International Journal of Computers and Applications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjca20

Ensemble ResDenseNet: Alzheimer's disease staging from brain MRI using deep weighted ensemble transfer learning

Md Rabiul Hasan, A. B. M. Aowlad Hossain & Shah Muhammad Azmat Ullah

To cite this article: Md Rabiul Hasan, A. B. M. Aowlad Hossain & Shah Muhammad Azmat Ullah (2024) Ensemble ResDenseNet: Alzheimer's disease staging from brain MRI using deep weighted ensemble transfer learning, International Journal of Computers and Applications, 46:7, 539-554, DOI: 10.1080/1206212X.2024.2380648

To link to this article: https://doi.org/10.1080/1206212X.2024.2380648

INTERNATIONAL JOURNAL OF COMPUTERS AND APPLICATIONS
2024, VOL. 46, NO. 7, 539–554
https://doi.org/10.1080/1206212X.2024.2380648

Taylor & Francis
Taylor & Francis Group

Ensemble ResDenseNet: Alzheimer's disease staging from brain MRI using deep weighted ensemble transfer learning

Md Rabiul Hasan, A. B. M. Aowlad Hossain and Shah Muhammad Azmat Ullah

Department of Electronics and Communication Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh

ABSTRACT

Alzheimer's disease (AD) is a persistent neurological disorder with almost no current cure, although available medications can mitigate its progression. The timely identification and staging of AD is pivotal in impeding and managing its advancement. This study aims to develop a comprehensive framework for the early detection of AD and the classification of medical images across four AD stages. We have proposed a weighted ensemble deep transfer learning framework using two pretrained CNN architectures, namely ResNet152V2 and DenseNet201. Additionally, gradient-weighted class activation mapping (Grad-CAM) is utilized to enhance interpretability by capturing gradients associated with AD in brain MRI images, propagating them to the final convolutional layer. This technique narrows the interpretability gap in deep learning models, enhancing accessibility and understanding of their decision-making processes, particularly in the context of AD diagnosis. Two different MRI datasets have been used to train and evaluate the performances of the proposed framework focusing on generalization capability with diverse data. The performances of the framework are analyzed meticulously under different case studies. The experimental results demonstrate that the weighted ensemble architecture exhibits superior performance characteristics over base models showing accuracy of 98%, which is compatible with related state of the art algorithms.

ARTICLE HISTORY
Received 6 May 2024
Accepted 1 July 2024

KEYWORDS
Alzheimer disease staging; Weighted ensemble transfer learning; ResNet152V2; DenseNet201; Grad-CAM

1. Introduction

The on-going aging of the population has led to an increased prevalence of age-related neurodegenerative diseases as a significant consequence. Alzheimer's Disease International (ADI) reports that globally, over 50 million individuals are grappling with dementia. By 2050, Alzheimer's will affect 152 million individuals with new cases occurring every three seconds [1]. In Bangladesh, 0.9 million people had Alzheimer's disease (AD) in 2016 and 2.5 million in 2019 [2].

Mild cognitive impairment (MCI) is the initial phase of AD, a neurodegenerative dementia that worsens gradually over time due to cell damage in the brain. This causes memory loss, impaired cognitive functioning, and problems with doing regular activities. Information regarding this medical disorder is still lacking, although some well-documented reports say that mood swings and abnormalities in the brain's cortex are the symptoms. The main causes of this disease are still not clear, but most professionals agree that environmental factors and genetics are responsible for AD. Particularly, some research has also pointed out a connection between periodontitis and the herpes simplex virus type 1, along with other possible examples [3]. AD is viewed as related to age [4], and reports say there is sexual dimorphism in relation to the presence of this condition [5]. The social consequences and insufficient consensus on this health condition require intensive examination.

For experts in this domain, the primary obstacle they face is the absence of a well-established treatment for AD. As for this medical condition there is no proven cure, current medical care can lessen symptoms or deaccelerate its progression. Thus, Alzheimer's detection in its initial stage is very important. In order to mitigate the projected increase in AD treatment costs, computer-aided diagnosis systems are employed to facilitate accurate and early diagnosis [6–28]. Conventional machine learning techniques apply two types of features for AD: the first approach is based on regions of interest (ROIs) and the second approach is based on individual voxels [7,8]. Conventional methods rely on repetitive and subjective manual feature extraction, requiring technical expertise and iterative efforts. To solve this problem, an effective way is utilizing deep learning-based convolutional neural network (CNN) architectures. CNNs greatly improve efficiency, showing excellent outcomes in diagnosing AD [9–28]. Additionally, CNNs automatically extract key features, eliminating the need for human intervention.

However, it is found in the related literature survey that the performances of the classifiers degrade when number of class increases and many of the previous works considered just binary problems like healthy vs. ADs. Furthermore, number of subjects and size of the dataset of many of the previous works are quite smaller. Therefore, there are still scopes to conduct researches on multi-class AD staging using diverse and larger datasets which is a motivation behind this work. To fulfill this objective, design of a deep weighted ensemble transfer learning-based framework and focusing on the explainability of the classifier in terms of its localization capability of the ROI for feature extraction are the major concerns of this study. In fact, ensembling of multiple models improve overall performance taking the benefits of individual models [29]. Furthermore, we planned to emphasize on the data preprocessing strategies and different investigations considering two different datasets towards the model's generalization capability and performance improvement. For the purpose of early AD detection and classification from brain MRI, this study proposes a weighted ensemble deep transfer learning-based framework. Clinical studies frequently use magnetic resonance imaging (MRI) brain images as the main data source due to the significant

CONTACT A. B. M. Aowlad Hossain aowlad0403@ece.kuet.ac.bd Department of Electronics and Communication Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh

correlation with brain architecture and capacity to show morphological changes. The study classifies AD into four stages – (I) mild demented, (II) moderate demented, (III) non-demented, and (IV) very mild demented – employing a multi-classification approach. Two distinct datasets are utilized to assess model performance. The work can be summarized as follows:

- We have proposed 'Ensemble ResDenseNet,' a deep learning framework that uses a weighted ensemble strategy for four-class AD staging from brain MRI images, combining ResNet152V2 and DenseNet201 CNNs.
- ResDenseNet is applied to both original and augmented MRI images from two dataset (Dataset1 and Dataset2) to improve generalization, feature learning, robustness, and overall performance. The framework's performance in multi-class classification of AD is improved by optimizing the hyperparameters of the CNN models.
- A thorough comparative analysis and performance evaluation of the proposed ResDenseNet has been conducted. We employ various performance metrics and incorporate Grad-CAM image analysis to provide visual insights into its effectiveness.

2. Literature review

The detection of AD has been a domain of significant research, encompassing numerous issues and challenges. Helay et al. [10] applied a CNN approach for early AD detection. The authors undertook oversampling by adding rotations and flips to the MRIs in order to address class imbalance issues in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, producing an expanded dataset of 48,000 pictures. Subsequently, they developed both 2D and 3D CNN models, achieving accuracy rates of 93.61% and 97%, respectively. Oktavian et al. [11] tackled the class imbalance issue within a dataset consisting MRIs of 306. To address this challenge, they employed the ResNet-18 architecture along with a weighted loss function and activation function. This strategy led to a notable improvement in accuracy, with the performance increasing from 69.1% to 88.3%. Working with the identical dataset, Lu et al. [12] devised an innovative deep neural network employing a multistage methodology for AD identification, achieving notable results. Their model attained 86.4% accuracy in predicting MCI, demonstrating a sensitivity of 94.23% for this condition. Furthermore, this high sensitivity was beneficial for identifying patients who later progressed to AD within a three-year timeframe. Ge et al. [13] presented a deep learning 3D multiscale framework and applied it to a subject-specific dataset of brain scans, randomly partitioned. The model demonstrated an average accuracy of 87.24% and demonstrated the highest test accuracy of 93.53% on subject-specific data. While the primary goal of this approach is not individual categorization, it leverages multiple biomarkers to discern group disparities related to AD. Khagi et al. [14] recommended a superficial fine-tuning approach involving established CNN architectures like AlexNet, GoogleNet, and ResNet50. The focus of this investigation was to assess the impact of each layer section on the outcomes of both natural image and medical image classification tasks. Wang et al. [15] introduced an innovative CNN structure that leveraged a multimodal MRI analytical strategy alongside diffusion tensor imaging (DTI) data. This framework was utilized for classifying individuals into groups of patients with AD, normal controls (NC), and those with amnestic mild cognitive impairment (aMCI). While the model exhibited impressive classification accuracy, it was hypothesized that employing 3D convolution could potentially yield even better performance compared to the use of 2D convolution. Spasov et al. [16] introduced the APOe4 model, focusing on Apolipoprotein E expression level 4 (APOE4). This model integrated MRI scans, genetic testing, and clinical assessments as input factors. This approach aimed to reduce computational complexity, mitigate overfitting, lower memory demands, accelerate prototyping, and employ a streamlined parameter set in contrast to pretrained models like VGGNet and AlexNet. Korolev et al. [17] illustrated that it was feasible to achieve a similar level of performance. The results obtained when applying both residual networks and standard 3D CNN architectures to 3D structural MRI brain data indicated that these networks exhibited excessive depth and complexity. Consequently, their performance fell short of expectations. Goenka et al. have used 708 T1-weighted MRI of 69 subjects from the MIRIAD datasets for binary classification of AD and healthy case classification [18]. Using 3D volumetric CNN, they have got an accuracy of 97% for original images and 100% with rotation+scaling augmented images. However, number of subjects and images of the MIRIAD dataset are quite smaller than other datasets. Maringanti et al. [19] employed a recurrent neural network (RNN) model within a neural network framework. The research focused on analyzing MRI data of ADNI dataset, achieving a noteworthy accuracy of 90%. Abbas et al. [20] introduces an innovative approach to the diagnosis of AD by proposing a 3D Jacobian domain convolutional neural network (JD-CNN). This novel model achieves outstanding classification results, reaching an accuracy of 94.20%. The study uses baseline T1-weighted structural MRI data from the ADNI database. It turns spatial information into the Jacobian domain for feature generation and shows that the JD-CNN is effective at identifying AD patients. Mamun et al. [21] presented a framework for the development of deep learning models leveraging a dataset comprising 6219 MRI images representing varying degrees of demented and non-demented brains. Employing CNN, DenseNet121, ResNet101, and VGG16, the study revealed that CNN emerged as the most effective model, showcasing remarkable performance with an accuracy of 97.60%, recall of 97%, and an impressive AUC of 99.26%. Research by Allioui et al. [22] sought to automate the identification of damaged regions and the diagnosis of AD. Their suggested approach properly segments MRI images, detects brain lesions, and differentiates between AD stages. Their approach had a 94.73% accuracy rate, 93.82% recall rate, and 92.8% F1 score. Hussain et al. [23] introduced a 12-layer CNN to perform AD classification by analyzing brain MRI data. Previous CNN models were compared to the proposed model, which did better performance on the Open Access Series of Imaging Studies (OASIS) dataset with an impressive accuracy of 97.75%. Xiaojun et al. [24] proposed a multi-task learning strategy for electroencephalogram (EEG) spectral image classification applying a discriminative convolutional high-order Boltzmann machine with hybrid feature maps. The authors developed the DCssCDBM model. This innovative method reduces overfitting, improves inter-subject variances, and reduces intra-subject variations to improve AD detection. Ebrahimi et al. [25] utilized a 2D CNN to interpret ADNI MRI data. The accuracy rate they achieved was 98%. Arijit et al. [26] used 3D DTI for four-class classification. VoxCNNs were trained using a unique approach using echo planar imaging intensities, fractional anisotropy (FA), and mean diffusivity (MD) from DTI images. Each brain region's average FA and MD values were fed into a random forest classifier. The classification accuracy was 92.6%, achieved by using modified rank averaging at the decision level. Tamer et al. [27] introduced and evaluated two novel hybrid deep learning architectures for the AD detection. These architectures use a number of BiLSTM fusions. The first architecture takes the form of an interpretable multi-task regression model, predicting seven essential cognitive scores for patients 2.5 years after their last observations. The forecasted scores contribute to the establishment of an interpretable clinical decision support system, built upon a transparent glass-box model. Fang et al. [28] introduced an efficient system

Table 1. Overview of studies using different methods for AD detection.

