Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 2 (2025) 239 - 249 e-ISSN: 2580-0760

Detecting Alzheimer's Based on MRI Medical Images by Using External Attention Transformer

Farrel Ardannur Deswanto[1*], Isman Kurniawan[2]
[1,2]School of Computing, Telkom University, Bandung, Indonesia
[1]farrelardes@student.telkomuniversity.ac.id, [2]ismankrn@telkomuniversity.ac.id

Abstract

Alzheimer's disease is one of the major challenges in medical care this century, affecting millions of people worldwide. Alzheimer's damages neurons and connections in brain areas responsible for memory, language, reasoning, and social behavior. Early detection of this disease enables more effective treatment and proper care planning. Unfortunately, the traditional method of detecting Alzheimer's has several limitations, such as subjective analysis and delayed diagnosis. One commonly used method is visual inspection, which uses magnetic resonance imaging (MRI). The limitations of visual inspection include subjectivity and its time-consuming nature, especially with large or complex MRI datasets, making accurate interpretation a significant challenge. Therefore, an alternative for detecting Alzheimer's disease is to use deep learning-based MRI image analysis. One promising approach is to implement the External Attention Transformer (EAT) model. It enhances image classification by using two shared external memories and an attention mechanism that filters out redundant information for improved performance and efficiency. The aim of this research is to evaluate and compare the performance of the baseline Convolutional Neural Network (CNN) model, the Vision Transformer (ViT) model, and the EAT model in detecting Alzheimer's using a dataset of 6400 brain MRI images. The EAT model outperforms the baseline CNN model and ViT model in detecting Alzheimer's, achieving its best results with an accuracy of 0.965 and an F1-score of 0.747 for the test data. Our results could be integrated with clinical analysis to assist in the faster diagnosis of Alzheimer's.

Keywords: *alzheimer's disease; detection; MRI; CNN; external attention transformer*

How to Cite: F. Ardannur Deswanto and I. Kurniawan, "Detecting Alzheimer's Based on MRI Medical Images by Using External Attention Transformer", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 2, pp. 239 - 249, Mar. 2025.
DOI: https://doi.org/10.29207/resti.v9i2.6257

1. Introduction

Alzheimer's disease (AD) is one of the biggest challenges in medical care this century and is the leading cause of dementia, affecting around 40 million people worldwide [1]. Dementia is a medical condition in which a person experiences difficulties with various cognitive aspects, especially in terms of memory, but also in terms of language, attention, orientation, judgment, and planning [2]. Alzheimer's damages neurons and their connections in areas of the brain that are responsible for memory, language, reasoning, and social behavior [3]. Alzheimer's disease is a progressive disease, which means it will get worse over time, how fast it progresses, and what abilities are affected vary from person to person [4]. The most relevant risk factor for the development of this disease is age, with a prevalence of 10% in people over 65 years of age and 40% in people over 80 years of age [5].

Alzheimer's disease is the most common type of dementia, accounting for approximately 70% of all dementia cases [6]. Around 1 in 85 people in the world is expected to suffer from Alzheimer's disease by 2050 [7]. Alzheimer's detection can provide patients with the opportunity to collaborate on the development of an advanced care plan with family, caregivers, doctors, and other members of the support team. Alzheimer's detection also allows patients to begin seeking treatments that help manage symptoms, make lifestyle changes to maintain quality of life, and reduce the risk of cognitive, functional, and behavioral decline [8]. One of the most used detection methods is through medical images based on magnetic resonance imaging (MRI).

Received: 24-12-2025 | Accepted: 14-03-2025 | Published Online: 19-03-2025

MRI is an imaging technique that can be used to visualize the anatomy and physiology of the body in both disease and health conditions [9]. MRI allows users to view detailed images from inside the body with good contrast and high resolution. This technology uses the principles of physics to create images showing various physical and physiological aspects of the body, such as the structure of tissues and the changes that occur within them. The advantage of MRI lies in its non-invasive nature, enabling safe and repeatable scans [10].

Visual inspection is an important process for evaluating the quality of data generated by an MRI machine. It involves manual inspection by an expert to detect artifacts, distortions, or other anomalies that may appear in the images. One of the drawbacks of visual inspection is its risk of observer subjectivity, where interpretation of data quality may vary between different observers. In addition, visual inspection requires considerable time and effort, especially in the case of large or complex datasets [11]. Alternatively, deep learning-based inspection has become a major attraction in disease detection, especially Alzheimer's disease. Deep learning allows the system to automatically learn meaningful features from MRI images, improving the speed and accuracy of the diagnosis. Therefore, this method offers the potential to improve speed, accuracy, and consistency in MRI image-based disease detection [12].

There have been several studies related to the implementation of Alzheimer's disease detection on MRI based on deep learning. In 2019, Ji and his team studied early diagnosis of Alzheimer's disease using MRI-based ConvNets, achieving 97.65% accuracy for Alzheimer's Disease (AD)/Mild Cognitive Impairment (MCI) and 88.37% for Mild Cognitive Impairment/Cognitively Normal (CN) using ensemble learning after convolution operations [13]. In 2021, Ebrahimi and Luo compared several models, including 2D and 3D CNNs and RNNs, and found that ImageNet's transfer learning-based 3D voxel method achieved the highest accuracy of 96.88% [14]. In the same year, Helaly and colleagues designed an end-to-end framework using CNNs for early Alzheimer's detection, achieving 93.61% and 95.17% accuracy for 2D and 3D images, respectively [15]. In 2022, Houria and his team developed a multi-modality MRI fusion strategy, using 2D deep CNN as the feature extractor and SVM as the classifier, and achieved 99.79% accuracy for AD/CN classification, 99.6% for AD/MCI classification, and 97% for MCI/CN classification [16]. In 2023, Hoang and colleagues explored the MCI-to-AD prediction method using Vision Transformers for structural MRI, achieving 83.27% accuracy [17].

Although MRI has proven its usefulness in disease diagnosis with a prominent level of detail, accurate interpretation of MRI images remains a challenge. This interpretation challenge may hinder efforts for early detection of Alzheimer's and proper structural analysis of the brain. One promising solution is to utilize the External Attention Transformer (EAT) model. External Attention is a mechanism in data processing that allows consideration of correlations between all data samples in a dataset implicitly. Its advantages include strong regularization and generalization, linear computational complexity, and discriminative feature selection. This allows the model to capture the most informative part of the data and ignore distracting information from other samples [18]. This provides the superiority of obtaining understanding of the whole picture, which can be particularly useful in analyzing the brain structure for the detection of Alzheimer's.

