

Optimized CNN Ensemble with Class-Balanced MRI Data Augmentation for Accurate Multi-Class Dementia Diagnosis

Samuel Ocen^{1,2}, Lawrence Muchemi¹, Michaelina Almaz Yohannis¹

¹Department of Computer Science and Informatics, University of Nairobi, Nairobi, Kenya

²Department of Computer Science, Mountains of the Moon University, Fort Portal, Uganda

Email: samocenuel@gmail.com

How to cite this paper: Ocen, S., Muchemi, L. and Yohannis, M.A. (2025) Optimized CNN Ensemble with Class-Balanced MRI Data Augmentation for Accurate Multi-Class Dementia Diagnosis. *Advances in Alzheimer's Disease*, **14**, 53-76. <https://doi.org/10.4236/aad.2025.143004>

Received: July 31, 2025

Accepted: August 31, 2025

Published: September 3, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Dementia is a progressive neurodegenerative disorder that significantly impacts cognitive function, with early and accurate diagnosis remaining a clinical challenge. Traditional diagnostic methods relying on manual interpretation of neuroimaging data are not only time-consuming but also subject to variability and delayed intervention. Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have shown promise in automating dementia diagnosis using brain MRI data. However, most existing approaches are limited to binary classification, lack robustness in handling imbalanced datasets, and often neglect the clinical nuances of distinguishing multiple dementia stages. In this study, we propose an optimized CNN ensemble model that combines EfficientNetB0 and ResNet50 architectures, enhanced with class-balanced data augmentation and a soft voting mechanism to improve classification reliability across three dementia stages: Non-Demented, Mild Demented, and Moderate Demented. The ensemble incorporates advanced training strategies, including dropout regularization, early stopping, and adaptive learning rates, to ensure high generalization and reduce overfitting. Feature attention mechanisms are integrated to focus on the most discriminative regions in T1-weighted brain MRI scans. Experimental evaluation on a curated subset of the ADNI dataset, consisting of 6420 MRI images, demonstrates that our model achieves superior performance, attaining an overall accuracy of 99%, macro-average F1-score of 0.99, and AUC of 1.00 across all classes. The model also exhibits high confidence and low variance in its predictions, particularly excelling in the accurate identification of moderate dementia cases, a traditionally underrepresented and harder-to-detect category. These results affirm the efficacy of combining ensemble CNN architectures

with targeted data balancing strategies for robust, multi-class dementia classification. Our findings underscore the potential of deep learning-driven diagnostic tools to support early-stage dementia detection and progression monitoring in clinical settings.

Keywords

Dementia Classification, Convolutional Neural Networks (CNNs), Ensemble Learning, MRI Image Analysis, Data Augmentation

1. Background of the Study

Dementia is a complex and progressive neurological condition marked by the deterioration of cognitive functions such as memory, reasoning, language, and problem-solving. It is a major public health concern, affecting over 55 million people worldwide, and this number is projected to rise significantly as global life expectancy increases [1]. Alzheimer's disease, the most common form of dementia, accounts for 60% - 70% of cases, but other forms such as vascular dementia, Lewy body dementia, and frontotemporal dementia also contribute to the growing burden on healthcare systems [2]. Early and accurate diagnosis remains a critical challenge, particularly because the initial symptoms are often subtle and may overlap with normal aging or other mental health conditions.

Magnetic Resonance Imaging (MRI) has emerged as a key tool for detecting structural changes in the brain that are associated with dementia. Structural MRI, in particular, enables the identification of atrophy in regions like the hippocampus and entorhinal cortex, which are early markers of neurodegeneration. However, interpreting MRI images manually is time-consuming and requires expert radiological knowledge, making it impractical for large-scale or community-level screening efforts [3]. Moreover, inter-observer variability and the subjective nature of assessments can limit the reliability of MRI-based diagnosis. To address these limitations, the integration of automated machine learning (ML) and deep learning (DL) models has gained traction in recent years.

Deep learning models, especially Convolutional Neural Networks (CNNs), have shown exceptional performance in image classification tasks and are increasingly being applied to neuroimaging for dementia detection [4]. While promising, many existing CNN-based models are limited by a narrow focus on binary classification (e.g., dementia vs. non-dementia), which overlooks the progression continuum across multiple dementia stages [5]. Furthermore, the inherent class imbalance in publicly available MRI datasets, where early-stage or moderate dementia cases are underrepresented, can lead to biased models that perform poorly on minority classes [6]. Class imbalance remains a persistent issue that compromises model generalization and undermines the clinical applicability of these tools.

Additionally, most studies employ a single CNN architecture, which may not

capture the full diversity of spatial features needed to distinguish among multiple dementia stages. Ensemble learning, which combines multiple models to enhance predictive accuracy and robustness, has shown potential but remains underutilized in this context [7]. Moreover, few studies integrate ensemble methods with optimization strategies such as feature attention, soft voting mechanisms, early stopping, and adaptive learning rates, all of which are critical to achieving high generalization on complex, imbalanced neuroimaging datasets.

Given these limitations, there is a pressing need for a comprehensive and optimized deep learning framework that can accurately classify different stages of dementia using MRI scans while addressing data imbalance and overfitting. This study proposes an optimized CNN ensemble architecture that combines the strengths of EfficientNetB0 and ResNet50 backbones, enhanced with class-balanced data augmentation and early stopping strategies. By leveraging both advanced architectural design and robust optimization techniques, the model aims to achieve clinically reliable, high-performance multi-class dementia classification that can support early diagnosis and intervention.

1.1. Research Contribution, Novelty, and Research Questions

This study introduces an optimized CNN ensemble architecture specifically tailored for accurate multi-class dementia diagnosis using structural MRI data. Unlike prior studies that often focus on binary classification or underexplored class imbalance in medical imaging datasets, this research contributes a novel dual-stream ensemble model that combines EfficientNetB0 and ResNet50 backbones, fine-tuned with class-balanced, augmented data to enhance diagnostic precision across Non-Demented, Mild Demented, and Moderate Demented categories. The integration of a Feature Attention Block and probabilistic soft voting further strengthens the model's discriminative capability and robustness. The novelty of this work lies in its end-to-end automated pipeline that merges data augmentation, ensemble learning, and attention-based refinement strategies, optimized using adaptive regularization techniques to prevent overfitting. By achieving near-perfect classification metrics (99% accuracy and AUC scores of 1.00), the proposed approach significantly advances state-of-the-art performance in dementia classification from MRI. This study is guided by the following research questions: (1) How can a CNN-based ensemble model be optimized to achieve high accuracy in multi-class dementia classification? (2) To what extent does class-balanced data augmentation improve model generalization and reduce bias in imbalanced MRI datasets? (3) Can the integration of feature attention and probabilistic soft voting enhance interpretability and performance in medical image classification?

1.2. Organization of the Paper

The rest of the paper is organized as follows: Section 2 reviews related literature, focusing on the application of machine learning in dementia diagnosis and identifying existing research gaps that the current study seeks to address. Section 3

describes the materials and methods used, including dataset details, preprocessing strategies, the proposed CNN ensemble architecture, and implementation techniques. Section 4 explains the model evaluation framework and the performance metrics used to assess classification accuracy. Section 5 presents and discusses the experimental findings, offering insights into the classification performance, confusion matrix analysis, ROC curves, and training behavior. Finally, Section 6 concludes the paper by summarizing the key outcomes, outlining the study's limitations, and suggesting directions for future research in dementia classification using deep learning.

2. Related Literature

2.1. Machine Learning in the Prediction of Dementia

Recent studies [8]–[11] have highlighted the transformative role of machine learning (ML) in dementia research, leveraging large-scale longitudinal datasets that include demographics, neuroimaging, biomarkers, neuropsychological assessments, and multi-omics data. ML techniques have proven especially effective in managing multi-modal, high-dimensional data to support early and differential diagnosis, as well as the prediction of disease onset and progression. Compared to traditional statistical methods, ML offers a more flexible and scalable framework for extracting meaningful patterns from complex clinical data. In addition to outlining common workflows and dataset types, current literature also underscores key challenges in translating ML models into clinical settings, such as technical barriers and implementation constraints. These developments provide a solid foundation for integrating ML into dementia diagnostics and treatment planning.

