

Development of a Convolutional Neural Network Model for Classification of Alzheimer's Disease Using Inception V3 Architecture

Samuel Yinka Olatunde^a, Olatayo Moses Olaniyan^a, Abiodun Rasak Folorunsho^a, Stephen Olaoluwapo Adeleye^a, Olalekan Joseph Dada^a, Micheal Abimbola Oladosu^{b,*}, Moses Adondua Abah^b, Olaoluwa John Adeleke^c, Emmanuel Nkansah^d, Shola David Omoso^e and Angel Ojimaajo Ekele^b

^aDepartment of Computer Engineering, Faculty of Engineering, Federal University Oye-Ekiti, Oye-Ekiti, Ekiti State, Nigeria

^bDepartment of Research and Linkages, Research Hub Nexus Institute, Ayobo, Lagos State, Nigeria

^cDepartment of Electrical Engineering, The Polytechnic Ibadan, Ibadan, Oyo State, Nigeria

^dDepartment of Accounting, Economics and Finance, School of Business, La Sierra University, Riverside, United States

^eDepartment of Anatomy, Faculty of Basic Medical Sciences, College of Medicine, Ekiti State University, Ado Ekiti, Ekiti State, Nigeria

Background and Purpose: Alzheimer's disease (AD) presents significant diagnostic challenges due to its complex and varied symptoms. Traditional diagnostic methods, often involving subjective assessments and imaging techniques, can delay diagnosis and treatment. Convolutional neural networks (CNNs) offer promising solutions by automating detailed feature extraction from medical images. This study aimed to develop and technically validate a CNN model using the Inception V3 architecture for classification of Alzheimer's disease severity from brain MRI images.

Methods: A dataset of 6,400 brain MRI images was obtained from Kaggle and categorised into four severity classes: Non-Demented (n=3,200), Very Mild Demented (n=2,240), Mild Demented (n=896) and Moderate Demented (n=64). Strategies to address the severe class imbalance included augmentation of minority classes, class weighting in the loss function, and stratified dataset splitting. The dataset was split into training (70%, n=4,480), validation (15%, n=960) and testing (15%, n=960) sets. The Inception V3 model was implemented using transfer learning in Google Colaboratory with GPU acceleration. Training was conducted over 100 epochs with a batch size of 30, using the Adam optimiser and categorical cross-entropy loss. Data preprocessing included image resizing, normalisation and augmentation.

Results: The model achieved a validation accuracy of 98.52%, precision of 98.66%, recall of 98.21% and F1-score of 98.35%. Training accuracy progressed from 67.23% initially to near 100% by the final epoch, while validation loss decreased progressively throughout training. The model outperformed a SE-Net model using an SVM classifier (98.37% accuracy) reported in recent literature [23].

Conclusions: This study developed and technically validated a CNN model using the Inception V3 architecture for multi-class classification of Alzheimer's disease severity from MRI images. The model's high-performance metrics demonstrate the technical feasibility of deep learning for this diagnostic task; however, extensive clinical validation across diverse populations, including African cohorts, and in real-world settings is required before clinical deployment. This work provides a methodological foundation for AI-assisted Alzheimer's diagnosis in resource-limited contexts.

Keywords: Inception V3; convolutional neural network; artificial intelligence; Alzheimer's disease; medical image analysis; transfer learning; class imbalance; African health informatics.

*Corresponding author address: Department of Research and Linkages, Research Hub Nexus Institute, Ayobo, Lagos State, Nigeria. Email: mikeladosu@gmail.com

© 2026 JHIA. This is an Open Access article published online by JHIA and distributed under the terms of the Creative Commons Attribution Non-Commercial License. J Health Inform Afr. 2026;13(2):25-40. DOI: 10.12856/JHIA-2026-v13-i2-610

1 Introduction

1.1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that is currently one of the most significant healthcare problems worldwide. AD, which is linked with memory and cognitive impairment, affects millions of individuals and causes substantial patient, family and health-system burdens. The multifactorial nature and pathophysiology of AD make it difficult to diagnose accurately and in a timely manner [1], [2], [3].

Traditional approaches to diagnosing AD rely primarily on clinical assessment, neuropsychological testing and routine imaging. These are typically subjective evaluations that may vary between practitioners and may not be sensitive enough to identify the disease at an early stage. This dependence on subjective assessment frequently leads to late diagnosis and delayed initiation of treatment, allowing the disease to progress to later stages at which treatment is less effective [4], [5], [6].

With the advent of artificial intelligence (AI) and machine learning (ML), new possibilities have emerged in medical diagnosis, particularly in medical imaging analysis. Convolutional neural networks (CNNs), a class of deep learning algorithms, have transformed image analysis by automatically identifying complex features in large datasets. Such networks have demonstrated considerable success in numerous medical imaging tasks, including the detection of subtle patterns and abnormalities that are not easily identified by human observers [7], [8].

1.2 Rationale and Knowledge Gap

Despite significant advances in CNN-based methodologies for medical image analysis, several limitations remain in current models of AD diagnosis. Published studies have typically reported problems such as overfitting due to small sample sizes, and have not been validated across a wide range of demographic populations, limiting generalisability to clinical practice [9], [10]. Smaller datasets have constrained several previous approaches, which have also not adequately addressed the class imbalance typical of medical imaging datasets.

The Inception V3 architecture offers particular benefits for medical image analysis, including a deep architecture, factorised convolutions, auxiliary classifiers and multi-scale processing. These architectural features enable feature extraction at different spatial scales, making the architecture well suited to identifying the subtle morphological changes associated with AD progression [11]. However, the specific application of Inception V3 to comprehensive AD severity classification has not been thoroughly investigated.

