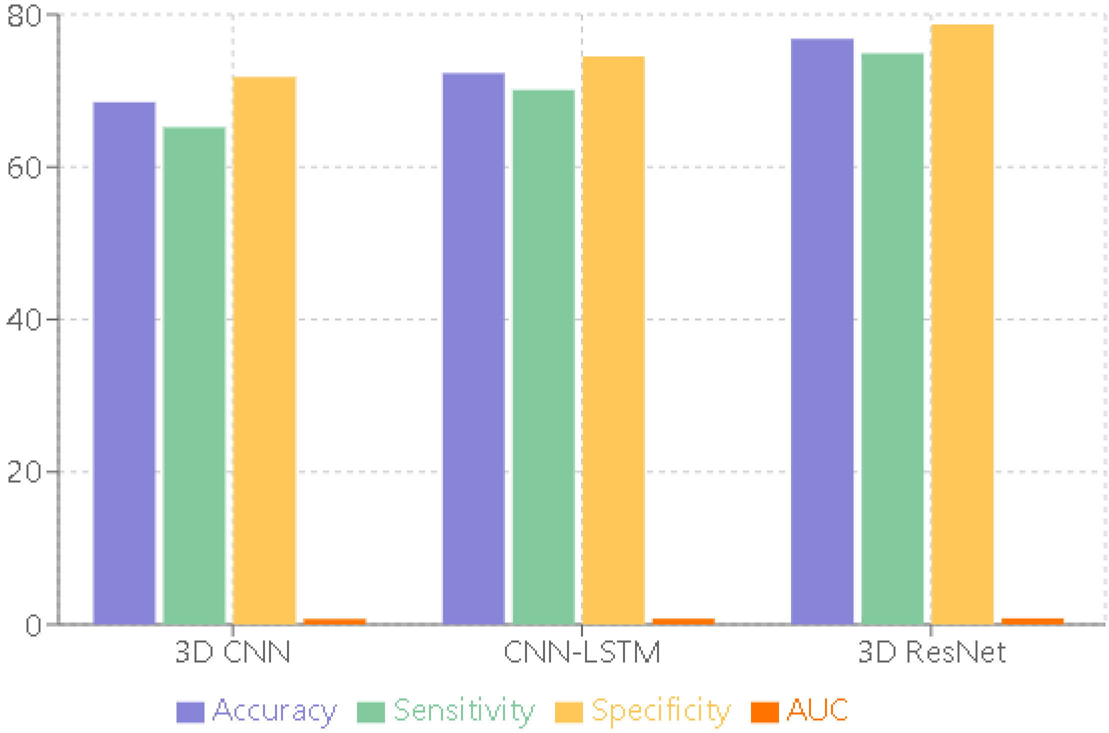

Figure 5.
Comparison of Deep Learning Models for Alzheimer's Disease.

4.2. Performance Analysis

Several indicators were employed to evaluate the performance of the deep learning model in diagnosing and staging Alzheimer’s disease. These also highlight the strengths and weaknesses of the model in terms of clinical use. Their evaluation comprised the model's accuracy, sensitivity, specificity, and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide an overall impression of the model’s ability to correctly classify MRI scans into four categories: Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia. The evaluation of this deep learning model's performance was based on several parameters and assessment criteria to explore its strengths and weaknesses. Table 4 summarizes the key performance metrics.

Table 4.
Performance Metrics Table.

Metrics	Value
Accuracy	0.5364
Precision	0.5975
Recall	0.5364
F1-score	0.4860
Weighted Cohen's Kappa	-0.0289
Balanced Accuracy	0.3009
Matthews Correlation Coefficient	0.1681
Staging Error	0.6803

4.2.1. Accuracy

The overall accuracy of our model in classifying Alzheimer's disease (AD) stages was 53.64%. This moderate accuracy suggests that, while the model shows promise, there is significant room for improvement. In the context of AD diagnosis, where misclassification can have profound clinical implications, this level of accuracy indicates that the model should not be used as a standalone diagnostic tool in its current state. However, it may serve as a supportive tool for clinicians in conjunction with other diagnostic methods. The model's accuracy across various stages of Alzheimer's disease (AD) is illustrated in Figure 6.