S/N	Author	Method	Biomarker	Performance
1	Helaly et al. [10]	2D CNN, 3D CNN and VGG19	MRI	The accuracy of 2D CNN, 3D CNN, and VGG19 is 93.61%, 95.17%, and 97%, respectively.
2	Oktavian et al. [11]	ResNet18	MRI	Accuracy = 88.3%
3	Lu et al. [12]	Deep Neural Network	MRI, FDG-PET	Overall accuracy, sensitivity, and specificity are 86.4%, 94.23%, and 86.3%, respectively.
4	Ge et al. [13]	3D CNN	MRI	Best and average test accuracy of 93.53% and 87.24% on subject-separated dataset, and 99.44% and 98.80% on random brain scan-partitioned dataset
5	Khagi et al. [14]	AlexNet, GoogleNet, and ResNet50	MRI	For AlexNet, the accuracy of Q1, Q2, Q3, and Q4 tuning is 0.94, 0.93, 0.91, and 0.84. For GoogleNet, the accuracy of Q1, Q2, Q3, and Q4 tuning is 0.88, 0.88, 0.81, and 0.79. For ResNet50, the accuracy of Q1, Q2, Q3, and Q4 tuning is 094, 0.93, 0.92, and 0.91.
6	Wang et al. [15]	CNN	MRI	Accuracy = 92.06%
7	Spasov et al. [16]	CNN	MRI	Accuracy = 99%, Sensitivity = 98%, Specificity = 100% and AUC = 1
8	Korolev et al. [17]	ResNet and VoxCNN	MRI	For VoxCNN the accuracy is AD vs NC = .79 ± .08, AD vs EMCI = .64 ± .07, AD vs LMCI = .62 ± .08, LMCI vs NC = .63 ± .10, LMCI vs EMCI = .56 ± .11, EMCI vs NC = .54 ± .09 and For ResNet the accuracy is AD vs NC = .80 ± .07, AD vs EMCI = .63 ± .09, AD vs LMCI = .59 ± .11, LMCI vs NC = .61 ± .10, LMCI vs EMCI = .52 ± .09, EMCI vs NC = .56 ± .07
9	Goenka et al. [18]	3D CNN	MRI	Accuracy = 97% (without augmentation)
10	Maringanti et al. [19]	RNN, Neural Network	MRI	Accuracy = 90%
11	Abbas et al. [20]	JD-CNN	MRI	Accuracy = 94.20%
12	Mamun et al. [21]	ResNet, DenseNet, VGG16	MRI	Accuracy = 97%
13	Allioui et al. [22]	3D CNN	MRI	Accuracy = 94.73%, Recall = 93.82%, and F1-score = 92.8%
14	Hussain et al. [23]	12-Layer CNN	MRI	Accuracy = 97.75%
15	Xiaojun et al. [24]	Convolutional96 Deep Boltzmann Machine	EEG	Accuracy = 96%
16	Ebrahimi et al. [25]	2D CNN	MRI	Accuracy = 98%
17	Arijit et al. [26]	VoxCNN	DTI	Accuracy = 92.6%
18	Tamer et al. [27]	BiLSTM	MRI	Accuracy for MRBL and DFBL is 84.95% and 86.08%, respectively
19	Fang et al. [28]	AlexNet	MRI	Accuracy = 92.85%

employing transfer learning to classify images through fine-tuning a pre-trained convolutional network, specifically AlexNet. The architecture was trained and tested on both segmented (grey matter, white matter, and cerebral spinal fluid) and un-segmented images for both one-class and multiple-class classification. When tested on the OASIS dataset, the algorithm showed great promise, with an overall success rate of 92.85% for multi-class classification of un-segmented images. Table 1shows an overview of studies using different methods for AD detection. To avoid repetition, a number of related state of the art works on deep learning-based techniques [30–46] are described in Table 9 of section 4.3 (which is given there for comparative analysis) mentioning theirs used datasets, methodology, and performance. From the literatures survey, it is found that performances of the classifiers degrade when number of class increases and many of the previous works considered only binary problems like healthy vs. ADs. Furthermore, number of subjects and size of the dataset of many of the previous works are quite smaller. Therefore, there are still scopes to conduct researches on multi-class AD staging using diverse and larger dataset which is a motivation behind this work. Furthermore, focusing on the explainability of the classifier in terms of its localization capability of the ROI for feature extraction is the major concern of this study.

3. Proposed method

The principal goal of this study is to design a framework for identifying AD stages from MRI scans. To achieve this objective, we leveraged the ensemble of the ResNet152V2 and DenseNet201 models. The foundational step involved meticulous data preparation, encompassing essential MRI image preprocessing techniques such as scaling, normalization, augmentation, and one-hot encoding. The dataset was then divided into two portions: testing and training with validation. Subsequently, an iterative process was undertaken to fine-tune the hyperparameters of the models, seeking to minimize error. The optimized models, thus obtained, were employed for AD detection. A visual representation of our proposed methodology is shown in Figure 1.

3.1. Dataset description

We have used two distinct datasets to train and evaluate our proposed model. The datasets have been collected from Kaggle repository [47,48]. We have chosen these datasets since, both of them consist of four-class data of healthy and ADs stages namely, mild demented, moderate demented, non-demented, and very mild demented and hence, they are useful for our targeted multi-class problem and data diversity investigation. To improve the generalization, feature learning, robustness of the classifier as well as for overall performance investigation, we think this intentional data diversity strategy is important. Figure 2 shows a comparative view of the size of the two datasets.

Dataset1: This dataset consists of 6400 original brain MRI of four AD stages, which are readily accessible on Kaggle [47]. Where 896

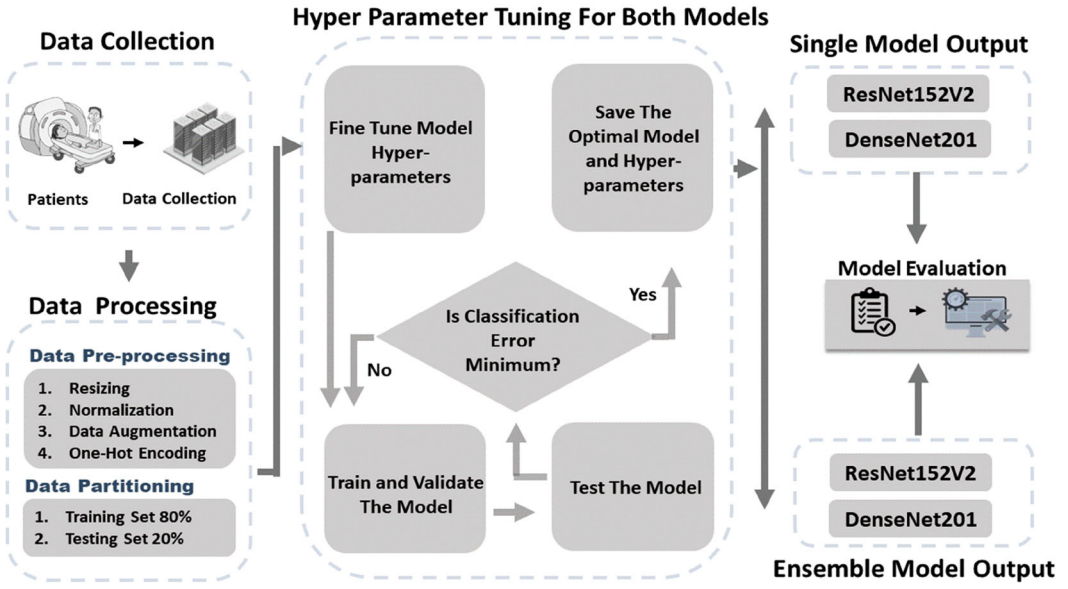


Figure 1. Proposed methodology for Alzheimer staging.

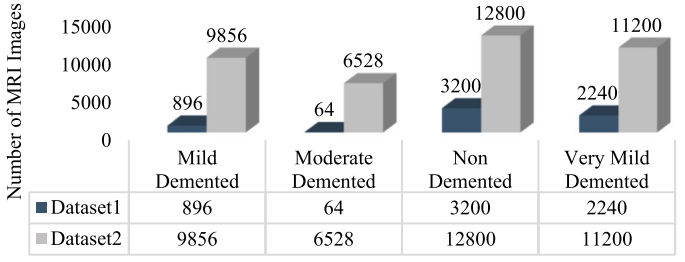


Figure 2. Comparison of the size of Dataset1 and Dataset2.

images are for mild demented, 64 images are for moderate demented, 3200 images are for non-demented, and 2240 images fare or very mild demented. Since this dataset is quite small, we will use them for the blind testing of the classifiers in this study. Whereas, Dataset2 will be used for training and validation along with testing.

Dataset2: This dataset is also collected from Kaggle [48] which consists of 40,384 images of the four AD classes. The dataset is structured into two directories, one containing original images exclusively and the other containing augmented versions, as depicted in Figures 3 and 4. Table 2 shows the augmentation techniques employed in Dataset2. The original images are reserved solely for the purpose of model testing. The dataset was partitioned for training, validation, and testing, comprising 27,187, 6,797, and 6,400 images, respectively. During the training and validation phase, we employed 8960 images for the mild demented category, 6464 for the moderate demented category, 9600 for the non-demented category, and 8960 images for the very mild demented category. In the testing phase, considering the same size of Dataset1, we also utilized randomly chosen 896 images for mild demented, 64 images for moderate demented, 3200 images for non-demented, and 2240 images for very mild demented from Dataset2.

3.2. Image preprocessing

Preprocessing is a key procedure that entails transforming raw data into a more suitable format for subsequent analysis and improved user understanding. Resizing, normalization, augmentation, and encoding are the four consecutive steps that make up the preprocessing of Alzheimer's MRI images.

Table 2. The augmentation technique employed in Dataset2.

Technique	Settings
Image Rescaling	Rescale = 1/255
Width Shift	Width_shift_range = 0.1
Height Shift	Height_shift_range = 0.1
Shear Transformation	Shear_range = 0.2
Zooming	Zoom_range = 0.1
Horizontal Flipping	Horizontal_flip = True
Vertical Flipping	Vertical_flip = True

3.2.1. Resizing

Biomedical image classification relies on image resizing to ensure image proportions are uniform. As the datasets contain MRI images of various resolutions, all the MRI images have been transformed from their original formats. We resized all the brain MRI images to 128 by 128 for seamless integration in the ResNet152V2 and DenseNet201 CNN models. For improved generalization and correctly diagnosing AD, resizing the brain MRI image is very crucial.

3.2.2. Normalization

Alzheimer's brain MRI normalization involves transforming pixel values from 0 to 255 to fall in the range [0,1] or [−1,1]. Normalization step is necessary for MRI analysis and it helps the ResNet152V2 and DenseNet201 CNN architectures work better. This study emphasizes the importance of brain MRI normalization for diverse healthcare image standards. The model's outcome suggests that normalizing MRI images enhances the model's performance by making CNN architectures less vulnerable to changes in image intensity. This research demonstrates the considerable advantages of proper image normalization in enhancing the performance of deep learning-based diagnostics for AD detection and classification.

3.2.3. Data augmentation

For augmenting brain MRI images, width shifts, height shifts, flips, rotations, and zooms are considered. Image augmentation is one of the fundamental ways of enriching and broadening the training images for CNN models. This technique enhances the generalization and robustness of deep CNN architectures and reduces the chance of overfitting conditions by increasing the dataset volume.

