International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predicting Alzheimer's Onset: Leveraging Pretrained Deep Neural Networks and Transfer Learning for Early Detection

Naveen. N [1*], Nagaraj. G. Cholli [2]

Submitted: 02/05/2024 **Revised**: 15/06/2024 **Accepted**: 22/06/2024

Abstract Over the past few years, medical image processing has increased in use of deep learning algorithms, especially in the analysis of magnetic resonance (MR) scans. MRI is a crucial diagnostic tool for Alzheimer's disease (AD), a prevalent type of dementia that ranks seventh among fatal illnesses globally. As there is no known cure for Alzheimer's disease, early detection and intervention are vital to prevent its irreversible progression. This study proposes a comprehensive framework for detecting Alzheimer's disease that employs convolutional neural networks (CNNs) and deep learning approaches. We applied transfer learning to pretrained deep learning models rather than training them from scratch. Three distinct pretrained CNN models (VGG-19, ResNet-50, and Inception V3) with a fine-tuned transfer learning approach were used for five-way classification of AD. We employed the ADNI dataset, which includes MRI scans from 608 patients across five classes: Alzheimer's disease (AD), early mild cognitive impairment (EMCI), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI), and normal control (NC). The models' performance was evaluated based on eight metrics: accuracy, precision, sensitivity/recall, specificity, error rate, false positive rate, F1-score, and kappa. Our findings indicate that the ResNet-50 architecture outperformed other pretrained models, achieving the highest overall accuracy of 98.7% for multiclass AD classification. Additionally, the ResNet-50 model excelled in classifying the EMCI category with an accuracy of 99.25%, indicating its effectiveness in detecting early signs of memory impairment. The proposed framework surpasses the performance of previous studies in terms of overall accuracy, sensitivity, and specificity, setting a new benchmark for five-way AD classification. The outcomes of this study will contribute significantly to early prevention efforts by enabling Alzheimer's disease to be detected before it progresses irreversibly. Furthermore, this research represents a promising approach for improving the early detection and classification of Alzheimer's disease using deep learning methods with MRI data.

Keywords: *Deep Neural Network (DNN), Transfer Learning (TL), Convolutional Neural Network (CNN), VGG, Inception, ResNet, Magnetic Resonance Imaging (MRI).*

1. Introduction

Alzheimer's disease (AD) is a progressive neurological illness that gradually hampers cognitive ability, leading to a decline in thought processes and consciousness in affected individuals. This disease has a direct impact on cognitive abilities and overall neurocognition [1], which makes it the leading cause of cognitive impairment, accounting for up to 80% of all dementia cases [2]. Approximately 10 million people are diagnosed with dementia each year worldwide, with more than 55 million individuals affected in 2020. Forecasts indicate an increase to 78 million by 2030 and a staggering increase to 139 million by 2050 [3]. AD dementia accounts for 60–90% of all cases of neurodegenerative diseases [4].

Despite advances in medical research, a specific cure for AD has not been identified [5]. Dementia usually develops gradually over time and tends to worsen progressively, beginning with MCI and culminating in AD. The onset of mild cognitive impairment often serves as an early indicator that an individual is at risk of developing AD [6]. Early recognition of MCI serves as a vital indicator of the risk of developing Alzheimer's disease. Timely diagnosis is crucial for comprehending the progressive nature of the disease and its significant impact on individuals with Alzheimer's disease [7, 8]. Although identifying AD symptoms using clinical measures is feasible, identifying symptoms is a labor-intensive process that requires specialized expertise. Early diagnosis is often difficult for experts unless symptoms are evident.

Advances in neuroimaging methods such as positron emission tomography (PET) and magnetic resonance imaging (MRI) have aided in identifying biological markers linked to AD [9]. MRI, in particular, is widely employed for screening, identifying, and distinguishing Alzheimer's disease [10–12]. MRI provides detailed brain structure information, allowing researchers and doctors to detect irregularities, monitor disease development, and support accurate AD diagnosis. However, these advances have resulted in a vast amount of complex, multidimensional data that requires comprehensive analysis. Computer-aided

[1] *Department of Computer Science & Engineering, M. S. Ramaiah University of Applied Sciences, Visvesvaraya Technological University, Belagavi, India. E-mail: naveenn.cs.et@msruas.ac.in, ORCID iD: https://orcid.org/0000-0001-9146-3251*
[2] *Department of Information Science & Engineering, RV College of Engineering, Visvesvaraya Technological University, Belagavi, India. E-mail: nagaraj.cholli@rvce.edu.in, ORCID iD: https://orcid.org/0000-0001-7409-8272*

machine learning techniques are becoming increasingly popular for this purpose. Medical imaging and computer-aided techniques are the most reliable methods for detecting AD in its early stages. Although most related research has focused on binary classification, distinguishing patients with or without AD, precise classification of patients into different categories, such as cognitively normal (CN), early mild cognitive impairment (EMCI), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI), and Alzheimer's disease (AD), is necessary for effective Alzheimer's disease treatment. While this represents a significant challenge, we are confident that ongoing research will lead to the full resolution of this issue.

With remarkable advancements in technology, we now have a tremendous opportunity to improve the diagnosis of diseases significantly. Significant progress has recently been made in image processing because of the abundance of large-scale and labeled image datasets. ImageNet [13], a dataset containing more than 1.2 million labeled images, has played a critical role in this growth. Convolutional neural networks (CNNs) have achieved remarkable accuracy and improved the categorization of medical images by leveraging these datasets during training [14]. These networks excel at classifying medical images when trained on extensive datasets such as ImageNet, which consists of more than 1000 different data classes. They can be utilized for this task in various ways [15, 16], including training a pretrained network on a vast dataset and tuning it to the specific dataset to be classified. Traditional image descriptors based on primitives have also shown promising performance. To further optimize the process, a more efficient method is to use transfer learning to fine-tune a pretrained CNN on smaller datasets [17, 18]. Such networks are frequently employed in tasks related to computer vision, particularly those involving the detection of objects. This technique has also been proven to be exceptionally effective in cross-domain classification tasks where a CNN trained on natural images is subsequently used to classify medical images [19].

AD is a complex condition that progresses through different stages and can be challenging to diagnose accurately. Close observation by a radiologist and careful clinical assessment are needed but can be time-consuming and costly. We aimed to develop a more efficient and cost-effective method for diagnosing AD in its initial stages using deep CNN models (VGG-19, ResNet-50, and Inception V3) and transfer learning (TL). This approach will not only help to slow the progression of the illness but also reduce the participation of radiologists and the overall cost of diagnosis. This study presents a framework that uses deep neural networks (DNNs) to predict Alzheimer's disease from MRI scans. The framework captures essential visual features crucial for distinguishing between five different AD stages: cognitively normal, early mild cognitive impairment, mild cognitive impairment, late mild cognitive impairment, and Alzheimer's disease. By identifying the disease in its early stages, we can overcome the shortcomings of conventional machine learning approaches and improve the effectiveness of treatment.

The primary research contributions are outlined below:

- We propose and evaluate a fine-tuned transfer learning-based multiclass (five-way) classification framework for the early diagnosis of AD.
- MRI images from the ADNI dataset were analyzed using second-generation DNNs to differentiate between the stages of cognitive decline, including CN, EMCI, MCI, LMCI, and AD.
- Resampling techniques such as oversampling and downsampling are used to overcome the imbalance of the acquired dataset classes.
- To improve the transfer learning technique, we also used several data augmentation methods to increase the diversity of the input data and extract more robust and distinctive features.
- Using eight different performance measures, ResNet-50 provided the highest overall accuracy of 98.7%, representing a new benchmark for the five-way classification of AD.

The remaining sections of the paper are structured as follows. Section 2 reviews the existing related research. Section 3 describes the proposed method. Section 4 presents the experimental setup and results. Section 5 discusses the proposed work, and finally, section 6 presents the research conclusions and outlines possible future directions.

2. Previous Works

AbdulAzeem et al. [20] developed a CNN model from scratch for binary classification of Alzheimer's disease. The model consists of three sets of convolution and pooling layers trained using the Adam optimizer. The final classification layer of the model features the SoftMax activation function. The researchers tested the model's performance using different image sizes, 128 × 128 and 64 × 64, and applied various data augmentation methods with and without dropout. The best results were obtained with images resized to 128 × 128 and dropout disabled images. Model validation was performed by arbitrarily partitioning the dataset into training and testing sets with ratios ranging from 0.1–0.5. Increasing the batch size resulted in better training accuracy, although the accuracy decreased when the batch size exceeded 64. Their model achieved an accuracy of 95.6% in categorizing images into two categories

A study by Liu et al. [21] introduced a novel framework for AD classification. The researchers constructed a CNN with

three convolutional layers, three pooling layers, and two fully connected layers, with SoftMax activation in the final layer. To improve the classification accuracy and prevent overfitting, researchers have also employed transfer learning with the GoogLeNet and AlexNet models. However, the models achieved only moderate accuracy. The GoogLeNet and AlexNet models were subjected to 5-fold cross-validation and 500 epochs of training during the transfer learning process. The GoogLeNet, AlexNet, and CNN models achieved classification accuracy rates of 93.02%, 91.4%, and 78.02%, respectively.

Al-Adhaileh [22] conducted a study to compare the diagnostic performance of two pretrained convolutional neural networks, AlexNet and ResNet50, in detecting AD. The study used the Kaggle Alzheimer's disease dataset, which was partitioned into four classes, and the images were resized to 224 × 224 pixels. The study results showed that, with 34 layers and five max pooling layers, AlexNet achieved a high accuracy rate of 94.53%. ResNet50, which has 177 layers and five max pooling layers, uses the RMSprop optimizer and achieves an accuracy rate of 58.07%. Both models implemented ReLU activation functions and SoftMax for classification. The study showed that AlexNet outperformed ResNet50 in diagnosing AD.

Helaly et al. [23] developed deep CNN models to differentiate between the four stages of AD. The authors employed different approaches: basic CNNs (2D and 3D) and transfer learning. The model's classification accuracy rates are 93.61% for 2D and 95.17% for 3D multiclass AD stage classifications. After fine-tuning, the pretrained VGG-19 model achieved an even higher accuracy of 97% for the four-way classification of AD.

Savaş et al. [24] conducted a study using various CNN models to classify 2182 images in the ADNI dataset. They developed a framework to evaluate the performance of 29 pretrained models on these images. Preprocessing steps included formatting, cleaning the data, and splitting the images before they were fed into the models. During testing, the EfficientNetB0 model was the best performing model, with an impressive accuracy of 92.98%. The EfficientNetB2 and EfficientNetB3 models displayed the best precision, specificity, and sensitivity for AD, with 94.42% and 97.28%, respectively, based on the confusion matrix from the comparison stage.