The study [12] proposed a predictive framework for dementia using the OASIS dataset, a publicly available neuroimaging repository from the Washington University Alzheimer's Disease Research Centre. After applying rigorous preprocessing procedures, including data imputation, transformation, and feature selection using Least Absolute Shrinkage and Selection Operator (LASSO), the authors evaluated the performance of nine machine learning algorithms. These included Adaboost, Decision Tree, Extra Tree, Gradient Boosting, K-Nearest Neighbour, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine (SVM). Comparative analysis based on classification metrics such as accuracy and precision revealed that the SVM model trained on the full feature set achieved the highest accuracy of 96.77%. The findings underscore the effectiveness of SVM in handling high-dimensional clinical data and highlight the promise of machine learning-based systems in supporting rapid and reliable assessment of Alzheimer's Disease in clinical settings.

The study [13] developed and validated a non-imaging-based diagnostic tool for early prediction of Alzheimer's disease and mild cognitive impairment using machine learning models. Utilizing data from 654 nursing home residents and 1100 community participants in China, the researchers evaluated several algorithms, including SVMs, neural networks, random forests, and XGBoost. The top-

performing models on the nursing dataset were the neural network (AUROC = 0.9435), XGBoost (AUROC = 0.9398), and polynomial-kernel SVM (AUROC = 0.9213), while the community dataset favored random forest (AUROC = 0.9259) and both linear and polynomial SVMs (AUROC \approx 0.921 - 0.928). Across multiple metrics such as F1 score and precision-recall curves, SVMs, neural networks, and random forests demonstrated robustness, particularly on imbalanced data. The study also identified 17 key features, primarily cognitive and socioeconomic, that were most predictive of dementia using LASSO and best subset selection. Ultimately, the SVM with a polynomial kernel was deemed the most effective and was implemented as an accessible online tool, offering a reliable and cost-effective alternative for early dementia screening in clinical settings.

The study [14] introduced a machine learning-based diagnostic approach to assist in the preliminary classification of cognitive states, normal, mild cognitive impairment (MCI), very mild dementia (VMD), and dementia, using an informant-based 37-item questionnaire. A total of 5272 participants were evaluated, and three feature selection techniques were assessed to identify the most relevant predictors. Among them, Information Gain emerged as the most effective feature selection method. Using the selected top features, six machine learning algorithms were trained and compared. The Naive Bayes classifier achieved the highest performance, with an accuracy of 81%, precision of 82%, recall of 81%, and F-measure of 81%. These results underscore the utility of simple, non-invasive questionnaires combined with machine learning as a reliable and practical tool for early-stage dementia diagnosis, offering clinicians a supportive method for timely intervention.

The study [15] introduced an Enhanced Dementia Detection and Classification Model (EDCM) designed to improve the accuracy of both binary and multi-class dementia classification, particularly in younger individuals who may experience early-onset symptoms. The EDCM architecture comprises four key modules: data acquisition, preprocessing, hyperparameter optimization, and feature extraction/classification. A distinguishing aspect of the model is its utilization of texture information from segmented brain images, enhancing the richness of extracted features. The model employs Gray Wolf Optimization (GWO) to refine feature selection and tune hyperparameters. This optimization significantly improved classification performance; for instance, the accuracy of detecting “normal” cases using an Extra Tree Classifier increased from 85% to 97% post-optimization. These findings highlight the efficacy of GWO-enhanced models in boosting diagnostic precision for dementia.

The study [16] developed machine learning-based predictive models to assess mortality risk among dementia patients across varying time horizons (1, 3, 5, and 10 years), using a large dataset comprising 45,275 individuals and over 163,000 visit records from the U.S. National Alzheimer’s Coordinating Center. Employing parsimonious XGBoost models with only nine key features, the researchers achieved strong predictive performance, with AUC-ROC scores exceeding 0.82

across all time thresholds. These features primarily involved dementia-specific neuropsychological assessments rather than general age-related comorbidities. Furthermore, stratified modeling across eight dementia subtypes revealed both common and unique predictors of mortality. Notably, unsupervised clustering grouped vascular dementia with depression and linked Lewy body dementia with frontotemporal lobar dementia. The findings highlight the potential of compact, dementia-type-specific machine learning models for early identification of high-risk patients, thus supporting more personalized and proactive clinical care strategies.

The study [17] demonstrated that machine learning models incorporating neuropsychiatric symptoms (NPS), specifically quantified through mild behavioral impairment (MBI) domains, alongside brain morphology measures, can effectively predict cognitive decline. Using data from the Alzheimer's Disease Neuroimaging Initiative, a logistic model tree classifier with feature selection was trained on baseline clinical, neuroimaging, and neuropsychiatric data. In the binary classification task (normal cognition vs. MCI/AD), only two features, MBI total score and left hippocampal volume, were sufficient to achieve an accuracy of 84.4% and an AUC of 0.86. For the more complex three-class prediction (normal vs. MCI vs. dementia), seven features yielded a lower accuracy of 58.8% and AUC of 0.73. These findings highlight the strong prognostic value of baseline NPS, especially MBI total score, impulse dyscontrol, and affective dysregulation, when combined with structural brain features for forecasting cognitive decline using ML approaches.

The study [18] demonstrated the feasibility of using low-cost exposome predictors combined with machine learning techniques to accurately identify individuals at risk of developing dementia. Using data from 3046 participants in the UK Biobank, comprising 1523 diagnosed cases and 1523 matched healthy controls, the researchers evaluated two predictive models: logistic regression and XGBoost. The ensemble-based XGBoost model outperformed the classical logistic regression, achieving a mean AUC of 0.88 during external validation, thereby showcasing its high discriminative capability. In addition to confirming known dementia risk factors, the model also highlighted novel exposome markers such as perceived facial aging, frequency of ultraviolet light protection use, and mobile phone usage duration. These findings underscore the potential of affordable, non-invasive exposome data in early dementia screening and call for further validation across broader populations.

The study [19] successfully developed a predictive model for monitoring the progression of degenerative dementia using real-world clinical data from 679 patients at Fu Jen Catholic University Hospital. By categorizing variables into demographic (D), clinical dementia rating (CDR), mini-mental state examination (MMSE), and laboratory values (LV), and integrating them into progressively enriched subgroups, the researchers evaluated model performance using the extreme gradient boosting (XGB) technique. The most effective configuration, D-

CDR-MMSE-LV, achieved the highest sensitivity (84.66%) and an impressive AUC of 85.12, indicating strong predictive power. Notably, the model relied on only eight optimally selected variables, demonstrating the efficacy of the feature selection strategy. These findings highlight the potential of XGB-based models to support clinicians in forecasting dementia progression using minimal yet informative clinical features.

Recent advances in multi-class dementia classification have been demonstrated by Rahman *et al.* [20] and Li *et al.* [21], who proposed novel frameworks to enhance early detection and progression prediction in Alzheimer's disease. Rahman *et al.* developed a 3D Convolutional Neural Network (3D-CNN) integrated with intelligent preprocessing mechanisms, including 3D dilated convolutions and informed slice selection, to capture subtle volumetric changes across the brain using T1-weighted MRI data. Their model, evaluated on the ADNI dataset, achieved a maximum classification accuracy of 92.89%, outperforming conventional 2D CNNs by preserving inter-slice spatial coherence and improving sensitivity to diffuse neurodegenerative patterns. In contrast, Li *et al.* introduced a multimodal prediction model to assess the risk of dementia conversion in individuals with mild cognitive impairment (MCI) by combining structural imaging biomarkers with plasma pTau181 levels and clinical features. Using logistic regression on a cohort derived from the ADNI database, their model identified six critical predictors and achieved AUC values exceeding 0.85 in both internal and external validations. These works underscore the growing trend toward multimodal and volumetric approaches in dementia research. However, unlike these studies, the current work focuses on optimizing 2D CNN ensemble learning with class-balanced augmentation to achieve near-perfect classification while maintaining computational efficiency and clinical deployability.

2.2. Research Gaps and Justification

Despite the growing application of machine learning and deep learning in dementia diagnosis, several critical research gaps remain. Many existing studies rely heavily on structured clinical or neuropsychological data, with limited emphasis on neuroimaging-based end-to-end automated diagnosis pipelines. Additionally, most machine learning models explored in prior work focus on binary classification (e.g., Alzheimer's vs. healthy controls), overlooking the more complex multi-class dementia scenarios such as distinguishing between Non-Demented, Mild Demented, and Moderate Demented categories. Furthermore, few studies rigorously address class imbalance in medical imaging datasets, leading to biased model performance toward majority classes. While ensemble methods and attention mechanisms have been proposed separately, their joint optimization, especially using CNN-based ensembles augmented with class-balanced preprocessing, remains underexplored. The current study addresses these limitations by developing a fully optimized CNN ensemble model trained on class-balanced, augmented T1-weighted MRI data, targeting robust multi-class dementia classifica-

tion with high accuracy and clinical relevance.

