ACADEMIA International Journal for Social Sciences
Volume 5, Issue 3(s6), 2026 ISSN-L (Online): 3006-6638

A Hybrid CNN Ensemble Approach for Intelligent Alzheimer's Disease Detection in MRI Imaging

Rubab Haider
rubabhaiderawan@gmail.com
Department of Computer Science, The University of Faisalabad, Pakistan

Benish William
komalbenishwilliam@gmail.com
Department of Computer Science, The University of Faisalabad, Pakistan

Munazzah Munwer
Department of Computer Science, The University of Faisalabad, Pakistan

Abdul Rauf
abdulrauf2000.pk@gmail.com
Department of Computer Science, The University of Faisalabad, Pakistan

Majid Hussain
majidhussain1976@gmail.com
Department of Computer Science, The University of Faisalabad, Pakistan

Corresponding Author: Majid Hussain majidhussain1976@gmail.com

Received: 18-01-2026 **Revised:** 01-02-2026 **Accepted:** 16-02-2026 **Published:** 02-03-2026

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative condition, and diagnosing it early and accurately is critical for better clinical care. This study introduces a deep learning framework for multiclass AD detection using brain MRI images. The model classifies patients into four groups: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The approach combines standardized MRI preprocessing, oversampling of the training set to address class imbalance, and transfer learning based on an EfficientNetV2S backbone with a custom classification head. Using a dataset of 6,400 MRI images, the model was trained in two stages to strengthen task-specific feature learning.When tested on a separate set of 640 images, the framework reached 93% accuracy, with a macro precision of 0.80, macro recall of 0.85, and macro F1-score of 0.81. AUC values were above 0.90 across all classes. The model showed strong sensitivity in identifying early and intermediate stages of the disease. Most errors appeared between neighboring classes, which aligns with the gradual progression of Alzheimer's. These results suggest that the framework is effective at detecting clinically meaningful MRI patterns and has strong potential as an AI-assisted decision-support system for Alzheimer's diagnosis. The study still has a few limitations. Evaluation was carried out on only one dataset, the Moderate Demented class had a limited number of samples, and external validation was not included. Even with these constraints, the findings show that deep learning-based MRI analysis holds strong promise for making Alzheimer's detection more reliable and scalable.

Keywords: *Hybrid CNN***,** *Alzheimer's disease, MRI images*

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia and one of the most serious neurodegenerative conditions affecting older adults across the world. Its impact keeps growing. According

to the World Health Organization, 57 million people were living with dementia in 2021, and nearly 10 million new cases are diagnosed each year. Since the biological changes linked to Alzheimer's often begin years before major cognitive symptoms appear, early and accurate diagnosis plays a key role in identifying risk, starting treatment sooner, and improving long-term care planning.

Among the imaging tools used to study Alzheimer's, structural magnetic resonance imaging (MRI) stands out because it is noninvasive, widely available in clinical settings, and effective for detecting brain atrophy linked to the disease. These changes are often seen in areas such as the hippocampus, the medial temporal lobe, and related cortical regions. Because of this, structural MRI has become an important biomarker for tracking disease stage and progression. In research, the Alzheimer's Disease Neuroimaging Initiative (ADNI) has been especially valuable, offering a large longitudinal multicenter dataset for building and testing computational models for Alzheimer's diagnosis.

Deep learning, especially convolutional neural networks (CNNs), has improved MRI-based Alzheimer's analysis by reducing reliance on hand-designed features and learning subtle disease patterns directly from brain scans. Earlier studies showed that CNN-based models can distinguish between AD, mild cognitive impairment (MCI), and cognitively normal individuals from a single structural MRI scan. A 2024 meta-analysis also reported strong overall performance for MRI-based deep learning models, with a pooled sensitivity of 0.84, specificity of 0.86, and an AUROC of 0.92. These findings show the growing value of deep learning for detecting both AD and MCI.

Even with this progress, several problems still limit real clinical use. MRI-based CNN studies are often hard to compare because they differ in cohort selection, preprocessing steps, validation methods, and reporting practices. Wen et al. also pointed out that many published studies may include data leakage or weak validation procedures, which leads to overly optimistic results. On top of that, models trained on one dataset often perform poorly on another with different demographics or inclusion rules. Recent reviews continue to point to the same concerns, especially dataset heterogeneity, limited reproducibility, and the need for AI pipelines that are more transparent and dependable for Alzheimer's prediction.

For this reason, a hybrid CNN ensemble approach offers a strong path forward for MRI-based Alzheimer's detection. By combining multiple CNN backbones or complementary feature extractors, ensemble models reduce variance, improve robustness, and capture both local and global structural changes better than a single network. When paired with interpretability tools such as Grad-CAM, this type of framework also gives clinicians clearer insight into which brain regions influence the final prediction. Based on this motivation, the present study proposes a hybrid CNN ensemble framework for MRI-based Alzheimer's detection with the goal of improving diagnostic accuracy, strengthening generalization across disease stages, and providing more interpretable decision support in clinical practice.