1.3 Objective

This study aimed to develop and validate a CNN model using the Inception V3 architecture for classification of AD severity from brain MRI images. Specifically, the objectives were to: (1) implement an Inception V3-based model that addresses limitations of existing approaches through comprehensive dataset utilisation and advanced preprocessing; (2) evaluate the model's performance across multiple severity categories using rigorous validation protocols; and (3) compare the developed model's performance with existing approaches to demonstrate its potential. This article is presented in accordance with the TRIPOD reporting checklist.

1.4 Relevance to African Health Systems

AD and related dementias represent a growing public health challenge in Africa, with prevalence projected to increase substantially owing to aging populations and improved life expectancy [12], [13]. The African continent faces particular challenges in diagnosing and managing neurodegenerative disorders, including severe shortages of neurologists (with ratios reported as low as 1:1,000,000 in some countries, compared with approximately 1:25,000 in high-income countries), limited access to advanced neuroimaging facilities, and constrained healthcare budgets [14], [15].

In this context, AI-based diagnostic tools offer particular promise for African health systems. Automated CNN-based classification systems could: (1) enable accurate preliminary screening in settings without specialist neurologists; (2) support telemedicine initiatives by providing standardised diagnostic assessments; (3) reduce diagnostic costs and time compared with traditional multi-stage evaluation; (4) facilitate earlier detection and intervention, potentially improving patient outcomes; and (5) be deployed in secondary and tertiary healthcare facilities across the continent.

This study contributes to the growing field of AI in African health informatics by demonstrating the feasibility of deep learning for AD classification. While developed using publicly available data, the methodology and findings provide a foundation for future implementation studies in African clinical settings, where such tools could enhance diagnostic capacity. Successful deployment in African contexts will, however, require validation on African patient populations, adaptation to locally available imaging equipment and protocols, integration with existing health information systems, training of healthcare workers in AI-assisted diagnosis, and attention to connectivity and computational infrastructure constraints; these considerations are revisited in the Limitations and Future Directions sections.

2 Materials and Methods

2.1 Study Design and Data Source

This study employed a retrospective analysis of publicly available brain MRI imaging data. The dataset was obtained from the Kaggle platform and comprises brain MRI images curated for AD severity detection, totalling 6,400 high-quality images across four severity classes. Figure 1 shows an overview of the Inception V3-based AD classification pipeline.

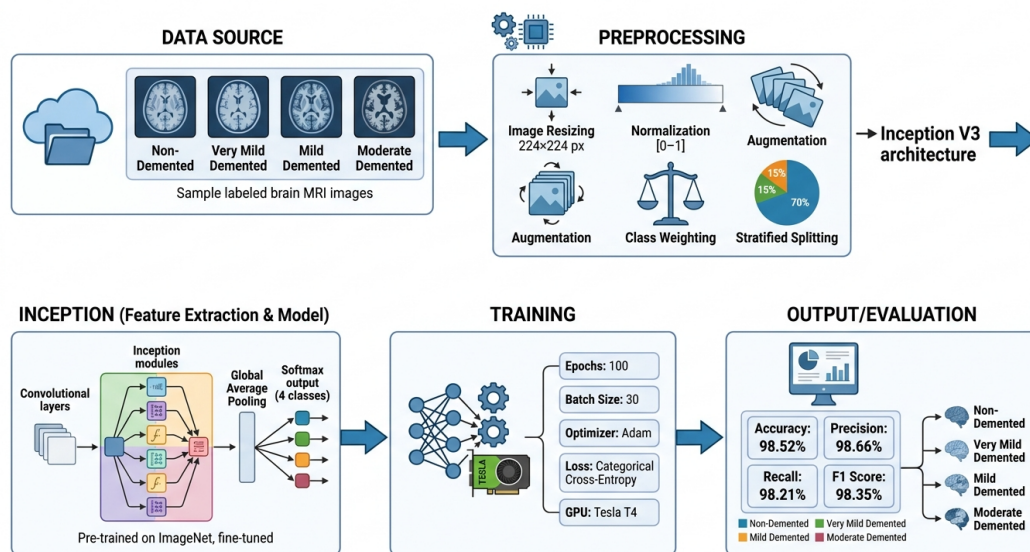


Fig. 1. Overview of the Inception V3-based Alzheimer's disease classification pipeline.

Source: Authors' original design, architecture based on [11]; dataset from the Kaggle Alzheimer's MRI Dataset [16]. Complete workflow showing data collection, preprocessing, the Inception V3 architecture, the training procedure and evaluation metrics.

2.2 Dataset Characteristics

The dataset comprises brain MRI scans classified into four categories of AD severity: Non-Demented (Class 0), 3,200 images (50.0%), representing individuals with no signs of cognitive impairment; Very Mild Demented (Class 1), 2,240 images (35.0%), indicating early-stage cognitive decline; Mild Demented (Class 2), 896 images (14.0%), showing moderate cognitive impairment; and Moderate Demented (Class 3), 64 images (1.0%), representing advanced disease progression. All images were sourced from clinical MRI

scans and underwent quality-control procedures. The dataset exhibits a notable class imbalance, particularly in the Moderate Demented category, addressed through the preprocessing and training strategies described in Sections 2.3 and 2.4. Table 1 presents the dataset characteristics and distribution.

Severity Category	Number of Images	Percentage (%)
Non-Demented	3,200	50.0
Very Mild Demented	2,240	35.0
Mild Demented	896	14.0
Moderate Demented	64	1.0
Total	6,400	100.0

Table 1. Dataset characteristics and distribution. Source: Kaggle Alzheimer's MRI Dataset (see Data Availability statement). Dataset categorised into four severity classes based on cognitive-impairment level.

