diagnostics

MDPI

Article

An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning

Tanjim Mahmud [1,*,†], Koushick Barua [1,*,†], Sultana Umme Habiba [2], Nahed Sharmen [3], Mohammad Shahadat Hossain [4] and Karl Andersson [5]

[1] Department of Computer Science and Engineering, Rangamati Science and Technology University, Rangamati 4500, Bangladesh
[2] Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh; habiba.kuet@gmail.com
[3] Department of Obstetrics and Gynecology, Chattogram Maa-O-Shishu Hospital Medical College, Chittagong 4100, Bangladesh; dr.nahedsharmen@gmail.com
[4] Department of Computer Science and Engineering, University of Chittagong, Chittagong 4331, Bangladesh; hossain_ms@cu.ac.bd
[5] Pervasive and Mobile Computing Laboratory, Luleå University of Technology, 97187 Luleå, Sweden; karl.andersson@ltu.se
* Correspondence: tanjim_cse@yahoo.com (T.M.); koushickbarua111@gmail.com (K.B.)
† These authors contributed equally to this work.

Abstract: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects millions of individuals worldwide, causing severe cognitive decline and memory impairment. The early and accurate diagnosis of AD is crucial for effective intervention and disease management. In recent years, deep learning techniques have shown promising results in medical image analysis, including AD diagnosis from neuroimaging data. However, the lack of interpretability in deep learning models hinders their adoption in clinical settings, where explainability is essential for gaining trust and acceptance from healthcare professionals. In this study, we propose an explainable AI (XAI)-based approach for the diagnosis of Alzheimer's disease, leveraging the power of deep transfer learning and ensemble modeling. The proposed framework aims to enhance the interpretability of deep learning models by incorporating XAI techniques, allowing clinicians to understand the decision-making process and providing valuable insights into disease diagnosis. By leveraging popular pre-trained convolutional neural networks (CNNs) such as VGG16, VGG19, DenseNet169, and DenseNet201, we conducted extensive experiments to evaluate their individual performances on a comprehensive dataset. The proposed ensembles, Ensemble-1 (VGG16 and VGG19) and Ensemble-2 (DenseNet169 and DenseNet201), demonstrated superior accuracy, precision, recall, and F1 scores compared to individual models, reaching up to 95%. In order to enhance interpretability and transparency in Alzheimer's diagnosis, we introduced a novel model achieving an impressive accuracy of 96%. This model incorporates explainable AI techniques, including saliency maps and grad-CAM (gradient-weighted class activation mapping). The integration of these techniques not only contributes to the model's exceptional accuracy but also provides clinicians and researchers with visual insights into the neural regions influencing the diagnosis. Our findings showcase the potential of combining deep transfer learning with explainable AI in the realm of Alzheimer's disease diagnosis, paving the way for more interpretable and clinically relevant AI models in healthcare.

Keywords: Alzheimer's disease; transfer learning; explainable AI (XAI); saliency maps; grad-CAM

check for updates

Citation: Mahmud, T.; Barua, K.; Habiba, S.U.; Sharmen, N.; Hossain, M.S.; Andersson, K. An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning. *Diagnostics* **2024**, *14*, 345. https://doi.org/10.3390/diagnostics14030345

Academic Editor: Dania Cioni

Received: 25 December 2023
Revised: 29 January 2024
Accepted: 31 January 2024
Published: 5 February 2024

1. Introduction

Alzheimer's disease is a progressive neurological disorder that primarily affects older adults, gradually impairing memory, cognitive functions, and eventually the ability to perform daily activities. It is the most common cause of dementia, a syndrome characterized by a decline in memory, thinking, behavior, and the ability to perform everyday activities [1].

According to the World Alzheimer's Association [2]. The prevalence of Alzheimer's disease in the United States is on the rise, with more than 6 million Americans currently affected by the condition. Among Americans aged 65 and older, an estimated 6.7 million individuals are living with Alzheimer's as of 2023, with the majority being 75 years or older. Approximately 10.7% of people aged 65 and older have Alzheimer's, and women make up nearly two-thirds of those affected. Furthermore, older Black and Hispanic Americans are disproportionately affected by Alzheimer's compared to older White Americans. With the aging population, it is projected that the number of individuals with Alzheimer's will continue to increase, potentially reaching 12.7 million by 2050 if no medical breakthroughs are made to prevent or cure the disease. Moreover, it is a formidable challenge in contemporary healthcare, affecting millions globally and imposing significant burdens on healthcare systems and societies [3,4]. These statistics underscore the urgent need for advancements in Alzheimer's research and care to address the growing impact of the disease on individuals, families, and society as a whole. The timely and accurate diagnosis of AD is imperative for facilitating early intervention and potential disease-modifying therapies. However, the intricacies of AD, coupled with the overlap of symptoms with other neurodegenerative conditions, render the accurate identification and classification of the disease, especially in its early stages, intricate tasks [5].

We know that machine learning is currently one of the methods to solve many kinds of problems, such as disease detection [6,7], sports [8], natural language processing [9–11], and so on. Based on this, this paper introduces a revolutionary approach to address the complexities of Alzheimer's disease diagnosis through the fusion of deep transfer learning and explainable artificial intelligence (XAI) techniques [12]. Our investigation involved the utilization of well-established pretrained convolutional neural networks (CNNs), including VGG16, VGG19, DenseNet169, and DenseNet201. Through extensive experimentation, we assessed their individual performances on a comprehensive dataset. Notably, our proposed ensembles, Ensemble-1 (VGG16 and VGG19) and Ensemble-2 (DenseNet169 and DenseNet201), exhibited superior metrics—accuracy, precision, recall, and F1 scores—outperforming individual models, with achievements reaching up to 95%.

In order to fortify interpretability and transparency in Alzheimer's diagnosis, we present a novel model attaining an impressive accuracy of 96%. This model integrates cutting-edge explainable AI techniques, such as saliency maps [13] and grad-CAM (gradient-weighted class activation mapping) [14].

The integration of these techniques not only enhances the model's outstanding accuracy but also provides clinicians and researchers with valuable visual insights into the neural regions pivotal in the diagnostic process.

The primary objectives of this paper are as follows:

1. To develop reliable Alzheimer's classification using a deep transfer learning ensemble;
2. To introduce a novel model for transparent Alzheimer's diagnosis with high accuracy;
3. To enhance interpretability using XAI methods such as saliency maps and grad-CAM;
4. To evaluate the proposed approach against benchmarks, highlighting superior accuracy and interpretability.

The subsequent sections of this paper are organized as follows: Section 2 provides an overview of related works in the field of AI-based AD diagnosis, transfer learning, and XAI. Section 3 outlines the methodology employed in developing the deep transfer learning ensemble and integrating XAI techniques. In Section 4, we present our experimental setup and evaluation results. The discussion of the findings is presented in Section 5, and finally, Section 6 concludes the paper with future directions for research in this domain.

2. Related Works

The field of Alzheimer's disease diagnosis has witnessed growing interest in leveraging artificial intelligence (AI) techniques, particularly deep learning [1], for enhanced accuracy and efficiency. Numerous studies have contributed valuable insights into various aspects of Alzheimer's disease classification, utilizing diverse data types, machine learning methodologies, and XAI techniques.

Ref. [15] focuses on a classification task distinguishing between healthy controls (HCs) and Alzheimer's disease (AD) using numeric data. It employs LIME and SHAP as XAI frameworks, revealing significant features, such as whole brain volume, years of education, and socio-economic status. The classifiers include support vector machine (SVM), k-nearest neighbors (kNN), and multilayer perceptron (MLP).

Ref. [16] utilizes image data for HCs vs. AD classification, incorporating XAI frameworks such as HAM and PCR. Significant features include salient aspects related to AD, like cerebral cortex and hippocampus atrophy. The classifier employed is a deep learning (DL) convolutional neural network (CNN), achieving an impressive accuracy of 95.4%.

By examining HCs, MCI, and AD, Ref. [17] employs image data and the XAI framework GNNExplainer. The significant features encompass volume, area of cortical regions, and vertex-based thickness measures. The classifier is a DL graph neural network (GNN), yielding an accuracy of 53.5 ± 4.5%.

By using image data, Ref. [18] adopts the XAI framework of occlusion sensitivity mapping, focusing on white matter hyperintensities (WMHs). The classifier is a DL EfficientNet-B0, achieving an accuracy of 80.0%.

Incorporating image data and XAI frameworks such as saliency maps and layer-wise relevance propagation (LRP), Ref. [19] considers MRI, 3D PET, biological markers, and assessments. The classifier is a DL 3D CNN AD.

Ref. [20] involves image data and XAI frameworks utilizing decision trees (DTs). The significant features include demographic data, cognitive factors, and brain metabolism data. The classifiers include Bernoulli naive Bayes (NB), SVM, kNN, random forest (RF), AdaBoost, and gradient boosting (GBoost), achieving an accuracy of 91.0%.

Utilizing image data, Ref. [21] employs XAI frameworks such as 3D Ultrametric Contour Map, 3D Class Activation Map, and 3D GradCAM. Significant features encompass 3D MRI features, and the classifier is a DL 3D CNN, achieving an accuracy of 76.6%.

Ref. [22], involving image data, utilizes Sensitivity Analysis and Occlusion as XAI frameworks. Significant features include 3D Image features, and the classifier is a DL 3D CNN, achieving an accuracy of 77.0%.

3. Materials and Methods

The section offers resources for the diagnosis of Alzheimer's disease using MRI images, as illustrated in Figure 1, in order to identify the disease at an early stage of development. The VGGNets (very deep convolutional networks) and DenseNets models were fed with the enhanced MRIs of Alzheimer's illness using the same technique across all platforms. First of all, we trained the AD dataset with VGGNets, such as VGG16 and VGG19. Then, we trained the AD dataset with DenseNets, such as DenseNet169 and DenseNet201. After examining the above net results, we integrated the VGGNets (VGG16 and VGG19) and DenseNets (DenseNet169 and DenseNet201) together.

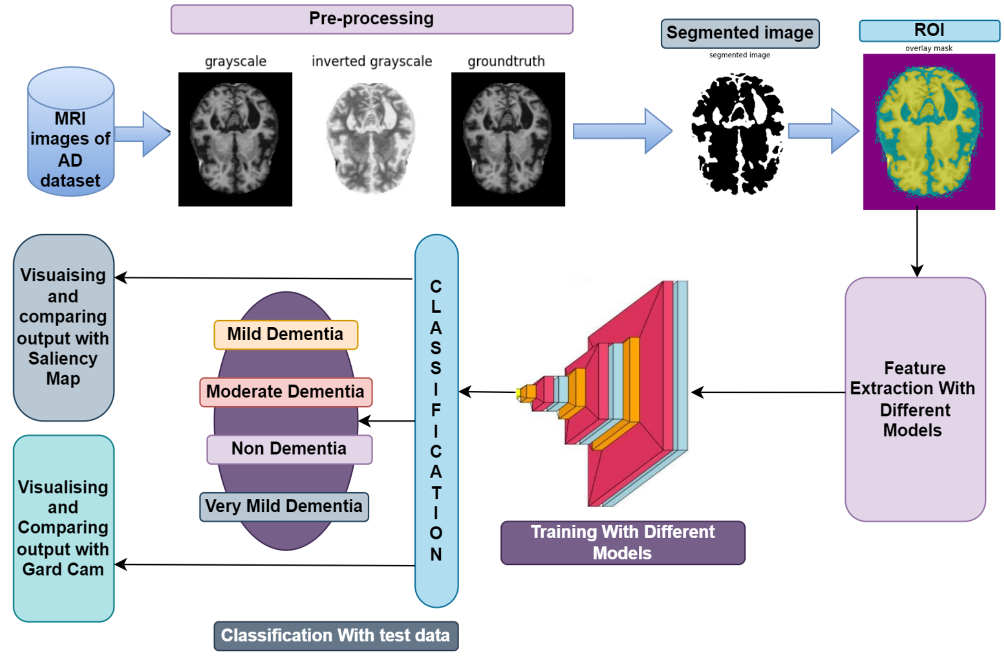


Figure 1. Workflow diagram for diagnosing MRI images for the detection of AD.

