Kathford Journal of Engineering and Management,
ISSN: 2661-6106.
© 2024 *Kathford Journal of Engineering and Management* VOL. 4 ISSUE 1 DECEMBER 2024, PP. 31-47

Leveraging Vision Transformers for Early Detection of Alzheimer's disease Through MRI Image Analysis

Ayush Pandey[1,], Ayushma Pandey[1], Shashanka Poudel[2]*

[1]*Department of Computer and Electronics, Communication, and Information Engineering, Kathford International College of Engineering and Management (Affilitiated to Tribhuvan University), Balkumari, Lalitpur, Nepal*
[2]*Department of Computer and Electronics, Communication, and Information Engineering, Kantipur Engineering College (Affilitiated to Tribhuvan University), Dhapakhel, Lalitpur, Nepal*

Corresponding author: *ayushpandey.076@kathford.edu.np*

ABSTRACT—Alzheimer's disease (AD) poses challenges for early recognition due to subtle and overlapping symptoms, making timely care and intervention difficult. This work presents a new method for identifying Alzheimer's disease by combining magnetic resonance imaging (MRI) data with a Vision Transformer (ViT) model. In this research, we develop a method that leverages MRI scans to classify Alzheimer's disease (AD) into four distinct stages: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. To improve the identification of clinical and structural alterations in the brain linked to AD, sophisticated characteristics are retrieved, including voxel-intensity patterns, grey matter volume, and cortical thickness. These characteristics allow for a more accurate representation of MRI data when paired with sophisticated processing methods. Additionally, even in situations where data is scarce, transfer learning is used to maximize the Vision Transformer (ViT) model's classification accuracy, especially when it comes to differentiating between neighboring stages of dementia. The ViT model, a cutting-edge deep learning technique, shows strong performance and dependability in identifying AD phases. Our findings highlight its effectiveness, showcasing a significant improvement in diagnostic accuracy compared to existing methods. The accuracy with which the ViT model can distinguish between different phases of AD emphasizes both its suitability for handling complicated MRI data and its promise for early diagnosis and detection.This paper offers a promising method for enhancing MRI image processing for Alzheimer's disease diagnosis through sophisticated feature extraction, transfer learning, and the use of Vision Transformers (ViTs). Present authors believe this method successfully tackles the difficulties of early and precise detection, making it essential for appropriate medical image analysis.

KEYWORDS—ViT, MRI, Demented, Non-Demented, Mild Demented, Clinical contexts

1. INTRODUCTION

Individually and over time, the pace of decline with age appears to remain constant: it appears that people are living longer and in better health as they age (Vaupel, 2010). Projections on a global scale indicate that by 2050, over 21% of the population will be over 60, creating a sizable older demographic of two billion (Samir K. C., 2017) (Mohammed G. Alsubaie, 2024). This trend reflects a significant global increase in the number of elderly individuals. After receiving an Alzheimer's diagnosis, individuals typically have only four to eight years left in life. This disorder affects 1 in 10 individuals over the age of 65 on average, although rare cases have been diagnosed in younger individuals, even in their 20s. It remains the leading cause of dementia in the elderly (Pradhan, 2021). The number of persons suffering from Alzheimer's disease, which is characterized by memory loss and cognitive impairment, rises along with the aging population. Alzheimer's disease (AD) is a chronic neurodegenerative illness that progressively damages brain tissue, impairing thinking and memory abilities, and ultimately making them unable to carry out even the most basic tasks. AD patients gradually lose their physiological capabilities and develop dementia, which ultimately results in death (Odusami, 2023) (DeTure, 2019). Understanding the pathophysiology of AD is still exceedingly difficult (Odusami, 2023). The term "very mild demented" refers to the early symptoms. It is difficult for those who have this illness to carry out daily tasks like driving and cooking. The symptoms are not immediately apparent in the early stages and can include trouble remembering names, misplacing key belongings, difficulty planning, etc. The longest stage of Alzheimer's disease is the middle stage, during which time patients may experience extreme mood swings, disorientation, impulsivity, lack of focus, difficulty recognizing objects, etc. The final phase is the most intense (Pradhan, 2021). The most obvious signs are impaired judgment, poor sense of direction, visual issues, inability to properly connect with others, increased susceptibility to infections, and short-term memory loss (Pradhan, 2021). As the symptoms worsen, a mildly demented stage is reached, followed by a moderately demented stage.

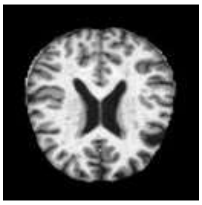
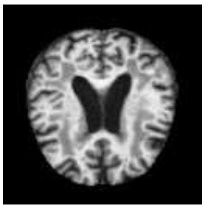
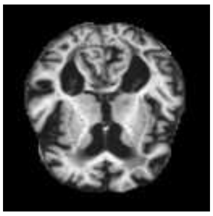
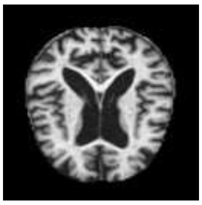
Although there is currently no known cure for AD, there are a number of treatments that can help control symptoms and slow the illness's progression (Almufareh, 2023). Because it can assist people and their families with long-term care planning, access to therapies and support services, and future planning, early detection of AD is crucial. The symptoms of AD can mimic those of other illnesses, and they may even be viewed as a typical aspect of aging, making early detection difficult (Almufareh, 2023). Deep learning is an artificial intelligence technology that has been used to forecast and diagnose diseases by recognizing important aspects of medical images (Aggarwal, 2021) (Varoquaux, 2022) (Salahuddin, 2022). Convolutional neural networks (CNNs) are a well-known example of deep learning applications. While they are mostly employed for picture identification tasks across several domains (Ayush Pandey, 2023), they have also proven to be a useful diagnostic tool for AD (Acquarelli, 2022) (Samhan, 2022) (Kang, 2021). Prior research has primarily employed CNN-based backbones for feature extraction; however, Vision Transformers (ViT) have lately surfaced as a potent substitute for CNNs in image diagnostic classification. ViTs use a self-attention mechanism to extract global contextual information and long-range relationships from images, in contrast to CNNs, which rely on localised receptive fields. ViTs perform better thanks to this feature, which helps them better analyse subtle structural and pathological changes in medical imaging, especially when tasks involving complicated or sparse datasets are involved.