A study conducted by Antony et al. [25] used VGG-16 and VGG-19 models to detect AD using 780 MR images from the ADNI database. These images were preprocessed through skull removal and augmentation and then resized to 224×224 pixels. The sigmoid function for binary classification activated the last layer of VGG-16, while VGG-19 utilized the softmax activation function.

In a recent study conducted by Raza et al. [26], a DenseNet DL model and transfer learning were used to analyze brain MRI scans for Alzheimer's disease staging via gray matter (GM) measurements. The scans were preprocessed using SPM12 to divide the brain into GM, white matter, and CSF sections. The GM slices were then turned into 2D slices for model training and evaluation. The study successfully modified a pretrained DenseNet model and retrained the final two blocks, achieving an outstanding accuracy of 97.84% in accurately classifying Alzheimer's disease stages.

Hazarika et al. [27] conducted a study to assess the effectiveness of different DNN models in classifying Alzheimer's disease. This study focused on models such as DenseNet-121, Inception-V1/V2/V3, ResNet-50, EfficientNet-B0, VGG-16, MobileNet-V1, AlexNet, VGG-19, LeNet, and Xception. The study showed that DenseNet-121 outperformed the other models in terms of accuracy, recall, precision, and F1 score, despite having the longest computing time. The model achieved an accuracy rate of 86.55%. However, due to the computational complexity of the model, the authors introduced a hybrid approach that combined LeNet and AlexNet, which achieved an overall accuracy of 93.58% and outperformed DenseNet.

A study conducted by Mujahid et al. [28] reported a remarkable deep learning ensemble model that utilized transfer learning methods to identify AD cases from a multiclass dataset. To achieve these goals, researchers have employed various models, including DenseNet-121, EfficientNet-B2, CNN, Xception, and VGG-16. They utilized adaptive synthetic oversampling (ADASYN) to address dataset imbalances. The proposed model achieved an impressive accuracy of 97.35% in detecting disease cases. Notably, combining the DenseNet-121 and Xception models (DenseNet-121+Xception) resulted in better performance than did the individual models, with an 18% improvement in accuracy. Another ensemble model also showed a 1.46% improvement over the individual EfficientNet models.

A recent study by Fathi et al. [29] proposed an ensemble approach for stage classification of Alzheimer's disease. The researchers conducted a comparative analysis of various ensemble scenarios and selected six leading CNN-based classifiers for the ensemble model, namely, DenseNet201, Inception-ResNet V2, DenseNet121, ResNet50, DenseNet169, and VGG-19. The effectiveness of the ensemble model was assessed using accuracy, sensitivity, and specificity metrics. The results showed that the ensemble achieved an impressive accuracy of 93.88% for 4-way AD classification and 93.82% for 3-way AD classification.

Table 1 summarizes recent research on the classification of AD using DNN models. Current studies have focused mainly on classifying AD stages into three or four categories. However, many of these studies have

encountered challenges related to class imbalance in detecting AD. Unbalanced datasets often lead to overfitting, imprecise results, and lower accuracy in deep learning models. Additionally, the inadequate data availability for training deep learning models poses another obstacle. Our study addresses these challenges by focusing on the five-way classification of Alzheimer's disease using pretrained deep learning architectures and transfer learning techniques.

3. Proposed Methodology

We present a multiclass classification system for the early detection of Alzheimer's disease using deep neural networks and transfer learning. Figure 1 shows the workflow of the proposed model. The proposed method involves various steps, such as dataset acquisition, preprocessing, feature extraction, training, and classification. Our approach uses data augmentation and fine-tuned feature extraction to train three CNN architectures (VGG-19, Inception V3, and ResNet-50). We apply fine-tuned transfer learning by replacing the final layers of pretrained CNNs to fine-tune the models to target classes present within our ADNI dataset.

First, we preprocess and augment the data by converting DICOM and single-channel images to PNG and three channels while removing any extra surrounding area. Additionally, we apply various data augmentation techniques to improve the representation of the input space for the classifier. Next, we retrain the pretrained CNNs on the target dataset using their weights. Once the training process is complete, we evaluate the models' effectiveness using previously unseen data. The subsequent subsections provide a detailed discussion of our methodology.

3.1. Data Preparation

3.1.1. Dataset acquisition

We conducted our study using ADNI data, available at http://adni.loni.usc.edu/. The main aim of the ADNI is to develop more precise and sensitive methods for diagnosing AD in its initial phases. We obtained baseline T1-weighted structural MRI (sMRI) scans in the NIFTI (.nii) format for our experiments. The dataset comprises 608 subjects (292 males and 316 females) separated into five classes: AD, EMCI, MCI, LMCI, and CN. According to our data, multiple scans were taken at different times, with varying numbers of scans per subject—at least three and up to fifteen. The images were distributed among the classes as follows: AD (1648 images), EMCI (1184 images), MCI (1527 images), LMCI (1098 images), and NC (1648 images). More information on the subjects can be found in Table 2.

The dataset summary provides essential information about the data. The figure shows the number of individuals in each category, the breakdown of male and female individuals, and the average ages, along with their corresponding standard deviations (STDs). Moreover, as shown in Figure 2, sample MRI scans were obtained from all five classes.

Table 1. Summary of the literature related to AD classification using DNN models

Author(s)	Dataset & Image count	Modality	Classifier(s)	Classification	Accuracy
AbdulAzeem et al. [20]	ADNI (2,11,655 after DA)	MRI	CNN from scratch	2-way (AD/CN)	95.60%
Liu et al. [21]	ADNI (–)	MRI	CNN from scratch AlexNet GoogLeNet	3-way (CN/MCI/AD)	78.02% 91.40% 93.02%,
Al-Adhaileh [22]	ADNI (1279)	MRI	AlexNet ResNet50	3-way (CN/MCI/AD)	94.53% 58.07%
Helaly et al. [23]	ADNI (48000 after DA)	MRI	2D CNN 3D CNN Fine-tuned VGG-19	4-way (CN/EMCI/LMCI/AD)	93.61% 95.17% 97.0%
Savaş [24]	ADNI (2182)	MRI	EfficientNet-B0 EfficientNet-B2 EfficientNet-B3	3-way (CN/MCI/AD)	92.98% 94.42%, 97.28%
Antony et al.	ADNI (780)	MRI	Random forest	2-way (AD/CN)	68.0%

Author(s)	Dataset & Image count	Modality	Classifier(s)	Classification	Accuracy
[25]			VGG-16 VGG-19		81.0% 84.0%
Raza et al. [26]	ADNI (5016 after DA)	MRI	CNN from scratch	4-way (CN/MCI/LMCI/AD)	97.84%
Hazarika et al. [27]	ADNI (11000 after DA)	MRI	Hybrid model (LeNet + AlexNet)	3-way (CN/MCI/AD)	93.58%
Mujahid et al. [28]	ADNI (12,846 using ADASYN)	MRI	Ensemble model	4-way (CN/EMCI/MCI/AD)	97.35% (VGG-16 +EfficientNet-B2)
Fathi et al. [29]	ADNI (14,241)	MRI	Ensemble Model using WPBEM	3-way (NC/MCI/AD) 4-way (NC/EMCI/LMCI/AD)	93.92% 93.88%

DA- Data Augmentation, ADASYN- Adaptive synthetic technique, weighted probability-based ensemble method (WPBEM)

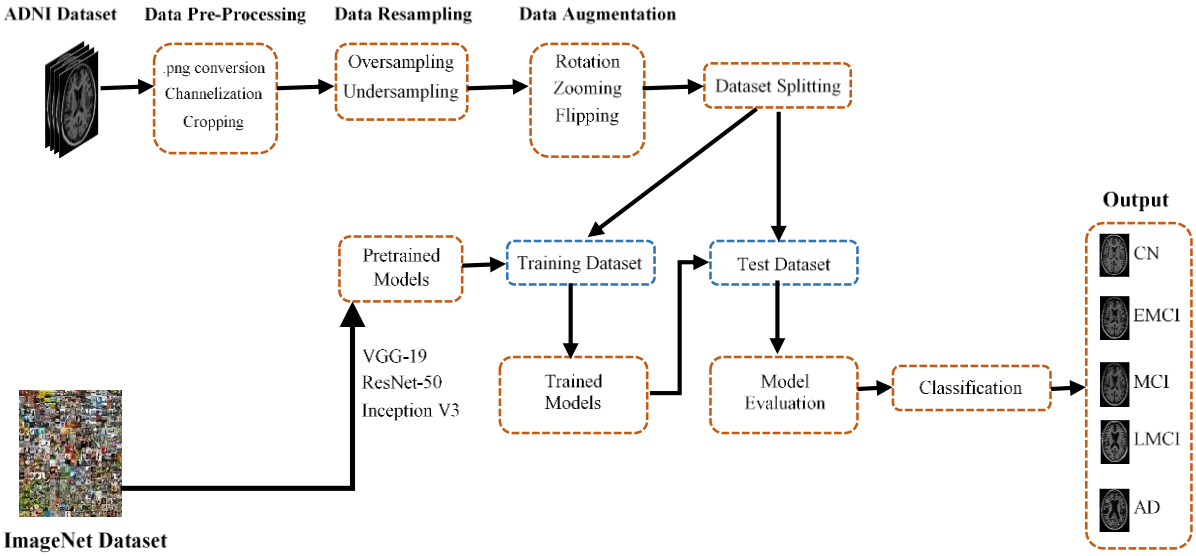


Fig. 1. Workflow of the proposed model

Table 2. Demographic Data

Alzheimer Stages	CN	EMCI	MCI	LMCI	AD
Count	128	124	131	119	106
Male/female	57/71	59/65	63/68	64/55	49/57
Age (mean ± STD)	76.78±0.48	75.84±0.86	75.81±0.79	76.89±1.21	76.21±0.90
Total Image Scans	1261	1184	1527	1098	1648

3.1.2. Data Preprocessing

Accurate data input is crucial for learning-based technologies to provide reliable predictions. Therefore, data preprocessing that enhances image contrast and pixel intensity is a critical stage in improving the effectiveness of these models. In the initial data preprocessing phase, we transformed the images from NIFTI to PNG format. However, the images obtained were in single-channel format and different in dimensions, which is unsuitable for deep learning models. To overcome this issue, we converted the images from single-channel to 3-channel format (RGB), which is necessary for all deep learning models.

