

Efficient multimodel method based on transformers and CoAtNet for Alzheimer's diagnosis

Rahma Kadri ^{a,*}, Bassem Bouaziz ^b, Mohamed Tmar ^b, Faiez Gargouri ^b

^a Faculty of Economics and Management of Sfax, University of Sfax, Tunisia

^b Multimedia, Information systems and Advanced Computing Laboratory, University of Sfax, Tunisia

ARTICLE INFO

Keywords:

CNN
ViT
CoAtNet
Swin transformer
Alzheimer's diagnosis
Multimodal

ABSTRACT

Convolutional neural networks (CNNs) have been widely used in medical imaging applications, including brain diseases such as Alzheimer's disease (AD) classification based on neuroimaging data. Researchers extract the potential brain regions related to AD disease using CNN from various imaging modalities due to its architectural inductive bias. The major limitation of the current CNN-based model is that it doesn't capture long-range relationships and long-distance correlation within the image features. Vision transformers (ViT) have proven an astounding performance in encoding long-range relationships with strong modeling capacity and global feature extraction due to the self attention mechanism. However, ViT doesn't model the spatial information or the local features within the image and is hard to train. Researchers have demonstrated that combining CNN and a transformer yields outstanding results. In this study, two new methods are proposed for Alzheimer's disease diagnosis. The first method combines the Swin transformer with an enhanced EfficientNet with multi-head attention and a Depthwise Over-Parameterized Convolutional Layer (DO-Conv). The second method consists of modifying the CoAtNet network with ECA-Net and fused inverted residuals blocks. We evaluated the effectiveness of our proposed methods based on the Open Access Series of Imaging Studies (OASIS) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). Further, we evaluated the proposed methods using the Gradient-based Localization (Grad-CAM) method. The first method achieved 93.23% accuracy of classification on the OASIS dataset. The second method achieved 97.33% accuracy of classification on the OASIS dataset. We applied different multimodal image fusion methods (MRI and PET, MRI and CT) using our proposed method. The experimental results demonstrate that the fusion method based on PET and MRI outperforms the fusion method based on MRI and CT achieving 99.42% accuracy. Our methods outperform some traditional CNN models and the recent methods that are based on transformer for AD classification.

1. Introduction

Neurodegenerative diseases are a range of disorders that are characterized by a progressive structural and functional deterioration of brain neurons, which causes cognitive decline, cognitive impairment, mental functioning, the loss of balance and movement control. These diseases are becoming more prevalent across the world. Alzheimer's disease is one of the most challenging and complex neurodegenerative diseases. It is vital to building comprehensive information and an accurate disease characterization for an early diagnosis. This disease has different stages of cognitive decline and impairment, such as mild demented, moderate demented, non-demented, and very mild de-

mented. These stages represent the progression of the disease. Another well known stages are cognitively normal (CN), Mild cognitive impairment (MCI), and Alzheimer's disease (AD). Data collection about the patient is a critical step, this involves the patient's medical history, environment, genetic information, and various medical tests that extract the patient's functional and anatomical structures through diagnostic imaging methods. Neuroimaging is an attractive field that is devoted to studying brain structure and discriminative brain biomarkers using imaging tools. The imaging modalities present powerful tools that illustrate the potential biomarkers that enhance the understanding of the anatomical and functional neural changes related to the disease. Magnetic resonance images (MRI) and positron emission tomography (PET)

* Corresponding author.

E-mail address: elkadri rahma@gmail.com (R. Kadri).

