EARLY DETECTION SAVES LIFE: COMPARATIVE EVALUATION OF TRADITIONAL CONVOLUTIONAL NEURAL NETWORK MODEL AND THE VISION TRANSFORMER MODEL

A DEEP LEARNING APPROACH TO CLASSIFY DEMENTIA STAGES

TUNAHAN TALU

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2065427

COMMITTEE

prof. Dr. Eric Postma
Dr. Eriko Fukuda

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2024

WORD COUNT

6023

ACKNOWLEDGMENTS

No journey is undertaken alone, and I am very grateful to those who supported me both practically and psychologically.
First, I extend my deepest gratitude to my supervisor, Prof. Dr. Eric Postma. Your inclusion in my research project and the challenging tasks you assigned expanded my knowledge and pushed my skills. Your guidance and support were crucial in completing my Bachelor thesis.
I am also deeply thankful to my family. Despite the distance, their limitless support and encouragement kept me motivated through every challenge.
Lastly, I thank my housemates, Perfecto and Efe, for creating a welcoming home away from home. Your companionship and our endless nerve stimulating conversations were invaluable. Thank you all for your support and encouragement.

EARLY DETECTION SAVES LIFE: COMPARATIVE EVALUATION OF TRADITIONAL CONVOLUTIONAL NEURAL NETWORK MODEL AND THE VISION TRANSFORMER MODEL

A DEEP LEARNING APPROACH TO CLASSIFY DEMENTIA STAGES

TUNAHAN TALU

Abstract

Alzheimer's disease (AD) represents the most prevalent cause of cognitive decline among the elderly, posing significant global health, social, and economic challenges. Early and accurate diagnosis is critical for mitigating the impact of AD. This study investigates the effectiveness of Transformer model, initially developed for Natural Language Processing, in image classification of MRI scans to detect AD at various stages. Utilising the Open Access Series of Imaging Studies (OASIS) dataset, comprising 80,000 MRI samples categorised into four dementia stages, the study aims to enhance diagnostic accuracy through advanced deep learning techniques. Traditional diagnostic methods, including MRI and PET scans, although effective, suffer from subjectivity and resource constraints, limiting their widespread applicability. The Vision Transformer (ViT) model, by leveraging its attention mechanism, processes images as sequences of patches, focusing on crucial regions to grasp contextual relationships within the data. This approach contrasts with the conventional Convolutional Neural Network (CNN) models, which apply filters across the entire image. Both models demonstrated exceptional success from the test results, achieving between 95%-99% from testing metrics. Key metrics, including precision, recall, and F1-score, underscore the ViT model's proficiency in distinguishing between dementia stages. The study concludes that ViT models hold significant promise in improving AD diagnosis, emphasising the need for further optimization to reduce computational demands. The findings advocate for integrating ViT models into clinical settings, potentially transforming

early AD detection and enhancing patient outcomes through timely intervention.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The OASIS dataset has been acquired from the (Open Access Series of Imaging Studies) through online access. The dataset was publicly available in the kaggle website in the following link Aithal (2024). The obtained data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. However, the institution was informed about the use of this data for this thesis and potential research publications. All the tables (table 1, table 2, table 3) belong to the author. The thesis code can be accessed through following links; 'Singh (2023), Angyalfold (2021), Google (2023), Face (2023). The reused/adapted code fragments are clearly indicated in the notebook. In terms of writing, the author used assistance with the language of the paper. Thesaurus used for synonyms of the words used. No other typesetting tools or services were used.

2 INTRODUCTION

Alzheimer's disease (AD) stands out as the prevailing etiology of cognitive decline in elderly individuals, presenting a complex problem within the field of neurodegenerative conditions. With the progressive ageing of populations across the world, the incidence of Alzheimer's disease exhibits an upward trajectory, thereby presenting intimidating health, societal, and economic obstacles on a global scale (X. Li et al., 2022). The escalating prevalence of Alzheimer's disease highlights the acute need for collaborative efforts in research, healthcare, and policymaking to address the multifaceted implications of this debilitating condition. The global estimate provided by the World Health Organization suggests that currently more than 55 million individuals worldwide are affected by dementia, with a significant portion of cases, ranging from 60% to 70%, being attributed to AD, making it the most prevalent form of dementia (International, 2023). The projected increase in this numerical value is anticipated to experience a threefold growth by the year 2050, emphasising the critical importance of promptly implementing efficient diagnostic and therapeutic approaches to address the escalating demand for healthcare services.

The pathophysiological mechanisms underlying AD associate deterioration in cognitive functions, particularly affecting memory retention and executive functioning, which are crucial for decision making and problem

solving processes in individuals diagnosed with AD (DeTure & Dickson, 2019). The decline in these cognitive abilities is a strong indicative feature of AD, reflecting the progressive degeneration of brain structures and neural networks responsible for cognitive processing and information retrieval in AD patients. The progression of this disease is emphasised by the accumulation of tau protein tangles and amyloid-beta plaques within the brain, resulting in neuronal injury and depletion, which serves as the foundation for the deteriorative course of the disease (Uddin et al., 2020). Despite the progress made in the field of research, Alzheimer's disease continues to be without a cure, with existing treatments only able to provide temporary slowing of its progression.