This paper proposes a model that addresses class imbalance using the class weight technique and extracts features through the Vision Transformer (ViT). Despite being trained on a relatively small dataset, the model achieves a noteworthy 71% accuracy when classifying MRI scans into four dementia stages: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented, as summarized in Table 1. Furthermore, our approach demonstrates competitive performance with reduced parameters and computational cost, making it a viable alternative to previous models reliant on more resource-intensive designs. The model's effectiveness is further increased by methods like crop non-zero preprocessing, which makes it more effective than more conventional CNN-based models like VGG19 and DenseNet169, which frequently need bigger datasets to attain comparable performance.

1.1 LITERATURE REVIEWS

Some researchers employ the traditional machine learning approach for classification after using CNN to extract features such as texture patterns, gray matter density, and structural atrophy from MRI scans. Suk et al. (2017) used CNN to apply clinical decision-making to the target-level representations produced by sparse regression (Suk, 2017). Feng et al. (2020) used MRI and 3D CNN to classify AD based on MRI pictures. An SVM was used in place of SoftMax as the classifier, and the 3D-CNN-SVM model outperformed 2D-CNN and 3D-CNN in terms of classification performance. Researchers have contributed multiple CNN backbones that achieve state-of-the-art performance in numerous tasks,contributing to CNN's flourishing in computer vision (Feng, 2020). Helaly et al. (2022) proposed a deep learning approach that utilizes CNNs for the

Table 1. MRI Images for five Alzheimer's stages and their features

Stages	Name	Explanation	Sample image
1	Non_Demented	No indications of a deterioration in cognition. The person is not exhibiting any obvious cognitive or memory issues and is functioning properly.	
2	Very_Mild_Demented	Early indicators of deteriorating cognitive function. Minor memory loss, particularly when it comes to names or recent events, but not enough to interfere with day-to-day activities.	
3	Mild_Demented	Noticeable memory problems, disorientation, and difficulty performing intricate tasks like organising or planning. The majority of personal care tasks can still be completed by the person, albeit daily living is impacted.	
4	Moderate_Demented	More severe memory loss, disorientation regarding location and time, trouble identifying known faces, and a greater need on others for daily care. There could be behavioural shifts, such agitation or roaming.	

early detection of AD, achieving significant accuracy in multi-class classification tasks using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Their study emphasizes the importance of reducing computational complexity while maintaining high accuracy, aligning with the findings of other researchers in this field (Helaly, 2022). LeNet-5, a widely used CNN architecture, was employed by Sarraf et al. to classify AD from the NC brain (binary classification) (Sarraf, 2016). Wang et al. examined an eight-layer CNN architecture. Convolutional layers used six layers to extract features, and classification used two completely linked layers. The outcomes demonstrated the outstanding performance of the Leaky Rectified Linear Unit (LReLU) with max-pooling (Wang, 2018). For the diagnosis of AD, Khvostikov et al. employed a 3D Inception-based CNN. The technique relied on the integration of Diffusion Tensor Imaging (DTI) and Structural Magnetic Resonance Imaging (SMRI) on hippocampus Regions of Interest (RoI). They contrasted that method's performance with the network based on AlexNet. 3D Inception reported better performance than AlexNet (Khvostikov, 2018). Kundaram and Pathak (2021) proposed a deep convolutional neural network (DCNN) model for Alzheimer's disease classification using MRI samples, achieving a remarkable 98.57% accuracy, which underscores the efficacy of deep learning in medical image analysis (Kundaram, 2021). Pradhan et al. (2021) proposed a model using VGG19 and DenseNet169 for Alzheimer's disease classification based on MRI images, providing a comparative analysis of these CNN architectures for classifying patients with mild, moderate, or no Alzheimer's disease (Pradhan, 2021).Suresha et al. used a rectified adam optimizer and a deep neural network to classify images into Normal, AD, and MCI, respectively. They achieved a high accuracy of 99.5% by using a Histogram of Oriented Gradients to extract features (Suresha, 2020). Almadhoun and Abu-Naser (2021) utilized the Xception CNN architecture combined with traditional classifiers to detect early-stage Alzheimer's disease from brain imaging data (Almadhoun, 2021).

From the literature review, it was observed that traditional machine learning methods often rely on handcrafted feature extraction, such as texture patterns and structural atrophy, which can be limited in capturing subtle changes in brain morphology. Studies utilizing CNNs have demonstrated improved performance, but they are often constrained by computationally intensive architectures and require large datasets for effective training. Moreover, class imbalance and difficulty in distinguishing adjacent stages of dementia remain significant challenges. These observations emphasize the need for a more precise and efficient approach. By leveraging Vision Transformers (ViTs) with advanced feature extraction capabilities and transfer learning, this study addresses these gaps, offering improved accuracy even with limited data.

2. COMPUTATIONAL AND THEORETICAL DETAILS

2.1. Vision Transformer (ViT)

In this paper, we apply the Vision Transformer (ViT) model as shown in Figure 1 sophisticated capabilities to the analysis and interpretation of medical imaging data in the task of Alzheimer's disease identification. The ViT model takes a different approach from traditional convolutional neural networks (CNNs) by using self-attention mechanisms in place of convolutional layers. This methodological shift enables the ViT to achieve superior performance across various computer vision applications, including medical image analysis. Because CNNs can capture local spatial data through convolutional processes, they have historically been the architecture of choice for image analysis. Nevertheless, by processing images using self-attention mechanisms, the Vision Transformer marks a paradigm change. This method is inspired by natural language processing (NLP) transformers, which are particularly good at managing text sequences. Rather than treating images as a grid of individual pixels, the ViT model interprets them as a series of fixed-size patches. With this modification, the model can process images similarly to how NLP processes text sequences. Through the process of picture segmentation and discrete token treatment, the ViT model can employ self-attention mechanisms to effectively capture long-range linkages and global dependencies inside the image. This capacity is especially useful for medical imaging since accurate diagnosis depends on the ability to comprehend context across different sections of a picture.

2.1.1 Model Architecture and Preprocessing

We used a pre-trained ViT model with a 16x16 pixel patch size for Alzheimer's detection. The amount of context and detail that the model can collect depends on the patch size selection, which is crucial. While smaller patches might concentrate on finer details, larger patches might collect more contextual information. A 16x16 patch size is a good compromise between preserving global context and capturing sufficient detail. PyTorch, a popular deep-learning framework renowned for its adaptability and extensive feature set, was utilized to create the ViT model. We took advantage of the ViT model's prior training on extensive datasets like ImageNet by utilizing a pre-trained version of the model. When training data is scarce, this transfer learning strategy is very helpful because it enables the model to build upon pre-existing knowledge.