3.1. Dataset Collection

Methods with enhanced capabilities for the detection of AD were established in this study by analyzing MRI images from the AD dataset. The dataset was gathered and downloaded from kaggle [23]. It contained four classes and 6400 images. The 6400 images in the OASIS-2 collection have a size of 176 × 208. The images for each class are shown in Figure 2, with the labels and classes verified. The Alzheimer's dataset contains the following images for each class: 896 MRIs for mild dementia, 64 MRIs for moderate dementia, 3200 MRIs for non-dementia, and 2240 MRIs for very mild dementia. The samples from the Alzheimer's dataset are shown in Figure 2.

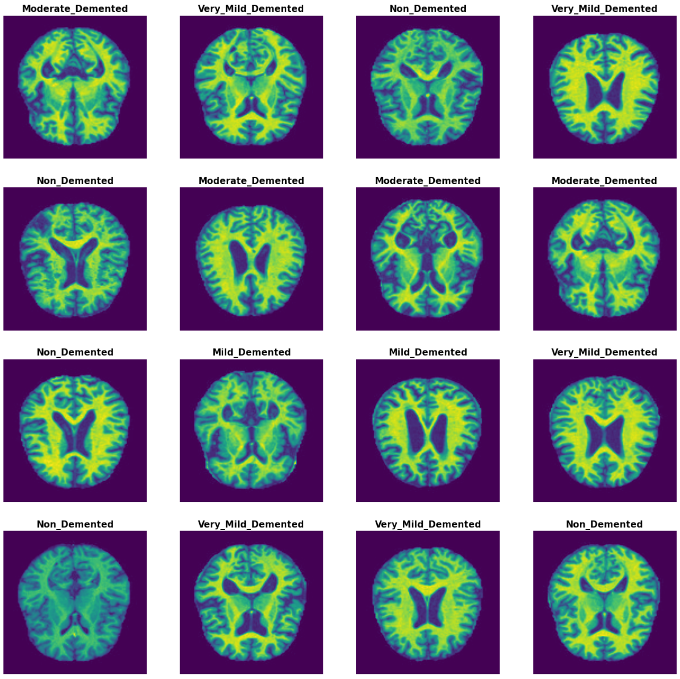


Figure 2. MRI image samples from AD dataset.

Table 1 demonstrates how the four classes of images are divided into an imbalanced configuration. According to the table, the dataset looks to be significantly imbalanced across all classes. The biased level contains 3200 images and traditional learning makes it difficult to classify all of the levels correctly because they are not all the same number. As a result, we employ data augmentation to label the images with an equal number, where every class has the same number of images. The procedure is described in the data augmentation section. Multiple transfer learning models and ensemble models were used to appropriately categorize all of the levels.

Table 1. Dataset statistics.

Class Name	Number of Images
Mild Dementia	896
Moderate Dementia	64
Non-Dementia	3200
Very Mild Dementia	2240

3.2. Image Preprocessing

Image preprocessing techniques are needed because of the noise that MRI images exhibit for a variety of reasons, including imaging procedures, brightness, reflections, low contrast, and acquiring MRI images from many sources. As a result, enhancing MRI images results in image enhancement, which aids in obtaining high performance and collecting the right information from the images [24].

3.2.1. Image Resizing

The complete set of images was shrunk to 224 × 224, including training, testing, and validation, when the primitive size was 176 × 208. The image size for transfer learning models was 224 × 224, and RGB values were used for ensemble models as well [25].

3.2.2. Normalization

Normalization is the process of transforming features to be of the same scale. This increases the model's performance and training stability. We projected into a preset range (i.e., usually [0, 1] or [−1, 1]) to normalize all image data, and we also used the same techniques (transfer learning and ensembling) on them, as described in Equation (1) and typically performed as follows:

$$img = \frac{1}{255.0} \tag{1}$$

3.2.3. Removing Noise by Median Filtering

Median filtering is utilized similarly to an averaging filter. In contrast, during median filtering, a pixel's value is determined by the neighborhood's median of pixels, as opposed to an average. Compared to the average, which is substantially less sensitive to outliers or extreme values, median filtering is more capable of removing these without compromising the image's clarity [26].

3.2.4. Image Segmentation

For a better image experience, we go through numerous segmentation procedures to extract the most useful information from the images. Compared to gray images, which are one-dimensional, RGB images have three color channels. First of all, we converted the RGB images to grayscale. In order to convert images from RGB to gray, we used the OpenCV library's built-in Canny edge detection function. The *Canny*() function can be used to conduct this process on an image; the function's syntax is as follows:

$$Canny(image, edges, T\_lower, T\_upper)$$

We set the lower threshold and upper threshold to 175.5 and 207.5, respectively. After that, we inverted the gray scale image, setting the value '0' to '1' and '1' to '0'. Now, "true and accurate" segmentations, which are typically made by one or more human experts, were used to convert the inverted grayscale images into ground truth. Following the preprocessing stage, the subsequent step involves segmentation, where a threshold value is determined. In the process of thresholding an image, the selected thresholds aim to optimize the distinction between the average gray levels of the foreground and background regions, along with the overall average gray level across the entire image. This optimization is reflected in the variance of the regions. Otsu's maximum variance method, derived from the least squares principle of discriminant analysis, is commonly employed in stable threshold segmentation approaches [27–29].

For a given input image, I, with a height of H and width of W, the image's normalized gray histogram is represented by the $histogram_1$. Here, $histogram_1(k)$ denotes the ratio of pixels in the image with a gray value equal to k, where k is in the range [0, 255]. The algorithm's detailed steps include straightforward calculations based on Otsu's method, ensuring an effective and widely utilized threshold segmentation approach.

Step 1: Calculate the zero-order cumulative moments, also known as cumulative histograms, of the gray histogram described in Equation (2).

$$zeroCumuMoment(k) = \sum_{i=0}^{k} histogram_1(i), k \in [0, 255] \tag{2}$$

Step 2: Compute the cumulative first-order moments of the gray histogram shown in Equation (3).

$$oneCumuMoment(k) = \sum_{i=0}^{k}((i) * histogramI(i)), k \in [0, 255] \tag{3}$$

Step 3: Compute the mean gray level of image I by calculating the first-order cumulative distance when k = 255, expressed as

$$meanoneCumuMoment = (255)$$

Step 4: In the process of determining each gray level as a threshold in Equation (4), the calculations involve the computation of the average gray level for the foreground area, the average gray level for the background area, and the variance of the overall average gray level for the entire image. The variance is measured using the following metrics:

$$\sigma^2(k) = \frac{(mean * zeroCumuMoment(k) - oneCumuMoment(k))^2}{zeroCumuMoment(k) * (1 - zeroCumuMoment(k))}, k \in [0, 255] \tag{4}$$

Step 5: From Equation (4) mentioned above, the corresponding k is the threshold of Ostu automatic selection, that is

$$thresh = arg_{k \in [0,255]} max(\sigma^2(k))$$

Following segmentation, we employed image masks in the form of (num_masks, H, W) or (H, W), with an alpha value between 0 and 1 indicating the masks' transparency. A value of 1 indicates complete transparency, whereas 0 indicates none. There are lists of colors that contain the mask colors or use a single color for all masks. PIL strings or RGB tuples, such (240, 10, 157), can be used to represent the color. Each mask generates random colors by default [30].

3.3. Balancing the Dataset through Data Augmentation

A dataset with limited images is one of the drawbacks of deep learning (pretrained) models. In order to be able to train the systems and provide them with enough data during the training phase, deep learning requires a dataset with a large number of images. Additionally, because the accuracy is skewed toward the class that contains the majority of images, the unbalanced dataset affects model performance and is one of its drawbacks. The data augmentation technique addresses the aforementioned issues by artificially boosting the dataset's image count with additional images from the same dataset [31].

In order to address the unbalanced dataset, the data augmentation technique [4] generates images from the minority class by using a more significant percentage than the rise in the majority class in order to balance out the unbalanced dataset. As a result, this method balances the dataset while increasing image quality. Images from the dataset are enhanced by a variety of processes, including horizontal and vertical shifting, horizontal and vertical flipping, random rotating, random zooming, random brightness, and others.

Figure 3 depicts images distributed within the dataset classes, where the image counts prior to data augmentation are relatively small and imbalanced. The diagram also depicts the dispersion of MRI. After performing data augmentation, the images were transferred between the dataset classes. Table 2 summarizes the MRI images amongst the dataset classes before and after training data augmentation is used. According to the table, Moderate dementia has the lowest number of images at 52, and the non-dementia class has the highest number of images at 2560. We chose a middle number of 1280, which was easily achievable for each class, with each having fewer or more images regarding their image class [32].

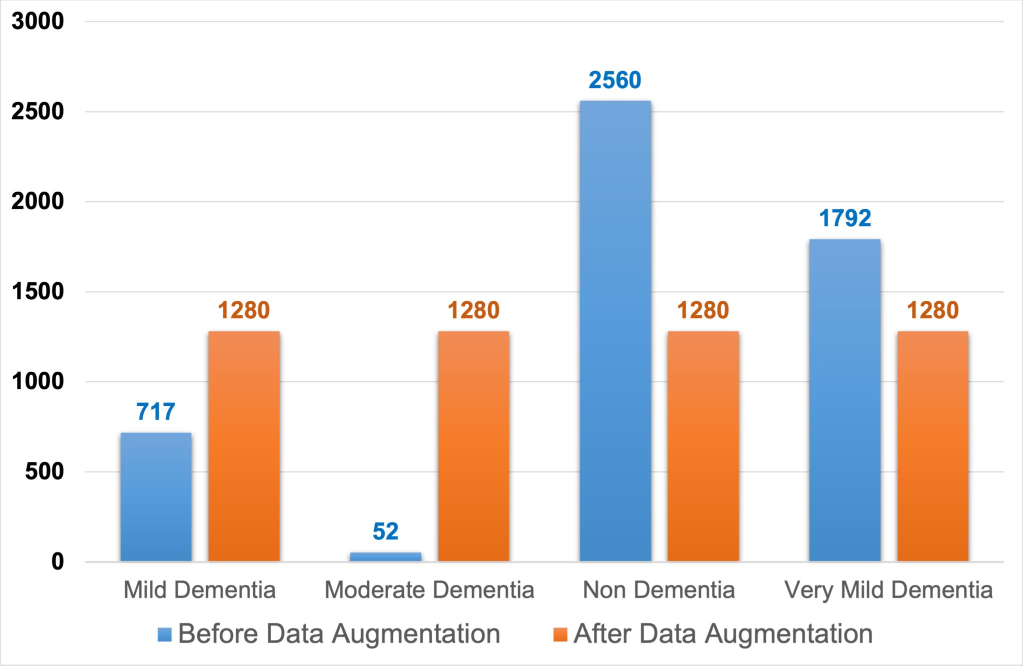


Figure 3. Training dataset before and after data augmentation.