LITERATURE REVIEW

Early work on Alzheimer's diagnosis from MRI focused on hand-crafted imaging features, including hippocampal volume, cortical thickness, voxel-based morphometry, and region-of-interest measures. Researchers then fed those features into traditional machine learning models such as support vector machines, k-nearest neighbors, and random forests. These studies helped show that MRI carries structural patterns linked to Alzheimer's disease. Still, the approach had clear limits. It depended on heavy preprocessing, manual feature design, and precise anatomical segmentation, which made it harder to scale and harder to apply consistently across different cohorts.

More recent reviews paint a similar picture. Wang et al.'s 2024 meta-analysis found that MRI-based deep learning shows strong overall performance for classifying AD and MCI, but the study also pointed to major

variation in study design, MRI sequence selection, and overall method quality. Givian and Calbimonte's 2024 review, along with Kaur et al.'s 2024 systematic review, shows how fast the field has grown across both machine learning and deep learning. At the same time, the same weaknesses keep appearing: small datasets, class imbalance, inconsistent preprocessing pipelines, limited external validation, and poor reporting of outcomes that matter in clinical settings. In other words, better accuracy isn't enough on its own. Reliability and reproducibility matter just as much.

Because of the limits of single-network models, many researchers have turned to hybrid and ensemble systems. Recent MRI-based studies often combine multiple CNN branches, transfer-learning backbones, or added classifiers to capture different aspects of disease-related brain structure. Zolfaghari et al., for instance, introduced a dual-CNN system paired with an ensemble classifier for Alzheimer's staging from MRI.

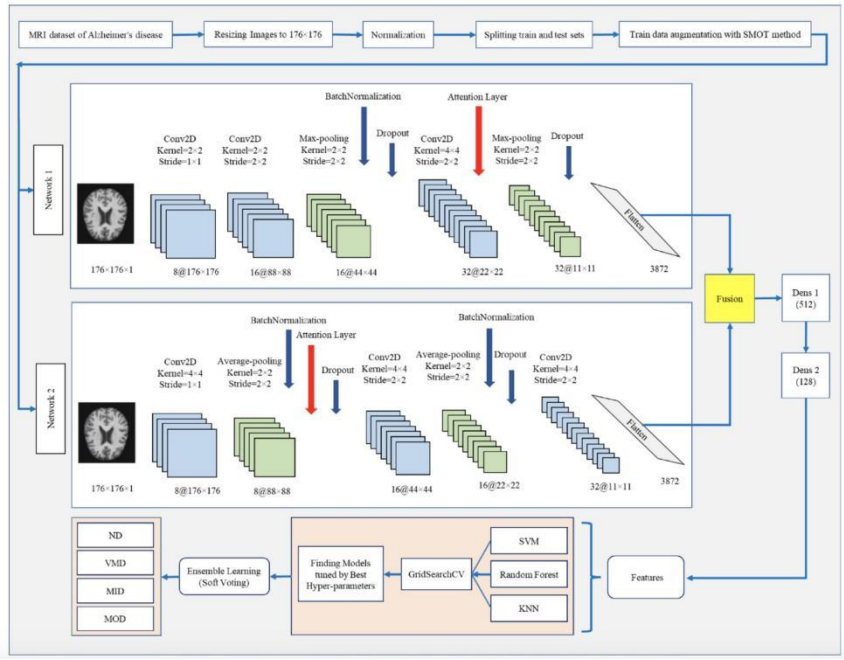

Figure 1: The framework of the proposed method, including pre-processing, feature extraction via a parallel CNN, and classification of AD stages using an ensemble learning method consisting of SVM, RF, and KNN classifiers optimized by GridSearchCV.

Sriram et al. combined InceptionResNetV2, ResNet50, and a custom CNN with weighted averaging, and they reported stronger results than any of the individual models alone. These studies support the view that ensemble learning can strengthen robustness and multistage discrimination, especially when MRI findings are subtle and class boundaries overlap.

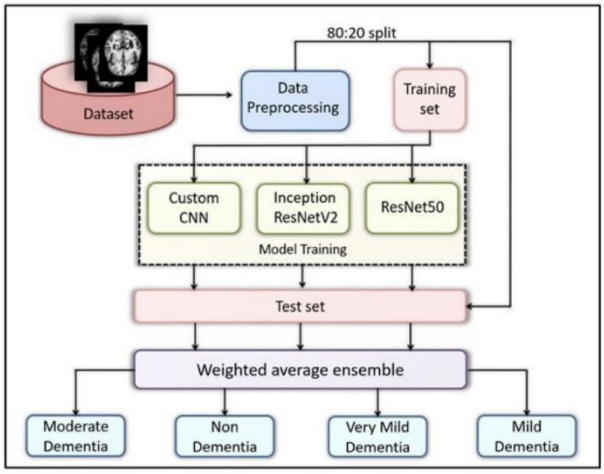

Figure 2: Model Architect

Custom trained CNN

This study uses a CNN model for image classification. The architecture begins with a fixed-size input image of 176 × 176 × 3 and passes it through several convolutional blocks. Each block contains convolutional layers that extract and learn image features, followed by a max-pooling layer that reduces the spatial dimensions by half. This design helps the model capture important visual patterns efficiently for classification by using Eq.1.