2.1 *Diagnostic Challenges*

Correctly diagnosing and especially early treatment in preventing AD to progress is critical and should not be underestimated. Timely detection of the disease enables the application of treatment procedures that have the potential to effectively alter its progression and influence on individuals' well-being (Cummings et al., 2022). Up to date, there are some effective neuroimaging techniques to examine the brain either structurally or functionally. These neuroimaging techniques differ from each other in different ways such as invasiveness, temporal resolution, spatial resolution, and accessibility. Every technique has its own specific requirements and mechanisms to be working as desired for different types of examination from the brain. In the scope of diagnosing dementia stages, structural Magnetic Resonance Imaging (sMRI) and Positron Emission Tomography (PET) are widely used. In the 1980s PET was the dominant invasive hemodynamic imaging method which has a unique contribution relative to functional Magnetic Resonance Imaging (fMRI) in such by measuring metabolism and detecting biomarkers and neurotransmitter concentrations via injection of radioactive tracers into our bloodstream (Finnema et al., 2016). On the other hand, structural MRI is a non-invasive diagnostic tool that uses strong magnetic fields and radio waves to generate images of the brain's structure which help in diagnosing a variety of conditions. These generated images help to investigate the brain's structural compounds with relations to some neurological conditions. MRI can identify and characterise brain tumours based on their location and the possible damage they have given on surrounding tissues (Venere, Zadeh, Puduvalli, & Haynes, 2020). For effective results, these imaging methods require a high level of expertise for interpretation and are often wrongly guiding the patients because of subjectivity in assessment, resulting in inconsistent diagnostic outcomes. Additionally, these techniques require a lot of resources that might not be

available in all healthcare settings, especially in low-resource environments such as mediocre countries, underlining the necessity for more accessible and standardised diagnostic instruments (DeStigter et al., 2021).

2.2 *AI in Healthcare*

In recent years, artificial intelligence (AI) has started to revolutionise various fields especially in healthcare and medical practices. AI works as a significant beneficiary in different aspects for the medical industry, delivering the potential of robust, impartial, and effective solutions in the realm of diagnosing intricate diseases like Alzheimer's Disease (Basu, Sinha, Ong, & Basu, 2020). Deep learning methodologies, which have the nature of training artificial neural networks on large datasets to carry out tasks like image recognition and classification, have proven their ability to diagnose accurately and efficiently (Aggarwal et al., 2021).

This research aims to address the critical gap in Alzheimer's disease diagnosis by using deep learning techniques, specially Vision Transformer (ViT) which is initially built for Natural Language Processing (NLP) tasks, for classification of MRI scans and early detection of AD. The ViT model, which employs mechanisms from NLP in image classification, presents a promising framework due to its capacity to concentrate on different regions of an image as an input and grasp contextual relationships within the data. As pointing out the critical medical concerns, these advancements in early diagnosis could have a significant impact on treatment outcomes and the quality of life for patients with Alzheimer's disease (Diogo, Ferreira, Prata, & Initiative, 2022). Through using artificial intelligence tools and cognitive science, study seeks to detect AD early, highlighting the exciting potential of AI in the medical field. It provides an important advancement in the fight against this severe condition and aligns with the larger goal of utilising technology to address difficult conditions.

2.3 *Dataset*

In the beginning of dataset searching, a dataset containing 6000 samples of MRI scans were found on open source website Kaggle. However, due to limited dataset sample size, we shifted our scope to OASIS dataset containing 80,000 samples which was also free to access via Kaggle. By utilising OASIS dataset comprising 80,000 samples categorised into four dementia stages (Moderate, Non-Demented, Mild, and Very Mild), the objective of this research is to test a model's capability of precisely identifying and classifying AD stages through analysis of brain scans. This study entails the potential to enhance the objectivity and accuracy of AD diagnostics

while increasing the awareness of early detection in various healthcare environments.

2.4 *Patients and Society*

Social and personal impact that AD creates expands over the patient himself and the family of the patient. The progressive decline in cognitive functions affects the patients' daily life behaviours and the interactions with patients' environment, sparking off a comprehensive care essential yet challenging (Grabher, 2018). Families experience devastating emotional and financial load, usually without adequate support neither from authorities nor from patients. Patients tend to be more aggressive and unresponsive to face reality and accept the disease. This mental state of the patients makes the process even harder for both sides, resulting in late diagnosis of the stage and unguided treatment (Wattmo & Wallin, 2017). On a societal level, public awareness promotes the importance of early diagnosis which can help individuals to recognize the early signs. Providing resources and training camps for caregivers can make both parties life easier and healthier

The adaptation of Transformer models to image classification, specifically medical image classification is an innovative improvement in AI technology. Convolutional Neural Networks (CNN) typically require fixed input sizes, which can be limitation for medical imaging where images can vary in dimensions. Unlike some architecture of CNN, ViTs do not process the input data at once instead, they divide the image into patches and process the patches simultaneously utilising self-attention mechanism, which allows model to capture any part of the image regardless of the spatial distance between different features (Matsoukas, Haslum, Söderberg, & Smith, 2023). This method suggests higher effectiveness in handling complex image data such as MRI scans where subtle changes might indicate early stages of Alzheimer's disease. Innovatively, for improved generalization and performance, Vision Transformers can benefit from pre-training on large datasets and fine-tuning on smaller medical datasets using transfer learning. ViT's ability to scale effectively with more data and computational resources offer a potential for significant performance improvements in image related tasks as more high-quality labelled images become available. This scalability is crucial in medical settings where accumulating vast amounts of data over time is inevitable.

