Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(6), 1732-1751
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Vision Transformers (VIT's) for Early Identification of Alzheimer's Disease

Aayush Rajesh Jadhav

Department of Computer Science, Liverpool Johns Moores University, Liverpool, England
Email: aayushrj22@gmail.com

How to cite this paper: Jadhav, A.R. (2025) Vision Transformers (VIT's) for Early Identification of Alzheimer's Disease. *Open Journal of Applied Sciences*, **15**, 1732-1751.
https://doi.org/10.4236/ojapps.2025.156119

Received: April 17, 2025
Accepted: June 24, 2025
Published: June 27, 2025

Abstract

This thesis focuses on leveraging Image Processing, Computer Vision, Machine Learning, and Deep Learning, particularly the Vision Transformer (ViT) model, for early identification of Alzheimer's disease (AD), the most common form of dementia progressing from mild memory loss to significant impairment in daily interactions. Unlike prior studies that rely on conventional Convolutional or Recurrent Neural Networks, this research integrates ViT with a unique pre-processing strategy that includes sagittal-plane slicing, PCA-based feature reduction, and watershed segmentation to enhance regional interpretability. A key novelty lies in training on a clean, pre-augmented Kaggle dataset and testing on the real-world, imbalanced OASIS-3 dataset—demonstrating the model's ability to generalize from curated to noisy clinical data. The study details the ViT model's architecture, pretraining, and fine-tuning processes, employing a two-step training approach for efficient classification. The ViT-Base-Patch16-224 model undergoes pretraining on ImageNet-21k and fine-tuning on ImageNet 2012, incorporating data preprocessing with image partitioning, positional embeddings, and various transformations. The training process involves optimization with the AdamW optimizer, learning rate adjustments, exponential moving averages, and early stopping callbacks. Evaluation on Kaggle Alzheimer's and OASIS-3 datasets reveals promising performance, achieving 97.34% accuracy on Kaggle and 81.25% on OASIS-3. The confusion matrix and F1 score analyses highlight the model's strengths and areas for improvement, demonstrating high precision and recall for different classes, particularly in Alzheimer's disease identification. This study contributes to medical image analysis by emphasizing the ViT model's accuracy in classifying Alzheimer's cases, highlighting a novel framework adaptable to varied MRI datasets and offering interpretable, transferable results for clinical use.

Keywords

Alzheimer's Disease, Computer Tomography, Magnetic Resonance Imaging, Deep Neural Networks, Recurrent Neural Networks, Convolutional Neural Networks, Vision Transformers

1. Introduction

Alzheimer's disease, affecting 55.2 million people globally, poses a significant health challenge, particularly in low and middle-income countries. Its early detection is crucial for effective management, but current diagnostic challenges include the subtle onset of symptoms and lack of specific biomarkers. This study explores the use of Vision Transformers and Deep Learning in MRI brain image analysis for early detection, addressing the urgent need for improved diagnostic methods.

Given the remarkable effectiveness exhibited by existing frameworks, it becomes imperative to explore harnessing recent advancements in the field of computer vision to formulate an innovative framework aimed at the timely identification of Alzheimer's disease. This study also tries to focus on how we can leverage a pre-processed MRI dataset to extract relevant features to classify AD on a raw dataset like the OASIS-3 dataset.

The Vision Transformer (ViT) has emerged as a trendy and modern design within computer vision, as evidenced by the work of [1]. In a study by [2], explicitly focusing on MRI brain images, the findings of this study indicate that frameworks leveraging Vision Transformers (ViTs) and transformers exhibit superior performance compared to alternative models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

As demonstrated in the study by [3], image-processing techniques proved effective in identifying Alzheimer's Disease (AD) by analyzing multiple brain MRI scans. Incorporating the watershed method, picture segmentation, and thresholding techniques into the proposed framework presents potential avenues for enhancing accuracy in conjunction with the Vision Transformer (ViT) network.

The Vision Transformer (ViT) model architecture was introduced in a conference paper [4]. According to a study conducted by [5] transformers are already altering the field of computer vision, with tremendous development in research employing transformers in medical image processing, where most existing transformer-based algorithms may be adapted to medical imaging issues without major adjustments. Transformers have become very popular across a wide spectrum of CV tasks, as seen in [6] for segmentation, [4] for image classification, and [7] for object detection.

Numerous deep learning algorithms have been studied for identifying Alzheimer's disease (AD) using medical imaging data, as seen in [8]-[16]. The field of computer vision and deep learning has witnessed significant advancements in recent years, particularly with the introduction of the Vision Transformers (ViT) model architecture. ViT has emerged as a powerful tool for image classification tasks,

demonstrating remarkable performance by leveraging the self-attention mechanism of transformers, as seen in [16]-[21].

Vision Transformers (ViTs) present potential resolutions to many research deficiencies in computer vision and medical imaging like computational limitations [22], few shot learning [23], absence of interpretability [24]. The capacity to effectively generalize, manage extensive and diverse datasets, offer interpretability, and process images fast renders them invaluable for tackling various practical issues in multiple domains. This study emphasizes the need for advanced image processing and deep learning techniques, particularly Vision Transformers, to address these challenges and improve early detection and management of Alzheimer's Disease.

2. Materials

The dataset selected for the classification of Alzheimer's disease (AD) consists of two primary sources: OASIS-3 [25] and the Alzheimer's MRI Pre-processed Dataset [26].

The OASIS-3 dataset is a longitudinal dataset that includes MRI scans of 150 individuals aged 60 - 96 years. The dataset offers a diverse range of brain scans, including those of healthy individuals and those with varying degrees of AD, from mild to severe cases. The availability of longitudinal data allows for tracking disease progression over time. OASIS-3 provides a valuable resource for validating the VIT model. (**Figure 1**) shows the sagittal, coronal, and axial view of the Oasis-3 dataset once the MRI is sliced across different planes.

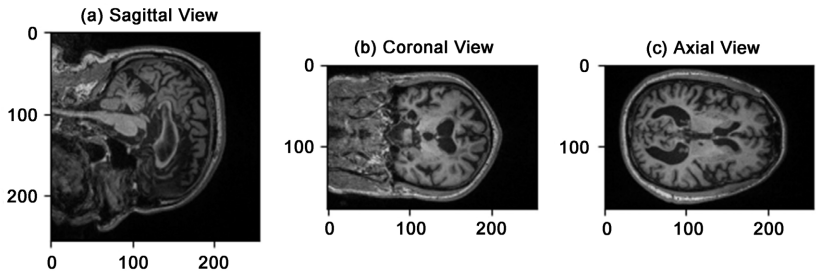

Figure 1. Sagittal, coronal, axial view of Oasis-3 Dataset.