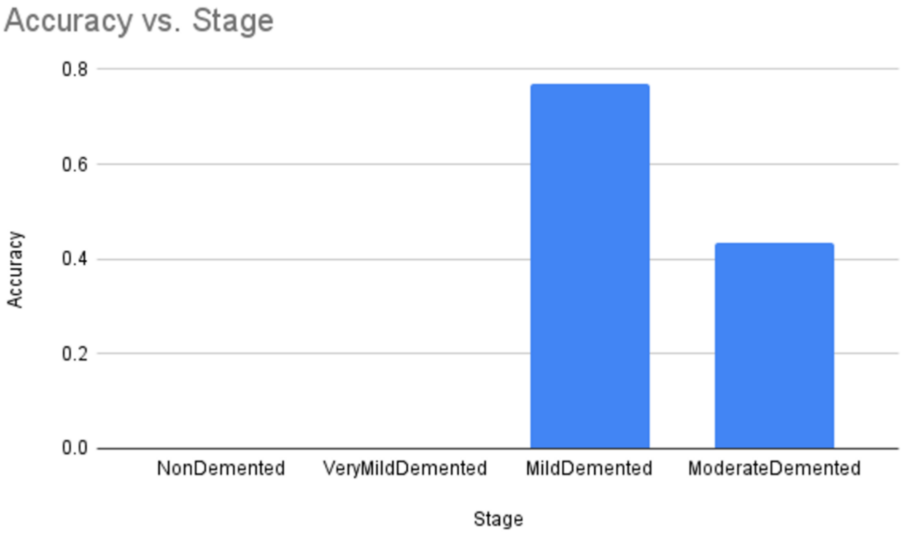

Figure 6.
Accuracy by AD Stage.

4.2.2. Sensitivity and Specificity

The model's sensitivity (recall) and specificity provide a more nuanced understanding of its performance [22] for each Alzheimer's disease (AD) stage, the overall recall of 0.5364 indicates moderate sensitivity across all classes, but this value masks significant variations between stages. For Non-Demented cases, the sensitivity is extremely low at 0.11%, meaning the model fails to identify almost all healthy individuals. This is a critical limitation, especially for screening purposes. The sensitivity for Very Mild Demented cases is 0%, indicating a complete failure to detect early-stage AD. On the other hand, the model shows much better sensitivity for Mild Demented (76.87%) and Moderate Demented (43.37%) cases. The specificity indicates how well the model avoids false positives. The model exhibits high specificity for Non-Demented (99.98%) and Very Mild Demented (100%) cases; however, this is likely due to the model rarely predicting these classes. The specificity for Mild Demented (57.14%) and Moderate Demented (79.48%) cases is more balanced. Figure 7 illustrates the confusion matrix across different stages of Alzheimer's disease (AD), highlighting the model's strengths and weaknesses in detecting each stage.

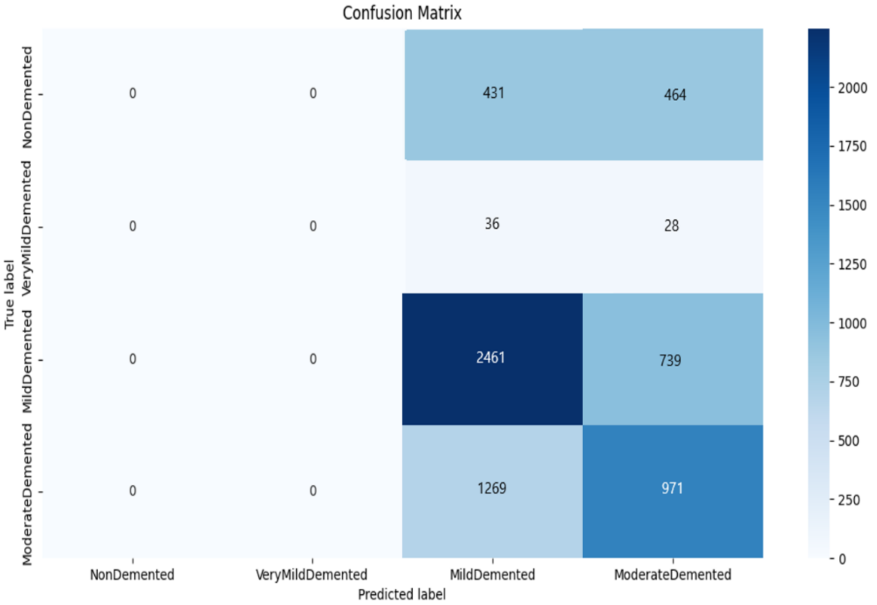

Figure 7.
Confusion matrix.