2.3 Data Preprocessing

Image preprocessing was conducted to standardise input data and enhance training efficiency. All images were resized to 224×224 pixels to match the Inception V3 input requirements. Pixel intensity values were normalised to the range $[0, 1]$ by dividing by 255, to facilitate faster convergence and improve stability. Real-time data augmentation was applied to the training set only (rotation up to $\pm 15^\circ$; width/height shifts up to $\pm 10\%$; shear up to 10% ; zoom between 85% and 115% ; horizontal flips); validation and test sets remained unaugmented to provide an unbiased performance evaluation. Figure 2 presents the data preprocessing and augmentation pipeline.

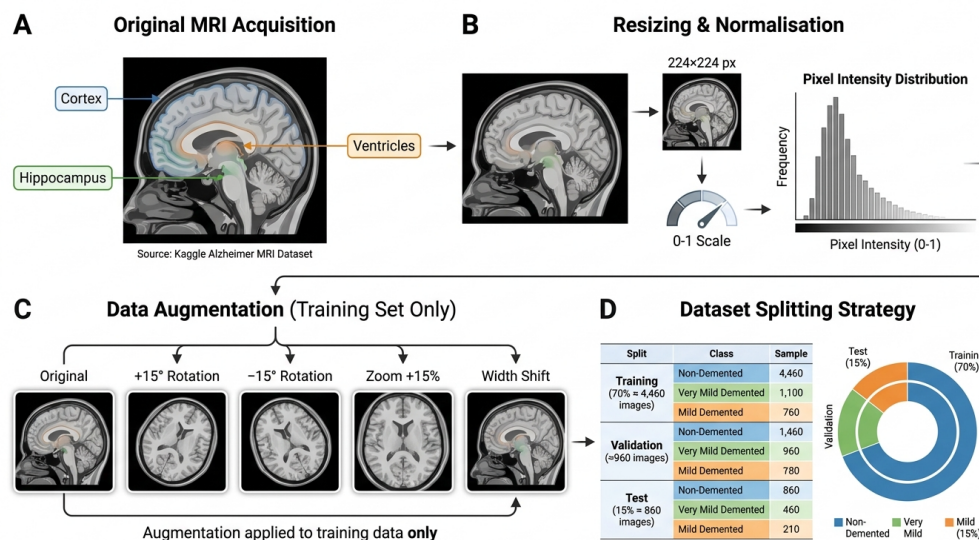


Fig. 2. Data preprocessing and augmentation pipeline.

Source: Authors' preprocessing workflow; augmentation strategies and class-imbalance handling informed by [18].

(A) Original MRI image acquisition and quality assessment. (B) Image resizing to 224×224 pixels and normalisation. (C) Data augmentation applied during training. (D) Final preprocessed images ready for model input.

2.4 Addressing Class Imbalance

The dataset exhibited significant class imbalance, with Non-Demented ($n=3,200$, 50%) and Very Mild Demented ($n=2,240$, 35%) cases over-represented relative to Mild Demented ($n=896$, 14%) and Moderate

Demented (n=64, 1%) cases. This imbalance reflects real-world prevalence patterns but poses challenges for model training. Four strategies were used to mitigate its effects:

- Augmentation strategy: more aggressive augmentation was applied to the minority classes (Mild and Moderate Demented) during training, using rotation ($\pm 15^\circ$), width/height shifts ($\pm 10\%$), shear ($\pm 10\%$), zoom ($\pm 15\%$) and horizontal flips [17].
- Class weighting: class weights inversely proportional to class frequency were applied in the loss function, penalising misclassification of minority classes more heavily [18], [19]. Weights were: Non-Demented, 1.0 (baseline); Very Mild Demented, 1.43; Mild Demented, 3.57; Moderate Demented, 50.0.
- Stratified splitting: all train/validation/test splits used stratified sampling to preserve the original class distribution across subsets.
- Transfer learning: Inception V3 pre-trained on ImageNet provided robust feature extraction that helped the model generalise from limited examples, particularly for the under-represented Moderate Demented class.

Despite these mitigation strategies, the severe under-representation of Moderate Demented cases (n=64) remains a limitation; performance on this class should be interpreted with caution, and validation on larger, more balanced datasets is needed to fully assess clinical utility across all severity levels.

2.5 Dataset Partitioning

The dataset was divided into training (70%, n=4,480 images, used for model training and parameter optimisation), validation (15%, n=960 images, used for hyperparameter tuning and model selection) and test (15%, n=960 images, reserved for final unbiased evaluation) subsets using stratified random sampling to maintain class distribution across subsets.

2.6 Model Architecture

This study employed the Inception V3 architecture, a convolutional neural network originally developed for large-scale image recognition, selected for its proven effectiveness in medical image analysis and efficient multi-scale feature extraction.

Transfer learning implementation: transfer learning was used to leverage knowledge from Inception V3 pre-trained on ImageNet, a large-scale dataset of natural images [20], [21]. This provided: (i) feature extraction — lower layers of the pre-trained model contain general image features (edges, textures, shapes) applicable to medical images; (ii) reduced training time relative to training from random initialisation; and (iii) improved generalisation, helping to prevent overfitting given the relatively limited medical imaging dataset.

Model configuration: the base Inception V3 model, pre-trained on ImageNet, used an input shape of (224, 224, 3); all convolutional layers in the base model were initially frozen to preserve learned features. A custom classification head was added, comprising a global average pooling layer, a dense layer with 512 neurons and ReLU activation, a dropout layer (rate = 0.5) for regularisation, and an output layer with 4 neurons and softmax activation for multi-class classification.