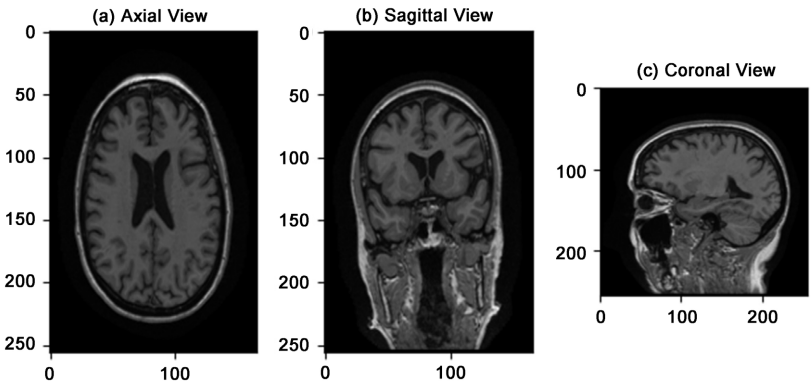

Figure 2. Controlled normal class (Oasis-3).

(Figures 2-4) provide an illustrative portrayal of the distinct categories of Alzheimer's disease (AD) as observed within the OASIS-3 dataset.

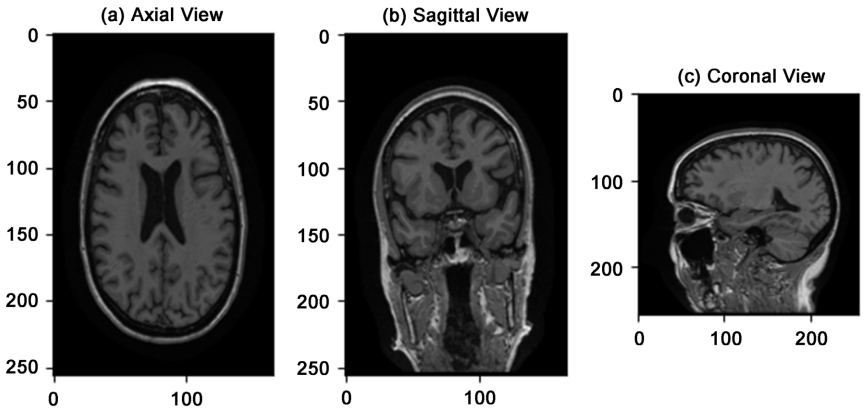

Figure 3. Mild cognitive impairment class (Oasis-3).

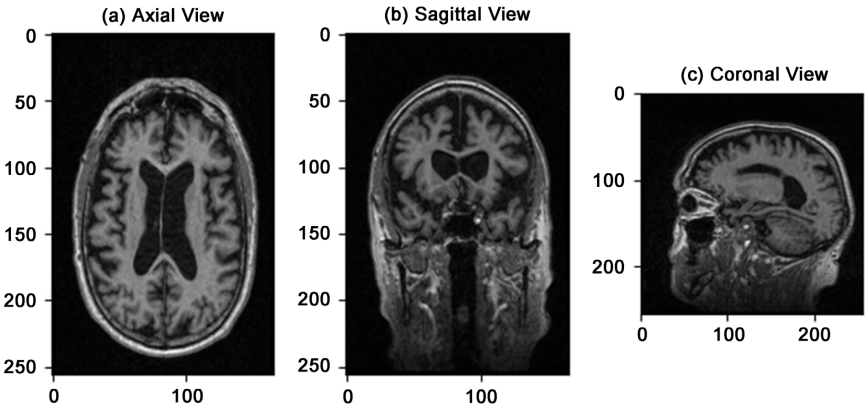

Figure 4. Alzheimer's disease class (Oasis-3).

The Alzheimer's MRI Pre-Processed Dataset was carefully collected and pre-processed from various websites, hospitals, and public repositories. It consists of pre-labeled and pre-processed MRI brain scan images. The dataset is valuable for training and testing both the ViT models in the proposed framework, allowing for a comprehensive evaluation of the framework's performance. (Figure 5) gives an illustrative example of the different classes of dementia that the dataset offers.

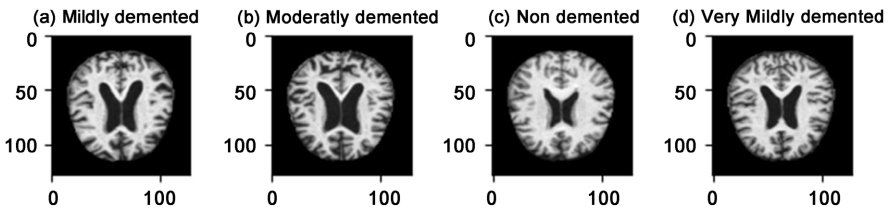

Figure 5. Different classes in Kaggle Dataset.

The chosen datasets provide a rich and varied collection of MRI scans and de-

mographic information, making them well-suited for training, validating, and testing the proposed framework. This comprehensive approach enhances the potential of the framework to accurately detect Alzheimer's disease at an early stage, leading to improved diagnostic capabilities and potential advancements in AD research and treatment.

3. Methods

This study aims to develop a reliable framework for detecting Alzheimer's disease (AD) in brain MRI scans. The framework will be accomplished by leveraging these two cutting-edge technologies.

The study will use cutting-edge strategies, procedures, and methodologies, in addition to multiple datasets containing MRI scans. The framework that has been suggested is organized into three primary layers, which are referred to as the Pre-Processing Layer, the Model Building Layer, and the Model Evaluation Layer, respectively. Each layer contributes to the overall process of AD detection and plays an essential part in the overall success of the framework's implementation.

3.1. Data Pre-Processing

The OASIS-3 dataset consists of raw, un-processed NIfTI files. NIfTI (Neuroimaging Informatics Technology Initiative) files, colloquially referred to as .nii files, have gained significant traction as a prevalent format for the storage of neuroimaging data about brain scans, including magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI) scans.

The 3D volumetric MRI scans from the OASIS-3 dataset were converted into 2D sagittal views by extracting slices along the sagittal plane at a specific z-coordinate ($z = -22$) as mentioned in [11], thereby rendering them suitable for analysis and interpretation. In addition, it is imperative to resize the images as mentioned above to dimensions of 128×128 to facilitate their utilization in the training process of the Vision Transformer.

The subsequent step involves the application of an image segmentation technique on the individual slices of the three-dimensional images obtained from the OASIS-3 dataset and the Kaggle dataset, explicitly employing the watershed algorithm.

(**Figure 6**) depicts the watershed algorithm used over the dataset to segment the image and bring out the regions of focus. Six types of watershed algorithms were used, and the binary thresholding algorithm was selected.

Watershed algorithms employ binary thresholding to simplify segmentation. By transforming the grayscale image to a binary image, each pixel is assigned to the foreground (object), or background (non-object) based on a threshold value. Implementing a threshold is simple and effective. Hierarchical watershed techniques demand more marker identification and region merging. Complex image structures can be handled by these techniques, but they require more processing and parameter optimization. Basic picture segmentation using binary threshold-

ing works well for items with obvious boundaries and background contrast. The simplicity, computational economy, and capacity to deliver good results in varied settings make this technique popular in image processing. Picture segmentation's watershed algorithm depends on picture properties and task requirements.