Table 2. Training dataset before and after data augmentation.

Class Name	Before Data Augmentation	After Data Augmentation
Mild Dementia	717	1280
Moderate Dementia	52	1280
Non-Dementia	2560	1280
Very Mild Dementia	1792	1280

3.4. Feature Extraction

There are some difficulties with using deep learning (pretrained) models to train a dataset, including the need for powerful computers and the cost; as a result, we chose some pretrained models, such as VGG16 and VGG19 (VGGnets) and DenseNet169 and DenseNet201 (DenseNets) for feature extraction. At first, we went for VGG16 to extract features (with 512 features) and added a fully connected layer. Unlike VGG16, we constructed VGG19 for extracting features at 512 after connecting a fully connected layer. Therefore, the features matrix had a size of 5120 × 512 for each model. Then, we extracted features through DenseNet169 and DenseNet201, where 1664 features were extracted with a fully connected layer and 1920 features were extracted through a fully connected layer, respectively. Therefore, the features matrix had a size of 5120 × 1664 for DenseNet169 and 5120 × 1920 for DenseNet201. For feature extraction, by leveraging its ability to capture high-level features from input images, such as edges, textures, and shapes, we used EfficientNet as a feature extractor [1], where we realized a 6400 × 1280 feature matrix [33].

3.5. Addition of New Layers

After extracting the features from the above-mentioned feature extraction subsection, we added some layers for training (see Figure 4). For simplicity, we used the same layers for all pretrained models and ensemble models also.

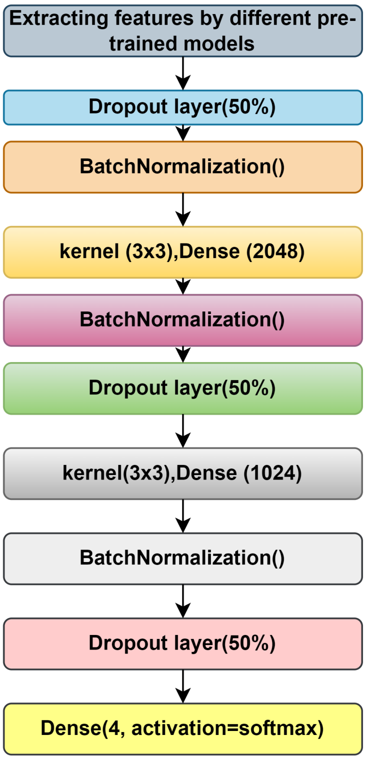


Figure 4. Addition of new layers after feature extraction.

We used a dropout layer for batch normalization, two dense layers with a kernel size of 3 × 3, the activation function of which is 'RelU', and finally, an output layer with four nodes, with the activation function being 'softmax' [34].

3.6. Training with Pretrained Models

In this section, we will describe the materials needed to train the training dataset using the 5120 images contained within the training data. First of all, we focus on the VGG16 and VGG19 models, then DenseNet169, and finally, DenseNet201 [35].

3.6.1. VGG19 and VGG19

The architecture of VGG16 and VGG19 is Google's pretrained architecture, which uses fine-tuning across all layers and replaces the top layers with 53,540,868 trainable parameters (as is the case for both the models shown in Figures 5 and 6). Some new layers were added, which are described in Section 3.5.

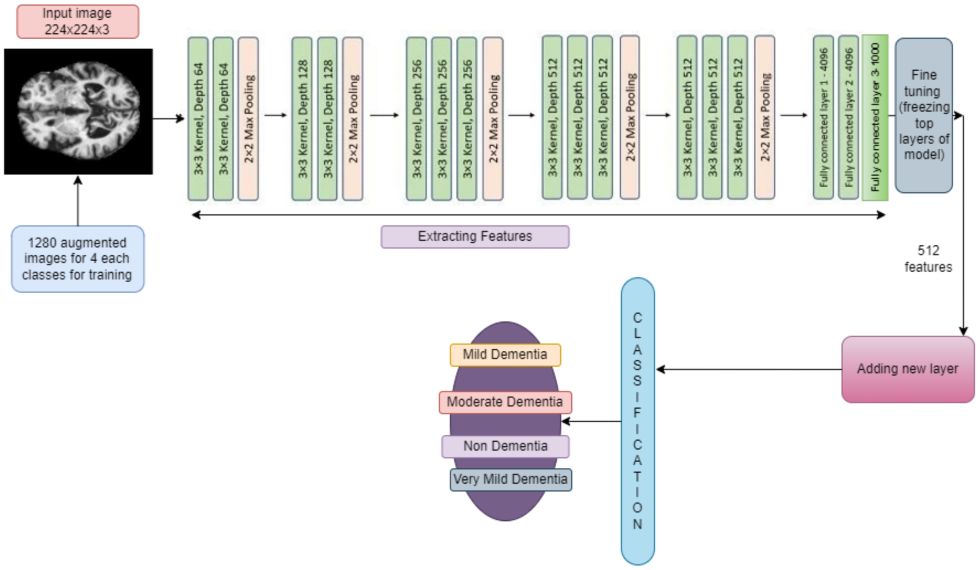


Figure 5. Addition of new layers after feature extraction using VGG16.

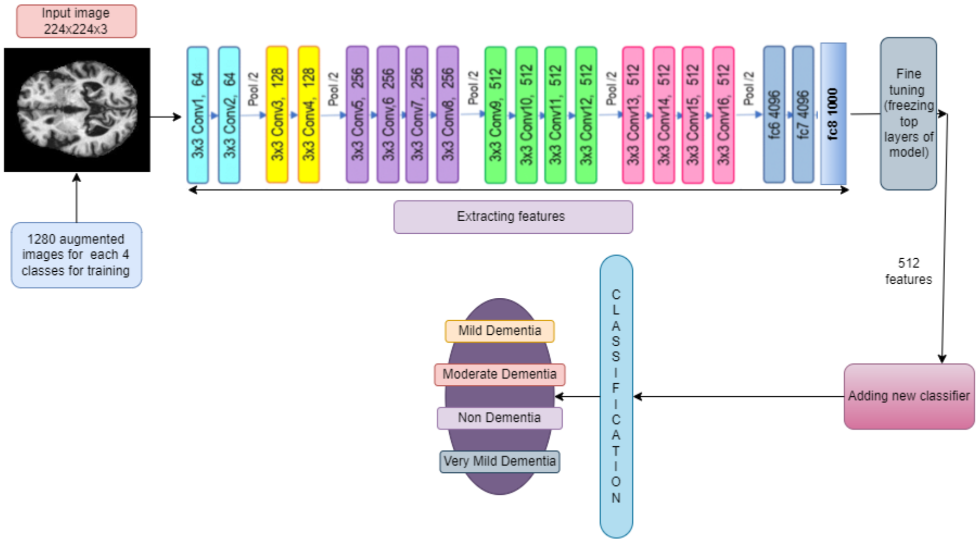


Figure 6. Addition of new layers after feature extraction using VGG19.

The input images were all resized to (224, 224) to be compatible with this model. The learning rate was set to 0.001, and the Adam optimizer algorithm was used as the optimizer.

3.6.2. DenseNet169 and DenseNet201

A DenseNet, short for dense convolutional network, is a type of convolutional neural network that incorporates dense connections between layers through dense blocks. These blocks establish direct connections between all layers, ensuring matching feature map sizes. In order to preserve the feed-forward structure, each layer not only receives additional inputs from all preceding layers but also transmits its own feature maps to all subsequent layers. This design facilitates enhanced information flow and gradient propagation throughout the network, contributing to improved learning capabilities. Fine-tuning across all layers and replacing top layers with 169,259,268 and 194,974,468 trainable parameters for DenseNet169 and DenseNet201, respectively, is shown in Figures 7 and 8. Some new layers were added, which are described in Section 3.5.

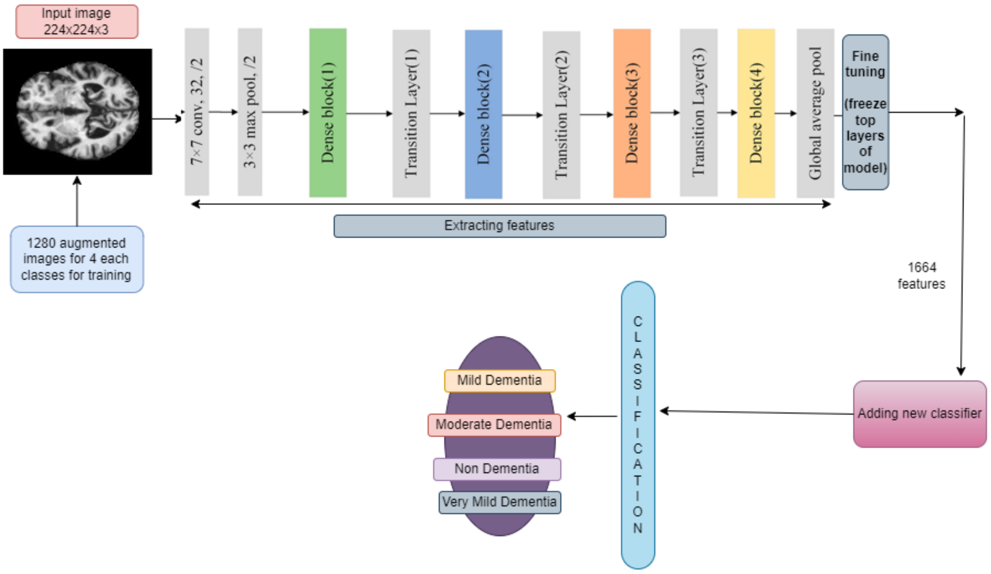


Figure 7. Addition of new layers after feature extraction using DenseNet169.

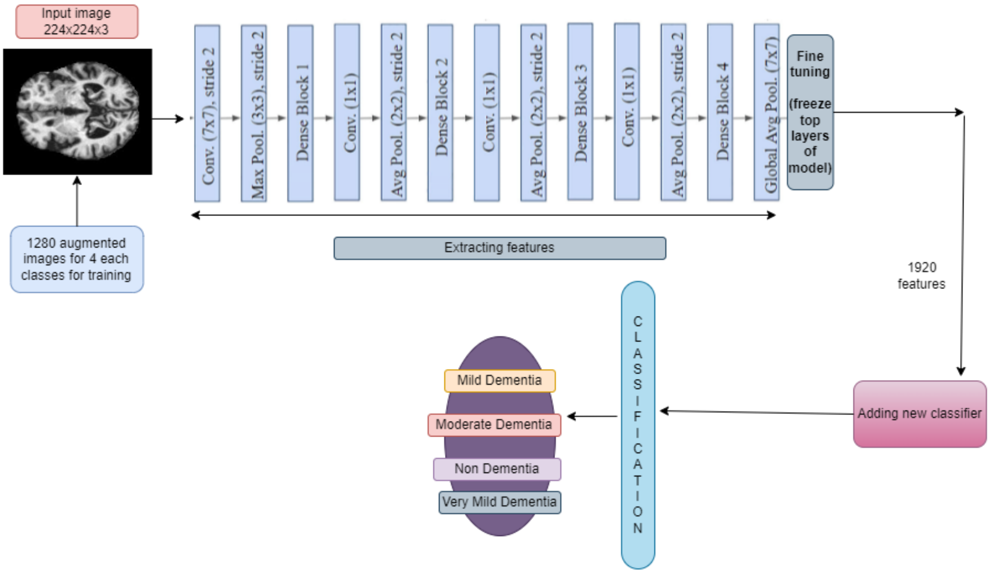


Figure 8. Addition of new layers after feature extraction using DenseNet201.

