

Article

Enhancing Alzheimer's Disease Detection through Ensemble Learning of Fine-Tuned Pre-Trained Neural Networks

Oguzhan Topsakal ^{1,*}  and Swetha Lenkala ²¹ Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA² Computer Science Department, Florida Polytechnic University, Lakeland, FL 33805, USA

* Correspondence: topsakal@usf.edu

Abstract: Alzheimer's Disease, a progressive brain disorder that impairs memory, thinking, and behavior, has started to benefit from advancements in deep learning. However, the application of deep learning in medicine faces the challenge of limited data resources for training models. Transfer learning offers a solution by leveraging pre-trained models from similar tasks, reducing the data and computational requirements to achieve high performance. Additionally, data augmentation techniques, such as rotation and scaling, help increase the dataset size. In this study, we worked with magnetic resonance imaging (MRI) datasets and applied various pre-processing and augmentation techniques including intensity normalization, affine registration, skull stripping, entropy-based slicing, flipping, zooming, shifting, and rotating to clean and expand the dataset. We applied transfer learning to high-performing pre-trained models—ResNet-50, DenseNet-201, Xception, EfficientNetB0, and Inception V3, originally trained on ImageNet. We fine-tuned these models using the feature extraction technique on augmented data. Furthermore, we implemented ensemble learning techniques, such as stacking and boosting, to enhance the final prediction performance. The novel methodology we applied achieved high precision (95%), recall (94%), F1 score (95%), and accuracy (95%) for Alzheimer's disease detection. Overall, this study establishes a robust framework for applying machine learning to diagnose Alzheimer's using MRI scans. The combination of transfer learning, via pre-trained neural networks fine-tuned on a processed and augmented dataset, with ensemble learning, has proven highly effective, marking a significant advancement in medical diagnostics.



Citation: Topsakal, O.; Lenkala, S. Enhancing Alzheimer's Disease Detection through Ensemble Learning of Fine-Tuned Pre-Trained Neural Networks. *Electronics* **2024**, *13*, 3452. <https://doi.org/10.3390/electronics13173452>

Academic Editors: Gemma Piella, Luca Mesin, Chiara Zucco and Barbara Calabrese

Received: 4 July 2024

Revised: 22 August 2024

Accepted: 29 August 2024

Published: 30 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: Alzheimer's; pre-trained; fine tune; ensemble learning; neural network; deep learning; stacking; boosting; diagnosis; MRI; disease; ADNI; ResNet-50; DenseNet-201; Xception; EfficientNetB0; Inception V3

1. Introduction

Alzheimer's disease (AD) is a chronic neurological condition that begins with memory loss and progressively leads to death. As one of the most severe diseases, the World Health Organization (WHO) estimates that the number of AD cases will reach 82 million in the next decade [1]. The disease primarily affects individuals over the age of 65. While its progression varies among individuals, the expected lifespan after diagnosis ranges from three to nine years, depending on a person's health. Early diagnosis of AD is crucial due to the various abnormalities, such as memory loss, mood swings, metabolic decline, and behavioral issues, that arise during its progression [2].

Even though AD is irreversible, early detection offers several potential benefits. Patients diagnosed with AD can receive treatments to address cognitive losses, which are more effective when administered in the early stages. Accurate diagnosis can be difficult, as AD symptoms can resemble those of normal aging. A thorough examination of brain cells is necessary for a precise diagnosis [3]. Early diagnosis can help extend a person's lifespan despite the disease's irreversibility. Several non-invasive techniques, including

magnetic resonance imaging (MRI), positron emission tomography (PET), functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), and functional near-infrared spectroscopy (fNIRS), can be used for diagnosing AD [4]. However, accurate diagnosis can be challenging due to the need for expert analysis of complex medical data, such as brain scans and cognitive assessments. Therefore, there is demand for algorithms that can accurately diagnose AD [5].

Numerous studies have been conducted in recent years on AD detection using artificial intelligence [6]. Established machine learning classification methods have been applied to various brain scans, such as MRI. However, this remains a challenging problem due to the complexity of medical images, as a successful classification approach must be able to identify specific features among similar brain image patterns. Deep learning, particularly for classification tasks, has emerged as an ideal solution, addressing these concerns and yielding impressive results. As a result, deep learning can efficiently classify brain images at different stages of the disease [7].

Deep learning methods like convolutional neural networks (CNNs) have shown superior performance over traditional machine learning techniques. However, training CNNs from scratch requires large datasets and significant computational power. Transfer learning addresses this by using pre-trained networks, which need smaller datasets and less computational power [8]. While transfer learning enhances model accuracy, it can lead to overfitting on smaller datasets [9]. To mitigate this, data augmentation increases dataset size by generating new samples from existing data, and intensity normalization improves image quality by standardizing image intensities.

Ensemble learning enhances machine learning performance by combining multiple models to improve accuracy and robustness [10]. By integrating diverse predictive insights, ensemble methods reduce individual model errors and minimize overfitting. This approach leverages the strengths of various models, leading to superior performance in complex tasks such as medical diagnostics.

In this study, we utilized transfer learning techniques by fine-tuning high-performing pre-trained models such as ResNet-50, DenseNet-201, Xception, EfficientNetB0, and Inception V3 with the pre-processed and augmented datasets. We applied ensemble learning techniques such as stacking and boosting to increase the performance of final predictions. To the best of our knowledge, no study utilizes data augmentation, transfer learning, and ensemble learning in one solution to provide better performance for Alzheimer's diagnosis. This study aimed to evaluate the effectiveness of transfer and ensemble learning techniques combined with data augmentation techniques for diagnosing Alzheimer's disease by applying deep learning utilizing brain MRI scan datasets.

2. Background and Related Research

Alzheimer's disease (AD) is a neurological disorder characterized by issues with thinking, memory, and behavior. The primary features of Alzheimer's include abnormal clumps (amyloid plaques), loss of neuronal connections, and tangled bundles of fibers (neurofibrillary or tau tangles). The hippocampus is often the first area affected, making it a valuable biomarker in Alzheimer's patients. As the disease progresses and more neurons die, other brain regions become affected. In the final stage, significant damage occurs, and the brain tissue extensively shrinks. Unfortunately, there is no cure for Alzheimer's, as it is an irreversible neurodegenerative dementia. To improve prognosis, it is essential to develop reliable automatic detection systems to identify and diagnose the disease at its earliest stage [9].

Biomarkers are medical indicators that can be objectively measured from outside the patient and accurately evaluated in a reproducible manner. For Alzheimer's disease (AD), various biomarkers are available, including magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), tau positron emission tomography (tau-PET), electroencephalography (EEG), speech transcripts, genetic measures, and cerebrospinal fluid (CSF) measures. Among these, MRI is particularly valuable for diagnosing AD. MRI is a

non-invasive imaging technique that can detect structural brain changes associated with Alzheimer's disease [2]. It can identify shrinkage in the hippocampus, a key characteristic of AD, and changes in white matter tracts that connect different brain regions. Diffusion tensor imaging (DTI), a type of MRI, measures the integrity of white matter tracts and can detect early signs of damage in AD. MRI provides useful biomarkers even before clinical signs or significant brain damage become evident.

The most extensively used datasets for AD diagnosis include the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [11]; the Open Access Series of Imaging Studies (OASIS) dataset [12]; the Dementia Bank [1]; the Harvard Aging Brain Study (HABS) dataset [3]; the Mayo Clinic Study of Aging (MCSA) dataset [13]; the Australian Imaging, Biomarkers, and Lifestyle (AIBL) dataset [14]; and the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset [15].