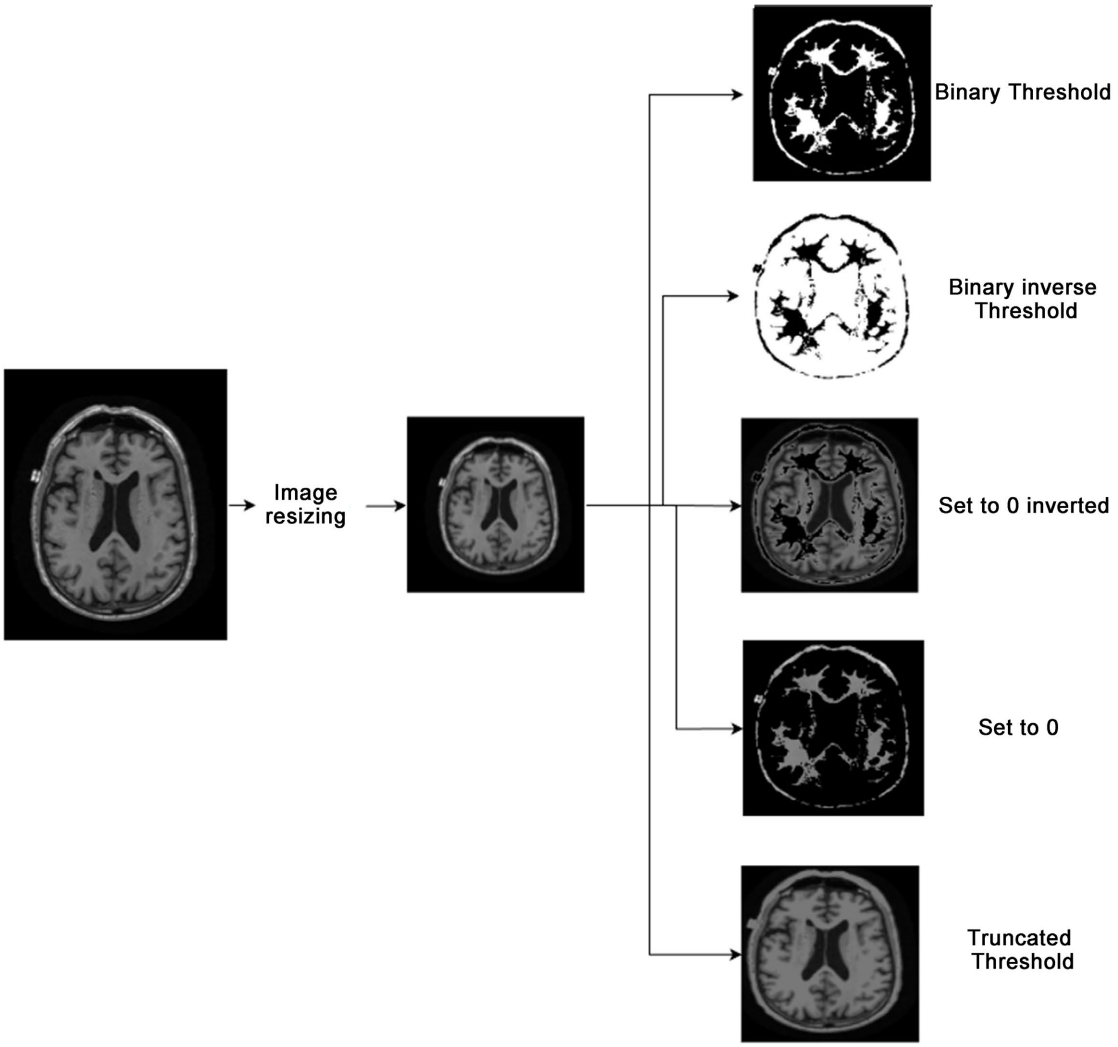

Figure 6. Watershed algorithms applied.

3.2. Model Building

The model building layer consists of the ViT model (Figure 7). The ViT model will be used to classify AD using brain MRI scans. This model will be trained and tested on the Kaggle dataset and validated over the OASIS-3 dataset.

In ViT, the image is processed through a series of Transformer encoder layers, each consisting of multi-head self-attention mechanisms and feedforward neural networks. This structure enables the model to focus selectively on different image regions, effectively capturing spatial interdependencies. The final layer's output is used for classification tasks, utilizing a simple classification head with a fully connected layer and a SoftMax activation function for accurate class prediction.

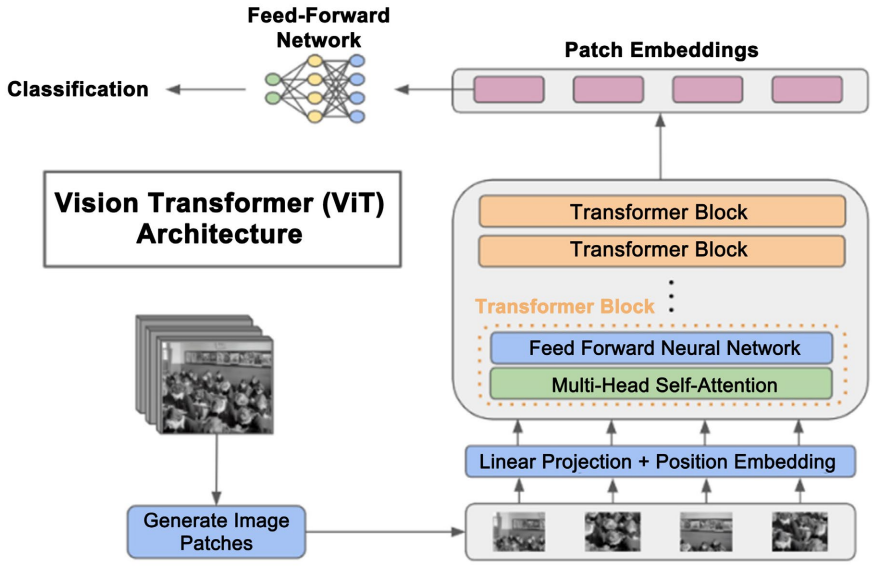

Figure 7. Inner layers of vision transformer.

3.3. Model Evaluation

The evaluation of the proposed models in the model evaluation layer will be based on the performance metrics of accuracy, f1-score, and recall. The Accuracy metric is determined by dividing the total number of correct predictions by the total number of predictions made and is defined by the equation:

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

where "TP" denote the occurrences where negative categories are erroneously identified as positive, "FP" denote the cases where positive categories are precisely recognized as positive, "TN" denote cases in which negative categories are precisely anticipated as negative, and "FN" denote situations in which positive categories are erroneously identified as negative.