The input images underwent resizing to dimensions (224, 224) to align with the compatibility requirements of the model. A learning rate of 0.001 was chosen, and the optimization process employed the Adam optimizer algorithm.

3.7. Ensemble Models

An ensemble model is a machine learning strategy that integrates several other models in the prediction process. These models are known as base estimators. The technical obstacles experienced when developing a single estimator can be overcome by using ensemble models [36]. We employed two ensemble models, which are detailed below.

3.7.1. Ensemble-1

In order to improve accuracy, we investigated two ensemble models, Ensemble-1 and Ensemble-2, with the Ensemble-1 model combining VGG16 and the VGG19 models, comprising 107,081,736 trainable parameters. The model architecture of Ensemble-1 is shown in Figure 9. For Ensemble-2, the Input images were all resized to (224, 224) to be compatible with this model. The learning rate was set to 0.001, and the Adam optimizer algorithm was used as the optimizer.

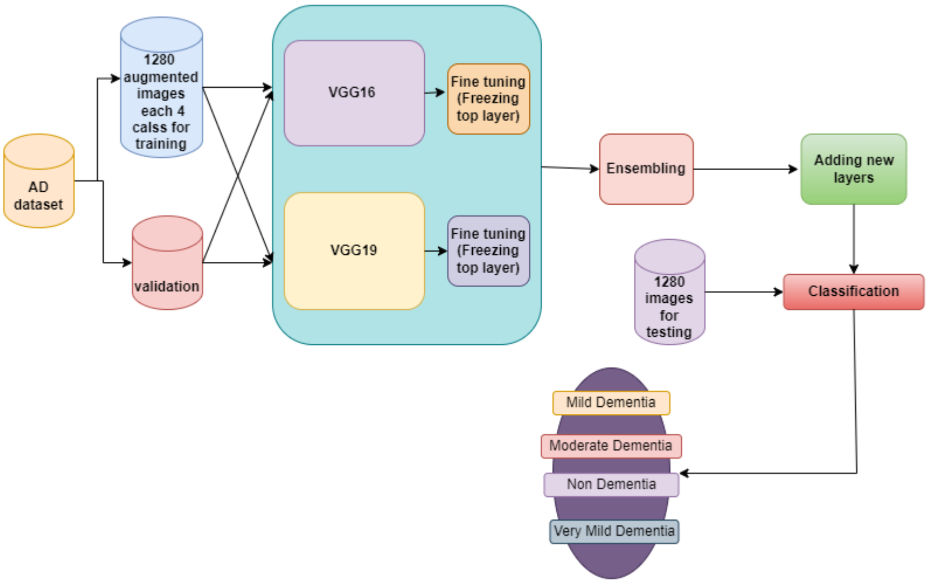


Figure 9. Ensemble-1 model architecture.

3.7.2. Ensemble-2

The Ensemble-2 model incorporates the DenseNet169 and DenseNet201 models together with 364,233,736 trainable parameters. Figure 10 shows the model description, with input images all resized to (224, 224) to be compatible with this model. The learning rate was set to 0.001, and the Adam optimizer algorithm was used as the optimizer.

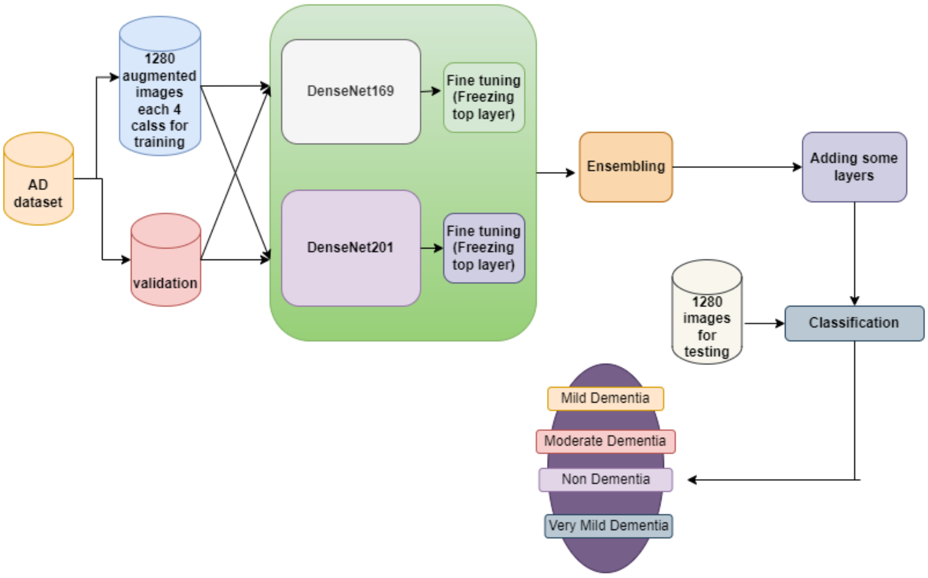


Figure 10. Ensemble-2 model architecture.

3.8. Proposed Model

Instead of traditional learning, deep learning, like other convolutional neural networks, is used for the training, testing, and validation, with the input layer of all images having a size of 224x224 for the RGB values. Then, a convolution and pooling layer, such as max pooling and min pooling, is used, with a fully connected layer for classifying all classes. This figure is a basic CNN model after successful feature extraction by using EfficientNet. A convolution layer (32, 64, 128, 256, and 512, with a kernel size 3 × 3, a pooling layer with flatten, and a dense layer with 512 units) and a final output of four dense layers were added. Table 3 shows a description of the architecture of the proposed model.

Table 3. Description of the proposed model architecture.

Model Content	Details
Input Image Size	224 × 224 × 3, with 5120 training images and 1280 images in each class
Feature extraction	Using EfficientNet with 1280 features
First Convolution Layer	32 filters; size = 3 × 3; ReLu; Padding = 'Same'
First Max Pooling Layer	Pooling Size: 2 × 2
Second Convolution Layer	64 filters; size = 3 × 3; ReLu; Padding = 'Same'
Second Max Pooling Layer	Pooling size: 2 × 2
Third Convolution Layer	128 filters; size = 3 × 3; ReLu; Padding = 'Same'
Third Max Pooling Layer	Pooling size: 2 × 2
Fourth Convolution Layer	256 filters; size = 3 × 3; ReLu; Padding = 'Same'
Fourth Max Pooling Layer	Pooling Size: 2 × 2
Fifth Convolution Layer	512 filters; size = 3 × 3; ReLu; Padding = 'Same'
Fifth Max Pooling Layer	Pooling Size: 2 × 2
Fully Connected Layer	4096 nodes; ReLU
Dropout Layer	50% Neurons dropped randomly
Dense_1 Layer	8320 nodes; ReLu
Dense_2 Layer	516 nodes; ReLu
Output Layer	Four nodes; Softmax activation
Optimization Function	Adam optimization
Learning Rate	0.001
Loss Function	Categorical cross entropy

3.9. Saliency Map

In computer vision, a saliency map is a visualization approach that highlights the most essential regions or features within an image. It aids in the identification of areas of interest or the significance of a certain job, such as object detection or image classification. The saliency map shows the image regions that get the most attention, pixel by pixel.

There are numerous approaches for creating saliency maps, but one typical method is to analyze the gradients of a pretrained convolutional neural network (CNN). The goal is to compute the gradients of the output class score with respect to the pixels in the input image. High gradients show that a minor change in a specific pixel has a substantial impact on the output score, implying that the relevant region is important [37].

The typical process for creating a saliency map using gradient-based techniques is summarized below:

Select a CNN model that has already been trained using a sizable dataset, such as ImageNet. The network must forward-pass an image in order to determine the expected class probabilities. Then, it must determine the gradients of the projected class score in relation to the pixels of the input image. By using the gradients, it calculates the importance scores for each pixel (for example, by taking the absolute values or squaring them). In order to have the importance ratings fall inside a particular range, such as [0, 1], the values are normalized. Finally, the network maps the importance scores to the image dimensions to create the saliency map, emphasizing the salient areas [38].

Saliency maps can be used for a variety of purposes, such as object localization, image captioning, and human attention modeling. They provide insights into deep learning models' visual attention mechanisms and aid in understanding which components of an image contribute the most to the model's decision-making process. The input image x is passed through the network to determine the network output value, f (x), which is used to compute the saliency map. The gradient of $f(x)_y$ with respect to the input image x is then calculated using a backward pass, where y is the ground truth label corresponding to the input image x. Formally, Formula (6) is used to determine the gradient $G(x)$, which estimates the significance of each pixel in the image, x. $G(x)$ has the same dimensions as the image, x, and is a tensor. $G(x)$ is a tensor with a dimension of 3, $W \times H$ and is indexed by three indexes: i for indexing channels and j, k for i. This is the case if x has a width, W, height H, and three channels [39].

$$G(x) = \frac{df(x)_y}{dx} \tag{5}$$

The maximum of the absolute values across channels is determined to estimate the relevance of a pixel $x(i, j)$. As a result, the created matrix with the dimension $W \times H$ is known as saliency map(SM).

$$SM(i,j) = Max[|G(0,i,j), |G(1,i,j), |G(2,i,j)|] \tag{6}$$

The saliency map can localize (with good precision) the infected regions in the input leaf image.

3.10. Grad-CAM

The grad-CAM method leverages gradients between the classification score and the ultimate convolutional feature map to identify specific regions within an input image that exert the greatest impact on the classification score. The significance of these areas in influencing the final score is heightened in locations where the gradient is pronounced [40]. When we focus this on the images, it detects their regions with the help of the accuracy they achieved. Now, we move to grad-cam, where an activation map that localizes the identified object to a specific area of the image is known as the *Grad-CAM* output. It has both width, u, and height, v, for class c.

$$L^{e}_{Grad-CAM} \in R^{(uxv)}$$

The gradients take the shape (u,v,Z), where (u,v) is the width and height of the 2D convolution filter, and Z is the number of filters. The following phase averages each of the filters to produce a single value, resulting in a final shape of Z or the number of filters. This corresponds to the global average pooling 2D layer.

$$\alpha^{c}_{k} = \frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A^{k}_{ij}} \tag{7}$$

Each one of these gradients represents the connection from one of the pixels in the 2D array to the neuron/output representing the target class. This is accomplished by the global average pooling 2D layer; the following layer in the model flattens Z and averages the number of form filters (u $\times v$) to single numbers. In order to establish a link between the final prediction outputs and the fully linked (Dense) layers, this is required. Subsequently, the gradients can be increased, which indicates the significance of the provided feature map or filter by using the feature map or filter that it truly reflects [41].

$$ReLU\left(\sum_{k} a^{c}_{k} A^{k}\right) \tag{8}$$

We can generate a map of the regions that would reduce the network's confidence in its prediction by negating the value of

$$\frac{\partial y^c}{\partial A^k}.$$

4. Experimental Results

This section provides an overview of the results obtained from our model. The entire experiment was conducted on "Google-Colab", utilizing the provided GPU, specifically the "Tesla-K80". The training of the model involved using 80% of the available information, and the evaluation was performed on the remaining 20%. A comprehensive comparison of various models is presented in this section. In order to assess the model objectively, classification matrices, including accuracy, precision, recall, and the macro F1 score, were computed. These metrics provide a thorough evaluation of the model's performance across different aspects without introducing bias.