We noticed that the dataset we collected had a significant imbalance in the number of instances across different classes. We implemented two resampling methods to address this issue: oversampling and undersampling. In the oversampling approach, we duplicated instances from underrepresented classes, namely, CN, EMCI, and LMCI. Conversely, in the undersampling approach, we removed instances from overrepresented classes, namely, AD and MCI. After applying these resampling methods, we ensured that all the AD classes contained 1500 MR images, which led to an expanded dataset of 7500 images.

The MRI scans were subjected to spatial normalization to the Montreal Neurological Institute (MNI) space. This was achieved using Statistical Parametric Mapping (SPM12) software (available at HTTP://WWW.FIL.ION.UCL.AC.UK/SPM/) and the Diffeomorphic Anatomical Registration Exponentiated Lie Algebra (DARTEL) registration process. The intensity values of the MRI scans were adjusted to fall within the range of [0, 1] through normalization. A denoising process was also performed on the images using a nonlocal means algorithm. This step introduces image blurring to lessen the impact of noise in the images.

3.1.3. Data Augmentation

Deep neural networks (DNNs) enhance model performance, especially when trained with larger datasets. Data augmentation is a technique that involves generating more images to expand the original dataset. Introducing modifications to the images boosts a model's ability to classify more accurately. As a result, the model can become more generalized and less prone to overfitting.

A large dataset is crucial for training deep neural networks effectively. However, when the dataset is small, such as in the case of MRI, data augmentation techniques can significantly diversify the data for neural network training. In our study, we compiled a dataset of images from 608 patients. Unfortunately, the dataset was insufficient for practical deep neural network training and optimal performance. To address this issue, we used image rotation (90°, 180°, and 270°), flipping/reflection (both horizontal and vertical flipping), and zooming in and out techniques to expand the dataset. As a result of implementing these techniques, we enlarged the dataset from 7500 to 39,980 images. Figure 3 displays the outcomes of the data augmentation procedure. The dataset was then split into three sets for training (80%), validation (10%), and testing (10%). Table 3 provides a summary of the dataset used in our study.

3.1.4. Cross-Validation

Cross-validation is a popular method for evaluating the effectiveness of deep learning models. It involves partitioning the available data into several subsets, typically two or more, with one subset chosen for model training and the other for performance assessment. K-fold cross-validation is one of the most commonly used

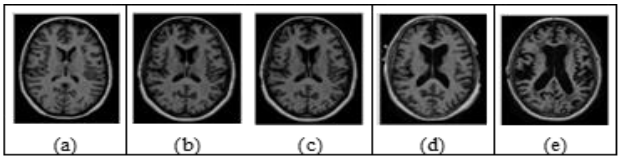


Fig. 2. Stages of AD: (a) CN, (b) EMCI, (c) MCI, (d) LMCI, and (d) AD

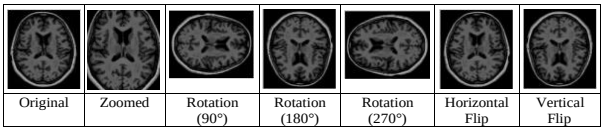


Fig. 3. Results of data augmentation

Table 3. Description of the training, validation, and test datasets

Class Label	CN	EMCI	MCI	LMCI	AD	Total
Image count	7886	8058	8031	7894	8111	39980
Train set	6308	6446	6425	6316	6489	31984
Validation set	789	806	803	789	811	3998
Test set	789	806	803	789	811	3998

methods in deep learning [30]. It partitions the dataset into k subsets, or folds, of approximately the same size. The model undergoes training k times, with each fold acting as the validation set and the remaining folds serving as the training set. Finally, the results from these k-fold models are averaged to derive a final estimate of the model's performance.

3.2. Network Architecture

3.2.1. Convolution Neural Networks (CNNs)

Our work uses deep neural network models constructed using the CNN algorithm to examine the data effectively. Unlike traditional machine learning techniques, which involve three fundamental steps, CNNs combine all these stages, and the feature extraction process can be automated. However, in real-world scenarios, gathering sufficient data for training CNN models can be challenging and can lead to difficulties such as overfitting and convergence during deep CNN training. In such cases, researchers often opt for the transfer learning method, which utilizes pretrained models and their weights to mitigate data scarcity and address potential training challenges effectively [31]. An example of a CNN structure is shown in Figure 4. Three layers are used: the convolution layer performs feature extraction, the pooling layer reduces dimensionality, and the fully connected layer performs the classification task with two-dimensional matrices from the previous layers.

In a convolutional neural network, the convolutional layer

applies a learnable filter (kernel) to an input image, which allows the network to identify and extract pertinent features from the provided data. The following is a breakdown of the dimensions involved:

Input image: H × W × C, where H, W, and C denote the height, width, and number of channels, respectively.

Filter (Kernel): FH × FW × FC, where FH, FW, and FC denote the height, width, and number of channels of the filter, respectively.

Output activation map: AH × AW, where AH is the activation height and AW is the activation width. The dimensions of the output activation map are dictated by the convolution operation and chosen hyperparameters, including the stride and padding.

We slide a filter across the input image to obtain the output activation map. At each position between the filter and the corresponding input image patch, we calculate the elementwise dot product. The resulting product is then passed through an activation function. This process occurs across the entire input image, resulting in the output activation map. The stride and padding parameters affect the dimensions of the output activation map.

The stride parameter dictates the pixel displacement that occurs as the filter moves. Padding involves adding extra pixels around the input image, maintaining the spatial properties of the resulting feature maps. By adjusting these parameters, we can control the spatial size of the output activation map.

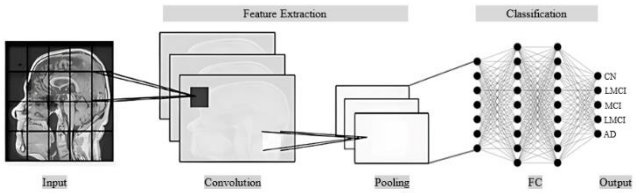


Fig. 4. Structure of a CNN

The following formulas can be used to determine the spatial size of the output activation map:

$$AH = (H - FH + 2 * padding) / stride + 1 \quad (1)$$

$$AW = (W - FW + 2 * padding) / stride + 1 \quad (2)$$

Activation functions are an essential component of neural networks. They introduce nonlinearity to the model by nonlinearly modifying the output of neurons. The nonlinearity added by activation functions is crucial because it empowers the network to learn complex relationships between inputs and outputs. The network would remain limited to linear models without these functions, even with multiple layers. Therefore, activation functions play a crucial role in the success of neural networks, enabling them to model complex relationships and achieve high accuracy.

An activation function may include one or more of the following:

ReLU (rectified linear unit): $f(x) = max(0, x)$ (3)

Sigmoid: $f(x) = 1 / (1 + exp(-x))$ (4)

Tanh (Hyperbolic Tangent):

$$f(x) = (exp(x) - exp(-x)) / (exp(x) + exp(-x)) \quad (5)$$

Leaky ReLU: $f(x) = max(ax, x)$, where 'a' is a slight positive slope for negative inputs. (6)

The proposed multiclassifier uses the SoftMax function. Based on the likelihood specified by Eq. (7), the data points are assigned to various classes.

$$f(v_i) = \frac{e^{v_i}}{\sum_{j=1}^{K} e^{v_j}} \text{ For i=1, 2……, K and v= } [v_1, v_2 \ldots v_k] \quad (7)$$

where v denotes the input vector, e^{v}_{i} is the standard exponential function for the input vector, K denotes the number of classes in the multiclass classifier, and e^{v}_{j} is the standard exponential function for the output vector.

Overfitting occurs when a model learns to perform exceptionally well on the training data but fails to generalize to new, unseen data. To address this issue, a regularization technique called dropout is used. Dropout randomly deactivates or drops out some neurons during each training cycle, ensuring that the network does not rely heavily on specific neurons. This forces the network to learn more diverse and robust features that are not tied to the presence of particular neurons. By doing so, the network becomes more capable of generalizing to new inputs and improving its overall performance.

Deep learning models are often designed for specific tasks, but their true potential lies in their ability to handle various problems. Over time, there has been a shift toward using pretrained models as a starting point for new investigations. Recently, the use of these established models has increased. These approaches involve using the learned weights from pretrained models or adding additional layers to these models via transfer learning methods. Our research uses pretrained CNN models, namely, Inception V3, ResNet-50, and VGG-19.

The process of hyperparameter tuning in deep learning often involves a combination of trial and error, domain knowledge, and insights gained from previous studies. No single approach fits all the scenarios, and finding the correct hyperparameter values usually requires experience and experimentation. In this study, hyperparameters were established through an iterative process involving multiple alternatives to identify optimal choices that enhance

performance without altering the core structure of the pretrained models.

To adapt the final layers of the pretrained models for the classification of images as AD, CN, EMCI, MCI, or LMCI, the output data were gradually reduced, and additional dense layers were added to the network structure, with sizes of 512 and 3. To prevent overfitting, we incorporated dropout layers between these layers, with a dropout rate of 0.5 chosen to ensure stable outcomes. We selected the stochastic gradient descent method (SGDM) as an optimizer, with the default learning rate set at 1e-4. For the loss parameter, categorical cross-entropy was adopted, and accuracy was designated as the chosen metric.

The models underwent training processes lasting for a standard 100 epochs. This specific number of epochs was established through repeated experimentation to optimize the hyperparameters before encountering a situation of overfitting. The framework for model training and testing is outlined in Figure 5, and the training choices and hyperparameters utilized for training the various network architectures are summarized in Table 4.