$$(\boldsymbol{I} * \boldsymbol{K})\,(\boldsymbol{i},\boldsymbol{j}) \;=\; \sum_{\boldsymbol{m}}\sum_{\boldsymbol{n}} \boldsymbol{I}\,(\boldsymbol{m},\boldsymbol{n}).\boldsymbol{K}\,(\boldsymbol{i}-\boldsymbol{m},\boldsymbol{j}-\boldsymbol{n}) \qquad \boldsymbol{Eq.1}$$

where I, K ,i, j are the input image, kernel and positions on the output feature map,respectively. Following the convolutional layers, batch normalization and pooling are applied to increase the model's performance. The max pooling operation is done using the given formula Eq.2.

$$\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j}) \;= \boldsymbol{max}_{\boldsymbol{m},\boldsymbol{n}}(\boldsymbol{I}(\boldsymbol{i}+\boldsymbol{m},\boldsymbol{j}+\boldsymbol{n}) \qquad \boldsymbol{Eq.2}$$

where the output is the maximum value from input regions defined by filter size.

The custom CNN is specifically designed for Alzheimer's disease detection from MRI images. It uses convolutional blocks with batch normalization, dropout, and max-pooling to learn hierarchical features while reducing overfitting. After feature extraction, a flatten layer and dense layers capture higher-level patterns, and a softmax layer produces multiclass predictions. Unlike general pre-trained models such as InceptionResNetV2 and ResNet50, this custom CNN focuses on Alzheimer's-related biomarkers, including hippocampal shrinkage and cortical atrophy. Its early layers learn low-level features such as edges, textures, and shapes, helping it detect subtle structural brain changes.

This specialization improves its effectiveness on MRI data, especially with smaller datasets. The model also showed stronger performance in severe dementia classification, achieving an F1-score of 0.91, compared with 0.76 for InceptionResNetV2 and 0.83 for ResNet50. Combined with the complementary

strengths of the pre-trained models, the custom CNN helps create a more balanced and accurate diagnostic system for detecting all stages of Alzheimer's disease.

InceptionResNetV2

InceptionResNetV2 is a convolutional neural network that incorporates both the idea of an inception and residual connections. InceptionResNetv2 is a neural network architecture designed for performing convolutional operations. It gained knowledge from a massive repository of several millions of photos found in the ImageNet database. It was a 164-layer network capable of correctly classifying images into an exact representation comprising 1000 different characteristics.

The input image fed into the network is 299 x 299 pixels, while on the output side, the network returns a list of the predicted probabilities over all classes. During creation, it merges the initial structure with the residual connection. Following the input layer is the stem block, which reduces the spatial dimensions of the image and correspondingly increases its depth. The InceptionResNet block combines convolutional filters of different sizes and merges them using residual connections. This strategy greatly reduces the training time required. Following this block, there is an average pooling layer that decreases the dimensionality of each feature map to a single layer. The Average Pooling computation is done using Eq. 3.

$$\boldsymbol{P(i,j)} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{m,n}} \boldsymbol{I}\,(\boldsymbol{i+m, j+n}) \qquad \boldsymbol{Eq.3}$$

which calculates average of the elements in input region, reducing spatial dimensions. This is followed by a fully connected layer, which is a dense layer with softmax activation which gives the output of the classification probabilities of each class. The softmax activation function is calculated as given in Eq. (4)

$$\mathbf{softmax}(\boldsymbol{z_i}) = \frac{\boldsymbol{e^{z_i}}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{e^{z_j}}} \qquad \boldsymbol{Eq.4}$$

Where z_i is the i-th element of input vector z. InceptionResNetv2 is distinguished by its ability to process large images quickly while retaining the fine-grained information required for classification. The use of residual connections, which allow information to be bypassed via shortcut connections, reduces the vanishing gradient problem and speeds up training convergence.

ResNet50

ResNet-50 is an already trained deep learning model on the CNN architecture for processing visual input. The ResNet-50 model has a total of 50 layers. This model was trained with one million photographs from the ImageNet database, which has 1000 categories of objects. ResNet50 is a bit different from other pre-trained models such as AlexNet, GoogleNet, and VGG19, due to its great generalization ability and minimum error rates in recognition tests. The system will work on the exact classification of photos into 1000 categories, including keyboards, mice, pencils, and animals, using a 224 × 224-pixel input. It starts with the convolutional layer with many filters in the ResNet50 model, followed by a max pooling that directly comes after the input layer. The residual block contains a sequence of convolutional layers and shortcut connections. After residual blocks, the global average pooling layer decreases each feature map to individual layer, succeeded by the fully connected layer. The equation for calculating global average pooling is as given in Eq. 5.

$$P = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} I(i,j) \qquad Eq.5$$

Researchers wanting to improve the efficiency of ResNet-50 may benefit from investigating various fine-tuning processes, such as changing top layers, tweaking learning rates, and applying data augmentation approaches.