We use ADNI datasets for AD diagnosis. ADNI, a longitudinal multicenter study initiated in 2003 by Dr. Michael W. Weiner, aims to develop biomarkers for early Alzheimer's disease (AD) diagnosis using clinical, genetic, imaging, and biospecimen data [11]. The study has four phases: ADNI-1 followed 800 participants with MRI and FDG-PET biomarkers for 2–3 years; ADNI-GO included the existing cohort and 200 additional early mild cognitive impairment (EMCI) subjects; ADNI-2 expanded the sample size, including 107 significant memory concern (SMC) subjects and amyloid PET images; ADNI-3, running from 2016 to 2022, which added tau-PET as a key indicator for AD. Each subject had between three and fifteen scans. The dataset is extensively used in research, with 92% of studies utilizing it either alone or in combination with other datasets [9].

2.1. Related Work on Utilizing Transfer Learning for Alzheimer's Disease

Ghazal et al. developed the ADDTLA model for classifying Alzheimer's disease into stages, non-demented (ND), very mild demented (VMD), mild demented (MD), and moderate demented (MOD), using transfer learning with a customized AlexNet model. The system includes pre-processing MRI images and using AlexNet's layers to achieve 91.7% validation accuracy and 98% training accuracy with SGDM optimization [7].

Acharya et al. used pre-trained CNNs (VGG-16, ResNet, and AlexNet) to classify Alzheimer's into four stages on a dataset of 6400 MRI images. A modified AlexNet with two convolutional layers and the Adam optimizer achieved the highest accuracy of 95.07% [9].

Ashraf et al. evaluated 13 pre-trained CNN models, including AlexNet, VGG, and ResNet, using the ADNI dataset. DenseNet achieved the highest accuracy (99.05%) on augmented data. They also proposed a ResNet-18 model for 2D and 3D MRI scans, achieving up to 96.88% accuracy with transfer learning [16].

Hon et al. applied transfer learning with VGG16 and Inception on the OASIS dataset. Their Inception model achieved the highest accuracy of 96.13%, demonstrating effective classification with limited training data using image entropy selection [17].

2.2. Related Work That Utilized Both Transfer Learning and Ensemble Learning

During our literature review, we identified studies utilizing transfer learning and ensemble learning for Alzheimer's disease diagnosis. In this section, we compare our study with three notable studies found in the literature.

Tanveer et al. proposed a deep learning architecture using transfer learning combined with random search hyperparameter tuning and snapshot ensemble strategies [18]. They used two pre-processing pipelines for large and small MRI datasets and applied random search to select the best M models from N trials, minimizing generalization error through cross-validation. The models, with diverse hyperparameter sets, were ensembled to boost performance using two strategies: 'Within dataset ensemble' (DTE-W) and 'Across dataset ensemble' (DTE-AC). DTE-W created an ensemble of M VGG16 models with random hyperparameters for each dataset, while DTE-AC ensembled the best M models from each dataset for final classification.

Yang et al. introduced METL, a framework for multi-source transfer learning with two phases: single-source tri-transfer learning and mutual information-based multi-source ensemble learning [19]. In the first phase, a single source domain is combined with the target domain to create a new training dataset, training three heterogeneous classifiers (Softmax, SVM, and DNN) iteratively to enhance transferability. In the second phase, classifiers from multiple source domains are ensembled using weight assignments based on domain correlations, producing a robust final classifier for the target domain.

Colbaugh et al. proposed a machine learning approach to diagnose Alzheimer’s disease using MicroRNA data, employing an ensemble transfer learning (ETL) algorithm [20]. The ETL algorithm includes three steps: feature extraction, ensemble-based prediction, and prediction refinement via unsupervised clustering. Feature extraction uses a stacked autoencoder to transform data representations. The ensemble-based prediction trains decision tree classifiers on labeled source data, generating initial diagnoses for target patients. The prediction refinement step uses unsupervised clustering to enhance initial predictions, integrating unsupervised feature learning, supervised ensemble learning, and unsupervised clustering.

Table 1 compares our proposed methodology with the approaches of Tanveer et al., Yang et al., and Colbaugh et al. Unlike Tanveer et al., who use only MRI images, we incorporate both MRI and PET images, apply intensity normalization and data augmentation for pre-processing, and utilize various pre-trained architectures in transfer learning, achieving diversity through stacking different architectures. Compared to Yang et al., who use only MRI without pre-trained models or pre-processing techniques, our method also uses PET images and combines predictions using stacking. In contrast to Colbaugh et al., who utilize only MicroRNA data and a stacked autoencoder for feature extraction, our method employs MRI and PET images, fine-tunes pre-trained architectures, and achieves diversity with an ensemble of different architectures.

Table 1. Comparison of this study with the studies by Tanveer et al., Yang et al. and Colbaugh et al.

	Tanveer et al. [17]	Yang et al. [18]	Colbaugh et al. [19]	This Study
Dataset	MRI	MRI	MicroRNA	MRI
Pre-processing techniques	Image segmentation using DARTEL, registration of images into MNI space	No pre-processing	No pre-processing technique used	Intensity normalization, data augmentation
Transfer learning	Only VGG-16 pre-trained architecture	No pre-trained model used. They claim to transfer learning from SoftMax, SVM, and DNN models.	SAE (stacked autoencoder); the authors call this transfer learning; however, this is not the transfer learning methodology we are using.	Pre-trained architectures: ResNet-50/101/152, DenseNet-201, Xception, EfficientNet-B6/B7, Inception-ResNet-v2
Ensemble learning	Diversity is achieved by utilizing different datasets and randomly assigned hyperparameters	Diversity is achieved by using different (not pre-trained) models. To combine the predictions, they use weight assignment based on the correlation between the source domains and the target domain.	Diversity is achieved by ensembling of 500 decision trees on the dataset	Diversity will be achieved utilizing different pre-trained architectures and combining the predictions using stacking methodology.

3. Methodology

Our methodology involves several steps, depicted in Figure 1. First, we pre-process the MRI dataset using various techniques, detailed below, to ensure a robust and clean dataset. The processed dataset is then used with pre-trained neural networks to perform a hyperparameter search, identifying the best-performing hyperparameters. These hyperparameters are applied in the fine-tuning phase, along with an augmented dataset. The results from the fine-tuned pre-trained datasets are then combined using stacking and boosting techniques to obtain the final results. The details of each step are explained in the following sections. The code for this methodology has been shared on GitHub [21].

