Alzheimer's Disease Detection Using Deep Learning

Abhinaya Puvvada
Department of Computer Science
Texas Tech University
Lubbock, United States
Email: apuvvada@ttu.edu

Gowtham Aviligonda
Department of Computer Science
Texas Tech University
Lubbock, United States
Email: gaviligo@ttu.edu

Preethi Pavani Akurathi
Department of Computer Science
Texas Tech University
Lubbock, United States
Email: pakurath@ttu.edu

Keerthiga Kalidas
Department of Computer Science
Texas Tech University
Lubbock, United States
Email: keerthiga.kalidas@ttu.edu

Sampath Kumar Vuppala
Department of Computer Science
Texas Tech University
Lubbock, United States
Email: savuppal@ttu.edu

Abstract— Alzheimer's disease is a chronic progressive neurodegenerative disorder characterized by a severe deformation of the structure and function of the brain. Certain proteins are deposited into the brain, which in turn leads to overall reduced brain volume and ultimately death of brain cells. The disease predominantly affects the elderly population, and patients diagnosed with dementia are in most cases Alzheimer's disease patients. Although dementia has become a major societal concern, the diagnosis is difficult, particularly in the early stages. Despite considerable research efforts undertaken to identify various risk factors and relevant biomarkers, the current methods in use do not always adequately detect the early stages of the disease. Based upon a broad range of recent studies in this field, this paper aims to discover if deep learning can be used to diagnose Alzheimer's disease at the early stages using magnetic resonance imaging (MRI) scans. It compares the results from four cutting-edge convolutional neural network (CNN) architectures: VGG-16, VGG-19, Inception V3, and EfficientNet-B3. These models all have their advantages: VGG-16 and VGG-19 have deep architectures that lead to improved feature complexities; Inception V3 enables us to have a balanced trade-off between the complexity of the pattern recognition and the computational efficiency; EfficientNet-B3 is trained in a balanced way to obtain a model with improved accuracy and optimized model efficiency, with well-balanced depth, width, and resolution. Trained on a large database of MRI scans categorized into four stages of cognitive impairment (non-demented, mildly demented, moderately demented, and very mildly demented), these models all exhibit good performance, with the capacity to pick up on patterns in MRI data that we humans cannot easily see. Deploying deep learning technologies to assist in the process of Alzheimer's disease diagnosis will potentially enable faster and more objective diagnosis, which is a fundamental step to better medical treatment. Moreover, as deep learning technologies continue to be further developed, it is safe to say that deep learning has a role to play in neurology, and will continue to provide diagnostic support to neurologists, enabling them to better manage their patients, enhance clinical outcomes, and support optimal treatment of this disease.

Index terms: Alzheimer's disease(AD), deep learning, convolutional neural network (CNN), Visual Geometry Group 16 and 19 (VGG16, VGG19), InceptionV3, EfficientNet-B3, Disease detection, Accuracy, Loss value.

I. Introduction

Alzheimer's disease (AD) represents the most common form of neurodegenerative disease, characterized by the insidious progressive decline of cognitive functions, such as memory, reasoning, and language, primarily affecting individuals over the age of 65. It is the main cause of dementia, defined as a syndrome due to a disease affecting the brain, characterized by a decline in daily functioning and quality of life. The increase of the world population which is ageing will lead to an unparalleled upsurge in the number of people affected by Alzheimer's disease in the next few decades. The risk of developing the disease doubles every five years after age 65, and in its advanced stage, it usually leads to death. Estimates from the World Health Organization suggest that the number of people living with dementia will reach 152 million by 2050. With the enormous social and healthcare costs imposed by this disease, the need for diagnostic and therapeutic strategies is ever more imperative. Almost every modern image classification problem has been outperformed by deep learning models, in particular Convolutional Neural Networks (CNNs). They are well-suited for medical imaging tasks because of their ability to automatically learn features that are hierarchically nested to detect complex patterns in MRI scans. While the standardized

CNN architecture is very effective and can extract features from the image with reasonable performance, it also comes with some limitations. More advanced models, such as VGG-16, VGG-19, Inception V3, and EfficientNet-B3, can be added to further improve the performance of the Alzheimer's detection model by increasing the accuracy and robustness of the final prediction.

The research uses multiple cutting-edge deep learning architectures. CNNs are the backbone of many modern image classification systems, built from several layers of convolutions, usually followed by some pooling and activation of the results. Working with images of the brain, CNNs can learn to recognize subtle features of the underlying pathology, such as atrophy of specific regions, that are characteristic of Alzheimer's. Standard CNNs, however, can require large amounts of training data and can be prone to overfitting, especially when working with small datasets.

VGG-16 and VGG-19 are simple-to-understand deep learning architectures that were particularly successful at feature learning. The architectures use a sequence of convolutional layers followed by fully connected layers. VGG-16 and VGG-19 are 16 and 19 layers deep, respectively. The depth allows for learning of higher-level features which can be more abstract. This abstraction can help discriminate between normal brains and brains with Alzheimer's. Since these models have been trained on ImageNet, they can be transferred to the task of Alzheimer's detection. For this, there is a technique called transfer learning. With transfer learning, fine-tuning the model on the smaller label data is performed, instead of re-training the whole model.

The multi-scale architecture of Inception V3, which has filters of different sizes in the same layer, makes the model more expressive. In conjunction with the residual connections, the model is better able to find features at different scales (each scale captures different patterns), helping us detect the disease earlier. The efficient architecture that saves resources also makes the model faster and adaptable for clinical applications in real-time. EfficientNet-B3 is One of the most advanced models that trade-off accuracy and efficiency to scale the depth, width, and resolution of the network in a balanced way.EfficientNet-B3 achieved high performance with fewer parameters than CNNs without decreasing accuracy as much, which made it less prone to overfitting and more generalizable – a crucial asset for an Alzheimer's detection system where variations in MRI scans can arise from differences in patient demographics or stages of the disease. Compared with standard CNN models, highly advanced architectures such as VGG-16, VGG-19, Inception V3, and EfficientNet-B3 have many advantages. They help in extracting more complex and abstract features from the MRI images. Ultimately, that helps in better distinguishing healthy brains from those affected by Alzheimer's disease. As a result, the use of 'deeper' architectures and advanced techniques can lead to higher classification accuracy, which is a very important aspect of early diagnosis. More importantly, the risk of overfitting is greatly reduced by using transfer learning and engineering efficient architectures, even with limited data. Also, the models can be easily reconfigured to incorporate additional data or imaging modalities to increase their robustness and their applicability in real-world scenarios.