2.1.2 Model Fine-Tuning and Training

By repeating the training process on a fresh dataset, a pre-trained model can be fine-tuned to a new, targeted task. The ViT model is optimized for 50 epochs in our Alzheimer's detection task. To better match the patterns and features of our dataset, this step entails modifying the model's parameters. During fine-tuning, several strategies are used to improve the model's performance.

- **Data augmentation** is the process of creating different versions of the training images by applying techniques like color correction, flipping, and rotation. By providing the model with a wide variety of input variables, data augmentation helps to improve the model's capacity to generalize and lowers the risk of overfitting.
- **Learning Rate Scheduling:** This technique entails changing the training's learning rate. The learning rate usually starts off high and then steadily declines 2.1.over time. This tactic keeps the model from overshooting its ideal weights and contributes to more steady convergence.

2.2 Data Transformation Pipeline

A critical step in getting picture data ready for the Vision Transformer model is data transformation. By following this procedure, photos are guaranteed to be scaled and formatted correctly for efficient model training and assessment. Several crucial operations are included in the transformation pipeline:

- **Cropping Non-Zero Regions:** A custom cropping function is used to eliminate any empty or unnecessary areas surrounding the image's content so that the viewer can concentrate on the image's pertinent material. The process involves locating and obtaining the bounding box of non-zero areas in the picture.
- **Resizing:** To guarantee consistency in input size for the ViT model, photos are scaled to a fixed dimension, usually 224x224 pixels, after cropping. For photos to be processed consistently, this standardization is necessary. The images used in this study are sourced from Kaggle's and all necessary courtesies and permissions have been obtained for their use.
- **Conversion to Tensor:** Tensors are created from images, as this is the format that PyTorch models require as input. To prepare the data for training,

this phase additionally scales pixel values from the range [0, 255] to [0, 1].

- **Normalization:** To aid in stabilizing and speeding up the training process, pixel values are normalized to have a zero mean and a unit variance. By applying normalization, you can make sure that the model gets data that is scalable and fits the particular features of the dataset.

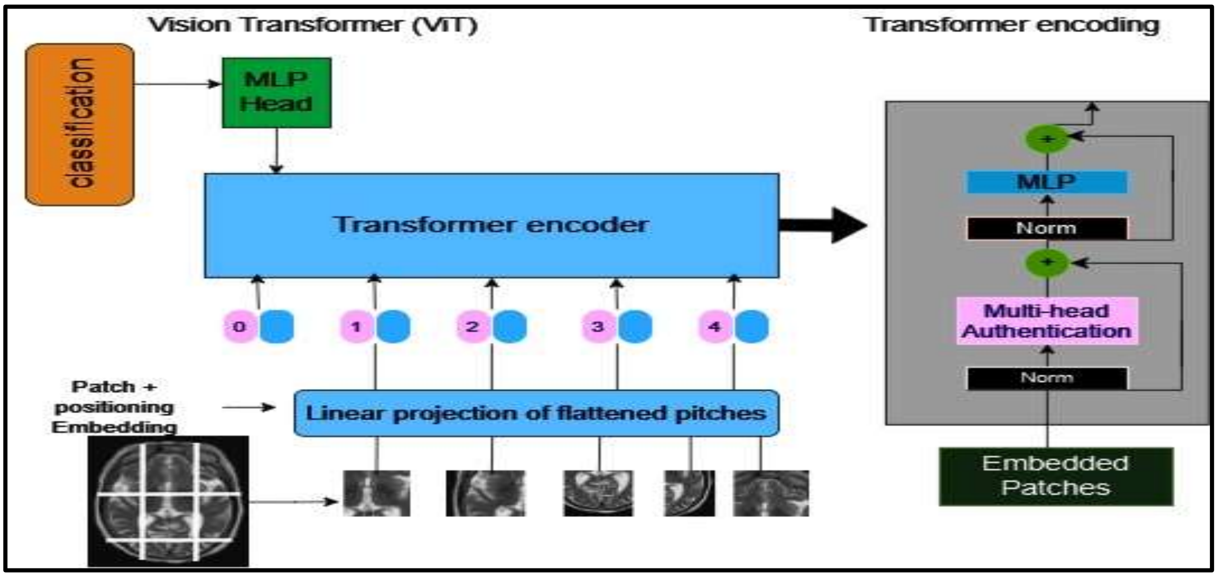


Figure 1. A ViT basic block

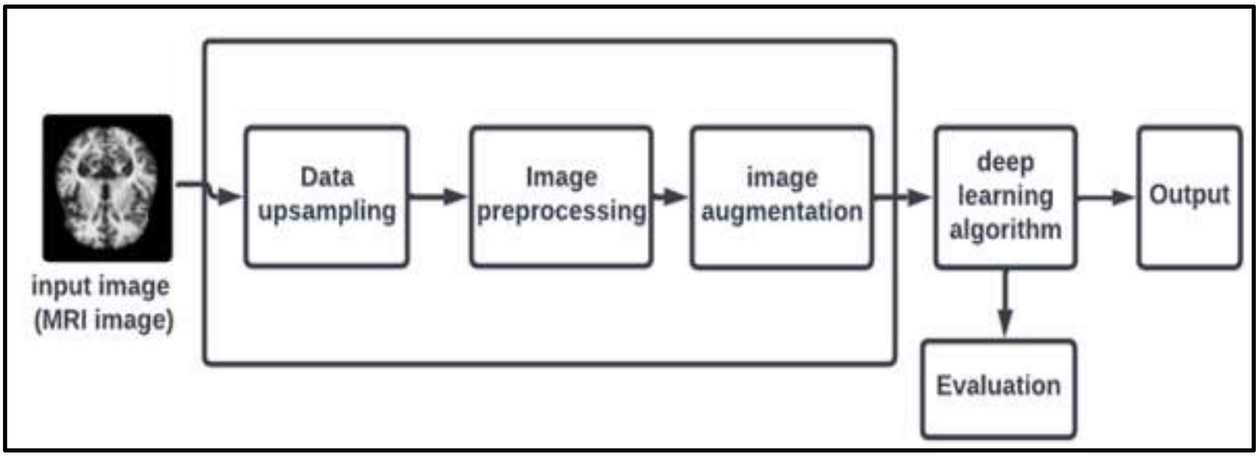


Figure 2. Process Model

The Hugging Face Transformers library, which offers an effective and user-friendly interface for managing transformer models, allows the Vision Transformer (ViT) model to be smoothly incorporated into our workflow. This library supports different stages of model training and evaluation and makes the process of adding pre-trained models easier. We use common metrics like confusion matrices and accuracy to assess the effectiveness of our improved ViT model. A comprehensive evaluation of the model's performance is provided by accuracy, which quantifies the percentage of correctly classified cases in relation to the total number of samples. Confusion matrices, on the other hand, offer a thorough analysis of the model's predictions made for various classes, making it easier to pinpoint the precise areas where the model works well or where it still needs to be improved.