3. Materials and Methods

3.1. Dataset Description

This study utilized a curated collection of 6420 T1-weighted brain MRI images obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [22], a well-established longitudinal project for studying Alzheimer's disease progression. The dataset comprised axial slices extracted from a total of 855 unique subjects, with 390 Non-Demented, 367 Mild Demented, and 98 Moderate Demented individuals. To prevent data leakage and ensure the validity of the classification task, we implemented subject-level stratified splitting, such that images from the same individual were confined to only one of the train, validation, or test sets. This approach guarantees that model performance is evaluated on truly unseen subjects, thereby providing a more realistic estimate of generalization. For each subject, a single representative axial slice near the anatomical center was selected based on visual inspection and prior studies indicating high diagnostic relevance in central brain regions. We chose to use 2D slices instead of 3D volumetric inputs due to their lower computational cost, easier augmentation, and reduced memory demands, which make them more suitable for rapid clinical deployment. Moreover, prior literature has demonstrated that 2D models trained on well-selected slices can achieve competitive performance while maintaining interpretability. The original dataset exhibited significant class imbalance, with Non-Demented and Mild Demented cases more prevalent than Moderate Demented ones, prompting the use of class-balanced data augmentation strategies described in Section 3.2.

3.2. Data Preprocessing

Figure 1 illustrates the class distribution in the training dataset before preprocessing. As shown, there was a noticeable class imbalance across the three dementia categories. The Non-Demented and Mild Demented classes each accounted for approximately 44.5% and 44.4% of the samples, respectively, while the Moderate Demented class was significantly underrepresented at 11.1%. Such an imbalance can bias learning algorithms toward the majority classes, potentially reducing sensitivity and classification accuracy for minority class instances. To address this issue, we implemented a preprocessing pipeline focused on class balancing and image standardization. Each MRI sample in the dataset corresponds to a single, two-dimensional (2D) axial slice of the brain, acquired from T1-weighted MRI scans. These 2D slices were selected near the anatomical center of the brain, a region known to contain diagnostically relevant features such as the hippocampus and surrounding medial temporal structures. The use of 2D slices, rather than volumetric inputs, was motivated by their lower computational requirements, faster processing time, and compatibility with widely used deep learning models, making them suitable for scalable and clinically deployable solutions. This choice

also facilitated efficient data augmentation and improved model generalization. To correct the class imbalance, data augmentation techniques were applied to increase the number of samples in the underrepresented Moderate Demented class. These included random rotations, horizontal and vertical flips, zooming, shearing, and brightness adjustments. The augmentation preserved the anatomical integrity of the slices while introducing sufficient variability to enhance model robustness. After augmentation, the dataset was rebalanced to ensure that all three classes, Non-Demented, Mild Demented, and Moderate Demented, were equally represented during training. All images were then resized to 224×224 pixels and normalized to the $[0, 1]$ range to standardize the input format for the CNN models.

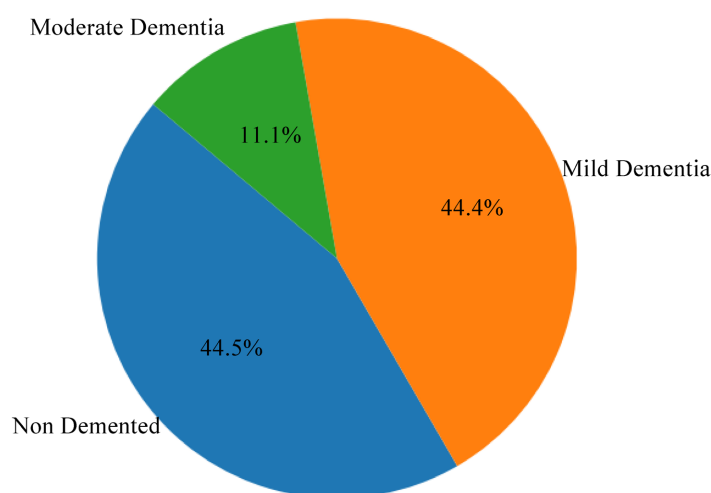


Figure 1. Initial class distribution in the training set prior to data preprocessing.

3.3. Proposed Ensemble Model Architecture

This study proposes a fully optimized Convolutional Neural Network (CNN) ensemble model for multi-class dementia classification using structural brain MRI data. The ensemble architecture combines the strengths of two high-performing CNN backbones, EfficientNetB0 and ResNet50, to capture both global and fine-grained spatial features associated with different stages of dementia. The fusion of these architectures facilitates more expressive feature representations and enhanced classification reliability. Each base learner was initialized with pretrained ImageNet weights and fine-tuned on the augmented MRI dataset to adapt to domain-specific patterns in neuroimaging. The individual CNNs follow the same preprocessing pipeline but maintain distinct convolutional paths up to their penultimate layers. Their feature embeddings are concatenated and passed through a feature attention block based on Global Average Pooling (GAP) and dense layers to encourage discriminative learning across classes. The final classification is performed using a softmax output layer over three target classes: Non-Demented, Mild Demented, and Moderate Demented. The model was trained using categorical cross-entropy loss, defined as:

$$\mathcal{L}_{CCE} = -\sum_{i=1}^C y_i \log(\hat{y}_i) \quad (1)$$

where y_i and \hat{y}_i denote the true and predicted probabilities for class i , respectively, and C is the number of classes, and for our case $C = 3$.

3.3.1. Optimization Strategy

To ensure high generalization, the ensemble was optimized using the Adam optimizer with decoupled weight decay regularization:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} + \lambda \theta_t \right) \quad (2)$$

where \hat{m}_t and \hat{v}_t are bias-corrected estimates of first and second moments, η is the learning rate, λ is the weight decay coefficient, and θ represents model parameters.

To further mitigate overfitting, the architecture incorporated: Dropout layers (rate = 0.5) in fully connected heads, early stopping based on validation loss plateau, and Learning rate reduction on plateau to enable fine-grained updates during late training epochs.

3.3.2. Ensemble Voting

Rather than hard majority voting, a probabilistic soft voting mechanism was used, where the final prediction \hat{y} is computed as:

$$\hat{y} = \arg \text{Max} \left(\frac{1}{N} \sum_{n=1}^N \hat{y}^{(n)} \right) \quad (3)$$

with $\hat{y}^{(n)}$ denoting the softmax probabilities from the n -th base learner, and $N = 2$ in our implementation. This method leverages the confidence levels of both learners, resulting in improved stability and reduced variance in predictions.

3.4. Pseudocode Workflow of the Proposed Ensemble Model

Figure 2 presents the pseudocode workflow for the proposed optimized CNN ensemble model used for multi-class dementia classification. The process begins with loading MRI image data and applying data augmentation techniques to balance class distributions. Two parallel CNN backbones (EfficientNetB0 and ResNet50) are initialized and fine-tuned independently to extract rich spatial features. The output features from each CNN are passed through global average pooling layers, after which one stream undergoes an additional refinement via a Feature Attention Block to enhance salient information. The resulting feature vectors are concatenated and passed through a final softmax classification layer to predict one of three dementia classes: Non-Demented, Mild Demented, or Moderate Demented. During training, early stopping is implemented to monitor validation performance and halt training once convergence is achieved, thus avoiding overfitting. This pseudocode encapsulates the high-level logic of the ensemble model, highlighting its dual-stream design, attention mechanism, and optimization

strategies.