This research aims to develop an innovative approach that integrates the EAT model to detect Alzheimer's disease through MRI medical image analysis. In addition, this research also aims to evaluate the effectiveness of the EAT model in improving the accuracy and consistency of Alzheimer's disease detection compared to the baseline Convolutional Neural Network (CNN) model and Vision Transformer (ViT) model. It is expected that this approach will not only make an important contribution to the detection of Alzheimer's disease but can also open opportunities for the development of more advanced and effective medical image analysis methods in the future.

2. Research Methods

2.1 Data Preparation

The dataset utilized in this research was sourced from Kaggle, consisting of a total of 6400 brain MRI images manually collected from various websites. Originally, the dataset contained four classes: Mild Demented, Moderate Demented, Non Demented, and Very Mild Demented. However, the sample number of the Moderate Dementia class (only 2 subjects) and Mild Dementia class (28 subjects), which is much lower than that of other classes, lead to class imbalance condition and risk of models that are biased towards the majority classes. Hence, we restructured the data set into two main classes: Demented and Non Demented. Also, in clinical practice, the primary concern is detecting the presence of dementia. This modified dataset will be utilized to evaluate the performance of the baseline CNN model, the ViT model, and the EAT model. The classes represent the cognitive state of Alzheimer 's-related patients. The dataset can be accessed from the Alzheimer MRI 4 classes dataset.

The first class, Demented, includes MRI images of patients who exhibit signs or symptoms of dementia. These individuals often experience a significant decline in cognitive function, which may manifest as impaired memory, difficulties in thinking and language, or challenges in performing daily activities. Example images can be seen in Figure 1.

The second class, Non Demented, includes MRI images of patients who do not exhibit any signs or symptoms of dementia. These individuals typically have normal cognitive functions, although they may display mild symptoms associated with normal ageing. Example images can be seen in Figure 2.

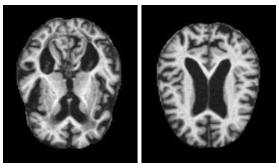

Figure 1. Example images for the Demented class

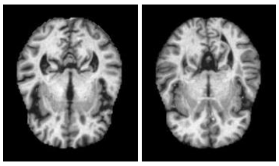

Figure 2. Example images for the Non Demented class

Data normalization is a preprocessing technique that involves adjusting the scale of the data, ensuring that each feature contributes equally to the model's performance [19]. The dataset is normalized by converting the image pixel values from the range [0, 255] to the range [0, 1] by dividing each pixel value by 255. The purpose of this normalization is to ensure that all data values are on a consistent scale, which makes the modeling process more efficient and stable.

The dataset is split into two main subsets: the training dataset and the test dataset. The training dataset consists of 5121 images used to train the model, while the test dataset consists of 1279 images used to test the performance of the model after the training process is complete. This splitting aims to balance learning and validation, as shown in Table 1.

Table 1. Data Splitting Table

Class	Train	Test
Demented	2561	639
Non Demented	2560	640
Total	5121	1279

2.2 Data Augmentation

Data augmentation's main goal is to create new datasets to increase the diversity and sufficiency of training data. More comprehensive features can then be represented by the augmented dataset [20]. The augmentation techniques used include zooming to the image in the range of 10% to enhance resilience to resolution variations, horizontal shifting up to 10% of the image width, and vertical shifting up to 10% of the image height to simulate slight variations in patient positioning during MRI scans. When conducting image shifting, the missing areas were filled in using reflect mode to prevent artificial artifacts that could mislead the model. In addition, an image rotation of up to 10 degrees was implemented to account for slight head orientation changes. Lastly, a brightness adjustment in the range of 80% to 120% of the original brightness was implemented to handle scanner and patient-related brightness variations. The selected augmentation techniques were performed to ensure that the model is robust to common variations without introducing unrealistic distortions.

2.3 Convolutional Neural Network (CNN)

Convolutional Neural Network is a type of deep learning model specifically designed to process data with grid patterns, such as images. CNN was developed to learn a spatial hierarchy of features automatically and adaptively, from low-level to high-level patterns [21].

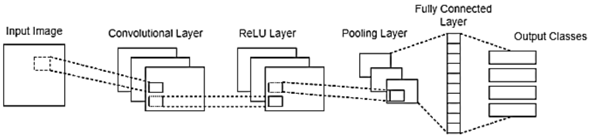

Figure 3. CNN architecture

The CNN architecture in Figure 3 is used for Alzheimer's detection through MRI image processing. The CNN architecture consists of several layers: convolutional, activation, pooling, and fully connected layers. The convolutional layer identifies the image features using a kernel (ReLU), followed by the activation layer, which applies ReLU to capture non-linear patterns and enhance training speed. The pooling layer reduces the dimensionality of the features, and finally, the fully connected layer flattens the output matrix into a vector, connecting the features to their corresponding categories, and improving overall classification performance [22].

Several CNN models with varying numbers of convolutional layers were used as baseline models in comparison with the research results, as presented in Table 2. These baseline models were evaluated to determine the impact of the depth of convolutional layers on model performance, which provides a reference for analyzing the effectiveness of the proposed approach.

Table 2. CNN Baseline Model Parameters

Scheme	Conv. Layer	Number of Kernels
Baseline 1	1	32
Baseline 2	2	32, 64
Baseline 3	3	32, 64, 128

Table 3. CNN Fixed Parameters

Parameters	
Kernel size	3 × 3
Pooling method	Max-Pooling
Pool Size	2 × 2
Fully Connected Layer	256

Table 3 presents the fixed parameters used across each scheme. In all schemes presented, the kernel size in each convolutional layer is 3 × 3, the pooling method used is max-pooling, and the pooling size is 2 × 2. In addition, the fully connected layer has 256 neurons for all schemes. The CNN model is trained with 100 epochs, a batch size of 32, and a learning rate of 0.001. The Adam optimizer is used, and the loss function is Categorical Crossentropy with accuracy as the evaluation metric. To improve training efficiency, the ReduceLROnPlateau technique is applied. This method reduces the learning rate by a factor of $\sqrt{5}$ if the validation loss does not show improvement over 5 consecutive epochs, with a minimum learning rate set to 1e-6.