Based on the successful performance of deep neural networks in image classification tasks and the Vision Transformer models' image processing abilities, this paper aims to compare the performances of traditional convolutional neural networks model and the Vision Transformer model using

Magnetic Resonance Imaging samples from Open Access Series of Imaging Studies. The research objectives are to find answers the following research questions:

- Can the Vision Transformer model outperform the traditional convolutional neural network models?
- Is the Vision Transformer model worthwhile for integration into medical image analysis?

3 RELATED WORK

Deep learning's application in medical diagnostics, especially the diagnosis of Alzheimer's disease (AD), represents a significant advancement in the field's understanding of neurodegenerative illnesses. Deep learning is an exceptional tool for autonomously discovering features in large complex datasets, especially considering that AD is characterised by subtle, early-stage biomarkers that are often missed by traditional detection techniques (Helaly, Badawy, & Haikal, 2022). The potential for better diagnostic accuracy has been highlighted by recent studies using deep learning for AD stage categorization (Ftoutou, Majdoub, & Ladhari, 2023; Jagadeeswari, Priya, Athira, Dhanalakshmi, & Shree, 2022; Mggdadi, Al-Aiad, Al-Ayyad, & Darabseh, 2021). However, the results also highlight the need for additional optimization to improve performance on different datasets and types of imaging. In their study, El-Assy, Amer, Ibrahim, and Mohamed (2024) proposes a novel CNN architecture by concatenating two simple CNN models from scratch, each simple model achieving close to the excellent results by 95% from every performance metrics. Individually, the simple CNNs achieve substantially high scores from the metrics; however, their proposed model outperforms the simple ones in every performance metric namely, accuracy, recall, and precision.

The adaptation of Vision Transformer models, uniquely built for NLP tasks, have recently taken substantial attention in the medical imaging field because of their ability to be trained for learning robust features without the need for extensive pre-processing procedures required for traditional Convolutional Neural Networks (Shamshad et al., 2023). As a result of processing the images into multiple patches simultaneously, ViT models are highly being used for especially Alzehimer's disease detection. Wang, Chen, Zhang, and Wang (2024) demonstrate that ViT achieves exceptional performance in identifying profound neuroanatomical changes in MRIs, also in other neuroimaging techniques such as CT scans, compared to standard CNNs, assigning a suggestive shift towards more innovative neural architectures in medical imaging.

As one of the newly innovated neural architecture for enhancing the computational framework, Swin Transformers have also been integrated, which is a variant of ViTs, adapts the Transformer architecture to engage more effectively on image data. Swin Transformers also have shown significant progress in segmenting complex patterns from image data such as MRI and PET scans which are essential types of imaging techniques for specifically early AD diagnosis. Zhao et al. (2024) highlighted how Swin Transformers outperform traditional methods in training metrics such as accuracy and also computationally efficiency, offering a scalable solution for analysing high-dimensional medical image data.

In order to reduce time and computational resources for training a model a common strategy is being used in the deep learning field. Transfer learning is an effective technique for taking advantage of related previous tasks where a developed model for a task is reused as a starting point for another task. Integration of transfer learning and deep learning models like ViT models creates a way for improving AD diagnostics (Ghaffari, Tavakoli, & Pirzad Jahromi, 2022). However, as Krishnapriya and Karuna (2023) pointed out, while transfer learning can reduce the need for large labelled datasets, it also reproduces a challenge in model generalisation across diverse clinical environments. Making contemporary research in this area a necessity for exploiting the full potential of transfer learning, in their research Filipiuk and Singh (2022) are investigating adaptive algorithms that alter pre-trained models to better fit specific medical applications.

A vital drawback that affects deep learning models to be used in clinical application is that there is a lack of interpretability with the analysis of proposed models (M. Li, Jiang, Zhang, & Zhu, 2023). Interpretability is the ability of researchers to understand and explain how a model makes decisions and comprehend the factors and processes that lead to particular results from a model. Complex models like deep learning models provide high accuracy scores but their "black box" character makes interpretability difficult for clinicians and patients to trust their predicted scan results without understanding how they arrive at this specific conclusion. S. A. Kumar and Sasikala (2023) and Arafa, Moustafa, Ali, Ali-Eldin, and Saraya (2024) addressed these concerns by developing frameworks that improve transparency of the decision making process of deep learning models. During recent research, Liu et al. (2021) implements transfer learning method in order to solve the insufficient data problem in the brain imaging domain by introducing the Vision Transformer model. Even though expected and acquired results agree that Transfer learning can alleviate the problem of data, sample size still has an influence on the transfer performance which might cause an increase of model complexity but also overfitting to pre-trained features. These progressions are important for clinicians to

be more confident with the analyses and patients to gain trust towards innovative developments in the field of medical AI. The trust towards AI applications makes adapting possible for deep learning models in clinical settings, where understanding the reasoning behind the diagnostic suggestions is crucial for patients' mental state in order to fight back the disease ideally.

In a literature review, Maurício, Domingues, and Bernardino (2023) discuss and compare the Vision Transformer model and Convolutional Neural Network for image classification. In the review, different variants of CNN models were compared with the Vision Transformer model in different medical image classification tasks. Stating the both neural networks advantages in different settings, the paper expresses the need for interpretability in such models. Filipiuk and Singh (2022) investigates the Vision transformer in comparison and conjunction of CNNs, exploring whether ViT and CNN can be used together for more accurate results in image classification. One of the most emphasised aspects of this study is the integration of the unique network design of the ViT in safety critical systems for computer vision, underlying the ViTs being more resilient than CNNs. While CNN achieves 0.972 in the top 5 accuracy scores, ViT achieved 0.974 in the top 5 accuracy scores. Although the ensemble of the CNN and the ViT increases the accuracy up to 10% higher than individual performance results, up to 0.981, from the ImageNet domain.