II. LITERATURE REVIEW

To enhance diagnostic accuracy, Liu et al. (2018) also constructed multi-modality cascaded convolutional neural networks (CNNs) for Alzheimer's disease diagnosis integrating different MRI data sources. The benefit of combining different imaging modalities against using data from a single modality for the diagnosis of Alzheimer's disease was compared in the study. Using multiple imaging modalities significantly improved the classification performance of Alzheimer's disease. The suggested approach achieved state-of-the-art results on several benchmark datasets, emphasizing the importance of incorporating a multitude of MRI features for reliable diagnosis [1].

One of the drawbacks acknowledged by the study is the incomplete treatment of the dimension overload problem during the integration process, which requires further optimization. Basaia et al. (2019) were able to utilize deep neural networks and a single MRI to develop an automated classification system that performed with promising accuracy without having to rely on a large number of scans. The diagnostic process becomes easier to explain and it takes fewer resources. In paper[2] the authors were able to obtain promising results in distinguishing healthy controls from moderate cognitive impairment (MCI) and Alzheimer's disease by using a deep learning framework that was able to record the subtle anatomical changes Alzheimer's Their research also pointed out datasets to improve the generalizability of the model. Amoroso et al. (2018) depended on deep learning to distinguish the onset of Alzheimer's in MCI patients. The results of this international challenge demonstrated how deep learning models can be used for early diagnosis, one of the most important aspects of timely intervention and treatment. The authors' participation in the CAD Dementia challenge proved the utility of their deep learning approach in clinical use, demonstrating its ability to reliably predict the onset of Alzheimer's disease [3] while highlighting that further research is needed to make the model more robust over a variety of populations. In their study on 3D brain MRI classification using residual and plain CNNs, Korolev et al (2017) demonstrated that residual networks outperform plain CNNs, because of their ability to work well even when the depth of the network increases. The analysis of their results showed that residual networks allow addressing the vanishing gradient problem and, therefore, go deeper into the network, discovering more complex structures in the MRI data. This essentially enhanced the reliability of Alzheimer's disease classification [2].

For example, at the end of his summary of algorithms – most of which demonstrated impressive results – Al-Shoukry et al. (2020) discuss possible future improvements and list already existing prospective avenues for research. Similarly, another

systematic review of state-of-the-art deep learning algorithms for detecting Alzheimer's by Al-Shoukry, Raiham, and Gil (2020) summarizes the algorithms' performance while also pointing out issues relevant to future development – such as points for future research, the lack of data, the absence of interpretable models, and the need for standardized evaluation procedures [3].

With how these models make their decisions still largely unknown, the authors call for AI in the field to be made Explainable yet again. Another review appeared last year (2019) which described both from the paper Liu et al. 's team have written. It deals with how deep learning methods might be used on the problem of Alzheimer's Disease diagnosis and brings out possible future study directions as well as external advantages or disadvantages for each way. The review also drew attention to the potential of deep learning to improve diagnostic accuracy, and the necessity of large, representative datasets being employed in training such models [8]. Tong et al. (2024) investigate the use of spatial context convolutional neural networks (CNNs) for the early diagnosis of Alzheimer's disease [9].

Published in the Journal of Supercomputing, their study explores how incorporating spatial context into CNNs improves the accuracy and effectiveness of early Alzheimer's diagnosis by leveraging spatial relationships in brain MRI scans. In the paper [10] the author wrote a review about the diagnosis of Alzheimer's, a severe illness, using deep learning methods. In this article, the authors provided an overview of the latest developments in this field and described the main challenges. For example, they listed the requirements of large-scale annotated datasets, as well as the integration of clinical data. Moreover, they described different CNN architectures that are used for the diagnosis of Alzheimer's, as well as the corresponding results, so the readers will understand why deep learning is a good match for the medical diagnosis problem Using a multimodal learning method developed by the authors of the paper [12], Vu et al. (2017), the aim was to increase learning from a combination of multiple modalities – for example, images and text data – using CNNs combined with sparse autoencoders, which have the potential to increase feature extraction and representation quality. The ensemble approach of Wang et al. on mild cognitive impairment and Alzheimer's disease diagnosis using 3D densely connected convolutional networks (2019) improves the identification of the presence of disease at the early stage of mild cognitive impairment. They establish different network models to obtain enhanced diagnostic accuracy and robustness. The effectiveness of this advanced architecture can be proven by the ability to distinguish cognitive stages [13]. In 2019, Huang et al trained a multimodal 3D convolutional neural network for Alzheimer's disease using multiple different types of imaging modalities for diagnosis. By using some of the latest advances in deep learning, the authors are improving the detection and classification of Alzheimer's disease based on MRI data. One of the key findings of their work involves more robust diagnosis using data from multiple sources [14]. The deep structure VGG-16 and VGG-19 design was to begin with proposed by Simonyan and Zisserman in 2014. Since then, their structures with fewer parameters and richer data have been widely used in the field of deep learning in identifying images. The classification of Alzheimer's disease is efficient in medical imaging [26] Szegedy et al. (2016) used this architecture called Inception V3, which employs a series of inception modules to extract multi-scale image features. In the paper [27] For MRI images, the same configuration has been shown to capture complex features with significant success in the detection of Alzheimer's disease. Another important model is EfficientNet-B3 which is a scalable CNN architecture that strikes a good balance between performance and computational cost, introduced by Tan and Le in 2019. It is a notable alternative to traditional deep learning models and was found that it achieved promising results for Alzheimer's disease classification [20] In this way, the multi-method assessment strategy employed by Tondelli et al (2018) enables the identification of neural correlates in AD, also in patients with mild cognitive impairment, and in those with anosognosia. The study's goal is to 'contribute to the literature on the brain mechanisms underlying the unawareness of cognitive deficits in AD and mild cognitive impairment', a process that is unpicking how, and to what extent, the brain leads to, or facilitates, anosognosia happening [30].

III. Problem statement

The goal is to detect Alzheimer's disease using deep learning, but it is a difficult task, and there are several challenging issues to be overcome to achieve early diagnosis The key to successful intervention is to build a robust system, the data need to be comprehensive for a model to accurately identify early signs in MRI scans. Importantly, the data need to be 'de-noised' to overcome the heterogeneity issue, which is related to imaging protocols and the diversity of patient demographics. In this way, the model will be able to perform better when transferring from a commonly used dataset to other patient populations.

Model overfitting is another major concern, particularly when training on small or imbalanced datasets. To mitigate this, the study will make use of transfer learning in addition to well-tested fine-tuned architectures such as EfficientNet-B3 and Inception V3, which are designed to enhance model generalization. It will be measured in terms of how well it performs against key metrics for any diagnostic tool, namely sensitivity and specificity. In terms of classification, sensitivity (also known as the true positive rate) can be defined as:

$$\text{Sensitivity} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The above formula helps ensure that the model accurately identifies individuals with Alzheimer's disease.