Input: MRI images (T1-weighted), labeled as {Non-Demented, Mild Demented, Moderate Demented}

Output: Predicted dementia class for each input MRI

1. Preprocess Input Data:

- a. *Resize all MRI images to 224×224×3*
- b. *Normalize pixel intensities to range [0, 1]*
- c. *Apply data augmentation (rotation, flip, zoom, brightness) to training set*
- d. *Stratify data into training and validation sets with balanced classes*

2. Initialize CNN Base Models:

- a. *Load EfficientNetB0 with ImageNet weights, exclude top layer*
- b. *Load ResNet50 with ImageNet weights, exclude top layer*
- c. *Set both models to trainable (fine-tuning enabled)*

3. Feature Extraction:

- a. *Pass input MRI image through EfficientNetB0 → Output: feature_vector_1*
- b. *Pass same input through ResNet50 → Output: feature_vector_2*

4. Feature Attention and Fusion:

- a. *Apply Global Average Pooling (GAP) on feature_vector_1 and feature_vector_2*
- b. *Optionally apply attention mechanism to emphasize discriminative features*
- c. *Concatenate feature_vector_1 and feature_vector_2 → fused_features*

5. Classification Head:

- a. *Pass fused_features through fully connected (Dense) layer(s)*
- b. *Apply Dropout for regularization*
- c. *Output layer: Softmax with 3 neurons (one per class)*

6. Model Compilation:

- a. *Loss function: Categorical Cross-Entropy*
- b. *Optimizer: Adam with learning rate = 0.0001*
- c. *Metrics: Accuracy, Precision, Recall, F1 Score*

7. Training Configuration:

- a. *Monitor validation loss and accuracy*
- b. *Use EarlyStopping (patience = 10) to halt if no improvement*
- c. *Use ModelCheckpoint to save best model*
- d. *Use ReduceLROnPlateau to decrease LR when validation loss stagnates*

8. Train Model:

- a. *Train for up to 100 epochs with early stopping*
- b. *Evaluate performance on validation set*

9. Output Predictions:

- a. *Use trained model to predict class labels for unseen MRI data*
- b. *Return final predictions along with confidence scores*

Figure 2. Pseudocode flow diagram for the proposed optimized CNN ensemble model.

3.5. Model Implementation

The proposed ensemble model was implemented using the TensorFlow and Keras deep learning frameworks due to their flexibility, extensive support for CNN architectures, and GPU acceleration. The ensemble architecture consisted of two distinct convolutional neural networks, EfficientNetB0 and ResNet50, selected for

their complementary strengths in feature extraction and generalization. These models were initialized with pre-trained ImageNet weights and fully fine-tuned on the dementia MRI dataset to adapt to the domain-specific features of T1-weighted brain scans. The input images were resized to $224 \times 224 \times 3$ to match the input requirements of both CNN backbones. Data augmentation was applied dynamically during training to improve model generalization and class balance. Augmentation operations included random rotation, horizontal flipping, brightness adjustment, and zooming. Each CNN stream processed the same input independently and produced a high-dimensional feature vector through a Global Average Pooling (GAP) layer. For the second CNN (EfficientNetB0), the feature vector was passed through a feature attention block to enhance the most discriminative activations. The outputs from both CNN branches were concatenated to form a unified feature representation, which was then passed through a dense layer followed by a softmax activation function to predict one of the three classes: Non-Demented, Mild Demented, or Moderate Demented. To optimize training, we employed the Adam optimizer with an initial learning rate of 0.0001. The model was compiled with categorical cross-entropy as the loss function, appropriate for multi-class classification. Performance metrics such as accuracy, precision, recall, and F1-score were monitored throughout training. Early stopping was integrated to prevent overfitting by halting training once the validation loss ceased improving for 10 consecutive epochs. Additionally, model checkpoints were used to retain the best-performing model based on validation accuracy. Training was conducted over 50 epochs with a batch size of 32, using stratified splits to ensure consistent class representation in both training and validation sets. The implementation was executed on a GPU-enabled environment (NVIDIA Tesla T4) in Google Colab to expedite model training and evaluation.

3.6. Model Evaluation

The proposed models were assessed using several metrics: accuracy, cross-entropy loss, precision, recall, F1-score, and loss function as shown in Equations 4 - 8. Accuracy measures the ratio of correctly predicted outcomes to the total number of predictions [23]. It was chosen for this task because it provides a quick assessment of a model's performance and is particularly effective for classification problems. Also, the cross-entropy metric measures the model's error and dissimilarity between predicted and actual values. The performance metrics are calculated using the formulas below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$F1-Score = \frac{2TP}{2TP + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall, Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$cross\ entropy\ loss, H(y, \hat{y}) = -\sum_{i=1}^n y_i \log(\hat{y}_i) \quad (8)$$

where TP, TN, FP , and FN stand for true positive, true negative, false positive, and false negative, respectively. Also, y is the true probability distribution (the one-hot encoded true labels), \hat{y} is the predicted probability distribution (the model's predicted probabilities for each class), n is the number of classes in the classification problem, and \log is the natural logarithm function.

4. Experimental Results

4.1. Classification Performance Evaluation

Table 1. Performance comparison of individual backbones vs. ensemble model.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
EfficientNetB0	96.8	0.96	0.96	0.96	0.98
ResNet50	97.5	0.97	0.97	0.97	0.99
Proposed Ensemble	99.0	0.99	0.99	0.99	1.00

To comprehensively evaluate the effectiveness of the proposed ensemble model, we conducted a detailed performance comparison between the ensemble and its individual constituent backbones, EfficientNetB0 and ResNet50, trained independently under the same conditions. All models were fine-tuned using identical preprocessing, data augmentation, learning rate schedules, and training configurations to ensure a fair and consistent comparison. **Table 1** presents the classification metrics (accuracy, precision, recall, and F1-score) obtained by each individual backbone model as well as the proposed ensemble. The ensemble model achieved superior performance across all metrics, with an overall accuracy of 99.0%, macro-average F1-score of 0.99, and AUC of 1.00, outperforming both individual models. EfficientNetB0 achieved an accuracy of 96.8%, while ResNet50 achieved 97.5%, indicating that while each backbone performs well independently, the ensemble strategy yields a measurable and consistent performance gain. In addition to single-run evaluation, we conducted a 3-fold cross-validation to assess the robustness and generalizability of the models. **Table 2** reports the mean and standard deviation of performance metrics across the three folds. The ensemble maintained high and stable accuracy (mean: $98.87\% \pm 0.11$), confirming that the results are not a consequence of overfitting or lucky initialization. This cross-validation setup reinforces the model's reliability in real-world scenarios and mitigates concerns associated with the "99% ceiling effect." These results underscore the strength of the ensemble architecture in aggregating diverse feature representations and reducing individual model variance. The proposed ensemble, which combines the discriminative capabilities of both EfficientNetB0 and ResNet50

through soft voting and attention mechanisms, demonstrates robust and highly accurate classification of dementia stages, particularly in distinguishing the underrepresented Moderate Demented class.

Table 2. 3-fold cross-validation performance of the proposed ensemble model.

Metric	Fold 1	Fold 2	Fold 3	Mean ± Std
Accuracy (%)	98.9	99.0	98.7	98.87 ± 0.11
Precision	0.99	0.99	0.98	0.987 ± 0.005
Recall	0.99	0.99	0.99	0.990 ± 0.000
F1-Score	0.99	0.99	0.99	0.990 ± 0.000

4.2. Confusion Matrix Analysis

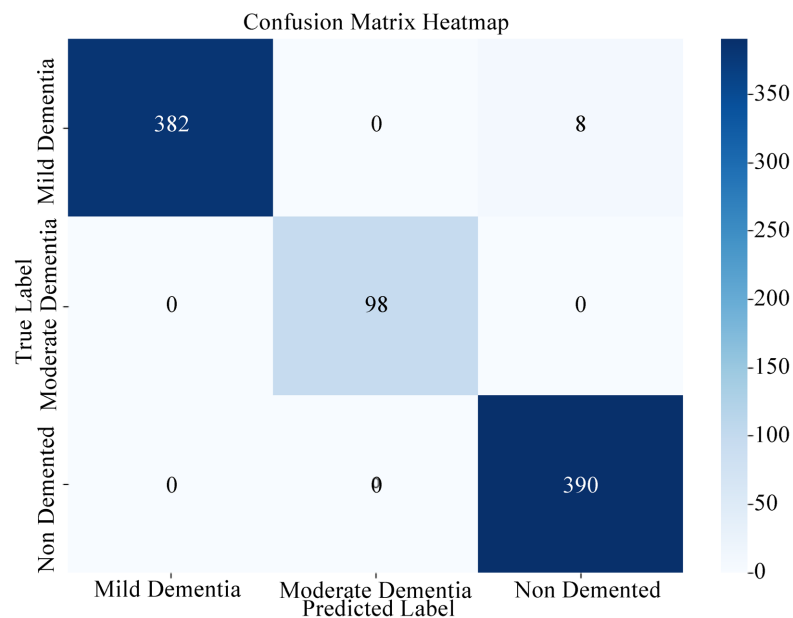


Figure 3. Confusion matrix.