4.1. Splitting Dataset

In order to facilitate the categorization or prediction tasks in a machine learning domain, the OASIS dataset is partitioned into three distinct sections: training, testing, and validation. Examining Table 1 reveals a notable imbalance in the dataset. During the dataset split, 80% of the total images are allocated for training and validation purposes, while the remaining 20% is designated for testing.

4.2. Accuracy

Accuracy is the most intuitive performance measure, and it is simply a ratio of the correctly classified observations to the total observations. It is said that if accuracy is high, then the model is accurate in predicting or classifying. The equation for accuracy is shown below in Equation (9).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

4.3. Precision

Precision is the ratio of the correctly predicted positive observations to the total predicted positive observations. The equation for precision is shown below in Equation (10).

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

4.4. Recall

Recall is the ratio of the correctly predicted positive observations to all observations in an actual class. The equation for recall is given below in Equation (11).

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

4.5. F1 Score

The $F1$ score is also alluded to as the $F1$ measure. It is nothing but the weighted harmonic mean of recall and precision. It is calculated in Equation as follows (12).

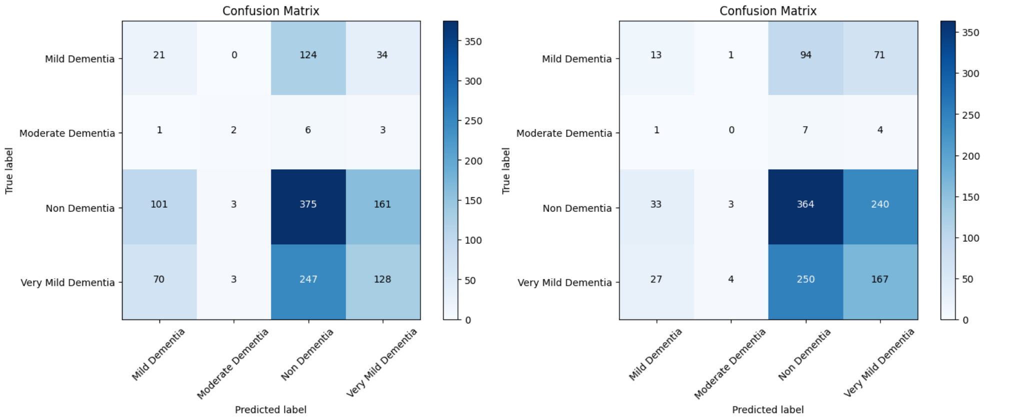
$$F1 - score = 2 * \frac{PRECISION * RECALL}{PRECISION + RECALL} \tag{12}$$

TP denotes true positive, TN denotes true negative, FP denotes fake positive, and FN denotes false negative.

4.6. Results

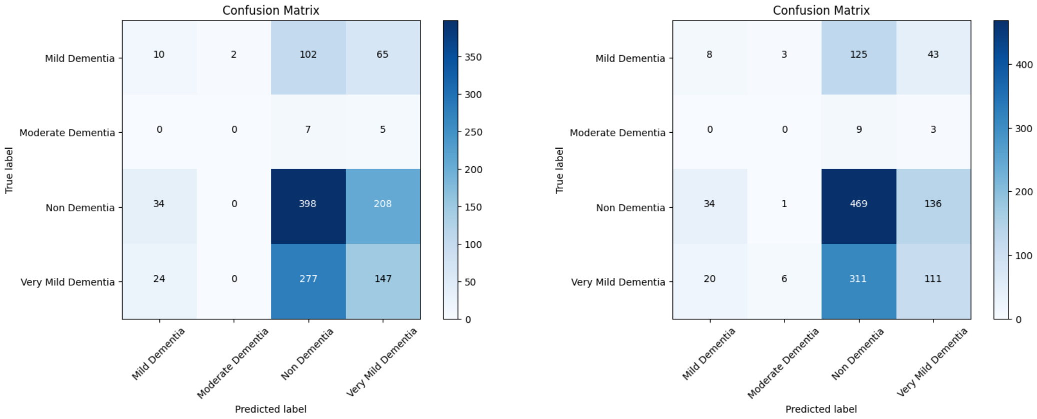
The section includes a confusion matrix for evaluating different network performances by merging the pretrained model and the ensemble model with CNN features with the features for AD progression diagnosis. Because of the similarity in the MRI images during the evolution of Alzheimer's disease, the features are critical when distinguishing between the phases of the disease. Thus, in this phase, the VGGNet model's deep features were combined with handcrafted features, and this was saved in a features matrix; the deep features of DenseNet169 and DenseNet201 are coupled with the handcrafted features and saved in a features matrix. Finally, the CNN was given two 15 out of the 22 feature matrixes, which divided the feature matrix into 80% for training and validation and 20% for performance.

The confusion matrix for different network performances in the evolution of AD detection is shown in Figures 11–13 . For each class—mild dementia, moderate dementia, non-dementia, and very mild—the models using the merged features between the VGGNets and ensembles obtained a true predictive of 492 images, which was the highest score, and 365 was the lowest for non-demntia. In comparison, the ensembles combining features from VGGNets and DenseNets found 192 and 227 for non-dementia, respectively; for the proposed model, the number is 170, as shown in Figure 13.



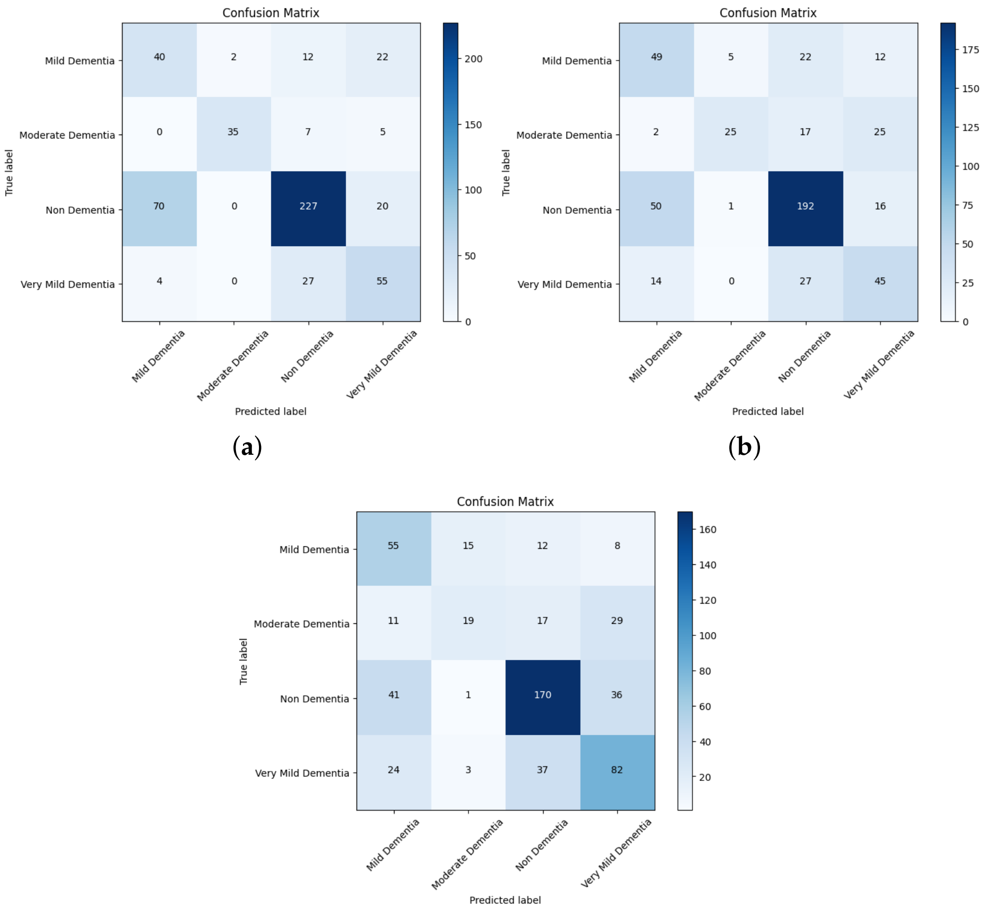
(**a**) (**b**)

Figure 11. Confusion matrix for VGG16 and VGG19. (**a**) VGG16. (**b**) VGG19.



(**a**) (**b**)

Figure 12. Confusion matrix for DenseNet169 and Densenet201. (**a**) DenseNet169. (**b**) DenseNet201.

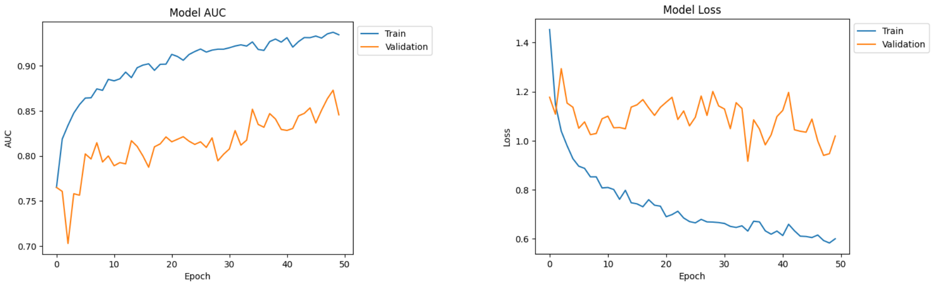


(**a**) (**b**)

(**c**)

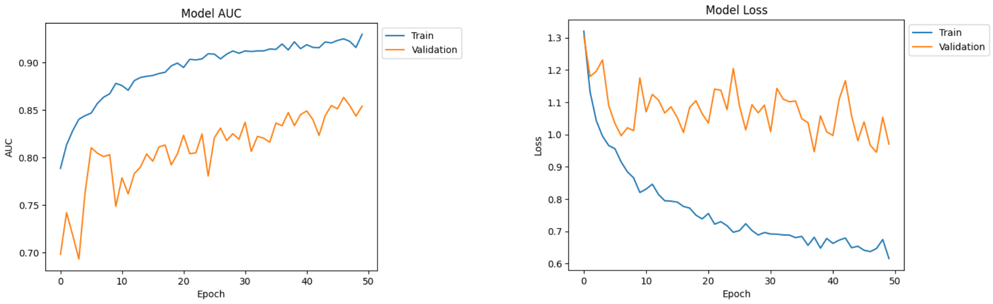
Figure 13. Confusion matrix for Ensemble-1, Ensemble-2, and the proposed model. (**a**) Ensemble-1. (**b**) Ensemble-2. (**c**) Proposed model.