AI-driven diagnostics in the medical field carry out important adjustments in the future developments, regarding limitations for handling large volumes of data, human errors, and for enhancing personalised medicine. Current developments in computational efficiency and data processing are supposed to improve the practical implementation of complex models like ViTs (Y. Kumar, Koul, Singla, & et al., 2023). Ongoing evolutions of AI technologies, specially over the adaptation of Vision Transformers and their varieties, are accommodating the prospect of Alzheimer's disease diagnostics (Marshall & Uchegbu, 2022). Addressing the current challenges related to data dependency, interpretability, and clinical integration, further research has the potential to unravel more reliable and accessible tools for early and accurate detection of AD.

4 METHODS

This study will investigate the adaptation of the Transformer model, initially created for NLP tasks, for the early detection of Alzheimer's disease (AD) using medical images such as MRI. Vision Transformer models address the need for efficient processing of ever-expanding datasets by offering versatile computations suitable for large and diverse datasets. This

capability is especially useful in medical imaging, where datasets can be limited due to privacy concerns or lack of data. ViT models offer promising solutions through the use of transfer learning from pre-trained models on bigger natural image datasets such as ImageNet, in spite of the challenges encountered when working with smaller datasets in medical imaging. Addition to that, ViT models can adapt and fine-tune their parameters to extract revealing features even from small medical image datasets, thereby enhancing their usefulness for diagnostic tasks. Furthermore, ViT models provide interpretability through attention maps, allowing analysts to identify fundamental regions in medical images contributing to diagnoses (Komorowski, Baniecki, & Biecek, 2023). This clarity reinforces trust in AI-assisted diagnostics and aids in error analysis for model improvement. The model's rate of processing and analysis of MRI data is significantly increased by using the L4 GPU for training. This option not only increases processing capacity but also goes above the limitations offered by traditional CPUs, enabling a faster and more efficient diagnostic procedure.

4.1 *Dataset Description*

In this section, the preprocessing pipeline created to fit a ViT model for a dataset containing different stages of dementia is described. The preprocessing steps are designed to handle class imbalance and improve the model's generalisation over a variety of data points, in addition to transforming images into a format that is compatible with model intake. The study utilises the publicly available MRI dataset from the Open Access Series of Imaging Studies (OASIS), comprising images categorised into four classes: 'Mild Dementia', 'Moderate Dementia', 'Non Demented', and 'Very mild Dementia'. The dataset contains 86,437 samples of MRI scans combined and exhibits significant class imbalance. 'Non Demented' has the highest sample size with 67,222, followed by 'Very mild Dementia' with 13,725, 'Mild Dementia' with 5,002, and 'Moderate Dementia' with just 488 samples. Class imbalance makes it challenging to train a model that can accurately and unbiasedly identify and classify all stages of dementia.

4.2 *Image Handling*

Accurate handling of images plays an essential role in the preprocessing pipeline of the Vision Transformer model to ensure that the images meet the requirements for model input. In order to open and convert image files into a consistent RGB format, the Python Imaging Library (PIL) is imported and used. This standardisation is critical for eliminating the

variations in image formats throughout the dataset, ensuring that every input is handled uniformly.

The images, when in RGB format, are passed through ViTImageProcessor, which has been specially trained for image classification tasks using Vision Transformer architectures. By scaling the images to 224x224 pixels, which is a prerequisite for the input dimension of the 'google/vit-base-patch 16-224-in21k' model, this feature extractor standardises the image sizes. Furthermore, by maintaining input values within a scale that the model is accustomed with, it normalises pixel values based on the mean and standard deviation of pixel values across the ImageNet dataset, thus helping in stabilising the learning process of the model.

Processed images are also transformed into PyTorch tensors by the feature extractor, making it easier to use them directly for model training. Rearranging the image dimensions to match the model's expected channel sequence is a part of this tensor transformation, which is essential to the functioning of the Vision Transformer. In addition, these actions will ensure every image is set to be processed by the neural network in the most efficient possible way, preserving consistency throughout the dataset and improving the model's capacity to identify patterns in the images.

4.3 *Balancing the Dataset*

Resampling method was used to alleviate the notable class imbalance and to make sure that the model is trained with equal class sizes of all stages of dementia. 'Non Demented' class was undersampled to match the second highest class size of 13,725 samples and two other low sized classes were oversampled in order to match the 'Very mild Dementia' class as the second highest class. This method maximises the variety of training examples that the Vision Transformer model gets to work with while also correcting the imbalance. Following resampling, 13,725 samples were evenly distributed among the four classes – 'Mild Dementia', 'Moderate Dementia', 'Very mild Dementia', and 'Non Demented'. This allows for a more impartial learning process.

For a robust model training, it requires efficient dataset management. In this scope, the balanced dataset was deliberately splitted to allow thorough training and evaluation of the ViT model. Explicitly, the dataset was split into 70% for training, 15% for validation, and 15% for testing. Along with offering enough data for testing and validating the model's performance across unseen images, this specific division enables comprehensive learning. Further, experiments involved conditioning the model with different proportions of the training set, using full size and subsets containing 80% and 60% of the balanced data. This methodology empowers the evalua-

tion of the model's performance sensitivity with respect to size of dataset, providing convincing insights into the robustness and scalability of the learned features.