2.4 Vision Transformer (ViT)

Vision Transformer (ViT) is an architecture derived from the vanilla Transformer, which was originally designed for natural language processing tasks. ViT utilizes the Transformer's encoder module and self-attention mechanism to process image data. This mechanism captures long-range dependencies by attending to different regions of the image and integrating information across the image. By dividing images into patches and mapping them to semantic labels, ViT generalizes the Transformer’s capabilities for image classification without relying on data-specific architectures, demonstrating its versatility across modalities beyond text [23].

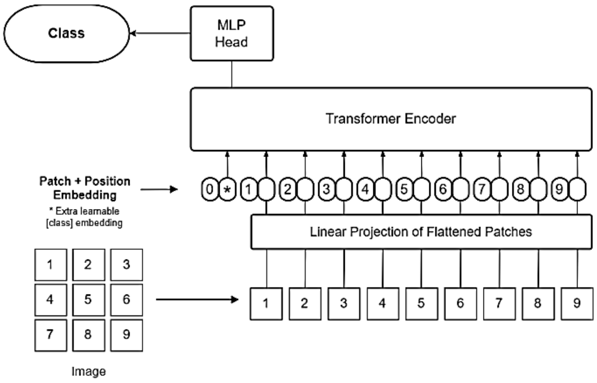
Class
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9
Image

Figure 4. Vision Transformer framework

The ViT framework, as shown in Figure 4, processes 2D images by dividing them into small fixed-size patches. These patches are flattened and linearly projected into a higher-dimensional space to create patch embeddings. A learnable embedding token is added to the sequence of patch embeddings, which represents the image. To retain positional information, 1D position embeddings are added to the patch embeddings. ViT utilizes only the encoder part of the transformer architecture, and the output of the encoder is passed through an MLP head for classification tasks [24].

Table 4. ViT Model Parameters

Parameters	
Project dimensions	64
Attention heads	4
Transformer layers	4

Table 4 presents the parameters used in the Vision Transformer (ViT) model. The model's architecture includes a project dimension of 64, which represents the size of the feature embeddings projected from the image patches. The ViT model utilizes 4 attention heads and consists of 8 transformer layers. The ViT model is trained with 100 epochs, a batch size of 32, a patch size of 6, a learning rate of 0.001, and a weight decay of 0.0001. The model training employs the Adam optimizer and uses Categorical Crossentropy as the loss function, with accuracy serving as the evaluation metric. To enhance training efficiency, the ReduceLROnPlateau technique is implemented, which reduces the learning rate by a factor of $\sqrt{5}$ if the validation loss does not show improvement over 5 consecutive epochs, with a minimum learning rate set to 1e-6.

2.5 External Attention Transformer (EAT)

External Attention Transformer is a model built on two small, teachable, shared external memories. It introduces an external attention mechanism that improves performance and computational efficiency in image classification tasks by eliminating patches with repetitive and unnecessary information [25]. In External Attention, all samples share two memory units M_k and M_v, which represents the most essential information of the entire dataset [26].

$$A = \mathrm{Norm}(FM_k^T\ ) \tag{1}$$

Equation 1 is the first step in calculating the attention value. It multiplies the input feature F with the transposed external memory M_k, resulting in the attention to weight A [18].

$$F_{out} = AM_v \tag{2}$$

Equation 2 describes how the attention weights computed in the first step are used to weight another external memory, M_v, to produce an updated feature output, F_{out} [18].

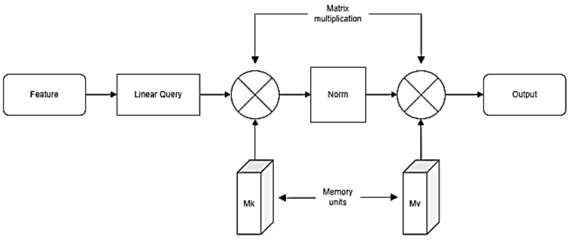

Figure 5. External attention for the EAT model

In Figure 5, external attention first computes the attention map by measuring the affinities between the self-query vector and the external learnable key memory. Once the attention map is obtained, a more refined feature map is generated by multiplying the attention map with the external learnable value memory, rather than weighing the self-value vector. This approach introduces an external learnable memory for more efficient and scalable attention computation [18].

Furthermore, the configuration of the EAT parameters is also detailed in Table 5. Sequential tuning was employed to optimize the parameters for the EAT model. Initially, 12 different combinations of hyperparameters were tested. However, after evaluating the performance and training time, 4 configurations were selected for further evaluation. The selected parameters for the final 4 schemes are listed in Table 6.

Table 5. EAT Model Parameters

Parameters	
Image size	150 × 150 × 3
Optimizer	Adam
Loss function	Categorical Crossentropy
Learning rate	1×10^{-3}
Weight decay	0.0001
Activation function	Softmax
Patch Size	2
Batch Size	32
Epochs	100

Table 6. EAT Model Scheme Experiments

Scheme	MLP Dimensions	Attention Heads	Transformer Blocks
1	64	4	4
2	64	8	4
3	128	4	4
4	128	8	4

2.6 Model Evaluation

Model validation is performed to collect, analyze, and evaluate the performance of the model. Performance is measured by accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), and Cohen's kappa values, which can be calculated using the confusion matrix in Table 7.

Table 7. Confusion Matrix Table

Class	Actual YES	Actual NO
Predict YES	True Positive (TP)	False Positive (FP)
Predict NO	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{7}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

Equation 3 calculates the overall accuracy by determining the proportion of correct predictions across the entire dataset. Equation 4 calculates the ratio of accurately predicted positive cases. Equation 5 measures the model's ability to find all positive cases in the dataset. Equation 6 is the harmonic mean of precision and recall, balancing their contributions equally [27]. Equation 7 measures the quality of a classifier by considering both positive and negative cases, accounting for class imbalance, and ensuring invariance to class swapping [28]. Equation 8 measures the level of agreement between two classifications by considering the possibility of agreement occurring by chance [29].