The paper also proposes a deep learning model for enhancing the diagnostic precision of Alzheimer's disease through imaging data from MRI, and clinical and genomic data. This is especially true in the case when hereditary factors, together with environmental ones, are firmly linked. It suggests that

a deep learning model could harness varied data, including clinical and genomic data, for a better understanding of the diseases by deep learning than a single MRI scan. Such a model would greatly enhance deep learning for the diagnosis of Alzheimer's, thereby helping in the early identification of the condition and reducing the societal cost of the disease.

IV. METHODOLOGY

A. Data Collection

Collecting an exhaustive collection of MRI scans is the first step to identifying the early Alzheimer's Disease(AD)stage. MRI scans come from healthy controls and AD patients with a wide range of brain images. The diversity of the dataset is essential in helping the model learn minor differences associated with AD. This is achieved by exposing the model to a diverse range of brain structures and conditions, such as seizures. AD generally develops through four stages: non-demented, very mildly demented, mildly demented, and moderately demented. For this supervised learning, diagnostic labels for each stage of the disease are assigned to each MRI image, helping the model learn AD diagnosis and the relations between certain MRI features and the presence or absence of AD in a particular phase, to enhance its ability to pinpoint and categorize the illness accurately.

B. Segregation of Data

The data must be segmented into Training, Validation, and testing to train a reliable and generalizable machine learning model.

1) Training Set: The model is trained on this subset, which is typically 80 percent of the entire dataset. Because classes are already represented in similar numbers (eg, an equal number of MRI scans from people with Alzheimer's and healthy controls), the likelihood of the machine leaning towards a particular class is minimized. The model is far more likely to learn the unique features of AD because the training set is well-balanced.

2) Validation Set: 10% of the data, which has been retained in the training phase. This allows us to tune the hyperparameters and models periodically. It is important to not overfit and selection of one's best model configuration.

3) Test Set: This has another 10% of the dataset. The test set is kept idol until the model is executed to make independent judgments of it and get an unbiased estimate of how well it works and generalizes new data.

C. Data Preprocessing

Preprocessing is a vital step that serves as the first processing stage before extracting useful features from unprocessed MRI data to train machine learning models. The ensuing steps include:

1) Normalization: MRI scan pixel values are normalized to give a uniform range. Normally this is 0-1. This standardization is important to make the different elements of the data set unified enough to allow training of the model. The normalizing formula is below:

$$I_{\text{norm}} = \frac{I - I_{\min}}{I_{\max} - I_{\min}}$$

$$I = Original\ pixel\ value$$

$$I_{\min} = minimum\ pixel\ value$$

$$I_{\max} = maximum\ pixel\ value$$

2) Data Augmentation: These make it possible to mitigate the limitations of having a relatively small medical dataset and avoid overfitting, since millions of additional training samples are generated from the original images through data augmentation techniques, such as:
1) Rotation: Different orientations are simulated by rotation in different angles. Rotation by an angle represented as

$$I_{\text{rot}}(\theta) = R(\theta).I$$

2) Flipping: Horizontal and vertical flipping is performed to simulate different perspectives:

$$I_{\text{flip}} = flip(I)$$

3) Zooming: Zooming in and out of images adjusts the scale, which is represented as::

$$I_{\text{zoom}} = S(a).I$$

4) Shifting: Zooming in and out of images adjusts the scale, which is represented as::

$$I_{\text{shift}}(\delta x, \delta y) = T(\delta x, \delta y).I$$

3) Rescaling and Resizing: MRI scans are often captured at differing dimensions/resolutions, and so must be rescaled (resized) so that the dimensions of input are the same. In addition, images are typically resized to a fixed dimension (for example, 224x224 pixels), which is the input requirement of most deep learning models, to have consistency across all images.

4) Noise Reduction: An MRI scan might contain noise, obscuring some important features. Gaussian blurring and median filtering might cut down noise, and improve the

image, and hence the performance accuracy of the model in identifying the features it thinks are relevant to AD.

5) Image Cropping: A region from the data frame inside the image from the MRI scan the proposed model works by cropping an image on certain parts. For instance, there could be certain regions in the hippocampus of the brain that are more informative of whether the patient has AD or not, and cropping those regions out of the MRI scans cab be done to reduce the compute intensity and ask the model to focus its attention on the right region.
6) Histogram equalization: It increases the level of contrast between intensity values of pixels in an image by shifting them to a more equal distribution, which helps the model detect subtle changes in the volume of folding in the brain associated with AD.

D. Selection of models

1) VGG16: The VGG16 model is an artificially deep network known for its depth and simplicity and trained to recognize patterns in colored photographs of common objects and its specifications. It has 13 convolutional layers with 2×2 max-pooling and 16×16 kernel sizes and three dense, or fully connected, layers. The small 3×3 filters and 2×2 max-pooling layers help to capture details and hierarchical features in the MRI scans (eg, subtle structural changes associated with the cause of Alzheimer's) in a deep network.

2) VGG19: Starting with the VGG16 model, the Addition of three more convolutional layers to VGG16 makes the network 19 layers deep. These extra layers in VGG19 compared to VGG16 enable it to learn more fine-grained features from the MRI scans. The increased feature-extraction capacity allowed VGG19 to better distinguish early Alzheimer's Disease from healthy subjects.

3) Inception V3: Inception V3 uses inception modules in the same layer with different filter sizes (1x1, 3x3, 5x5), which provides a multi-scale way to extract features of different sizes and patterns from MRI scans. Features of different sizes and complexities greatly benefit medical imaging.

4) EfficientNet-B3: EfficientNet-B3 uses a compound scaling scheme that scales the depth, width, and resolution of the network uniformly, which is good for finding a robust accuracy-computation trade-off. The requirement is for the model to perform well with many images but fewer parameters to accelerate the process With a highly accurate deep-learning model holding good.

E. Model Training and Hyperparameter Tuning

1) Training Process: Lastly, the training set of MRI scans is passed through selected network architectures, and the prediction of the model is compared with the ground truth labels via some function (here, the cross-entropy function), to obtain a loss value. The model weights are then tweaked via backpropagation, along the gradient of the loss function, back through the network until it reaches the weights, and tweaked again. This repeats iteratively over many, many thousands of iterations – until some loss value (here, that of the number of errors in the training example) to be minimized, is as small as possible. In reality, the process runs through many 'epochs'. Here, the training of the models with 30 epoch values is finalized by evaluation validation loss value. The dataset is broken into many thousands of small sections of, around 200 examples that are known as 'batches', and each of those batches is passed through the network en masse, followed by the whole process again, until convergence.