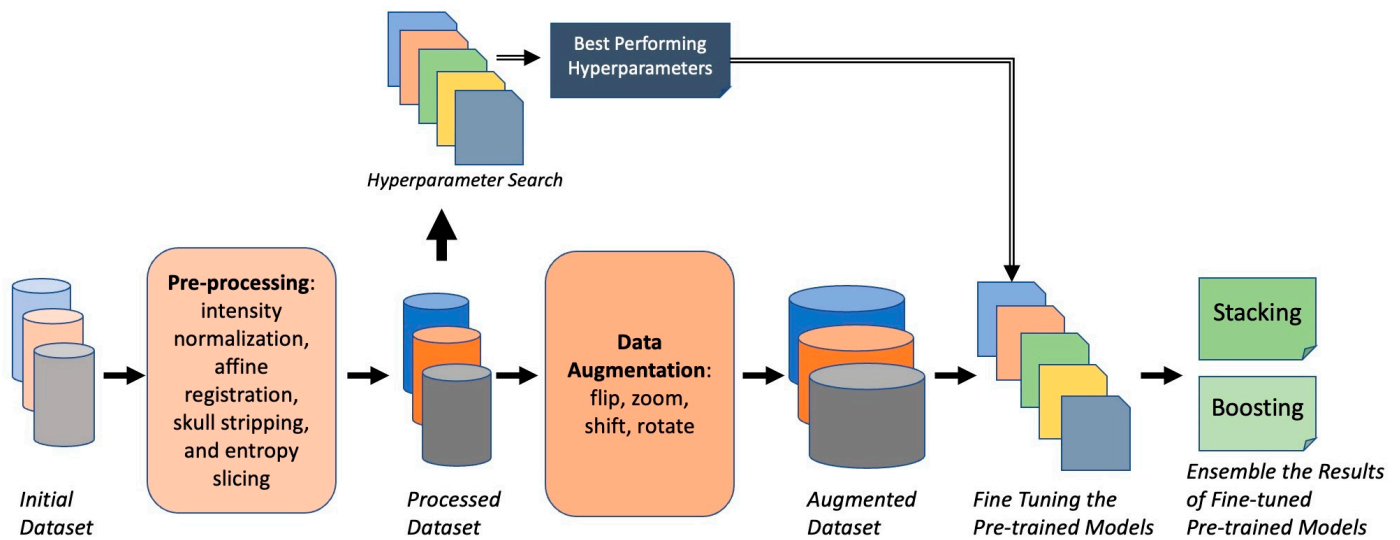


Figure 1. Workflow for enhancing performance by ensembling fine-tuned pre-trained models for Alzheimer's disease detection.

3.1. Initial Dataset

We sourced our primary dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [11]. The ADNI dataset is instrumental to our research, providing a comprehensive suite of MRI images essential for the diagnosis and analysis of Alzheimer's disease. It includes 432 images of Alzheimer's disease (AD), 663 images of cognitively normal (CN), and 1293 images of mild cognitive impairment (MCI).

3.2. Pre-Processing and Data Augmentation Techniques Used

Key pre-processing techniques in this pipeline include intensity normalization, affine registration, skull stripping, entropy-based slicing, and data augmentation (flipping, zooming, shifting, and rotating).

Intensity normalization is implemented using MONAI-1.3.0 (Medical Open Network for AI) [22], an open-source framework for medical image analysis. MONAI standardizes voxel intensities across MRI images, minimizing contrast variations and enhancing scan comparability. This step ensures that the neural network models trained on the dataset are robust to intensity variations, improving their generalization performance.

Affine registration is performed using the FSL-6.0.6 (FMRIB Software Library) FLIRT (FMRIB's Linear Image Registration Tool) module [23]. This step aligns MRI images with a standardized reference template, enabling consistent spatial mapping across scans. FLIRT employs a twelve-degree-of-freedom (dof) transformation model to optimize registration, correcting for differences in position, orientation, and scale between images.

Skull stripping is executed using the BET (Brain Extraction Tool) module from FSL [23]. This step removes non-brain tissues such as the skull and scalp, leaving only the brain

tissue for analysis. The BET module accurately delineates the brain region, minimizing the presence of extraneous structures and artifacts in the MRI images.

Entropy-based slicing extracts relevant features from the pre-processed MRI data. This technique partitions the brain volume into smaller regions based on voxel intensity distribution, identifying areas of high entropy associated with structural abnormalities or tissue degeneration. The algorithm computes voxel intensity entropy across multiple slices, selecting regions with the highest entropy values for further analysis. We extracted the 16 most informative slices and arranged them in a 4×4 grid, as shown in Figure 2.

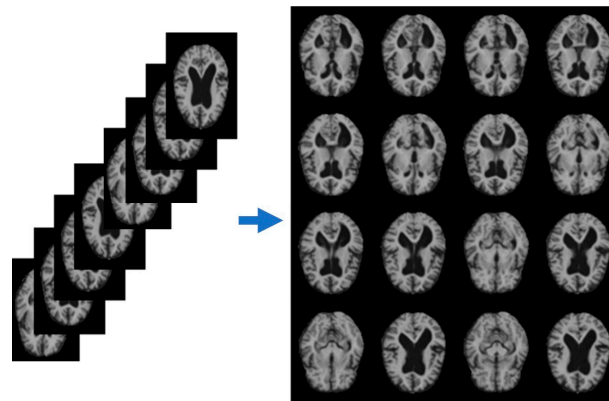


Figure 2. Pre-processed image after applying entropy slicing.

The pre-processing techniques—affine registration, skull stripping, and slicing—are orchestrated using Python scripts and the Nipype-1.8.6 package, providing a flexible and efficient framework for neuroimaging data processing [24].

After pre-processing, the dataset was split into training, validation, and testing sets, with 80% allocated for training and validation, and the remaining 20% reserved for testing. This division ensured balanced class representation, resulting in 1054 MCI, 515 AD, and 332 CN images for training and validation, and 91 AD, 148 CN, and 239 MCI images for testing. The ‘train_test_split’ function is used prior to cross-validation to separate out a test set, dividing the original data into 80% for training and 20% for testing. During cross-validation, these training data (80% of the total dataset) are further divided into 5 folds. In each fold, 16% of the total dataset is used for validation, and the remaining 64% for training.

Using the pre-processed dataset, we performed hyperparameter tuning to identify the best-performing settings. A grid search technique was applied to optimize hyperparameters such as epoch number, batch size, learning rate, dropout rate, and optimizer. The optimal hyperparameters were recorded and used in subsequent steps.

Data augmentation techniques were applied to increase the dataset size and variety, utilizing the Keras framework from TensorFlow-2.7.0 [25]. This involved geometric transformations such as flipping, rotating, shifting, and zooming, as shown in Figure 3. These transformations diversified the training dataset, improving the generalization capabilities of the machine learning models, reducing the risk of overfitting, and enhancing the models’ ability to learn meaningful features from the MRI data. The total number of images increased to 11,509, with 3198 CN, 2126 AD, and 5735 MCI images. The pre-processed and augmented images were stored in PNG format.

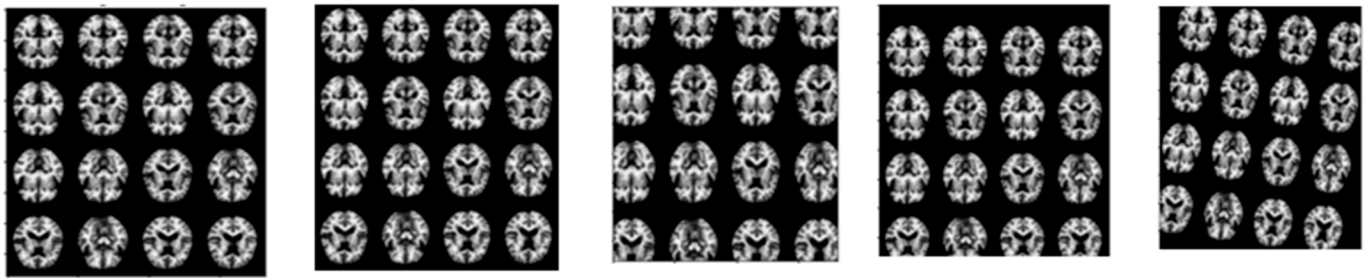


Figure 3. Transformation of images after augmentation, from left to right: original image, flipped, zoomed, shifted, and rotated.