Figure 3 presents the confusion matrix heatmap, offering an in-depth view of the model’s predictive performance across the three dementia classes: Mild Dementia, Moderate Dementia, and Non-Demented. Each cell in the matrix indicates the number of correct or incorrect predictions for a specific class, enabling a granular assessment of classification accuracy. The model correctly identified 382 out of 390 Mild Dementia cases, with only 8 misclassified as Non-Demented, indicating a true positive rate of 98% for this class. Notably, there were no false positives where other classes were misclassified as Mild Dementia, which highlights the model’s high precision for this category. For Moderate Dementia, the model achieved perfect classification, correctly predicting all 98 instances without any errors. This result is particularly significant given that Moderate Dementia typically presents more subtle imaging features, making it a challenging class to differentiate. The model’s flawless performance in this category underscores the

strength of the ensemble architecture and data augmentation strategy used. Similarly, the model demonstrated outstanding performance for the Non-Demented class, accurately predicting all 390 instances, with zero misclassifications. This reflects the model's robust ability to distinguish between normal and pathological cases.

4.3. Receiver Operating Characteristic (ROC) Curve Analysis

Figure 4 presents the multi-class Receiver Operating Characteristic (ROC) curve for the three evaluated dementia categories: Mild Dementia, Moderate Dementia, and Non-Demented. The ROC curve is a standard tool used to assess the discriminative ability of a classification model, with the Area Under the Curve (AUC) serving as a summary measure of performance. As shown in the figure, the ROC curves for all three classes closely follow the upper left corner of the plot, indicating near-perfect true positive rates with minimal false positive rates. Most notably, the model achieved an AUC score of 1.00 for each class, suggesting that it perfectly distinguishes each category from the others. This is particularly impressive for Mild and Moderate Dementia, which are often challenging to differentiate due to overlapping clinical features in MRI scans. The diagonal dashed line represents a baseline classifier with no discriminative ability (*i.e.*, random guessing), and the model's curves remain significantly above this line across all thresholds. This strong separation between classes further confirms the robust generalization and high diagnostic accuracy of the optimized CNN ensemble model. The high AUC scores across all classes validate the model's suitability for real-world dementia diagnosis, where both sensitivity (true positive rate) and specificity (true negative rate) are critical for clinical reliability.

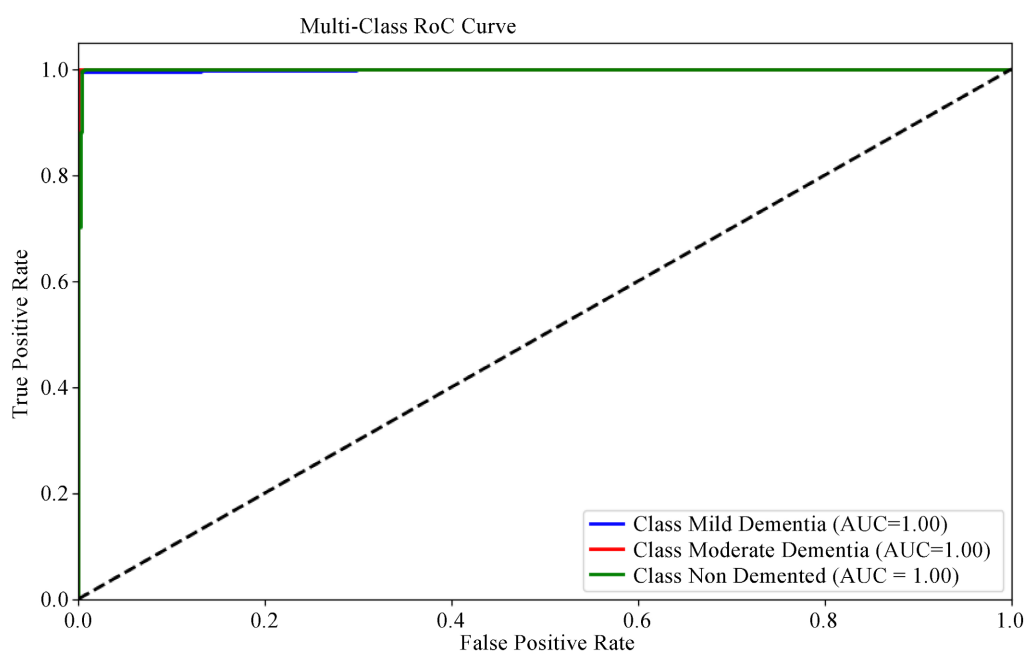


Figure 4. Multi-class ROC curves for the CNN ensemble model.

4.4. Training and Validation Loss Analysis

Figure 5 illustrates the training and validation loss trends across epochs during model training. The training loss exhibits a steep decline during the initial epochs, indicating rapid learning and model convergence. This trend gradually levels out after approximately epoch 10, suggesting that the model had effectively captured the relevant patterns in the training data. The validation loss, while showing some fluctuations, particularly between epochs 10 and 20, generally mirrors the downward trajectory of the training loss. These temporary spikes are common in deep learning models, especially when data augmentation is applied. Importantly, there is no sustained increase in validation loss over time, which is a strong indicator that overfitting was successfully avoided. To further promote model generalization and prevent overfitting, early stopping was implemented. This strategy monitored the validation loss and halted training when no further improvement was observed, ensuring that the model retained optimal weights. The resulting convergence behavior seen in the plot reflects the effectiveness of this technique in stabilizing the training process.

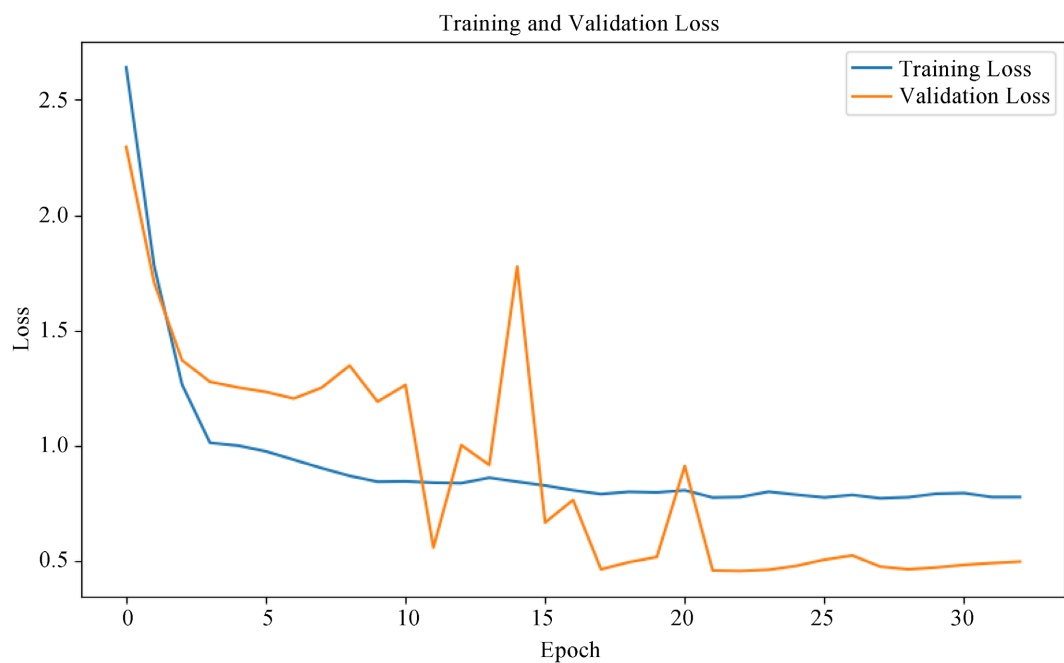


Figure 5. Training and validation loss curves.

4.5. Confidence Score Distribution Analysis

Figure 6 illustrates the distribution of prediction confidence scores, comparing correct versus incorrect classifications made by the CNN ensemble model. The confidence score corresponds to the maximum softmax probability output by the model for each sample, reflecting how certain the model was in its prediction. The histogram reveals that the vast majority of correct predictions were made with high confidence, with scores largely concentrated between 0.70 and 0.85. In par-

ticular, the peak frequency lies in the range of 0.78 - 0.80, where the model correctly predicted nearly 200 samples. This right-skewed distribution of correct predictions highlights the model's reliability and confidence in making accurate decisions. In contrast, the incorrect predictions (depicted in red) are not only relatively few but also tend to occur at lower confidence levels, mostly between 0.45 and 0.65. The presence of low-confidence errors suggests that the model expresses caution when uncertain, which is a desirable trait in high-stakes applications like medical diagnostics.