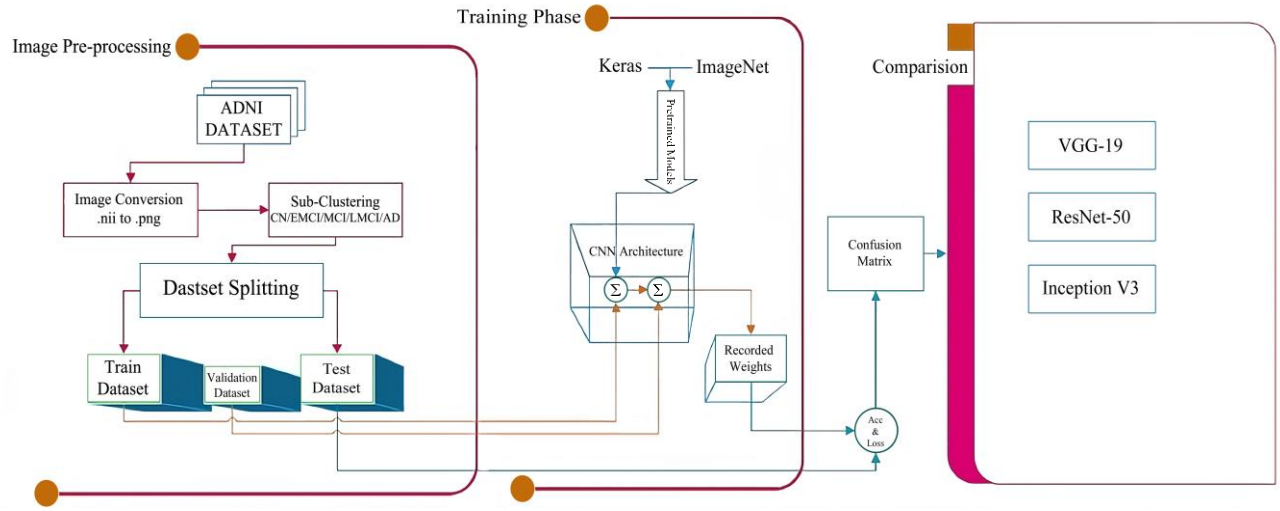


Fig. 5. Framework for model training, testing, and comparison

Table 4. Hyperparameters adopted

Parameter	Value
Optimization Algorithm	SGDM
Momentum	0.9
Loss Function	Categorical Cross Entropy
Initial Learning Rate	1e-4
Maximum Number of Epochs	100
Dropout Rate	0.5

Two platforms were used to utilize deep learning models for analyzing MRI data: Kaggle and Google Colaboratory (Colab). Kaggle is a collaborative platform that supports data science problem solving, while Colab is a platform that facilitates writing and executing ML and data analytics tasks using resources such as GPUs. The Keras deep learning library module, which was implemented with TensorFlow, was used to train the DL models. After the model operations, graphics and weights were systematically documented, and the models' accuracy and loss parameters were visually depicted. Confusion matrices were generated to assess model performance.

3.2.2. VGG-19

The VGG-19 model is a framework that uses a deep CNN architecture. It was first introduced in 2014 in "Very Deep Convolutional Networks for Large-Scale Image Recognition" by Karen Simonyan and Andrew Zisserman [32]. Figure 6 shows the architecture of VGG-19.

Key characteristics of the VGG-19 architecture

Depth: VGG-19 is characterized by its depth. It encompasses nineteen layers, sixteen of which are convolutional and three of which are fully connected. This approach provides the capacity to comprehend intricate patterns and features present within images.

Convolutional Layers: VGG-19 uses a downsampling scheme that consists of a max pooling layer that comes after each convolutional layer. Using small 3 × 3 filters in the convolutional layers combined with max pooling leads to a consistent receptive field size and a more accessible architecture.

Filter size: All the convolutional layers in VGG-19 use 3 × 3 filters, which helps to maintain a compact architecture. This results in fewer parameters than larger filters, making the model computationally efficient.

Fully Connected Layers: VGG-19 comprises three fully connected layers after the convolutional and pooling layers,

culminating in a softmax layer for classification.

VGG-19 is frequently employed as a pretrained model for TL approaches due to its performance in image recognition tasks. It can be fine-tuned on specific datasets or used to extract distinctive features by eliminating terminal layers and adding new features tailored to a particular task.

3.2.3. ResNet-50

ResNet-50 belongs to the ResNet family of CNNs. It was developed by Kaiming He et al. [33] in 2016 and emerged as the winner of the ILSVRC 2015, with an error rate of 3.6%. ResNet-50 was created by Microsoft Research 2015 to address the challenge of vanishing gradients in neural networks with a significant depth. The key features of ResNet-50 are its depth and the use of residual blocks, which enable the training of networks with greater depths without suffering from degradation. Figure 7 depicts the structural design of ResNet-50.

Key features of ResNet-50:

Identity Shortcut Connections (Skip Connections): ResNet-50 represents the novel notion of incorporating identity shortcut connections or skip connections. Instead of the conventional approach where layers stack sequentially, the input to a layer is fused with the output of a later layer in a residual block. Identity mapping alleviates the challenge of the vanishing gradient problem, consequently facilitating the training of networks with significantly deep architectures.

Bottleneck Architecture: Each residual block in ResNet-50 employs a bottleneck architecture. Initially, the network consists of a 1 × 1 convolutional layer, followed by a 3 × 3 convolutional layer and a 1 × 1 convolutional layer. The computational complexity is minimized by utilizing 1 × 1 convolutions to reduce the number of channels, thus improving the efficiency.

Skip Connections: Skip connections provide a path for the gradient to flow more directly during backpropagation. If a residual block learns an identity function, the gradients can flow unimpeded through the shortcut connections, making training more accessible.

Pooling Layers: Average pooling in ResNet-50 before the final fully connected layer diminishes the spatial dimensions and quantity of parameters in the model, thereby augmenting the model's ability to generalize.

Global average pooling: ResNet-50 uses global average pooling as a replacement for the fully connected layers for the ultimate classification process, thereby reducing overfitting and parameter count in the model.

Pretrained models, such as ResNet-50, are widely used for transfer learning. These models are trained on extensive image datasets such as ImageNet. During transfer learning, the lower layers of the pretrained model are utilized to extract features, while the upper layers are modified according to specific tasks. ResNet-50 was a significant breakthrough in deep learning. Its depth and residual blocks enabled the creation of even deeper architectures. The ResNet model proved that expanding the neural network depth can lead to poorer performance. Instead, this approach can enhance results when appropriately designed.

3.2.4. Inception v3

Inception v3 [34] is a cutting-edge CNN framework created by Google researchers that significantly improves the performance and effectiveness of DL models for image recognition tasks. Building upon the existing Inception models, Inception v3 enhances efficiency and delivers superior results. The architecture of Inception v3 is presented in Figure 8.

Key features of Inception-v3:

Multiple Pathways (Parallel Convolutions): Inception V3 employs a unique structure called "Inception modules." These modules consist of multiple convolutional layers with diverse filter sizes (1 × 1, 3 × 3, and 5 × 5) and max pooling layers. These pathways run in parallel, capturing features at multiple spatial scales and abstraction levels. The idea is to allow the network to acquire fine-grained and high-level features simultaneously.

1×1 Convolutions: Inception V3 heavily uses 1×1 convolutions to diminish the dimensionality of the source feature maps. This approach mitigates the computational complexity and allows for efficient parallel processing.

Bottleneck Layers: To further improve efficiency, Inception V3 often incorporates bottleneck layers, which use 1×1 convolutions to diminish the number of input channels before larger filters are applied. The computational load is reduced without sacrificing the model's capacity to identify intricate features.

Auxiliary Classifiers: Training is carried out with auxiliary classifiers at intermediate layers in Inception V3, which serve as extra branches to the main classification layer. They help to combat the vanishing gradient problem during training, ultimately improving convergence and gradient flow.

Global average pooling: Inception V3 uses global average pooling in place of completely connected layers at the end. This technique replaces the traditional flattening and fully connected layers with a single average pooling layer that generates predictions directly from spatial feature maps.

Regularization Techniques: Inception V3 incorporates batch normalization and dropout techniques to enhance generalization and reduce overfitting.

Inception V3 is a popular pretrained model for transfer

learning. It captures generic image features in its lower layers, which remain unchanged, while the top layers are customized for specific tasks. Its outstanding performance and efficient use of resources have made it a standard benchmark in image recognition competitions.

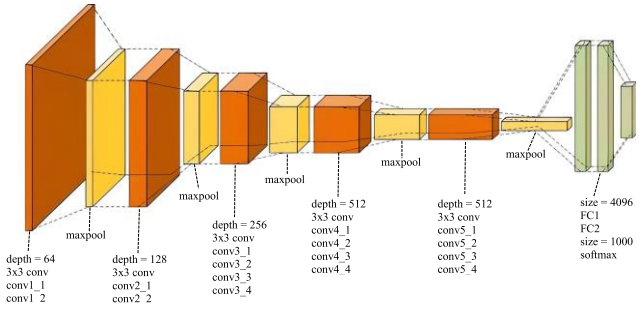


Fig. 6. VGG-19 architecture

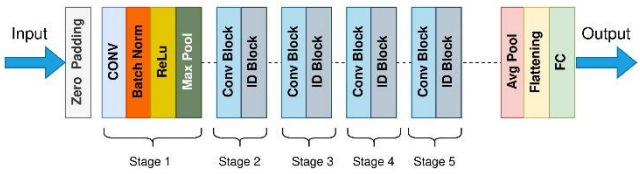


Fig. 7. ResNet-50 architecture

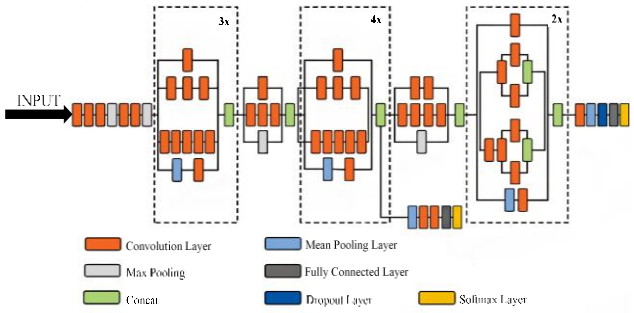


Fig. 8. Inception V3 architecture

3.3. Transfer Learning Framework

Transfer learning is a machine learning technique that leverages knowledge acquired from one task to enhance the performance of another related but nonidentical task. In mathematical terms, transfer learning can be defined as follows.

Domain: A domain

$$\mathfrak{D} = \{X, (X)\} \quad (8)$$

is delineated by two constituents:

- A feature space X
- A marginal probability distribution $P(X)$ where $X = \{x_1, x_2, x_3, \ldots, x_n\} \in X$

If two domains exhibit dissimilarity, it follows that they possess disparate feature spaces ($Xt \neq Xs$) or separate marginal distributions ((Xt) $\neq$ (Xs)).

Task: Two components make up a task $\mathcal{T} = \{Y, (.)\}$, given a specific domain D:

- Y denotes the label space
- A predictive function, $f(.)$ can be learned from training data, even though it is not observed

$\{(x_i, y_i) \mid i \in \{1, 2, 3\ldots N\}$, where $x_i \in X$ and $y_i \in Y\}$.

Given that (x_i) can alternatively be expressed as $p(y_i|x_i)$ from a probabilistic perspective, we can reformulate task $\mathcal{T}$ as

$$\mathcal{T} = \{Y, (Y|X)\} \quad (9)$$

If two tasks exhibit dissimilarity, then possess disparate label spaces ($Yt \neq Ys$) or dissimilar conditional probability distributions ($P(Yt|Xt) \neq P(Ys|Xs)$).

