Original article

Journal of High School Science

Comparing different Convolutional Neural Networks for the classification of Alzheimer's Disease

Eroltu K

Submitted: June 5, 2023, Revised: version 1, July 12, 2023, version 2, July 19, 2023, version 3, July 20, 2023
Accepted: July 20, 2023

Abstract

Alzheimer's disease (AD) is a degenerative, incurable neurological disorder that progressively damages cognitive abilities. AD affects millions of people worldwide. The biopsy method is the most reliable method to identify AD, but it has the chance to cause irreparable harm. There are many non-invasive alternatives to biopsies that can be used to diagnose AD without undue risk. One of these alternatives comprises computer-assisted diagnostic systems, which enable the identification of brain-impairing illnesses/diseases. This paper used Magnetic Resonance Imaging scans of brains with four different classes to create a model to detect AD. The open-source OASIS dataset, which serves as the basis for the study's data, was split into an 80% training set and a 20% test set. The dataset consisted of classes of moderate AD, mild AD, very mild AD, and non-AD scans. Five different Convolutional Neural Network methods were used for classification. The DenseNet-121, ResNet-50, ResNet-18, and AlexNet methods achieved detection accuracies of 90.5%, 95.1%, 88.4% and 70.5% respectively. The EfficientNetB-7 method failed to identify many cases of AD.

Keywords

Alzheimer's disease, CNN, ResNet-50, ResNet-18, DenseNet-121, MRI, AlexNet, EfficientNetB-7, Confusion Matrix, Neural network

Kaan Eroltu, United World College of the Adriatic, Frazione Duino, 29, 34011 Duino, Trieste, Italy, ad22kero@uwcad.it

Introduction

Alzheimer's disease (AD) is a severe degenerative brain ailment that affects memory and language abilities (1, 2). AD is one of the most commonly diagnosed forms of dementia; for which, there are no definitive solutions (3). AD ensues due to many different protein deposits in the brain, specifically beta-amyloid plaques and tau tangles (4). The build-up of these proteins triggers inflammatory responses in the brain, damaging the neuronal networks, which worsens the cognitive faculties, causes memory impairment, and deterioration of motor functions (5).

The approaches investigated to stop AD may not be able to reverse the process (6). However, they can be used to control its effects and also limit its symptoms, thus granting a better quality of life for those suffering from AD (6). It is widely accepted that early diagnosis can slow down the progression of AD (7). However, there is no evidence to suggest that such early diagnosis can permanently halt potential worsening of the condition (8). There are currently over 6 million people in the United States diagnosed with AD; this figure is expected to nearly triple by the year 2050 (9).

There are many ways to diagnose AD; some of the methods involve physical and neurological examinations, laboratory tests, and neuroimaging (10). Doctors utilize these approaches to attain a definitive diagnosis of the condition. For instance, positron emission tomography (PET) and magnetic resonance imaging (MRI) can identify the regions affected by AD and amyloid plaques (10). Other cognitive examinations, such as mini-mental state examination (MMSE) and Montreal cognitive assessment (MoCA), are used to assess a patient's memory and other intellectual abilities (11). MRI scans help to precisely capture soft tissue parameters (12). The use of neuroimaging tools in the field of engineering, particularly in the area of image processing, has become more common (13, 14). A new field of study has emerged around creating computer-aided diagnostic (CAD) systems based on these neuroimaging tools (15). With computer-aided diagnostic systems, doctors can get more accurate information that helps them make a complete and accurate diagnosis (12).

AD can result in the whole brain shrinking, pushing the ventricles out and causing them to appear larger than usual (16). Studies have demonstrated unusual gray matter in individuals suffering from AD. Compared to healthy participants, patients with AD had significantly lower overall gray matter volume, reduced total brain volume, and increased ventricles (17, 18). As the brain tissue shrinks and becomes atrophied, the spaces filled with cerebrospinal fluid, known as the ventricles, may appear enlarged or expanded. This is because the reduction in brain volume creates more space for the ventricles to occupy (17, 18). The CNN models focus on the thickness of the gray matter and the size of the ventricles in order to classify and detect the severity of AD.

Conventional methods and deep-learning architecture are commonly used for CAD (19). Conventional methods convert images into matrices from which attributes can be extracted (20). Feature extraction techniques serve two

primary purposes: finding the key differences between targets and reducing the data's dimensionality while preserving its essential characteristics (12). Examples include linear regression, logistic regression, k-nearest neighbor, and support vector machines. This research paper does not address conventional methods, instead, using deep learning architecture to evaluate and analyze MRI images. Early layers of a deep network can detect features, while later layers combine these components into more complex input attributes (21). This paper focuses more on deep-learning architectures, to which belong ResNet-50, ResNet-18, DenseNet-121, AlexNet, and EfficientNetB-7.

Literature Review

Much research has been done on detecting and classifying AD for more than ten years using a wide variety of approaches. Park et. al. proposed a deep learning-based model capable of predicting AD by utilizing large-scale gene expression and DNA methylation data (22). The purpose was to increase performance using an alternative feature selection method. Hence the study achieved an 82.3% validation performance (22). Jo et. al. used deep learning techniques that were applied to neuroimaging data without requiring pre-processing for feature selection and demonstrated an accuracy of up to 96.0% in the classification of AD (23). Allioui et. al. used the segmentation method to detect AD (24). The U-Net model was used for the segmentation, and this methodology achieved an accuracy rate of 92.7% (24). Gupta et. al. utilized several approaches for AD detection from MRI images (25). The team achieved 94.7% accuracy in binary classification and 85.00% accuracy in three-way classification (25). Ozic et. al. used 140 MRI images for their model, and the experiments yielded among the highest accuracy rates of 79.3% for white matter classification (26). Bi and Wang used a multi-task learning strategy based on the Spike Convolutional Deep Boltzmann machine and achieved an accuracy of 95% (27). Kim and Kim utilized Deep Neural Network with four hidden layers, which they extracted features from relative power and attained an accuracy of 75% (28). On the other hand, Ieracitano et. al. used CNN with two hidden layers, and the extracted feature was 2D grayscale Periodogram images, from which the group achieved the highest accuracy of 92% (29). He and Zhao used Restricted Boltzmann Machine with three hidden layers, for which they used raw data. The highest accuracy that they achieved was 92%. (30). J Huggins et. al. utilized AlexNet as a CNN model, and the extracted feature was 2 D RGB of Scalogram images, which achieved 98.9% accuracy (31). Alvi et. al. used Long Short-Term Memory, Gated Recurrent Unit, k-Nearest Neighbors, and Support Vector Machine (32). The team achieved an accuracy of over 95%, and the extracted feature was Raw Electroencephalogram Data (32). Liu et. al. formulated a sophisticated deep-learning structure that included stacked auto-encoders and a SoftMax output tier (33). The highest accuracy they achieved was 45.28% in the 4-class classification. The mean values of the binary classification performance of MRI and PET images were found to be 77.3% (33). Liu et. al. implemented SAE along with a SoftMax logistic regressor and a zero-mask approach for