Another strong trend in recent research is the push toward explainable AI, or XAI, in MRI-based dementia diagnosis. CNNs often face criticism for functioning like black boxes, so researchers are giving more attention to tools that make model decisions easier to understand. Methods such as Grad-CAM, occlusion sensitivity analysis, SHAP, and saliency maps are now widely used to highlight the brain regions driving a model's prediction.

Chattopadhyay et al. found that Grad-CAM and similar techniques were able to point to patterns aligned with known Alzheimer's-related brain abnormalities. Recent review studies also stress the same point: interpretability matters for clinical trust, accountability, and real-world use. Seen from this angle, the main gap in the literature is not a lack of accurate models. The bigger issue is the limited number of frameworks that combine strong performance, solid generalization, multistage classification, and clinically useful interpretability in one system. That gap is what motivates the proposed hybrid CNN ensemble for intelligent Alzheimer's disease detection from MRI scans.

METHODOLOGY

This study presents an MRI-based deep learning framework for multiclass Alzheimer's disease classification through transfer learning. The system is built to classify brain MRI scans into four diagnostic groups: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented, as shown in the figure. The full workflow includes dataset preparation, image preprocessing, class balancing, transfer-learning-based model development, two-phase training, and performance evaluation. The overall aim is to strengthen classification performance while dealing with common challenges in medical imaging, including limited data, class imbalance, and overfitting. For this work, the Alzheimer MRI 4 Classes Dataset has used as shown in Figure 3.

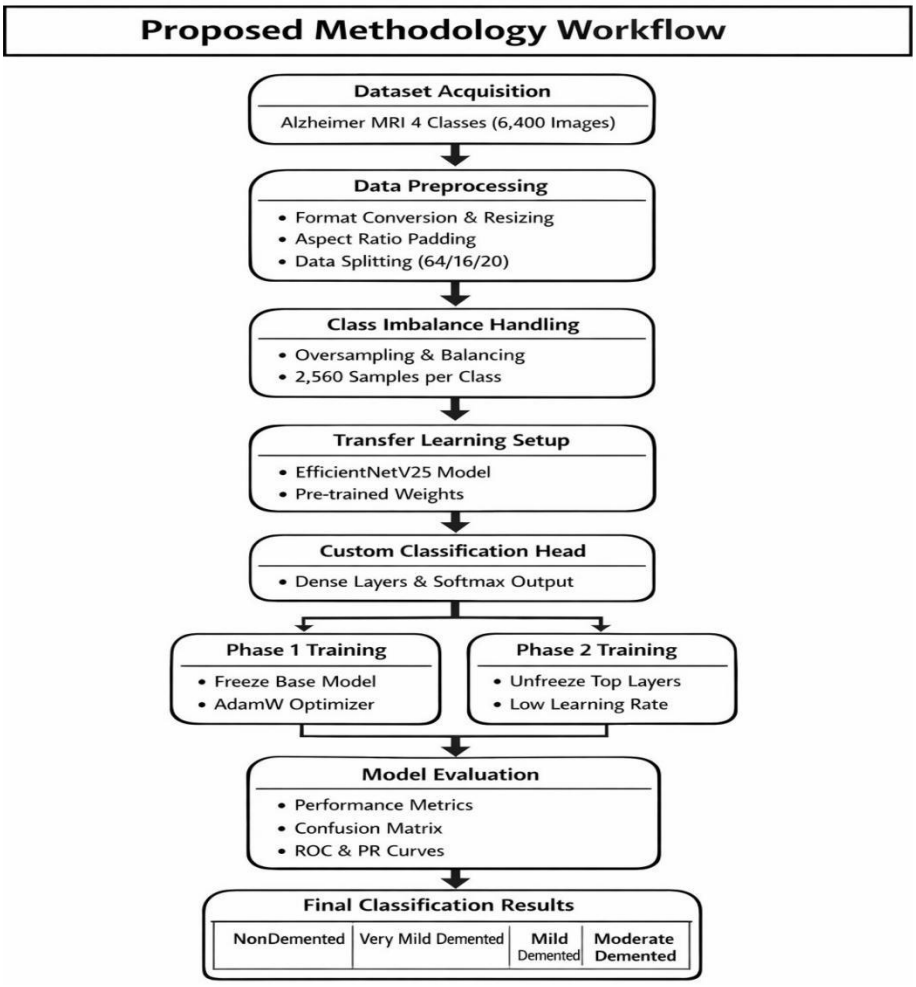

Figure 3: Proposed methodology workflow for MRI-based Alzheimer's disease classification.

The dataset contains 6,400 MRI images across four classes representing different stages of disease progression. Based on the manuscript, the class distribution includes 3,200 Non-Demented images, 2,240 Very Mild Demented images, 896 Mild Demented images, and only 64 Moderate Demented images. This uneven distribution shows a serious class imbalance, especially in the Moderate Demented category, and that issue shaped the preprocessing steps and training strategy used in the study.