In this study, the best model is selected based on F1-score, which provides a balance between precision and recall. In Alzheimer's detection, it is important to minimize both types of errors, namely false positives and false negatives. The F1-score calculates the harmonic mean of precision and recall, thus providing a more comprehensive picture of the model's performance in a medical context. Therefore, F1-score was chosen to ensure that the selected model is not only accurate, but can also provide more consistent and reliable results in clinical applications.

3. Results and Discussions

The results of this research focus on evaluating the performance of the baseline CNN model, ViT model, and EAT model in classifying brain MRI images, with a particular emphasis on the Demented class. To tackle the class imbalance and guide the model's focus effectively on correctly classifying the Demented class, class weights were applied during training. Key metrics are analyzed to provide a comprehensive comparison between the two models.

3.1 Augmentation Process

The data augmentation process produced a diverse set of transformed images, demonstrating the effectiveness of the applied techniques. These transformations include zooming up to 10 percent, horizontal and vertical shifting by 10 percent of the image dimensions, rotation up to 10 degrees, and brightness adjustments ranging from 80 percent to 120 percent. The reflected mode was used during image shifting to fill the missing areas, ensuring a seamless and natural appearance in the augmented images.

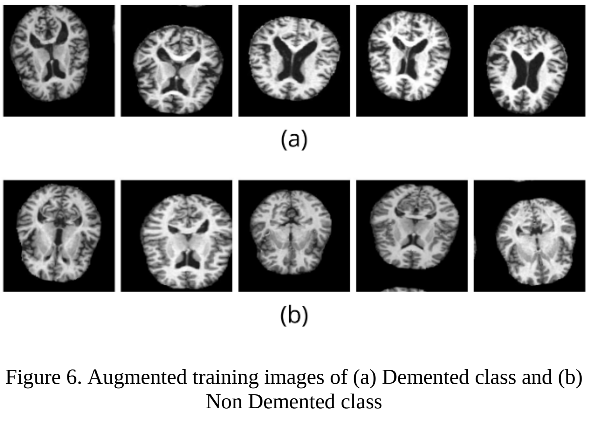

Figure 6. Augmented training images of (a) Demented class and (b) Non Demented class

The augmented images, which can be seen in Figure 6, display various changes in scale, position, orientation, and lighting conditions, demonstrating the effectiveness of the augmentation process. This diversity enhances the model's capacity to adapt to new, unseen data by presenting it with a wider variety of features and scenarios. By visualizing the results of augmentation, it becomes evident how these transformations improve the dataset's quality. The augmentation process ensures that the core characteristics of the original data are retained while introducing meaningful variations, making it a vital step toward building a robust and reliable model.

3.2 Baseline CNN Model Result

A set of experiments was carried out on the CNN baseline model to evaluate its performance. Three baseline schemes were developed, as outlined in Table 2. Each scheme is configured with a different number of layers, specifically using 1, 2, and 3 layers, respectively, to explore the impact of depth on performance.

Figure 7 illustrates the learning curves for the CNN baseline model. The graph presents the accuracy and loss metrics for the training and validation datasets. The x-axis represents the training iterations in terms of epochs, while the y-axis indicates the loss value. A lower loss value corresponds to a higher level of predictive accuracy achieved by the CNN model.

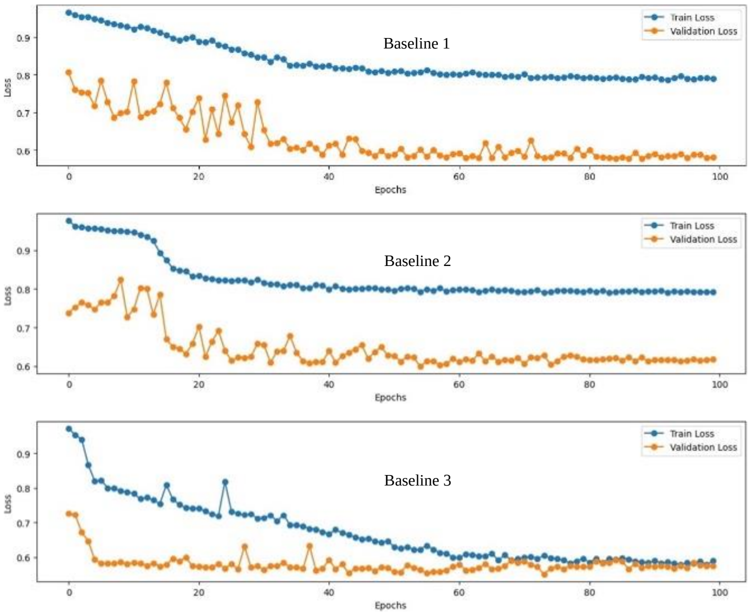

Figure 7. The plot of a learning curve for each CNN baseline scheme

Table 8 presents the evaluation results of the three CNN baseline schemes in classifying the Demented class. Each scheme is evaluated based on key metrics such as TP, FN, FP, TN, accuracy, precision, average precision, recall, F1-score, MCC, and Cohen’s kappa.

Table 8. CNN Result

Metrics	Baseline 1	Baseline 2	Baseline 3
TP	503	555	495
FN	136	84	144
FP	252	316	199
TN	388	324	441
Accuracy	0.787	0.868	0.774
Precision	0.666	0.637	0.713
Avg Precision	0.703	0.716	0.734
Recall	0.767	0.869	0.775
F1-Score	0.722	0.735	0.743
MCC	0.400	0.402	0.465
Cohen’s Kappa	0.393	0.375	0.463

In Baseline 1, the model correctly classified 503 samples as positive, while 136 positive samples were misclassified as negative. The scheme recorded 252 false positives and 388 true negatives, with an accuracy of 0.787. This scheme has a precision of 0.666, average precision of 0.703, recall of 0.787, and F1-score of 0.722. The MCC value of 0.400 indicates modest consistency between predicted results and real classifications, while Cohen's Kappa value of 0.393 confirms moderate agreement, suggesting the model could improve in handling class imbalance and distinguishing classes. This performance indicates that while the model is capable of capturing basic features, its shallow architecture limits its ability to represent complex patterns in the data. The relatively high false positive count and false negative count suggest that the single layer struggles to generalize well across all classes.