2) Hyperparameter Tuning: Hyperparameter tuning, on the other hand, hyperparameters such as the learning rate, batch size, number of epochs, regularization techniques, and so on. Learning rate is the parameter that decides the step size of the weight updates; batch size is the parameter that decides the number of samples to take at a time together to calculate the weight updates and both of them, put together, reduces the chances of a wrong or low accuracy. Regularisation techniques such as dropout and L2 regularization decrease the chances of over-fitting. Dropout randomly drops some random neurons while training it. L2 regularization is a penalty based on the weight magnitudes and it works like magic on the mishap of an over-fitting because it curbs the model from over-fitting.

F. Custom Modifications

1) Transfer Learning: Transfer learning would save us time by starting with weights from models that have been trained on large datasets, and transferring those weights into the task, even if it is different, because some features learned from the previous model will be also relevant to this task. The application of a process called fine-tuning is included which adds a bit of complexity to the transfer-learning process. It involves changing the last layer of the model to fit it better to the Alzheimer's Disease MRI dataset, helping us to be even better at recognizing relevant features.

2) Ensemble Learning: Ensemble learning is a type of prediction that uses many different types of models and then produces a final decision based on the predictions of the individual models. In this case, weighted voting is used to aggregate the predictions of models like VGG16, VGG19, Inception V3, and EfficientNet-B3. The final prediction is:

$$\mathrm{P} = \frac{1}{N} \sum_{i=1}^{N} w_i . p_i$$

3) Grad-CAM for Interpretability: Grad-CAM (Gradient-weighted Class Activation Mapping), is a way of visualizing the parts of an MRI scan that are most responsible for the model's predictions. By amplifying these parts, Grad-CAM

forces us to see the decision-making of the model and lets us check that the model is focusing on the relevant Alzheimer's brain regions.

G. Metrics

Several principal metrics and techniques are needed for the evaluation of Alzheimer's disease detection. The accuracy will return the percentage of correctly classified MRI scans; it, however, has unintuitive behavior in the case of class imbalance. Precisions and recall provide more insights: precision is the measure of the reliability of positive predictions, and recall goes to a model's ability to detect true positive cases. The F1 score is a harmonic mean of precision and recall; thus, it can provide a well-balanced performance metric. Besides, the ROC curve and AUC compute the trade-off between the true positive rate versus the false positive rate; the latter should also lie close to 1 for high performance. It has utilized cross-validation techniques, including k-fold cross-validation, to avoid overfitting or underfitting and ensure that the model is robust in its evaluation; this has yielded the most realistic measure of model performance. All of these metrics and techniques are a very comprehensive framework of effectiveness evaluation in models for the detection of Alzheimer's disease.

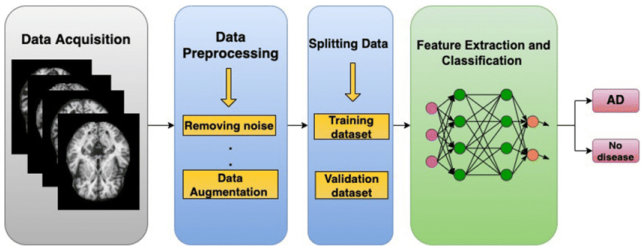

Fig. 1: Alzheimer's disease detection workflow

V. EXPERIMENT

(a).In this experiment, the comparison of the performance of VGG-16, VGG-19, Inception V3, and EfficientNet-B3 models, as well as a standard CNN that was trained for 85 epochs. The computation of the confusion matrices for each of the models which were used to find TP, TN, FP, and FN. This information was utilized to calculate accuracy, specificity, sensitivity, and precision. Moreover, the F1 score was measured to provide an outline of the execution of each model. The VGG-16, VGG-19, Inception V3, and EfficientNet-B3 models were trained for 30 epochs, with identical batch size, and learning rates throughout the training process, to ensure a level playing field. The results show how each of the models compares to each other, in terms of accuracy and adherence to domain inference.

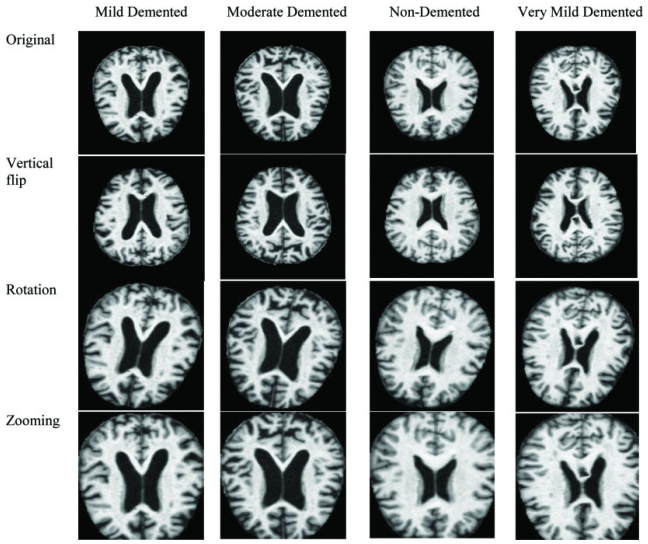

Fig. 2: Dataset Representation

A. Dataset

The dataset used for model development is the "Alzheimer's Dataset 4 Class of Images" on Kaggle. It is composed of 6,400 MRI images annotated into four different classes: Mild Dementia, Moderate Dementia, Non-Demented, and Very Mild Dementia. Every image in this dataset is standardized to have a resolution of 224 × 224 pixels and be composed of three color channels (RGB) so it can feed the deep learning models used in the study. As a part of preprocessing, resizing was done to 224 x 224 x 3, and all images were normalized so that consistent data was fed to all models. The applied preprocessing techniques were resizing and normalization; data augmentation was also carried out to enhance model performance since there may be class imbalances. The total dataset was then split into three subsets: training, validation, and test sets. The training set is for model learning, and the validation set helps fine-tune parameters to avoid overfitting the model. In the end, a subset is kept reserved for estimating model performance. It will result in a robust and reliable result. Using this dataset enables a thorough evaluation of the models' detection and categorization of Alzheimer's Disease, as well as a fair comparison of such performance across different deep learning architectures.

B. Evaluation Details

This section will provide an overview of the evaluation of the proposed methodology. So the evaluation consists of splitting data and comparison of results. it consists of splitting data and comparison of results.

1) Model Training Details: The models VGG-16, VGG-19, Inception V3, and EfficientNet-B3 are carefully trained for 30 epochs each, while a standard CNN underwent 85 epochs of training in the experiments. This meant 80% of the images were used in training and validation,

where the training set would train the models, while the validation subset was used to tune the hyperparameters of the models so that it didn't result in overfitting. The remaining 20% was then left solely for testing, therefore obtaining unbiased performance measures on unknown data. All images in the dataset were resized to 224 × 224 pixels with three color channels (RGB), compatible with the input requirement for both models. Data augmentation techniques were designed, including rotation and flipping, accompanied by normalization for better training of the models and improving their generalization ability. This provided absolute assurance that the effective training of models would follow, further providing a well-balanced and robust evaluation of their performance across different stages of Alzheimer's disease.