The objective of transfer learning is to enhance the learning of a conditional probability distribution denoted by ($Yt|Xt$) by leveraging the knowledge obtained from a source domain $\mathfrak{D}s$ and a matching learning task $\mathcal{T}s$. This knowledge is then applied to a target domain $\mathfrak{D}t$ and learning task $\mathcal{T}t$ such that

$$\mathfrak{D}t \neq \mathfrak{D}s \text{ or } \mathcal{T}t \neq \mathcal{T}s \quad (10)$$

3.4. Performance analysis

Evaluation metrics refer to the numerical measures utilized to assess the effectiveness of a model in addressing a specific task. Our study used various evaluation metrics: accuracy, precision, error, false positivity rate, kappa, sensitivity/recall, specificity, and F-measure. The confusion matrix complements these metrics, as it visually represents the classification model's performance across various classes. The model is organized with actual class labels forming the rows and predicted class labels forming the columns. Table 5 shows this information.

Table 5. Confusion matrix

	Predicted Positive (P)	Predicted Negative (N)
Actual Positive (P)	True Positive (TP)	False Negative (FN)
Actual Negative (N)	False Positive (FP)	True Negative (TN)

Table Description

Positive (P): indicates that the patient suffers from dementia

Negative (N): indicates that the patient has a normal cognitive status

True Positive (TP): Correctly predicted positive instances.

True Negative (TN): Correctly predicted negative instances.

False Positive (FP): Incorrectly predicted positive instances (Type I error).

False Negative (FN): Incorrectly predicted negative instances (Type II error).

Accuracy: Accuracy measures the proportion of correctly predicted instances among all instances. It is calculated as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (11)$$

Accuracy Error: The accuracy error, often called the misclassification error, calculates the proportion of instances a machine learning model has about the number of cases incorrectly classified, calculated as

$$Accuracy\ Error = 1 - Accuracy \quad (12)$$

Precision: The precision of a prediction is the ratio of true positives to all positive predictions. It is calculated as follows:

$$Precision = TP/(TP + FP) \quad (13)$$

Sensitivity (recall or true positive rate): Sensitivity measures how well positive instances are predicted from actual instances. It is calculated as follows:

$$Sensitivity\ (Recall) = True\ Positive\ Rate = Recall = TP/(TP + FN) \quad (14)$$

Specificity: Specificity, also referred to as the selectivity of a network, is a measure that indicates the ratio of accurately detected negative instances to those that are negative. The true negative rate represents this parameter and can be calculated using the following formula:

$$Specificity = True\ Negative\ Rate = TN/(TN + FP) \quad (15)$$

False positive rate: The false positive rate, or fall-out rate, pertains to the frequency of positive predictions generated by a network that is not correct. This is calculated as the ratio of incorrectly predicted positives to negatives observed. This metric helps assess the model's tendency to incorrectly classify negative instances as positive.

$$FPR\ (1 - Specificity) = FP/(TN + FP) \quad (16)$$

F-Measure (F1-Score): The F1-Measure is the mean of precision and recall, weighted harmonically. It is calculated as

$$F1 - score = (2 \times Precision \times Recall)/(Precission + Recall) \quad (17)$$

The F1 score is a valuable metric for evaluating the accuracy of a classifier's predictions in machine learning, especially when there is an imbalance in the class distribution. It considers both false positives and false negatives equally.

Kappa: The kappa coefficient, or Cohen's kappa score, is a valuable metric in machine learning evaluation considering the possibility of random agreement. Moreover, this approach is beneficial in scenarios with imbalanced datasets or disparate class probabilities, providing insights into the performance of classification models beyond simple accuracy. The higher the kappa coefficient is, the more consistent the model's predictions are with the results.

$$Kappa = (2 * (TP * TN - FN * FP))/((TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)) \quad (18)$$

4. Experimental Setup and Results

Training a CNN from scratch can be challenging for medical image analysis tasks due to the unavailability of large-scale datasets. To overcome this, a common approach is to use CNNs that have been pretrained on a vast dataset, such as ImageNet. Transfer learning adapts this pretrained knowledge to a new task, significantly accelerating the learning process. In our work, we improved upon our model's ability to learn using a fine-tuning method. We first trained a deep neural network (CNN) on a large dataset of images (ImageNet) and then fine-tuned this pretrained CNN on a different dataset (ADNI) for a specific task. This approach involves two datasets: the "source" dataset (ImageNet) and the "target" dataset (ADNI). We utilized pretrained deep CNN models on the given dataset. The input images were processed through multiple convolutional layers acting as feature extractors. The convolutional layers apply kernels or convolutional matrices to the input, extracting relevant features. This process updates the layer weights and stores them in a vector. Subsequently, the weights were utilized by fully connected and soft-max layers to classify the images into distinct AD classes (CN, EMCI, MCI, LMCI, and AD).

For our study, pretrained CNNs, such as VGG-19, ResNet-50, and Inception V3, were employed to classify Alzheimer's disease into five categories. We used the ImageNet dataset as the pretraining dataset for all these models. We employed transfer learning techniques to distinguish between the AD classes and fine-tuned these pretrained networks on the target dataset (ADNI). Table 6 details the pretrained CNN models we used in our study.

We adjusted the size of the input images to use different image classification models. Specifically, we used 299 × 299 pixels for the inception V3 model and 224 × 224 pixels for the VGG-19 and ResNet-50 models. We kept the parameters identical during the training process for all the experiments. We split the dataset into three parts: training (80%), validation (10%), and testing (10%). We ran five experiments for each model, both on the original and augmented datasets. We used various data augmentation methods to increase the dataset size, which resulted in different sample representations.

Since training CNNs from scratch is challenging and requires large-scale datasets for optimal prediction via transfer learning fine-tuning, we augmented the dataset to

39,980 images. After augmentation, we trained deep convolutional neural networks using the same data split and other parameters. We focused on 608 participants from the ADNI dataset. Patients were categorized into five classes: AD, EMCI, MCI, LMCI, and CN. Our primary goal was to make predictions and establish the effectiveness of the fine-tuned networks as a benchmark. We tested the fine-tuned networks on the ADNI dataset to evaluate their performance.

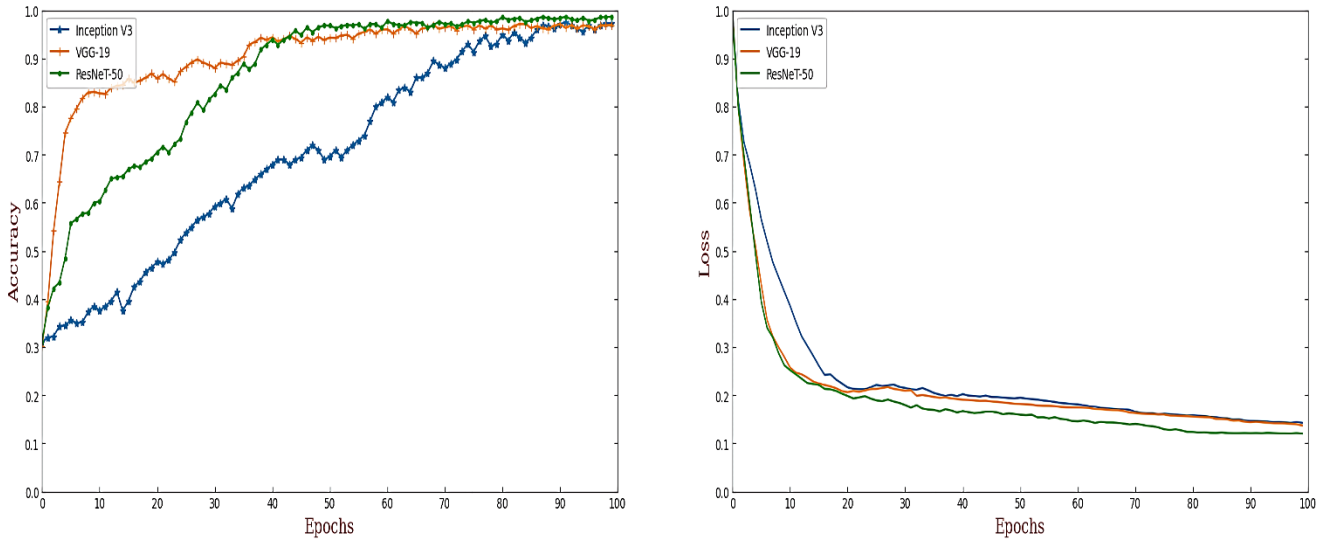
4.1. Training Progress

We used a specific training dataset, and predetermined settings were used to train three network architectures: VGG-19, ResNet-50, and Inception V3. The objective was to compare the performance of these methods in terms of accuracy and other relevant metrics. The training process involved fine-tuning the pretrained networks with the same training parameters for a fixed number of epochs. Figure 9 visually displays the improvement in the performance of the network architectures during the training process. This helps us understand how well each architecture optimized and enhanced its performance with increasing training epochs.

According to the plot, VGG-19 and ResNet-50 performed better in terms of optimization than Inception-V3. Additionally, ResNet-50 exhibited better optimization performance than VGG-19. This is indicated by the accuracy or performance metrics on the vertical axis increasing more steeply for VGG-19 and ResNet-50 than for Inception V3, suggesting faster and more efficient learning. Moreover, the plot shows that Inception V3 required more than 50 epochs to achieve optimal performance. The performance of the Inception-v3 model continued to improve even after 50 training epochs, indicating that additional epochs may be needed to achieve its optimal accuracy. More complex AI models, such as Inception V3, typically require more training time to reach their full potential. Thus, it is crucial to adjust the training duration based on the model's complexity and performance during the training process.

Table 6. Details of the pretrained CNN architectures

Model	Depth (in layers)	Input dimensions	Parameter Count (in millions)	Size (in MB)
VGG-19	19	224 x 224	144	549 MB
ResNet-50	50	224 x 224	25.6	98 MB
Inception V3	42	299 x 299	27	92 MB



(a) Accuracy (b) Loss

Fig. 9. Accuracy and loss across different epochs

4.2. Evaluation of the Proposed Model

4.2.1. Performance analysis of the pretrained networks for AD classification

4.2.1.1. Performance at baseline without data augmentation

We assessed the performance of each model based on the original dataset without any data augmentation techniques. We conducted various tests to measure the accuracy, precision, recall, specificity, error rate, false positive rate, F1 score, and kappa coefficient of several pretrained networks on the original dataset. The results of our analysis are presented in Table 7, where ResNet-50 achieved the highest accuracy (93.39%), followed by Inception-V3

(93.42%) and VGG-19 (92.19%). It is worth noting that ResNet-50 uses fewer parameters than the VGG-19 and Inception V3 models.