data amalgamation to extract supplementary information from multiple modes of neuroimaging data (34). In 2018, Lu and colleagues applied a sparse autoencoder (SAE) for initial training and employed deep neural networks (DNN) in the final stage. They achieved a classification accuracy of 84.6% for differentiating AD and healthy control cases and a prediction accuracy of 82.9% for mild cognitive impairment (MCI) conversion (35). Suk et. al. initialized SAE parameters with target-unrelated samples and tuned the optimal parameters with target-related samples to obtain a 98.8% accuracy for AD and cognitive normal classification and 83.7% accuracy for the prediction of MCI to AD conversion (36). Islam and Zhang achieved a 73.45% accuracy with a 416 data count (37). They concluded that AD can be classified into three major stages (37). Additionally, they illustrated how hyper-parameters from a deep convolutional neural network could aid in extracting features from inadequate medical image datasets (37). Their model was inspired by the Inception-V4 network; the network that they provided accepted an MRI scan, then processed it by gathering the features of each layer from the initial stem layer to the final drop-out layer (37). For the classification tasks, Subramoniam et. al. employed Residual Neural Networks (ResNet-101) in the architecture; they classified AD using a dense neural network with a vanilla structure (38). They obtained the highest accuracy in the moderate class, of 100% (38). Subramoniam and his team recorded an average accuracy of 99.70% on the OASIS dataset (38). Ghazal et. al. utilized a four-class dataset in their research; a system model was suggested based on 6400 MR images of Alzheimer's patients, resulting in a 91.70% success rate (39). Guerrero et. al. utilized the data gathered from both ADNI and ADNI-GO datasets (40). The ADNI dataset contained 511 images classified into four distinct categories, whereas the ADNI-GO dataset included 363 images divided into two classes. The ADNI dataset achieved a success rate of 71%, with the ADNI-GO dataset reaching 65% (40). Eskildsen et. al. conducted a study to find if patterns of cortical thickness measurements could be used to predict Alzheimer's Disease in those with Mild Cognitive Impairment (MCI) (41). Distinct patterns of deterioration were identified, with certain features selected as regions of focus from those patterns; The accuracy was 81% (41). Plant et. al. utilized a data mining framework joined with three types of classifiers - Support Vector Machine (SVM), Bayesian estimates, and Voting Feature Intervals (VFI) - to create a numerical index for anticipating outcomes in their work (42). Data were acquired from 32 Alzheimer's Disease (AD) patients, 24 MCI individuals, and 18 healthy controls (42). Results from this research indicated that pattern matching using multivariate techniques achieved a highly accurate rate of 92% in a clinical setting (42). MRI surface morphometry mapping was employed by Devanand et. al. in order to examine and identify any local distortions of the hippocampus, parahippocampal gyrus, and entorhinal cortex that could indicate potential conversion from MCI to AD (43). MRI of the brain of 130 people with MCI, labeled as broadly defined, and 61 healthy individuals was performed using surface morphological analysis (43). These individuals were

monitored for approximately four years at one site as part of the study (43). A research study by Zhang et. al. explored the use of Multimodal multi-task (M3T) learning to make multiple predictions from various data sources (44). The multi-task feature selection process was employed to identify the overlapping group of pertinent features when analyzing multiple variables from different sources (44). This same grouping of features was then combined with a multimodal support vector to make predictions for multiple (regression and classification) tasks (44). Using MRI, the highest performance achieved by a single modality was only 85% (44). SVM is one of the most commonly used machine learning techniques, which allows for extracting high dimensional and meaningful features to derive classification models for automated clinical diagnosis (45). Deep learning, a rapidly advancing branch of machine learning that utilizes raw neuroimaging information to generate features, is gaining substantial interest in the area of extensive, high-dimensional medical imaging exploration as proposed by Plis et. al. (46). Deep learning allows for an optimal representation of the data to be derived from the raw images without requiring initial image pre-processing, resulting in a more unbiased and impartial process (46).

Methods and Dataset

The flowchart of the work process is shown in Figure 1.