4.4 *Data Loaders for Model Training*

From the PyTorch library, DataLoader class was used to generate iterators for the training, validation, and test datasets in order to handle and perform training on the images promptly. The loaders were constructed with a batch size of 32, enhancing memory usage and computational efficiency. Concerning to prevent the model from learning any unintended biases that may arise from the order of the data, the training dataset loader was set to shuffle. On the contrary, the loaders used for validation and testing were set up without shuffling to make sure consistent outcomes throughout the model evaluation stages.

4.5 *Model Architecture*

Vision Transformer (ViT) (Angyalfold, 2021) is a transformer encoder model, works like the BERT language model, which was pre-trained at 224x224 pixel resolution on a large set of supervised images with ImageNet-21k. The transformer principle, which is firstly and commonly applied in Natural Language Processing, is extended to the field of image classification by the ViT model more specifically medical image classification. The architecture handles the images as if they were words, by tokenizing images into a series of patches and processing them using the transformer's self-attention mechanism, which allows the model to capture global dependencies between various image segments.

The "vit-base-patch16-224-in21k" (Dosovitskiy et al., 2020), (Google, 2023) model splits each input image into patches of 16x16 pixels. After the division, positional encodings are added to these patches so that we can linearly introduce them in order to preserve locational information. The resulting sequence of embeddings is fed into a series of transformer layers, which analyse the image data by using self-attention. The model, which was originally pre-trained on the ImageNet-21k dataset (14 million images, 21,843 classes), is capable of classifying dementia stages from medical imaging and has a comprehensive learning of a wide range of image attributes.

4.6 *Training Process*

Prior to training, all images were scaled to 224x224 pixels and normalised by setting the mean and standard deviation for each RGB channel to (0.5, 0.5, 0,5). This normalisation was for the purpose of ensuring a constant distribution of input data, which aims to stabilise the model's training process. The pre-trained model was fine-tuned during 5 epochs with a batch size of 32.

4.7 *Comparison Model*

As a comparison model, we have used a simple Convolutional Neural Network (CNN) (Singh, 2023) with the full size of the dataset, which was obtained from the Open Access Series of Imaging Studies. The comparison model was published in Kaggle web site under title of "OASIS Alzheimer's Detection" however, the model was trained with 400 samples from the OASIS dataset. This dataset was resampled to the same target sample size used for the Vision Transformer Model.

Compared to our proposed model, we used the same proportions for training, validation, and test procedures: 70%, 15%, and 15%, respectively. Nevertheless, the preprocessing requirements for the CNN model differ from those of the ViT model. While images were configured to specific input dimensions of 224x224 pixels for processing by the ViT model, the CNN model that we used as an example was resized to 128x128 pixels for the Convolutional Neural Network model. Before the training and comparison, this input size resized to 224x224, as in Inception model (Szegedy et al., 2014), in order to make the comparison fair. This architecture allows CNN model to directly apply convolutional filters to the input image, leveraging spatial hierarchies. The model employs a sequential convolutional architecture that consists of convolutional layers with activation and batch normalisation, pooling layers to reduce the dimensionality and dropout layers for preventing overfitting.

Conversely, ViT models possess Natural Language Processing (NLP) based Transformer architecture that is adapted for image processing which utilises a sequence of self-attention layers where each patch of the input image is treated as a token. These tokens are processed in a similar way to words in a sentence in NLP tasks. The ViT model, which takes as inputs of 224x224 pixels, splits images into patches of 16x16, then processes these patches through multiple layers of self-attention, enabling the model to focus on the most informative parts of the image across all patches. By means of both computational and memory efficiency, CNN model generally offers better memory efficiency and faster performance for small

or medium sized datasets due to fewer parameters (Zhao et al., 2024). In contrast, the ViT model requires significantly higher computational resources in terms of GPU memory and processing power because of using a large number of parameters and especially the global nature of self-attention mechanism. Although it requires quite a lot more computational resources, the ViT model provides a better understanding when used with larger datasets (Heo, Seo, & Kang, 2023).

Typically, the Vision Transformer excels in environments where extensive training data is available and computational resources are not a limiting factor, showing strong ability to capture broader dependencies across the image. Contrariwise, CNNs are favored for faster classification in environments that offer limited data and resources.

Evaluation Metrics

To evaluate how well the ViT model classified the stages of dementia comparing to CNN model, several metrics were used:

- Accuracy: this metric offers a clear way for evaluating the model's overall performance in all classes by using true positives and true negatives
- Precision and Recall: Essential metrics for medical applications where false negatives and positives might have significant effects. Recall evaluates the model's capacity to find all relevant occurrences, whereas precision measures the accuracy of positive predictions.
- F-1 score: the harmonic mean of precision and recall, offering a balance between the two when evaluating model performance.

When analyzing as a whole, these metrics provide a thorough evaluation of the model's diagnostic abilities, which is crucial for ensuring accuracy and dependability when classifying medical images.