3.3. Pre-Trained Neural Networks Used

The pre-trained models selected for our study are ResNet-50, DenseNet-201, Xception, EfficientNet2L, and Inception V3 [26–29]. These models were chosen based on their proven high performance and accuracy on similar tasks for which they were initially trained. This selection criterion ensures that the models have a robust foundation of knowledge, acquired from their prior training on extensive and diverse datasets.

The rationale behind incorporating multiple pre-trained models lies in their unique architectures and methodologies. Each model offers distinct approaches to feature extraction and representation, enabling a comprehensive exploration of the image feature space. By utilizing a variety of models, we aimed to maintain heterogeneity in the training process, effectively capturing the multifaceted nature of MRI images and enhancing the overall robustness of the classification system.

Incorporating multiple pre-trained models—ResNet-50, DenseNet-201, Xception, EfficientNet2L, and Inception V3—serves as a foundational strategy for employing ensemble learning techniques. Ensemble learning capitalizes on the diversity among the models, leveraging their unique strengths and mitigating individual weaknesses through a collaborative decision-making process. The architectural variability and feature extraction capabilities of these models facilitate a more nuanced and comprehensive interpretation of MRI images than any single model could achieve alone.

The diversity in model selection aids in capturing different aspects of image features. For instance, ResNet-50 excels in leveraging residual connections to mitigate the vanishing gradient problem, while EfficientNet dynamically scales model depth, width, and resolution for superior efficiency. DenseNet-201 enhances feature propagation with densely connected layers, and Inception V3 and Xception use innovative designs such as inception modules and depthwise separable convolutions, respectively, to optimize feature extraction and enhance performance. The models selected for our study are listed in Table 2, which also presents the datasets they were trained on, their initial training accuracy, input image size, and their parameters.

Table 2. Specifications about the pre-trained models selected for the study and their resources.

Model	Resource	Dataset	Accuracy for the Dataset	Input Image Size	Parameters (Millions)
ResNet-50	Keras	ImageNet	74.9%	224 × 224	25.6 M
DenseNet-201	Keras	CIFAR-10	77.3%	224 × 224	20.2 M
Xception	Keras	ImageNet	79.0%	299 × 299	22.9 M
EfficientNetB0	Keras	COCO-17	85.7%	224 × 224	5.3 M
Inception V3	Keras	ImageNet	77.9%	299 × 299	23.9 M

3.4. Fine Tuning

We utilized the feature extraction [30] technique of fine-tuning by replacing the classification layer. This approach customizes the model for a new task without disturbing the

learned features of the convolutional base. By retaining the convolutional layers—trained on a large and diverse dataset like ImageNet—and only replacing the top, task-specific layers, this method ensures that the model captures universal features like textures, shapes, and edges. This technique is particularly useful when the new dataset is relatively small and not diverse enough to train a complex model from scratch without overfitting.

To implement this technique, we loaded the pre-trained model without its original classification layers, which are tailored for 1000 ImageNet classes. All pre-existing layers were set as non-trainable (frozen) to preserve their learned weights. New fully connected layers were then added to create a new top suited to the target task, such as a different number of output classes. This top typically consists of one or more dense layers and a final softmax activation that outputs probabilities across the new classes. The model was then compiled with an appropriate optimizer, loss function, and metrics, and trained on the new dataset. This setup limits training to the newly added layers, making it computationally efficient while leveraging robust pre-learned features.

We also experimented with fine-tuning by unfreezing the last three layers before the classification layer [31], allowing for these layers to adjust their weights and fine-tune the abstract representations captured by the model to better suit the new task. This method requires more data and is beneficial when the new data have similarities to the original training data but also possess unique characteristics. However, our results were not as favorable as those obtained through feature extraction, likely due to insufficient data for effectively fine-tuning the unfrozen layers. Consequently, we chose not to pursue this technique further due to its lower performance. In summary, replacing the classification layer proved to be the most effective method for our dataset, balancing computational efficiency with robust feature extraction.

3.5. Ensemble Techniques

After training five transfer learning models, namely, ResNet-50, InceptionV3, EfficientNetB0, DenseNet201, and Xception, on augmented images with tuned hyperparameters, we employed an ensemble learning approach using both stacking and boosting techniques to enhance the final predictions.

Boosting, specifically gradient boosting, was utilized to aggregate predictions from the individual models [32]. Gradient boosting iteratively builds an ensemble of weak learners—in this case, the five pre-trained models—and adjusts subsequent models to correct errors made by previous ones, as shown in Figure 4. By combining the predictions of multiple models through boosting, the ensemble can capture diverse patterns in the data, potentially improving accuracy beyond what individual models can achieve alone. Gradient boosting, being an iterative technique, learns from the errors of previous model predictions to progressively refine subsequent predictions [33]. For each image, the gradient boosting model sequentially updates its predictions. The ensemble's prediction for class c after k iterations is calculated as shown in Equation (1):

$$F_{k,c}(x) = F_{0,c}(x) + \eta \sum_{i=1}^k h_{i,c}(x) \quad (1)$$

where $F_{0,c}(x)$ is the initial prediction for class c , $h_{i,c}(x)$ is the contribution of the i th tree for class c , η is the learning rate that controls the contribution of each tree, and x is the input data, which in this context is an image. The output after the final iteration K is then passed through a SoftMax function for classification to ensure valid probability distribution as formulated in Equation (2):

$$P(c|x) = \frac{\exp(F_{k,c}(x))}{\sum_j \exp(F_{k,c}(x))} \quad (2)$$

where the sum in the denominator is over all classes j , normalizing the raw scores into probabilities. The final predicted class for the image is then chosen based on the highest probability after the final iteration as given in Equation (3):

$$\hat{c} = \operatorname{argmax}_c P(c|x) \quad (3)$$

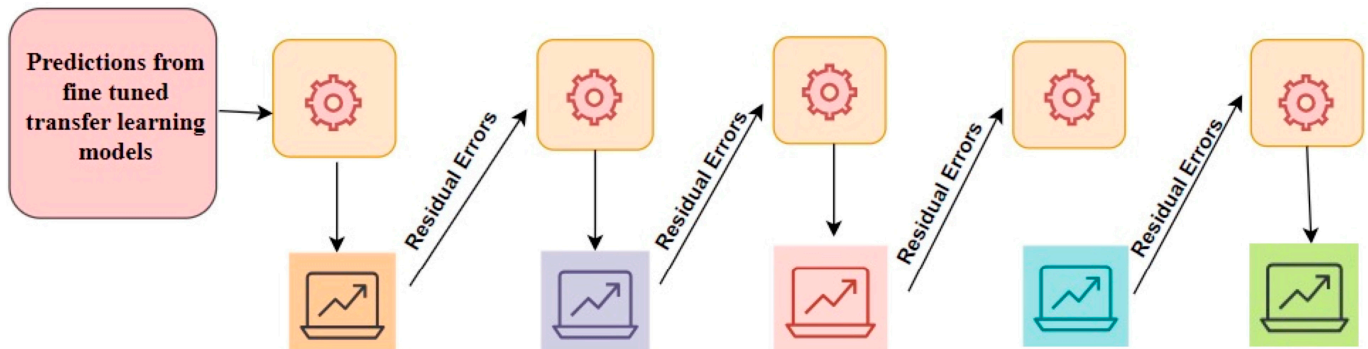


Figure 4. Illustration of boosting classifier using gradient boosting classifier.