4.2.1.2. Augmented data for enhanced performance

In our research, we faced the challenge of limited data, which could have negatively impacted our models' performance and generalization accuracy. However, we implemented data augmentation methods to overcome this limitation. Before training, we applied rotation, zooming, and horizontal and vertical flipping to the original images to expand our dataset. The performance of our models was then tested using the augmented data. The results, which are presented in Table 8, are impressive. Among the pretrained networks, ResNet-50 was the most effective, achieving an average accuracy of 98.7%, outperforming VGG-19 (97.16%) and Inception V3 (97.54%). Additionally, ResNet-50 demonstrated superior performance in several other key metrics, including precision, recall, specificity, error rate, false positive rate, F1 score, and kappa. Overall, our study highlights the effectiveness of data augmentation methods and the superiority of ResNet-50 in producing consistently better results than the other two pretrained networks.

4.2.2. Confusion Matrices

The confusion matrices for three network models, VGG-19, ResNet-50, and Inception V3, are shown in Figures 10-12, with and without data augmentation.

The VGG-19 model had 81% accuracy in predicting cases before implementing data augmentation. The accuracy rate was similar for the CN, EMCI, and AD classes. However, the accuracy rates were slightly lower at 80% and 79% for the MCI and LMCI classes, respectively. Following data augmentation, the accuracy rates for the CN class, EMCI and MCI classes, and LMCI and AD classes improved to 97%, 93%, 92%, and 90%, respectively. These results indicate a significant improvement in the model's prediction capabilities.

The ResNet-50 model was able to accurately predict 84% of the CN and EMCI patients, 83% of the MCI and AD patients, and 82% of the LMCI patients using the original dataset. However, upon using the augmented dataset, the model's performance improved significantly, achieving accuracy rates of 99% for the CN and EMCI classes, 98% for the MCI and LMCI classes, and 91% for the AD class.

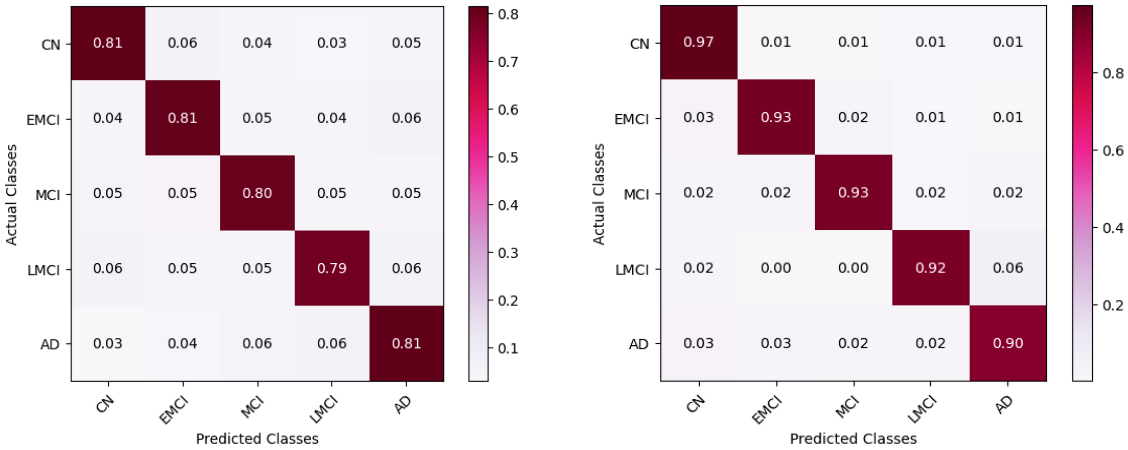
Moreover, the Inception-v3 model also delivered remarkable results. Initially, it accurately predicted 82% of the CN, EMCI, and AD subjects; 80% of the MCI subjects; and 79% of the LMCI subjects in the original dataset. However, after data augmentation, the model's performance improved even further, with accuracy rates of 96% for the CN, EMCI, and MCI classes; 92% for the LMCI class; and 90% for the AD class.

Table 7. Overall performance of the pretrained networks on the original dataset

Metric	VGG-19	ResNet-50	Inception V3
Accuracy	92.19	93.39	92.42
Accuracy Error	7.81	6.61	7.58
Sensitivity/Recall	80.47	83.47	81.05
Specificity	95.12	95.87	95.26
Precision	80.48	83.47	81.06
False Positive Rate	4.88	4.13	4.74
F1-Score	80.47	83.47	81.04
Kappa	75.58	79.33	76.30

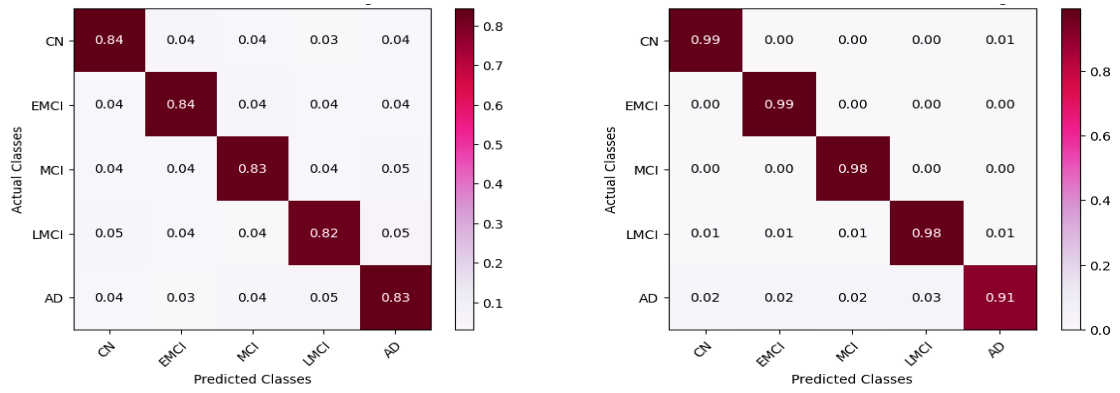
Table 8. Overall performance of the pretrained networks on the augmented dataset

Metric	VGG-19	ResNet-50	Inception V3
Accuracy	97.16	98.70	97.54
Accuracy Error	2.84	1.30	2.46
Sensitivity/Recall	92.90	96.90	93.80
Specificity	98.23	99.15	98.48
Precision	92.96	96.65	93.93
False Positive Rate	1.78	0.85	1.53
F1-Score	92.89	96.73	93.84
Kappa	91.12	95.92	92.31



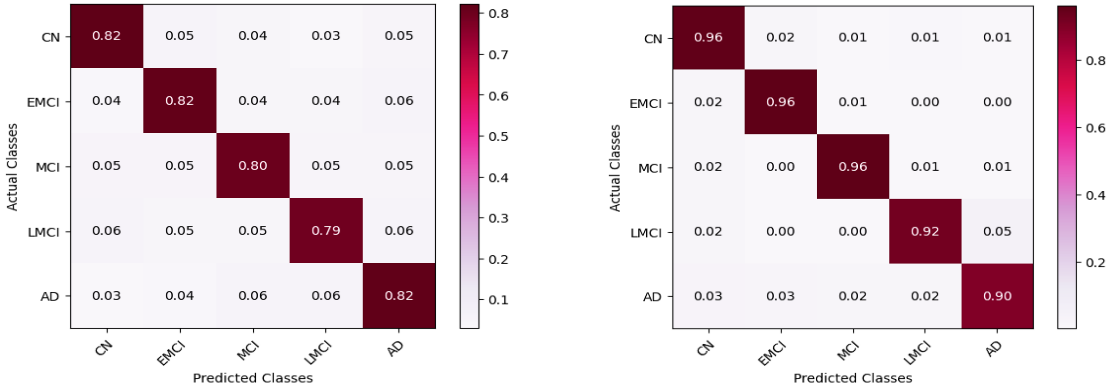
(a) (b)

Fig. 10. Confusion matrix results for VGG-19: (a) Original dataset and (b) Augmented dataset



(a) (b)

Fig. 11. Results of the ResNet-50 confusion matrix for the (a) original dataset and (b) augmented dataset



(a) (b)

Fig. 12. Confusion matrix results for Inception-v3: (a) original dataset and (b) augmented dataset

Upon analyzing the confusion matrices, it is evident that data augmentation generally improves the accuracy of all three models (VGG-19, ResNet-50, and Inception V3) during the transfer learning process.

4.2.3. Classwise performance of individual fine-tuned deep learning models

Confusion matrices were generated for all the pretrained models during the final testing phase. This helped us analyze each model's performance using images from each class. We computed the accuracy, error rate, recall, specificity, precision, false positive rate, and F1-score for each model in each class. The results of these parameters are presented in Table 9 for the original dataset and Table 10 for

the augmented dataset.

Table 9. Class-based performance metrics for the original dataset

Model	Predicted Class	Accuracy	Precision	Recall	Specificity	F1-score	Error Rate	False Positive Rate
VGG-19	CN	92.67	81.47	81.37	95.45	81.42	7.33	4.55
	EMCI	92.27	80.87	80.77	95.18	80.82	7.73	4.82
	MCI	92.05	79.98	80.57	94.93	80.27	7.95	5.07
	LMCI	91.90	78.60	80.99	94.58	79.78	8.10	5.42
	AD	92.05	81.48	78.67	95.45	80.05	7.95	4.55
ResNet-50	CN	**93.65**	**84.43**	**83.14**	**96.23**	**83.78**	**6.35**	**3.77**
	EMCI	**93.65**	**83.91**	**84.74**	**95.90**	**84.32**	**6.35**	**4.10**
	MCI	**93.32**	**83.17**	**83.69**	**95.74**	**83.43**	**6.68**	**4.26**
	LMCI	**93.20**	**82.35**	**83.40**	**95.61**	**82.87**	**6.80**	**4.39**
	AD	**93.12**	**83.50**	**82.37**	**95.86**	**82.93**	**6.88**	**4.14**
Inception V3	CN	92.95	82.29	81.88	95.67	82.08	7.05	4.33
	EMCI	92.57	81.61	81.51	95.36	81.56	7.43	4.64
	MCI	92.20	80.12	81.32	94.93	80.72	7.80	5.07
	LMCI	92.15	79.28	81.50	94.76	80.38	7.85	5.24
	AD	92.22	81.97	79.04	95.58	80.48	7.78	4.42