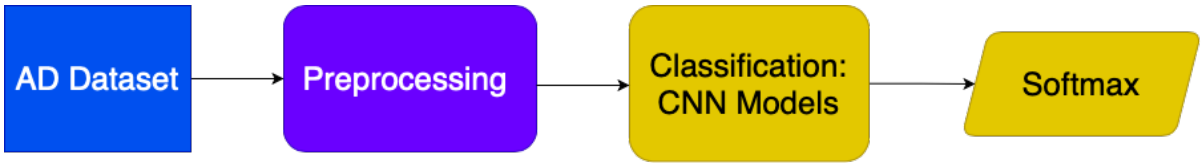


Figure 1. Flow diagram of the work process

Various CNN models are used for classification, including ResNet-18, DenseNet-121, AlexNet, EfficientNetB-7, and ResNet-50, which will be compared and evaluated. AlexNet's seven layers of activation functions, called ReLU, pass the positive output and suppress the negative output in feature maps and set them to zero (47, 48). The deep feature maps' dimensions are reduced by three max-pooling layers (47, 48). ResNet-50 is a deep learning model that is derived from the ResNet family (48, 49). Other ResNet models are ResNet-18, ResNet-34, ResNet-101, and ResNet-152, which demonstrate varying levels of accuracy in classification problems. The ResNet-50 architecture retains 50 layers of the convolutional neural network, with 48 of them being convolutional layers used to extract deep feature maps (48, 49). Two pooling layers are employed to reduce the dimensions of deep feature maps (48, 49). DenseNet-121 consists of 121 layers, a substantial proportion of which

are convolutional layers instrumental in extracting complex feature maps (50). A pair of pooling layers are implemented to control the dimensionality of these feature maps. EfficientNetB-7 has a total of 66 million parameters (51). Also, it is the most accurate model in the EfficientNet family, being 8.4x smaller than most of the existing CNNs (52). These CNNs have shown significant success in image recognition tasks and have been used profusely in research papers. Previously these models were used in a research paper that compared eleven CNNs, including ResNet-18, ResNet-50, and DenseNet-121, to investigate their merits in detecting lung abnormalities in small datasets of COVID-19 patients (52). Also, in another research paper, sixteen different CNNs were compared using the CheXpert and COVID-19 Image datasets (53).

The data for this paper was obtained from the OASIS dataset (54). A total of 6400 MR images are in the dataset, consisting of four different classes (54). The dataset is classified as mild AD, moderate AD, non-AD, and very mild AD (54). Some examples of the dataset can be found in Table 1.

The dataset consists of 896 images of patients with mild AD, 64 images of patients with moderate AD, 3200 images of individuals without AD (non-AD), and 2240 images of patients with very mild AD (54). The dataset was unbalanced; hence different weights were asssigned to each class. Also, accuracy is not a sufficient indicator when dealing with imbalanced datasets because it can be misleading; thus, other metrics, such as recall, precision, f-1 score, specificity, and Matthew's correlation coefficient, were used in this research paper.

All the images in the dataset have 128x128 pixels; images are resized accordingly with optimal image size for each architecture. For instance, ResNet-50 uses images of 224x224 pixels. The initial learning rate was set to 0.001 for all deep learning architectures. The epoch number was set to 50, because, after reaching that epoch, there were no further changes in the accuracy. Indeed, after reaching the 100^{th} epoch, a decrease in accuracy became apparent; perhaps because of overtraining. The dataset was allocated 80% for training and 20% for testing. The data was not age-adjusted or age-standardized.

Table 1: Some images from OASIS Dataset (54)

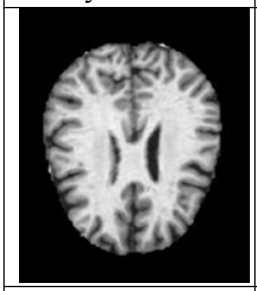
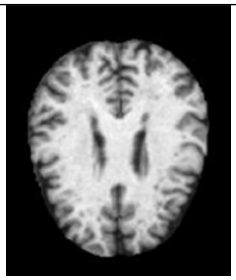
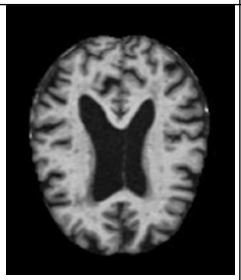
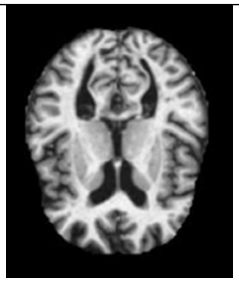
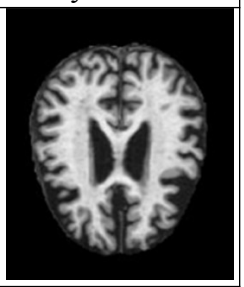
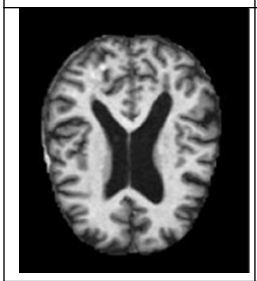
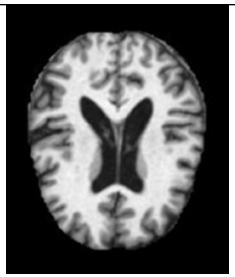
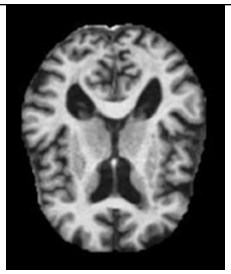
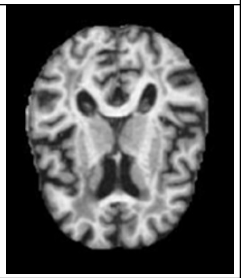
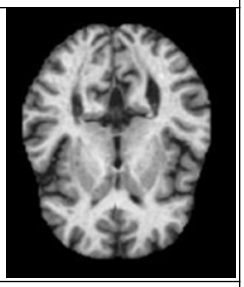
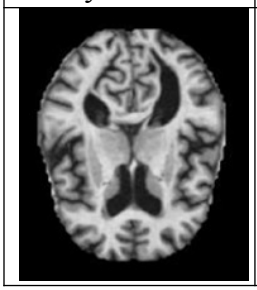
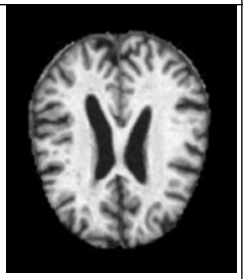
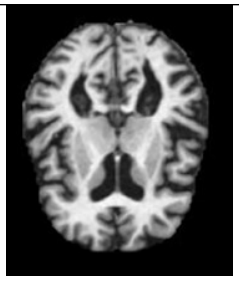
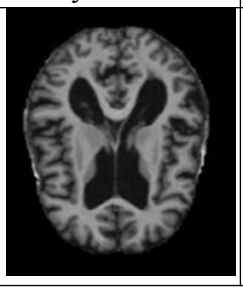
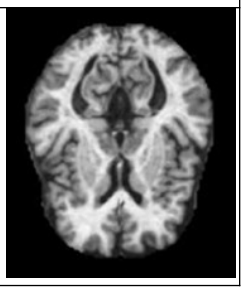
Very Mild AD	Mild AD	Mild AD	Non-AD	Very Mild AD
				