C. Evaluation Metrics

1) Training Accuracy: Quite clearly, the training accuracy is positive across the epochs, which means evident learning by the model on the training data. Nevertheless, one's increased performance does hint toward an increasing familiarity of the model with the training set. Performance does hint toward an increasing familiarity of the model with the training set. Validation Accuracy: Validation Accuracy: That's much lower than the training accuracy and the difference is how badly it has generalized to unseen data. The difference between train and validation accuracies points towards a possibility of overfitting. Alternatively, a lack of proper techniques to improve generalization could also be a factor.

2) Area Under the ROC Curve:
   (a) Training AUC: This is high; thus, it means it's good to separate classes according to the training data. This metric gives insight into how well the model is performing during training in identifying and ranking positive and negative classes.

   (b) Validation AUC: The validation AUC itself is always very volatile, but in general, it remains lower than the training AUC. This fluctuation could be an indication that, though the model is good in training, it is not that robust in the distinction of the classes within the validation set. One region of change that can be taken note of from the values of AUC is the improvement in the capability of the show to generalize.

D. F1 Score

1) Training F1 score: The training F1 score will be such that it will be middling concerning balancing precision and recall, showing that it is reasonable at both metrics concerning the training data. It finds the positive cases with minimizing false positives.

2) Validation F1 Score: The validation F1 score is far below that of the training F1 score. The drop suggests possible challenges in creating the same balance between precision and recall on the validation data. That would tell that the performance may not be so stable or reliable if applied to unseen data.

Although it performs quite well at train time, reaching good accuracy, AUC, and F1 score with the model, there exist large gaps in train and validation metrics that could be indicative of potential overfitting or issues in generalization. There is, therefore, a need for further tuning for the model's architecture, regularization techniques, or data augmentation and class-balancing strategies. It can achieve better generalization capability by augmenting diversity in the training set data, balancing it further, improving the model against overfitting, and gaining stability on unseen data.

E. Compared models

At epoch 30, a comparison of these models, with their corresponding metrics is shown below: Overall, the standard CNN performs strongly as it records the best metrics. This is evident in its weighted precision, recall, and f1-score of 0.98. They are showing that it has performed strongly in this task. However, VGG16 performed the best among the models scoring the highest metrics at 0.99. Which is strong, showing that it has been able to balance out precision and recall well, making it the best model at epoch 30. Performing second-best is EfficientNet-B3 with respectable metrics of 0.96. Showing that it has performed strong, though less than the leading model VGG16. Lastly, VGG19 has performed competitively with metrics of 0.94, though also less than the leading model. Hence, showing that it is also a solid model. On the flip side, Inception V3 performed least effectively with metrics of 0.67,0.62, and 0.62 respectively for its weight f1-score. Therefore, comparing the metrics from the different models at epoch 30, the leading models are VGG16 and EfficientNet-B3. With the VGG16 being the best model with better overall metrics.

Table I- Compared Models Weighted Results

Model	Epoch	Weighted Precision	Weighted Recall	Weighted F1 score
Standard CNN	85	0.86	0.83	0.84
VGG 16	30	0.99	0.99	0.99
VGG 19	30	0.96	0.96	0.96
Inception V3	30	0.67	0.67	0.67
EfficientNet-B3	30	0.94	0.94	0.94

VI. RESULTS

The testing of five different CNN architectures – Standard CNN, VGG16, VGG19, Inception V3, and EfficientNet-B3 – where each of them was compared based on their accuracy on

training data, accuracy on test data, loss on the training data, and loss on the test data. The accuracy of each model during their training was observed to understand their capability of generalization from the training data to the unseen test data. For each of the models, some callback functions and early stopping mechanisms were employed to train all the models for a suitable time. Owing to the gradient descent used here, the loss function continually tries to decrease the loss score. For the majority of the models, early stopping was done at around 30 epochs. The Standard CNN, however, performed well only after 85 epochs, which is probably the reason for its aforementioned characteristic of performance.

1) Standard CNN:

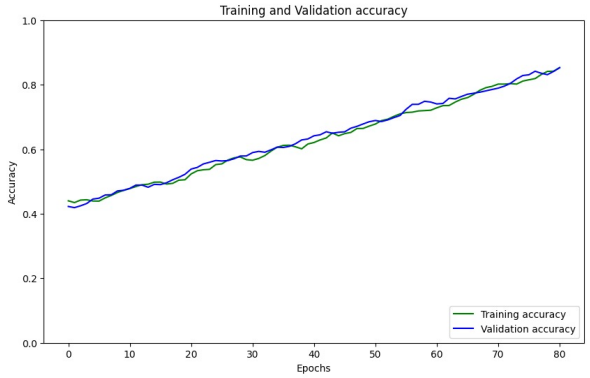

Fig. 3: Accuracy Graph for CNN

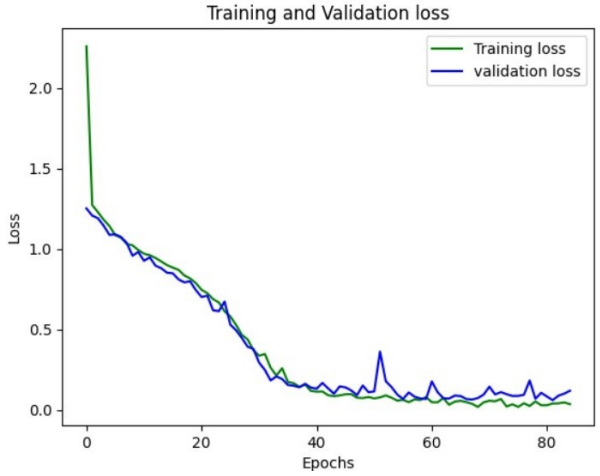

Fig. 4: Loss Graph for CNN

A more balanced result is produced by the Standard CNN, which gives training accuracy of 86.34 percent and test accuracy of 86.45 percent, along with training loss of 0.189 and test loss of 0.234. The model displays a moderate accuracy across both the training and test datasets. The difference between training loss and test loss on the other hand implies a margin of improvement needed for generalization with the model choice. The performance of the model on unseen data can be improved by avoiding this imbalance, and the training of 85 epochs, since the model optimized the training itself for the training set for more than two times the training time compared to the other three models. In effect, this led it to become over-focused on the training set, as it achieved very high training loss but stagnated on the validation performance.