5 RESULTS

After testing the test proportion of unseen data from the dataset, the ViT model classified Moderate Dementia class without errors, achieving 100% from precision, recall and F-1 score. Similarly, other classes performed high scores from the F-1 metric ranging 96%-98% across the classes. In the CNN models classification report, Mild Dementia and Moderate Dementia classes performed well and scored 100% from the metrics, while the Very mild Dementia class presents 99% from precision and Non Demented from recall metric. As another testing metric for the Vision Transformer model, the confusion matrix displayed in table 3, is an exemplary performance

Classification Report	Precision	Recall	F-1 Score	Support
Mild Dementia	1.00	0.97	0.98	2108
Moderate Dementia	1.00	1.00	1.00	1998
Non Demented	0.98	0.94	0.96	2057
Very mild Dementia	0.92	1.00	0.96	2072
Accuracy			0.97	8235
Macro Avg	0.98	0.97	0.97	8235
Weighted Avg	0.98	0.97	0.97	8235

Table 1: ViT model Classification Report

Classification Report	Precision	Recall	F-1 Score	Support
Mild Dementia	1.00	1.00	1.00	2108
Moderate Dementia	1.00	1.00	1.00	1998
Non Demented	1.00	0.99	1.00	2057
Very mild Dementia	0.99	1.00	1.00	2072
Accuracy			1.00	8235
Macro Avg	1.00	1.00	1.00	8235
Weighted Avg	1.00	1.00	1.00	8235

Table 2: CNN model Classification Report

in differentiating between non-demented, very mild, mild, and moderate dementia stages. However, there were observable misclassifications primarily between the very mild and mild dementia categories, pinpointing an area ripe for model refinement. When classifying dementia stages into non-demented, very mild, mild, and moderate dementia phases, the confusion matrix represents outstanding results but also noteworthy misclassifications, mostly in the very mild and mild dementia groups, indicating a potential scope for model improvement.

	Predicted			
Actual	Non Demented	Mild Dementia	Very mild Dementia	Moderate Dementia
Non Demented	1928	5	124	0
Mild Dementia	5	2037	46	0
Very mild Dementia	5	1	2066	0
Moderate Dementia	5	0	2	1991

Table 3: The Vision Transformer model's Confusion Matrix

The model's accuracy recall across a range of sample sizes, 13,725, 10,980, and 8,235 is indicative of the model's scalability and adaptability. The Vision Transformer (ViT) model's training procedure showed rapid improvement over the initial epochs. Starting at an accuracy of 84.2% in the first epoch, the model's accuracy spiked to 98.44% by the third epoch.

This early performance spike demonstrates the ViT model's ability to rapidly learn from the training data and effectively adjust its weights and biases, which is crucial for tasks such as detecting the stages of dementia. Similarly, the validation accuracy improved, starting at 95.11% and peaking at 98.07% in the final epoch. The model obtained a peak training accuracy of 99.53% and validation accuracy of 99.22% with the full size dataset. This suggests that the ViT model consistently remains effective in terms of performance with respect to the dataset size, which is essential for real-world applications where researchers may face some limitations such as data availability.

CNN model's training process demonstrated astonishingly faster performance than ViT model and similar training and validation accuracy scores. While both models are showing high accuracy scores from the first epoch, loss metric for ViT model is noticeably lower than CNN. In the confusion matrix, CNN outperforms ViT by classifying the classes with lower mistakes between classes. The most misclassified class was 'very mild dementia' class as it shows similarity with ViT model.

6 DISCUSSION

The expectations from the Vision Transformer model, the adaptability of a neural network architecture satisfies at large scope. As the Vision Transformer model requires slightly different pre-processing techniques than the traditional Convolutional Neural Network model, some of the pre-processing steps used were similar. More than the differences in the architecture and image processing procedures, the idea behind why the proposed model is used for medical image classification stands strong due to the underlying mechanism that Vision Transformer holds. Self-attention mechanism not only presents innovation for image classification and recognition but also proves the adaptability of different task based architectures into the computer vision field (Guo et al., 2022). The consistent performance of the Vision Transformer model across different sample sizes of the OASIS dataset– from 13,725 to 8,235– demonstrates its robustness in the aspects of scalability and adaptability.

Represented feature of the proposed model is a valuable remark in medical imaging because of varying dataset availability in the field. One of the model's strengths is to maintain high accuracy from both training and validation processes since initial epochs. By achieving nearing 99% from these processes, the model proves its potential of effectiveness in diverse clinical environments. Obtained loss metrics results from training and validation making us sure that the model does not overfit or underfit the training set and validation set. This captures the model's ability of

well-generalisation of unseen data, meaning the proposed model does not memorise the patterns of the images but learns from them. This also indicates that the model is neither too complex for the task nor too simple which is an indicative feature of a well-suited architecture for medical image classification. Vision Transformer model demonstrates remarkable learning efficiency by starting its accuracy for correct classification of data points from 84.2% to 98.44% within three epochs of training. This significant increase in accuracy underlines ViT's capability to adjust its parameters efficiently which is essential for tasks requiring high precision such as the detection and staging of dementia in Alzheimer's disease patients.

In critical cases where early and precise diagnosis is vital, the proposed model's swift adaptation offers an important potential in timely medical interventions. On the other hand, CNN models happen to be showing similarly high accuracy for the dataset from the first epoch during the training process with accuracy of 82% and increasing sharply to 98% in the fourth epoch. Acquiring 0.85 from loss function of the first epoch of training, substantial decrease on the third epoch for loss function is a strong indicator of CNN model that it successfully achieves to learn the data rapidly. High accuracy and low loss score from the first epoch of the validation process also prove the model's architecture to be well-suited for the task. In the testing process, performance analysis metrics show equally successful results for classifying the unseen data by the model into four desired classes for the Vision Transformer model. As evaluated metrics, precision rate for each class demonstrates significant high scores from 92% to 100% across the classes, especially an excellent performance for the classes Moderate Dementia and Mild Dementia. While the lowest precision rate is from Very mild Dementia, recall rate obtained from Very mild Dementia class and Moderate Dementia are excellent with 100%. Also, Mild Dementia and Non Demented classes follow high recall rates of Very mild and Moderate Dementia classes with respectively, 97% and 94%. These high scores from precision and recall yields overall high scores from F-1 score as it provides a harmonic mean of precision and recall, ranging from 96% to 100% across the classes. CNN model also demonstrates impressive results from precision, recall, and F-1, proving the model to be highly effective for such tasks with specified data size.