There are several benefits of employing the Vision Transformer to identify Alzheimer's. Through the use of convolutional layers, standard convolutional neural networks (CNNs) mostly learn local features; however, ViT's self-attention mechanism enables a thorough comprehension of the complete visual context. This global viewpoint is especially crucial in medical imaging, where precise diagnosis depends on the ability to identify minute patterns and correlations between various image regions. Furthermore, the intrinsic flexibility and scalability of the transformer design allow it to be customize for a wide range of jobs and datasets. ViT may be tailored for certain applications by changing factors like model depth and patch size, which increases its

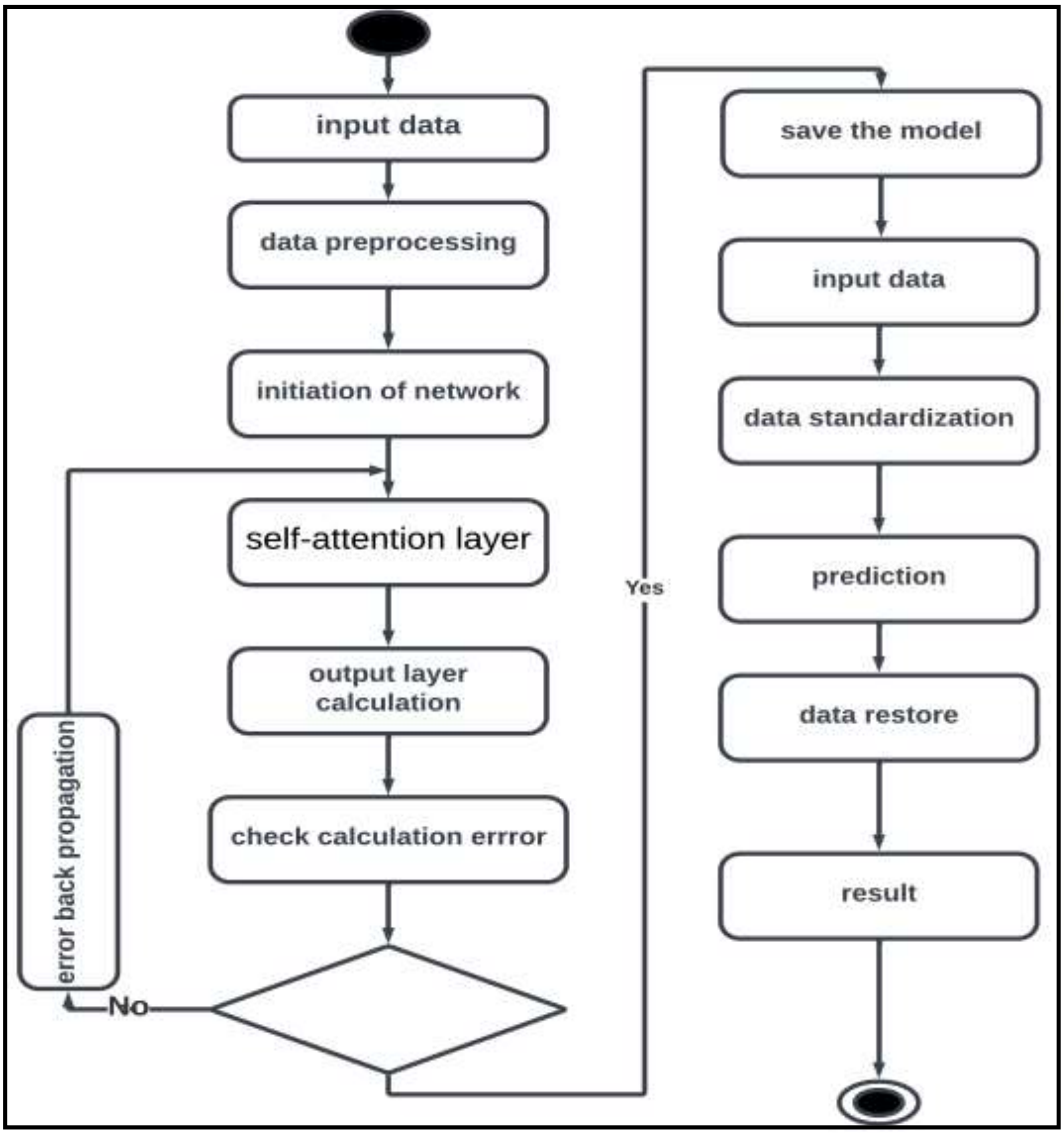


Figure 3. Activity Diagram

adaptability in Alzheimer's detection. Moreover, ViT's self-attention mechanism helps competitors perform better in picture classification tasks by identifying intricate patterns and long-range dependencies, which improves the precision and dependability of illness diagnosis.

2.3 Process Model

As shown in Figure 2, the process model of Alzheimer's disease is explained below:

- **Input Image (MRI Images):**

To diagnose Alzheimer's disease, the model requires an input image (brain scan). This could include MRI scans or other pertinent brain imaging for the diagnosis of Alzheimer's.

- **Data Upsampling:**

Data upsampling is done to make sure the model gets enough samples from each class to train from, due to possible class imbalances in the dataset (e.g., different stages of Alzheimer's disease or different retinal diseases). By balancing the datasets, this phase makes sure that every class is fairly represented during training.

- **Image Preprocessing:**

Preprocessing includes clearing and getting the photos ready for additional processing. When detecting Alzheimer's disease:

Cropping Non-Zero Regions: To draw attention to the information that has useful properties, non-relevant portions are cropped from the image.

Resizing: Pictures are shrunk to 224 by 224 pixels, which is the normal size. This guarantees that the Vision Transformer model's input size is consistent. Conversion to Tensor: Pixel values are scaled to the range [0, 1] after the resized photos are transformed into tensors, the PyTorch model input format.

Normalization: To stabilize the training process, the pixel values are normalized to have a zero mean and a unit variance.

Image Augmentation:

To create variants of the training images, data augmentation techniques including flipping, rotating, and altering colours are used. By guaranteeing the model sees a wide variety of input data during training, this helps avoid overfitting.