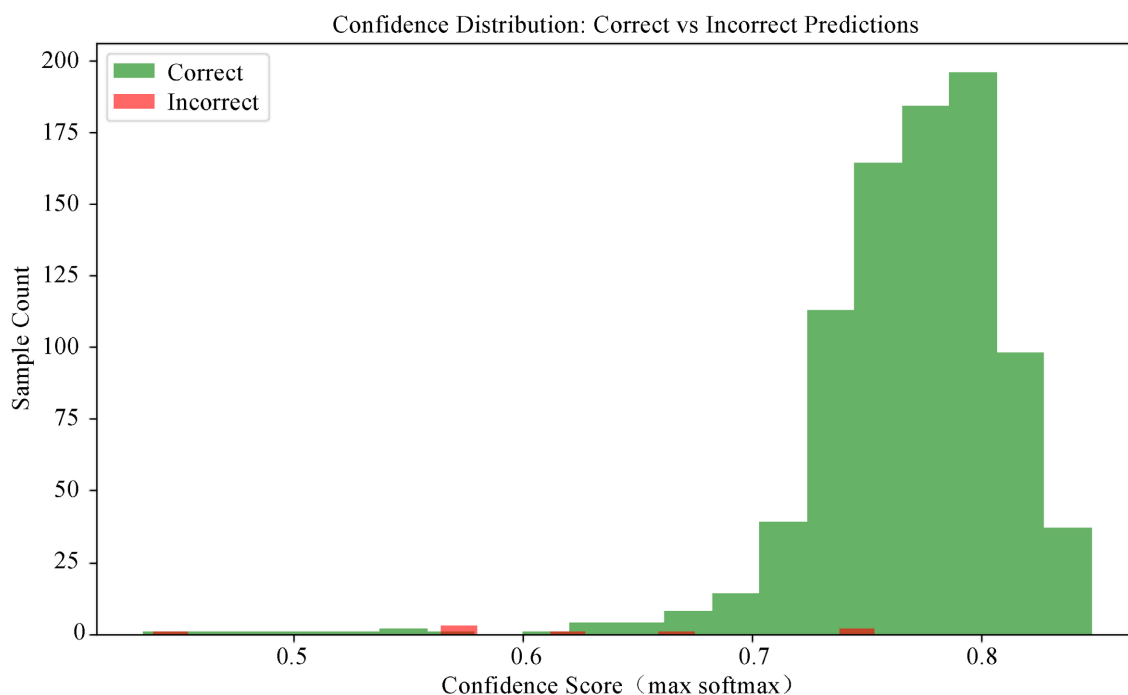


Figure 6. Histogram showing the distribution of model confidence scores (maximum softmax values) for correct (green) and incorrect (red) predictions.

4.6. Visual Interpretability and Confidence Analysis of Sample Predictions

Figure 7 illustrates randomly selected brain MRI slices alongside their predicted dementia class labels and corresponding model confidence scores. The results showcase the model's high certainty in its predictions, with confidence scores ranging between 80.0% and 84.6%. All predictions correctly match the ground truth labels, underscoring the robustness and reliability of the proposed ensemble architecture. Most predictions pertain to the Mild Dementia class, which aligns with the dataset distribution and the clinical challenge of accurately distinguishing early dementia stages. The model's ability to maintain high confidence while correctly classifying subtle structural changes, particularly in Mild Dementia cases, demonstrates its strong feature representation capabilities. Furthermore, the visual consistency across slices affirms that the model generalizes well across diverse anatomical presentations. Such interpretability, combined with confidence scores,

enhances the model's potential for clinical decision support by signaling both classification outcome and the associated certainty level.

Random Sample Predictions with Confidence Scores

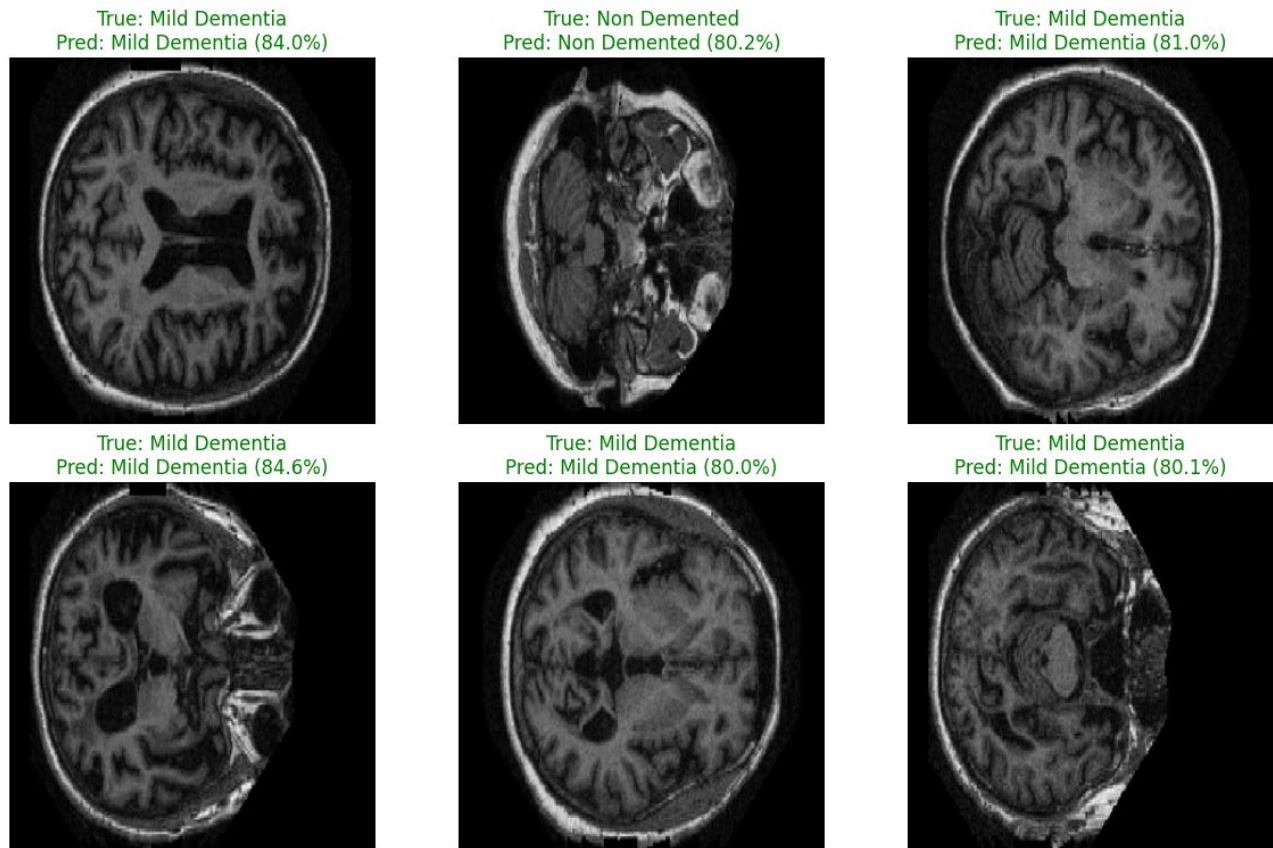


Figure 7. Sample MRI predictions with confidence scores.

5. Discussion of Results in Relation to Existing Literature

The results of this study demonstrate that the proposed optimized CNN ensemble model, which integrates EfficientNetB0 and ResNet50 architectures with class-balanced data augmentation, delivers exceptional performance in the multi-class classification of dementia. Achieving an overall accuracy of 99%, with F1-scores of 0.99 - 1.00 across all classes, the model significantly outperforms several prior approaches in similar domains.

These findings align with, yet also extend, earlier studies that employed machine learning for dementia diagnosis. For instance, Ahmed *et al.* [24] demonstrated a maximum accuracy of 96.77% using an SVM trained on full feature sets derived from clinical neuroimaging data. While impressive, their work primarily addressed binary classification and did not incorporate ensemble learning or address class imbalance. Similarly, Lin *et al.* [25] used neural networks and SVMs on non-imaging features and achieved the C-index being 0.79 (95% confidence interval = 0.75 - 0.83) and the corrected C-index for internal validation being 0.79, but

their approach lacked direct integration with structural MRI data, which is often more informative for disease staging.