In Baseline 2, there was A notable rise in the count of true positives to 555, while false negatives were reduced to 84. However, this scheme recorded higher false positives of 316, with true negatives of 324. Achieved results included 0.868 accuracy, 0.637 precision, 0.716 average precision, 0.869 recall, and 0.735 F1-score. The MCC value of 0.402 and Cohen's Kappa of 0.375 suggest moderate agreement, implying that despite improvements in performance, the model still encounters issues with false positives. The increased depth enhances the model's capacity to recognize correlations within the data, resulting in more true positives and fewer false negatives. However, the added complexity increases false positives. The second layer's improved sensitivity to positive cases makes the model incorrectly label negative samples as positive.

The Baseline 3 scheme displayed a true positive count of 495 and a false negative count of 144. This scheme recorded a lower number of false positives than Baselines 1 and 2, which was 199, with a true negative of 441. The accuracy achieved was 0.774, precision 0.713, average precision 0.734, recall 0.775, and F1-score 0.743. The MCC value of 0.465 and Cohen's Kappa of 0.463 further indicate a better level of agreement compared to the previous baselines, reflecting improved model performance. The additional layer enhances the model's proficiency in differentiating between classes., leading to a higher precision and F1 score. While its accuracy and recall are slightly lower than Baseline 2, the overall balance between precision and recall makes Baseline 3 the most robust of the three.

Overall, the Baseline 3 scheme model is the best based on the highest F1-score of 0.743. Although the Baseline 2 scheme has higher accuracy, the Baseline 3 scheme Achieves a more balanced trade-off between precision and recall, thus providing a more reliable performance in overseeing the trade-off between the two. This is due to the increased depth of the model, with three convolution layers that allow it to capture more complex features. As a result, Baseline 3 is superior in Demented class classification compared to other schemes.

3.3 ViT Model Result

An experiment was carried out on the Vision Transformer (ViT) model to evaluate its performance. The ViT model, as outlined in Table 4, is configured with a specific number of project dimensions, attention heads, and transformer layers.

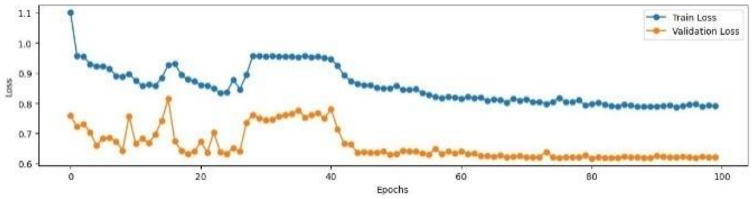

Figure 8. The plot of a learning curve for the ViT model

Figure 8 depicts the learning curves of the ViT model for its experiment, presenting accuracy and loss metrics for the training and validation datasets. The x-axis represents the epochs, while the y-axis shows the loss values. A decrease in loss values signifies improved accuracy achieved by the ViT model.

Table 9 presents the evaluation results of the ViT model in classifying the Demented class. The model is evaluated based on key metrics such as TP, FN, FP, TN, accuracy, precision, average precision, recall, F1-score, MCC, and Cohen’s kappa.

Table 9. ViT Result

Metrics	Value
TP	498
FN	141
FP	284
TN	356
Accuracy	0.779
Precision	0.637
Avg Precision	0.677
Recall	0.779
F1-Score	0.701
MCC	0.344
Cohen’s Kappa	0.335

The ViT model achieved an accuracy of 0.779, with 498 true positives and 356 true negatives correctly classified. However, it also generated 141 false negatives and 284 false positives. The model has a precision of 0.637, which reflects moderate effectiveness in avoiding false positives, as well as a recall of 0.779, showing strong sensitivity in detecting true positives. With an F1-score of 0.701, this model shows a balanced performance between precision and recall. In addition, the MCC value of 0.344 and Cohen's Kappa of 0.335 indicate a fair agreement between predictions and actual results, emphasizing the need for further refinement to improve overall classification accuracy and reliability.

3.4 EAT Model Result

A series of experiments on the EAT model were conducted to analyze its performance with different parameter settings. Four optimization schemes were developed for EAT, as outlined in Table 6. Each scheme has a unique configuration of EAT parameters, focusing on variations in MLP dimensions and the number of attention heads.

Figure 9 displays the learning curves for each EAT scheme. The graph shows the accuracy and loss figures for the training and validation data sets. The training iterations are represented by the epochs on the x-axis. Meanwhile, the y-axis displays the loss amount. A lower loss number indicated a higher degree of accuracy in the model's predictions. Table 10 presents the evaluation results of the four EAT model schemes in classifying the Demented class. Each scheme is evaluated based on key metrics such as TP, FN, FP, TN, accuracy, precision, average precision, recall, F1-score, MCC, and Cohen’s kappa.

Table 10. EAT Result

Metrics	1	2	3	4
TP	628	622	617	624
FN	11	17	22	15
FP	162	234	243	196
TN	478	406	397	444
Accuracy	0.982	0.973	0.965	0.976
Precision	0.568	0.605	0.608	0.584
Avg Precision	0.752	0.769	0.763	0.757
Recall	0.983	0.973	0.966	0.977
F1-Score	0.720	0.746	0.747	0.731
MCC	0.344	0.426	0.425	0.380
Cohen’s Kappa	0.235	0.338	0.345	0.282