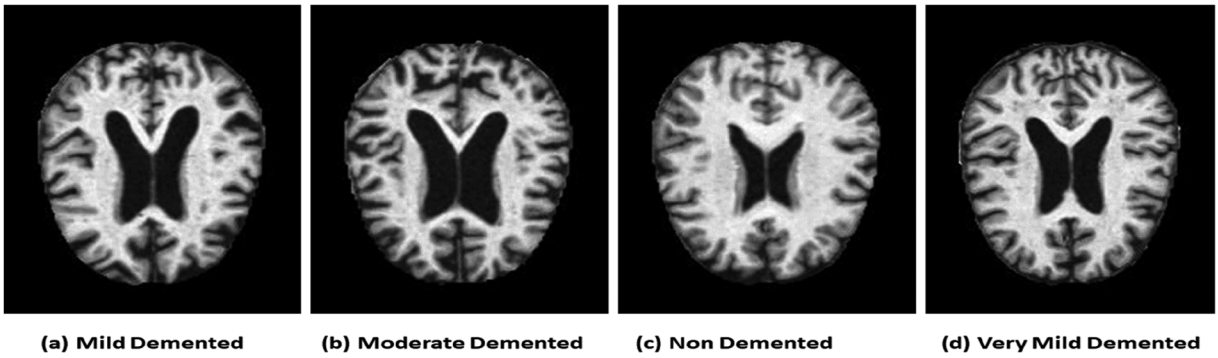


Figure 3. Original MRI images of different Alzheimer's stages.

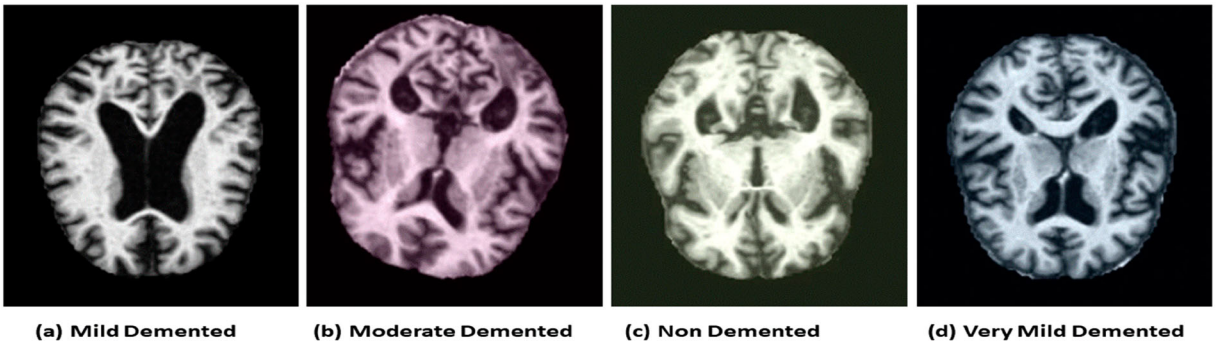


Figure 4. Sample augmented MRI images of different Alzheimer's stages.

3.2.4. One-hot encoding

This study employs a mathematical way of transforming categorical data into binary vectors called one-hot encoding. The four different categories – mild, very mild, moderate, and non-demented – are converted to unique binary vectors using this technique. For performing one-hot encoding, we have applied the 'tensorflow.keras.utils.to_categorical' function to ensure unique binary code for each category. After that, the ResNet152V2 and DenseNet201 CNN architectures are trained using these binary vectors as labels.

3.3. Deep weighted ensemble transfer learning model architecture

The CNN stands as one of the most prevalent deep neural network architectures. It primarily aims to synergize computer vision and deep learning techniques, showcasing notable efficacy in handling image data, particularly in tasks of image classification, within the realm of deep learning challenges. The CNN architecture comprises several concealed layers, including convolutional layers, pooling layers, and fully connected layers. These intricate layers collaboratively contribute to the network's hierarchical feature extraction. Ultimately, the CNN generates its output by flattening and aggregating the outputs from these various layers. Figure 5 shows the basic architecture of CNN.

A primary function of the convolution layer is to extract features from the data. First, it applies a convolution function, and then it applies an activation function to the output of the convolution function. This is how it handles the feature extraction process. In the process of convolution, it extracts the features by employing a linear function that is referred to as the kernel function, which might be considered as filter. Let's say, we have an input image that is characterized by tensor I and has dimensions of $m_1 \times m_2 \times m_c$. Here, m_1 = height of image, m_2 = width of image, and m_c = number of channels. We employ a tensor filter with dimensions $n_1 \times n_2 \times n_c$, where n_c corresponds to the number of channels in the kernel, mirroring the input image. The filter horizontally traverses the image, conducting element-wise multiplication between the specified region of the input (I) and the filter (K), then summing these results. The stride parameter dictates the interval at which the filter advances across the image. The outcome of this operation, represented as the tensor product of I and K, results in another tensor with dimensions $(m_1 - n_1 + 1) \times (m_2 - n_2 + 1) \times 1$. Here, Dimension of $I = m_1 \times m_2 \times m_c$, Dimension of $K = n_1 \times n_2 \times n_c$, and Dimension of $F = (m_1 - n_1 + 1) \times (m_2 - n_2 + 1) \times 1$. And, $F[i,j] = (I \times K)_{[i,j]}$.

The following is the entry for the ij-th position on the feature map:

$$F[i,j] = \sum_{x}^{m_1} \sum_{y}^{m_2} \sum_{z}^{m_c} K_{[x,y,z]} I_{[i+x-1,j+y-1,z]} \tag{1}$$

The pooling process involves the extraction of features to enhance the model's accuracy, followed by pooling operations aimed at reducing the dimensionality of the convolutional layer. A pooling function is applied to the convolution layer's result in this stage. Let us make the assumption that: $\text{conv}(I,K) = C$ and $P = \emptyset_p(C)$, where $\emptyset_p$ is a pooling function. The dimension of the pooled section can be expressed as follows:

$$P = \left(\frac{m_1 + 2p - n_1}{s}\right) \times \left(\frac{m_2 + 2p - n_2}{s}\right) \times m_c \tag{2}$$

where s and p stand for stride and padding, $m_1 \times m_2 =$ input image dimensions, and $n_1 \times n_2 =$ padding kernel dimensions. Pooling encompasses various techniques, including sum pooling, average pooling, and max pooling.

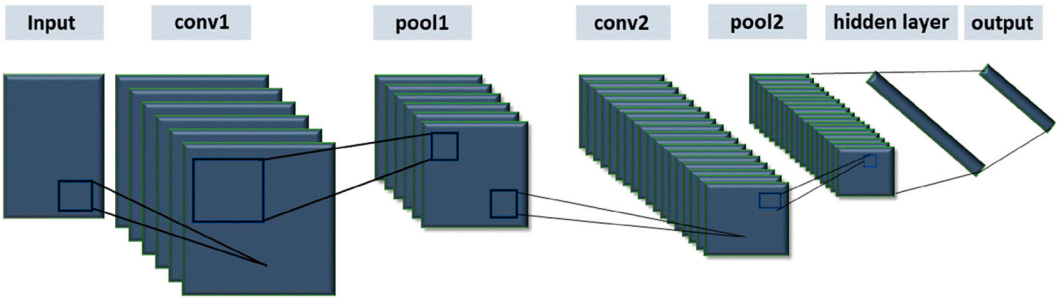


Figure 5. Basic CNN architecture.

A series of concealed layers, comprising both convolutional and pooling layers, is employed for feature extraction. Following feature extraction, the resulting features are flattened into a singular vector. Subsequently, this singular vector serves as the input for the fully connected layer responsible for the classification task. The fully connected layer processes the flattened vector, yielding another vector. In machine learning models, there is the possibility of one class being more predominant than another. To address this imbalance, balanced weights are incorporated with the pooled component, a bias term is introduced, and an activation function is applied. This process is repeated across each layer, where weights are added to the pooled segments and the activation function is applied. The final layer employing an activation function that conducts classification by computing the probability for each class.

We have chosen two CNN-based widely used pretrained models namely ResNet152V2 and DenseNet201 as feature extractor and customized the fully connected parts according to the AD staging problem which are then ensembled. We have chosen these two powerful pretrained models for our study, since both of the models have large number of layers and hence they can effectively extract the prominent features. ResNet152V2 having 152 layers and DenseNet201 having 201 layers can extract complex and hierarchical features from brain MRI images and fuse low-level features (like edges and textures) with high-level features (like anatomical structures). This is useful for distinguishing between small variations in MRI images linked to various stages of AD. Moreover, they have the capability to solve the vanishing gradient problem. As an improved version of ResNet, the pre-activation residual units, enhanced normalization, and initialization methods in ResNetV2 result in quicker convergence and superior generalization on new MRI data. Furthermore, compared to the conventional CNNs, DenseNet201 is more effective since it reuses features, which lowers parameters and enhances generalization on new data.

3.3.1. ResNet152V2

The residual network, commonly referred to as ResNet, was initially introduced by Kaiming He [49]. ResNet was conceptualized as a solution to address two significant challenges in CNNs: the problem of network degradation and the issue of vanishing gradients, where gradients tend to approach zero, hindering weight updates. ResNet152V2, an advanced variant, boasts an impressive depth with a total of 152 layers in its neural network architecture. The ResNet152V2 architecture is shown in Figure 6. The ResNet152V2 model represents an evolution of the original ResNet design, incorporating state-of-the-art techniques such as batch normalization and additional residual blocks. These innovations play a pivotal role in enhancing training stability and accelerating convergence during the training process. ResNet152V2 is widely adopted across various computer vision applications, including but not limited to image classification, object detection, and image segmentation. Its utilization is particularly advantageous in scenarios that demand precise feature extraction and recognition, owing to its exceptional depth and accuracy.

Residual learning: Let $H(X)$ be regarded as an underlying mapping intended to be modeled by a series of stacked layers, not necessarily encompassing the entire network, where X signifies the input to the initial layer. If one assumes that multiple non-linear layers can asymptotically replicate intricate functions, this is tantamount to assuming their ability to asymptotically mimic the residual functions, denoted as $H(X) - X$ (assuming consistent dimensions for the input and output). Instead of expecting stacked layers to precisely emulate $H(X)$, we explicitly permit these layers to imitate a residual function $F(X) := H(X) - X$. Consequently, the original function is reformulated as $F(X) + X$. While both formulations theoretically possess the capability to asymptotically approximate the desired functions, the ease of learning may exhibit variations.

Identity mapping by shortcuts: Residual learning is employed in a periodic manner over multiple stacked layers. Figure 7 depicts a building block. A building block can be defined as:

$$Y = F(X, \{W_I\}) + X \tag{3}$$

Here, X and Y denote the input and output vectors of the layers under consideration. The function $F(X, \{W_I\})$ signifies the residual mapping targeted for learning. The function $F(X, \{W_I\})$ has the ability to represent many convolutional layers and form the ResNet152V2 architecture. For the specific case illustrated in Figure 7, involving two layers, $F = W_2\sigma(W_1X)$, where σ denotes the rectified linear unit (ReLU), and biases are excluded for notational simplicity, The operation $F+X$ is executed through a shortcut connection and element-wise addition. Following the addition, we incorporate the second nonlinearity (i.e. $\sigma(Y)$) for further processing.

3.3.2. DenseNet201

The term 'DenseNet' characterizes a family of CNNs characterized by their adoption of dense connections within their layers [50]. Variants within this category include DenseNet 201 and DenseNet 169. In the case of DenseNet201, for instance, when processing images of dimensions 128 × 128 pixels, dense blocks are employed, establishing direct connections between all 201 layers as shown in Figure 8. This architectural design facilitates the continuous flow of information, wherein each layer both acquires new insights from the layers beneath it and conveys its own feature maps to the layers above it, preserving the feed-forward nature of the network.

Feature concatenation mathematical expression is described as follows:

$$Z^I = H_I([Z^0, Z^1, \ldots, Z^{I-1}]) \tag{4}$$

Here, the non-linear transformation denoted as $H_I(.)$ can be expressed as a composite function consisting of batch normalization (BN) followed by a ReLU and a (3 × 3) convolution.