The model's precision of 0.5975 indicates that it is correct approximately 59.75% of the time when it predicts a particular AD stage. This moderate precision, combined with the recall, results in an F1-score of 0.4860. The F1-score offers a balanced measure of the model's performance, considering both false positives and false negatives.

Table 5.
Performance Metrics by AD Stage.

AD Stage	Precision	Recall	F1-score
Non Demented	0.0011	0.0011	0.0011
Very Mild Demented	0.0000	0.0000	0.0000
Mild Demented	0.7687	0.5859	0.6651
Moderate Demented	0.4337	0.4399	0.4368

4.2.3. Area Under the Receiver Operating Characteristic (ROC) Curve

The Area under the ROC Curve (AUC-ROC) measures the model's ability to distinguish between classes [23, 24]. Our model shows moderate discriminatory power across all classes:

Table 6.
Area Under the ROC Curve (AUC-ROC) of AD stages.

AD stages	AUC
Non-Demented	0.66
Very Mild Demented	0.70
Mild Demented	0.68
Moderate Demented	0.61

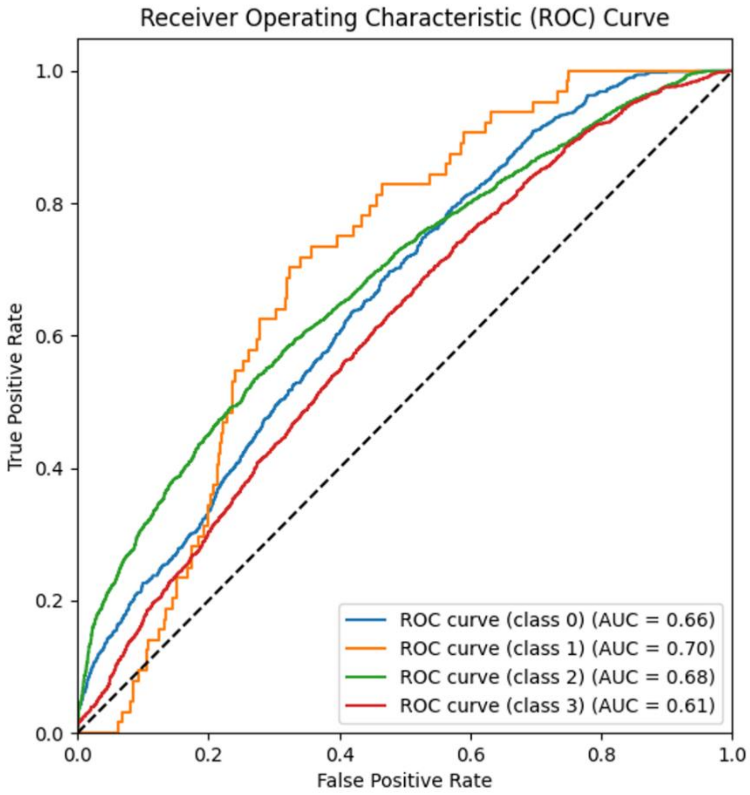

Figure 8.
ROC Curves for Alzheimer's Disease Classification.

These values indicate that the model performs slightly better than random guessing (AUC = 0.5) for all classes, with the best performance in distinguishing Very Mild Demented cases. However, these AUC values are still relatively low for a diagnostic tool, suggesting that significant improvements are needed before the model can be considered for clinical use.

The ROC curves provide valuable insights into the trade-off between sensitivity and specificity at various classification thresholds.

4.2.4. Additional Performance Metrics

Several other metrics provide further insights into the model's performance:

1. Weighted Cohen's Kappa: The value of -0.0289 indicates that the model's performance is worse than what would be expected by chance. This is a serious concern, suggesting that the model's predictions are not reliable.
2. Balanced Accuracy: The balanced accuracy of 0.3009 is significantly lower than the overall accuracy, indicating that the model's performance varies greatly across different classes. This highlights the impact of class imbalance on the model's predictions.