The confusion matrix represents exceptional accurate classification results when it's observed for each class individually. Although the model's successful performance, there are some misclassified data points that need to be pointed out. These misclassifications are mostly observable in the Non Demented class with 124 data points that were predicted as Very mild Dementia, following this class, some of Mild Dementia data points

were misclassified as Non Demented or Very mild Dementia. Similar misclassifications were obtained from CNN, as mostly misclassifying Non Demented class to Very mild Dementia class with 16 data samples. These observed misclassifications from Mild Dementia and Non Dementia stages point out an important challenge for the ViT model, highlighting the need for further improvements in the ViT model's feature extraction competence.

Enhancing the model's sensitivity for subtle neuroanatomical differences is crucial for early detection of early-stage dementia could offer more improved diagnostic accuracy and patient outcomes. While both ViT and CNN models present high accuracy, the ViT model predominates in handling complex image data with its lower loss metrics which indicates more efficient error minimization. Nevertheless, CNN's computational efficiency and faster performance suggest it is an applicable option in resource-constrained settings. This difference highlights the trade-off between computational efficiency and diagnostic precision that must be considered in clinical implementations. In the medical aspect, brain scanning methods require high computational resources as well as financial resources where it is not possible to easily find and access for most of the patients that suffer from Alzheimer's disease. ViT's substantial computational requirements compared to traditional CNNs presents a practical challenge in resource-limited settings. Feasibility of employing such computational demanding models may not be easily implemented without required resources.

7 CONCLUSION

The study affirms the potential of Vision Transformers in improving the diagnostic processes for Alzheimer's disease through exceptional image analysis ability. Even though requiring higher computational resources, ViT models propose a significant advancement in detecting subtle changes in neuroanatomy associated with various dementia stages. Integration of Transformer models into diagnostics of clinical neuroimaging methods potentially improve the accuracy of Alzheimer's staging detection, thereby alleviating earlier and more precise interventions. This intervention can substantially alter the disease's management, underscoring patients care and therapeutic results.

Regarding these conclusions, research objectives that have been proposed can be answered in such a way that the ViT model has the potential of outperforming the CNN model while being used on a larger dataset and having high computational resource settings. For small dataset settings, CNN models still offer memory and computational efficient solutions. Subtle changes in neuroanatomical level can be decisive for diagnosis of some

of the neurodegenerative diseases due to the ViT model's underlying mechanisms, this aspect definitely should be taken into consideration for the ViT model to be used in medical image analysis. Further research should focus on optimization of ViT models to reduce computational demands and increase the accuracy of differentiating similar stages of dementia. In other neurodegenerative disease research, application of ViT models could also expand the impact of innovative Transformer models in neurological analyses.

REFERENCES

Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S., Karthikesalingam, A., King, D., ... Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, *4*(1), 65.

Aithal, N. (2024). *Images oasis dataset.* https://www.kaggle.com/datasets/ninadaithal/imagesoasis. (Accessed: 2024-06-23)

Angyalfold. (2021). *Hugging face bert with custom classifier (pytorch).* Retrieved from https://www.kaggle.com/code/angyalfold/hugging-face-bert-with-custom-classifier-pytorch?scriptVersionId=67388905&cellId=19 (Kaggle Kernel)

Arafa, D. A., Moustafa, H. E.-D., Ali, H. A., Ali-Eldin, A. M., & Saraya, S. F. (2024). A deep learning framework for early diagnosis of alzheimer's disease on mri images. *Multimedia Tools and Applications*, *83*(2), 3767–3799.

Basu, K., Sinha, R., Ong, A., & Basu, T. (2020). Artificial intelligence: How is it changing medical sciences and its future? *Indian journal of dermatology*, *65*(5), 365–370.

Cummings, J., Lee, G., Nahed, P., Kambar, M. E. Z. N., Zhong, K., Fonseca, J., & Taghva, K. (2022). Alzheimer's disease drug development pipeline: 2022. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *8*(1), e12295.

DeStigter, K., Pool, K.-L., Leslie, A., Hussain, S., Tan, B. S., Donoso-Bach, L., & Andronikou, S. (2021). Optimizing integrated imaging service delivery by tier in low-resource health systems. *Insights into Imaging*, *12*, 1–11.

DeTure, M. A., & Dickson, D. W. (2019). The neuropathological diagnosis of alzheimer's disease. *Molecular neurodegeneration*, *14*(1), 32.

Diogo, V. S., Ferreira, H. A., Prata, D., & Initiative, A. D. N. (2022). Early diagnosis of alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy*, *14*(1), 107.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

El-Assy, A., Amer, H. M., Ibrahim, H., & Mohamed, M. (2024). A novel cnn architecture for accurate early detection and classification of alzheimer's disease using mri data. *Scientific Reports*, *14*(1), 3463.