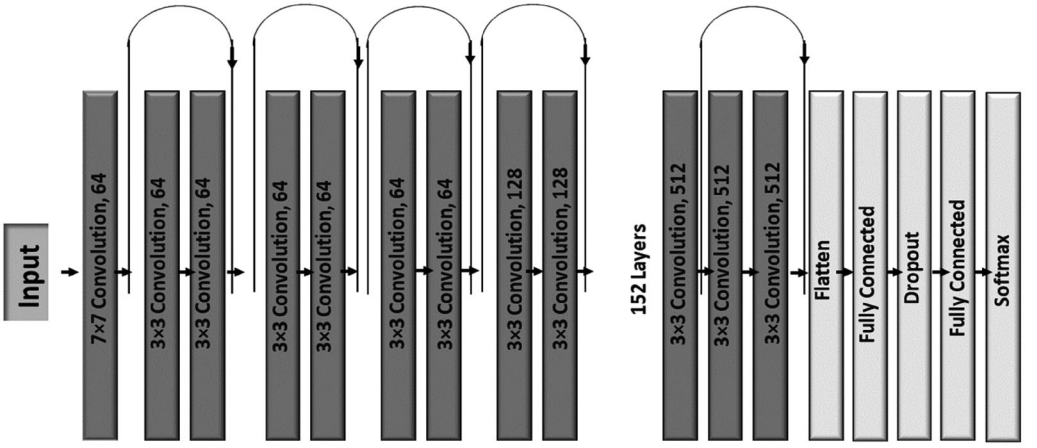


Figure 6. ResNet152V2 architecture.

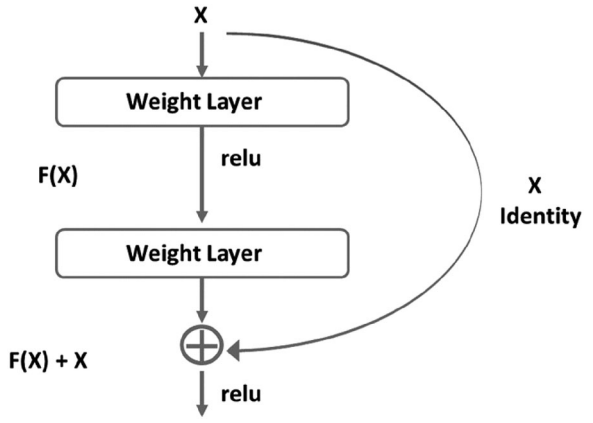


Figure 7. Residual learning building block.

$[Z^0, Z^1, \ldots, Z^{I-1}]$ represents the amalgamation of feature maps from layer 0 to l – 1, consolidated into a singular tensor for simplified implementation. In the network architecture, dense blocks are established for down-sampling, interspersed with transition layers. These transition layers include batch normalization, succeeded by a 1 × 1 convolution layer, and culminate in a 2 × 2 average pooling layer.

The hyperparameter *k*, which stands for growth rate, shows how well DenseNet201's dense architecture works at getting state-of-the-art results. Remarkably, DenseNet201 demonstrates robust performance even with a reduced growth rate, attributed to its unique architecture, which treats feature maps as a global state of the network. This design ensures that every subsequent layer has access to all feature maps from preceding layers. Specifically, each layer contributes k feature maps to the global state, where the total number of input feature maps at the I^{th} layer (denoted as $(FM)^1$) is precisely defined.

$$(FM)^1 = K^0 + K(I-1) \tag{5}$$

Here, k^0 stands for the channels in the input layer. To enhance computational efficiency, a 1 × 1 convolution layer is incorporated before each 3 × 3 convolution layer. This inclusion serves to reduce the number of input feature maps, which are generally more abundant than the output feature maps (k). The introduced 1 × 1 convolution layer is termed the bottleneck layer and yields 4 K feature maps.

Every neuron within the fully connected dense layer establishes complete connectivity with each neuron in the preceding layer. This connectivity can be expressed mathematically through a fully connected layer, wherein the input 2D feature map is transformed into a 1D feature vector. The role of the softmax activation function is to transform non-normalized outputs into multi-class outputs. Consequently, it facilitates the conclusive classification of AD into categories such as mildly, very mildly, moderately, and non-demented. The softmax function can be formally defined as:

$$S(Z)_I = \frac{e^{Z^I}}{\sum_{J=1}^{K} e^{Z_J}} \tag{6}$$

where $I = 1, \ldots, K$ and $Z = (Z_1, \ldots, Z_K) \epsilon \mathbb{R}^K$

3.3.3. Transfer learning

The primary objective of transfer learning is to utilize acquired information obtained from solving a particular problem and use it to another. Pre-trained models are initially developed using one dataset and subsequently adapted to address challenges in a different dataset as shown in Figure 9. ResNet and DenseNet stand out as two prominent CNNs frequently employed in the realm of medical image analysis. These models were originally trained on a dataset comprising 14 million images and 1000 distinct classes, known as ImageNet. A conventional practice involves fine-tuning the higher-level layers of the model while keeping the lower layers fixed. This approach facilitates the transfer of information between the pre-trained model and the new model within the same domain. Grid search technique is used for hyperparamaets tuning. The hyperparameters chosen as common for both the ResNet152V2 and DenseNet201 models training for keeping uniformity are given in Table 3.

3.3.4. Weighted ensemble

In this study, we have used the weighted average ensemble method to the ResNet152V2 and DenseNet201 model architectures. Let M_1 represents the ResNet152V2 model, and M_2 represents the DenseNet201 model. The output of each individual model M_i for a given input sample x is denoted as $P_i(x)$ representing the predicted probability distribution over the four classes: mild, very mild, moderate and non-dement [51].

The weighted ensemble method combines these individual model predictions to obtain the final ensemble prediction $P_{\text{ensemble}}(x)$ as

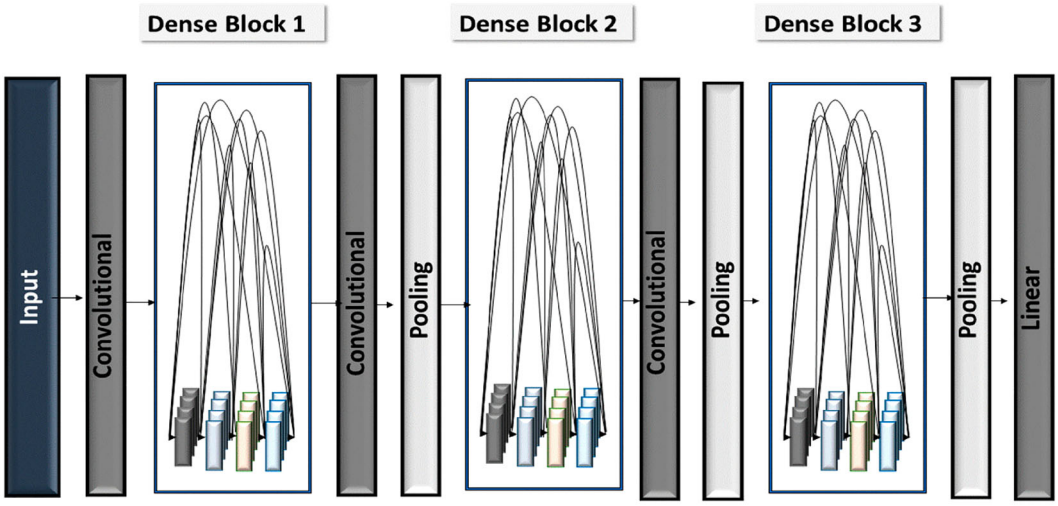


Figure 8. DenseNet201 architecture.

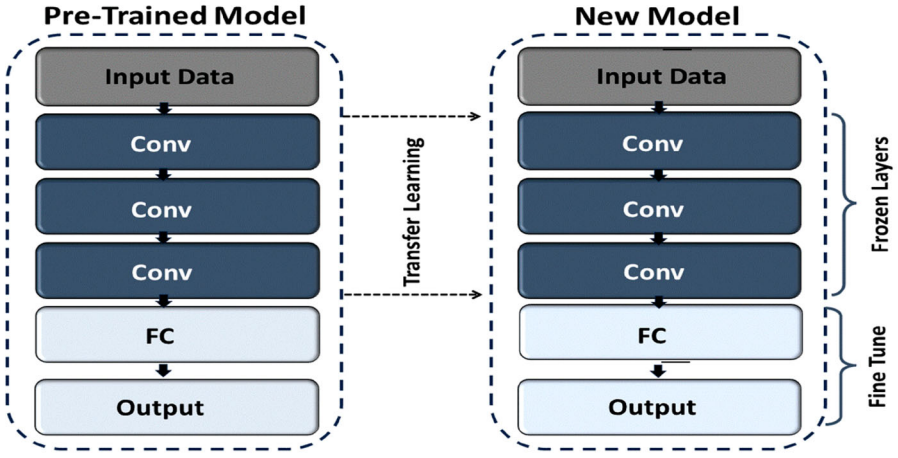


Figure 9. Transfer learning concept.

Table 3. List of hyperparameters used during ResNet152V2 and DenseNet201 models training.

Hyperparameter	Values	Description
Image size	128 × 128	Input image dimensions (width × height)
Batch size	32	Number of MRI images per batch during training
Depth	3	Number of color channels in input images (RGB)
Epochs	100	Number of training epochs
Learning Rate	0.001	Initial learning rate for Adam optimizer
Optimizer	Adam	Optimization algorithm
Loss Function	Categorical Crossentropy	Loss function used for training
Early Stopping	10	Number of epochs with no improvement before stopping early
Dropout Rate	0.5	To avoid overfitting, the dropout rate for fully connected layers
Activation Function	ReLU for hidden layers, Softmax for output	Activation functions utilized in the network
Weight Initialization	He Normal	Method for initializing weights
Call Backs	EarlyStopping, ModelCheckpoint	Callbacks used during training

follows:

$$P_{\text{ensemble}}(x) = w_1 \cdot P_1(x) + w_2 \cdot P_2(x) \tag{7}$$

where w_1 and w_2 represent the weights assigned to the ResNet152V2 and DenseNet291 models, respectively. The weights w_1 and w_2 are determined through a training process and are typically optimized to maximize the overall ensemble performance. These weights reflect the importance of each model in contributing to the final prediction. To ensure that the ensemble weights sum to 1 ($w_1 + w_2 = 1$), one common normalization techniques is to use softmax weights:

$$w_i = \frac{e^{z_i}}{\sum_{j-1}^{2} e^{z_j}} \tag{8}$$

where z_i is a learnable parameter associated with the i-th model. The training of the weighted ensemble involves optimizing the weights and model parameters jointly to minimize the loss function for the classification problem with four classes. During the testing phase, the final ensemble prediction is obtained using the weighted combination as described above. This mathematical formulation allows for the effective utilization of both ResNet152V2 and DenseNet201

Algorithm: The proposed Ensemble ResDenseNet architecture pseudocode

```
1-4: Input: Dataset Dir, Categories C, Hyperparameters E, B, InputShape
5-6: Output: Best Model, performance metrics
7: SplitDataset(Dir) into training T, validation V, and testing S
Model Training:
8: Initialize ImageNet Pre-trained CNN Models (ResNet152V2, DenseNet201)
9: For each CNN model M in (ResNet152V2, DenseNet201) do
   a: If M = = ResNet152V2 then
      i: Initialize resnet_model with ImageNet weights, include_top = False,
  input_shape = InputShape
      ii: Freeze first 100 layers, unfreeze remaining layers
      iii: Add GlobalAveragePooling2D, Dropout, Dense, and output layers to
  resnet_model
   b: Else if M = = DenseNet201 then
      i: Initialize densenet_model with ImageNet weights, include_top = False,
  input_shape = InputShape
      ii: Add GlobalAveragePooling2D, Dropout, Dense, and output layers to
  densenet_model
   c: Fine-tune model hyperparameters (Learning Rate, Batch Size, Number of
  Epochs, Optimizer, Loss Function, Dropout Rate)
   d: Train model M on train_generator with early stopping and ModelCheckpoint
  callbacks
   e: Compile model M with categorical_crossentropy loss and adam optimizer
   f: Do while training is not converged
      i: Train model M on T with batch size B and epochs E
      ii: Monitor classification error on V
      iii: If classification error < minimum then
         save the optimal model and hyperparameters
      end if
   end while
end for
Testing and Evaluation:
10: Test the optimal model on the testing set S
11: Evaluate model performance (Accuracy, Specificity, Sensitivity)
12: For each test image in S do
   a: Generate Grad-CAM visualization
end for
Ensemble Learning:13: Combine predictions from ResNet152V2 and DenseNet201
   using Weighted Ensemble method
14: Evaluate ensemble model performance
Performance Evaluation:
15: For each test image in S do
   a: Classify images into categories (Mild Demented, Moderate Demented, Very
  Mild Demented, Non     Demented)
   b: Visualize classification results using Grad-CAM
end for
16: Output the best model and performance metrics (Accuracy, Specificity,
   Sensitivity)
17: Return Best Model, Accuracy, Specificity, Sensitivity
```

models, leveraging their complementary strengths for enhanced AD detection. The pseudocode of the algorithm of the proposed Ensemble ResDenseNet architecture is given below:

3.4. Evaluation

The present study utilized two separate datasets in order to detect cases of AD. Both datasets have the same four-class labels: mild, very mild, moderate, and non-demented. In this subsection, an overview of the evaluation metrics utilized for the four classes of AD classification has been provided along with eight evaluation metrics for assessing the model's performance. The following is a concise explanation of each of these eight metrics, accompanied by their respective mathematical equations.