- **Deep Learning Algorithm (ViT Model):**

Images that have been enhanced and preprocessed are sent into a Vision Transformer (ViT) model. A pre-trained ViT model with a 16x16 patch size is utilized to balance context and detail capture for Alzheimer's identification. Unlike conventional convolutional neural networks (CNNs), which concentrate on local features, the ViT model makes

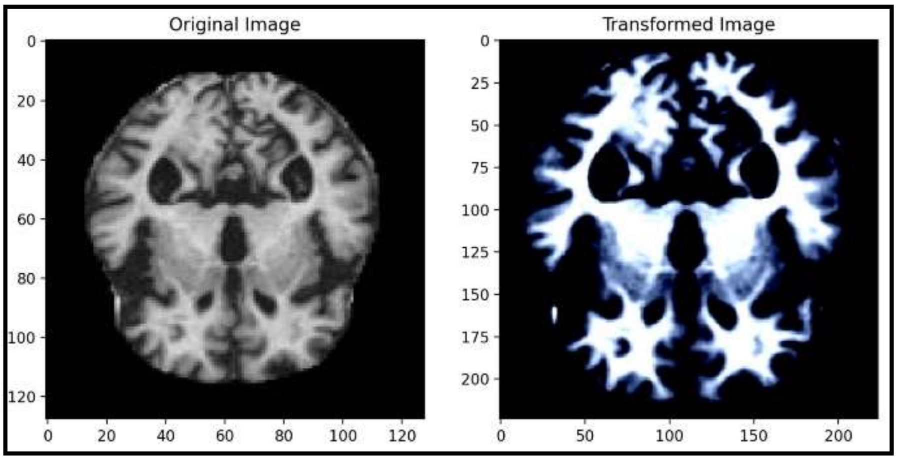


Figure 4. Image before and after transforming

use of self-attention processes to take into account the image's global context.

Model Fine-Tuning: To maximize the training process, the ViT model is tweaked using learning rate scheduling over 50 epochs of Alzheimer's detection data.

- **Evaluation:**

The performance of the model is assessed using common-metrics:

Accuracy: Calculates the percentage of photos that are correctly classified.

Confusion Matrix: This tool helps determine where the model excels and where it can falter by breaking down predictions made for various classes.

Additional Metrics: The model may also be evaluated using the F1-score, precision, and recall.

- **Output:**

A prediction (such as a diagnosis or classification) based on the input image is the model's output. Based on the brain imaging, this might be a sign of the disease's stage for Alzheimer's identification.

2.4 Activity Diagram

The activity diagram of the proposed model is illustrated in Figure 3.

- **Input Data (MRI Images):**

The first input the system receives is MRI brain scans, which are the main source of information for diagnosing Alzheimer's disease.

- **Data Preprocessing:**

To get them ready for the ViT model, the MRI images are preprocessed. One of the possible preprocessing steps is :

  - Resize the MRI pictures to a standard dimension.
  - Normalizing the values of pixel intensity.
  - Techniques for smoothing or reducing noise to eliminate unnecessary information.
  - Augmentation (if required) to add rotations or flips to the training set and hence boost its variability. By doing this, the model is guaranteed to have clean, consistent data.

- **Initiation of the Network (ViT Architecture Initialization):**

The Vision Transformer model is initialized at this point. To process the MRI pictures through transformer layers, they must first be divided into smaller patches, embedded, and ready for processing. Multiple layers of self-attention mechanisms, which are essential for comprehending the connections between various MRI regions, will make up the ViT model.

- **Self-Attention Layer:**

The self-attention mechanism, which enables the model to concentrate on particular brain regions that are more pertinent for Alzheimer's disease detection, is the central component of the Vision Transformer design. These areas may serve as early warning signs of plaques, brain shrinkage, or other neurodegenerative abnormalities.

By focusing on distinct areas of the image, the model gains an understanding of which regions of the MRI scan are essential for identifying Alzheimer's early warning indicators.

- **Output Layer Calculation:**

A series of self-attention layers are applied to the image before the final output layer produces predictions. In this instance, the forecast is a categorization of the patient's state of health or Alzheimer's disease symptoms.

- **Check Calculation Error:**

By contrasting the model's projected classification—whether or not Alzheimer's—with the ground truth labels from the dataset, the system calculates the error. This stage calculates the model's deviation from accurate prediction.

Table 2. Confusion Matrix

	Predicted: Non_Demented	Predicted: Mild_Demented	Predicted: Very_Mild_Demented	Predicted: Moderate_Demented
True: Non_Demented	202	11	75	0
True: Mild_Demented	2	55	23	0
True: Very_Mild_Demented	38	17	146	1
True: Moderate_Demented	0	0	1	5

- **Error Backpropagation (if error exists):**

Backpropagation takes place if the prediction contains a sizable error. Gradient descent is used to update the weights in the ViT model, which helps the model become more predictive over time.

- **Save the Model (if error is minimized):**

The trained model is saved after it has learned well, which is after multiple training epochs and minimal error. You can utilize this saved model for deployment or additional inference.

Deployment/Prediction Phase:

- **Input Data (New MRI pictures):** During this stage, newly acquired MRI pictures of the brain are predicted using the learned model.
- **Data Standardization:** To guarantee that the input MRI data matches the format and preprocessing processes used during the training phase, the data is standardized before making predictions.
- **Prediction:** After analyzing the fresh MRI scans, the ViT model forecasts whether Alzheimer's disease will manifest or not.
- **Data Restoration:** A human-readable format is created by decoding the anticipated outcomes.

This could entail binaryly classifying the model's output, such as "Alzheimer's Detected" or "No Alzheimer's Detected."

- **Result:** The outcome is shown, giving the

medical practitioner or system access to the classification output. This might be examined in more detail or included in a tool that helps doctors diagnose the patients.

3. RESULTS AND DISCUSSIONS

Using MRI scans, the Vision Transformer (ViT) model was trained across 50 epochs using a batch size of 40, an Adamax optimizer, and categorical cross-entropy loss in this work to diagnose Alzheimer's disease. 10% was set aside for validation, 10% for testing, and 80% was used for training. Effective learning was proven by the loss and accuracy curves, where the validation loss stabilized towards later epochs, indicating less overfitting, while the training loss dropped from 1.2 to 0.5. Strong generalization was demonstrated by the training accuracy, which reached 82%, and the validation accuracy, which reached approximately 71%. Cropping, scaling photos to 224 by 224 pixels, transforming them into tensors, and normalizing pixel values were the preprocessing procedures where the image looked as shown in Figure 4. Data augmentation methods like rotation and flipping were applied to increase the variety of training images and avoid overfitting. With a patch size of 16x16, the ViT model was fine-tuned to capture both local and global information from the MRI scans. This showed the model's effectiveness in identifying the stages of Alzheimer's disease, while there is still potential for improvement in terms of early-stage differentiation.