Non-AD	Very Mild AD	Non-AD	Non-AD	Non-AD
				
Very Mild AD	Very Mild AD	Mild AD	Very Mild AD	Mild AD
				

Table 2: Image counts of the brain MRI dataset

Dataset	Non-AD	Very mild AD	Mild AD	Moderate AD	Total
Train	2566	1791	715	48	5120
Test	634	449	181	16	1280
Total	3200	2240	896	64	6400

Model Performance Metrics

Performance metrics of the model were calculated to ascertain the reliability of the study. To measure the performance metrics, six methods were used: Accuracy, Recall (equation 1), Precision (equation 2), f-1 score (equation 3), and specificity (equation 4) (55, 56). An illustrative example of the "True Class" confusion matrix (Table 3) was prepared. There are two distinct classes, denoted P and N. The predicted output of these classes was either true or false.

Table 3: True Class

	Positive (P)	Negative (N)
True (T)	True Positive (TP)	False Positive (FP)
False (F)	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ Equation 1

$$Recall = \frac{TP}{TP + FN}$$ Equation 2

$$Precision = \frac{TP}{TP + FP}$$ Equation 3

$$F1 - score = \frac{2\,xPrecisionxRecall}{Precision + Recall}$$ Equation 4

$$Specificity = \frac{TN}{FP + TN}$$ Equation 5

$$Matthew's\,Correlation\,Coefficient = \frac{(TP\,x\,TN) - (FP\,x\,FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$ Equation 6

A confusion matrix is a tool for evaluating the performance of multi-class classification models, such as a 3x3 square (Table 4). In this matrix type, the model's predicted and true labels are compared between all classes to determine the number of accurate and inaccurate predictions for each.

Table 4: Confusion Matrix for a Multi-class Classification Test

	True Class			
Predicted		Y	Z	K
	Y	TP_Y	E_{ZY}	E_{KY}
	Z	E_{YZ}	TP_Z	E_{KZ}
	K	E_{YK}	E_{ZK}	TP_K

The TP_Y and TP_Z and TP_K in the Table 4 shows the number of true positives in classes Y, Z and K respectively (55, 56). E_{YK}, on the other hand, is in the Y class and shows the number of those misclassified as K class (55, 56). Furthermore, E_{ZY} is in the Z class, indicating the number of those misclassified as Y class (55, 56). Similarly, E_{KY} shows the number of those misclassified as Y class in the K class. The number of false negatives for class Y is the sum of E_{YZ} and E_{YK} (55, 56). The number of false negatives for class Z is the sum of E_{ZY} and E_{ZK} (55, 56). The number of false negatives for class K is the sum of E_{KY} and E_{KZ} (55, 56). The number of false positives for class Y is the sum of E_{ZY} and E_{KY} (55, 56). The number of false positives for class Z is the sum of E_{YZ} and E_{KZ} (55, 56). The number of false positives for class K is the sum of E_{YK} and E_{ZK} (55, 56). An important example of a true negative can be given from Y. The number of true negatives for class Y equals the total number of TP_Z, E_{KZ}, E_{ZK}, and TP_K (55, 56). The number of true negatives for class Z is equal to the total number of TP_K, TP_Y, E_{YK}, E_{ZK} (55, 56). Lastly, the number of true negatives for class K is equal to the total number of TP_Y, TP_Z, E_{YZ}, E_{YZ} (55, 56).

Results and Discussion

DenseNet 121, EfficientNetB-7, ResNet-50, ResNet-18, and AlexNet were used and evaluated in this section. In addition, the classification accuracy (equation 1), recall (equation 2), precision (equation 3), f-1 score (equation 4), and specificity (equation 5) results derived from the proposed models are reported in Table 5.

Table 5: Model Performance Metrics. Mild, moderate, non-AD and very mild AD are arranged from top to bottom for each model.

Model	Precision	Recall	F-1 Score	Sensitivity	Specificity	Total Accuracy
AlexNet	0.51	0.69	0.59	0.69	0.92	70.5%
	0.38	0.75	0.50	0.75	0.99	
	0.72	0.84	0.77	0.84	0.75	
	0.78	0.59	0.59	0.59	0.85	
EfficientNetB-7	0.099	0.18	0.13	0.18	0.86	37.0%
	0.063	0.0070	0.013	0.0073	0.99	
	0.60	0.49	0.54	0.49	0.50	
	0.17	0.28	0.21	0.28	0.63	
ResNet-50	0.92	0.97	0.94	0.97	0.99	95.1%
	0.75	1.00	0.86	1.00	1.00	
	0.99	0.93	0.96	0.93	0.99	
	0.92	0.98	0.95	0.98	0.95	
ResNet-18	0.83	0.92	0.87	0.92	0.97	88.4%
	0.71	1.0	0.83	1.00	1.00	
	0.93	0.88	0.91	0.88	0.93	
	0.84	0.87	0.86	0.87	0.98	
DenseNet-121	0.87	0.91	0.89	0.91	0.98	90.5%
	0.57	1.0	0.73	1.0	0.99	
	0.97	0.88	0.93	0.88	1.0	
	0.84	0.94	0.89	0.94	0.92	

AlexNet has been used as one of the deep learning models, and it received the second lowest specificity when applied to non-AD. A high- specificity test accurately identifies unaffected individuals at a low false-positive rate. This makes it particularly useful when a false- positive result could lead to unnecessary treatment or undue anxiety. However, a high-specificity test may miss some true positive cases, so it is important to consider both sensitivity and specificity when evaluating the performance of a diagnostic test. The Matthews Correlation Coefficient (MCC) is a quantitative measure of the efficiency of binary classifications (57). It considers true positives, true negatives, false positives, and false negatives and is bounded by the range -1 to +1. An MCC score of +1 indicates a perfect prediction, 0 corresponds to a random prediction, while -1 illustrates total disagreement between predicted and observed results (57). AlexNet demonstrated an MCC of 0.538, 0.587, 0.530, and 0.463 for AD, moderate AD, non-AD and very mild AD, respectively.