All MRI images undergo a standardized preprocessing pipeline before training. First, each image is loaded and decoded from storage. To ensure compatibility with the deep learning model, all images are resized to 208 × 208 pixels while preserving the original aspect ratio through padding where necessary. Because the selected pre-trained model expects three-channel input, grayscale MRI scans are converted to RGB format (208 × 208 × 3). Pixel values are then normalized using the preprocessing function associated with the EfficientNetV2S model so that the input distribution matches the ImageNet-based pre-trained weights. This preprocessing stage ensures input consistency and reduces unnecessary distortion that could affect diagnostic features.

This paper reports stratified splitting so that each subset preserves the original class proportions. Based on the class-count tables, the working split appears to be 5,120 training images, 640 validation images, and

640 testing images. Since the dataset is heavily imbalanced, especially in the minority Moderate Demented class, the training set is balanced through random oversampling. Specifically, all minority classes are replicated until each training class reaches 2,560 samples, resulting in a balanced training set of 10,240 images. This strategy aims to reduce the model's bias toward majority classes and improve sensitivity for the less represented stages of Alzheimer's disease.

To make training more efficient, the study uses several data pipeline optimizations. The training samples are shuffled at every epoch so the model does not pick up misleading patterns from data order. Images are loaded in mini-batches of 64, and prefetching keeps data preparation running alongside GPU computation. Image processing is also parallelized to increase throughput and limit I/O delays. Together, these steps improve hardware efficiency and reduce training time.

The proposed classifier is based on EfficientNetV2S, a pre-trained convolutional neural network first developed on the ImageNet dataset. Transfer learning is chosen because annotated medical datasets are often small, and pre-trained networks already capture useful low-level visual patterns such as edges, textures, and shapes. In this architecture, EfficientNetV2S acts as the backbone feature extractor, using an input size of 208 × 208 × 3, about 20.3 million parameters, and a 1280-dimensional feature output.

The full model includes an input layer, an EfficientNetV2S preprocessing layer, the EfficientNetV2S backbone, a Global Average Pooling layer, a Dropout layer with a rate of 0.3, and a final Dense layer with four softmax outputs for the four Alzheimer's classes. In the first stage of training, the EfficientNetV2S backbone is frozen, so only the classification head is updated. This setup helps limit overfitting and gives the model time to adapt to MRI-specific patterns in a controlled way.

Training follows a two-phase procedure. In Phase 1, the backbone remains frozen and only the newly added classification head is trained. This stage uses the AdamW optimizer with a learning rate of 3×10^{-4}, a weight decay of 1×10^{-4}, a batch size of 64, and training for up to 70 epochs. In Phase 2, selected upper layers of the backbone are unfrozen for fine-tuning, and the learning rate is reduced to 1×10^{-5}. This lower rate helps prevent catastrophic forgetting while still allowing the network to adapt to Alzheimer-related MRI features. At this point, the number of trainable parameters rises from 5,124 in the classification head to about 5.2 million.

The model uses sparse categorical cross-entropy loss, which fits integer-labeled multiclass classification tasks well. To improve generalization and limit overfitting, three callbacks are included in training: Model Checkpoint to save the best-performing weights, Early Stopping with a patience of five epochs, and ReduceLROnPlateau, which cuts the learning rate in half when validation loss does not improve for two consecutive epochs, with a minimum learning rate of 1×10^{-6}.

The manuscript reports that the implementation was developed in TensorFlow/Keras with mixed precision training enabled to reduce memory use and speed up computation. Mixed precision relies on 16-bit floating-point operations where appropriate, while loss scaling preserves numerical stability. The reported setup includes TensorFlow 2.19.0, GPU hardware compatible with H100 or A100, a batch size of 64, and a fixed random seed of 42 to support reproducibility.

Model performance is assessed through several classification metrics, including accuracy, precision, recall, F1-score, confusion matrix, ROC-AUC, and precision-recall AUC, as presented in Table 3.1. The manuscript also uses test-time augmentation during evaluation. In this step, slightly transformed versions of each test image are generated, and the prediction probabilities are averaged to produce a more stable final output. This reduces prediction variance and strengthens robustness against small input changes.

Table 1. Evaluation metrics used in the study.

Evaluation Category	Metrics
Classification Metrics	Accuracy, Precision, Recall, F1-score
Error Analysis	Confusion Matrix
Discrimination Analysis	ROC Curve, AUC
Imbalance-Sensitive Analysis	Precision–Recall Curve
Training Behaviour	Training/Validation Accuracy and Loss Curves
Robustness	Test-Time Augmentation (TTA)

My work proposes a multiclass Alzheimer's disease detection framework based on MRI images, standardized image preprocessing, oversampling-based class balancing, EfficientNetV2S transfer learning, two-phase fine-tuning, and comprehensive metric-based evaluation. The method is designed to address the major challenges of limited data, class imbalance, and overfitting in medical image classification.