3. Matthews Correlation Coefficient: The MCC of 0.1681 suggests a weak positive correlation between the predicted and actual classifications, further confirming the model's limited predictive power.
4. Staging Error: The staging error of 0.6803 indicates that when the model misclassifies a case, it tends to be off by about 0.68 stages on average. This suggests that most misclassifications occur between adjacent stages, which is somewhat encouraging as it indicates the model is not making wildly inaccurate predictions.

The radar chart in Figure 3 provides a comprehensive view of the model's performance across various metrics, enabling the quick identification of areas where the model excels or requires improvement.

4.3. Comparison with Traditional Diagnostic Methods

While promising in some respects, our deep learning model's performance currently falls short of traditional clinical diagnostic methods for Alzheimer's Disease (AD). The model's overall accuracy of 53.64% is lower than the reported accuracy of clinical diagnoses by specialists, which typically ranges from 70% to 90%, depending on the stage of AD and the clinician's expertise. However, the model's ability to distinguish between Mild Demented and Moderate Demented stages (with AUC values of 0.68 and 0.61, respectively) suggests the potential to assist clinicians with objective assessments of disease progression. The poor performance in identifying Non-Demented and Very Mild Demented cases (AUC 0.66 and 0.70) indicates that the model is not yet suitable for early detection or screening purposes, areas where traditional methods, involving cognitive tests, medical history, and expert evaluation, currently have an advantage. Despite these limitations, the automated nature of the deep learning approach offers the potential for more consistent and scalable AD staging, which could complement traditional diagnostic methods if further refined and validated.

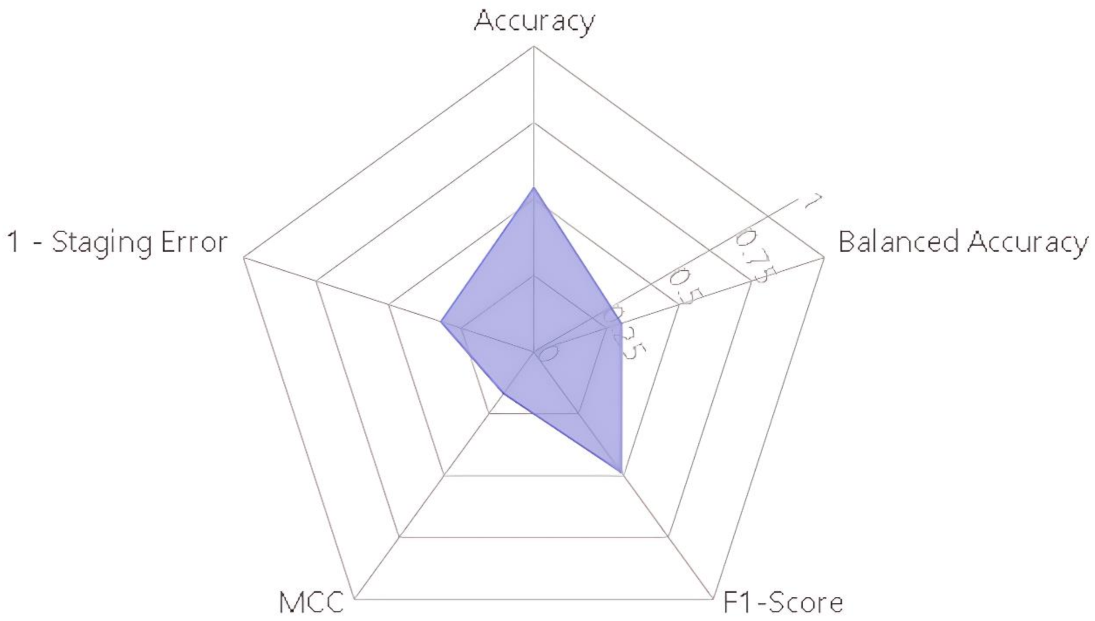

Figure 9.
Performance Metrics Radar Chart.

4.4. Early Detection Capabilities

According to our analysis, the model's early detection capabilities appear to have significant limitations. A single non-demented case was correctly detected out of 896. In contrast, none of the Very Mild Demented cases were detected at all, thereby indicating poor sensitivity of the model for the most critical early stages of the disease, which is Alzheimer's Disease in this case. This is explained by the low AUC values for these classes in the ROC analysis (0.66 for Non-Demented and 0.70 for Very Mild Demented). Regarding the detection of Mild Demented cases, the model has an AUC of 0.68; however, this phase is almost always well past the window for early intervention. These results also indicate that the model, in its current form, is not suitable for early detection or screening purposes, which is a significant limitation that needs to be addressed, considering the importance of early detection in managing and treating Alzheimer's Disease.