- t_{pi} : True Positive Rate of i^{th} Class.
- f_{pi} : False Positive Rate of i^{th} Class.
- t_{ni} : True Negative Rate of i^{th} Class.
- f_{pi} : False Negative Rate of i^{th} Class.
- l: Total dataset classes.

I. Overall accuracy

Overall accuracy measures themodel's rate of accurate predictions.

$$\text{Overall Accuracy} = \frac{TP + TN}{\text{Test Set}}$$

II. Macro-averaged precision

Macro precision measures average-class precision.

$$\text{Precision}_{\text{M}} = \frac{\frac{\sum_{i=1}^{l} t_{p_i}}{t_{p_i} + f_{p_i}}}{l}$$

III. Weighted precision

The weighted mean of the precision is interpreted using class probability-based weights.

$$\text{Precision}_{\text{W}} = \frac{\sum_{i=1}^{l} \text{Precision}_i + \text{Test Set}_i}{\text{Test Set}}$$

IV. Macro-averaged recall

Macro recall measures average-class recall.

$$\text{Recall}_{\text{M}} = \frac{\frac{\sum_{i=1}^{l} t_{p_i}}{t_{p_i} + f_{n_i}}}{l}$$

V. Weighted recall

The weighted mean of the recall is interpreted using class probability-based weights.

$$\text{Recall}_{W} = \frac{\sum_{i=1}^{l} \text{Recall}_i + \text{Test Set}_i}{\text{Test Set}}$$

VI. Macro-averaged F1 score

The metric is referred to the mean F1 score per class.

$$\text{F1Score}_{\text{M}} = \frac{2 \times \text{Precision}_{\text{M}} \times \text{Recall}_{\text{M}}}{\text{Precision}_{\text{M}} + \text{Recall}_{\text{M}}}$$

VII. Weighted F1 score

The mean of each class' F1 score is averaged to determine the weighted average F1 score.

$$\text{F1Score}_{\text{W}} = \frac{\sum_{i=1}^{l} F1\text{Score}_i + \text{Test Set}_i}{\text{Test Set}}$$

VIII. Overall error rate

It is a measurement of how many incorrect predictions the model made about the test data as a whole.

$$\text{Overall error rate} = \frac{FP + FN}{\text{Test Set}}$$

4. Results analysis and discussions

The effectiveness of deep learning CNN models in AD diagnosis can be assessed from their performance metrics. For better investigations, the overall outcomes of this study have been formatted into

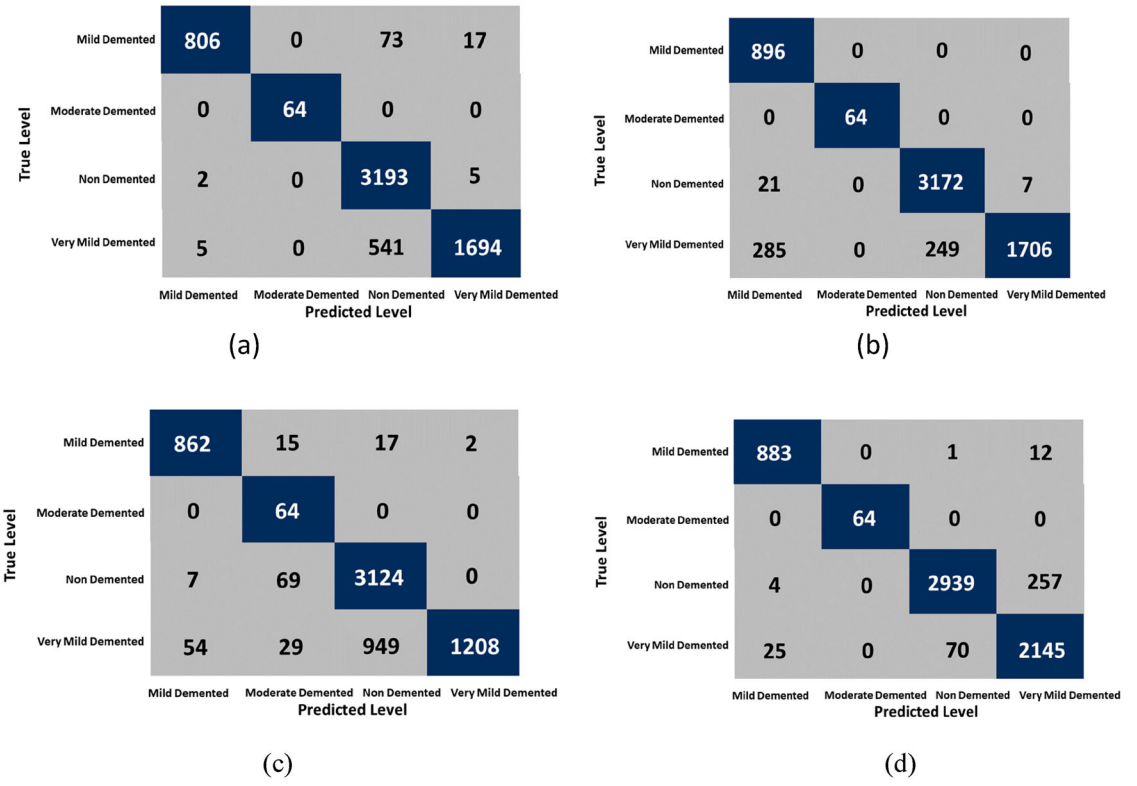


Figure 10. (a, b) Confusion matrices of ResNet152V2 for Dataset1 and Dataset2, respectively, and (c, d) confusion matrices of DenseNet201 for Dataset1 and Dataset2, respectively.

four subsequent phases. Phase 1 represents the individual model's performance on both datasets. Phase 2 demonstrates the ensemble model's performance on both datasets. Phase 3 provides a comparison between individual and ensemble models for both datasets and the state of the art approaches. Finally, Phase 4 provides a Grad-Cam image analysis to better understand how the deep learning models localized the lesions for feature extraction.

The confusion matrix as depicted in Figure 10 for the ResNet152 architecture in AD detection highlights the model's adeptness in identifying non-demented cases (3199 total, 3172 true positives) and moderate demented cases (64 true positives out of 64). While the model demonstrates proficiency in recognizing non-demented and moderate demented cases, it encounters difficulty distinguishing between mild and very mild demented cases. Similarly, for DenseNet201 model architecture, the model excels at accurately identifying non-demented cases, with a true positive rate of 2939 out of 3200. It effectively classifies all cases of moderate demented case, correctly identifying all 64 instances. The model demonstrates proficiency in distinguishing between non-demented and moderate demented cases.

4.1. Phase 1: individual model performance on both datasets

4.1.1. Performance of ResNet152V2 and DenseNet201 architecture on Dataset1

Dataset1 that includes 4096 training images and 1025 testing images. Using iterative adjustments, accuracy reached 89.95%, with precision, recall, and F1 score around 95%, 91%, and 92% for ResNet152V2 model and accuracy, precision, recall, and F1-score of 82.15%, 86%, 82% and 81% for DenseNet201 model. For ResNet152V2 architecture non-demented cases (3193 out of 3200 true positives) and all moderate demented cases (64 true positives) are accurately identified by the algorithm. Despite its ability to distinguish between non-demented and moderate demented, it struggles to recognize mild and very mild demented cases. Similarly, for DernseNet201 the model demonstrates accuracy in correctly identifying non-demented cases, with a true positive rate of 3124 out of 3131. However, it encounters challenges in accurately detecting cases of mild and very mild demented classes, with notable false positives in the former (15 out of 877 mild demented cases) and false negatives in the latter (949 out of 1950 very mild demented cases). Additionally, the model effectively classifies all cases of moderate demented, correctly identifying all 64 instances. Despite these challenges, the model exhibits proficiency in distinguishing cases of non-demented and moderate demented.

ResNet152V2 architecture's receiver operating characteristic (ROC) curve shows its ability to diagnose AD at different stages with having distinct ROC curve areas as shown in Figure 11. These values are 0.91 for mild demented, 0.96 for moderate, 0.88 for non-demented, and 0.83 for very mild. These large areas under the ROC curves demonstrate the model's accuracy in identifying AD. Also, for DenseNet201 model architecture the ROC curve areas are 0.81 for mild demented, 0.96 for moderate demented, 0.78 for non-demented, and 0.88 for very mild demented. The overall error rate of the ResNet152V2 and DenseNet201 models were found as 10.05% and 17.84% for Dataset1, respectively.

4.1.2. Performance of ResNet152V2 and DenseNet201 architecture on Dataset2

The ResNet152V2 and DenseNet201 architecture was applied to Dataset2, which encompasses 40,384 MRI images. The initial dataset underwent augmentation to enhance its real images, promoting improved generalization and mitigating overfitting to specific dataset patterns as mentioned in Table 2. The dataset was partitioned for training, validation, and testing, comprising 27,187, 6797, and

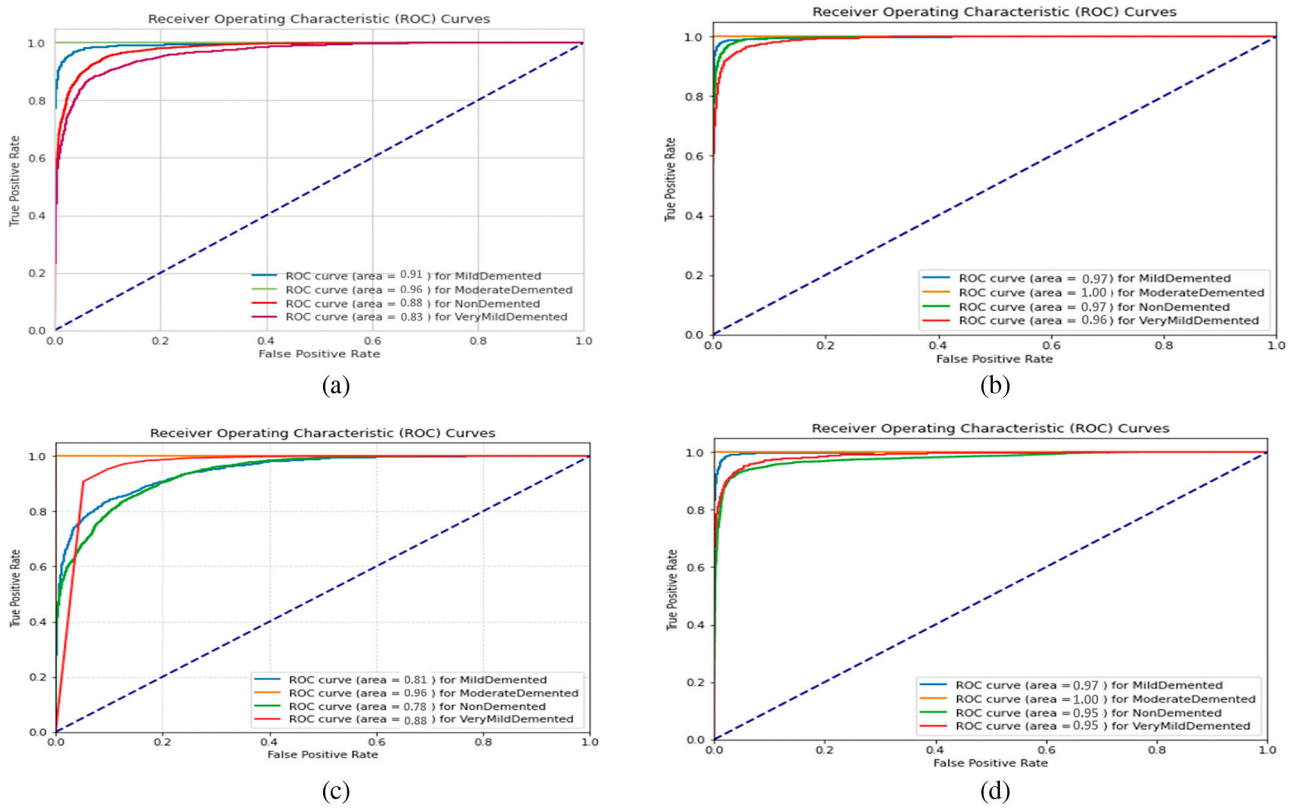


Figure 11. (a, b) ROC curves of ResNet152V2 for Dataset1 and Dataset2, respectively, and (c, d) ROC curves of DenseNet201 for Dataset1 and Dataset2, respectively.