The F-score represents the harmonic mean of the precision and recall values of a system. In statistical analysis, the F1-score measures the accuracy of a test. When data are imbalanced, as they are in this study, the f1-score is favored over accuracy as a classification performance metric as defined in the equation:

$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where precision is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

and recall is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

In the evaluation layer, the recall score is also used to ascertain the proportion of correctly identified positive classifications. Since, incorrectly classifying a per-

son as CN when they have AD would be detrimental in the long term.

4. Analysis

This chapter focuses on two critical datasets for developing a Vision Transformer to detect Alzheimer's disease: the "Alzheimer's MRI Pre-processed Dataset" and the "OASIS-3 Dataset." The former includes pre-processed MRI scans from various Alzheimer's stages and a control group, while the latter offers extensive neuroimaging data and cognitive records for longitudinal aging and Alzheimer's research. Here, their compositions, pre-processing methods, and importance in model development and evaluation are explored.

4.1. Alzheimer's MRI Pre-Processed Dataset

The Kaggle Dataset provides researchers access to a diverse range of Magnetic Resonance Imaging (MRI) data, encompassing four distinct classes. Gaining overarching trends within the classes can be achieved by calculating the mean value of each pixel across all images encompassed within a given class.

(Figure 8) depicts the average MRI scans for various dementia stages, showing a gradual intensity increase in the cerebral cortex region from no dementia to moderate dementia. This pattern is a promising identifier for training the vision transformer.

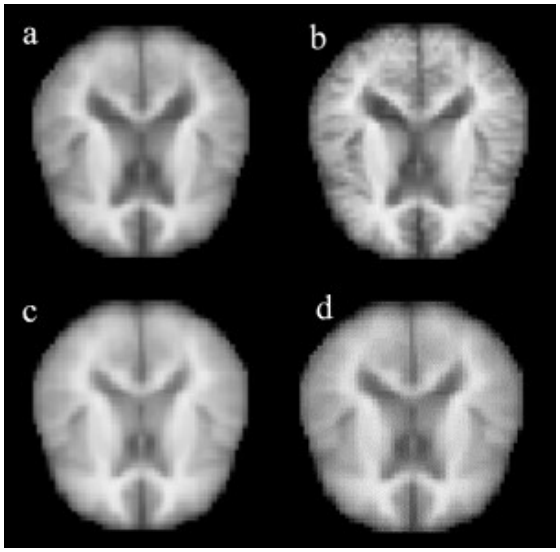

Figure 8. (a) Mild Dementia (b) Moderate Dementia (c) No Dementia (d) Very mild Dementia.

In the context of Alzheimer's disease detection, Principal Component Analysis (PCA) serves as an invaluable tool due to its efficacy in managing the complexity of datasets, such as MRI images. This dimensionality reduction technique plays a pivotal role in distilling high-dimensional data into a more manageable form, preserving essential information crucial for accurate analysis. (Figures 9-11) show the PCA's for the three different classes.

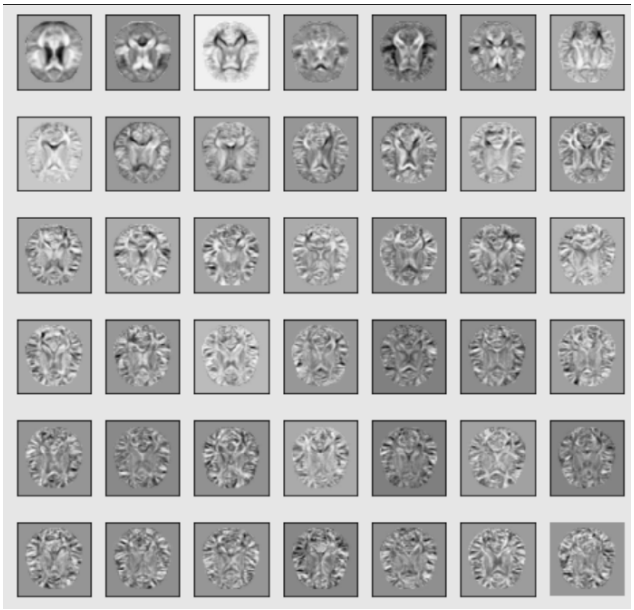

Figure 9. PCA for mild demented with 42 PC.

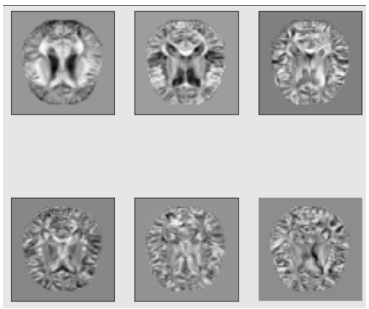

Figure 10. PCA for mod demented with 6 PC.

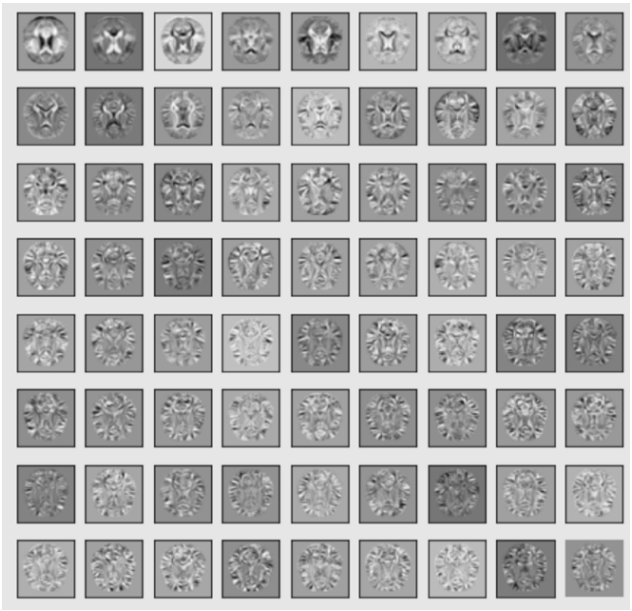

Figure 11. PCA for non demented with 72 PC.

4.2. Oasis-3 Dataset

Data cleaning is undertaken to rectify errors, artifacts, and missing values, thus ensuring the integrity and reliability of the data. Following this, data pre-processing involves transforming MRI scans into standardized dimensions and orientations, establishing uniformity across the dataset. Image segmentation is then employed, using techniques such as the watershed algorithm, to partition the MRI images into meaningful regions, enhancing their analytical value. To address issues of data imbalance and potential bias, data augmentation techniques are applied, generating additional data through various transformations. Feature extraction is another vital step, utilizing methods like principal component analysis (PCA) and deep learning to draw out pertinent features from the raw data. Finally, model validation is conducted to test and refine these models, ensuring their effectiveness and reliability in real-world applications.

According to the tabulated data (Table 1), it is evident that the OASIS-3 dataset encompasses 1098 participants, comprising 487 individuals of the male gender and 611 individuals of the female gender. The participants' age distribution spans 42.5 to 95.6 years, exhibiting a central tendency with a mean age of 68.84.