RESULTS AND DISCUSSION

The proposed MRI-based Alzheimer's disease classification framework was evaluated on a held-out test set of 640 images after two-phase training. The dataset remained highly imbalanced prior to oversampling, particularly for the Moderate Demented class, which had only 51 training samples compared with 2,560 Non-Demented samples. To reduce this imbalance, oversampling was applied only to the training set, increasing each class to 2,560 samples and producing a balanced training set of 10,240 images. The training process was conducted over 140 epochs in two phases.

During the first phase, when only the classifier head was trained, training accuracy rose from about 33% to 62%, while validation accuracy leveled off at around 60 to 62%. In the second phase, after unfreezing the upper layers of the EfficientNetV2S backbone, performance improved sharply. Training accuracy reached 97%, and validation accuracy increased to 94%. The loss curves followed the same pattern. In Phase 1, loss declined steadily from roughly 1.36 to 0.88. In Phase 2, it dropped quickly to below 0.1. Validation loss stayed slightly above training loss, but both curves settled over time, which points to model convergence with only mild overfitting.

On the test set, the model reached an overall accuracy of 93%, as reported in Table 4.1, with a macro precision of 0.80, macro recall of 0.85, and macro F1-score of 0.81. These results show strong multiclass discrimination, especially given the difficulty of the task, which involved four clinically related classes and substantial class imbalance. The manuscript reports 486 correctly classified samples in one table and 487 in another section. Even so, both figures reflect the same overall test accuracy of about 93%.

Table 2. Overall model performance on the test set

Metric	Value
Accuracy	93%
Precision (Macro Average)	0.80
Recall (Macro Average)	0.85
F1-Score (Macro Average)	0.81
Total Test Samples	640
Correctly Classified	486
Misclassified	154

Class-wise results give a clearer picture of model behavior, as shown in Table 2. For the Non-Demented class, the model achieved a precision of 0.93, recall of 0.67, and F1-score of 0.78. For Very Mild Demented, it reached 0.71 precision, 0.82 recall, and 0.76 F1-score. For Mild Demented, the scores were 0.56 precision, 0.91 recall, and 0.69 F1-score. The Moderate Demented class achieved perfect scores of 1.00 for precision, recall, and F1-score, though this result came from only six test images. Taken together, these values show that the model performed especially well in identifying disease-positive cases, particularly in the Mild and Moderate stages.

Table 3. Class-wise performance metrics.

Class	Precision	Recall	F1-Score	Support
Non-Demented	0.93	0.67	0.78	320
Very Mild Demented	0.71	0.82	0.76	224
Mild Demented	0.56	0.91	0.69	90
Moderate Demented	1.00	1.00	1.00	6
Macro Average	0.80	0.85	0.81	640

The confusion matrix, shown in Table 3, supports this pattern. Among Non-Demented cases, 67% were classified correctly, while 28% were misclassified as Very Mild Demented. For the Very Mild Demented class, 82% were identified correctly, with 8% misclassified as Non-Demented and 10% as Mild Demented. The Moderate Demented class was classified perfectly, with no confusion across other stages. The manuscript also reports no skip errors, meaning the model did not confuse Non-Demented cases with Moderate Demented cases. This suggests that the model learned the progression of disease in a clinically sensible way.

Table 4. Summary of confusion matrix findings.

Class	Correct Classification Rate	Main Misclassification Pattern	Interpretation
Non-Demented	67%	28% misclassified as Very Mild Demented	Conservative bias toward possible disease detection
Very Mild Demented	82%	8% as Non-Demented; 10% as Mild Demented	Early stage remains challenging but reasonably separated
Mild Demented	91%	9% as Very Mild Demented	Strong disease-stage recognition
Moderate Demented	100%	No confusion with other classes	Perfect classification, but based on very small sample size
Overall pattern	-	Errors mostly between adjacent stages	Clinically reasonable disease progression behavior
Non-Demented	67%	28% misclassified as Very Mild Demented	Conservative bias toward possible disease detection

ROC analysis also showed strong discriminative performance across all four classes. The reported AUC values were 0.93 for Non-Demented, 0.91 for Very Mild Demented, 0.96 for Mild Demented, and 1.00 for Moderate Demented. Precision-recall analysis showed the same trend, with average precision scores of 0.93, 0.83, 0.87, and 1.00, respectively. These findings show that the classifier maintained strong ranking performance across different decision thresholds, including in minority classes.

The mean prediction confidence was reported as 0.82, the median confidence as 0.89, and nearly 48.8% (Table 4.5) of all predictions were made with confidence greater than 0.90. Only 6.6% of predictions had confidence below 0.50, suggesting that the model was uncertain only in a relatively small number of cases. The manuscript also states that test-time augmentation (TTA) improved performance and increased macro F1-score to 0.83, although the numeric values in the TTA table should be checked because one entry appears inconsistent with the stated gain.

Table 5. Test-time augmentation (TTA) results.