On the other hand, stacking was employed using a random forest algorithm [34]. Stacking involves training a meta-model, as shown in Figure 5, often a simple classifier or regressor, on the predictions of base models. In this scenario, random forest was chosen as the meta-model due to its robustness and ability to handle high-dimensional data [35]. By training random forest on the predictions of the individual models, the ensemble effectively leverages the strengths of each base model while mitigating their weaknesses, leading to more accurate predictions. For an image with true class c , the random forest ensemble makes a prediction based on the averaged probabilities across all decision trees as shown in Equation (4):

$$P(c|image) = \frac{1}{N} \sum_{i=1}^N P_i(c|image) \quad (4)$$

where N is the number of trees and P_i is the predicted probability of each class c by the i th tree. The predicted class is calculated as shown in Equation (5):

$$\hat{c} = \operatorname{argmax}_c P(c|image) \quad (5)$$

where argmax selects the class with the highest mean probability. With stacking, the final predictions were obtained by utilizing the trained meta-models on the test data.

Overall, this ensemble learning approach harnesses the strengths of both boosting and stacking to create a more robust and accurate model. Boosting improves accuracy by iteratively correcting errors, while stacking combines diverse predictions from multiple models to further enhance performance. By leveraging these ensemble techniques, the final prediction model can better generalize to unseen data and achieve higher accuracy levels than any individual model could achieve on its own.

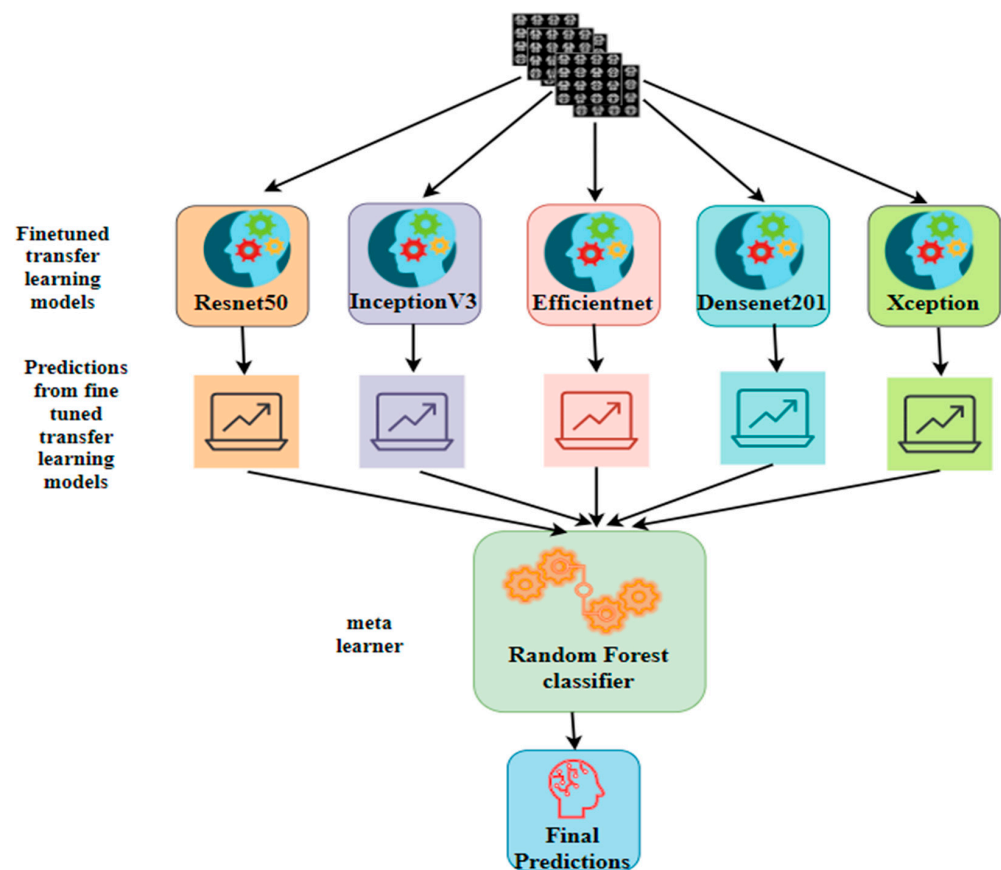


Figure 5. Illustration of stacking ensemble learning.

4. Results

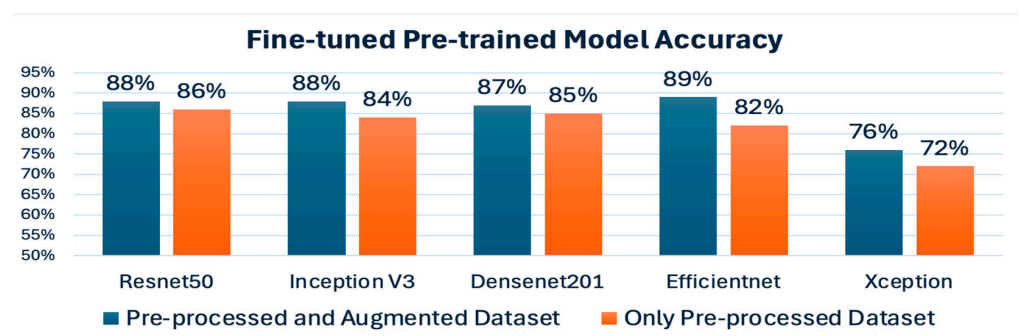
In our study, we utilized HiperGator’s computational resources to perform our experiments, leveraging a configuration that included 1 GPU, 1 CPU, and 7 GB of memory. The GPU (graphics processing unit) is particularly beneficial for parallelizable computations, offering significant acceleration for deep learning tasks.

We fine-tuned the pre-trained models (ResNet50, Inception V3, DenseNet201, EfficientNet, and Xception) using the feature extraction technique with the pre-processed and augmented dataset. The hyperparameters that yielded the best performance in the fivefold grid search were utilized during the fine-tuning process. These hyperparameters along with the accuracy achieved for each model after fine-tuning are listed in Table 3. Across all models, a batch size of 4 and a dropout rate of 0.4 were consistent. ResNet50 and Inception V3 both achieved an accuracy of 88%, with ResNet50 using the Adam optimizer and Inception V3 using RMSprop, both at a learning rate of 0.0001. EfficientNet achieved the highest accuracy at 89%, while Xception recorded the lowest at 76%. The DenseNet201 model had an accuracy of 87%. The accuracy metrics listed in Table 3 were achieved using the validation dataset. This was accomplished by performing a grid search with fivefold cross-validation during the fine-tuning of the hyperparameters.

Figure 6 shows that the performance of models fine-tuned with the augmented dataset improved compared to those using only the pre-processed dataset. The results in Figure 6 demonstrate the effectiveness of data augmentation in increasing dataset size and variability. The accuracy values listed in Figure 6 were achieved using grid search on the validation dataset.

Table 3. Hyperparameters used for fine-tuning pre-trained models with pre-processed and augmented dataset.