<https://doi.org/10.1016/j.dsp.2023.104229>

the most powerful tools used for disease characterization and to gain insights for accurate diagnosis. The use of these modalities generates a huge amount of unstructured, heterogeneous, and complex healthcare data. There is an unmet need for extracting insights from these data and individualized representations of the disease. Deep learning techniques such as convolutional neural networks (CNN) achieved impressive results and have shown outstanding performance on brain diseases such as AD diagnosis [1], [2]. The CNN has been widely used to classify, detect, and predict AD disease based on brain modalities such as MRI and PET. The CNN has been applied to 2D and 3D images to extract useful biomarkers in brain modalities, such as amyloid-beta plaques and tau protein tangles. In addition, it extracts the relevant brain region associated with the AD disease. Several studies have reported the success of the application of CNN to Alzheimer's disease (AD), including image classification and disease progression prediction. However, CNN requires a huge amount of data, and the main challenge is the limited training AD imaging data. Furthermore, the brain modalities are complex because they have big variations in quality due to various factors such as image acquisition technique, scanner type, and patient movement during the scan. These variations can make it challenging for CNNs to extract meaningful features from the images. Today, vision transformers represent an emerging method and a promising new application for deep learning in computer vision. ViT has shown promising results in medical image analysis. The key technique behind the ViT is the attention mechanism. The main advantage of transformers over CNNs is that they capture complex relationships and long-range dependencies between the image features due to the self attention mechanism. In contrast, CNN cannot handle the long dependencies between the image features, but it has great potential to extract local and translation-invariant features from the input image, allows the model to focus on different parts of the input when making predictions. This attention mechanism can be visualized and interpreted, making it easier to understand how the model is making predictions. In contrast, CNNs typically have more opaque representations that are harder to interpret due to their ability to extract complex and non-linear relationships within brain modalities. However, CNN cannot encode long range dependencies within the image. Vision transformers (ViTs) [3], [4], [5] incorporate more global information than CNN, which ensures the extraction of more features. However, ViT is a data-hungry that relies on a huge amount of data and is hard to optimize. Furthermore, ViT is lacks inductive bias. In this paper, we overcome these issues by proposing two new methods based on combining the strengths of CNN and ViT.

- The first method is an hybrid method based on the Swin transformer and an enhanced EfficientNet. We add a multihead attention layer and a DO-Conv layer to the EfficientNetb0 to capture global dependencies within the MRI image.
- The second method consists of modifying the CoAtNet network, we replace the SE module with the ECA module within the MB-Conv, and we add an improved fused MBConv as an early layers to enhance the generalization capability and reduce the model complexity.
- We applied different multimodal image fusion methods (MRI and PET, MRI and CT) using our proposed method. The first method combines the PET and MRI modality. The second method combines the MRI and CT modality.

The first method achieved 93.23% accuracy of classification on the OASIS dataset. The second method achieved 97.33% accuracy of classification on the OASIS dataset for Alzheimer's disease classification into four classes (mild demented, moderate demented, non-demented, and very mild demented) and 98.87% on ADNI dataset for Alzheimer's disease classification into three classes (AD, CN, and MCI). The experimental results demonstrate that our methods outperform some traditional CNN models and the recent methods that are based on transformers for AD classification.

2. Related work

Brain disease diagnosis and classification, such as Alzheimer's disease, using deep learning has shown promising results. In this section, we explore and investigate state-of-the-art research insights on Alzheimer's disease (AD) using various deep-learning methods. We compare methods based on deep networks with and without attention mechanisms and with recent architectures such as vision transformers and hybrid methods that combine CNN and transformers. CNN is widely adopted for Alzheimer's disease diagnosis. [6] performed 3D CNN on FDG PET for AD detection. They achieved an accuracy of 75% using the ADNI dataset [7] also used 2D CNN using the same modality. This model achieved an accuracy of 59.73%. [8] improved the CNN architecture and adopted a multi-stream convolutional neural network to classify the progression of the MCI stage into stable MCI and progressive MCI for an early diagnosis. This model obtained 85.96% accuracy. MRI [9] have developed an AD classification model based on VGG16 as a feature extractor using the OASIS dataset. This model achieved 71% accuracy. Within the same task, [10] figures out that the current CNN methods cannot encode the global features from the input image. They improve the VGG16 by integrating a convolutional block attention module to the model to extract more complex features. They obtained an accuracy of 98%.

[11] asserted that the sMRI modality cannot efficiently capture the structural changes related to AD disease. They overcome these issues by proposing an attention multi-instance deep learning network (DAMIDL) to extract more local and global features of abnormally changed behavior and improve feature representation for the whole brain using the ADNI dataset.