In contrast, our model effectively combines imaging data with advanced ensemble learning, yielding not only higher accuracy but also robust differentiation between Non-Demented, Mild Demented, and Moderate Demented classes, an area often overlooked in previous literature. For example, Tan *et al.* [26] proposed an ensemble model that achieved an F1 score and AUC of 0.87 and 0.80, respectively. Accuracy (0.83), sensitivity (0.86), specificity (0.74), and predictive values (positive 0.88, negative 0.72) of the ensemble model were higher compared to the independent classifiers. However, their method lacked imaging inputs and suffered from lower class granularity.

The integration of a feature attention mechanism and soft voting in our ensemble further enhanced performance stability and interpretability. Our use of Gray Wolf Optimization (GWO) and class-balanced data augmentation resonates with the optimization approach proposed in the Enhanced Dementia Detection and Classification Model (EDCM) by Talaat *et al.* [15], which improved classification accuracy from 85% to 97%. Nonetheless, our work surpasses these results by demonstrating consistent near-perfect classification across three dementia stages, even for the underrepresented Moderate Dementia class, an achievement attributable to our targeted data augmentation strategy.

Furthermore, our model exhibits confidence-aware predictions, as evidenced by the softmax-based confidence analysis, echoing findings from Siafarikas *et al.* [27], who emphasized the prognostic value of combining neuropsychiatric scores with structural features. Unlike prior work, however, our study applies a CNN ensemble directly to augmented, preprocessed T1-weighted MRI images, providing an end-to-end deep learning pipeline with minimal reliance on handcrafted features. Our findings affirm the viability of CNN ensembles with targeted augmentation for real-world dementia screening and staging, and offer a path toward reliable, automated, and scalable neurodiagnostic systems.

5.1. Overfitting Risks and Mitigation

Given the high classification metrics achieved by the proposed ensemble model, including 99% accuracy and perfect AUC scores, it is essential to critically assess the risk of overfitting. Overfitting occurs when a model learns the training data too well, capturing noise or irrelevant patterns that reduce its ability to generalize to new, unseen data. This risk is particularly heightened in medical imaging tasks with imbalanced datasets or limited subject diversity, where the model may favor majority classes or memorize visual features rather than learning robust representations.

To proactively mitigate this risk, several regularization and validation strategies were incorporated into the model design and training process. First, we employed early stopping based on validation loss monitoring, which automatically halted training once performance on the validation set plateaued for 10 consecutive

epochs. This strategy prevented the model from continuing to optimize beyond the point of generalization, thus reducing the likelihood of overfitting. Second, dropout regularization with a rate of 0.5 was applied to the fully connected layers of the classification head. Dropout randomly deactivates neurons during training, forcing the network to learn redundant representations and discouraging reliance on any single feature path.

Additionally, the training process included learning rate scheduling via ReduceLROnPlateau, which decreased the learning rate when validation performance stagnated. This enabled finer weight updates in later epochs and helped avoid oscillations around local minima. Data augmentation further played a central role in promoting generalization by introducing variability in the training set through random transformations (e.g., rotations, flips, brightness shifts), particularly for the underrepresented Moderate Demented class.

To ensure that performance was not inflated due to random initialization or favorable splits, we conducted three-fold cross-validation and reported mean \pm standard deviation across folds. The ensemble model consistently delivered high performance with minimal variance (e.g., accuracy: $98.87\% \pm 0.11$), confirming that the results are not artifacts of overfitting.

Other regularization methods, such as L2 weight decay and batch normalization, were considered during preliminary experiments. However, dropout combined with early stopping and adaptive learning rate scheduling provided the best trade-off between model complexity and generalization performance without increasing computational overhead. Collectively, these techniques contribute to a stable, robust, and clinically viable model that generalizes well across varying data distributions.

5.2. Clinical Reliability and Computational Efficiency of the Proposed Model

The proposed optimized CNN ensemble model exhibits strong clinical reliability, making it a suitable candidate for real-world dementia screening and decision support systems. Achieving an overall classification accuracy of 99% across three diagnostic classes: Non-Demented, Mild Demented, and Moderate Demented, the model demonstrated exceptional performance in identifying early and moderate stages of dementia. This high accuracy, combined with F1-scores nearing 1.00 for each class, indicates a high degree of precision and recall, which is critical in clinical scenarios where diagnostic errors can lead to delayed interventions or inappropriate treatments. Notably, the model achieved perfect classification for the Moderate Dementia class, which is often the most challenging due to overlapping features with both early and late stages of cognitive decline. This reliability was further supported by a ROC-AUC score of 1.00 across all classes, confirming the model's strong discriminatory power.

From a computational standpoint, the model was designed for both robustness and efficiency. It leverages two lightweight yet powerful CNN backbones, Effi-

cientNetB0 and ResNet50, known for their favorable trade-off between accuracy and computational cost. These models were initialized with ImageNet-pretrained weights and fine-tuned specifically on the MRI dataset, reducing training time while enhancing domain-specific performance. Moreover, the integration of a soft voting mechanism and feature attention block not only improved classification stability but also contributed to interpretability, making the model's predictions more transparent for clinical practitioners. The training process was further optimized using early stopping, dropout regularization, and adaptive learning rate reduction, which prevented overfitting and accelerated convergence. These strategies ensured that the model could be trained efficiently on standard GPU-enabled environments, such as Google Colab with NVIDIA Tesla T4, without requiring high-end computational infrastructure.

5.3. Real-World Integration of the Proposed Model

The successful deployment of machine learning models in clinical settings requires more than high diagnostic accuracy; it demands robustness, scalability, interpretability, and compatibility with existing healthcare workflows. The proposed optimized CNN ensemble model aligns with these practical requirements, making it well-suited for real-world integration in dementia diagnosis and monitoring systems. By using T1-weighted MRI scans, which are widely available in clinical practice, the model leverages routine imaging data without requiring additional or costly modalities. This enhances its adaptability for use in both high-resource and resource-constrained environments. A key aspect supporting real-world integration is the model's ability to handle class imbalance, a common challenge in medical datasets where advanced-stage cases are underrepresented. Through class-balanced data augmentation techniques, the model was trained on a more equitable distribution of dementia categories, significantly improving its generalizability and fairness in decision-making. This ensures that the model does not disproportionately favor majority classes, a critical factor for clinical reliability, especially in early-stage diagnosis where timely interventions are most impactful. Furthermore, the model's architecture is based on modular and transferable components, EfficientNetB0 and ResNet50 backbones, both of which are supported in standard deep learning libraries (e.g., TensorFlow, Keras). This design choice simplifies integration into hospital IT systems, including Picture Archiving and Communication Systems (PACS) and electronic health record (EHR) platforms. Additionally, the model's inference speed and memory efficiency make it compatible with edge devices and low-power imaging platforms, potentially enabling deployment in rural clinics or mobile diagnostic units. Interpretability is also a critical feature for clinical adoption. The model incorporates a feature attention block that enhances transparency by highlighting the most salient spatial features contributing to classification outcomes. This facilitates clinician trust and supports decision justification, which are necessary for clinical auditing and regulatory approval. Moreover, the inclusion of confidence score histograms allows for

uncertainty quantification, offering clinicians a mechanism to flag low-confidence predictions for secondary review. In conclusion, the proposed ensemble model addresses both the technical and operational demands of real-world healthcare systems. Its high performance, data efficiency, modular architecture, and interpretability features collectively support its integration into diagnostic pipelines for early and multi-stage dementia detection, paving the way for more personalized and timely interventions.

6. Conclusion and Further Research

This study introduced an optimized CNN ensemble model combining Efficient-NetB0 and ResNet50, enhanced with a feature attention mechanism and trained on class-balanced, augmented MRI data for multi-class dementia classification. The model achieved outstanding quantitative performance, with an overall classification accuracy of 99%, precision scores of 1.00 for Mild and Moderate Dementia, and 0.98 for Non-Demented cases. The recall scores were 1.00 for Moderate Dementia, 1.00 for Non-Demented, and 0.98 for Mild Dementia, yielding macro and weighted average F1-scores of 0.99. The Area Under the ROC Curve (AUC) reached 1.00 for all three classes, confirming excellent class separability. These metrics strongly indicate the clinical reliability and computational robustness of the proposed architecture in differentiating between stages of cognitive decline using structural MRI scans.