EfficientNetB-7 (Table 5), one of the CNN models, performed poorly in many respects. EfficientNetB-7 demonstrated an MCC of 0.03690, -0.01756, 0.00258, and -0.07622 for mild, moderate, non-AD and very mild AD, respectively. A negative MCC indicates that the classifier performs inferior to random guessing and the predictions are unreliable. This may stem from the classifier systematically making unidirectional errors when making predictions.

ResNet-50, deep convolutional neural network architecture, achieved the highest overall accuracy. The ResNet-50 model demonstrated an MCC of 0.9357, 0.8643, 0.9177 and 0.9223 for mild, moderate, non-AD and very mild AD respectively. The ResNet-50 model was effective in accurately detecting AD.

ResNet-18, a deep convolutional neural network architecture, attained one of the highest overall accuracy ratings. Nonetheless, accuracy rose again to 96.34% when applied on very mild AD. The ResNet-18 model achieved an MCC of 0.8544, 0.8410, 0.8150 and 0.8340 for mild, moderate, non-AD and very mild AD respectively.

DenseNet-121, a deep convolutional neural network architecture, achieved the second-highest overall accuracy ratings, with an MCC of 0.8718, 0.7523, 0.9280 and 0.8339 for mild, moderate, non-AD and very mild AD respectively.

Specificity refers to correctly identifying those without the disease, in other words, the true negative rate. A high specificity means that fewer false positives exist for people who don't have the disease but test positive. This is considered a good sign because it reduces unnecessary treatments. In all models, moderate AD had a high specificity, which means that all models accurately identified samples that did not belong to the moderate AD class. Sensitivity means correctly identifying those with the disease; i.e. a true positive rate. When this rate is low, the models miss a number of individuals who have the disease, which is a false negative. This

situation can lead to individuals who are actually sick might not receive the necessary treatment.

During times of a disease outbreak such as the COVID-19 pandemic, it becomes crucial to detect and identify as many infected individuals as possible in order to prevent further spread. This means that having a test with high sensitivity is preferable. A sensitive test can accurately identify the majority of infected individuals minimizing the risk of missing any infections. While false positives might inconvenience those who receive misdiagnoses, it is important to prioritize health and avoid the consequences of missing infected cases. On the other hand, when it comes to diagnosing conditions like cancer, having a test with high specificity is often desirable so as to minimize false positives and reduce instances where healthy individuals are wrongly diagnosed with the disease. False positives in cancer diagnosis can lead to treatments that may potentially harm patients, increase anxiety levels, and incur unnecessary healthcare costs. However, striking a balance is crucial so as not to overlook true cancer cases. When diagnosing AD, it is crucial to consider both sensitivity and specificity. Having a test with sensitivity is valuable as it can accurately detect the majority of individuals with Alzheimer's, reducing the risk of overlooking cases. This, enables more patients to receive treatment. Given that AD often shares symptoms with forms of dementia, a test with specificity proves helpful in distinguishing AD from other conditions. Enhancing the specificity of AD tests can help improve the overall reliability of AD classification, contributing to better research outcomes.

Additionally, the analysis of the model performance metrics is facilitated by the use of a confusion matrix (Tables 6-10). Confusion matrices are important because they help in performance evaluation, error analysis, and also imbalance class detection. When the classes are imbalanced in the dataset, the number of images/samples in each class can differ. Hence, examining the distribution of T_P, T_N, F_P, and F_N across classes can help with potential issues related to imbalanced data.

Table 6: AlexNet Confusion Matrix

	Mild AD	Moderate AD	Non-AD	Very mild AD
Mild AD	93	1	11	76
Moderate AD	4	6	0	6
Non-AD	15	1	454	164
Very mild AD	22	0	77	350

This confusion matrix shows that AlexNet was accurate in classifying very mild AD. However, it could not classify a higher percentage of the cases in moderate AD. This might be related to the number of images because the amount of training and test data in the moderate AD category was significantly lower in comparison to the other classes.

Table 7: EfficientNetB-7 Confusion Matrix

	Mild AD	Moderate AD	Non-AD	Very mild AD
Mild AD	18	22	115	26
Moderate AD	4	1	4	7
Non-AD	33	61	380	160
Very mild AD	44	59	271	75

As can be seen from the confusion matrix, EfficientNetB-7 has a lower accuracy than random guessing for mild, moderate, and very mild AD. The model was not compatible with EfficientNetB-7.

Table 8: ResNet-50 Confusion Matrix

	Mild AD	Moderate AD	Non-AD	Very mild AD
Mild AD	166	0	12	3
Moderate AD	0	15	0	1
Non-AD	0	0	629	5
Very mild AD	5	0	33	411

ResNet-50 was the most successful CNN for the model. It demonstrated high accuracy in the non-AD classification category. To provide a more comprehensive evaluation of the classification accuracy for moderate AD, an expanded set of training and testing data may be required.

Table 9: ResNet-18 Confusion Matrix

	Mild AD	Moderate AD	Non-AD	Very mild AD
Mild AD	149	0	13	19
Moderate AD	0	12	2	2
Non-AD	5	0	592	37
Very mild AD	8	0	62	379

While ResNet-18 did not achieve the same level of accuracy as ResNet-50, it still emerged as the third most effective model in this study.