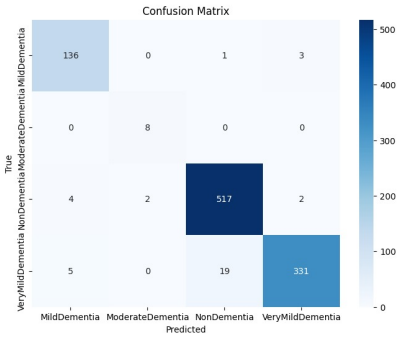

Fig. 5: Confusion Matrix for CNN

2) VGG 16:

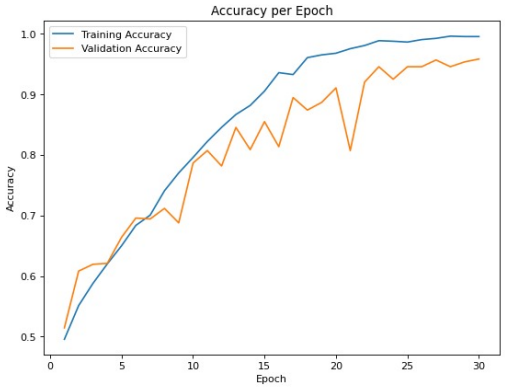

Fig. 6: Accuracy Graph for VGG16

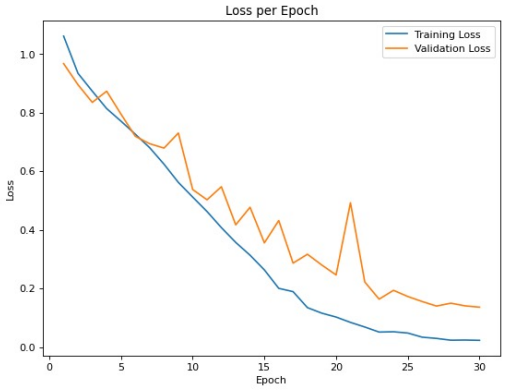

Fig. 7: Loss Graph for VGG16

The VGG16 model performance is given by a training accuracy of 98.70% and a test accuracy of 98.65%, while the training loss was 0.044 and the test loss was 0.445. This shows that the model has very good performance on the training as well as on the test datasets. Since the training loss was a bit low, it signifies that the model has learned from the training dataset properly and the test loss (the loss on the test dataset) is very little, hence showing that this model is one of the strong classification models. The consistency in the high accuracy over the training and test datasets proves that this model is reliable and would show good performance while applying it for real predictions. The early stopping feature of this implementation helped to stop the training at the optimal point, thus hindering any overfitting to the model and keeping it robust to new and general data.

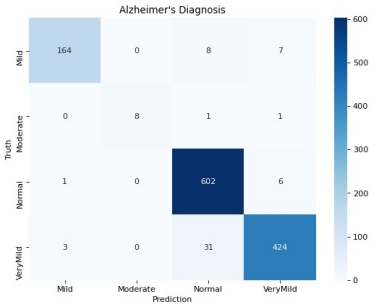

Fig. 8: Confusion Matrix for VGG16

3) VGG 19:

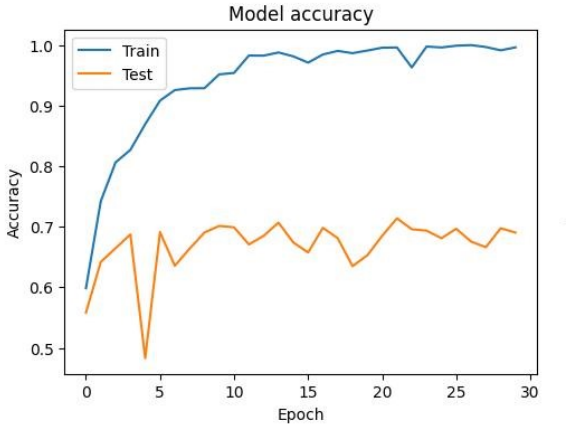

Fig. 9: Accuracy Graph for VGG19

The VGG19 model has a training accuracy of 99.50%, test accuracy of 70.32%, training loss of 0.136, and test loss of 0.85. We can see that while VGG19 has performed excellently on the training data, the test accuracy has significantly decreased, indicating massive generalization issues. The fact that training and test losses are much higher than that of VGG16 demonstrates that VGG19 is not generalizing at all. Although

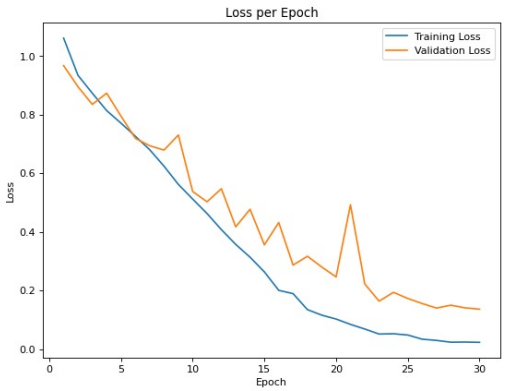

Fig. 10: Loss Graph for VGG19

VGG19 has learned the training data extremely accurately, when it comes to new, unseen data, it completely fails to perform at all.

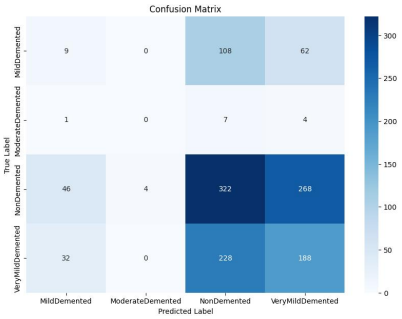

Fig. 11: Confusion Matrix for VGG19

4) Inception V3:

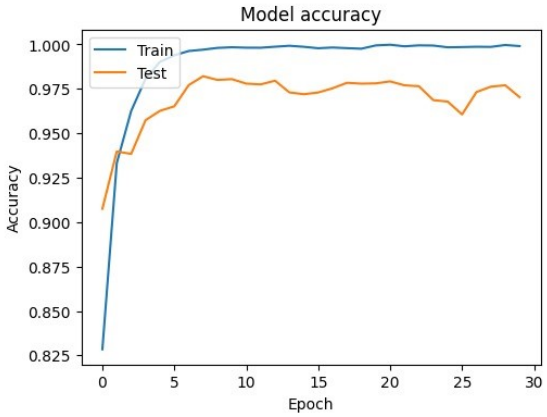

Fig. 12: Accuracy Graph for Inception V3

It's impressive that Inception V3 has a training accuracy of 99.90% and a test accuracy of 96.10%. However, its train loss is 0.29% and its test loss is 0.55%, which are high, as compared to models like VGG16 and EfficientNet-B3. While

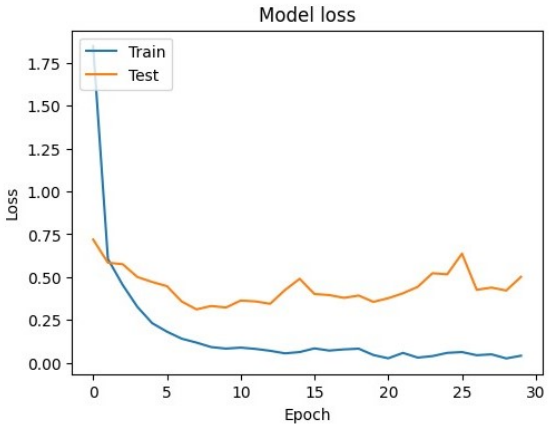

Fig. 13: Loss Graph for Inception V3

Inception V3 is a high-performance model from the accuracy perspective, high loss that it shows suggests that it may not be the best model in learning in comparison with these other models. Higher test loss, in particular, means that although it performs well in accuracy, it continues to have a possibility for further improvement in loss minimization.