Model	Epochs	Batch Size	Learning Rate	Loss Function	Dropout Rate	Optimizer	Accuracy
Resnet50	50	4	0.0001	Categorical cross-entropy	0.4	Adam	88%
Inception V3	50	4	0.0001	Categorical cross-entropy	0.4	RMSProp	88%
Densenet201	50	4	0.0001	Categorical cross-entropy	0.4	Adam	87%
EfficientNet	50	4	0.00001	Categorical cross-entropy	0.4	Adam	89%
Xception	50	4	0.000001	Categorical cross-entropy	0.4	RMSProp	76%

**Figure 6.** Comparison of accuracy of fine-tuned models when augmentation is utilized.

The implementation of ensemble learning techniques, namely, stacking and boosting, on the combined predictions of five augmented transfer learning models further enhanced performance, as listed in Table 4. The accuracy, precision, recall, and F1 score metrics listed in Table 4 were achieved using the test dataset. The boosting ensemble, which utilized gradient boosting—an algorithm that iteratively corrects previous prediction errors—demonstrated superior performance, achieving an overall accuracy of 95%. This model excels in handling high-dimensional data and naturally reduces overfitting, as evidenced by the high precision and recall scores for diagnosing Alzheimer’s disease (AD), cognitive normal (CN), and mild cognitive impairment (MCI). Stacking, while slightly less accurate overall with 92% accuracy, still showed robust performance, as detailed in Table 4. The meta-model employed in the stacking ensemble, random forest, was notably effective, as illustrated in Table 4. The F1 scores for all diagnoses under this model are impressive and indicate balanced performance. Table 5 presents the confusion matrix for the boosting and stacking approaches, which form the basis for the metrics listed in Table 4.

Table 4. Metrics of boosting and stacking ensemble learning models.

Ensemble Learning	Diagnosis	Precision	Recall	F1 Score	Accuracy
Boosting	AD	1.00	0.89	0.94	0.88
	CN	0.93	0.96	0.94	0.96
	MCI	0.94	0.98	0.96	0.97
Stacking	AD	0.92	0.85	0.88	0.85
	CN	0.89	0.92	0.91	0.96
	MCI	0.91	0.93	0.92	0.91

Table 5. Confusion matrix for the boosting and stacking ensemble learning models.

	Boosting				Stacking		
	AD	CN	MCI		AD	CN	MCI
AD	24	1	2	AD	23	1	3
CN	0	25	1	CN	1	24	1
MCI	0	1	45	MCI	1	2	43

5. Discussion

The results of the ensemble approach underscore the substantial improvement in accuracy and generalization achieved by both ensemble methods compared to the initial performance of individual transfer learning models. The strength of stacking lies in its collaborative approach, aggregating model predictions to capture diverse patterns within the data more effectively. This is evidenced by the model's precision, which reached 1.00 for Alzheimer's disease (AD), indicating an exceptional ability to minimize false positives.

Boosting's iterative refinement strategy is notable for progressively enhancing performance, which is particularly important in the dynamic context of medical diagnostics. Although it did not surpass stacking in overall accuracy, the gradient boosting model's ability to correct errors from individual predictions has proven valuable, as indicated by the high recall rates across diagnoses.

In clinical practice, where diagnostic precision is paramount, the F1 score emerges as a critical metric, encapsulating both precision and recall to provide a comprehensive measure of accuracy and consistency [31]. For instance, the boosting model's application to AD diagnosis achieved a precision score of 1.00, signifying impeccable accuracy—every instance labeled as AD was correct, eliminating false positives. Concurrently, the model exhibited a recall of 89% for AD, reflecting its capacity to correctly identify the majority of actual AD cases presented in the data.

The F1 score, derived as the harmonic mean of precision and recall, serves as a single metric that encapsulates the model's overall diagnostic accuracy. A high F1 score, such as the boosting model's 0.94 for AD, conveys a robust equilibrium between capturing genuine AD cases (high recall) and avoiding the misclassification of other conditions as AD (high precision). This balance is crucial in the medical domain, where the model's precision of 0.93 in identifying cognitively normal (CN) cases ensures that 93% of healthy individuals are correctly dismissed as not having AD, sparing them unnecessary medical procedures and associated psychological burdens.

Additionally, the model's high recall in recognizing cases of mild cognitive impairment (MCI), with a recall of 0.98, underscores its effectiveness in detecting nearly all instances of this condition, which is essential for early intervention and effective disease management. Moreover, the F1 score of 0.96 for MCI demonstrates the model's reliability and consistency across a diverse patient population, affirming its utility in various clinical settings and its potential as a standard tool for diagnosing varying degrees of cognitive impairment.

When comparing the ensemble learning results, as shown in Figure 7, stacking exhibits superior accuracy and a balanced precision–recall profile. This enhancement is primarily due to the stacking model's integration of the strengths of multiple transfer learning models, leveraging their collective insights.

By aggregating predictions from advanced transfer learning models such as ResNet50, Inception V3, EfficientNetB0, DenseNet201, and Xception, the stacking approach via the random forest meta-model offers a comprehensive analysis that captures diverse features extracted by these individual models. Random forest's ability to manage the ensemble of predictions and its inherent resistance to overfitting effectively captures and utilizes the intricate patterns of MRI images crucial for accurate Alzheimer's stage identification.

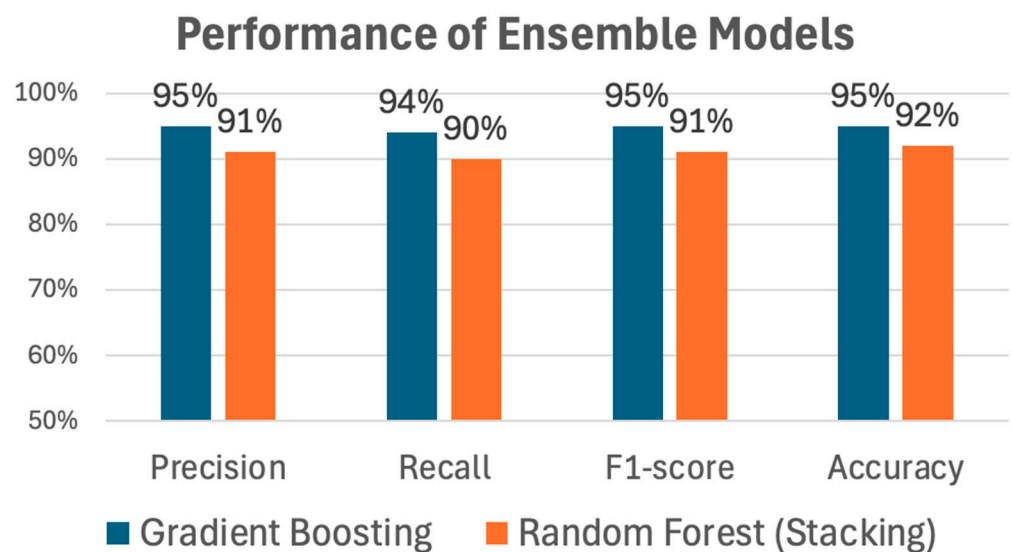


Figure 7. Comparison of metrics for ensemble learning models.

The random forest’s efficacy in the stacking model arises from its structure, which accommodates a wide array of decision-making processes through numerous trees. Each tree provides its perspective, rendering the final decision process more robust against errors that individual transfer learning models might make. This structure is particularly suited to high-dimensional data like MRI images.

In contrast, while gradient boosting is a robust method that iteratively improves upon its mistakes, it may not fully exploit the complex feature representations learned by different transfer learning models. Boosting’s sequential improvement strategy can sometimes overemphasize specific data points or features, potentially leading to overfitting.