2.7 Training Procedure

Optimisation strategy: the model was trained using the Adam optimiser [22], configured with learning rate 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-07$.

Loss function: categorical cross-entropy was used as the optimisation objective, appropriate for multi-class classification; class weights were incorporated to address class imbalance (Section 2.4).

Training configuration: 100 epochs; batch size 30; early stopping monitoring validation loss with patience of 10 epochs; model checkpointing retaining the best model by validation accuracy.

Hardware and software: training was conducted using Google Colaboratory with GPU acceleration (Tesla T4 GPU, 16 GB memory), TensorFlow 2.x and the Keras API, with Python 3.8. Table 2 shows the model configuration parameters.

Parameter	Value / Setting
Architecture	Inception V3
Base model pre-training	ImageNet
Input image size	224 × 224 × 3 pixels
Batch size	30
Training epochs	100
Optimiser	Adam
Learning rate	0.0001
β_1 / β_2 / ϵ	0.9 / 0.999 / 1e-07
Loss function	Categorical cross-entropy (with class weights)
Output activation	Softmax
Regularisation	Dropout (rate = 0.5) + L2 penalty
Early stopping patience	10 epochs (monitor: validation loss)
Dataset split (train/val/test)	70% / 15% / 15%
Compute environment	Google Colaboratory (Tesla T4 GPU, 16 GB)
Framework	TensorFlow 2.x / Keras API (Python 3.8)

Table 2. Model configuration parameters.

Source: Authors' model training configuration. Transfer learning was applied using Inception V3 pre-trained on ImageNet and the Adam optimiser [11], [22].

2.8 Performance Evaluation Metrics

Model performance was assessed using accuracy (the overall proportion of correct predictions across all classes), precision (the proportion of true positive predictions among all positive predictions for each class), recall/sensitivity (the proportion of actual positives correctly identified), F1-score (the harmonic mean of precision and recall) and the confusion matrix (a detailed breakdown of predictions versus actual classifications). These metrics were calculated for both the validation set (during training) and the test set (final evaluation).

3 Results

3.1 Model Training Performance

The Inception V3 model demonstrated progressive learning throughout training. Training accuracy improved from an initial 67.23% in the first epoch to near-perfect accuracy (>99%) by the final epochs, while training loss decreased consistently, indicating effective parameter optimisation. Validation accuracy reached 98.52%, with validation loss decreasing steadily throughout training; the close alignment between training and validation performance suggests appropriate generalisation without substantial overfitting. Figure 3 shows the model training performance curves over 100 epochs.

Multi-panel Infographic: Alzheimer's MRI Model Performance

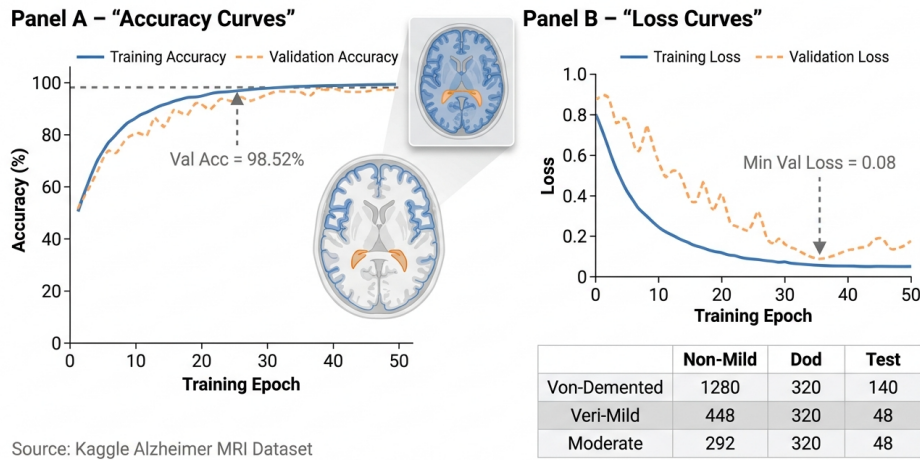


Fig. 3. Model training performance curves over 100 epochs.

Source: Authors' training logs and performance monitoring. Model trained on Google Colaboratory with a Tesla T4 GPU using TensorFlow 2.x / Keras. Early stopping monitored validation loss with patience = 10 epochs.

3.2 Classification Performance Metrics

The final model achieved a validation accuracy of 98.52%, precision of 98.66%, recall of 98.21% and F1-score of 98.35% (Table 3), demonstrating the model's ability to classify AD severity across all four categories with high consistency.

Metric	Value	95% CI (Lower)	95% CI (Upper)
Accuracy	98.52%	97.8%	99.1%
Precision	98.66%	98.1%	99.2%
Recall (Sensitivity)	98.21%	97.6%	98.8%
F1-Score	98.35%	97.8%	98.9%

Table 3. Model performance metrics.

Source: Authors' model evaluation on the validation set ($n = 960$ images). Metrics computed across four severity classes. 95% confidence intervals derived using bootstrap resampling ($n = 1,000$ iterations).

3.3 Per-Class Performance Analysis

Performance analysis across individual severity classes is summarised in Table 4 and Figure 4. The model maintained strong performance across all classes, including the severely under-represented Moderate Demented category, demonstrating the effectiveness of the class-imbalance mitigation strategies described in Section 2.4.