Note: Bold-marked entries represent the highest values

Table 10. Class-based performance metrics for the augmented dataset

Model	Predicted Class	Accuracy	Precision	Recall	Specificity	F1-score	Error Rate	False Positive Rate
VGG-19	CN	97.35	90.26	97.25	97.38	93.62	2.65	2.63
	EMCI	97.55	94.43	93.25	98.63	93.84	2.45	1.38
	MCI	97.45	94.63	92.5	98.69	93.55	2.55	1.31
	LMCI	97.25	94.34	91.75	98.63	93.03	2.75	1.38
	AD	96.2	91.12	89.75	97.81	90.43	3.8	2.19
ResNet-50	CN	**98.65**	**94.94**	**98.5**	**98.69**	**96.69**	**1.35**	**1.31**
	EMCI	**99.15**	**96.59**	**99.25**	**99.13**	**97.9**	**0.85**	**0.88**
	MCI	**98.95**	**96.56**	**98.25**	**99.13**	**97.4**	**1.05**	**0.88**
	LMCI	**98.9**	**96.78**	**97.75**	**99.19**	**97.26**	**1.1**	**0.81**
	AD	**97.85**	**98.37**	**90.75**	**99.63**	**94.41**	**2.15**	**0.38**
Inception V3	CN	97.15	90.74	95.5	97.56	96.69	2.85	2.44
	EMCI	98.25	95.29	96	98.81	97.9	1.75	1.19
	MCI	98.4	96.23	95.75	99.06	97.4	1.6	0.94
	LMCI	97.55	95.35	92.25	98.88	97.26	2.45	1.13
	AD	96.35	92.03	89.5	98.06	94.41	3.65	1.94

Note: Bold text represents the highest values

The top-performing models for all the classes were identified based on the confusion matrix values. Considering the accuracy parameter, the ResNet-50 model produced the highest results for all the classes, with scores of98.65% for CN, 99.15% for EMCI, 98.95% for MCI, 98.9% for LMCI, and 97.85% for AD. Inception V3 and VGG-19 followed in second and third place, respectively.

The ResNet-50 model was applied in terms of precision, achieving first place for all classes, with scores of 94.94% for the CN, 96.59% for the EMCI, 96.56% for the MCI, 96.78% for the LMCI, and 98.37% for the AD. Inception V3 and VGG-19 came in second and third place, respectively. Inception V3 ranked second, with scores of 90.74% for the CN, 95.29% for the EMCI, 96.23% for the MCI, 95.35% for the LMCI, and 92.03% for the AD. VGG-19 ranked third, with scores of 90.26% for the CN, 94.43% for the EMCI, 94.63% for the MCI, 94.34% for the LMCI, and 91.12% for the AD.

Finally, in terms of sensitivity/recall, the ResNet-50 model was the top-performing model for all the classes, with scores of 98.5% for CN, 99.25% for EMCI, 98.25% for MCI, 97.75% for LMCI, and 90.75% for AD. Inception V3 and VGG-19 ranked second and third, respectively. Inception V3 ranked second, with 95.5% for the CN, 96% for the EMCI, 95.75% for the MCI, 92.25% for the LMCI, and 89.5% for the AD. VGG-19 ranked third, with 97.25% for the CN, 93.25% for the EMCI, 92.5% for the MCI, 91.75% for the LMCI, and 89.75% for the AD.

The models that exhibited the highest results for all the classes were identified based on the values in the confusion matrix. Considering the accuracy parameter, the ResNet-50 model produced the highest results for all the classes (CN–98.65%, EMCI–99.15%, MCI–98.95%, LMCI–98.9%, AD–97.85%), Inception V3 (CN–97.15%, EMCI–98.25%, MCI–98.4%, LMCI–97.55%, AD–96.35%), and VGG–19 (CN–97.35%, EMCI–97.55%, MCI–97.45%, LMCI–97.25%, AD–96.2%).

Considering the precision parameter, the ResNet-50 model secured first place for all the classes (CN–94.94%, EMCI–96.59%, MCI–96.56%, LMCI–96.78%, AD–98.37%). Inception V3 was the second most common (CN–90.74%, EMCI–95.29%, MCI–96.23%, LMCI–95.35%, AD–92.03%), followed by VGG-19 (CN–90.26%, EMCI–94.43%, MCI–94.63%, LMCI–94.34%, AD–91.12%).

Finally, in terms of sensitivity/recall, ResNet-50 was the top-performing model for all the classes (CN–98.5%, EMCI–99.25%, MCI–98.25%, LMCI–97.75%, AD–90.75%). Again, Inception V3 was the second most common form (CN–95.5%, EMCI–96%, MCI–95.75%, LMCI– 92.25%, AD–89.5%), followed by VGG-19 (CN–97.25%, EMCI–93.25%, MCI–92.5%, LMCI–91.75%, AD– 89.75%).

Based on our findings, the ResNet-50 model outperformed other pretrained models (VGG-19 and Inception V3) in analyzing brain MR images. Although there were slight variations in accuracy across the different categories, the overall difference between them was trivial. The ResNet-50 model achieved the highest accuracy (99.25%) in classifying the "EMCI" category, indicating its exceptional ability to identify early signs of memory impairment. Additionally, the Inception-v3 model also produced promising results for specific categories. Therefore, using the ResNet-50 model directly or through transfer learning in future clinical studies is highly recommended and holds enormous potential for improved outcomes.

5. Discussion

Alzheimer's disease is a prevalent form of irreversible dementia that has a high mortality rate and ultimately leads to death. Detecting AD in its early stages can significantly improve patient survival and increase the effectiveness of drug interventions. Researchers have extensively researched Alzheimer's disease detection using several parameters and ADNI samples with one or more CNN architectures. To detect AD via a fine-tuned TL approach, we examined three deep neural network architectures (VGG-19, ResNet-50, and Inception V3) by applying the same parameters and using the same dataset samples from the ADNI database. Convolutional neural networks are robust tools for computer vision jobs, but training them from scratch can be difficult. Transfer learning leverages pretrained models for new data, significantly reducing training time while achieving superior results without overfitting. However, for optimal performance, fine-tuning via transfer learning necessitates extensive datasets. Gathering sufficient data in the relevant field is challenging, making data augmentation a valuable tool. To improve the performance and usability of the model, we utilized data augmentation methods to expand the dataset.

Data augmentation addresses issues such as imbalanced classes and overfitting, improving the model's robustness and tuning capabilities. In this study, we augmented the original dataset to include 39,980 images and retrained the CNN using the same data split and parameters. To assess the performance of the proposed AD detection framework, we conducted two separate studies on the original dataset and the augmented dataset. The performances of the fine-tuned deep learning models were tested on the original and augmented datasets. Table 7 and Table 8 present the results of the pretrained DNNs on the original and augmented datasets, respectively. Table 7 shows lower results for all the models because we needed to enhance the dataset. We applied rotation, flipping, scaling, zooming, and exclusion

techniques during preprocessing to increase the dataset size and improve accuracy. These steps are essential for enhancing image quality. We used data augmentation to create a broader range of data points, which reduced the gap between the training and validation sets. Table 8 shows the improved performance of the deep neural networks trained on the expanded dataset. By comparing the confusion matrices of the three pretrained networks, it can be found that data augmentation positively impacts their overall effectiveness during the transfer learning process. Remarkably, ResNet-50 boasts an impressive average accuracy of 98.7%, surpassing the performances of VGG-19 (97.16%) and Inception V3 (97.54%).

We used a leave-one-out cross-validation technique to assess our models. This involved training the model on all the data except for one instance, making a prediction based on that instance, and repeating this process for all the cases. We then calculated the average error to evaluate the models. We evaluated the models' performance based on sensitivity, specificity, and accuracy, as presented in Tables 4 and 5. The ResNet-50 model achieved the highest overall accuracy of 98.7% on the augmented dataset, representing the current state-of-the-art result for the five-way multiclass classification of Alzheimer's disease. Another significant aspect of the study is its emphasis on conducting genuine predictions. To ensure this, the model was trained without using the test data, preventing potential biases and overfitting that could compromise the credibility of the outcomes.

Our study involved a comprehensive review of various deep learning techniques that utilize brain imaging scans for Alzheimer's disease classification. We analyzed articles that employed individual deep learning methods and ensemble approaches that combined multiple deep learning models for classification. The timely and accurate diagnosis of Alzheimer's disease is crucial for effective treatment and therapy, and these approaches are invaluable in advancing the understanding and management of this disease. Table 11 compares our work's modalities, techniques, and accuracy with those found in the literature. All the approaches listed in the table employ the ADNI dataset. Among them, the ResNet-50 model yielded the best results, with an overall accuracy of 98.7% for 5-way classification (CN/EMCI/MCI/LMCI/AD) of AD stages. This model outperforms other methods, including those reported in the literature, in terms of accuracy. [20], [22], and [23] proposed a 4-way classification method using deep pretrained networks and transfer learning, which achieved accuracies of 97.5%, 94.53%, and 97%, respectively. Similarly, for 3-way classification, [21, 25, 35, 36] reported accuracies of 93.02%, 84%, 97.28%, and 95.23%, respectively. [28] and [29] proposed ensemble approaches for the 4-way classification of AD with accuracies of 97.35% and 93.88%, respectively. In our study, we employed deep pretrained models using a transfer learning approach, and the ResNet-50 model achieved the highest overall accuracy (98.7%), setting a new state-of-the-art benchmark for five-way multiclass classification of AD. Additionally, the ResNet-50 model classified the "EMCI" category with the highest accuracy, emphasizing that the model is particularly effective at identifying early signs of memory impairment, which is crucial for accurate and timely diagnosis of AD.

The field of deep learning is constantly evolving, and the importance of transfer learning is becoming increasingly apparent. In biomedical image analysis and classification, it is crucial to identify the most effective models from a range of successful models used in various image classification tasks. Rather than creating individualized models for specific scenarios or subjects, the focus should be on developing universally applicable models that promote consistency and reliability across different research studies.