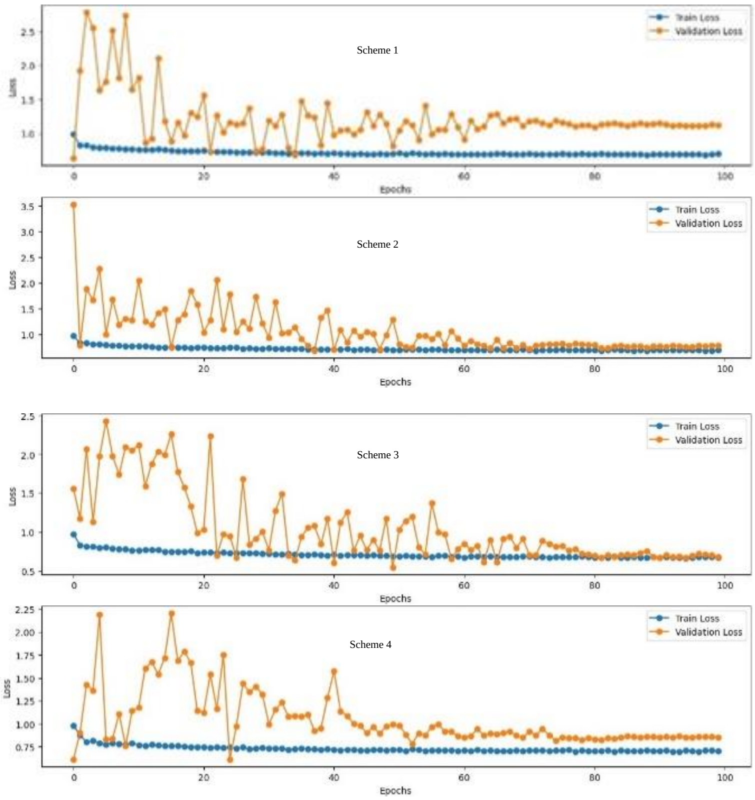

Figure 9. The plot of learning curve for each EAT scheme

In scheme 1, the model recorded 628 true positives and 11 false negatives, reflecting a high capacity to detect positive classes with a recall of 0.982. However, this scheme produces a high number of false positives, which is 478. Therefore, the precision only reached 0.568 with an average precision of 0.752 and an F1-score of 0.72. The MCC value of 0.344 and Cohen's Kappa of 0.235 suggest a modest agreement between the predicted and actual labels, with room for improvement. The model demonstrates strong positive class detection, as reflected by its high recall. However, the high count of false positives limits its precision and F1-score, indicating challenges in classification.

Scheme 2 shows an improvement in precision to 0.605 with a drop in false positives to 406. True positives decreased slightly to 622 with a false negative of 17. Model accuracy reached 0.973, while recall remained high at 0.973, resulting in an average precision of 0.769 and an F1-score of 0.746. The MCC value of 0.426 and Cohen's Kappa of 0.338 indicate better agreement between predicted and actual labels compared to Scheme 1, reflecting improved model performance. The model improves precision by lowering false positives while maintaining a high recall. A minor decrease in true positives reflects a trade-off for a better balance of detection accuracy and misclassification.

In scheme 3, the model successfully classified 617 data as true positives with only 22 false negatives. Although the number of false positives recorded was 397, the model achieved a precision of 0.608 and a recall of 0.966. The accuracy achieved was 0.965, with an average precision of 0.763 and an F1-score of 0.747, showing a good balance between positive detection ability and avoiding misclassification. The MCC value of 0.425 and Cohen's Kappa of 0.345 reflect a moderate level of agreement between predicted and actual labels. The model achieves an ideal balance, with slightly better precision and F1-score than previous schemes. Its ability to effectively manage false positives contributes to ensuring overall strong performance.

The model in scheme 4 performs exceptionally well in detecting positive data, with 624 true positives and only 15 false negatives. However, the high number of false positives, which is 444, causes the precision of this model to be 0.584. However, 0.976 accuracy and 0.977 recall were achieved. Average precision was recorded at 0.757, while the F1-score reached 0.731, which shows a solid performance despite the drop in precision. The MCC value of 0.380 and Cohen's Kappa of 0.282 are lower than those observed in Scheme 3, indicating a decline in agreement between predicted and actual labels. The model excels in recall and overall accuracy, indicating its strength in positive case detection. However, the increase in false positives impacts precision, showcasing a trade-off in balancing both sensitivity and specificity.

To determine the best model, we consider F1-score metrics as an overall measurement because the F1-score harmonizes the trade-off between precision and recall, offering a single, robust metric that reflects the model's effectiveness. Overall, the model in scheme 3 is the best based on the highest F1-score of 0.747. Although scheme 4 achieves the highest accuracy and recall, scheme 3 demonstrates a better balance between precision and recall, providing more reliable performance in managing the trade-off between the two. This is due to the optimal combination of 128 MLP dimensions and 4 attention heads in scheme 3. The larger MLP dimensions enhance learning capacity, while the 4 attention heads help capture dependencies effectively without overfitting. Despite a slight difference in the F1-score compared to scheme 2, scheme 3 still outperforms the other schemes in overall classification performance.

3.5 Overall Comparison

The performance of the Convolutional Neural Network (CNN), Vision Transformer (ViT), and External Attention Transformer (EAT) is compared using three key evaluation metrics including accuracy, average precision, and F1-score.

Table 11. CNN, ViT, and EAT Comparison

Metrics	CNN Baseline 3	ViT Model	EAT Scheme 3
Accuracy	0.774	0.779	0.965
Avg Precision	0.734	0.677	0.763
F1-Score	0.743	0.701	0.747

The results in Table 11 indicate a substantial performance improvement for the EAT scheme over both the CNN model and the ViT model. The EAT scheme achieved an accuracy of 0.965, significantly higher than the CNN accuracy of 0.774 and the ViT accuracy of 0.779. This suggests that EAT is better at capturing long-range dependencies and key features in the data, resulting in more accurate predictions compared to both CNN and ViT.

For average precision, the EAT scheme also outperformed both CNN and ViT, with a value of 0.763 compared to 0.734 for CNN and 0.677 for ViT. This highlights the superior ability of EAT to prioritize correct prediction results, a critical advantage for tasks involving unbalanced data. The ViT model’s relatively lower average precision indicates that it may struggle to maintain precision across all classes compared to the EAT and CNN models.

The F1-score metric, in which precision and recall are balanced, shows a slight advantage for the EAT scheme with a score of 0.747 compared to 0.743 for CNN and 0.701 for ViT. Although the difference is modest between EAT and CNN, the gap is more pronounced when compared to ViT, further emphasizing EAT's ability to provide a more balanced and reliable prediction performance.

The External Attention Transformer (EAT) consistently outperformed both the Convolutional Neural Network (CNN) and the Vision Transformer (ViT) across all three metrics evaluated. The superiority in accuracy, average precision, and F1 score underscores the strengths of the external attention mechanism in recognizing complex relationships and patterns in the data. These results affirm that EAT is a more robust and efficient model set for tasks that require both precision and accuracy, exceeding the abilities of CNN and ViT.

3.6 Computational Considerations

Overall, the training time for the CNN model is quite efficient, with a duration of about 55 to 65 minutes per training session. The training time tends to increase as the number of layers increases, although the difference is relatively small. Meanwhile, the training time for the ViT model is considerably longer, taking almost 2 hours. On the other hand, for the EAT model, the training time is strongly influenced by the number of significant parameters in each scheme. Schemes with more parameters require more time for computation. The EAT model took between 2 to 3 hours of training time, depending on the scheme used. After the model was developed, the prediction of Alzheimer's using a single image takes less than 2 minutes.