Table 4 depicts the accuracy of the proposed model and all the transfer learning models and ensemble models. With more epochs, better accuracy can be obtained. After 100 successful epochs for the proposed model, it provided an accuracy of 96%, a precision of 89%, a recall of 93% and an F1 score of 91%. The transfer learning models have the highest accuracy for VGG16 at 90%.The other model—VGG19, DenseNet169, and DenseNet20—had an accuracy of 89%, 87%, and 88% respectively. For the two ensemble models, 95% accuracy was achieved; Ensemble-1 had an accuracy of 92%. Figures 14–17 depict the training and validation curves for VGG16, VGG19, DenseNet169, and DenseNet201. The accuracy and loss curves for the two ensemble models are also shown in Figures 18 and 19. Finally, the accuracy–loss curve for the proposed model can be seen in Figure 20.



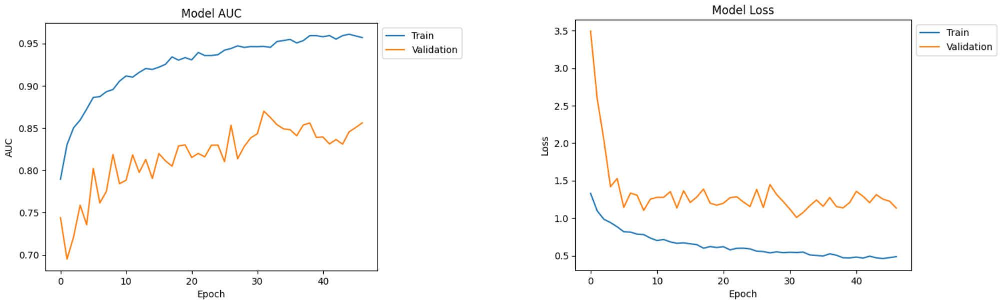
(**a**) (**b**)

Figure 14. Accuracy and loss curve for VGG16. (**a**) Accuracy. (**b**) Loss.



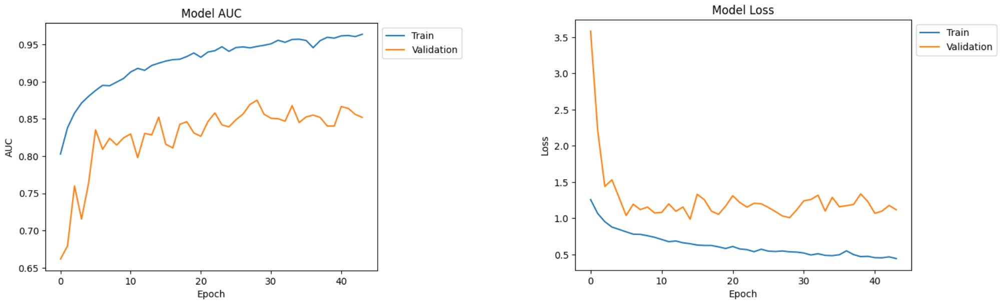
(**a**) (**b**)

Figure 15. Accuracy and loss curve for VGG19. (**a**) Accuracy. (**b**) Loss.



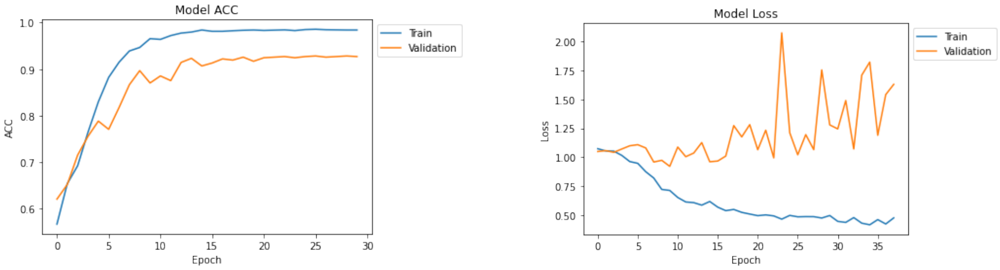
(**a**) (**b**)

Figure 16. Accuracy and loss curve for DenseNet169. (**a**) Accuracy. (**b**) Loss.



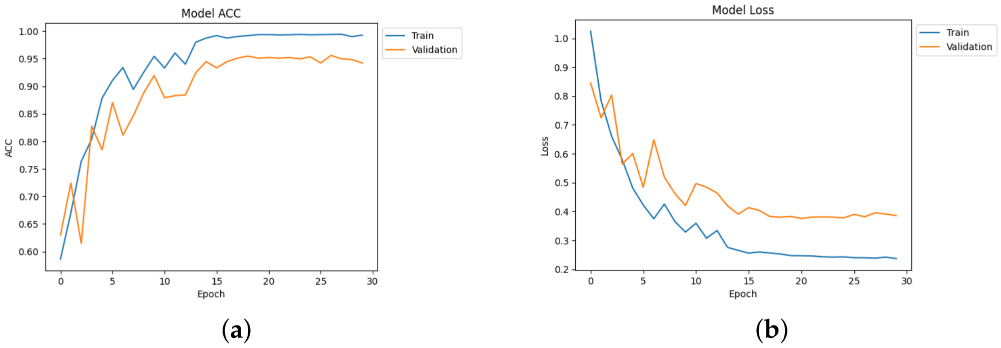
(**a**) (**b**)

Figure 17. Accuracy and loss curve for DenseNet201. (**a**) Accuracy. (**b**) Loss.



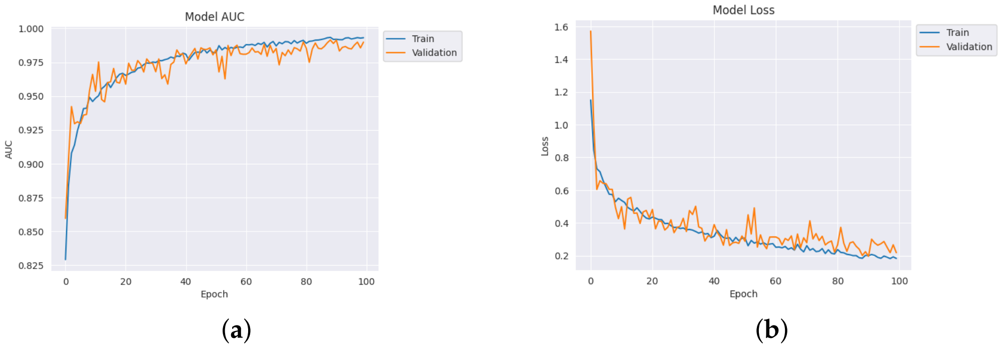
(**a**) (**b**)

Figure 18. Accuracy and loss curve for Ensemble-1. (**a**) Accuracy. (**b**) Loss.



(**a**) (**b**)

Figure 19. Accuracy and loss curve for Ensemble-2. (**a**) Accuracy. (**b**) Loss.



(**a**) (**b**)

Figure 20. Accuracy and loss curve for the proposed model. (**a**) Accuracy. (**b**) Loss.

Table 4. Model performance.

Model_Name	Accuracy	Precision	Recall	F1-Score
VGG16	90%	89%	84%	86%
VGG19	89%	86%	82%	84%
DenseNet169	87%	85%	84%	83%
DenseNet201	88%	86%	81%	85%
Ensemble-1	92%	90%	88%	87%
Ensemble-2	95%	91%	90%	89%
Proposed model	96%	89 %	93%	91%

4.7. Exploring Saliency Maps and Grad-CAM

As previously mentioned in Sections 3.9 and 3.10, the relevant studies have focused on computational techniques such as saliency maps and grad-CAM. These techniques serve as analytical tools that estimate the significance of each pixel solely by traversing through the network.

The underlying concept of this approach lies in identifying pixels that have a substantial impact on the node corresponding to the input image. If altering the values of a particular pixel leads to a noticeable change in that node, it is considered significant. Conversely, pixels with gradients close to zero are deemed unimportant, as their fluctuations do not affect the output node associated with the input images. Aggregate analysis across channels is employed to determine the collective relevance of the pixels rather than assessing the importance of each pixel channel in isolation.

For Alzheimers disease classification, we used saliency maps and grad-CAM. A saliency map and grad-cam can be used in a similar way as in occlusion experiments; they aid users in recognizing illness symptoms. In addition, this approach lacks sensitivity to the scattered critical areas since pixel importance is determined analytically rather than through occluding pixels.

This method might be thought of as an analytical form of occlusion experiments. The saliency map matrix is numerically computed in occlusion tests by modifying pixels and analyzing the output changes.

As a result, calculating a saliency map is not as computationally expensive as calculating a heat map in occlusion experiments because calculating a gradient in a numerically discrete manner necessitates the modification of each pixel or region in the image in order to approximate the gradient.

However, when calculating a gradient analytically, only one backward pass is required to calculate all derivatives with respect to all pixels.

The saliency map and grad-CAM are capable of accurately locating the diseased areas within the supplied AD image. Good examples of how the visualization of saliency maps precisely marked the infected sections of leaves are shown in Figures 21 and 22. Furthermore, the two distributed AD mold disease patches are localized, whereas the occlusion trials only reveal one diseased zone. Even with these encouraging outcomes, saliency maps frequently have noisy activations and lack clarity, which can be frustrating to users. For instance, in addition to the infected locations, the visualizations can display numerous activated regions. We have gone through several processes of saliency mapping operations, where some images are taken as input and are used to visualize the images, as was the case for the model we used throughout our paper. We used all the models represented in Figures 21 and 22 to explore the maps. As mentioned earlier in Table 4, the models acquired some accuracy, and we tried to visualize the accurate combination and how the images behave when it is time to predict using saliency maps and gard-cam. It seems the edges are correctly identified, and there are some yellow dots in the saliency maps, as well as when there are two competing factors in the image; this is helpful. With these areas hidden, we can create a “counterfactual” image that should increase confidence in the first prediction. By combining this with guided backpropagation, which zeroes the gradient parts that have a detrimental impact on choice, the grad-CAM output can be further enhanced. This method more accurately reflects producing a high-resolution map with the same resolution as the input image using the guided backpropagation methodology. The high-resolution map is then masked using the grad-CAM heatmap to concentrate solely on the details that contributed to the prediction outcome.

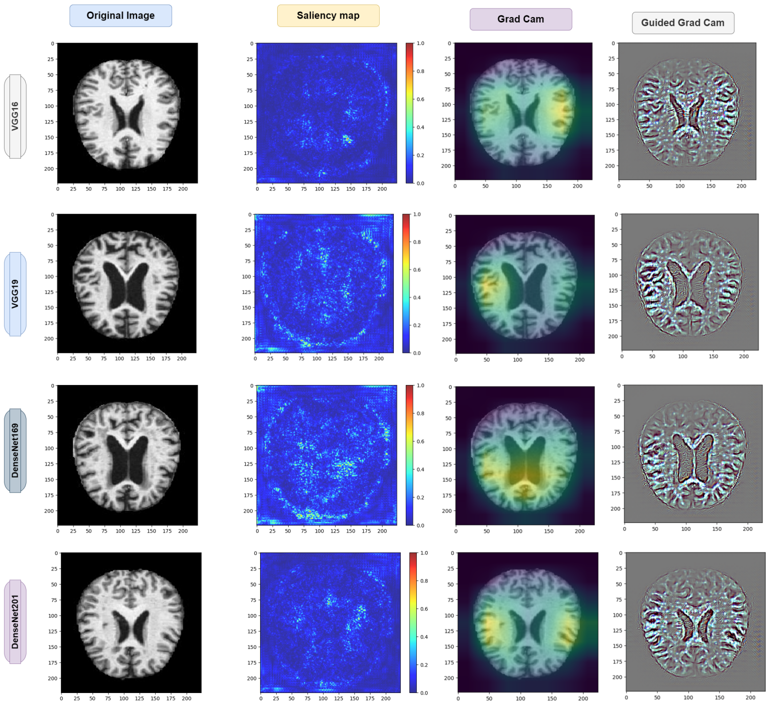


Figure 21. Visual depiction of the saliency map and grad-CAM results for the pretrained model.