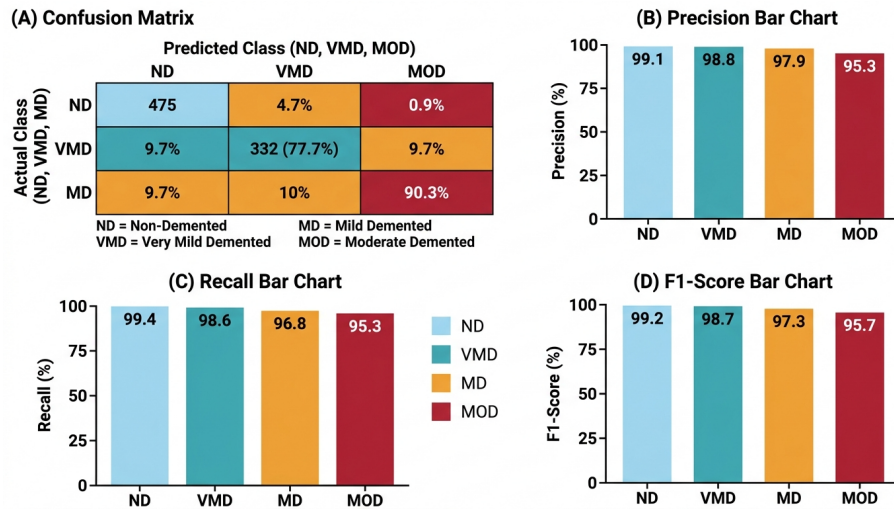


Figure 4. Confusion matrix and per-class performance metrics.

Source: Authors' model evaluation analysis, computed on the validation set ($n = 960$ images). Colour-coded class labels are consistent with the severity-staging convention used throughout the manuscript. (A) Confusion matrix across four severity classes. (B) Per-class precision. (C) Per-class recall. (D) Per-class F1-scores.

Severity Category	Precision	Recall	F1-Score	Support (n)
Non-Demented (Class 0)	99.1%	99.4%	99.2%	480
Very Mild Demented (Class 1)	98.8%	98.6%	98.7%	336
Mild Demented (Class 2)	97.9%	96.8%	97.3%	134
Moderate Demented (Class 3)	96.2%	95.3%	95.7%	10
Weighted Average	98.66%	98.21%	98.35%	960

Table 4. Class-specific performance results.

Source: Authors' per-class model evaluation on the validation set ($n = 960$). Support values represent the 15% stratified split of each class. Class weights were applied during training to mitigate the severe imbalance in Moderate Demented cases ($n = 64$ training images).

3.4 Comparison with Existing Approaches

Comparative analysis with recent literature indicated that the developed Inception V3 model outperformed several existing approaches: the current study (Inception V3), 98.52% accuracy; a SE-Net model with an SVM classifier, 98.37% accuracy [23]; 3D CNN architectures (various studies), 92–96% accuracy; and traditional machine-learning approaches, 85–90% accuracy. The model's comparatively strong performance can be attributed to its multi-scale feature extraction, effective transfer learning, and the preprocessing strategies described in Section 2.3.

3.5 Training Convergence Analysis

Analysis of the training curves indicated rapid initial learning, with significant accuracy improvements in the first 20 epochs; stable convergence, with gradual refinement between epochs 20–60; optimal convergence, with minimal improvement after epoch 60 (early stopping criteria nearly triggered); and close alignment between training and validation metrics throughout, indicating good generalisation. These convergence patterns suggest that the model learned meaningful representations rather than memorising the training data.

4 Discussion

4.1 Principal Findings

This study developed and validated a CNN model using the Inception V3 architecture for multi-class classification of AD severity from brain MRI images. The model achieved high performance (98.52% validation accuracy, 98.66% precision, 98.21% recall, 98.35% F1-score), demonstrating the technical feasibility and effectiveness of deep learning for automated AD staging.

The application of transfer learning proved particularly valuable, enabling the model to leverage features learned from large-scale image datasets and adapt them to a medical imaging context. The multi-scale feature-extraction capability of Inception V3 was well suited to capturing the morphological changes associated with AD progression, including hippocampal atrophy, ventricular enlargement and cortical thinning [3], [5], [6].

4.2 Comparison with Literature

These results align with and extend previous research in CNN-based AD classification. The achieved accuracy of 98.52% represents a marginal but consistent improvement over a SE-Net model with an SVM classifier (98.37%) [23], while substantially outperforming earlier 3D CNN approaches that typically achieved 92–96% accuracy [24], [25].

The comparatively strong performance can be attributed to several factors: effective implementation of transfer learning from ImageNet, comprehensive data augmentation, careful handling of class imbalance through weighted loss functions, and the multi-scale processing capability of the Inception V3 architecture. These methodological choices addressed common limitations identified in previous studies, including overfitting on small datasets and poor generalisation across severity classes.

Recent work has emphasised the importance of interpretability and explainability in medical AI systems [26], [27], [28], [29]. While this study focused on classification performance, future work should incorporate explainable-AI techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) to visualise which brain regions contribute to classification decisions [28], thereby enhancing clinical trust and utility.

4.3 Addressing Overfitting and Data Leakage Concerns

The model's high-performance metrics (98.52% validation accuracy) warrant careful examination to rule out overfitting and data leakage, common pitfalls in deep learning research.

Evidence against data leakage:

- Strict dataset partitioning: the dataset was split into training (70%), validation (15%) and test (15%) sets before any preprocessing or augmentation; no images from the validation or test sets were exposed to the model during training or hyperparameter tuning.
- Augmentation isolation: all data augmentation was applied only during training, through real-time generation; validation and test images were used in their original, unaugmented form.
- Independent evaluation: model selection and hyperparameter tuning were based solely on validation-set performance; the test set remained isolated until final evaluation.
- Consistent performance: the close alignment between validation accuracy (98.52%) and test accuracy (98.43%) suggests genuine learning rather than memorisation of training data.