[12] add a wise attention mechanism to the densely connected neural network to extract deep, high-level, and multi-scale features from MRI images from the ADNI dataset. They pointed out that the attention mechanism improves the denseNet. This model obtained 87.28% accuracy of classification.

[13] proposed a model that combines representation learning, feature distilling, and classifier modeling. They adopted a multi-head Pro-Sparse self-attention block to select the most discriminative features that represent the disease. To reduce the space complexity, they used a patch merging technique during the distillation operation. This model reached 92.8% as accuracy of classification. [14] adopted the Squeeze and Excitation (SE) mechanism with the Pyramid Squeeze Attention (PSA) mechanism combined with a Fully Convolutional Network (FCN) model to extract the feature information of the disease probability map for each MRI image. The authors demonstrated that image-filtering approaches and attention mechanisms improve Alzheimer's disease diagnosis. In [15] a new model based on 3D CNN that combines the kernel attention mechanism with a new global context was proposed. The attention mechanism enforces the CNN feature extraction compared to the previous works, reaching an accuracy of 97.28%. Recent studies addressed the issues of CNN and applied transformers to extract global features and encode long-range relationships among image features. [16] implemented ViT architecture on the PET modality for AD diagnosis reaching accuracy of 91%.

[17] add a multi-layer perception head to the ViT for AD detection using MRI modality. They obtained an accuracy of 89.58%. [18] noticed that CNN methods do not encode and extract the changes of these regions related to the disease. To address this issue, they incorporated 3D deformable self-attention module into the ViT architecture. The key feature of this network is that can encode the position of the selected patch based on the structural changes in sMRI related to the AD disease. They selected data from the ADNI and AIBL datasets. Transformer extracted global features from the image. However, it requires a huge amount of data and these architectures are hard to train. Recent studies combined CNN and transformers to address these issues. For example [19] combined VGG16 and Swin transformer for AD diagnosis. [20]

combine ViT, 2D CNN and 3D CNN to capture global dependencies and ensure good generalization. They achieved an accuracy of 93%.

[21] outlined that the main limitation of the recent studies based on the vision transformer on AD detection is that they don't extract the local features and don't encode the low-level feature interactions between the brain modalities within the feature extraction. They address this issue with a new model that combines dual-branch vision transformer with cross-attention and graph pooling. They obtained an accuracy of 98% using an MRI modality from the ADNI dataset. [22] proposed a multiple instance learning and self-supervised Data-Efficient Image Transformer (DeiT) for AD diagnosis. They obtained higher accuracy (93%) than the vision transformer. [23] combined ConvNeXt with an ensemble of machine learning classifiers. They reached an accuracy of 92%.

[24] proposed a multimodal fusion method that combines the MRI and PET modality using the harnessing demon algorithm and discrete wavelet transform. They perform ResNet-50 to extract features from the fused image, and then these features are classified using SVM. This method has an accuracy of 73%. [25] have implemented fusion models that combine the PET and MRI modality and extract deep multi-features using 3D convolutional neural networks. [26] applied a fusion modality that combined PET and MRI modality using 3D CNN. They achieved an accuracy of 71%. [27] developed a new fusion method between the MEG and MRI modalities. Here, the authors fused the features from the MEG and MRI modalities, and for classification, they used support vector machines (SVM). This model obtained an accuracy of 77%. [28] applied a contrastive self-supervised fusion of fMRI and MRI for AD classification.