Table 4. ResNet152V2 model performance on Dataset1 and Dataset2.

Model	Dataset	Classes	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet152V2	Dataset1	MildDemented	89.95	99.13	89.95	94.32
		ModerateDemented		100	100	100
		NonDemented		83.87	99.78	91.13
		VeryMildDemted		98.71	75.62	85.64
	Dataset2	MildDemented	97.92	100	94.86	97.36
		ModerateDemented		100	100	100
		ModerateDemented		97.69	99.28	98.48
		VeryMildDemted		97.40	97.14	97.27

6400 images, respectively. Training concluded after 70 epochs for ResNet152V2 model and 56 epochs for DenseNet201 model due to no discernible accuracy improvement. Through iterative adjustments, for ResNet152V2 model we got an accuracy of 97.92%, with precision, recall, and F1 score hovering at approximately 98%, 97%, and 98%, respectively. Similarly, for DenseNet201 model we got an accuracy of 95.03%, with precision, recall, and F1 score hovering at approximately 96%, 95%, and 95%, respectively. The overall error rate of the ResNet152V2 and DenseNet201 models were found as 8.78% and 5.77% for Dataset2, respectively. Tables 4 and 5 show the performance of ResNet152V2 and DenseNet201 models on Dataset1 and Dataset2.

The analysis of the ROC curves of the ResNet152V2 architecture signifies its robust discriminatory capability in discerning distinct stages of AD for Dataset2 also as shown in Figure 11. Notably, the ROC curve areas are excellent, recording values of 0.97 for mild demented, 1.00 for moderate demented, 0.97 for non-demented, and 0.96 for very mild demented. These substantial areas under the ROC curves underscore the model's efficacy in accurately distinguishing between various stages of AD. Similarly, for the DenseNet210 model, the ROC curve areas are perfect (AUC = 0.97) for mild demented and (AUC = 1.00) moderate demented, signifying flawless discriminatory power in distinguishing these categories. Additionally, the model achieves a near-perfect AUC for non-demented with AUC of 0.95. For very mild demented, the AUC is also 0.95, denoting a robust capacity to differentiate this stage of AD. These AUC values collectively underscore the DenseNet201 architecture's effectiveness in accurately distinguishing between various stages of AD, highlighting its potential for reliable clinical applications.

4.2. Phase 2: ensemble model performance on both datasets

The Ensemble architecture, which is made up of ResNet152V2 and DenseNet201 using a weighted ensemble technique, is very good at finding the AD stage. Figure 12 shows the confusion metrics for Ensemble model for Dataset1 and Dataset2 and Table 6 shows the corresponding performance matrices. It has high accuracy of 98%, precision of 99%, recall of 98%, and F1 score of 99% on Dataset 2, while on Dataset1 the performance metrics are accuracy of 93%, precision of 94%, recall of 93%, and F1 score of 93%. The overall error rate of the ensemble model was found as 7.38% and 1.56% for Dataset1 and Dataset2, respectively. Despite encountering challenges

Table 5. DenseNet201 model performance on Dataset1 and Dataset2.

Model	Dataset	Classes	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DenseNet201	Dataset1	MildDemented	82.15	93.39	96.20	94.77
		ModerateDemented		36.15	100	53.1
		NonDemented		76.38	97.62	85.70
		VeryMildDemted		99.83	53.92	70.02
	Dataset2	MildDemented	95.03	99.88	96.87	98.35
		ModerateDemented		100	100	100
		ModerateDemented		99.62	91.25	95.25
		VeryMildDemted		87.93	99.55	93.38

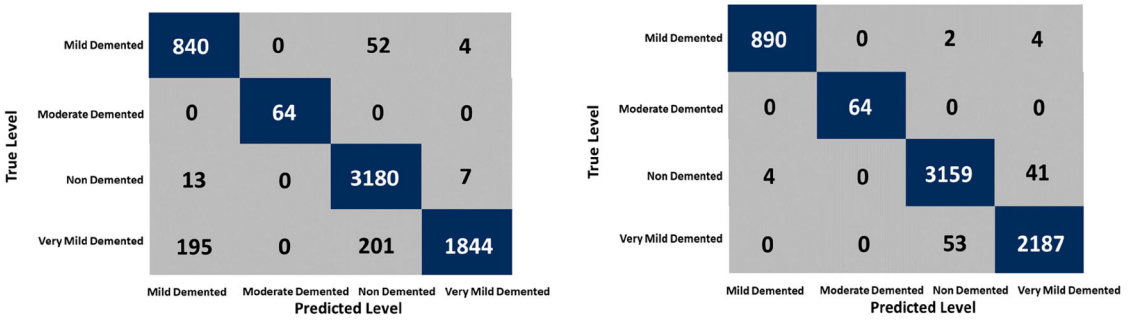


Figure 12. (a, b) Confusion matrices for Ensemble model for Dataset1 and Dataset2.

in detecting mild and very mild demented cases, the ensemble1 model for Dataset1 excels in recognizing non-demented cases (3159 out of 3200 true positives) and accurately identifies all moderate demented instances (64 true positives out of 64).

For Dataset 2, the ensemble model exhibits slightly low proficiency in correctly identifying mild demented cases, achieving a precision of 0.94 with 840 true positives, while encountering 52 false positives and 4 false negatives. For moderate demented cases, the model demonstrates perfect precision, recall, and F1 score, accurately classifying all 64 instances. In the non-demented category, the Ensemble model achieves precision of 0.99, for identifying 3180 true positives, with 13 false positives and 7 false negatives. However, challenges emerge in accurately distinguishing between mild and very mild dementia, with a higher number of misclassifications observed in the latter category.

Figure 13 shows the ROC curves of the Ensemble model for Dataset1 and Dataset2. The ROC curve analysis for Dataset 2 shows that Ensemble is very good at finding the difference between groups, with perfect ROC curves for each category: mildly demented, moderately demented, non-demented, and very mildly demented. This uniform perfection underscores the architecture's exceptional accuracy and reliability, positioning Ensemble as a promising tool for AD detection in clinical applications. On the other hand, the ROC curve areas for Dataset1 are 0.92 for mild demented, 0.84 for moderate demented, 0.91 for non-demented, and 0.89 for very mild demented. These AUC values signify the model's less effectiveness in distinguishing between various stages of AD. The reason behind the lower performance for Dataset1 is that it has not used during the training of the model.

4.3. Phase 3: comparison between individual and ensemble models on both dataset and state of the art

In this subsection, we conduct a comparative analysis of the individual performance of the ResNet152V2 and DenseNet201 models, as well as their collective ensemble model, with a focus on metrics such as accuracy, precision, recall, and F1 score. On Dataset2, the ResNet152V2, DenseNet201, and ensemble1 models achieve accuracies of 97.92%, 95.03%, and 98%, respectively, which is impressive performance shown in Table 7. The accuracy rates on Dataset 1, on the other hand, are significantly lower at 89.95%, 82.15%, and 93%. This performance gap between the two MRI datasets highlights the significance of more diversity, regularization effects, and better generalization. Incorporating a large number of augmented MRI images alongside real MRI images and model training using Dataset 2 are the reason for its better performance.

We have conducted a statistical analysis for performance significance of the proposed Ensemble ResDenseNet using the analysis of variance (ANOVA) approach which is shown in Table 8. It is observed that, for all the performance metrics (accuracy, precision, recall, F1 score), the p-values are greater than 0.05, and the F-statistics are less than the F-critical value. This indicates that there are slight differences in these metrics among the values tested.

Table 9 presents a comprehensive overview of our method's assessment in relation to that of others. After conducting the research, we have selected the Ensemble ResDenseNet model of our study for comparison with other related state of the art works. Various recent deep learning-based techniques [21,30–46 reported to classify and identify AD as described in this table mentioning theirs used datasets, methodology, and performance. It is seen that our proposed Ensemble ResDenseNet model framework performed significantly comparing with the other state of the art works.

4.4. Phase 4: Grad-Cam image analysis

To instill confidence in intelligent systems and facilitate their meaningful integration into practical applications, it is imperative to construct 'understandable' models capable of elucidating the rationale behind their predictions. Striking a balance between accuracy and interpretability is a recurrent challenge, with traditional rule-based or expert systems being highly interpretable but often lacking in accuracy or robustness. On the other hand, decomposable pipelines, in which each step is carefully planned, are thought to be easier to understand because they have parts that can be explained separately

Table 6. Ensemble model performance on dataset1 and dataset2.

Model	Dataset	Classes	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Ensemble ResDenseNet	Dataset1	MildDemented	93	81	94	87
		ModerateDemented		100	100	100
		NonDemented		93	99	96
		VeryMildDemted		99	85	91
	Dataset2	MildDemented	98	100	99	100
		ModerateDemented		100	100	100
		ModerateDemented		98	99	99
		VeryMildDemted		98	98	98

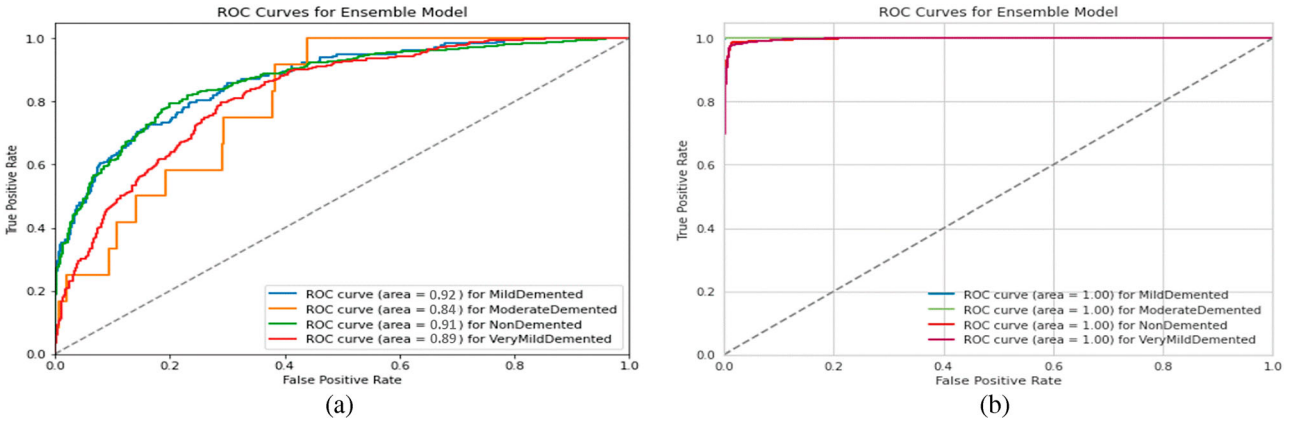


Figure 13. ROC curves of Ensemble model for Dataset1 and Dataset2.