Metric	Standard	With TTA	Reported Improvement
Accuracy	93%	78%	+2.6%
Macro F1	0.81	0.83	+2.5%

The experimental findings indicate that the proposed framework performs well for multiclass Alzheimer's disease classification from MRI images. An overall accuracy of 93% together with a macro F1-score of 0.81 suggests that the EfficientNetV2S-based transfer-learning pipeline successfully extracted diagnostically useful features from brain MRI scans despite the complexity of the four-class problem. The strong improvement observed after backbone fine-tuning also shows that transfer learning from ImageNet was beneficial, but adaptation of higher-level features to Alzheimer-related structures was still necessary for best performance.

Table 6. Prediction confidence analysis.

Confidence Metric	**Value**
Mean Confidence	0.82
Median Confidence	0.89
High Confidence Predictions (> 0.90)	48.8%
Low Confidence Predictions (< 0.50)	6.6%

The model showed strong sensitivity to disease-positive cases, with recall values of 0.82 for Very Mild Demented and 0.91 for Mild Demented, indicating effective detection of early and intermediate Alzheimer-related changes. Alzheimer's disease usually develops step by step, so this pattern makes clinical sense. MRI differences between nearby stages are often subtle, which explains why most errors happened between adjacent classes. The fact that the model did not make skip-stage errors is encouraging because it suggests it learned the progression of the disease in a meaningful way. The perfect performance in the Moderate Demented class also looks promising, but it needs to be treated with caution. That result came from only 6 test samples and 51 original training images before oversampling, so it may not hold up as well on outside datasets or in real clinical settings.

The ROC curves, precision-recall results, and prediction confidence also point to a reliable model. AUC values above 0.90 show strong separation between classes, and the confidence scores suggest most predictions were made with solid certainty. Test-time augmentation seems to add another layer of robustness, though the TTA table should be fixed before publication. Even with these strengths, the study still has clear limitations. It relies on a single dataset, does not include external validation, offers limited interpretability, and includes a very small Moderate Demented class. The slight gap between the training and validation curves also suggests some overfitting. Future work should use multi-center datasets, include explainable AI methods, and test the model on independent data.

CONCLUSION

This study presents a deep learning framework for detecting Alzheimer's disease from MRI images across four stages: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. By combining standardized preprocessing, oversampling, and transfer learning, the model handled class imbalance and the subtle differences between stages well. It reached 93% test accuracy, a macro F1-score of 0.81, and AUC values above 0.90 for every class. These results suggest the model is effective at identifying MRI patterns linked to clinically relevant stages of the disease, especially in the early and middle stages.

Most of the mistakes appeared between neighboring stages, which fits the gradual progression of Alzheimer's disease and supports the idea that the model learned stage-related patterns in a clinically sensible way. Even though the framework showed strong performance and looks promising as a decision-support tool, the findings are still limited by the use of one dataset, the small Moderate Demented sample, the lack of outside validation, and limited interpretability. Overall, the study shows the potential of AI-assisted MRI analysis for earlier and more dependable Alzheimer's diagnosis, while also making it clear that larger multi-center studies and better explainability are needed before clinical use.

LITERATURE CITED

Alok N, Krishan K and Chauhan P (2021) Deep learning-based image classifier for malaria cell detection. Machine Learning for Healthcare Applications 12: 187-197.

Ansari H, Vijayvergia A and Kumar K (2018) DCR-HMM: Depression Detection Based on Content Rating Using Hidden Markov Model. In: 2018 Conference on Information and Communication Technology (CICT). IEEE: 1-6.

Ansart M, Epelbaum S, Bassignana G et al., (2020) Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. Medical Image Analysis 67: 101848.

Basaia S, Agosta F, Wagner L et al., (2019) Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage: Clinical 21: 101645.

Bateman RJ, Aisen PS, De Strooper B et al., (2011) Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. Alzheimer's Research and Therapy 3(1): 1-13.

Battineni G, Chintalapudi N and Amenta F (2019) Machine learning in medicine: performance calculation of dementia prediction by support vector machines (SVM). Informatics in Medicine Unlocked 16: 100200.

Bharati S, Rahman MA and Podder P (2018) Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In: 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). IEEE: 581-584.

Bharati S, Podder P and Mondal MRH (2020) Hybrid deep learning for detecting lung diseases from X-ray images. Informatics in Medicine Unlocked 20: 100391.

Bharati S, Podder P, Mondal MRH and Prasath VBS (2021) Medical imaging with deep learning for COVID-19 diagnosis: a comprehensive review. International Journal of Computational Intelligence Systems 13: 91-112.

Bharati S, Podder P, Mondal MRH and Prasath VBS (2021) CO-ResNet: optimized ResNet model for COVID-19 diagnosis from X-ray images. International Journal of Hybrid Intelligent Systems 17(1-2): 71-85.

Boser BE, Guyon IM and Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory: 144-152.

Braak H, and Braak E (1997) Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiology of Aging 18(4): 351-357.

Brickman AM, Honig LS, Scarmeas N et al., (2008) Measuring cerebral atrophy and white matter hyperintensity burden to predict the rate of cognitive decline in Alzheimer disease. Archives of Neurology 65(9): 1202-1208.