Table 11. Evaluation of the proposed model against other CNNs employing transfer learning

Reference	Implemented Models	Dataset	Classification	Accuracy
[20]	CNN from scratch	ADNI	2-way (AD/CN)	95.6%
[21]	CNN from scratch	ADNI	3-way (CN/MCI/AD)	78.02%
	AlexNet			91.40%
	GoogLeNet			93.02%,
[22]	AlexNet	ADNI	3-way (CN/MCI/AD)	94.53%
	ResNet50			58.07%
[23]	2D CNN	ADNI	4-way (CN/EMCI/LMCI/AD)	93.61%
	3D CNN			95.17%
	Fine-tuned VGG-19			97%
[24]	EfficientNet-B0	ADNI	3-way (CN/MCI/AD)	92.98%
	EfficientNet-B2			94.42%,
	EfficientNet-B3			97.28%
[25]	Random forest	ADNI	2-way (AD/CN)	68%
	VGG-16			81%
	VGG-19			84%

[26]	CNN from scratch	ADNI	4-way (CN/MCI/LMCI/AD)	97.84%
[27]	Hybrid model (LeNet + AlexNet)	ADNI	3-way (CN/MCI/AD)	93.58%
[28]	Ensemble model	ADNI	4-way (CN/EMCI/MCI/AD)	97.35% (VGG-16 +EfficientNet-B2)
[29]	Ensemble Model using WPBEM	ADNI	3-way (NC/MCI/AD) 4-way (NC/EMCI/LMCI/AD)	93.92% 93.88%
Proposed Method	**VGG-19, ResNet-50, and Inception V3**	**ADNI**	**5-way (CN/EMCI/MCI/LMCI/AD)**	**VGG-19 (97.16) ResNet-50 (98.70) Inception V3 (97.54)**

This is especially relevant due to the difficulties of obtaining and processing medical scans and the limited availability of data. Identifying models that consistently produce reliable results across different imaging techniques is essential for advancing the field and improving the diagnosis and treatment of medical conditions such as Alzheimer's disease.

6. Conclusion

Detecting and classifying Alzheimer's disease in a multiclass setting is a difficult task, which emphasizes the need for classification frameworks and automated systems to effectively manage the disease and ease the burden on the healthcare system. This research aims to introduce a five-way Alzheimer's disease classification system using pretrained deep learning convolutional neural network (CNN) architectures with transfer learning methods. To address the issue of a highly imbalanced Alzheimer's disease dataset, we employed resampling methods such as oversampling and undersampling to balance the classes. We also applied various data augmentation techniques to enhance the extraction of salient features from MR images, which helped us achieve impressive results despite our limited dataset. We used three fine-tuned deep learning architectures, VGG-19, ResNet-50, and Inception V3, which were pretrained on the ImageNet dataset. We leveraged transfer learning and data augmentation techniques despite our limited dataset to achieve impressive results.

Our study used eight performance metrics to evaluate and compare the three models. We evaluated our method against existing state-of-the-art methods and found that the ResNet-50 model outperformed all the others. On augmented images, ResNet-50 achieved an exceptional classification accuracy of 98.70% across all five classes, making it the current state-of-the-art model for five-way AD classification.

We suggest exploring ensemble-based methods to further enhance these results. Given the exceptional performance of the proposed model compared to other state-of-the-art CNN models, there is potential for its application in detecting and classifying other diseases.

Acknowledgments:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments, which helped improve the quality of this work.

Funding Statement:

The author(s) received no specific funding for this study.

Authorship:

All the authors should have made substantial contributions to all of the following: (1) the conception and design of the study, acquisition of data, or analysis and interpretation of the data; (2) drafting the article or revising it critically for important intellectual content; and (3) final approval of the version to be submitted.

Availability of Data and Materials:

The data utilized for our study were sourced from the ADNI database, which is accessible at http://adni.loni.usc.edu/.

Conflict of interest:

The authors declare that they have no conflicts of interest.

References:

[1] Srivastava S, Ahmad R, Khare SK (2021) Alzheimer’s disease and its treatment by different approaches: A review. Eur J Med Chem 216:. https://doi.org/10.1016/J.EJMECH.2021.113320

[2] Gao S, Lima D (2022) A review of the application of deep learning in the detection of Alzheimer’s disease. Int J Cogn Comput Eng 3:1–8. https://doi.org/10.1016/J.IJCCE.2021.12.002

[3] Ferretti MT, Iulita MF, Cavedo E, et al (2018) Sex differences in Alzheimer disease - the gateway to precision medicine. Nat Rev Neurol 14:457–469. https://doi.org/10.1038/S41582-018-0032-9

[4] Naseer A, Zafar K (2022) Meta-feature based few-shot Siamese learning for Urdu optical character recognition. Comput Intell 38:1707–1727. https://doi.org/10.1111/COIN.12530

[5] Penney J, Ralvenius WT, Tsai LH (2020) Modeling Alzheimer's disease with ipsc-derived brain cells. Mol Psych 25:148–167. https://doi.org/10.1038/s41380-019-0468-3

[6] Palmer WC, Park SM, Levendovszky SR (2022) Brain state transition analysis using ultrafast fMRI differentiates MCI from cognitively normal controls. Front Neurosci 16:975305. https://doi.org/10.3389/FNINS.2022.975305/BIBTEX

[7] Dadar M, Manera AL, Ducharme S, Collins DL (2022) White matter hyperintensities are associated with gray matter atrophy and cognitive decline in Alzheimer's disease and frontotemporal dementia. Neurobiol Aging 111:54–63. https://doi.org/10.1016/J.NEUROBIOLAGING.2021.11.007

[8] Aderghal K, Khvostikov A, Krylov A, et al (2018) Classification of Alzheimer Disease on Imaging Modalities with Deep CNNs Using Cross-Modal Transfer Learning. Proc - IEEE Symp Comput Med Syst 2018-June:345–350. https://doi.org/10.1109/CBMS.2018.00067

[9] Veitch DP, Weiner MW, Aisen PS, et al (2019) Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. Alzheimers Dement 15:106–152. https://doi.org/10.1016/J.JALZ.2018.08.005

[10] Pan D, Zeng A, Jia L, et al (2020) Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning. Front Neurosci 14:501050. https://doi.org/10.3389/FNINS.2020.00259/BIBTEX

[11] Falahati F, Westman E, Simmons A (2014) Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimers Dis 41:685–708. https://doi.org/10.3233/JAD-131928

[12] Rathore S, Habes M, Iftikhar MA, et al (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. Neuroimage 155:530–548. https://doi.org/10.1016/J.NEUROIMAGE.2017.03.057

[13] Deng J, Dong W, Socher R, et al (2010) ImageNet: A large-scale hierarchical image database. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[14] Zeng N, Li H, Peng Y (2023) A new deep belief network-based multitask learning for diagnosis of Alzheimer's disease. Neural Comput Appl 35:11599–11610. https://doi.org/10.1007/S00521-021-06149-6

[15] Frid-Adar M, Diamant I, Klang E, et al (2018) GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321:321–331. https://doi.org/10.1016/J.NEUCOM.2018.09.013

[16] Bukowy JD, Dayton A, Cloutier D, et al (2018) Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. J Am Soc Nephrol 29:2081–2088. https://doi.org/10.1681/ASN.2017111210

[17] Tanveer M, Richhariya B, Khan RU, et al (2020) Machine learning techniques for the diagnosis of alzheimer's disease: A review. ACM Trans Multimed Comput Commun Appl 16:. https://doi.org/10.1145/3344998

[18] Liu X, Wang C, Bai J, Liao G (2020) Fine-tuning Pretrained Convolutional Neural Networks for Gastric Precancerous Disease Classification on Magnification Narrow-band Imaging Images. Neurocomputing 392:253–267. https://doi.org/10.1016/J.NEUCOM.2018.10.100

[19] Gao M, Bagci U, Lu L, et al (2018) Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput Methods Biomech Biomed Eng Imaging Vis 6:1–6. https://doi.org/10.1080/21681163.2015.1124249

[20] AbdulAzeem Y, Bahgat WM, Badawy M (2021) A CNN based framework for classification of Alzheimer's disease. Neural Comput Applic 33:10415–10428. https://doi.org/10.1007/s00521-021-05799-w

[21] Liu J, Li M, Luo Y, et al (2021) Alzheimer's disease detection using depthwise separable convolutional neural networks. Comput Methods Prog Biomed 203:. https://doi.org/10.1016/j.cmpb.2021.106032

[22] Al-Adhaileh MH (2022) Diagnosis and classification of Alzheimer's disease by using a convolution neural network algorithm. Soft Comput 26:7751–7762. https://doi.org/10.1007/s00500-022-06762-0

[23] Helaly HA, Badawy M, Haikal AY (2022) Deep Learning Approach for Early Detection of Alzheimer's Disease. Cognit Comput 14:1711–1727.

https://doi.org/10.1007/S12559-021-09946-2

[24] Savaş S (2022) Detecting the Stages of Alzheimer's Disease with Pretrained Deep Learning Architectures. Arab J Sci Eng 47:2201–2218. https://doi.org/10.1007/s13369-021-06131-3

[25] Antony F, Anita HB, George JA (2023) Classification on Alzheimer's Disease MRI Images with VGG-16 and VGG-19. Smart Innov Syst Technol 312:199–207. https://doi.org/10.1007/978-981-19-3575-6_22

[26] Raza N, Naseer A, Tamoor M, Zafar K (2023) Alzheimer Disease Classification through Transfer Learning Approach. Diagnostics 13:. https://doi.org/10.3390/DIAGNOSTICS13040801

[27] Hazarika RA, Maji AK, Kandar D, et al (2023) An Approach for Classification of Alzheimer's Disease Using Deep Neural Network and Brain Magnetic Resonance Imaging (MRI). Electron 12:676. https://doi.org/10.3390/electronics12030676

[28] Mujahid M, Rehman A, Alam T, et al (2023) An Efficient Ensemble Approach for Alzheimer's Disease Detection Using an Adaptive Synthetic Technique and Deep Learning. Diagnostics 2023, Vol 13, Page 2489 13:2489. https://doi.org/10.3390/DIAGNOSTICS13152489

[29] Fathi S, Ahmadi A, Dehnad A, et al (2023) A Deep Learning-Based Ensemble Method for Early Diagnosis of Alzheimer's Disease using MRI Images. Neuroinformatics 22:89–105. https://doi.org/10.1007/S12021-023-09646-2/TABLES/5

[30] Berrar D (2018) Cross-validation. Encycl Bioinforma Comput Biol ABC Bioinforma 1–3:542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

[31] Daldal N, Cömert Z, Polat K (2020) Automatic determination of digital modulation types with different noises using Convolutional Neural Network based on time–frequency information. Appl Soft Comput J 86:. https://doi.org/10.1016/J.ASOC.2019.105834

[32] Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc

[33] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-December:770–778. https://doi.org/10.1109/CVPR.2016.90

[34] Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the Inception Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-December:2818–2826. https://doi.org/10.1109/CVPR.2016.308

[35] Savaş S, Topaloğlu N, Kazcı Ö, Koşar PN (2019) Classification of carotid artery intima media thickness ultrasound images with deep learning. J Med Syst 43:273. https://doi.org/10.1007/s10916-019-1406-2

[36] Hazarika RA, Maji AK, Sur SN, et al (2021) A Survey on Classification Algorithms of Brain Images in Alzheimer's Disease Based on Feature Extraction Techniques. IEEE Access 9:58503–58536. https://doi.org/10.1109/ACCESS.2021.3072559