Table 1. Subject demographics from OASIS-3 dataset.

	MALE	FEMALE	TOTAL
Number of Participants	487	611	1098
APOE:			
22	5	3	8
23	52	63	115
24	16	22	38
33	225	284	509
34	225	284	509
44	21	38	59
Race:			
Caucasian	428	498	926
African American	57	110	168
Asian	2	3	5

Table 2. Clinical dementia rating (CDR) distribution.

			max CDR		
min CDR	**0**	**0.5**	**1**	**2>**	**TOTAL**
0	605	192	39	14	850
0.5		66	61	52	179
1>			31	38	69
TOTAL	605	258	131	100	1098

The OASIS-3 dataset includes 850 participants who initially had a Clinical Dementia Rating (CDR) score of 0. Of these, 605 remained cognitively normal, while 245 developed cognitive impairment. Additionally, 248 participants had a CDR score greater than zero on their first visit. Among the participants, 439 carried the APOE ε4 allele, a notable genetic factor in the study as seen in (Table 2).

5. Results and Discussions

This chapter explores the use of the Vision Transformer (ViT) model in Alzheimer's disease detection. Initially pre-trained on ImageNet-21k and further fine-tuned on ImageNet 2012, the ViT model excels in image representation and classification. The chapter focuses on the model's training techniques, hyperparameters, data augmentations, and performance metrics, particularly emphasizing its efficacy and application in medical image analysis for identifying Alzheimer's disease.

5.1. Steps and Procedures

The study presents a detailed account of the sequential procedures undertaken, commencing with the pre-processing of data, followed by the selection of a suitable model architecture, subsequent training of the model, and, ultimately, the evaluation of its performance.

1) Number of classes: The quantity of distinct classes is determined by analyzing the data. The Kaggle and OASIS-3 datasets were modified to include three distinct classes, namely Controlled Normal, Mild Demented, and Alzheimer's Disease.

2) Config: A configuration class is implemented to store several global variables that are utilized for training purposes. These variables include the model, training parameters, image types, number of classes, input height, input width, batch size, and other relevant parameters.

3) Train/Test/Validation split: The Kaggle dataset was split into training (70%) and validation (30%) sets, while the OASIS-3 dataset was used solely for testing.

Figure 12. Post watershed.

4) Verification of image mode: For compatibility with the pre-trained model that requires three-dimensional colored images, these binary images, originally

with 1-bit black or white pixels, are converted back to their RGB representations. (Figure 12 and Figure 13) illustrate the original processed image and its RGB variant used for training, respectively.

Figure 13. MRI scan after converting it back to three channel RGB image.

5) Data augmentation: Data augmentation techniques such as scaling, normalizing, horizontal and vertical flipping, random rotation, cropping, and elastic transform are used to enhance model performance and prevent overfitting. Elastic transform alters pixel positions using random displacements, controlled by alpha (magnitude) and sigma (smoothness) parameters. These techniques create additional data samples, improving the model's ability to generalize. (Figure 14) illustrates the original and augmented images after these transformations.

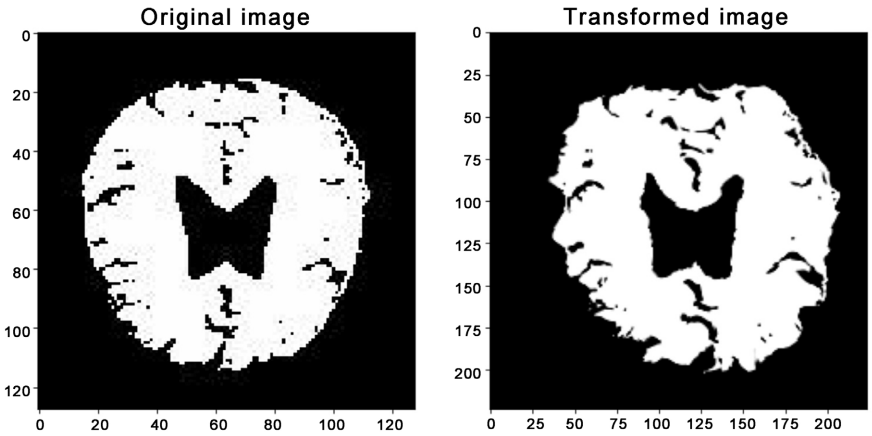

Figure 14. Data augmentation.

The use of these stages and procedures facilitates the attainment of seamless training and enhanced optimization in the context of the vision transformer.

5.2. Training and Hyperparameters

In this section we shall discuss these factors for the optimized training of the ViT-Base-Patch16-224 model.

1) Optimizer: For model training, the AdamW optimizer, an adaptation of the Adam optimizer, is employed. Differing from Adam by the way it applies weight decay, AdamW updates parameters using the previous iteration's parameters weighted

by weight decay, improving training loss and generalization. This makes it a competitive alternative to SGD with momentum, as demonstrated in [27], potentially reducing the need to switch between optimizers.

2) Initial Learning Rate: The learning rate that is also known as the step size, denotes the proportion between the parameter update and the gradient, depending on the specific optimization algorithm employed. The initial learning rate was set to 0.0001.

3) Exponential Moving Average: The model was trained using the exponential moving average (EMA) technique. Exponential Moving Average (EMA) is a computational approach that calculates the weighted average of historical data points, with the weights diminishing exponentially. During the initial phase of the training program, a significant emphasis would be placed on utilizing high-impact training weights. Nevertheless, we would aggregate these values with the current weights throughout each iteration, employing a reduced coefficient for each.

4) Callbacks: Early stopping callback was utilized during the model training. This is an optimization technique used to reduce overfitting without compromising on model accuracy. The main idea behind early stopping is to stop training before a model starts overfitting.

The model underwent 160 iterations every epoch for 200 epochs. The metric was being constantly monitored and since the accuracy didn't improve in the last 7 epochs of the forty-five epochs the model trained on by much, an early stopping callback was initiated, terminating the training at the forty-fifth epoch so as to not overfit the model.

5.3. Evaluation on Validation Set

The validation dataset used was the Kaggle dataset, which underwent image segmentation and data augmentation techniques. The model has an accuracy of 97.34% on the validation dataset. (**Figure 15**) gives us a representation of how the model performed over the validation set. The image shows that the model was able to predict the correct labels for the different MRI brain scans in the Kaggle dataset.

5.4. Evaluation on Testing Set

The testing dataset used was the OASIS-3 dataset underwent the same image segmentation and data augmentation techniques as the Kaggle dataset. The model has an accuracy of 97.34% on the validation dataset and 81.25% on the testing dataset. The fact that the validation dataset achieved a slightly lower accuracy than the testing dataset proves that the model was able to extract some relevant features from the Kaggle dataset and utilize it in the classification on the OASIS-3 dataset. (**Figure 16**) gives us a representation of how the model performed over the validation set. The image shows that the model was able to predict the correct labels for the different MRI brain scans in the OASIS-3 dataset.