Despite the strong performance of the gradient boosting model, evidenced by its high accuracy and F1 scores, its sequential nature might not aggregate the predictive capabilities of various transfer learning models as effectively as stacking. Consequently, the stacking approach with a random forest meta-model is better positioned to leverage the comprehensive, multi-dimensional insights provided by the transfer learning models, resulting in superior performance metrics for Alzheimer’s disease diagnosis through MRI analysis.

The integration of ensemble learning models, especially those harnessing the predictive power of transfer learning models, represents a significant advancement in Alzheimer’s disease classification and diagnosis. Transitioning from individual model predictions to a consolidated ensemble approach using stacking and boosting techniques markedly improved diagnostic capabilities, as evidenced by enhanced precision and recall metrics.

While our findings demonstrate the potential of deep learning in analyzing the ADNI MRI dataset, it is important to note that the dataset primarily comprises individuals from white backgrounds. This limitation restricts the generalizability of our results to the broader population with Alzheimer’s disease and mild cognitive impairment. Future studies should aim to include more ethnically diverse datasets to validate and potentially enhance the applicability of these findings across different populations.

To conduct the clinical validation of the proposed methodology and the resulting models, several steps will be necessary: Establishing partnerships with healthcare providers is essential for accessing real-world testing environments and diverse datasets. Navigating the regulatory framework, particularly through engagements with regulatory bodies like the FDA, and carrying out pilot studies are crucial to assess the effectiveness and safety of the models. Additionally, addressing challenges such as ensuring data privacy, achieving seamless integration of AI tools into existing healthcare systems, correcting potential biases in model performance, and securing adequate funding are all critical components. These

steps are integral to successfully transitioning the models from theoretical development to practical clinical application, ultimately enhancing patient care and treatment outcomes.

6. Conclusions

This study demonstrates a comprehensive approach to enhancing the classification of MRI images for diagnosing cognitive impairments, specifically normal cognitive (NC), Alzheimer's disease (AD), and mild cognitive impairment (MCI). By integrating advanced machine learning techniques, including pre-processing, fine-tuning of transfer learning models, and ensemble learning methods, this study achieves remarkable accuracy and reliability, crucial for clinical diagnostics.

Meticulous pre-processing steps such as intensity normalization, affine registration, skull stripping, and data augmentation significantly enhance the quality and consistency of MRI images. These foundational steps improve the neural network's ability to detect relevant features without distractions from irrelevant data variations. Data augmentation, in particular, has proven critical in enhancing model robustness against overfitting, reflected in the improved model performances.

The utilization of pre-trained models like ResNet-50, EfficientNet, DenseNet201, InceptionV3, and XceptionNet underscores the effectiveness of transfer learning in leveraging large, pre-existing datasets to extract complex features relevant to medical imaging tasks. Fine-tuning these models for the specific task ensures they adapt to the nuances of new medical data, evident from the accuracy improvements post-fine-tuning.

Ensemble learning techniques, such as stacking and boosting, further enhance diagnostic accuracy. Stacking, using a random forest meta-model, achieved particularly strong performance, illustrating the benefits of combining multiple predictive models to improve the final output. This technique significantly boosted model performance, achieving an accuracy of up to 95%, underscoring the value of a collaborative approach in predictive modeling.

The high precision, recall, and F1 scores achieved suggest that the models are not only capable of accurately identifying cognitive impairments but also reliable in distinguishing between different stages of cognitive diseases. Such high metric scores support the potential for these models to be implemented in clinical settings, where accurate and early diagnosis can significantly affect treatment outcomes.

The practical application of these ensemble learning models in clinical workflows could be transformative for Alzheimer's diagnostics. By facilitating earlier detection and providing richer diagnostic information, these models have the potential to significantly improve treatment planning. Moreover, their deployment could lead to a broader understanding of the disease through the identification of novel imaging biomarkers, contributing to the progression of medical research in Alzheimer's disease.

Clinically, the higher accuracy levels and the balanced precision and recall offered by the stacking model suggest its utility as an invaluable support tool for medical professionals. With the ability to deliver deep insights from imaging data, the stacking model is poised to provide a nuanced perspective on Alzheimer's at various stages, crucial for devising personalized and timely treatment strategies.

Overall, this study establishes a robust framework for applying machine learning to diagnose Alzheimer's and related cognitive conditions using MRI scans. The combination of transfer learning, via pre-trained neural networks fine-tuned on a processed and augmented dataset, with ensemble learning, proved highly effective, marking a significant advancement in medical diagnostics. This approach not only enhances diagnostic accuracy but also offers a scalable model adaptable to other medical imaging tasks, potentially transforming diagnostic procedures in healthcare settings.

Author Contributions: Conceptualization, O.T.; methodology, O.T.; software, S.L.; validation, S.L.; formal analysis, O.T. and S.L.; investigation, S.L. and O.T.; resources, O.T.; data curation, S.L.; writing—original draft preparation, S.L. and O.T.; visualization, O.T. and S.L.; supervision, O.T.; project administration, O.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Result data and the code can be downloaded from the GitHub page: <https://github.com/research-outcome/Alzheimers-Ensemble-Transfer-Learning> (accessed on 1 July 2024).

Acknowledgments: We thank Parisa Darbari for her feedback during this study.