In terms of resource requirements, the CNN model demonstrates high efficiency with relatively modest demands on computational resources. It requires approximately 2GB of RAM and around 2GB of GPU memory, making it suitable for environments with limited hardware capabilities. In contrast, the Vision Transformer (ViT) model necessitates slightly higher computational resources, utilizing about 2GB of RAM but significantly more GPU memory, approximately 11GB, reflecting its reliance on the attention mechanism to process image patches. The EAT model, being more complex, demands the highest resources among the three. It requires around 3GB of RAM and approximately 13GB of GPU memory, indicating its intensive computational requirements due to the high number of parameters and the complexity of its architecture. All model training processes were conducted using dual NVIDIA T4 GPUs provided by Kaggle, ensuring consistent hardware resources across experiments.

3.7 False Prediction Analysis

Visual examples of the model's predicted results are presented in Figure 10, which categorizes them as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These examples offer useful insights into the classification process of the model. Through these visual representations, the process of correctly identifying relevant instances is shown alongside cases where the model has failed to accurately distinguish between different categories, thereby highlighting potential areas for improvement in its performance.

The true positive (TP) image shows a brain with typical features of dementia, such as shrinkage of certain areas or significant structural changes. The model successfully identifies dementia as the visible changes are highly consistent with the training data present in dementia. On the other hand, the true negative (TN) image shows a brain that has no signs of dementia, with a normal structure without shrinkage or other abnormalities. The model was able to recognize the normal pattern of a healthy brain. This correct classification highlights the importance of including representative samples in the training dataset for the model to properly generalize unseen data.

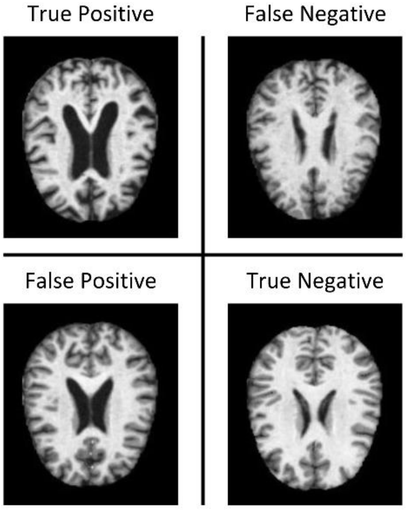

Figure 10. Sample images from each classification category

The false positive (FP) image may show normal variations of the brain or conditions that are similar to dementia but not directly related, such as slight shrinkage that can occur due to age. The model may be too sensitive to small changes similar to signs of dementia, thus misclassifying healthy brains as having dementia. This sensitivity can lead to a high rate of false positive results, with small variations in brain structure being considered as an indication of dementia. This can lead to unnecessary anxiety and potential misuse of medical interventions for individuals who do not actually have dementia. To mitigate this issue, integrating additional training data with examples of age-related variations and similar non-dementia conditions can help the model better distinguish between pathological and non-pathological changes. The consequence of FP is unnecessary medical treatment received by the patient due to an uncorrect positive demented condition.

The false negative (FN) image may show early or mild signs of dementia that are not obvious, such as slight shrinkage or small structural changes to the brain. Such changes are so subtle that they are difficult to detect. The model may not be sensitive enough to detect early or mild signs of dementia, thus failing to identify brains that actually have dementia. This could be caused by the small amount of variation in the training data that includes these small changes, or because the model is unable to identify subtle patterns that indicate structural changes in the brain associated with dementia. As a result, brains showing early signs of dementia may go undetected, which could hinder early diagnosis and effective treatment. To overcome this problem, it is possible to increase the diversity of training data by including a wider range of representative samples that can describe the characteristics of early-stage dementia. The consequence of FN leads to delayed treatment to misleading in the prediction of true demented condition.

These errors highlight significant ethical considerations, particularly in medical contexts. False positives could cause unnecessary stress and lead to patients undergoing unneeded diagnostic procedures, while false negatives might delay treatment, worsening the disease progression. In addition to the direct impact on patient well-being, such misclassifications can also strain healthcare resources and affect caregivers' mental and emotional health. Caregivers may experience heightened anxiety and uncertainty if a false positive result leads to unnecessary treatments or procedures, while a false negative could cause them to underestimate the severity of the disease, delaying proper care and interventions.

4. Conclusions

The External Attention Transformer (EAT) method is used to detect Alzheimer's from MRI medical images. After conducting several experiments with predefined schemes, the EAT model was trained with various configurations, testing different MLP dimensions and attention heads. All models used a learning rate of 0.001 and 100 epochs. While EAT achieved better accuracy and F1-scores, particularly in Scheme 3, which produced the best results with an accuracy of 0.965 and an F1-score of 0.747 on test data, the training process was more time-consuming compared to Convolutional Neural Networks (CNN). The Vision Transformer (ViT) model, on the other hand, achieved an accuracy of 0.779, with a lower F1-score of 0.701, but also exhibited long training times similar to the EAT model. CNN demonstrated faster training times and greater stability, making them a preferred choice in scenarios with limited computational resources or the need for rapid model development, despite slightly lower performance metrics. For future studies, it is recommended to explore some key areas to enhance the EAT model. Utilizing pre-trained transformer models can accelerate training and improve performance, especially with limited data. Optimizing the EAT architecture by refining the attention mechanism and testing various model depths would boost efficiency without sacrificing accuracy. In addition, addressing computational efficiency through techniques such as pruning, quantization, or mixed-precision training can reduce resource requirements, while techniques like Grad-CAM or SHAP can be applied to improve model interpretability. In addition, incorporating cross-validation could provide a more robust assessment of the model’s generalizability, particularly in medical applications where overfitting is a concern. Evaluating the model on multiple datasets or conducting external validation is also recommended to strengthen the findings and improve applicability in real-world scenarios. Lastly, investigating the impact of data augmentation strategies through ablation studies can improve model robustness and generalization. While basic augmentation is applied, exploring more advanced techniques, such as elastic deformations, which are particularly suitable for brain MRI data, could further enhance the model’s ability to capture anatomical variations. These approaches will make the EAT model more efficient and applicable to real-world scenarios, especially in medical image analysis. Furthermore, for real-world integration of the EAT model into clinical workflows, addressing challenges like computational demands and user interpretability is crucial. Optimizing the model's computational efficiency will be important for its application in resource-constrained settings. Additionally, improving the model's explainability will help doctors trust its predictions, allowing for easier integration into everyday medical practice. Overcoming these challenges will make the EAT model more feasible and impactful in actual clinical environments. Also, the consideration of 4 classes of Alzheimer might be possible by performing sampling approach, such overampling, and weighting approach, such as weighted loss

References

[1] K. G. Yiannopoulou and S. G. Papageorgiou, “Current and Future Treatments in Alzheimer Disease: An Update,” *Journal of Central Nervous System Disease*, vol. 12. pp. 1-12, 2020.