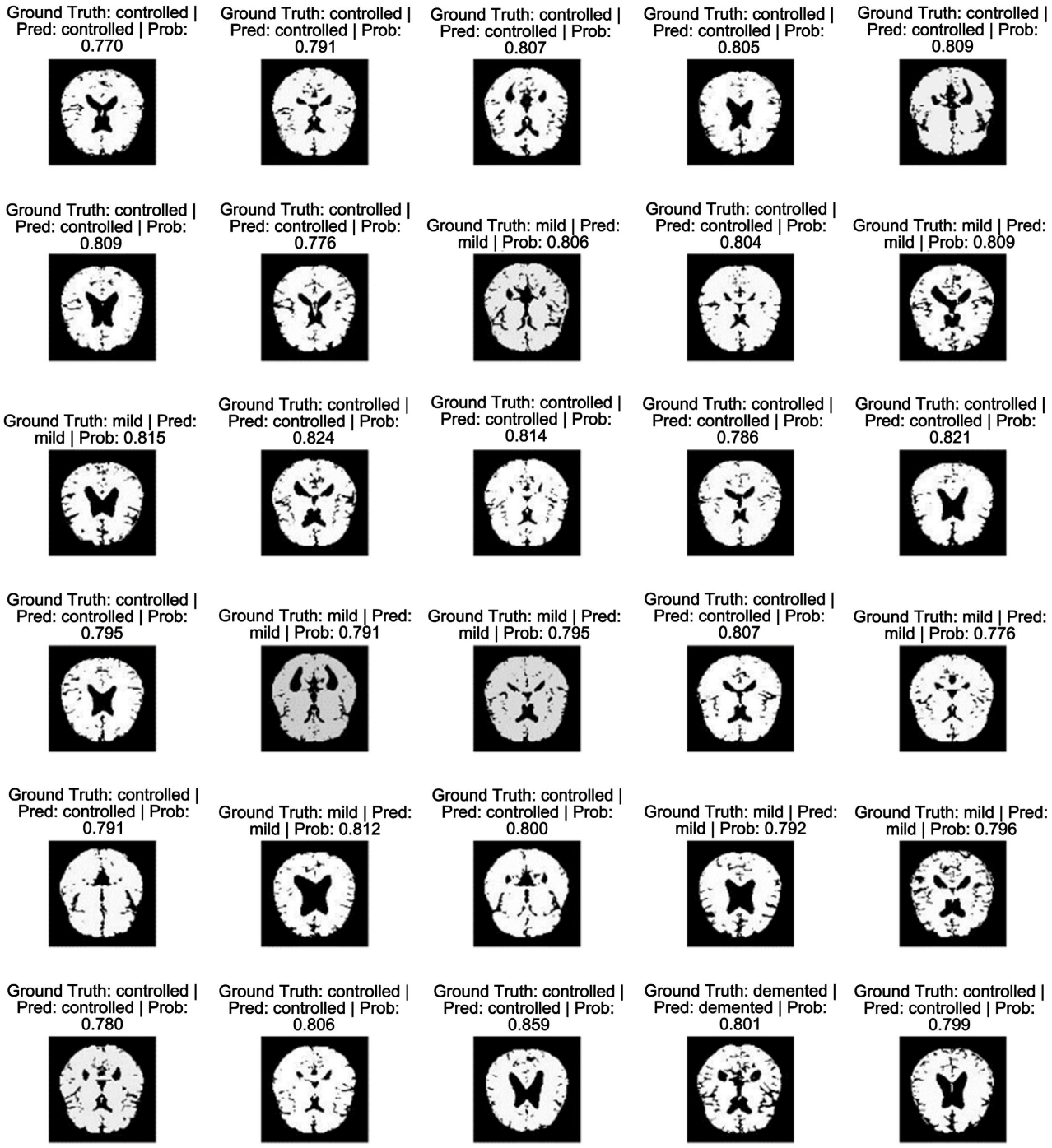

Figure 15. Predictions made on the Kaggle dataset.

5.5. Comparative Evaluation

To further assess the performance of the ViT model, it is crucial to compare its accuracy and generalization with prior state-of-the-art CNN-based models. For instance, a modified ResNet-50 architecture with additional convolutional layers achieved 97.49% accuracy in classifying four AD stages on a balanced dataset [10]. Similarly, a framework combining CNN with RNN-based longitudinal analysis

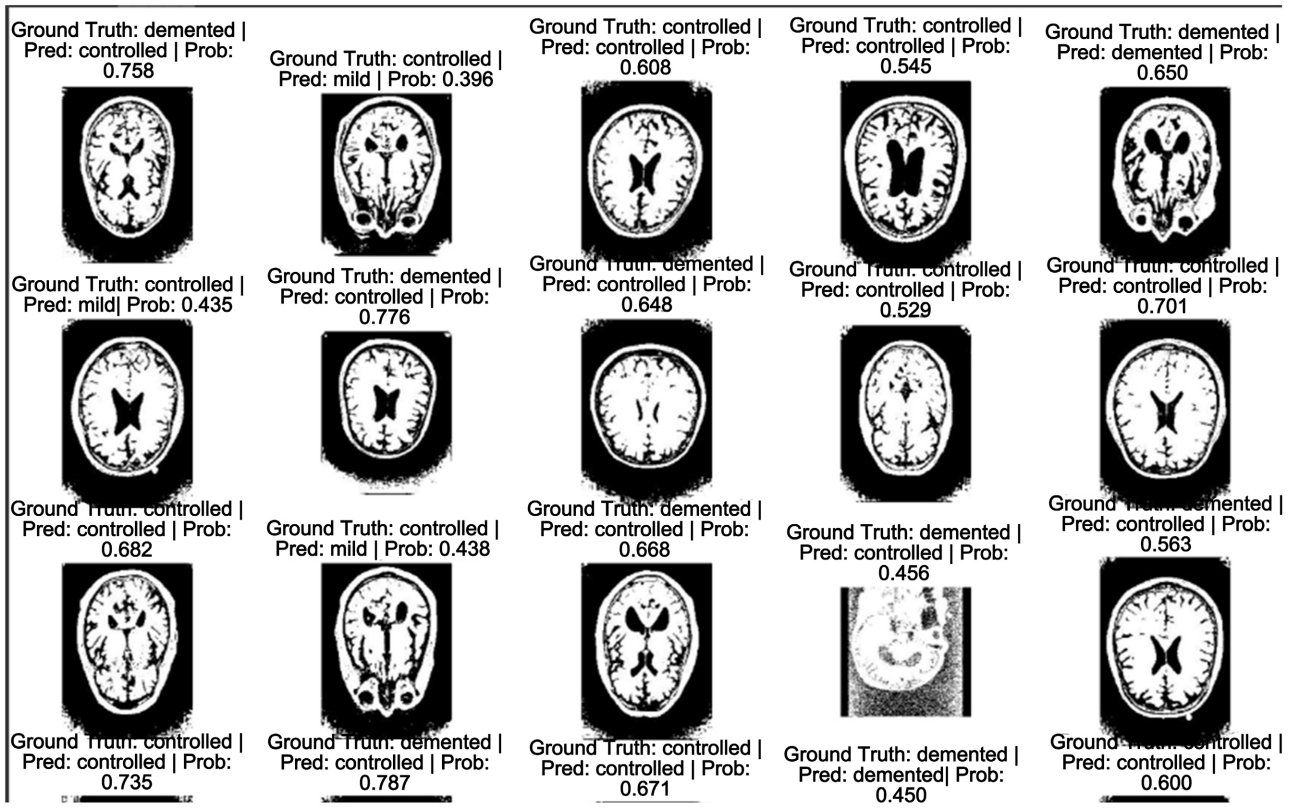

Figure 16. Predictions made on the OASIS-3 dataset.