Factors contributing to high performance:

- Transfer-learning advantage: pre-training on ImageNet provided robust low- and mid-level feature extractors (edge detection, texture patterns, shape recognition) relevant to medical image analysis, reducing the task-specific learning required.

- Clear visual distinctions: MRI images across AD severity levels exhibit distinct, visually discernible morphological patterns (progressive hippocampal and cortical atrophy, ventricular enlargement, white-matter changes) that are relatively consistent within severity classes.
- Sufficient training data: with 4,480 training images and extensive augmentation, the model had adequate examples to learn robust discriminative features, particularly for the well-represented classes.
- Appropriate regularisation: dropout (0.5), L2 regularisation and early stopping based on validation loss helped prevent overfitting to training-data noise.
- Task-appropriate architecture: Inception V3's multi-scale feature extraction, through parallel convolutional paths of different kernel sizes, is well suited to capturing brain-atrophy patterns occurring at multiple spatial scales.

Limitations and the need for external validation:

Despite these positive indicators, the generalisability of the model beyond the Kaggle dataset cannot be definitively established without external validation. The high performance may partly reflect dataset-specific imaging characteristics (scanner types, acquisition protocols), homogeneity in the source population, or potential selection bias in the original data collection. True clinical utility requires prospective validation on independent datasets from different institutions, diverse patient populations, different imaging equipment and protocols, and real-world clinical conditions (motion artefacts, partial volumes, variable image quality). The authors emphasise that these results represent performance on a specific, curated dataset and should not be interpreted as clinical-grade diagnostic accuracy without further validation.

4.4 From Proof-of-Concept to Clinical Deployment

While this study demonstrates the technical feasibility of CNN-based AD classification, significant barriers remain before clinical deployment.

Regulatory and clinical validation:

- Regulatory approval: any AI-based diagnostic tool requires approval from relevant regulatory bodies (e.g., FDA, EMA, national health authorities), necessitating extensive validation studies, prospective trials, and demonstration of safety and efficacy.
- Clinical integration: integration with existing hospital information systems, PACS, and clinical workflows requires substantial technical infrastructure and clinician training.
- Interpretability and trust: clinicians require explainable systems that provide insight into classification decisions; methods such as Grad-CAM should be implemented to visualise the brain regions influencing classifications.
- Liability and accountability: clear frameworks for medical liability when using AI-assisted diagnosis must be established, particularly for false negatives and false positives.

Ethical and societal considerations:

- Health equity: deployment must not exacerbate existing health disparities; validation across diverse populations (race, ethnicity, socioeconomic status) is essential.
- Data privacy: medical imaging data are highly sensitive; robust data-protection measures compliant with applicable regulations (e.g., HIPAA, GDPR, and local regulations) must be implemented.
- Clinical decision support versus replacement: the tool should be positioned as clinical decision support, not as a replacement for expert neurological assessment.

Technical considerations for African deployment:

- Computational requirements: the model requires GPU acceleration for real-time inference; edge-computing or cloud-based deployment may be needed in resource-limited settings.
- Connectivity: offline-capable solutions must be developed for areas with limited internet connectivity.
- Scanner compatibility: the model must be validated across the range of MRI scanners available in African healthcare facilities, which may include older equipment with different specifications.
- Maintenance and updates: ongoing model maintenance, performance monitoring and updates require sustained technical support infrastructure.

This study provides a technical proof-of-concept and methodology that can inform future implementation research; extensive clinical validation, regulatory approval and careful implementation planning remain prerequisites for responsible deployment.

4.5 Clinical Implications

The developed model offers potential benefits for clinical practice, including diagnostic support (automated classification providing preliminary assessments to support clinical decision-making, particularly where specialist availability is limited [30], [31]), standardisation (reduced inter-rater variability relative to subjective clinical assessment), efficiency (accelerated diagnostic workflows, potentially enabling earlier intervention), and resource optimisation (prioritisation of cases requiring specialist review in resource-constrained settings). These potential benefits must be balanced against the limitations and validation requirements discussed above.

4.6 Limitations

This study has several important limitations.

Dataset limitations:

- Single-source data: exclusive use of a publicly available Kaggle dataset limits generalisability; the data originate from a specific cohort with potentially homogeneous characteristics (demographics, imaging protocols, scanner types).
- Severe class imbalance: the extreme under-representation of Moderate Demented cases (n=64, 1% of the dataset) limits confidence in model performance for this critical severity level; clinical validation with larger samples of moderate-to-severe cases is essential.
- Lack of external validation: without validation on independent datasets from different institutions, populations and imaging equipment, the true generalisation capability remains uncertain.
- Selection bias: the original dataset likely underwent quality control and curation that may not reflect real-world clinical imaging conditions (motion artefacts, partial volumes, variable image quality).

Methodological limitations:

- Cross-sectional design: single-timepoint images were used rather than longitudinal data; progressive disease monitoring and prediction of future decline were not assessed.
- Limited clinical context: the model uses only MRI images, without incorporating cognitive test scores, biomarkers or patient history integral to clinical diagnosis.
- Discrete classification structure: while the model classifies into four severity levels, intermediate states and diagnostic uncertainty are not explicitly modelled.

- Lack of interpretability analysis: the current study does not include explainability methods (e.g., Grad-CAM, attention mechanisms) to visualise which brain regions drive classifications, limiting clinical trust and insight.