Cao J, Kwong S, Wang R et al., (2015) Class-specific soft voting based multiple extreme learning machines ensemble. Neurocomputing 149: 275-284.

Castellazzi G, Cuzzoni MG, Cotta Ramusino M et al. (2020) A machine learning approach for the differential diagnosis of Alzheimer and vascular dementia fed by MRI selected features. Frontiers in Neuroinformatics 14: 25.

Chen T, and Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785-794.

Chen R, and Herskovits EH (2010) Machine-learning techniques for building a diagnostic model for very mild dementia. NeuroImage 52(1): 234-244.

Cuijpers Y, and Van Lente H (2015) Early diagnostics and Alzheimer's disease: beyond 'cure' and 'care'. Technological Forecasting and Social Change 93: 54-67.

Dabral I, Singh M and Kumar K (2019) Cancer detection using convolutional neural network. In: International Conference on Deep Learning, Artificial Intelligence and Robotics. Springer, Cham: 290-298.

Darbari A, Kumar K, Darbari S and Patil PL (2021) Requirement of artificial intelligence technology awareness for thoracic surgeons. The Cardiothoracic Surgeon 29(1): 1-13.

Datta P, Shankle W and Pazzani M (1996) Applying machine learning to an Alzheimer's database. In: AAAI Spring Symposium on AI in Medicine, Stanford, USA: 25-27.

Davatzikos C, Resnick SM, Wu X et al. (2008) Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. NeuroImage 41(4): 1220-1227.

Delgado J, and Ishii N (1999) Memory-based weighted majority prediction. In: SIGIR Workshop on Recommender Systems. Citeseer: 85.

Den Heijer T, Geerlings MI, Hoebeek FE et al. (2006) Use of hippocampal and amygdalar volumes on magnetic resonance imaging to predict dementia in cognitively intact elderly people. Archives of General Psychiatry 63(1): 57-62.

Facal D, Valladares-Rodriguez S, Lojo-Seoane C et al. (2019) Machine learning approaches to studying the role of cognitive reserve in conversion from mild cognitive impairment to dementia. International Journal of Geriatric Psychiatry 34(7): 941-949.

Farid AA, Selim GI and Khater HAA (2020) Applying artificial intelligence techniques to improve clinical diagnosis of Alzheimer's disease. In: 9th International Conference on Research in Science and Technology, Berlin, Germany.

Filipovych R, and Davatzikos C (2011) Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). NeuroImage 55(3): 1109-1119.

Fox NC, Cousens S, Scahill R et al. (2000) Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: power calculations and estimates of sample size to detect treatment effects. Archives of Neurology 57(3): 339-344.

Friedman JH, (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics 29: 1189-1232.

Frozza RL, Lourenço MV and De Felice FG (2018) Challenges for Alzheimer's disease therapy: insights from novel mechanisms beyond memory defects. Frontiers in Neuroscience 12: 37.

Gill S, Mouches P, Hu S et al. (2020) Using machine learning to predict dementia from neuropsychiatric symptom and neuroimaging data. Journal of Alzheimer's Disease 75: 277-288.

González-Salvador T, Lyketsos CG, Baker A et al. (2000) Quality of life in dementia patients in long-term care. International Journal of Geriatric Psychiatry 15(2): 181-189.

Herzog NJ, and Magoulas GD (2021) Brain asymmetry detection and machine learning classification for diagnosis of early dementia. Sensors 21(3): 778.

Hosmer DW, Lemeshow S and Sturdivant RX (2013) Applied logistic regression. Wiley, New York.

Hussain, M., Shafeeq, M.F., Jabbar, S., Akbar, A.H. and Khalid, S., 2016. CRAM: a conditioned reflex action inspired adaptive model for context addition in wireless sensor networks. *Journal of Sensors, 2016(1),* p.6319830.

Jabbar, S., Akbar, A.H., Zafar, S., Quddoos, M.M. and Hussain, M., 2014. VISTA: achieving cumulative VIsion through energy efficient Silhouette recognition of mobile Targets through collAboration of visual sensor nodes. *EURASIP Journal on Image and Video Processing, 2014(1),* p.32.

Ubaid, S., Shafeeq, M.F., Hussain, M., Akbar, A.H., Abuarqoub, A., Zia, M.S. and Abbas, B., 2018. SCOUT: a sink camouflage and concealed data delivery paradigm for circumvention of sink-targeted cyber threats in wireless sensor networks. *The Journal of Supercomputing, 74(10),* pp.5022-5040.

Farooq, A., Javed, F., Hussain, M., Abbas, T. and Hussain, A., 2012. Open source content management systems: a canvass. *International Journal of Multidisciplinary Science and Engineering, 3(10),* pp.38-43.

Hussain, M., Kim, K.H., Akbar, A.H., Khalid, S., Bang, S.J., Javed, M. and Amjad, M., 2016. A gateway deployment heuristic for enhancing the availability of sensor grids. *International Journal of Distributed Sensor Networks, 12(8),* p.7595038.