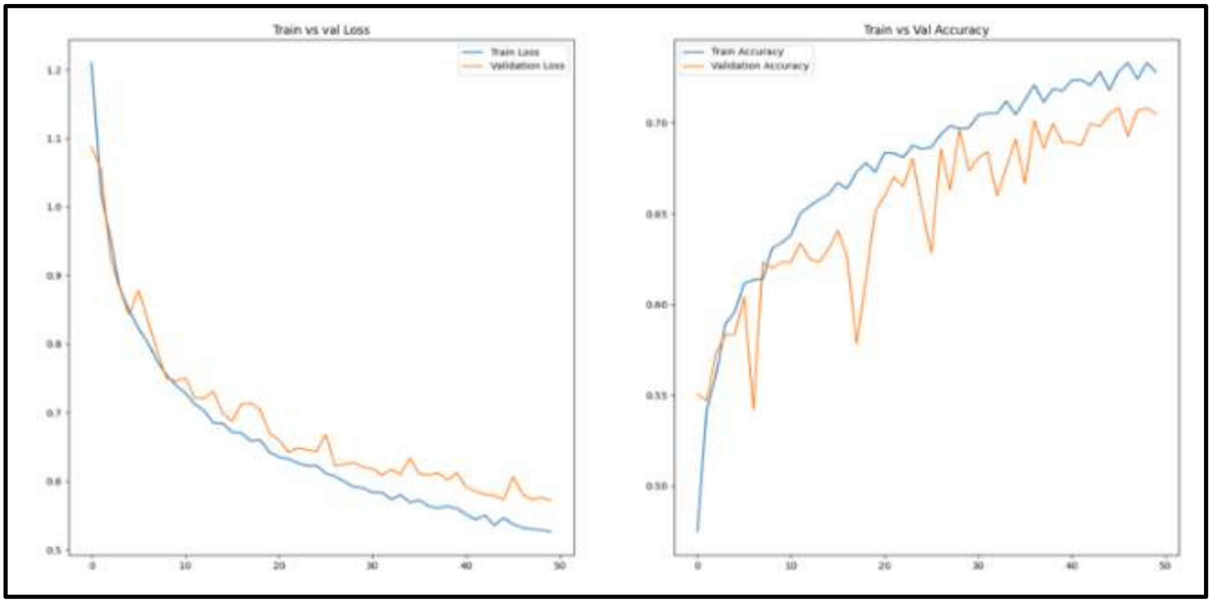


Figure 5. (a) Accuracy vs Epoch, **(b)** Loss vs Epoch of ViT

3.1. Visualization of Accuracy and Loss

For both training and validation datasets, the Vision Transformer (ViT) model's performance was assessed over a 50-epoch period. During training, two important metrics—accuracy and loss—are plotted to evaluate the model's performance and capacity for generalization as shown in Figure 5.

3.1.1. Loss vs Epoch (Figure 5(b)

Plotting the loss curves for the training and validation sets reveals a general downward trend, which is to be expected for successful training. One can make the following observations:

- **Training Loss:** By epoch 50, the training loss is estimated to be 0.5. It begins at 1.2 and steadily declines throughout training. This shows that the error on the training set is gradually being minimized by the model.
- **Validation Loss:** The validation loss initially has a similar pattern, but it varies more than the training loss, indicating the variability of the model's capacity to generalize to previously unobserved data. The validation loss stabilizes towards later

epochs, indicating less overfitting and improved generalization.

3.1.2. Accuracy vs Epoch (Figure 5(a))

The model's capacity to identify Alzheimer's disease in MRI images is shown by the accuracy curves for the training and validation sets:

- **Training Accuracy:** By epoch 50, the training accuracy had increased steadily over time to a maximum of almost 82%. This shows that the training data's patterns are being successfully learned by the model.

- **Validation Accuracy:** Achieving between 71 and 72% in the last epochs, the validation accuracy exhibits a similar trend to the training accuracy. The validation accuracy fluctuates a little bit, but overall, the trend is upward, indicating that the model is generalizing well to the validation data.

The learning curves indicate that the ViT model is doing a good job of classifying Alzheimer's disease. It is evident from the declining loss and rising accuracy that the model is successfully picking up the features required for prediction. The validation loss and accuracy stabilize in later epochs, therefore even with occasional volatility in the validation measures, the model does not exhibit significant overfitting symptoms. This indicates that learning and generalization have been well-balanced, which qualifies the model for additional refinement and testing on fresh MRI datasets.