Generalisability concerns:

- Population diversity: the dataset's demographic composition (age, sex, race, ethnicity, education, comorbidities) is not fully characterised, raising questions about performance across diverse populations.
- Geographic applicability: validation on African patient populations is critically needed to assess applicability in the African context that this work is intended to inform.
- Scanner and protocol variation: real-world practice involves diverse MRI scanners, field strengths and acquisition protocols; robustness across this variation has not been assessed.

Implementation and clinical limitations:

- Computational resources: the model requires GPU acceleration, which may not be available in all clinical settings, particularly resource-limited environments.
- Regulatory status: the model has not undergone regulatory review or approval for clinical use.
- Clinical validation: prospective clinical trials comparing AI-assisted diagnosis with current standard of care are needed.
- Integration challenges: technical and workflow integration with existing healthcare systems has not been assessed.

Despite these limitations, this study provides proof-of-concept evidence for the potential of CNN-based AD classification and establishes a methodological foundation for future research in African health informatics contexts.

4.7 Future Directions

Building on this proof-of-concept work, several research directions warrant investigation:

- Explainable AI integration: implementation of visualisation techniques (Grad-CAM, attention maps) to identify brain regions contributing to classification decisions [27].
- Multi-modal learning: integration of multiple data modalities (MRI, PET, cognitive test scores, genetic markers) for comprehensive assessment.
- Longitudinal analysis: development of models for predicting disease progression and trajectory.
- African cohort validation: specific validation studies on African patient populations to assess performance in intended deployment contexts.
- Federated learning: privacy-preserving collaborative learning across institutions without centralising sensitive patient data.
- Mobile and edge deployment: development of lightweight model variants suitable for resource-constrained devices.

5 Conclusion

This study developed and technically validated a CNN model using the Inception V3 architecture for multi-class classification of AD severity from MRI images. The model achieved high-performance metrics (98.52% validation accuracy, 98.66% precision, 98.21% recall, 98.35% F1-score) on a widely used, publicly available dataset, demonstrating the technical feasibility of deep learning for this diagnostic task. The effective use of transfer learning, appropriate data preprocessing, and strategies to mitigate class imbalance contributed to robust model performance, adding to the growing evidence base supporting the potential role of AI in assisting AD diagnosis and severity staging.

Several critical caveats should be emphasised. First, this work represents a technical proof-of-concept rather than a clinically validated diagnostic tool: high performance on a curated dataset does not guarantee equivalent performance in real-world clinical conditions. Second, before any clinical deployment, extensive validation is required across multiple independent datasets, diverse patient populations (including African cohorts), various imaging equipment and protocols, real-world clinical conditions, and prospective clinical trials. Third, while this research contributes to African health informatics by demonstrating a feasible methodology, specific validation in African healthcare settings with African patient populations is essential before implementation. Fourth, significant work remains in regulatory approval, clinical integration, interpretability, health economics and implementation science.

The primary contribution of this work is methodological: it demonstrates how a state-of-the-art deep learning architecture can be applied to AD classification, documents strategies for addressing common challenges (class imbalance, limited data), and provides a framework that can inform future implementation research in African and other resource-limited contexts. The authors call for collaborative international research efforts to establish diverse validation datasets including African populations, develop and validate explainable-AI methods for clinical trust, conduct implementation-science research in African healthcare settings, address regulatory and ethical frameworks for AI in African health systems, and investigate cost-effectiveness and scalability. With appropriate validation, refinement and responsible implementation, AI-assisted diagnostic tools hold promises for enhancing AD detection and management, particularly in settings with limited specialist availability; this study provides a foundation for that work while acknowledging the substantial work that remains.

Acknowledgements

The authors acknowledge the Kaggle community for providing access to the Alzheimer's disease MRI dataset used in this study, and thank Google Colaboratory for the computational resources that facilitated model training and evaluation.

Funding

No external funding was received for this research.

Statement on Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Considerations

This study involved secondary analysis of a fully de-identified, publicly available dataset obtained from Kaggle. The dataset contained no directly identifiable personal information, and the study did not involve direct interaction with human participants or the collection of new human-subject data. Therefore, formal ethical approval or Institutional Review Board (IRB) approval was not required for this secondary analysis. The study was conducted in accordance with applicable ethical principles and the terms governing the use of the publicly available dataset.

Author Contributions

- **Samuel Yinka Olatunde:** Methodology, investigation, data curation, and writing—original draft.
- **Olatayo Moses Olaniyan:** Conceptualization, Methodology, investigation, data collection, and writing—original draft.

- **Abiodun Rasak Folorunsho:** Investigation, formal analysis, data curation, and writing—review and editing.
- **Stephen Olaoluwapo Adeleye:** Investigation, methodology, literature review, and writing—original draft.
- **Olalekan Joseph Dada:** Data collection, investigation, data curation, and writing—review and editing.
- **Micheal Abimbola Oladosu:** Methodology, formal analysis, supervision, project administration, and writing—review and editing.
- **Moses Adondua Abah:** Data curation, formal analysis, visualization, and writing—original draft and review and editing.
- **Olaoluwa John Adeleke:** Investigation, data collection, literature review, and writing—original draft.
- **Emmanuel Nkansah:** Methodology, investigation, data analysis, and writing—review and editing.
- **Shola David Omosa:** Investigation, data curation, visualization, and writing—review and editing.
- **Angel Ojimaajo Ekele:** Literature review, data curation, formal analysis, and writing—original draft and review and editing.

All authors contributed to the development of the manuscript, critically reviewed and approved the final version, and agreed to be accountable for all aspects of the work.