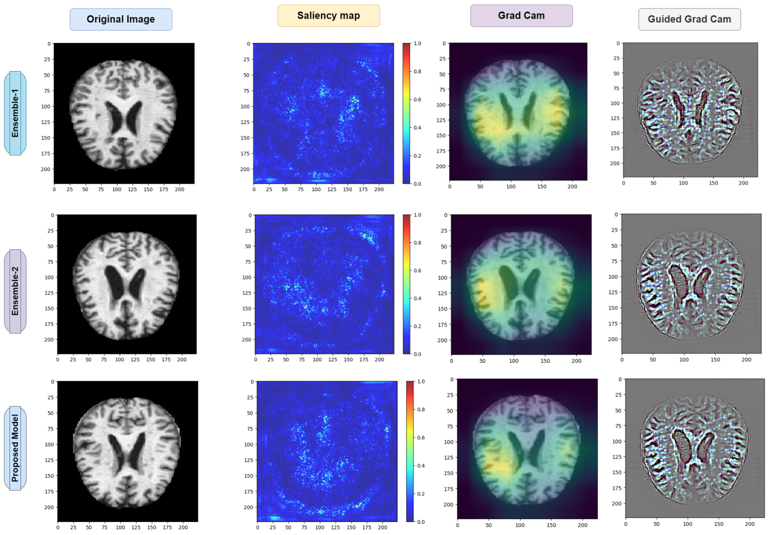


Figure 22. Visual depiction of the saliency map and grad-CAM results for the ensemble and proposed model.

5. Discussion

5.1. Interpretation of Results

Performance Evaluation: This study evaluated various deep transfer learning models, including VGG16, VGG19, DenseNet169, and DenseNet201, as well as two ensembles, termed Ensemble-1 and Ensemble-2. The proposed model outperformed all individual models and ensembles, achieving an impressive accuracy of 96%. The precision, recall, and F1-score metrics were also considered, indicating robust performance across different aspects of classification. Notably, the proposed model demonstrated a superior recall and F1 score, highlighting its effectiveness in correctly identifying Alzheimer's cases while minimizing false negatives.

Comparison with Baseline Models: The performance of the proposed model was compared against baseline models such as VGG16, VGG19, DenseNet169, and DenseNet201. While these models exhibited respectable accuracies, ranging from 87% to 90%, they were surpassed by the proposed model, emphasizing its efficacy in Alzheimer's diagnosis. Furthermore, the two ensemble models showed promising results, particularly Ensemble-2, which achieved a remarkable accuracy of 95%. This underscores the potential benefits of integrating multiple architectures for improved classification outcomes.

Interpretability Using XAI Methods: The integration of saliency maps and grad-CAM in the proposed model enhances its interpretability and transparency in Alzheimer's diagnosis. These XAI techniques provide valuable insights into the neural regions influencing diagnostic decisions, aiding clinicians and researchers in understanding the underlying mechanisms driving classification outcomes. By visualizing the regions of interest identified by the model, clinicians can gain confidence in its diagnostic recommendations and explore the potential biomarkers associated with Alzheimer's disease pathology.

5.2. Comparison with Previous Studies

This study stands at the forefront of Alzheimer's disease diagnosis research by integrating deep transfer learning and explainable artificial intelligence (XAI) techniques. A comparative analysis with prior studies in the field highlights the distinctive contributions and advancements of our proposed model (see Table 5).

The proposed method, which employs VGG16, VGG19, DenseNet169, and DenseNet201, as well as two ensembles (VGG16/VGG19 and DenseNet169/DenseNet201), along with EfficientNetB3 and CNNs, achieved a remarkable accuracy of 96% on MRI OASIS scans.

In order to fortify interpretability and transparency in Alzheimer's diagnosis, our study introduces a novel model that achieves an impressive accuracy of 96%. What distinguishes our model is the integration of cutting-edge XAI techniques, including saliency maps and grad-CAM (gradient-weighted class activation mapping). Unlike some of the existing studies that focus on specific XAI frameworks, such as LIME [15], SHAP [15], GNNExplainer [17], or occlusion sensitivity mapping [18], our research embraces a combination of techniques to provide a comprehensive and visually intuitive understanding of the model's decision-making process.

Table 5. Comparison of the proposed method with the state-of-the-art.

Paper	Classifier	Best Score (Accuracy)	XAI Method	Dataset
[15]	Support Vector Machines, KNN, MLP	91.4%	LIME, SHAP	Dementia dataset
[16]	CNN	95.4%	HAM, PCR	MRI scans ADNI
[17]	Graph Neural Network (GNN)	53.5 ± 4.5%	GNN Explainer	ADNI
[18]	EfficientNetB0	80%	Occlusion Sensitivity Mapping	MRI scans OASIS
[19]	3D CNN	-	Saliency Map, LRP	18F-FDG PET
[20]	KNN, RF, AdaBoost, Gradient Boosting Bernouli NB, SVM	91%	DT	Cognitive and and PET images
[21]	3D CNN	76.6%	3D Ultrametric Contour Map, 3D Class Activation Map, 3D GradCAM	ADNI
[22]	3D CNN	77%	Sensitivity Analysis Occlusion	MRI, PET
Proposed Method	VGG16, VGG19, DenseNet169, DenseNet201, Ensemble 1 (VGG16, VGG19) Ensemble 2 (DenseNet169, DenseNet20), Proposed model (EfficientNetB3 & CNN)	96%	Saliency maps, Grad-CAM (Gradient-weighted Class Activation Mapping)	MRI scans OASIS

When compared to the current methods, we see the following results:

Support vector machines, MLP, and KNN: Achieved 91.4% accuracy using LIME and SHAP on a dementia dataset [15].

CNN: Achieved 95.4% accuracy using HAM and PCR on MRI ADNI scans [16].

Graph neural network (GNN): Yielded an accuracy range of 53.5 ± 4.5% using GNN explainer on Ex-ADNI data [17].

EfficientNetB0: Attained 80% accuracy through occlusion sensitivity mapping on MRI OASIS scans [18].

3D CNN (multiple instances): Performance varies across different approaches and datasets, ranging from 18% to 77% accuracy, using various XAI methods on PET and MRI scans [19].

Various classifiers (KNN, RF, AdaBoost, gradient boosting, Bernoulli NB, and SVM): Achieved 91% accuracy using DT on cognitive and PET images [20].

3D CNN (using various XAI methods): Achieved 76.6% accuracy using 3D ultrametric contour maps, 3D class activation maps, and 3D gradCAM on ADNI scans [21].

3D CNN (with sensitivity analysis occlusion): Achieved 77% accuracy using sensitivity analysis occlusion on MRI and PET scans [22].

6. Conclusions and Future Work

6.1. Conclusions

This paper presents a pioneering approach to Alzheimer's disease diagnosis by integrating deep transfer learning and explainable artificial intelligence (XAI) techniques. The extensive experimentation with popular pretrained convolutional neural networks (CNNs) demonstrated the effectiveness of ensembles, specifically Ensemble-1 (VGG16 and VGG19) and Ensemble-2 (DenseNet169 and DenseNet201), in terms of achieving superior diagnostic performance, reaching up to 95% accuracy, precision, recall, and F1 scores. This signifies the potential of leveraging pretrained models and ensemble techniques for enhanced predictive capabilities in Alzheimer's diagnosis.

In order to address the crucial need for interpretability and transparency in medical diagnostics, we introduced a novel model with an impressive accuracy of 96%. The incorporation of explainable AI methodologies, including saliency maps and grad-CAM, played a pivotal role in enhancing the accuracy of our model. Furthermore, these techniques offered clinicians and researchers insightful visualizations of the neural regions, influencing diagnostic decisions.

6.2. Contributions of This Study

In summary, this research significantly contributes to the realm of Alzheimer's disease detection through the following key contributions:

1. The development of a robust deep transfer learning ensemble model for accurate Alzheimer's disease classification;
2. The introduction of a novel diagnostic model that achieved an impressive 96% accuracy;
3. The advancement of interpretability and transparency through the integration of explainable AI techniques such as saliency maps and grad-CAM;
4. The implementation of cutting-edge XAI methods to provide transparent and intuitive explanations for diagnostic predictions;
5. A rigorous performance evaluation using benchmark datasets showcasing the model's superiority in both accuracy and interpretability when compared to state-of-the-art methods.

6.3. Future Work

As we look ahead, there are several avenues for future research and development in this domain. Firstly, further exploration of deep transfer learning ensembles could involve investigating additional pretrained models and their combinations to optimize diagnostic performance. Additionally, refining the explainable AI techniques employed in our model and exploring newer methods might contribute to even more transparent and interpretable diagnostic systems.

Furthermore, the evaluation of the proposed model on diverse and larger datasets would enhance its generalizability and robustness. Collaborations with healthcare institutions for real-world validation and integration into clinical workflows would be crucial to assess the practical utility and impact of the developed model in a clinical setting.

The continuous evolution of AI and medical imaging technologies opens possibilities for incorporating multi-modal data and advanced feature extraction methods. Integrating genetic, demographic, or longitudinal data might potentially improve the accuracy and early detection capabilities of the diagnostic model.

Author Contributions: Conceptualization, T.M. and K.B.; methodology, T.M. and K.B.; validation, T.M., K.B., S.U.H., N.S., M.S.H. and K.A.; formal analysis, T.M., K.B., S.U.H., N.S., M.S.H. and K.A.; investigation, T.M., K.B., S.U.H., N.S., M.S.H. and K.A.; resources, T.M., K.B., S.U.H., M.S.H. and K.A.; data curation, T.M., K.B., S.U.H., N.S., M.S.H. and K.A.; writing—original draft preparation, T.M. and K.B.; writing—review and editing, T.M., K.B., S.U.H., M.S.H. and K.A.; visualization, T.M. and K.B.; supervision, T.M., M.S.H. and K.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used to support the findings of this study are available upon reasonable request to the corresponding author.