Table 7. Comparative analysis between Individual and ensemble models.

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Dataset1	ResNet152V2	89.95	95.43	91.34	92.27
	DenseNet201	82.15	76.44	86.93	75.90
	Ensemble1	93	93.25	94.5	93.2
Dataset2	ResNet152V2	97.92	98.77	97.82	98.27
	DenseNet201	95.03	96.86	96.91	96.74
	Ensemble1	**98**	**99**	**99**	**98**

Table 8. ANOVA on the performance of the proposed Ensemble ResDenseNet.

Parameters	Accuracy	Precision	Recall	F1-score
p-value	0.59	0.5	0.68	0.57
F-statistic	0.61	0.86	0.43	0.67
F-critical	9.55	9.55	9.55	9.55

[52]. However, the advent of deep learning introduces a trade-off whereby interpretability is sacrificed for heightened performance achieved through increased abstraction (more layers) and seamless integration (end-to-end training). Notably, recent models such as DenseNet201 and ResNet152V2 have exhibited remarkable proficiency across various intricate tasks. Yet, the intricate nature of these models makes them challenging to interpret. Consequently, the realm of deep learning is presently navigating the delicate equilibrium between interpretability and accuracy. Hence, in this study, we have used a technique to generate 'visual explanations' for decisions from the used CNN-based models ResNet152V2 and DenseNet201. We have used gradient-weighted class activation mapping (Grad-CAM) [52] method that leverages the gradients associated with AD in brain MRI images, which propagate to the final convolutional layer. Grad-CAM has proven usefulness to interpret the effective functionality of machine learning models in automatic medical image diagnosis [53,54]. This process generates a coarse localization map that highlights pivotal regions in the image for predicting the concept. As a consequence, crucial regions within Alzheimer's images, corresponding to any pertinent decision, are visualized in high-resolution detail, even in instances where the image encapsulates evidence for multiple conceivable concepts, as illustrated in Figure 14. The resultant heatmap serves as a visual guide, elucidating areas where the model concentrates its attention during the predictive process. Warmer hues in the heatmap denote regions of heightened activation and significance in predicting AD stage. This technique contributes to bridging the interpretability gap in deep learning models, rendering them more accessible and fostering a deeper understanding of their decision-making processes, particularly in the context of AD diagnosis.

4.5. Limitations and future perspectives

In this study, our proposed framework exhibits high efficacy in performing four-class classification tasks associated with AD. Notably, the weighted ensemble of ResNet152V2 and DenseNet201 within our framework, applied to Dataset2, demonstrates superior performance compared to both individual models and other ensemble methodologies.

Despite the promising outcomes achieved by our method, certain limitations persist. First, although the model exhibits high accuracy in AD detection, challenges arise in identifying mild and very mild demented cases, suggesting avenues for model refinement. This difficulty may stem from the subtle nature of anatomical changes in these cases, rendering them less observable. Second, the current methodology lacks the ability to directly discern brain lesion structures through up-sampling or deconvolution, given the reliance on the cascade of ResNet152V2 and DenseNet201 models. To address this, alternative models such as U-Net or other semantic segmentation models with up-sampling or deconvolution layers could be considered. A crucial aspect of our study explores whether further processing of features extracted by the CNN can

Table 9. Comparative analysis between our proposed model and previous studies on AD detection.

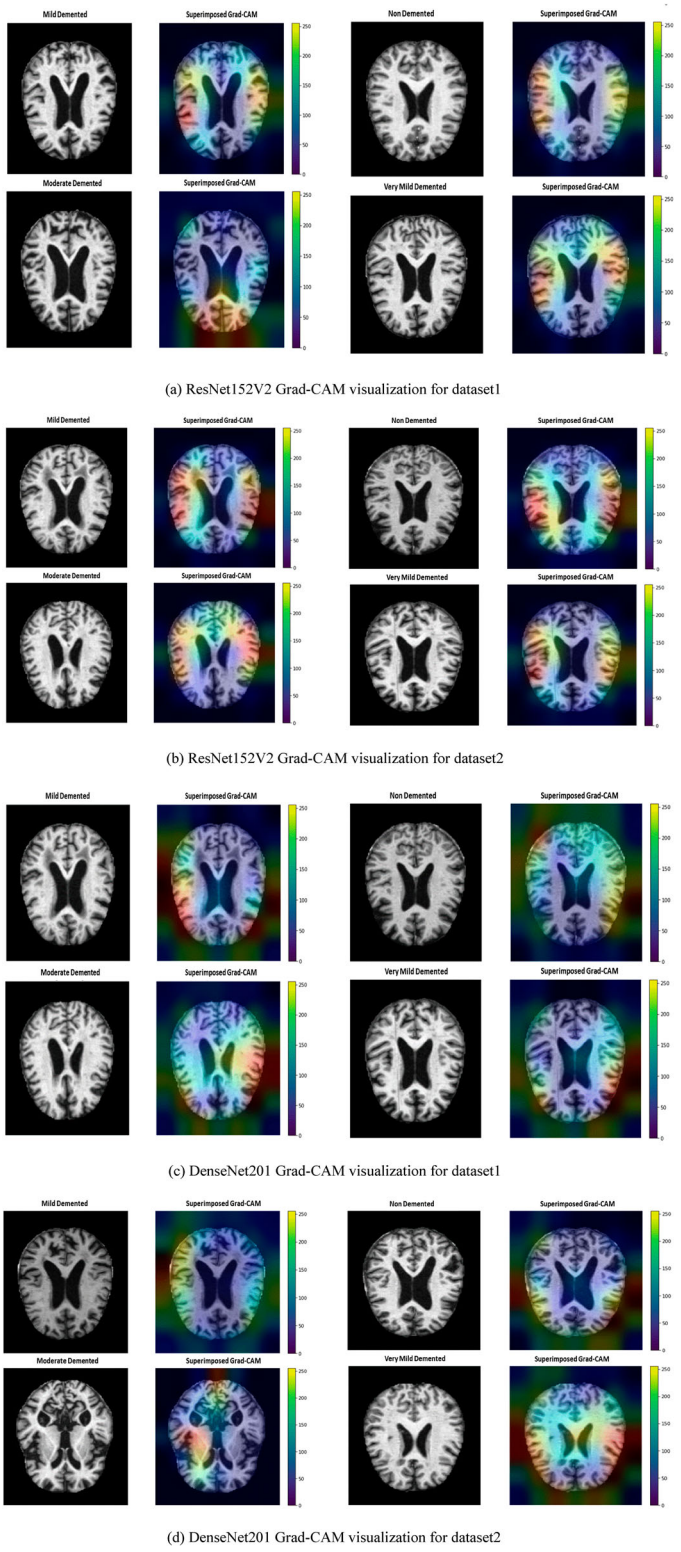
Author	Data	Dataset	Preprocessing Techniques	Method	Performance
Ortiz et al. [30]	MRI	ADNI	Resize, Normalization, Voxel preselection	Deep belief Network (DBN)	Accuracy = 90.00%
Shi et al. [31]	MRI	ADNI	Normalization, segmentation, T-test based patch extraction	De-noising Sparse Encoder (DASE)	Accuracy = 91.95%
Islam et al. [21]	MRI	OASIS	Data augmentation	CNN	Accuracy = 95.3%
Faturrahman et al. [32]	MRI	OASIS	Voxel-based morphometric (VBM) feature extraction	Deep Belief Network (DBN)	Accuracy = 91.76%
Nawaz et al. [33]	MRI	OASIS	Resize, feature extraction	AlexNet	Accuracy = 92.85%
Liang et al. [34]	MRI	KACD	Feature extraction	ADGNET	F1-score = 99.61%
Sun et al. [35]	MRI	ADNI	Normalization, encoding, K-fold ($k = 5$) cross validation and divide the dataset randomly	ResNet50	Accuracy = 97.10%, Precision = 95.5%, Recall = 95.3%
Hon et al. [36]	MRI	OASIS	Resize, Entropy based sorting algorithm for finding out most informative images	Inception V4	Accuracy = 96.25%
Katabathula et al. [37]	MRI	ADNI	Segmentation, K fold ($k = 5$) cross validation	DenseCNN2	Accuracy: 92.50%,
Liu et al. [38]	MRI, PET	ADNI + MRIAD	Normalization using non-linear image registration technique, t-test for dimensionality reduction	3D CNN	Accuracy: 91.40%
Ramzan et al. [39]	RS-fMRI	ADNI	Brain extraction, Motion correlation, Normalization, High pass filter, Image registration	ResNet18	Accuracy: 97.92% and 97.88 respectively (with and without fine-tuning)
Mendonca et al. [40]	MRI	ADNI	Segmentation, Texture feature extraction, graph edge removes	SVM	Accuracy: 91.40%
Ismail et al. [41]	MRI, PET	ADNI	Feature extraction,	Multiaz-Net	Accuracy: 93.3%
Raju et al. [42]	MRI	ADNI	Motion correction, Non-Uniform Intensity Normalization, Tailairach Transformation, Intensity Normalization, and Skull Stripping	CNN	Accuracy: 97.77%
Yildirim et al. [43]	MRI	Alzheimer's dataset (4 classes of images)	Resize, Normalization	DenseNet201, VGG16, AlexNet, ResNet50, Proposed model	Accuracy: 87%, 78%, 86%, 78% and 90% respectively
Balaji et al. [44]	MRI, PET	Alzheimer's Dataset1 (2 Class)	Resizing, reshaping, sharpening, ACO denoising	Hybrid CNN-LSTM	Accuracy = 92.80% (MRI)
Dong Nguyen et al. [45]	MRI	ADNI	Flip, crop, and rotation	Ensemble of 3D-ResNet and XGBoost	AUC score = 96%
Peixian et al. [46]	MRI	ADNI	Resampling, cropping, surrounding background removing	Deep Broad Ensemble Model	Accuracy = 93.58%
Proposed model Ensemble ResDenseNet	MRI	Alzheimer's dataset (4 classes and total 40384 images)	Resizing, normalization, augmentation, one-hot encoding	Ensemble	Accuracy: 98%

enhance early AD diagnosis. Structural and functional information gleaned from the brain using CNNs are subjected to weighted ensemble techniques might lead to improved generalization and interpretability. Furthermore, the integration of complementary features from individual models through assembling yields heightened performance in AD detection. Therefore, the exclusion of PET data in our study represents a potential limitation, as incorporating such data could offer valuable complementary information regarding disease evolution. We used here traditional augmentation technique while DCGAN can be a good option for further research.

Given our exclusive focus on brain MRI images, there is a concern that the evolving nature of the disease may not be fully captured. Future work may necessitate longitudinal studies or the inclusion of temporal information for a more precise assessment. Additionally, leveraging state-of-the-art techniques that integrate numerical and visual features through computer vision methodologies could prove beneficial. By maximizing the retention of structural data in the brain during processing, commonalities, and shared data among different features can be unveiled, facilitating prediction and diagnosis in practical healthcare applications. These insights and limitations pave the way for exploration in our future work.

5. Conclusion

This research presents a weighted ensemble framework for the categorization of medical images, with a specific focus on the detection of AD stages. The proposed framework leverages deep learning CNN architectures, namely ResNet152V2 and DenseNet201. The classification task encompasses four stages of AD, forming a multi-class classification paradigm. Two distinct datasets are employed for experimentation: Dataset1, comprising original images, and Dataset2, encompassing both original and augmented images to enhance dataset diversity and improve generalization. Transfer learning principles are applied to exploit the capabilities of pre-trained models, specifically fine-tuning ResNet152V2 and DenseNet201 models for multi-class medical image classifications. Additionally, the paper uses Grad-CAM image analysis for enhancing the interpretability of the decision-making process of the models. Evaluation and comparison of the two methods involve the utilization of various performance metrics. The rigorous investigational outcomes affirm



(a) ResNet152V2 Grad-CAM visualization for dataset1

(b) ResNet152V2 Grad-CAM visualization for dataset2

(c) DenseNet201 Grad-CAM visualization for dataset1

(d) DenseNet201 Grad-CAM visualization for dataset2

Figure 14. Grad-CAM visualization of ResNet152V2 and DenseNet201 models.

the suitability of the proposed architecture. Remarkably, the proposed framework achieves a high accuracy of 98%, accompanied by precision, recall, and F1 score values of 99%, 99%, and 98%, respectively, for classification tasks. Whereas, the fine-tuned ResNet152V2 and DenseNet201 models achieve accuracies of 89.95%, 97.92%, and 82.15%, 95.03%, indicating the effectiveness of the proposed framework for multi-class classification. The proposed framework might be useful for clinical applications of AD staging.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

In this study, two publicly available datasets have been used which are available at Kaggle repository and accessible at: https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images and https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset.