Table 10: DenseNet-121 Confusion Matrix

	Mild AD	Moderate AD	Non-AD	Very mild AD
Mild AD	157	0	15	9
Moderate AD	4	12	0	0
Non-AD	2	0	618	14
Very mild AD	7	0	64	378

DenseNet-121 ranked as the second most accurate convolutional neural network in this study. While it demonstrated significant success in classifying non-AD cases, it under performed in accurately classifying moderate AD.

AlexNet achieved accuracies of 51.3%, 37.5%, 71.6% and 77.95% in mild, moderate, non-AD and very mild AD detection respectively. EfficientNetB-7 achieved accuracies of 9.95%, 6.25%, 59.94% and 16.70% in mild, moderate, non-AD and very mild AD detection respectively. ResNet-50 achieved accuracies of 91.71%, 93.75%, 99.21% and 91.53% in mild, moderate, non-AD and very mild AD detection respectively. ResNet-18 achieved accuracies of 82.32%, 75.00%, 93.37% and 84.41% in mild, moderate, non-AD and very mild AD detection respectively. DenseNet-121 achieved accuracies of 86.74%, 75.00%, 97.47% and 84.19% in mild, moderate, non-AD and very mild AD detection respectively. The highest and lowest accuracies for mild AD were found in the ResNet-50 (91.7%) and EfficientNetB-7 (9.95%) models respectively. The highest and lowest accuracies for moderate AD were found in the ResNet-50 (93.75%) and EfficientNetB-7 (6.25%) models respectively. The highest and lowest accuracies for non-AD were found in ResNet-50 (99.21%), and in the EfficientNetB-7 (59.94%) models respectively. The highest and lowest accuracies in very mild AD were found in the ResNet-50 (91.53%) and in the EfficientNetB-7 (16.70%) models respectively. Although in overall accuracy, DenseNet-121 was more effective than ResNet-18, in very mild AD, ResNet-18 was found to be more effective than DenseNet-121. Moreover, ResNet-18 and DenseNet-121 were equally effective for detecting moderate AD.

Table 11: Comparison of similar studies in the literature

Study (reference)	Accuracy (%)	Data quantity
This paper	95.1	6400
Gupta et. al. (25)	94.7	4315
Ozic et. al. (26)	92.8	140
Liu et. al. (33)	45.3	311
Islam and Zhang (37)	73.5	416
Subramonium et. al. (38)	99.7	6400
Ghazal et. al. (39)	91.7	6400

Compared to other studies, the dataset used in this study appears to contain more images. On the basis of the classification performance, this study proposes an acceptable model.

Conclusion

Deep learning architectures were employed for AD detection using MR images from the OASIS dataset. Through experimentation with ResNet-18, DenseNet-121, AlexNet, EfficientNetB-7, and ResNet-50 architectures, the study demonstrated the critical role of image classification. The performance of the models was evaluated using several metrics, including accuracy, recall, precision, f-1 score, specificity, and Matthew's correlation coefficient. AlexNet achieved an average precision of 59.75%, making it the second worst. Its average recall and sensitivity scores were also second lowest at 71.75%. The f-1 score and average specificity were 61.25% and 87.75% respectively. EfficientNetB-7 performed the worst in all of the metrics. It achieved an average precision of 23.30%. Its average recall and sensitivity scores were also second lowest at 23.92%. The f-1 score and the average specificity were 22.33% and 74.50% respectively. This CNN model failed to accuretely detect or classify AD. The performance metrics of ResNet-50 were the best among all the measurements. It achieved a mean precision of 89.50% and a mean recall and sensitivity mean of 97.00%. The F-1 metric and average specificity were 92.75% and 98.25% respectively. In conclusion, this CNN model accurately classified all categories of AD. The performance metrics of ResNet-18 were one of the best among all the measurements. It achieved a mean precision of 82.75% and a mean recall and sensitivity score of 91.75%. The f-1 metric and average specificity were 86.75% and 97.00% respectively. Finally, the performance of DenseNet-121 proved to be exemplary when compared with the other metrics. It had a mean precision of 81.25%, a mean recall and sensitivity score of 93.25%, an f-1 metric rating of 86.00%, and an average specificity score of 97.25%.

In subsequent studies, the features obtained from the architectures can be evaluated in classical classifiers. Additionally, Principal

Component Analysis (PCA), Independent Component Analysis (ICA), and Locally Linear Embedding (LLE) can be employed for AD classification tasks. Moreover, using feature extraction methodologies such as back-propagation neural networks can be helpful for AD detection. Another excellent method could involve mapping the differences occurring in grey matter and white matter regions using 3D T1- weighted MRI with the Voxel-Based Morphometry.

While this study – and others – compared the performance of different CNNs, there is a paucity of research to determine *why* specific neural networks perform better than others for specific applications. The *mechanisms* behind the mere application of off-the-shelf CNNs need to be studied so that networks with specific architectures can be designed that are better suited for complex, esoteric or unconventional applications.

References

1. "Alzheimer's Disease." *Mayo Clinic*, 2 Feb. 2023, www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447.

2. Breijyeh, Zeinab, and Rafik Karaman. "Comprehensive Review on Alzheimer's Disease: Causes and Treatment." *Molecules (Basel, Switzerland)* vol. 25,24 5789. 8 Dec. 2020, http://doi.org/10.3390/molecules25245789

3. Haque, Rafi U., and Allan I. Levey. "Alzheimer's Disease: A Clinical Perspective and Future Nonhuman Primate Research Opportunities." *Proceedings of the National Academy of Sciences*, vol. 116, no. 52, 2019, pp. 26224–26229, https://doi.org/10.1073/pnas.1912954116.