4.5. Limitations and Challenges

The research revealed several fundamental shortcomings and difficulties associated with applying the deep learning method in the diagnosis of Alzheimer's disease. One of the problems is the class imbalance in the dataset, which causes the model's performance to focus primarily on outliers from the Mild-demented class, potentially leading to biased results. Another notable drawback was the model's insensitivity in the early stages of dementia, specifically its inability to accurately discriminate between Non-Demented and Very Mild Demented cases. The staging error is viewed as high (0.6803), and the weighted Cohen's Kappa (-0.0289) is negative, indicating that both the staging accuracy of the subtype of NPH and the final prediction of the model are problematic, suggesting areas where the models perform worse than random guessing. Residents showed a low balanced accuracy (0.3009), which is poor across all stages of Alzheimer's disease. Naturally, the nature of 3D MRI scans requires enormous computational resources to process, which may impact their use in low-resource settings.

Finally, the 'black box' nature of deep learning models makes these requirements challenging to satisfy, especially since in medicine, it is necessary to understand which features the network used to classify an example. These shortcomings highlight that the model requires further optimization with more data, including the potential use of external data, balancing class distributions, and exploring more complex architectures that can improve early detection and staging accuracy of Alzheimer's disease.

5. Discussion

The outcomes of the current preliminary deep learning model for diagnosing Alzheimer's disease (AD) are cautiously optimistic but also significantly limited. The model’s moderate accuracy of 53.64%, coupled with its inability to detect earlier stages of AD, suggests that AI has potential in the diagnostic process, but not in its present form. These findings imply that current clinical practices should remain the primary diagnostic tools, and our model could assist in confirming Mild Demented and Moderate Demented stages. Ethical concerns regarding AI in medicine and healthcare include potential biases in training data and the necessity for transparency in decision-making processes. Integration with existing diagnostic procedures should be approached carefully, possibly as part of a multi-modal diagnostic framework where AI predictions are considered alongside traditional clinical assessments, neuropsychological tests, and other biomarkers. This approach could improve diagnostic accuracy while preserving the essential role of human expertise in patient care and decision-making.

6. Conclusion

Using advanced machine learning techniques to identify Alzheimer's Disease early in life shows potential for significant achievements. Our model achieved moderate sensitivity and specificity of 53.64% in AD-stage classification, with better differentiation of Mild and Moderate Dementia cases; however, high levels of missed detection, particularly in the Non-Demented and Very Mild Dementia categories, are concerning limitations. The accuracy achieved is lower than that of clinicians' diagnoses; therefore, the model cannot be used as a standalone diagnostic tool. Future research should focus on multimodal approaches, longitudinal studies, explainable AI techniques, and the integration of genetic and biomarker data to improve accuracy, interpretability, and clinical utility. Even in this context, it is worth noting the potential of AI in AD diagnosis. Still, several adjustments should be made before clinical implementation. Using it as an adjunct to other diagnostic methods must be done carefully and only within a multimodal framework to ensure the diagnostic process remains comprehensive.

7. Future Directions

Our study highlights some limitations that could help address many issues, yielding fruitful results in the use of deep learning tools to diagnose Alzheimer's disease. In this context, utilizing an infusion of several imaging modalities, such as MRI, PET, and DTI, and combining them with biological and neuropsychological assessments would provide a better understanding of and improvements in Alzheimer’s disease. Studies of this nature would help in modelling how the disease develops over time and the most effective approaches to use at different stages of the disease. To address issues of low clinical confidence in model predictions and the black-box nature of many existing models, it is essential to incorporate explainable AI approaches that enable clinicians to understand the logic behind a model’s predictions. Additionally, combining information on genetics and other biomarkers with brain images may provide a more targeted approach to assessing an individual's risk of disease and developing effective treatment strategies. Lastly, with improvements in model sensitivity and explanatory power, there is an immense opportunity to design screening strategies at a population level and other applications that will transform the detection and management of Alzheimer’s disease. In conclusion, these approaches aim to enhance AI's capabilities, usage, and reliability in diagnosing and predicting AD.