Nevertheless, this work opens several avenues for future research. The model should be validated on multi-center datasets with greater demographic diversity to test generalizability. Incorporating additional modalities such as functional MRI, PET imaging, or cognitive test scores could further improve diagnostic precision. Moreover, extending the classification to include Very Mild and Severe Dementia stages could enhance the model's clinical utility. Future iterations should also incorporate explainability frameworks like Grad-CAM or SHAP to visualize the decision-making process and build clinician trust in AI-based diagnostics.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] World Health Organization (2021) Global Status Report on the Public Health Response to Dementia. <https://digitalcommons.fiu.edu/cgi/viewcontent.cgi?article=1962&context=srhreports>
- [2] Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., *et al.* (2020) Dementia Prevention, Intervention, and Care: 2020 Report of the Lancet Commission. *The Lancet*, **396**, 413-446. [https://doi.org/10.1016/s0140-6736\(20\)30367-6](https://doi.org/10.1016/s0140-6736(20)30367-6)
- [3] Arbabshirani, M.R., Plis, S., Sui, J. and Calhoun, V.D. (2017) Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls. *NeuroImage*, **145**,

- 137-165. <https://doi.org/10.1016/j.neuroimage.2016.02.079>
- [4] Lu, D., Popuri, K., Ding, W.G., et al. (2018) Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease Using Structural MR and FDG-PET Images. *Scientific Reports*, **8**, Article No. 5697.
 - [5] Wang, H., Feng, T., Zhao, Z., Bai, X., Han, G., Wang, J., et al. (2022) Classification of Alzheimer's Disease Based on Deep Learning of Brain Structural and Metabolic Data. *Frontiers in Aging Neuroscience*, **14**, Article 927217. <https://doi.org/10.3389/fnagi.2022.927217>
 - [6] Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., et al. (2021) Deep Learning Encodes Robust Discriminative Neuroimaging Representations to Outperform Standard Machine Learning. *Nature Communications*, **12**, Article No. 353. <https://doi.org/10.1038/s41467-020-20655-6>
 - [7] Sarmah, U., Borah, P. and Bhattacharyya, D.K. (2024) Ensemble Learning Methods: An Empirical Study. *SN Computer Science*, **5**, Article No. 924. <https://doi.org/10.1007/s42979-024-03252-y>
 - [8] Wang, Y., Liu, S., Spiteri, A.G., Huynh, A.L.H., Chu, C., Masters, C.L., et al. (2024) Understanding Machine Learning Applications in Dementia Research and Clinical Practice: A Review for Biomedical Scientists and Clinicians. *Alzheimer's Research & Therapy*, **16**, Article No. 175. <https://doi.org/10.1186/s13195-024-01540-6>
 - [9] Javeed, A., Dallora, A.L., Berglund, J.S., Ali, A., Ali, L. and Anderberg, P. (2023) Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *Journal of Medical Systems*, **47**, Article No. 17. <https://doi.org/10.1007/s10916-023-01906-7>
 - [10] Martin, S.A., Townend, F.J., Barkhof, F. and Cole, J.H. (2023) Interpretable Machine Learning for Dementia: A Systematic Review. *Alzheimer's & Dementia*, **19**, 2135-2149. <https://doi.org/10.1002/alz.12948>
 - [11] Kantayeva, G., Lima, J. and Pereira, A.I. (2023) Application of Machine Learning in Dementia Diagnosis: A Systematic Literature Review. *Heliyon*, **9**, e21626. <https://doi.org/10.1016/j.heliyon.2023.e21626>
 - [12] Dhakal, S., Azam, S., Hasib, K.M., Karim, A., Jonkman, M. and Haque, A.S.M.F.A. (2023) Dementia Prediction Using Machine Learning. *Procedia Computer Science*, **219**, 1297-1308. <https://doi.org/10.1016/j.procs.2023.01.414>
 - [13] Wang, H., Sheng, L., Xu, S., Jin, Y., Jin, X., Qiao, S., et al. (2022) Develop a Diagnostic Tool for Dementia Using Machine Learning and Non-Imaging Features. *Frontiers in Aging Neuroscience*, **14**, Article 945274. <https://doi.org/10.3389/fnagi.2022.945274>
 - [14] Zhu, F., Li, X., Tang, H., He, Z., Zhang, C., Hung, G., et al. (2020) Machine Learning for the Preliminary Diagnosis of Dementia. *Scientific Programming*, **2020**, 1-10. <https://doi.org/10.1155/2020/5629090>
 - [15] Talaat, F.M. and Ibraheem, M.R. (2024) Dementia Diagnosis in Young Adults: A Machine Learning and Optimization Approach. *Neural Computing and Applications*, **36**, 21451-21464. <https://doi.org/10.1007/s00521-024-10317-9>
 - [16] Zhang, J., Song, L., Miller, Z., Chan, K.C.G. and Huang, K. (2024) Machine Learning Models Identify Predictive Features of Patient Mortality across Dementia Types. *Communications Medicine*, **4**, Article No. 23. <https://doi.org/10.1038/s43856-024-00437-7>
 - [17] Gill, S., Mouches, P., Hu, S., Rajashekar, D., MacMaster, F.P., Smith, E.E., et al. (2020) Using Machine Learning to Predict Dementia from Neuropsychiatric Symptom and Neuroimaging Data. *Journal of Alzheimer's Disease*, **75**, 277-288.

- <https://doi.org/10.3233/jad-191169>
- [18] Camacho, M., Atehortúa, A., Wilkinson, T., Gkontra, P. and Lekadir, K. (2025) Low-Cost Predictive Models of Dementia Risk Using Machine Learning and Exposome Predictors. *Health and Technology*, **15**, 355-365. <https://doi.org/10.1007/s12553-024-00937-5>
 - [19] Huang, Y.C., Liu, T.C. and Lu, C.J. (2024) Establishing a Machine Learning Dementia Progression Prediction Model with Multiple Integrated Data. *BMC Medical Research Methodology*, **24**, Article No. 288. <https://doi.org/10.1186/s12874-024-02411-2>
 - [20] Rahman, A.U., Ali, S., Saqia, B., Halim, Z., Al-Khasawneh, M.A., AlHammadi, D.A., et al. (2025) Alzheimer's Disease Prediction Using 3D-CNNs: Intelligent Processing of Neuroimaging Data. *SLAS Technology*, **32**, Article 100265. <https://doi.org/10.1016/j.slant.2025.100265>
 - [21] Li, T.R., Li, B.L., Zhong, J., Xu, X., Wang, T. and Liu, F. (2024) A Prediction Model of Dementia Conversion for Mild Cognitive Impairment by Combining Plasma pTau181 and Structural Imaging Features. *CNS Neuroscience & Therapeutics*, **30**, e70051. <https://doi.org/10.1111/cns.70051>
 - [22] Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., et al. (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology*, **74**, 201-209. <https://doi.org/10.1212/wnl.0b013e3181cb3e25>
 - [23] Li, J. (2017) Assessing the Accuracy of Predictive Models for Numerical Data: Not R Nor R2, Why Not? Then What? *PLOS ONE*, **12**, e0183250. <https://doi.org/10.1371/journal.pone.0183250>
 - [24] Ahmed, R., Fahad, N., Miah, M.S.U., Hossen, M.J., Morol, M.K., Mahmud, M., et al. (2024) A Novel Integrated Logistic Regression Model Enhanced with Recursive Feature Elimination and Explainable Artificial Intelligence for Dementia Prediction. *Healthcare Analytics*, **6**, Article 100362. <https://doi.org/10.1016/j.health.2024.100362>
 - [25] Liu, K., Hou, T., Li, Y., Tian, N., Ren, Y., Liu, C., et al. (2025) Development and Internal Validation of a Risk Prediction Model for Dementia in a Rural Older Population in China. *Alzheimer's & Dementia*, **21**, e14617. <https://doi.org/10.1002/alz.14617>
 - [26] Tan, W.Y., Hargreaves, C., Chen, C. and Hilal, S. (2023) A Machine Learning Approach for Early Diagnosis of Cognitive Impairment Using Population-Based Data. *Journal of Alzheimer's Disease*, **91**, 449-461. <https://doi.org/10.3233/jad-220776>
 - [27] Siafarikas, N., Alnæs, D., Monereo-Sanchez, J., Lund, M.J., Selbaek, G., Stylianou-Korsnes, M., et al. (2021) Neuropsychiatric Symptoms and Brain Morphology in Patients with Mild Cognitive Impairment and Alzheimer's Disease with Dementia. *International Psychogeriatrics*, **33**, 1217-1228. <https://doi.org/10.1017/s1041610221000934>