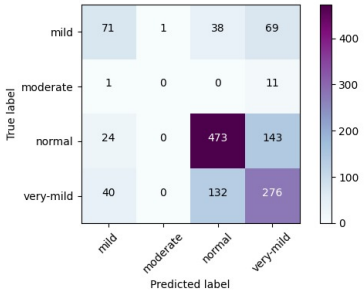

Fig. 14: Confusion Matrix for Inception V3

5) EfficientNet -B3:

Fig. 15: Accuracy Graph for EfficientNet-B3

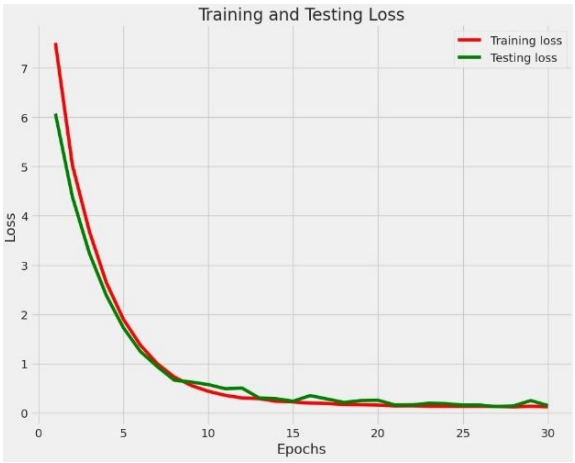

Fig. 16: Loss Graph for EfficientNet-B3

The EfficientNet-B3 achieves a high train accuracy of 99.68 percent and a high test accuracy of 96.21 percent, meanwhile, it has a low train loss of 0.084 and a low test loss of 0.129. The efficiency of this model is retained as it has a 96.21 percent accuracy on the test, while not overfitting and having a low loss. This indicates it learns efficiently and generalizes well. Thus, EfficientNet-B3 is ranked one of the best-performing models in this comparison.

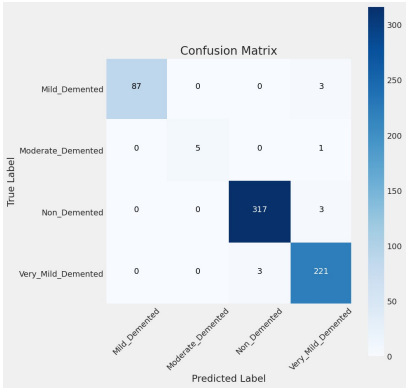

Fig. 17: Confusion Matrix for EfficientNet-B3

VII. Discussions

A trained model of these CNNs will converge on an answer much sooner for an Alzheimer's detection task compared to standard CNN. For example, VGG16, VGG19, Inception V3, and EfficientNet-B3 can reach good performance with about 30 epochs, while standard CNN may need over 85 epochs to learn and capture the tiny patterns in highly complex medical imaging data. With a longer training period, standard CNN can learn much more in-depth features; while other models converged in a much shorter time. Callbacks can be used during CNN training to control the training and address model overfitting. Callbacks monitor training for certain metrics in real-time and are used for early stopping, a technique where training ceases when an indicator of

performance in performing the given task stops improving (for example, validation loss or accuracy no longer improves for a specified number of epochs) to ensure that the models do not overfit to the training data, an issue that can limit their ability to generalize to new data. Callbacks can also be used during training to adjust learning rates and save models such as model checkpoints.

Table II- Accuracy table for Proposed Model

Model	Train Accuracy	Test Accuracy
Standard CNN	88.56%	88.42%
VGG 16	98.70%	98.65%
VGG 19	99.50%	70.32%
Inception V3	99.90%	96.10%
EfficientNet-B3	99.68%	96.21%

Table III- Loss table for Proposed Model

Model	Train Loss	Test Loss
Standard CNN	0.189	0.234
VGG 16	0.044	0.445
VGG 19	0.136	0.85
Inception V3	0.29	0.55
EfficientNet-B3	0.084	0.129

VIII. Conclusion

Finally, This paper demonstrates the extensive ability of various efficient deep CNN architecture to detect AD at an early stage. Deep CNN-based computational tools such as VGG16,VGG19, Inception V3 and EfficientNet-B3, are able to detect the minor functional and structural brain changes that are profoundly related to AD. This is because deep CNN models could learn the complex patterns from MRI data and facilitate a transfer learning strategy that exploits the highly structured nature of this data efficiently. This is promising for faster and more objective assessment of patients, leading to prompt early intervention.

Furthermore, these models are interpretable, which could help make them clinically adaptable, and also health professionals might feel comfortable using such tools to help improve outcomes for their patients. By making these models' logic visible, we might be able to find ways to apply the technology in the real world and improve diagnostic accuracy. Overall, the results show how deep learning technologies have the potential to transform Alzheimer's diagnosis, leading to improved outcomes for those at risk of or living with the disease.

IX. Future Works

Future works in Alzheimer's disease detection by deep learning models can take several innovative directions to make methods more diagnostic and clinically useful. One area involves the integration of multi-modal data, in particular, from the various imaging modalities, such as MRI, PET, or CT scans, with clinical data, including genetic information, patient history, and cognitive tests. It uses the strengths of all data types together to build up a fuller, more finely tuned understanding of the process of disease. In addition, the enrichment of data captured by devices and sensors can aid the continuance of monitoring patients' physiological and behavioral trends, hence facilitating timely and more accurate detection of their diseases. Noticeably, deep learning architectures, especially the recent ones, including models of transformers and graph neural networks, hold immense potential for performance improvement provided that rich complex temporal and spatial relationships within data are captured. Third, such models may deal more effectively with heterogeneous data sources, offering a more complete and deep pathology analysis. More importantly, the clinical adoption of AI will not be realized without developing explainable AI models for elucidating the decision-making processes of deep learning systems and thus building trust and transparency between clinicians and these deep learning systems. The higher the interpretability of AI models, the more one can understand and validate the diagnostic recommendations for improving health outcomes. Research in these lines shall develop the field of Alzheimer's disease detection and supply new tools and methodologies to deal with this challenging, wide-ranging condition.