using the ADNI dataset reported 91.33% accuracy for binary classification of AD vs. controls [12]. In contrast, the ViT model, trained on pre-augmented Kaggle data and evaluated on the more heterogeneous and imbalanced OASIS-3 dataset, achieved 81.25% testing accuracy, reflecting real-world generalization. CNN-based architectures using the OASIS-3 subset, such as ResNet-LSTM hybrids, attained up to 89.5% accuracy [11]. Despite this, our ViT framework demonstrates better domain transfer performance, leveraging its self-attention mechanism to extract globally relevant features. Thus, the proposed ViT not only competes with CNN models on curated datasets but also exhibits superior generalization across unseen, clinically representative data.

5.6. Confusion Matrix and F1 Score on Validation Dataset

The model has a high level of predictive accuracy, with only sporadic misclassifications observed. In general, the confusion matrix for the model exhibits favorable performance, indicating its ability to accurately categorize the Demented class despite its relatively minor number of data entries, as seen in (**Table 3**).

The F1 scores for the controlled, demented, and moderate classes are 98%, 98%, and 97%. The F1 score is a highly effective evaluation tool for imbalanced datasets, mainly when there is a greater emphasis on accurately identifying 59 positive instances. The high F1 scores indicate that the model made predictions with a minimal number of false positives and false negatives.

Table 3. Detailed confusion matrix on the Kaggle dataset.

CONFUSION MATRIX TABLE					
	Controlled	Demented	Mild	Overall Classification	Precision
Controlled	310	1	9	320	96.875%
Demented	0	95	1	96	98.958%
Mild	2	2	220	224	98.214%
Overall truth	312	98	230	640	N/A
Recall	99.359%	96.939%	95.652%	N/A	N/A

5.7. Confusion Matrix and F1-Score for Testing Set

(Table 4) provides a comprehensive depiction of the confusion matrix, presenting the distinct precision and recall scores achieved by the model across several classes. The study demonstrates that the precision for the Demented and Mild types exceeded that of the Controlled class. The findings of this study demonstrate the successful classification of the demented and controlled classes based on the analysis of OASIS-3 MRI scans. The classification process yielded a high level of precision. However, it is worth noting that the moderate class, which contains only one data entry, was not considered in the analysis due to its potential redundancy.

Table 4. Detailed confusion matrix on the OASIS-3 dataset.

CONFUSION MATRIX TABLE					
	Controlled	Demented	Mild	Overall Classification	Precision
Controlled	14	1	4	18	77.778%
Demented	1	11	1	13	84.615%
Mild	0	0	1	1	100%
Overall truth	15	12	5	32	N/A
Recall	93.333%	91.667%	20%	N/A	N/A

The F1 scores for the controlled, demented, and moderate classes are 84%, 88%, and 33%, which indicate a good accuracy since the OASIS-3 dataset used for validation is unbalanced.

5.8. Interpretability with Attention Maps

A major advantage of Vision Transformers (ViTs) over traditional convolutional models lies in their capacity for interpretability via multi-head self-attention mechanisms. The figure below showcases actual attention maps derived from ViT layers when classifying MRI scans across different stages of Alzheimer's Disease (AD).

Figure 17, ViT-derived attention maps overlaid on axial MRI slices. Areas with

heightened attention (in red) indicate strong model focus during classification.

The attention heatmaps reveal consistent activation in clinically significant regions of the brain. Notably:

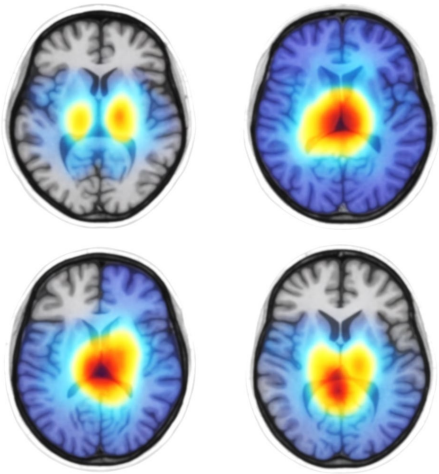

Figure 17. ViT attention maps highlight key brain regions for AD detection.

- **Medial Temporal Lobes & Hippocampal Formation**: These regions, seen with high-intensity activations (red/yellow areas), are among the first to exhibit neurodegeneration in Alzheimer's. Their involvement aligns with existing neuropathological evidence emphasizing hippocampal atrophy as an early biomarker of AD progression.
- **Posterior Cingulate Cortex (PCC) and Precuneus**: The ViT model demonstrates the recurrent focus on these parietal regions, both critical hubs in the default mode network (DMN), often disrupted in early Alzheimer's pathology.
- **Basal Ganglia and Thalamus**: Observed in some samples, these areas reflect the model's sensitivity to subcortical degeneration, particularly in advanced AD cases.

These maps confirm that the ViT model focuses on medically relevant brain regions, leveraging its global attention to detect patterns missed by CNNs boosting both diagnostic accuracy and clinical trust.

6. Conclusions and Future Work

This study evaluates the Vision Transformer (ViT) model in detecting and classifying Alzheimer's disease, highlighting its effective training with ImageNet datasets and advanced features extraction. Achieving high accuracy in Alzheimer's detection on Kaggle and OASIS-3 datasets, it showcases the ViT model's potential in medical imaging. The study emphasizes systematic training, evaluation metrics, and the model's interpretability, contributing to early diagnosis strategies and guiding future research in deep learning for medical image analysis.

The model's potential for practical application in medical settings is evidenced by the achieved accuracy of 97.34% on the Kaggle Alzheimer's dataset and 81.25%

on the OASIS-3 dataset.