Face, H. (2023). *Fine-tune vit.* https://huggingface.co/blog/fine-tune-vit. (Accessed: 2024-06-23)

Filipiuk, M., & Singh, V. (2022). Comparing vision transformers and convolutional nets for safety critical systems. In *Safeai@ aaai.*

Finnema, S. J., Nabulsi, N. B., Eid, T., Detyniecki, K., Lin, S.-f., Chen, M.-K., ... others (2016). Imaging synaptic density in the living human brain. *Science translational medicine*, *8*(348), 348ra96–348ra96.

Ftoutou, A., Majdoub, N., & Ladhari, T. (2023). Alzheimer's disease classification using deep learning. In *2023 ieee international conference on artificial intelligence & green energy (icaige)* (pp. 1–6).

Ghaffari, H., Tavakoli, H., & Pirzad Jahromi, G. (2022). Deep transfer learning–based fully automated detection and classification of alzheimer's disease on brain mri. *The British journal of radiology*, *95*(1136), 20211253.

Google. (2023). *Vit base patch16 224.* https://huggingface.co/google/vit-base-patch16-224-in21k. (Accessed: 2024-06-23)

Grabher, B. J. (2018). Alzheimer's disease and the effects it has on the patient and their family. *Journal of Nuclear Medicine Technology*.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., ... Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media*, *8*(3), 331–368.

Helaly, H. A., Badawy, M., & Haikal, A. Y. (2022). Deep learning approach for early detection of alzheimer's disease. *Cognitive computation*, *14*(5), 1711–1727.

Heo, J., Seo, S., & Kang, P. (2023). Exploring the differences in adversarial robustness between vit-and cnn-based models using novel metrics. *Computer Vision and Image Understanding*, *235*, 103800.

International, A. D. (2023). *World alzheimer report 2023: Continuing to live with dementia: New perspectives on resilience and the way forward.* London. Retrieved from https://www.alzint.org/u/WorldAlzheimerReport2023.pdf

Jagadeeswari, M., Priya, S. P., Athira, K., Dhanalakshmi, M., & Shree, P. G. (2022). Prediction and classification of alzheimer's disease using deep learning. In *2022 3rd international conference on electronics and sustainable communication systems (icesc)* (pp. 834–839).

Komorowski, P., Baniecki, H., & Biecek, P. (2023). Towards evaluating explanations of vision transformers for medical imaging. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3725–3731).

Krishnapriya, S., & Karuna, Y. (2023). Pre-trained deep learning models for brain mri image classification. *Frontiers in Human Neuroscience*, *17*, 1150120.

Kumar, S. A., & Sasikala, S. (2023). Enhanced alzheimer's disease classification using multilayer deep convolutional neural network-based experimentations. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *47*(4), 1595–1621.

Kumar, Y., Koul, A., Singla, R., & et al. (2023, 7). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, *14*, 8459–8486. Retrieved from https://doi.org/10.1007/s12652-021-03612-z (Received: 04 December 2020, Accepted: 18 November 2021, Published: 13 January 2022) doi: 10.1007/s12652-021-03612-z

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, *11*, 1273253.

Li, X., Feng, X., Sun, X., Hou, N., Han, F., & Liu, Y. (2022). Global, regional, and national burden of alzheimer's disease and other dementias, 1990–2019. *Frontiers in Aging Neuroscience*, *14*, 937486.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Marshall, C. R., & Uchegbu, I. (2022). Artificial intelligence for detection of alzheimer's disease: demonstration of real-world value is required to bridge the translational gap. *The Lancet Digital Health*, *4*(11), e768–e769.

Matsoukas, C., Haslum, J. F., Söderberg, M., & Smith, K. (2023). Pretrained vits yield versatile representations for medical images. *arXiv preprint arXiv:2303.07034*.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, *13*(9), 5521.

Mggdadi, E., Al-Aiad, A., Al-Ayyad, M. S., & Darabseh, A. (2021). Prediction alzheimer's disease from mri images using deep learning. In *2021 12th international conference on information and communication systems (icics)* (pp. 120–125).

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S.,

& Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, 102802.

Singh, R. (2023). *Oasis alzheimer's detection.* Retrieved from https://www.kaggle.com/code/rahulsingh787/oasis-alzheimer-s-detection/notebook (Kaggle Kernel)

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., ... Rabinovich, A. (2014). Going deeper with convolutions. *CoRR*, *abs/1409.4842*. Retrieved from http://arxiv.org/abs/1409.4842

Uddin, M. S., Tewari, D., Al Mamun, A., Kabir, M. T., Niaz, K., Wahed, M. I. I., ... Ashraf, G. M. (2020). Circadian and sleep dysfunction in alzheimer's disease. *Ageing Research Reviews*, *60*, 101046.

Venere, M., Zadeh, G., Puduvalli, V., & Haynes, C. (2020). *Sno 25th anniversary history series: Ancillary and satellite meetings of the society for neuro-oncology* (Vol. 22) (No. 9). Oxford University Press US.

Wang, Y., Chen, K., Zhang, Y., & Wang, H. (2024). Medtransformer: Accurate ad diagnosis for 3d mri images through 2d vision transformers. *arXiv preprint arXiv:2401.06349*.

Wattmo, C., & Wallin, Å. K. (2017). Early-versus late-onset alzheimer's disease in clinical practice: cognitive and global outcomes over 3 years. *Alzheimer's Research & Therapy*, *9*, 1–13.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, *57*(4), 1–43.