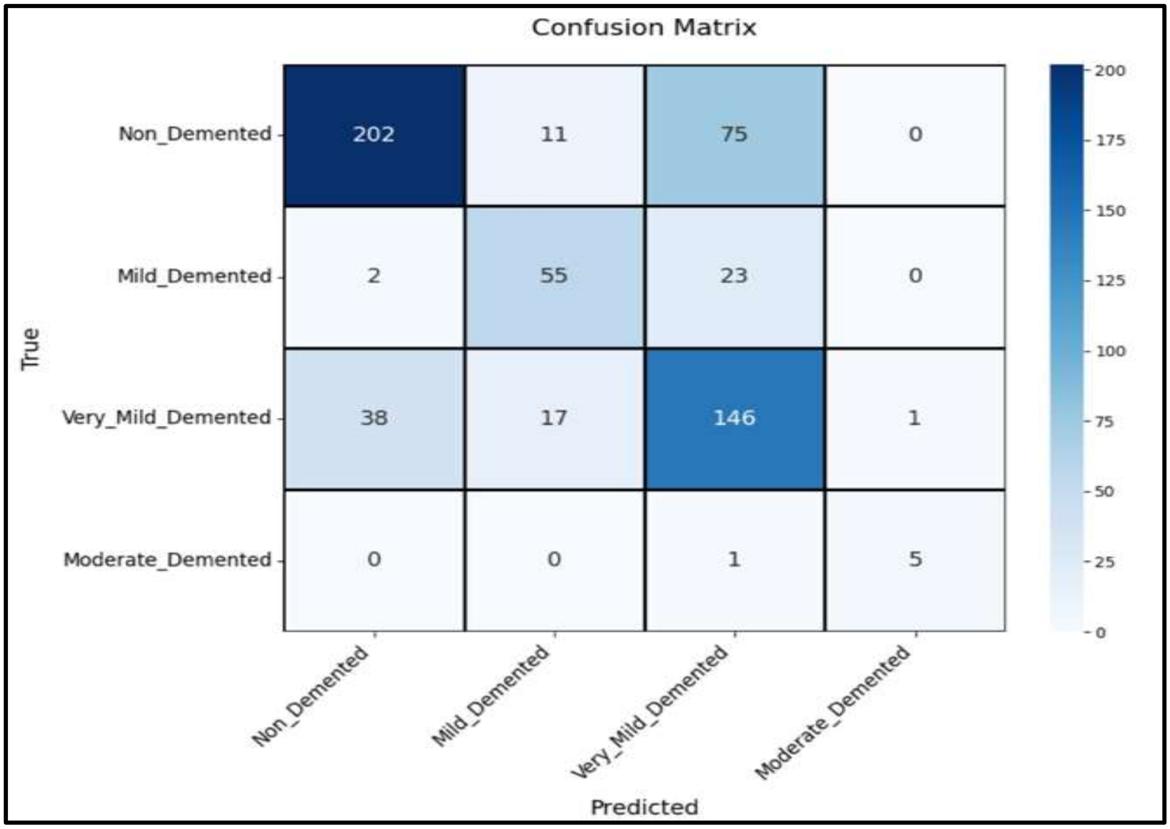


Figure 6. Confusion Matrix

3.2 Classification Report Analysis

A confusion matrix is used to evaluate the Vision Transformer (ViT) model's effectiveness in classifying Alzheimer's disease stages. It records the connection between true and predicted labels for four classes: Moderate Demented, Very Mild Demented, Non-Demented, and Mild Demented. In Table , the confusion matrix is displayed.

Non-Demented Patients: Of 288 patients without dementia, the model accurately identified 202, misclassifying the other 86 cases—75 were mistakenly projected to be Very Mild Demented, and 11 as Mild Demented. A well-known problem in neuroimaging-based diagnosis is the high rate of misclassifications into the Very Mild Demented class, indicating that there may be some overlap in the characteristics of the non-demented and early stages of dementia.

Table 3. Classification Report

	precision	**recall**	**f1-score**	**support**
0	0.83	0.70	0.76	288
1	0.66	0.69	0.67	80
2	0.60	0.72	0.65	202
3	0.83	0.83	0.83	6
Accuracy			0.71	576
Macro avg	0.73	0.74	0.73	576
Weighted avg	0.73	0.71	0.71	576

Mild Demented Patients: Of the 80 patients with mild dementia, the model accurately recognized 55 of them. Nevertheless, it misclassified two patients as non-demented and 23 patients as very mildly demented, suggesting some difficulties differentiating between similar degrees of cognitive impairment.

Very Mild Demented Patients: Out of 202 actual patients with very mild dementia, 146 were appropriately categorized. But as we can see, it can be difficult to distinguish between mild dementia and the early stages of dementia. Of these, 38 were wrongly classified as non-demented, and 17 as mildly demented.

Moderate Demented Patients: Of the six patients in the smallest class, the model correctly diagnosed five of them. Even with a small sample size, the model works well with more distinct stages of the disease, as seen by the fact that just one was mispredicted as Very Mild Demented. With the most data, the Vision Transformer (ViT) model performs well in differentiating between different stages of Alzheimer's disease, especially for the Non-Demented and Very Mild Demented classes. Moreover, the ViT model's comparatively high accuracy in identifying cases of moderate dementia shows that, even with fewer examples, it can distinguish between more severe stages of dementia. The model is not without flaws, though; the most frequent misclassifications occur between Very Mild Demented and Non-Demented, as well as between Mild Demented and Very Mild Demented. These inaccuracies suggest that the model has difficulty differentiating between minor cognitive and structural changes in early-stage dementia, a typical obstacle in neuroimaging tasks. Improving the model's capacity to distinguish between related phases of dementia could be accomplished via deeper feature extraction, multi-modal inputs, feature engineering, and additional model refinement.

3.3 Analysis using Confusion Matrix

The confusion matrix (Figure 6) demonstrates how well the ViT model can categorize Alzheimer's disease into its many stages. Nonetheless, there is a propensity to misidentify stages that are closely similar, especially between Very Mild Demented and Non-Demented and between Mild Demented and Very Mild Demented. These findings point to the necessity of additional model optimization, particularly for early-stage diagnosis, where the ability to differentiate between mild impairment and cognitive normality is essential for prompt intervention.

Where,

0 → Non_Demented
1 → Mild_Demented
2 → Very_Mild_Demented
3 → Moderate_Demented

To classify Alzheimer's disease stages, the Vision Transformer (ViT) model's classification performance was additionally assessed using metrics for accuracy, precision, recall, and F1-score. Table 3 displays the specific outcomes.

The following are the descriptions of the metrics:

• **Precision**: The precision numbers indicate how well the model can identify a certain class without accounting for false positives. Fewer false positives are indicative of higher precision.

Non-Demented: The model's low false positive rate for this class is indicated by its precision of 0.83.

Mild Demented: The model has a higher difficulty rate with false positives for this class, as indicated by its precision of 0.66.

Very Mild Demented: There are more false positives in this category, with a precision of 0.60.

Moderately Demented: Despite the small amount of data, the precision rate is 0.83, indicating a high precision rate for this class.

• **Recall**: Recall quantifies how well a model can recognize all positive examples of a particular class, hence reducing false negatives.

Non-Demented: A recall of 0.70 indicates that 70% of the Non-Demented cases are properly identified by the model.

Mild Demented: The model correctly detects 69% of the real instances of mild dementia, with a recall of 0.69.

Very Mild Demented: A recall of 0.72 indicates that real cases of Very Mild Demented patients were identified with a reasonably high degree of accuracy.

Moderately Demented: With an outstanding recall of 0.83, the model accurately diagnoses most individuals with moderate dementia.

• **F1-Score**: This score balances the trade-off between precision and recall by taking the harmonic mean of the two.

Non-Demented: For this class, an F1-score of 0.76 indicates a balance between precision and recall.

Mild Demented: An F1-score of 0.67 indicates that cases with mild dementia were classified with moderate accuracy.

Very Mild Demented: Despite having a higher recall, the model's difficulties with this class are indicated by its F1-score of 0.65.

Moderately Demented: Despite the limited sample size, an F1-score of 0.83 suggests high performance for this class.

- Overall Performance:

Accuracy: 71% of the samples in all classes are properly classified by the model, indicating that it has an overall accuracy of 71%.