4. Murphy, M Paul, and Harry LeVine 3rd. "Alzheimer's disease and the amyloid-beta peptide." *Journal of Alzheimer's disease : JAD* vol. 19,1 (2010): 311-23. http://doi.org/10.3233/JAD-2010- 1221

5. Ahmad, Md Afroz et al. "Neuroinflammation: A Potential Risk for Dementia." *International journal of molecular sciences* vol. 23,2 616. 6 Jan. 2022, http://doi.org/10.3390/ijms23020616

6. Yiannopoulou, Konstantina G, and Sokratis G Papageorgiou. "Current and future treatments for Alzheimer's disease." *Therapeutic advances in neurological disorders* vol. 6,1 (2013): 19-33. doi:10.1177/17562856124616794 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6935598/

7. Rasmussen, J., & Langerman, H. (2019). Alzheimer's Disease - Why We Need Early Diagnosis. *Degenerative neurological and neuromuscular disease*, *9*, 123–130. https://doi.org/10.2147/DNND.S228939

8. Dubois, Bruno et al. "Early detection of Alzheimer's disease: new diagnostic criteria." *Dialogues in clinical neuroscience* vol. 11,2 (2009): 135-9. http://doi.org/10.31887/DCNS.2009.11.2/bdubois

9. "Alzheimer's Disease Facts and Figures." *Alzheimer's Disease and Dementia*, www.alz.org/alzheimers-dementia/facts-figures.

10. "Learn How Alzheimer's Is Diagnosed." *Mayo Clinic*, 7 May 2022, www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048075.

11. Wang, Ge, et al. "Mini-Mental State Examination and Montreal Cognitive Assessment as Tools for Following Cognitive Changes in Alzheimer's Disease Neuroimaging Initiative Participants." *Journal of Alzheimer's Disease*, vol. 90, no. 1, 2022, pp. 263–270, https://doi.org/10.3233/jad-220397.

12. Wadhwa, A.; Bhardwaj, A.; Singh Verma, V. A review on brain tumor segmentation of MRI images. *Magnetic Resonance Imaging* **2019**, *61*, 247–259. https://doi.org/10.1016/j.mri.2019.05.043

13. Huntenburg, Julia M, et al. "Nighres: Processing Tools for High-Resolution Neuroimaging." *GigaScience*, vol. 7, no. 7, 2018, https://doi.org/10.1093/gigascience/giy082.

14. Man, Mei Yen et al. "A Review on the Bioinformatics Tools for Neuroimaging." *The Malaysian journal of medical sciences : MJMS* vol. 22,Spec Issue (2015): 9-19. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795522/

15. El-Dahshan, El-Sayed A., et al. "Computer-Aided Diagnosis of Human Brain Tumor through MRI: A Survey and a New Algorithm." *Expert Systems with Applications*, vol. 41, no. 11, 2014, pp. 5526–5545, https://doi.org/10.1016/j.eswa.2014.01.021.

16. Wu, Zhanxiong et al. "Gray Matter Deterioration Pattern During Alzheimer's Disease Progression: A Regions-of-Interest Based Surface Morphometry Study." *Frontiers in aging neuroscience* vol. 13 593898. 3 Feb. 2021, http://doi.org/10.3389/fnagi.2021.593898

17. Karas, G B et al. “Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease.” *NeuroImage* vol. 23,2 (2004): 708-16. http://doi.org/10.1016/j.neuroimage.2004.07.006

18. Guo, Xiaojuan et al. “Voxel-based assessment of gray and white matter volumes in Alzheimer's disease.” *Neuroscience letters* vol. 468,2 (2010): 146-50. http://doi.org/10.1016/j.neulet.2009.10.086

19. Chan, H. P., Hadjiiski, L. M., & Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical physics*, *47*(5), e218–e227. https://doi.org/10.1002/mp.13764

20. Minarno, Agus Eko et al. “Texture feature extraction using co-occurrence matrices of sub-band image for batik image classification.” *2014 2nd International Conference on Information and Communication Technology (ICoICT)* (2014): 249-254.52. http://dx.doi.org/10.1109/ICoICT.2014.6914074

21. Madhavan, S. and Jones, M. T. "Deep Learning Architectures." IBM Developer, IBM Corporation, 9 Jan. 2022, https://developer.ibm.com/articles/cc-machine-learning-deep-learning- architectures/

22. P Park, Chihyun, et al. “Prediction of Alzheimer’s Disease Based on Deep Neural Network by Integrating Gene Expression and DNA Methylation Dataset.” *Expert Systems with Applications*, vol. 140, 2020, p. 112873, https://doi.org/10.1016/j.eswa.2019.112873.

23. Jo, Taeho, et al. “Deep Learning in Alzheimer’s Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data.” *Frontiers in Aging Neuroscience*, vol. 11, 2019, https://doi.org/10.3389/fnagi.2019.00220.

24. Allioui, Hanane, et al. “Deep MRI Segmentation: A Convolutional Method Applied to Alzheimer Disease Detection.” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 11, 2019, https://doi.org/10.14569/ijacsa.2019.0101151.

25. Gupta, Ashish et al. “Natural Image Bases to Represent Neuroimaging Data.” *International Conference on Machine Learning* (2013). http://proceedings.mlr.press/v28/gupta13b.pdf

26. ÖZİÇ, Muhammet Üsame, and Seral ÖZŞEN. “3B Alzheimer Mr Görüntülerinin Hacimsel Kayıp Bölgelerindeki Voksel Değerleri Kullanılarak Sınıflandırılması.” *El-Cezeri Fen ve Mühendislik Dergisi*, 2020, https://doi.org/10.31202/ecjse.728049.