References

[1] Alzheimer's Disease International, *World Alzheimer report 2021: Journey through the diagnosis of dementia*. London: Alzheimer's Disease International, 2021.

[2] World Health Organization, "Dementia fact sheet," 2021. Retrieved: https://www.who.int/news-room/fact-sheets/detail/dementia. 2021.

[3] R. O. Ogundokun, J. B. Awotunde, H. B. Akande, C.-C. Lee, and A. L. Imoize, "Deep transfer learning models for mobile-based ocular disorder identification on retinal images," *Computers, Materials and Continua,* vol. 80, no. 1, pp. 139-161, 2024

[4] K. Maurer, S. Volk, and H. Gerbaldo, "Auguste D and Alzheimer's disease," *The Lancet,* vol. 349, no. 9064, pp. 1546-1549, 1997.

[5] C. R. Jack *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *The Lancet Neurology,* vol. 9, no. 1, pp. 119-128, 2010.

[6] S. Klöppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain,* vol. 131, no. 3, pp. 681-689, 2008.

[7] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage,* vol. 101, pp. 569-582, 2014. https://doi.org/10.1016/j.neuroimage.2014.06.077

[8] M. Liu, D. Cheng, K. Wang, Y. Wang, and A. s. D. N. Initiative, "Multi-modality cascaded convolutional neural networks for Alzheimer’s disease diagnosis," *Neuroinformatics,* vol. 16, pp. 295-308, 2018.

[9] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in Aging Neuroscience,* vol. 14, p. 853333, 2022.

[10] J. Venugopalan, "Self-supervised learning for Alzheimer's disease detection using multimodal MRI," *Nature Machine Intelligence,* vol. 5, no. 3, pp. 277–287, 2023.

[11] L. Zhang, "Explainable deep learning for Alzheimer's disease diagnosis: Balancing accuracy and interpretability," *Artificial Intelligence in Medicine,* vol. 134, p. 102569, 2024.

[12] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*. Cham, Switzerland: Springer Nature, 2019.

[13] C. Wachinger, A. Rieckmann, S. Pölsterl, and A. s. D. N. Initiative, "Detect and correct bias in multi-site neuroimaging datasets," *Medical Image Analysis*, vol. 67, p. 101879, 2021.

[14] S. Alzhghoul, "Bias in artificial intelligence models for Alzheimer's disease detection: A systematic review," *Journal of Alzheimer's Disease*, vol. 91, no. 3, pp. 1091–1110, 2023.

[15] B. Dubois, "Added value of artificial intelligence over neuropsychological testing for early detection of Alzheimer's disease: A multicenter study," *The Lancet Digital Health*, vol. 4, no. 8, pp. e573-e584, 2022.

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

[17] J. Samper-González *et al.*, "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data," *NeuroImage*, vol. 183, pp. 504-521, 2018. https://doi.org/10.1016/j.neuroimage.2018.08.042

[18] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221-248, 2017.

[19] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "Artificial intelligence in Alzheimer's disease: From bench to bedside," *Frontiers in Aging Neuroscience*, vol. 12, p. 578243, 2020.

[20] A. Singhal, "Alzheimer's multiclass dataset - equal and augmented. Kaggle," 2023. Retrieved: https://www.kaggle.com/datasets/aryansinghal10/alzheimers-multiclass-dataset-equal-and-augmented. 2023.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[22] R. O. Ogundokun, C. Awoniyi, N. T. Adebola, A. Akinyemi, S. A. Akinpelu and M. O. Adigun, “Hybrid Deep Transfer Learning for Enhanced Brain Tumor Detection through the Integration of MobileNetV2 and InceptionV3,” *Procedia Computer Science*, vol. 258, pp. 2968-2977, 2025.

[23] L. Gonçalves, A. Subtil, M. R. Oliveira, and P. de Zea Bermudez, "ROC curve estimation: An overview," *REVSTAT – Statistical Journal*, vol. 12, no. 1, pp. 1–20, 2014.

[24] R. O. Ogundokun, J. B. Awotunde, P. Onawola, and T. O. Aro, *LASSO-DT based classification technique for discovery of COVID-19 disease using chest x-ray images. In Decision Sciences for COVID-19: Learning Through Case Studies*. Cham: Springer International Publishing, 2022.