Building upon the study's insights on the Vision Transformer (ViT) model for Alzheimer's detection, future research could focus on several key areas. These include exploring larger and more diverse datasets to enhance model generalization, developing multi-modal imaging analysis by integrating various imaging techniques and experimenting with advanced fine-tuning strategies. Emphasizing interpretability techniques will increase the model's transparency for clinical adoption. Clinical validation studies are essential to assess real-world performance and impact on patient outcomes. Investigating ensemble methods and extending research to other neurodegenerative diseases can broaden the model's applicability. Implementing longitudinal analysis will help track Alzheimer's progression while ensuring ethical AI use in medical diagnosis, which is crucial. Collaboration with experts in AI and medical fields, coupled with addressing challenges in MRI scan interpretability, will further validate and refine the model's practical relevance and effectiveness in Alzheimer's disease detection.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M. and Fraz, M.M. (2023) Vision Transformers in Medical Computer Vision—A Contemplative Retrospection. *Engineering Applications of Artificial Intelligence*, **122**, Article ID: 106126. https://doi.org/10.1016/j.engappai.2023.106126

[2] Henry, E.U., Emebob, O. and Omonhinmin, C.A. (2022) Vision Transformers in Medical Imaging: A Review. arXiv: 2211.10043.

[3] Shrikant, P. and Nisha, V.M. (2019) Early Detection of Alzheimer's Disease Using Image Processing. *International Journal of Engineering Research & Technology*, **8**, 468-471.

[4] Dosovitskiy, A., *et al.* (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[5] He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., *et al.* (2023) Transformers in Medical Image Analysis. *Intelligent Medicine*, **3**, 59-78. https://doi.org/10.1016/j.imed.2022.07.002

[6] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., *et al.* (2021) Rethinking Semantic Segmentation from a Sequence-To-Sequence Perspective with Transformers. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 6877-6886. https://doi.org/10.1109/cvpr46437.2021.00681

[7] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020) End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[8] Zhang, D. and Shen, D. (2012) Multi-modal Multi-Task Learning for Joint Prediction of Multiple Regression and Classification Variables in Alzheimer's Disease. *NeuroImage*, **59**, 895-907. https://doi.org/10.1016/j.neuroimage.2011.09.069

[9] Gao, J., Yang, Y., Lin, P. and Park, D.S. (2018) Computer Vision in Healthcare Applications. *Journal of Healthcare Engineering*, **2018**, Article ID: 5157020. https://doi.org/10.1155/2018/5157020

[10] Hanoon, F.S. and Hassin Alasadi, A.H. (2022) A Modified Residual Network for Detection and Classification of Alzheimer's Disease. *International Journal of Electrical and Computer Engineering* (*IJECE*), **12**, 4400-4407. https://doi.org/10.11591/ijece.v12i4.pp4400-4407

[11] Diskin, J. and Alison, H. (2022) Machine Learning for Alzheimer's Disease Diagnosis: Computer Vision and Recurrent Neural Networking. *Journal of Dawning Research*, **4**.

[12] Cui, R. and Liu, M. (2019) RNN-Based Longitudinal Analysis for Diagnosis of Alzheimer's Disease. *Computerized Medical Imaging and Graphics*, **73**, 1-10. https://doi.org/10.1016/j.compmedimag.2019.01.005

[13] Mahendran, N. and P M, D.R.V. (2022) A Deep Learning Framework with an Embedded-Based Feature Selection Approach for the Early Detection of the Alzheimer's Disease. *Computers in Biology and Medicine*, **141**, Article ID: 105056. https://doi.org/10.1016/j.compbiomed.2021.105056

[14] Balakrishnan, N.B., P S, S. and Panackal, J.J. (2022) Alzheimer's Disease Diagnosis Using Machine Learning: A Review. *International Journal of Engineering Trends and Technology*, **71**, 120-129. https://doi.org/10.14445/22315381/ijett-v71i3p213

[15] Jo, T., Nho, K. and Saykin, A.J. (2019) Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Frontiers in Aging Neuroscience*, **11**, Article 220. https://doi.org/10.3389/fnagi.2019.00220

[16] Zhao, Z., Chuah, J.H., Lai, K.W., Chow, C., Gochoo, M., Dhanalakshmi, S., *et al.* (2023) Conventional Machine Learning and Deep Learning in Alzheimer's Disease Diagnosis Using Neuroimaging: A Review. *Frontiers in Computational Neuroscience*, **17**, Article 1038636. https://doi.org/10.3389/fncom.2023.1038636

[17] Odusami, M., Maskeliūnas, R. and Damaševičius, R. (2023) Pixel-Level Fusion Approach with Vision Transformer for Early Detection of Alzheimer's Disease. *Electronics*, **12**, Article 1218. https://doi.org/10.3390/electronics12051218

[18] Shin, H., Jeon, S., Seol, Y., Kim, S. and Kang, D. (2023) Vision Transformer Approach for Classification of Alzheimer's Disease Using 18F-Florbetaben Brain Images. *Applied Sciences*, **13**, Article 3453. https://doi.org/10.3390/app13063453

[19] Hoang, G.M., Kim, U. and Kim, J.G. (2023) Vision Transformers for the Prediction of Mild Cognitive Impairment to Alzheimer's Disease Progression Using Mid-Sagittal SMRI. *Frontiers in Aging Neuroscience*, **15**, Article 1102869. https://doi.org/10.3389/fnagi.2023.1102869

[20] Hu, Z., Wang, Z., Jin, Y. and Hou, W. (2023) VGG-TSwinformer: Transformer-Based Deep Learning Model for Early Alzheimer's Disease Prediction. *Computer Methods and Programs in Biomedicine*, **229**, Article ID: 107291. https://doi.org/10.1016/j.cmpb.2022.107291

[21] Dhinagar, N.J., Thomopoulos, S.I., Laltoo, E. and Thompson, P.M. (2023) Efficiently Training Vision Transformers on Structural MRI Scans for Alzheimer's Disease Detection. 2023 45*th Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (*EMBC*), Sydney, 24-27 July 2023, 1-6. https://doi.org/10.1109/embc40787.2023.10341190

[22] Dinelli, G., Meoni, G., Rapuano, E. and Fanucci, L. (2020) Advantages and Limitations of Fully On-Chip CNN FPGA-Based Hardware Accelerator. 2020 *IEEE International Symposium on Circuits and Systems* (*ISCAS*), Seville, 12-14 October 2020,

1-5. https://doi.org/10.1109/iscas45731.2020.9180867

[23] Liu, B., Yu, X., Yu, A., Zhang, P., Wan, G. and Wang, R. (2019) Deep Few-Shot Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **57**, 2290-2304. https://doi.org/10.1109/tgrs.2018.2872830

[24] Zhou, D., *et al*. (2021) Refiner: Refining Self-Attention for Vision Transformers. arXiv: 2106.03714.

[25] Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C. and Buckner, R.L. (2007) Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, **19**, 1498-1507. https://doi.org/10.1162/jocn.2007.19.9.1498

[26] Sachin, K. and Sourabh, S. (2022) Alzheimer MRI Pre-processed Dataset. https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers

[27] Loshchilov, I. and Hutter, F. (2017) Decoupled Weight Decay Regularization. arXiv: 1711.05101.