27. Bi, Xiaojun, and Haibo Wang. “Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning.” *Neural networks : the official journal of the International Neural Network Society* vol. 114 (2019): 119-135. http://doi.org/10.1016/j.neunet.2019.02.005

28. Kim, Donghyeon and Kiseon Kim. “Detection of Early Stage Alzheimer’s Disease using EEG Relative Power with Deep Neural Network.” *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018): 352-355. https://doi.org/10.1109/EMBC.2018.8512231

29. Ieracitano, C., Mammone, N., Bramanti, A., Hussain, A., & Morabito, F. C. (2019). A Convolutional Neural Network approach for classification of dementia stages based on 2D- spectral representation of EEG recordings. *Neurocomputing*, *323*, 96–107. https://doi.org/10.1016/j.neucom.2018.09.071

30. Zhao, Yilu and Lianghua He. “Deep Learning in the EEG Diagnosis of Alzheimer's Disease.” *ACCV Workshops* (2014). https://doi.org/10.1007/978-3-319-16628-5_25

31. Huggins, Cameron J et al. “Deep learning of resting-state electroencephalogram signals for three-class classification of Alzheimer's disease, mild cognitive impairment and healthy ageing.” *Journal of neural engineering vol.* 18,4 10.1088/1741-2552/ac05d8. 17 Jun. 2021, https://doi.org/10.1088/1741-2552/ac05d8

32. Gorji, Hamed Taheri, and Naima Kaabouch. “A Deep Learning approach for Diagnosis of Mild Cognitive Impairment Based on MRI Images.” *Brain sciences*vol. 9,9 217. 28 Aug. 2019, http://doi.org/10.3390/brainsci9090217

33. Liu, Siqi, et al. “Early Diagnosis of Alzheimer’s Disease with Deep Learning.” *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 2014, https://doi.org/10.1109/isbi.2014.6868045.

34. Liu, Siqi et al. “Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease.” *IEEE transactions on bio-medical engineering* vol. 62,4 (2015): 1132-40 https://doi.org/10.1109/tbme.2014.2372011

35. Lu, Donghuan, et al. “Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer’s Disease Using Structural MR and FDG-PET Images.” *Scientific Reports*, vol. 8, no. 1, 2018, https://doi.org/10.1038/s41598-018-22871-z

36. Suk, Heung-Il et al. “Latent feature representation with stacked auto-encoder for AD/MCI diagnosis.” *Brain structure & function* vol. 220,2 (2015): 841-59. http://doi.org/10.1007/s00429- 013-0687-3

37. Islam, Jyoti, and Yanqing Zhang. “A Novel Deep Learning Based Multi-Class Classification Method for Alzheimer’s Disease Detection Using Brain MRI Data.” *Brain Informatics*, 2017, pp. 213–222, https://doi.org/10.1007/978-3-319-70772-3_20.

38. Subramoniam, Manu, et al. “Deep Learning Based Prediction of Alzheimer’s Disease from Magnetic Resonance Images.” *arXiv.Org*, 14 May 2021, https://doi.org/10.48550/arXiv.2101.04961

39. M. Ghazal, Taher, et al. “Alzheimer Disease Detection Empowered with Transfer Learning.” *Computers, Materials & Continua*, vol. 70, no. 3, 11 Oct. 2021, pp. 5005–5019, https://doi.org/10.32604/cmc.2022.020866.

40. Guerrero, R et al. “Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO.” *NeuroImage* vol. 94 (2014): 275-286. http://doi.org/10.1016/j.neuroimage.2014.03.036

41. Eskildsen, Simon F et al. “Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning.” *NeuroImage* vol. 65 (2013): 511-21. http://doi.org/10.1016/j.neuroimage.2012.09.058.

42. Plant, Claudia et al. “Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease.” *NeuroImage* vol. 50,1 (2010): 162-74. https://doi.org/10.1016/j.neuroimage.2009.11.046

43. Devanand, D. P., Bansal, R., Liu, J., Hao, X., Pradhaban, G., & Peterson, B. S. (2012). MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease. *NeuroImage*, *60*(3), 1622–1629. https://doi.org/10.1016/j.neuroimage.2012.01.075

44. Zhang, Daoqiang et al. “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease.” *NeuroImage* vol. 59,2 (2012): 895-907. http://doi.org/10.1016/j.neuroimage.2011.09.069

45. Rathore, Saima et al. “A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages.” *NeuroImage* vol. 155 (2017): 530-548. http://doi.org/10.1016/j.neuroimage.2017.03.057

46. Plis, Sergey M et al. “Deep learning for neuroimaging: a validation study.” *Frontiers in neuroscience* vol. 8 229. 20 Aug. 2014, http://doi.org/10.3389/fnins.2014.00229

47. Krizhevsky, Alex, et al. “ImageNet Classification with Deep Convolutional Neural Networks.” *NIPS*, https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

48. Mohammed, Badiea Abdulkarem, et al. “Multi-Method Analysis of Medical Records and MRI Images for Early Diagnosis of Dementia and Alzheimer’s Disease Based on Deep Learning and Hybrid Methods.” *Electronics*, vol. 10, no. 22, 2021, p. 2860, https://doi.org/10.3390/electronics10222860.

49. He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” *arXiv.Org*, 10 Dec. 2015, https://doi.org/10.48550/arXiv.1512.03385

50. Huang, Gao, et al. “Densely Connected Convolutional Networks.” *arXiv.Org*, 28 Jan. 2018, https://doi.org/10.48550/arXiv.1608.06993

51. Tan, Mingxing. “EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling.” *Google Research Blog*, 29 May 2019, https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

52. Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking model scaling for convolutional Neural Networks*. https://doi.org/10.48550/ARXIV.1905.11946

53. Yang, Yuan et al. “A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions.” *Computers in biology and medicine* vol. 139 (2021): 104887. http://doi.org/10.1016/j.compbiomed.2021.104887

54. “Oasis Brains.” *OASIS Brains - Open Access Series of Imaging Studies*, www.oasis-brains.org/.

55. Powers, David M. W. “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation.” *arXiv.Org*, 11 Oct. 2020, https://doi.org/10.48550/arXiv.2010.16061

56. Tharwat, Alaa. “Classification Assessment Methods.” *Applied Computing and Informatics*, 17 Aug. 2018, www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html.

57. Chicco, Davide, and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.” *BMC genomics* vol. 21,1 6. 2 Jan. 2020, http://doi.org/10.1186/s12864-019-6413-7