ORCID

A. B. M. Aowlad Hossain http://orcid.org/0000-0002-2559-2781

References

[1] Su Z, Bentley BL, McDonnell D, et al. 6G and artificial intelligence technologies for dementia care: literature review and practical analysis. J Med Internet Res. 2022;24(4):e30503. doi:10.2196/30503

[2] Kadir DS, Sadik GM. Chronicles of Alzheimer's disease: a medicinal & therapeutic overview in Bangladeshi aspect. J Pharm Res Int. 2019;30(5):1–12. doi:10.9734/JPRI/2019/V30I530282

[3] Itzhaki R. Herpes simplex virus type 1, apolipoprotein E and Alzheimer 'disease. Herpes. 2004;11:77A–82A.

[4] Braak H, Braak E. Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiol Aging. 1997;18(4):351–357. doi:10.1016/S0197-4580(97)00056-0

[5] Fisher DW, Bennett DA, Dong H. Sexual dimorphism in predisposition to Alzheimer's disease. Neurobiol Aging. 2018;70:308–324. doi:10.1016/j.neurobiolaging.2018.04.004

[6] Bron EE, Smits M, Van Der Flier WM, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage. 2015;111:562–579. doi:10.1016/j.neuroimage.2015.01.048

[7] Klöppel S, Stonnington CM, Barnes J, et al. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. Brain. 2008;131(11):2969–2974. doi:10.1093/brain/awn239

[8] Yamanakkanavar N, Choi JY, Lee B. MRI segmentation and classification of human brain using deep learning for diagnosis of Alzheimer's disease: a survey. Sensors. 2020;20(11):3243. doi:10.3390/s20113243

[9] Noor MBT, Zenia NZ, Kaiser MS, et al. Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia. Brain Inform. 2020;7(1):1–21. doi:10.1186/s40708-020-00112-2.

[10] Helaly HA, Badawy M, Haikal AY. Deep learning approach for early detection of Alzheimer's disease. Cognit Comput. 2022;14(5):1711–1727. doi:10.1007/s12559-021-09946-2

[11] Oktavian MW, Yudistira N, Ridok A. Classification of Alzheimer's disease using the convolutional neural network (CNN) with transfer learning and weighted loss. IAENG Int J Comput Sci. 2023;50(3):1–10. doi: 10.48550/arXiv.2207.01584.

[12] Lu D, Popuri K, Ding GW, et al. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. Sci Rep. 2018;8(1):5697. doi:10.1038/s41598-018-22871-z

[13] Ge C, Qu Q, Gu IY-H, et al. Multiscale deep convolutional networks for characterization and detection of Alzheimer's disease using MR images. IEEE International Conference on Image Processing (ICIP); 2019. p. 789–793. doi:10.1109/ICIP.2019.8803731

[14] Khagi B, Lee B, Pyun J-Y, et al. CNN models performance analysis on MRI images of OASIS dataset for distinction between Healthy and Alzheimer's patient. International Conference on Electronics, Information, and Communication (ICEIC); 2019. p. 1–4. doi:10.23919/ELINFOCOM.2019.8706339

[15] Wang Y, Yang Y, Guo X, et al. A novel multimodal MRI analysis for Alzheimer's disease based on convolutional neural network. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018. p. 754–757. doi:10.1109/EMBC.2018.8512372

[16] Spasov SE, Passamonti L, Duggento A, et al. A multi-modal convolutional neural network framework for the prediction of Alzheimer's disease. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018. p. 1271–1274. doi:10.1109/EMBC.2018.8512468

[17] Korolev S, Safiullin A, Belyaev M, et al. Residual and plain convolutional neural networks for 3D brain MRI classification. IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017); 2017. p. 835–838. doi:10.1109/ISBI.2017.7950647

[18] Goenka N, Sharma AK, Tiwari S, et al. A regularized volumetric ConvNet based Alzheimer detection using T1-weighted MRI images. Cogent Eng. 2024;11(1):2314872. doi:10.1080/23311916.2024.2314872
[19] Maringanti HB, Mishra M, Pradhan S. Machine learning and deep learning models for early-stage detection of Alzheimer's disease and its proliferation in human brain. Artificial Intelligence for Neurological Disorders. In: Abraham A, Dash S, Pani SK, et al., editors. Artificial Intelligence for Neurological Disorders. Massachusetts, United States: Academic Press; 2023. p. 49–60. doi:10.1016/B978-0-323-90277-9.00024-9.
[20] Qasim Abbas S, Chi L, Chen YPP. Transformed domain convolutional neural network for Alzheimer's disease diagnosis using structural MRI. Pattern Recognit. 2023;133:109031. doi:10.1016/j.patcog.2022.109031
[21] Mamun M, Shawkat SB, Ahammed MS, et al. Deep learning based model for Alzheimer's disease detection using brain MRI images. IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON); 2022. p. 0510–0516. doi:10.1109/UEMCON54665.2022.9965730
[22] Allioui H, Sadgal M, Elfazziki A. Utilization of a convolutional method for Alzheimer disease diagnosis. Mach Vis Appl. 2020;31(4):1–19. doi:10.1007/s00138-020-01074-5.
[23] Hussain E, Hasan M, Hassan SZ, et al. Deep learning based binary classification for ALzheimer's disease detection using brain MRI images. 15th IEEE Conference on Industrial Electronics and Applications (ICIEA); 2020. p. 1115–1120. doi:10.1109/ICIEA48937.2020.9248213
[24] Bi X, Wang H. Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. Neural Netw. 2019;114:119–135. doi:10.1016/j.neunet.2019.02.005
[25] Ebrahimi-Ghahnavieh A, Luo S, Chiong R. Transfer learning for Alzheimer's disease detection on MRI images. IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT); 2019. p. 133–138. doi:10.1109/ICIAICT.2019.8784845
[26] De A, Chowdhury AS. DTI based Alzheimer's disease classification with rank modulated fusion of CNNs and random forest. Expert Syst Appl. 2021;169:114338. doi:10.1016/j.eswa.2020.114338
[27] Abuhmed T, El-Sappagh S, Alonso JM. Robust hybrid deep learning models for Alzheimer's progression detection. Knowl Based Syst 2021; 213:106688. doi:10.1016/j.knosys.2020.106688
[28] Fang X, Liu Z, Xu M. Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis. IET Image Proc. 2020; 14(2):318–326. doi:10.1049/iet-ipr.2019.0617
[29] Chaya JD, Usha RN. Predictive analysis by ensemble classifier with machine learning models. Int J Comput Appl. 2023;45(1):19–26. doi:10.1080/1206212X.2019.1675019
[30] Ortiz A, Munilla J, Gorriz JM, et al. Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. Int J Neural Syst. 2016;26(07):1650025. doi:10.1142/S0129065716500258
[31] Shi B, Chen Y, Zhang P, et al. Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis. Pattern Recognit. 2017;63:487–498. doi:10.1016/j.patcog.2016.09.032
[32] Faturrahman M, Wasito I, Hanifah N, et al. Structural MRI classification for Alzheimer's disease detection using deep belief network. 11th International Conference on Information & Communication Technology and System (ICTS); 2017. p. 37–42. doi:10.1109/ICTS.2017.8265643
[33] Nawaz H, Maqsood M, Afzal S, et al. A deep feature-based real-time system for Alzheimer disease stage detection. Multimed Tools Appl. 2021;80(28-29):35789–35807. doi:10.1007/s11042-020-09087-y
[34] Liang S, Gu Y. Computer-aided diagnosis of Alzheimer's disease through weak supervision deep learning framework with attention mechanism. Sensors. 2021;21(1):220. doi:10.3390/s21010220
[35] Sun H, Wang A, Wang W, et al. An improved deep residual network prediction model for the early diagnosis of Alzheimer's disease. Sensors. 2021;21(12):4182. doi:10.3390/s21124182
[36] Hon M, Khan NM. Towards Alzheimer's disease classification through transfer learning. 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017; vol. 2017-January. p. 1166–1169. doi:10.1109/BIBM.2017.8217822
[37] Katabathula S, Wang Q, Xu R. Predict Alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations. Alzheimers Res Ther. 2021;13(1):1–9. doi:10.1186/s13195-021-00837-0
[38] Liu M, Zhang J, Adeli E, et al. Landmark-based deep multi-instance learning for brain disease diagnosis. Med Image Anal. 2018;43:157–168. doi:10.1016/j.media.2017.10.005
[39] Ramzan F, Khan MUG, Rehmat A, et al. A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks. J Med Syst. 2020;44(2):1–16. doi:10.1007/s10916-019-1475-2
[40] De Mendonça LJC, Ferrari RJ, and Alzheimer's Disease Neuroimaging Initiative. Alzheimer's disease classification based on graph kernel SVMs constructed with 3D texture features extracted from MR images. Expert Syst Appl. 2023;211:118633. doi:10.1016/j.eswa.2022.118633
[41] Ismail WN, Fathimathul Rajeena PP, Ali MA. A meta-heuristic multi-objective optimization method for Alzheimer's disease detection based on multi-modal data. Mathematics. 2023;11(4):957. doi:10.3390/math11040957
[42] Raju M, Gopi VP, Anitha VS, et al. Multi-class diagnosis of Alzheimer's disease using cascaded three dimensional-convolutional neural network. Phys Eng Sci Med. 2020;43(4):1219–1228. doi:10.1007/s13246-020-00924-w
[43] Yildirim M, Cinar A. Classification of Alzheimer's disease MRI images with CNN based hybrid method. Ing Syst Inf. 2020;25(4):413–418. doi:10.18280/isi.250402
[44] Balaji P, Chaurasia MA, Bilfaqih SM, et al. Hybridized deep learning approach for detecting Alzheimer's disease. Biomedicines. 2023;11(1):149. doi:10.3390/biomedicines11010149
[45] Nguyen D, Nguyen H, Ong H, et al. Ensemble learning using traditional machine learning and deep neural network for diagnosis of Alzheimer's disease. IBRO Neurosci Rep. 2022;13:255–263. doi:10.1016/j.ibneur.2022.08.010
[46] Ma P, Wang J, Zhou Z, et al. Development and validation of a deep-broad ensemble model for early detection of Alzheimer's disease. Front Neurosci. 2023;17:1137557. doi:10.3389/fnins.2023.1137557
[47] Pinamonti M. Alzheimer MRI 4 classes dataset [dataset]. 2022 Jan 4. https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images.
[48] Uraninjo. Augmented Alzheimer MRI Dataset [dataset]. 2022 Sep 20. https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset.
[49] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition. p. 770–778.
[50] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 4700–4708.
[51] Hasan R, Azmat Ullah SM, Nandi A, et al. Improving pneumonia diagnosis: a deep transfer learning CNN ensemble approach for accurate chest x-ray image analysis. International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD); 2023. p. 109–113. doi:10.1109/ICICT4SD59951.2023.10303322
[52] Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision; 2017. p. 618–626.
[53] Foysal M, Aowlad Hossain ABM, Yassine A, et al. Detection of COVID-19 case from chest CT images using deformable deep convolutional neural network. J Healthc Eng. 2023;2023(4301745):1–12.
[54] Chien J-C, Lee J-D, Hu C-S, et al. The usefulness of gradient-weighted cam in assisting medical diagnoses. Appl Sci. 2022;12(15):7748. doi:10.3390/app12157748