Acknowledgments: The research was conducted at the Artificial Intelligence Laboratory (AI Lab) of Rangamati Science and Technology University in Bangladesh. The authors would like to extend their heartfelt appreciation to the head of the lab for offering invaluable guidance, consistent evaluation, and unwavering support throughout the duration of the research. Special recognition is also due to the dedicated lab members whose active involvement significantly contributed to the collection and refinement of the dataset. Additionally, we utilized ChatGPT for grammar refinement, which greatly enhanced the quality of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's disease
XAI	Explainable AI
CNNs	Convolutional neural networks
Grad-CAM	Gradient-weighted class activation mapping
VGGNets	Very deep convolutional networks
MRI	Magnetic resonance imaging

References

1. Umme Habiba, S.; Debnath, T.; Islam, M.K.; Nahar, L.; Shahadat Hossain, M.; Basnin, N.; Andersson, K. Transfer Learning-Assisted DementiaNet: A Four Layer Deep CNN for Accurate Alzheimer's Disease Detection from MRI Images. In Proceedings of the International Conference on Brain Informatics, Hoboken, NJ, USA, 1–3 August 2023; Springer: Cham, Switzerland, 2023; pp. 383–394.
2. Alzheimer's Association. Alzheimer's Disease Facts and Figures. Available online: https://www.alz.org/alzheimers-dementia/facts-figures (accessed on 10 January 2024).
3. Khalid, A.; Senan, E.M.; Al-Wagih, K.; Ali Al-Azzam, M.M.; Alkhraisha, Z.M. Automatic Analysis of MRI Images for Early Prediction of Alzheimer's Disease Stages Based on Hybrid Features of CNN and Handcrafted Features. *Diagnostics* **2023**, *13*, 1654. [CrossRef]
4. Mahmud, T.; Barua, K.; Barua, A.; Das, S.; Basnin, N.; Hossain, M.S.; Andersson, K.; Kaiser, M.S.; Sharmen, N. Exploring Deep Transfer Learning Ensemble for Improved Diagnosis and Classification of Alzheimer's Disease. In Proceedings of the International Conference on Brain Informatics, Hoboken, NJ, USA, 1–3 August 2023; Springer: Cham, Switzerland, 2023; pp. 109–120.
5. Balaji, P.; Chaurasia, M.A.; Bilfaqih, S.M.; Muniasamy, A.; Alsid, L.E.G. Hybridized deep learning approach for detecting Alzheimer's disease. *Biomedicines* **2023**, *11*, 149. [CrossRef] [PubMed]
6. Mahmud, T.; Barua, A.; Begum, M.; Chakma, E.; Das, S.; Sharmen, N. An Improved Framework for Reliable Cardiovascular Disease Prediction Using Hybrid Ensemble Learning. In Proceedings of the 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 23–25 February 2023; pp. 1–6.
7. Mahmud, T.; Barua, A.; Islam, D.; Hossain, M.S.; Chakma, R.; Barua, K.; Monju, M.; Andersson, K. Ensemble Deep Learning Approach for ECG-Based Cardiac Disease Detection: Signal and Image Analysis. In Proceedings of the 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 21–23 September 2023; pp. 70–74.
8. Das, S.; Mahmud, T.; Islam, D.; Begum, M.; Barua, A.; Tarek Aziz, M.; Nur Showan, E.; Dey, L.; Chakma, E. Deep Transfer Learning-Based Foot No-Ball Detection in Live Cricket Match. *Comput. Intell. Neurosci.* **2023**, *2023*, 2398121. [CrossRef] [PubMed]
9. Mahmud, T.; Ptaszynski, M.; Eronen, J.; Masui, F. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Inf. Process. Manag.* **2023**, *60*, 103454. [CrossRef]
10. Mahmud, T.; Ptaszynski, M.; Masui, F. Vulgar Remarks Detection in Chittagonian Dialect of Bangla. *arXiv* **2023**, arXiv:2308.15448.
11. Mahmud, T.; Das, S.; Ptaszynski, M.; Hossain, M.S.; Andersson, K.; Barua, K. Reason based machine learning approach to detect bangla abusive social media comments. In Proceedings of the International Conference on Intelligent Computing & Optimization, Hua Hai, Thailand, 27–28 October 2022; Springer: Cham, Switzerland, 2022; pp. 489–498.
12. Sudar, K.M.; Nagaraj, P.; Nithisaa, S.; Aishwarya, R.; Aakash, M.; Lakshmi, S.I. Alzheimer's Disease Analysis using Explainable Artificial Intelligence (XAI). In Proceedings of the 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 7–9 April 2022; pp. 419–423.
13. Lopes, M.K.S.; Cassani, R.; Falk, T.H. On the use of saliency maps and convolutional neural networks for improved Alzheimer's disease assessment. *Alzheimer's Dement.* **2021**, *17*, e054710. [CrossRef]
14. Özbay, F.A.; Özbay, E. An NCA-based Hybrid CNN Model for Classification of Alzheimer's Disease on Grad-CAM-enhanced Brain MRI Images. *Turk. J. Sci. Technol.* **2023**, *18*, 139–155. [CrossRef]
15. Loveleen, G.; Mohan, B.; Shikhar, B.S.; Nz, J.; Shorfuzzaman, M.; Masud, M. Explanation-driven hci model to examine the mini-mental state for alzheimer's disease. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 1–16. [CrossRef]
16. Yu, L.; Xiang, W.; Fang, J.; Chen, Y.P.P.; Zhu, R. A novel explainable neural network for Alzheimer's disease diagnosis. *Pattern Recognit.* **2022**, *131*, 108876. [CrossRef]
17. Kim, M.; Kim, J.; Qu, J.; Huang, H.; Long, Q.; Sohn, K.A.; Kim, D.; Shen, L. Interpretable temporal graph neural network for prognostic prediction of Alzheimer's disease using longitudinal neuroimaging data. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 1381–1384.
18. Bordin, V.; Coluzzi, D.; Rivolta, M.W.; Baselli, G. Explainable AI points to white matter hyperintensities for Alzheimer's disease identification: A preliminary study. In Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, 11–15 July 2022; pp. 484–487.

19. De Santi, L.A.; Pasini, E.; Santarelli, M.F.; Genovesi, D.; Positano, V. An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F-FDG PET. *J. Digit. Imaging* **2023**, *36*, 189–203. [CrossRef]
20. García-Gutierrez, F.; Díaz-Álvarez, J.; Matias-Guiu, J.A.; Pytel, V.; Matías-Guiu, J.; Cabrera-Martín, M.N.; Ayala, J.L. GA-MADRID: Design and validation of a machine learning tool for the diagnosis of Alzheimer's disease and frontotemporal dementia using genetic algorithms. *Med. Biol. Eng. Comput.* **2022**, *60*, 2737–2756. [CrossRef]
21. Yang, C.; Rangarajan, A.; Ranka, S. Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification. In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 3–7 November 2018; American Medical Informatics Association: Bethesda, MD, USA, 2018; Volume 2018, p. 1571.
22. Rieke, J.; Eitel, F.; Weygandt, M.; Haynes, J.D.; Ritter, K. Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. In Proceedings of the Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 16–20 September 2018; Springer: Cham, Switzerland, 2018; pp. 24–31.
23. Kumar, S.; Shastri, S. *Alzheimer MRI Preprocessed Dataset*; Kaggle: San Francisco, CA, USA, 2022. . [CrossRef]
24. Kavitha, C.; Mani, V.; Srividhya, S.; Khalaf, O.I.; Tavera Romero, C.A. Early-stage Alzheimer's disease prediction using machine learning models. *Front. Public Health* **2022**, *10*, 853294. [CrossRef] [PubMed]
25. Chui, K.T.; Gupta, B.B.; Alhalabi, W.; Alzahrani, F.S. An MRI scans-based Alzheimer's disease detection via convolutional neural network and transfer learning. *Diagnostics* **2022**, *12*, 1531. [CrossRef]
26. Sharma, S.; Guleria, K.; Tiwari, S.; Kumar, S. A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans. *Meas. Sensors* **2022**, *24*, 100506. [CrossRef]
27. Savaş, S. Detecting the stages of Alzheimer's disease with pre-trained deep learning architectures. *Arab. J. Sci. Eng.* **2022**, *47*, 2201–2218. [CrossRef]
28. Beata, B.K.; Wojciech, J.; Johannes, K.; Piotr, L.; Barbara, M. Alzheimer's disease—Biochemical and psychological background for diagnosis and treatment. *Int. J. Mol. Sci.* **2023**, *24*, 1059. [CrossRef] [PubMed]
29. Odusami, M.; Maskeliūnas, R.; Damaševičius, R. Pixel-Level Fusion Approach with Vision Transformer for Early Detection of Alzheimer's Disease. *Electronics* **2023**, *12*, 1218. [CrossRef]
30. Savonije, K.; Weaver, D.F. The Role of Tryptophan Metabolism in Alzheimer's Disease. *Brain Sci.* **2023**, *13*, 292. [CrossRef]
31. Sisodia, P.S.; Ameta, G.K.; Kumar, Y.; Chaplot, N. A review of deep transfer learning approaches for class-wise prediction of Alzheimer's disease using MRI images. *Arch. Comput. Methods Eng.* **2023**, *30*, 2409–2429. [CrossRef]
32. Van Dyck, C.H.; Swanson, C.J.; Aisen, P.; Bateman, R.J.; Chen, C.; Gee, M.; Kanekiyo, M.; Li, D.; Reyderman, L.; Cohen, S.; et al. Lecanemab in early Alzheimer's disease. *N. Engl. J. Med.* **2023**, *388*, 9–21. [CrossRef]
33. Varesi, A.; Pierella, E.; Romeo, M.; Piccini, G.B.; Alfano, C.; Bjørklund, G.; Oppong, A.; Ricevuti, G.; Esposito, C.; Chirumbolo, S.; et al. The potential role of gut microbiota in Alzheimer's disease: From diagnosis to treatment. *Nutrients* **2022**, *14*, 668. [CrossRef]
34. Xing, X.; Rafique, M.U.; Liang, G.; Blanton, H.; Zhang, Y.; Wang, C.; Jacobs, N.; Lin, A.L. Efficient training on Alzheimer's disease diagnosis with learnable weighted pooling for 3D PET brain image classification. *Electronics* **2023**, *12*, 467. [CrossRef]
35. Zeng, N.; Li, H.; Peng, Y. A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease. *Neural Comput. Appl.* **2023**, *35*, 11599–11610. [CrossRef]
36. Rallabandi, V.S.; Seetharaman, K.; Alzheimer's Disease Neuroimaging Initiative. Classification of cognitively normal controls, mild cognitive impairment and Alzheimer's disease using transfer learning approach. *Biomed. Signal Process. Control* **2023**, *79*, 104092.
37. Oktavian, M.W.; Yudistira, N.; Ridok, A. Classification of Alzheimer's Disease Using the Convolutional Neural Network (CNN) with Transfer Learning and Weighted Loss. *arXiv* **2022**, arXiv:2207.01584.
38. Hajamohideen, F.; Shaffi, N.; Mahmud, M.; Subramanian, K.; Al Sariri, A.; Vimbi, V.; Abdesselam, A. Four-way classification of Alzheimer's disease using deep Siamese convolutional neural network with triplet-loss function. *Brain Inform.* **2023**, *10*, 1–13. [CrossRef] [PubMed]
39. Brahimi, M.; Arsenovic, M.; Laraba, S.; Sladojevic, S.; Boukhalfa, K.; Moussaoui, A. Deep learning for plant diseases: Detection and saliency map visualisation. In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*; Springer: Cham, Switzerland, 2018; pp. 93–117.
40. Patil, V.; Madgi, M.; Kiran, A. Early prediction of Alzheimer's disease using convolutional neural network: A review. *Egypt. J. Neurol. Psychiatry Neurosurg.* **2022**, *58*, 1–10. [CrossRef]
41. Moujahid, H.; Cherradi, B.; Al-Sarem, M.; Bahatti, L.; Eljialy, A.B.A.M.Y.; Alsaeedi, A.; Saeed, F. Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation. *Intell. Autom. Soft Comput.* **2022**, *32*, 723–745. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.