- **Macro Average:** The precision, recall, and F1-score macro averages are roughly 0.73, indicating a uniform distribution of performance across all classes.
- **Weighted Average:** The precision, recall, and F1-scores are 0.73, 0.71, and 0.71 respectively, after adjusting for the various class sizes. This implies that the model works rather well, but it has problems with classes like Very Mild Demented and Mild Demented, which have smaller sample sizes or slight feature variations.

  According to the classification report, the model obtains good precision and recall when it comes to differentiating between the Non-Demented and Moderately Demented classes. However, it can be difficult to discern between the Mild Demented and Very Mild Demented classifications. This is probably because there are minor changes in brain structure and cognitive impairment that are hard to pick up from MRI pictures alone. These results imply the possibility of enhancing feature extraction or class balancing further to improve performance in the under-represented or more challenging-to-classify categories.

4. CONCLUSIONS

The primary motivation for this research stems from the growing need for early and accurate Alzheimer's disease (AD) diagnosis, which is crucial for effective intervention and improved patient care. As Alzheimer's disease progresses, timely detection becomes increasingly challenging, especially in its early stages, which is often characterized by subtle changes in brain structure. The most triggering factor for undertaking this work was the need to develop a more efficient and precise diagnostic method using medical imaging, particularly MRI, in conjunction with advanced machine learning techniques.

In this paper, we used MRI images to identify Alzheimer's disease using a pre-trained Vision Transformer (ViT) model, which has shown promising potential due to its ability to handle complex image patterns and efficiently process large-scale medical data. After 50 epochs of refinement, the model showed remarkable effectiveness during the training and validation stages. Training accuracy increased to almost 82%, and validation accuracy stabilized at 71–72%, indicating strong generalization to new data and efficient learning. The training loss steadily decreased, reaching approximately 0.5, and the validation loss stabilized, suggesting little over fitting, according to the loss measures. The impact of class imbalance on model performance was highlighted by the classification performance varying across different disease stages, with higher precision and recall found in recognizing early-stage Alzheimer's (Class 0) and some difficulties in identifying mid-stage Alzheimer's (Class 2). The Vision Transformer demonstrated an edge over conventional convolutional neural networks in its capacity to integrate both global and local data from MRI images. These results imply that Vision Transformers are a useful tool for medical image analysis, opening the door for additional improvements in hyperparameter optimization, hybrid model techniques, and class balancing to improve diagnostic accuracy and dependability in Alzheimer's disease identification.

The performance of the model can be further enhanced by integrating multi-modal data, such as combining MRI images with other biomarkers like genetic information or clinical data, providing a more comprehensive view of Alzheimer's disease and improving diagnostic accuracy. Additionally, refining the Vision Transformer (ViT) model using advanced techniques like fine-tuning, data augmentation, or larger, more diverse datasets could address class imbalance challenges and increase robustness. To ease medical analysis, this model can be implemented in clinical practice as a decision-support tool, assisting healthcare professionals in early-stage Alzheimer's detection by automating image interpretation, speeding up diagnostics, and minimizing human error. Expanding the model's scope could involve applying it to other neurodegenerative diseases, such as Parkinson's or Huntington's disease, by adapting it to detect specific structural brain changes. Furthermore, the development of real-time diagnostic systems and mobile platforms would enhance accessibility, especially in underserved areas with limited access to advanced medical imaging equipment.

REFERENCES

1. Vaupel, J. W. (2010). Biodemography of human ageing. *Nature*, *464*(7288), 536–542.
2. Samir K. C., &. W. (2017). The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100. *Global Environmental Change*, 42, 181–192.
3. Mohammed G. Alsubaie, S. L. (2024). Alzheimer’s disease detection using deep learning on neuroimaging: A systematic review. *Machine Learning and Knowledge Extraction*, *6*(1), 464–505.
4. Pradhan, A. J. (2021). Detection of Alzheimer’s disease (AD) in MRI images using deep learning. *International Journal of Engineering Research & Technology.*, 10(3), 580–585.
5. Odusami, M. M. (2023). Pixel-Level Fusion Approach with Vision Transformer for Early Detection of Alzheimer’s Disease. *Electronics*, *12*(5).
6. DeTure, M. A. (2019). he Neuropathological Diagnosis of Alzheimer’s Disease. *Molecular Neurodegeneration*, *14*(1).
7. Almufareh, M. F. (2023). Artificial Cognition for Detection of Mental Disability: A Vision Transformer Approach for Alzheimer’s Disease. *Healthcare*, *11*.
8. Aggarwal, R. W. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine*, *4*(1).
9. Varoquaux, G. C. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digital Medicine*, *5*(1).
10. Salahuddin, Z. K. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, *140*.
11. Ayush Pandey, A. P. (2023). Enhancing Waste Management: Automated Classification of Biodegradable and Non-biodegradable Waste using CNN. *ICT-CEEL*.

12. Acquarelli, J. S. (2022). Convolutional neural networks to predict brain tumor grades and Alzheimer’s disease with MR spectroscopic imaging data. *PLOS ONE, 17*(8).
13. Samhan, L. F.-N. (2022). Classification of Alzheimer's disease using convolutional neural networks.
14. Kang, W. L. (2021). Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer’s disease diagnosis. *Computers in Biology and Medicine, 136*.
15. Suk, H.-I. L.-J.-Y.-K. (2017). Deep ensemble learning of sparse regression models for brain disease diagnosis. *Medical Image Analysis, 37*, 101-113.
16. Feng, W. W. (2020). Automated MRI-based deep learning model for detection of Alzheimer’s disease process. *International Journal of Neural Systems, 30*(06).
17. Helaly, H. B. (2022). Deep Learning Approach for Early Detection of Alzheimer’s Disease. *Cognitive Computation, 14*, 1711–1727.
18. Sarraf, S. T. (2016). Classification of Alzheimer's disease structural MRI data by deep learning convolutional neural networks.
19. Wang, S.-H. Y.-J.-C. (2018). Classification of Alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *Journal of Medical Systems, 42*, 1-11.
20. Khvostikov, A. B. (2018). 3D inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer's disease diagnostics.
21. Kundaram, S. P. (2021). Deep Learning-Based Alzheimer Disease Detection. *Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems.* Springer, Singapore.
22. Suresha, H. S. (2020). lzheimer Disease Detection Based on Deep Neural Network with Rectified Adam Optimization Technique using MRI Analysis. *IEEEProceedings of the 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)* (pp. 1-6). IEEE.
23. Almadhoun, H. R.-N. (2021). Classification of Alzheimer’s Disease Using Traditional Classifiers with Pre-Trained CNN. *International Journal of Academic Health and Medical Research (IJAHMR), 5*(4), 17-21.