References

[1] M. Liu, D. Cheng, K. Wang, Y. Wang, and A. D. N. Initiative, "Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, pp. 295–308, 2018.

[2] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification," in *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pp. 835–838, IEEE, 2017.

[3] S. Al-Shoukry, T. H. Rassem, and N. M. Makbol, "Alzheimer's diseases detection by using deep learning algorithms: a mini-review," *IEEE Access*, vol. 8, pp. 77131–77141, 2020.

[4] Y. Çetin-Kaya and M. Kaya, "A novel ensemble framework for multi-classification of brain tumors using magnetic resonance imaging," *Diagnostics*, vol. 14, no. 4, p. 383, 2024.

[5] D. Sarwinda and A. Bustamam, "Detection of alzheimer's disease using advanced local binary pattern from hippocampus and whole brain of mr images," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 5051–5056, IEEE, 2016.

[6] U. Raghavendra, U. R. Acharya, and H. Adeli, "Artificial intelligence techniques for automated diagnosis of neurological disorders," *European neurology*, vol. 82, no. 1-3, pp. 41–64, 2020.

[7] S. A. Al-Majeed and M. S. Al-Tamimi, "Survey based study: Classification of patients with alzheimer's disease.," *Iraqi Journal of Science*, vol. 61, no. 11, 2020.

[8] G.-S. Fu, Y. Levin-Schwartz, Q.-H. Lin, and D. Zhang, "Machine learning for medical imaging," *Journal of healthcare engineering*, vol. 2019, 2019.

[9] Y. Tong, Z. Li, H. Huang, L. Gao, M. Xu, and Z. Hu, "Research of spatial context convolutional neural networks for early diagnosis of alzheimer's disease," *The Journal of Supercomputing*, vol. 80, no. 4, pp. 5279–5297, 2024.

[10] F. Zhang, B. Pan, P. Shao, P. Liu, S. Shen, P. Yao, R. X. Xu, A. D. N. Initiative, *et al.*, "A single model deep learning approach for alzheimer's disease diagnosis," *Neuroscience*, vol. 491, pp. 200–214, 2022.

[11] X. Bi, X. Zhao, H. Huang, D. Chen, and Y. Ma, "Functional brain network classification for alzheimer's disease detection with deep features and extreme learning machine," *Cognitive Computation*, vol. 12, pp. 513–527, 2020.

[12] T. D. Vu, H.-J. Yang, V. Q. Nguyen, A.-R. Oh, and M.-S. Kim, "Multimodal learning using convolution neural network and sparse autoencoder," in *2017 IEEE international conference on big data and smart computing (BigComp)*, pp. 309–312, IEEE, 2017.

[13] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, and X. Zhao, "Ensemble of 3d densely connected convolutional network for diagnosis of mild cognitive impairment and alzheimer's disease," *Neurocomputing*, vol. 333, pp. 145–156, 2019.

[14] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, and A. D. N. I. (ADNI), "Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network," *Frontiers in neuroscience*, vol. 13, p. 509, 2019.

[15] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline, "Classification of alzheimer disease on imaging modalities with deep cnns using cross-modal transfer learning," in *2018 IEEE 31st international symposium on computer-based medical systems (CBMS)*, pp. 345–350, IEEE, 2018.

[16] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[17] D. Agarwal, G. Marques, I. de la Torre-Díez, M. A. Franco Martin, B. García Zapiraín, and F. Martín Rodríguez, "Transfer learning for alzheimer's disease through neuroimaging biomarkers: a systematic review," *Sensors*, vol. 21, no. 21, p. 7259, 2021.

[18] W. R. PERDANI, R. MAGDALENA, and N. K. C. PRATIWI, "Deep learning untuk klasifikasi glaukoma dengan menggunakan arsitektur efficientnet," *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 10, no. 2, p. 322, 2022.

[19] S. Sharan, S. Kininmonth, U. V. Mehta, *et al.*, "Automated cnn based coral reef classification using image augmentation and deep learning," *International Journal of Engineering Intelligent Systems*, vol. 29, no. 4, pp. 253–261, 2021.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[21] A. W. Salehi, P. Baglat, B. B. Sharma, G. Gupta, and A. Upadhya, "A cnn model: earlier diagnosis and classification of alzheimer disease using mri," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 156–161, IEEE, 2020.

[22] R. M. Gemiralda and M. Marlaokta, "Efek neuroprotektor kunyit pada pasien alzheimer," *Jurnal Ilmu Keperawatan Jiwa*, vol. 2, no. 3, pp. 171–178, 2019.

[23] D. F. Santos, "Advancing automated diagnosis: Convolutional neural networks for alzheimer's disease classification through mri image processing," *Authorea Preprints*, 2023.

[24] T. Brosch, R. Tam, and A. D. N. Initiative, "Manifold learning of brain mris by deep learning," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pp. 633–640, Springer, 2013.

[25] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, *et al.*, "Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment," *Frontiers in neuroscience*, vol. 12, p. 777, 2018.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[28] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

[29] U. R. Acharya, S. L. Fernandes, J. E. WeiKoh, E. J. Ciaccio, M. K. M. Fabell, U. J. Tanik, V. Rajinikanth, and C. H. Yeong, "Automated detection of alzheimer's disease using brain mri images–a study with various feature extraction techniques," *Journal of medical systems*, vol. 43, pp. 1–14, 2019.

[30] M. Tondelli, A. M. Barbarulo, G. Vinceti, C. Vincenzi, A. Chiari, P. F. Nichelli, and G. Zamboni, "Neural correlates of anosognosia in alzheimer's disease and mild cognitive impairment: a multi-method assessment," *Frontiers in Behavioral Neuroscience*, vol. 12, p. 100, 2018.

[31] E.-G. Marwa, H. E.-D. Moustafa, F. Khalifa, H. Khater, and E. AbdElhalim, "An mri-based deep learning approach for accurate detection of alzheimer's disease," *Alexandria Engineering Journal*, vol. 63, pp. 211–221, 2023.

[32] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," *Artificial intelligence in medicine*, vol. 95, pp. 64–81, 2019.

[33] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, no. 1, p. 26286, 2016.

[34] N. Panic, E. Leoncini, G. de Belvis, W. Ricciardi, and S. Boccia, "Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (prisma) statement on the quality of published systematic review and meta-analyses," *PloS one*, vol. 8, no. 12, p. e83138, 2013.

[35] F. Razavi, M. J. Tarokh, and M. Alborzi, "An intelligent alzheimer's disease diagnosis method using unsupervised feature learning," *Journal of Big Data*, vol. 6, no. 1, p. 32, 2019.