[2] Z. Arvanitakis and D. A. Bennett, “What Is Dementia?,” *JAMA*, vol. 322, no. 17, p. 1728, 2019, doi: 10.1001/jama.2019.11653.

[3] A. S. V. Prasad, “Physiological basis of memory dysfunction in Alzheimer’s disease–an overview,” *International Journal of Biochemistry Research & Review*, vol. 29, no. 2, pp. 9–24, 2020.

[4] Alzheimer’s Association, “2023 Alzheimer’s disease facts and figures,” *Alzheimer’s & Dementia*, vol. 19, no. 4, pp. 1598–1695, 2023, doi: https://doi.org/10.1002/alz.13016.

[5] A. P. R. Machado, I. O. Carvalho, and H. M. da Rocha Sobrinho, “Neuroinflamação na doença de Alzheimer,” *Revista brasileira militar de ciências*, vol. 6, no. 14, 2020.

[6] J. M. Tublin, J. M. Adelstein, F. Del Monte, C. K. Combs, and L. E. Wold, “Getting to the heart of Alzheimer disease,” *Circulation research*, vol. 124, no. 1, pp. 142–149, 2019.

[7] M. Calabrò, C. Rinaldi, G. Santoro, and C. Crisafulli, “The biological pathways of Alzheimer disease: A review,” *AIMS neuroscience*, vol. 8, no. 1, p. 86, 2021.

[8] S. Afzal *et al.*, "Alzheimer disease detection techniques and methods: A review," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 26–38, 2021, doi: https://doi.org/10.9781/ijimai.2021.04.005.

[9] S. Hussain *et al.*, "Modern diagnostic imaging technique applications and risk factors in the medical field: A review," *Biomed Research International*, vol. 2022, pp 1-12, 2022.

[10] D. B. Plewes and W. Kucharczyk, “Physics of MRI: a primer,” *Journal of magnetic resonance imaging*, vol. 35, no. 5, pp. 1038–1054, 2012.

[11] R. J. Lepping *et al.*, “Quality control in resting-state fMRI: the benefits of visual inspection,” *Frontiers in Neuroscience*, vol. 17, pp. 1-9, 2023.

[12] E.-G. Marwa, H. E.-D. Moustafa, F. Khalifa, H. Khater, and E. AbdElhalim, “An MRI-based deep learning approach for accurate detection of Alzheimer’s disease,” *Alexandria Engineering Journal*, vol. 63, pp. 211–221, 2023.

[13] H. Ji, Z. Liu, W. Q. Yan, and R. Klette, “Early diagnosis of Alzheimer’s disease using deep learning,” in *Proceedings of the 2nd international conference on control and computer vision*, 2019, pp. 87–91.

[14] A. Ebrahimi and S. Luo, “Convolutional neural networks for Alzheimer’s disease detection on MRI images,” *Journal of Medical Imaging*, vol. 8, no. 2, pp. 1-18, 2021.

[15] H. A. Helaly, M. Badawy, and A. Y. Haikal, “Deep learning approach for early detection of Alzheimer’s disease,” *Cognitive computation*, vol. 14, no. 5, pp. 1711–1727, 2022.

[16] L. Houria, N. Belkhamsa, A. Cherfa, and Y. Cherfa, “Multi-modality MRI for Alzheimer’s disease detection using deep learning,” *Physical and Engineering Sciences in Medicine*, vol. 45, no. 4, pp. 1043–1053, 2022.

[17] G. M. Hoang, U.-H. Kim, and J. G. Kim, “Vision transformers for the prediction of mild cognitive impairment to Alzheimer’s disease progression using mid-sagittal sMRI,” *Frontiers in Aging Neuroscience*, vol. 15, pp. 1-11, 2023.

[18] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, “Beyond self-attention: External attention using two linear layers for visual tasks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436–5447, 2022.

[19] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Applied Soft Computing*, vol. 97, p. 105524, 2020, doi: https://doi.org/10.1016/j.asoc.2019.105524.

[20] S. Yang *et al.*, “Image data augmentation for deep learning: A survey,” *arXiv preprint arXiv:2204.08610*, 2022.

[21] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[22] A. W. Salehi, P. Baglat, B. B. Sharma, G. Gupta, and A. Upadhya, “A CNN model: earlier diagnosis and classification of Alzheimer disease using MRI,” presented at the 2020 International Conference on Smart Electronics and Communication (ICOSEC), IEEE, 2020, pp. 156–161.

[23] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, “Vision transformers for remote sensing image classification,” *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[24] K. Han *et al.*, “A survey on visual transformer,” *arXiv preprint arXiv:2012.12556*, 2020.

[25] S. Hossain, M. Tanzim Reza, A. Chakrabarty, and Y. J. Jung, “Aggregating Different Scales of Attention on Feature Variants for Tomato Leaf Disease Diagnosis from Image Data: A Transformer Driven Study,” *Sensors*, vol. 23, no. 7, p. 3751, 2023.

[26] H. Wang *et al.*, “Mixed transformer u-net for medical image segmentation,” in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2022, pp. 2390–2394.

[27] M. Grandini, E. Bagli, and G. Visani, “Metrics for multi-class classification: an overview,” *arXiv preprint arXiv:2008.05756*, 2020.

[28] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC genomics*, vol. 21, pp. 1–13, 2020.

[29] D. Chicco, M. J. Warrens, and G. Jurman, “The Matthews correlation coefficient (MCC) is more informative than Cohen’s Kappa and Brier score in binary classification assessment,” *IEEE Access*, vol. 9, pp. 78368–78381, 2021.