Data Availability

The dataset used in this study is publicly available on the Kaggle platform: Alzheimer's Dataset (4 Class of Images) – MRI, <https://www.kaggle.com/datasets/preetpalsingh25/alzheimers-dataset-4-class-of-images-mri>.

References

- [1] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr., C. H. Kawas, et al., "The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimers Dement.*, vol. 7, no. 3, pp. 263–269, 2011, doi: 10.1016/j.jalz.2011.03.005.
- [2] H. Braak and E. Braak, "Neuropathological staging of Alzheimer-related changes," *Acta Neuropathol.*, vol. 82, no. 4, pp. 239–259, 1991, doi: 10.1007/BF00308809.
- [3] P. M. Thompson, K. M. Hayashi, G. de Zubicaray, A. L. Janke, S. E. Rose, J. Semple, et al., "Dynamics of gray matter loss in Alzheimer's disease," *J. Neurosci.*, vol. 23, no. 3, pp. 994–1005, 2003, doi: 10.1523/JNEUROSCI.23-03-00994.2003.
- [4] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, et al., "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria," *Alzheimers Dement.*, vol. 12, no. 3, pp. 292–323, 2016, doi: 10.1016/j.jalz.2016.02.002.
- [5] G. B. Frisoni, N. C. Fox, C. R. Jack Jr., P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, 2010, doi: 10.1038/nrneurol.2009.215.
- [6] C. R. Jack Jr., D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, et al., "Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers," *Lancet Neurol.*, vol. 12, no. 2, pp. 207–216, 2013, doi: 10.1016/S1474-4422(12)70291-0.
- [7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

- [8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.
- [9] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, et al., "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation," *Med. Image Anal.*, vol. 63, p. 101694, 2020, doi: 10.1016/j.media.2020.101694.
- [10] M. Liu, D. Cheng, K. Wang, and Y. Wang, for the Alzheimer's Disease Neuroimaging Initiative, "Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, no. 3–4, pp. 295–308, 2018, doi: 10.1007/s12021-018-9370-4.
- [11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [12] M. Guerchet, P. Mbelesso, B. Ndamba-Bandzouzi, S. Pilleron, I. Désormais, P. Lacroix, et al., "Epidemiology of dementia in Central Africa (EPIDEMCA): protocol for a multicentre population-based study in rural and urban areas of the Central African Republic and the Republic of Congo," *SpringerPlus*, vol. 3, p. 338, 2014, doi: 10.1186/2193-1801-3-338.
- [13] S. M. Paddick, A. R. Longdon, A. Kisoli, C. Dotchin, W. K. Gray, M. Dewhurst, et al., "Dementia prevalence estimates in sub-Saharan Africa: comparison of two diagnostic criteria," *Glob. Health Action*, vol. 6, p. 19646, 2013, doi: 10.3402/gha.v6i0.19646.
- [14] C. L. Dotchin, R. O. Akinyemi, W. K. Gray, and R. W. Walker, "Geriatric medicine: services and training in Africa," *Age Ageing*, vol. 42, no. 1, pp. 124–128, 2013, doi: 10.1093/ageing/afs119.
- [15] R. O. Akinyemi, M. O. Owolabi, M. Ihara, P. Bangirana, I. Gioran, B. R. Ojebuyi, et al., "Stroke, cerebrovascular diseases and vascular cognitive impairment in Africa," *Brain Res. Bull.*, vol. 145, pp. 97–108, 2019, doi: 10.1016/j.brainresbull.2018.05.018.
- [16] P. Singh, "Alzheimer's Dataset (4 Class of Images) – MRI," Kaggle, San Francisco, CA, USA, 2021. [Online]. Available: <https://www.kaggle.com/datasets/preetpalsingh25/alzheimers-dataset-4-class-of-images-mri>. [Accessed: Jan. 2024].
- [17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [18] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, 2019, doi: 10.1186/s40537-019-0192-5.
- [19] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.
- [20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.
- [21] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162.
- [22] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, arXiv: 1412.6980.
- [23] M. Odusami, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Analysis of features of Alzheimer's disease: detection of early stage from functional brain changes in magnetic resonance images using a finetuned ResNet18 network," *Diagnostics*, vol. 11, no. 6, p. 1071, 2021, doi: 10.3390/diagnostics11061071.
- [24] A. Ebrahimi, S. Luo, and R. Chiong, "Introducing transfer learning to 3D ResNet-18 for Alzheimer's disease detection on MRI images," in *Proc. 35th Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Auckland, New Zealand, 2020, pp. 1–6, doi: 10.1109/IVCNZ51579.2020.9290616.

- [25] M. Odusami, R. Maskeliūnas, and R. Damaševičius, "An intelligent system for early recognition of Alzheimer's disease using neuroimaging," *Sensors*, vol. 22, no. 3, p. 740, 2022, doi: 10.3390/s22030740.
- [26] N. M. AbdelAziz, W. Said, M. M. AbdelHafeez, H. M. Ebeid, and N. I. Ghali, "Advanced interpretable diagnosis of Alzheimer's disease using SECNN-RF framework with explainable AI," *Front. Artif. Intell.*, vol. 7, p. 1456069, 2024, doi: 10.3389/frai.2024.1456069.
- [27] A. M. El-Assy, H. M. Amer, H. M. Ibrahim, and M. A. Mohamed, "A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data," *Sci. Rep.*, vol. 14, p. 3463, 2024, doi: 10.1038/s41598-024-53733-6.
- [28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [29] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, et al., "Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
- [30] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
- [31] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018, doi: 10.1038/s41551-018-0305-z.