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

References

- Agarwal, D.; Marques, G.; de la Torre-Díez, I.; Martin, M.A.F.; Zapirain, B.G.; Rodríguez, F.M. Transfer Learning for Alzheimer's Disease through Neuroimaging Biomarkers: A Systematic Review. *Sensors* **2021**, *21*, 7259. [CrossRef] [PubMed]
- Saleem, T.J.; Zahra, S.R.; Wu, F.; Alwakeel, A.; Alwakeel, M.; Jeribi, F.; Hijji, M. Deep learning-based diagnosis of Alzheimer's disease. *J. Pers. Med.* **2022**, *12*, 815. [CrossRef] [PubMed]
- Yang, K.; Mohammed, E.A. A review of artificial intelligence technologies for early prediction of Alzheimer's disease. *arXiv* **2020**, arXiv:2101.01781.
- Liampas, I.; Danga, F.; Kyriakouloupoulou, P.; Siokas, V.; Stamati, P.; Messinis, L.; Dardiotis, E.; Nasios, G. The Contribution of Functional Near-Infrared Spectroscopy (fNIRS) to the Study of Neurodegenerative Disorders: A Narrative Review. *Diagnostics* **2024**, *14*, 663. [CrossRef] [PubMed]
- Lenkala, S.; Marry, R.; Gopovaram, S.R.; Akinci, T.C.; Topsakal, O. Comparison of Automated Machine Learning (AutoML) Tools for Epileptic Seizure Detection Using Electroencephalograms (EEG). *Computers* **2023**, *12*, 197. [CrossRef]
- Ardalan, Z.; Subbian, V. Transfer Learning Approaches for Neuroimaging Analysis: A Scoping Review. *Front. Artif. Intell.* **2022**, *5*, 780405. [CrossRef]
- Ghazal, T.M.; Abbas, S.; Munir, S.; Ahmad, M.; Issa, G.F.; Zahra, S.B.; Khan, M.A.; Hasan, M.K. Alzheimer Disease Detection Empowered with Transfer Learning. *Comput. Mater. Contin.* **2022**, *70*, 5005–5019. [CrossRef]
- Topsakal, O.; Akinci, T.C. A review of transfer learning: Advantages, Strategies, and Types. In Proceedings of the International Conference on Modern and Advanced Research, Konya, Turkey, 29–31 July 2023.
- Acharya, H.; Mehta, R.; Singh, D.K. Alzheimer disease classification using transfer learning. In Proceedings of the 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021.
- Ganaie, M.; Hu, M.; Malik, A.; Tanveer, M.; Suganthan, P. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
- ADNI Alzheimer's Dataset. Available online: <https://adni.loni.usc.edu/> (accessed on 1 July 2024).
- Marcus, D.S.; Fotenos, A.F.; Csernansky, J.G.; Morris, J.C.; Buckner, R.L. Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. *J. Cogn. Neurosci.* **2010**, *22*, 2677–2684. [CrossRef]
- Roberts, R.O.; Geda, Y.E.; Knopman, D.S.; Cha, R.H.; Pankratz, V.S.; Boeve, B.F.; Ivnik, R.J.; Tangalos, E.G.; Petersen, R.C.; Rocca, W.A. The Mayo Clinic Study of Aging: Design and Sampling, Participation, Baseline Measures and Sample Characteristics. *Neuroepidemiology* **2008**, *30*, 58–69. [CrossRef]
- Fowler, C.; Rainey-Smith, S.R.; Bird, S.; Bomke, J.; Bourgeat, P.; Brown, B.M.; Burnham, S.C.; Bush, A.I.; Chadunow, C.; Collins, S.; et al. Fifteen Years of the Australian Imaging, Biomarkers and Lifestyle (AIBL) Study: Progress and Observations from 2,359 Older Adults Spanning the Spectrum from Cognitive Normality to Alzheimer's Disease. *J. Alzheimer's Dis. Rep.* **2021**, *5*, 443–468. [CrossRef] [PubMed] [PubMed Central]
- Malone, I.B.; Cash, D.; Ridgway, G.R.; MacManus, D.G.; Ourselin, S.; Fox, N.C.; Schott, J.M. MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage* **2013**, *70*, 33–36. [CrossRef] [PubMed]
- Ashraf, A.; Naz, S.; Shirazi, S.H.; Razzak, I.; Parsad, M. Deep transfer learning for Alzheimer neurological disorder detection. *Multimed. Tools Appl.* **2021**, *80*, 1–26. [CrossRef]
- Hon, M.; Khan, N.M. Towards Alzheimer's disease classification through transfer learning. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017.
- Tanveer, M.; Rashid, A.H.; Ganaie, M.A.; Reza, M.; Razzak, I.; Hua, K.-L. Classification of Alzheimer's Disease Using Ensemble of Deep Neural Networks Trained Through Transfer Learning. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 1453–1463. [CrossRef]
- Colbaugh, R.; Glass, K.; Gallegos, G. Ensemble transfer learning for Alzheimer's disease diagnosis. In Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Republic of Korea, 11–15 July 2017; pp. 3102–3105.
- Yang, Y.; Li, X.; Wang, P.; Xia, Y.; Ye, Q. Multi-Source Transfer Learning via Ensemble Approach for Initial Diagnosis of Alzheimer's Disease. *IEEE J. Transl. Eng. Health Med.* **2020**, *8*, 1–10. [CrossRef]
- Alzheimer's Ensemble Learning on Transfer Learning GitHub. Available online: <https://github.com/research-outcome/Alzheimers-Ensemble-Transfer-Learning> (accessed on 1 July 2024).
- MONAI Medical Open Network for AI. Available online: <https://github.com/Project-MONAI> (accessed on 1 July 2024).
- Woolrich, M.W.; Jbabdi, S.; Patenaude, B.; Chappell, M.; Makni, S.; Behrens, T.; Beckmann, C.F.; Jenkinson, M.; Smith, S.M. Bayesian analysis of neuroimaging data in FSL. *NeuroImage* **2009**, *45*, S173–S186. [CrossRef]
- Nipype. Available online: <https://nipype.org/packages/nipype/index.html> (accessed on 1 July 2024).

25. Manaswi, N.K.; Manaswi, N.K.; John, S. *Deep Learning with Applications Using Python. Chatbots and Face, Object, and Speech Recognition with TensorFlow and Keras*; Apress: Berkeley, CA, USA, 2018; pp. 31–43. [\[CrossRef\]](#)
26. Gupta, M.; Ghatak, S.; Gupta, A.; Mukherjee, A.L. Hyper parameters tuning resnet-50 for Alzheimer’s disease classification on neuroimaging data. In *Artificial Intelligence on Medical Data: Proceedings of International Symposium, ISCMM 2021*; Springer: Singapore, 2022.
27. Cui, Z.; Gao, Z.; Leng, J.; Zhang, T.; Quan, P.; Zhao, W. Alzheimer’s disease diagnosis using enhanced inception network based on brain magnetic resonance image. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 18–21 November 2019.
28. Jahan, S.; Saif Adib, M.R.; Mahmud, M.; Kaiser, M.S. Comparison between explainable AI algorithms for Alzheimer’s disease prediction using efficient net models. In *International Conference on Brain Informatics*; Springer: Cham, Switzerland, 2023; pp. 357–368.
29. Jismi, K.; Ananda Resmi, S.; Ramanujan, A. Deep Learning Enabled Classification of Cognitive Impairment Stages Using MRI Images. In *Proceedings of the 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Greater Noida, India, 28–30 April 2023.
30. Khan, R.; Akbar, S.; Mehmood, A.; Shahid, F.; Munir, K.; Ilyas, N.; Asif, M.; Zheng, Z. A transfer learning approach for multiclass classification of Alzheimer’s disease using MRI images. *Front. Neurosci.* **2023**, *16*, 1050777. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Maqsood, M.; Nazir, F.; Khan, U.; Aadil, F.; Jamal, H.; Mehmood, I.; Song, O.-Y. Transfer Learning Assisted Classification and Detection of Alzheimer’s Disease Stages Using 3D MRI Scans. *Sensors* **2019**, *19*, 2645. [\[CrossRef\]](#)
32. Syed, A.H.; Khan, T.; Hassan, A.; Alromema, N.A.; Binsawad, M.; Alsayed, A.O. An Ensemble-Learning Based Application to Predict the Earlier Stages of Alzheimer’s Disease (AD). *IEEE Access* **2020**, *8*, 222126–222143. [\[CrossRef\]](#)
33. Fulton, L.V.; Dolezel, D.; Harrop, J.; Yan, Y.; Fulton, C.P. Classification of Alzheimer’s Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50. *Brain Sci.* **2019**, *9*, 212. [\[CrossRef\]](#)
34. Khoei, T.T.; Labuhn, M.C.; Caleb, T.D.; Hu, W.C.; Kaabouch, N. A Stacking-based Ensemble Learning Model with Genetic Algorithm For detecting Early Stages of Alzheimer’s Disease. In *Proceedings of the IEEE International Conference on Electro Information Technology (EIT)*, Mt. Pleasant, MI, USA, 14–15 May 2021; pp. 215–222. [\[CrossRef\]](#)
35. Sarica, A.; Cerasa, A.; Quattrone, A. Random forest algorithm for the classification of neuroimaging data in Alzheimer’s disease: A systematic review. *Front. Aging Neurosci.* **2017**, *9*, 329. [\[CrossRef\]](#) [\[PubMed\]](#)

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.