[29] compare between CNN and the vision transformer and investigate the application of these models in Alzheimer's diagnosis. They applied the vanilla vision transformer, deep vision transformer, and class attention image transformer (CaiT). They pointed out that the vision transformers outperform CNN with a big dataset and lack inductive biases with a small dataset. Most of the multimodal methods that combine different brain modalities are based on CNN models. Thus, these methods didn't capture the long-term dependencies within each brain modality. Furthermore, these methods didn't extract the global features within the brain modalities and didn't capture the main interactions between the features extracted from the brain modalities. There are few studies [31] that used attention mechanism to address these issues. As depicted in the Table 1 the methods that are based on attention mechanism outperformed the methods that are based on CNN. The relevant choice of the modality and the architecture is crucial to obtain accurate results. For example, the studies that are based on the MRI modality outperformed the studies based on the PET modality. The studies that are based on vision transformer have an accurate accuracy compared to some CNN models. Furthermore, the studies that are based on the combination of CNN and transformers obtained more accurate results than ViT.

3. Methods

In this section, we describe the overall architecture of the proposed methods for AD diagnosis.

3.1. Vision transformer

The vision transformer is a subtype of transformer devoted to computer vision. The core blocks of the ViT are the patch embedding, transformer encoder, and multilayer perceptron (MLP) classifier. The self attention layer is the key compound of the vision transformer that overcomes the main limitation of CNN through dynamic feature reweighting. This mechanism enhances the network by learning the main connections and relationships between the patches.

The ViT split the input image into a grid of sub-image patches as depicted in Fig. 1. Given a 3D Image ($X \in resolution R^{H \times W \times C}$ and patch

size p . The image is divided into multiple patches of the same height and width. $x \in \mathbb{R}^{H \times W \times C} \Rightarrow x \in \mathbb{R}^{N \times P^2 \times C}$

$$N = \frac{HW}{p^2} \text{ is the number of patches}$$

P is the height and width of the patch with a patch dimension (P, P, C)

C is the number of channels

Then the patches are flattened and fed into a feed forward layer that takes each patch as input for patch projection and maps these features into a feature vector of large size, which refers to the patch embedding. The Fig. 2 illustrates the application of ViT for AD diagnosis.

Position embeddings are then linearly added to the sequence of image patches to retain positional information by adding information about the relative or absolute position of the image patches in the sequence. The output embedded patches are then passed through a vision encoder that involves a Multi-Head Self Attention (MSA) layer and a Multi-Layer Perceptron (MLP) layer. The Multi-Head Self Attention layer divides the input into various heads to extract the deep features from the image and obtain a global representation. This layer applies several times the self attention to each head. The outputs of all the heads are then concatenated and fed into the Multi-layer perceptron for classification. The layer norm (LN) is applied before every block. The Fig. 3 visualize the attention map within the application of ViT.

ViT has such limitations specifically with the high-resolution images. The computational complexity is quadratic to the input image size. This makes it hard to optimize. Another limitation is that ViT processes the input images as patches or tokens all with fixed sizes which are unsuitable. However, the visual elements of the input image are not unsuitable. The ViT is based on a global self-attention that measures the relationships between a token and all other tokens which leads to quadratic computational complexity.

3.1.1. Swin transformer with an improved EfficientNet B0 for AD diagnosis

The first method combines the Swin transformer with an improved EfficientNet B0 as depicted in 4.

Convolutional neural network models are commonly designed with manual tuning that consists of fixing the network parameters (Depth, Width, and Resolution) and then scaling these parameters to enhance accuracy. EfficientNet is a CNN model that is based on the compound scaling method that scales the dimensions of the network effectively. This method applies a grid search strategy to extract the main correlations between the different network parameters based on fixed constraint. The compound scaling method balances the scale in all three dimensions (width, depth, and image resolution) to improve the network as it flows:

$$\text{depth: } d = \alpha^\phi$$

$$\text{width: } w = \beta^\phi$$

$$\text{resolution: } r = \gamma^\phi$$

The EfficientNet consists of stacking multiple inverted bottleneck convolution (MBConv). The self-attention module is adopted to boost CNN and enable it to focus more on the relevant information within the input image rather than learning non-useful information such as background information that increases the model's complexity. Scaled dot-product attention is the core mechanism behind self-attention. It is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The scaled dot product attention is applied over a single sequence, which cannot emphasize all the important features within the input image. The solution is to apply the attention mechanism several times. Multihead self attention is an attention module that applies an attention mechanism several times in a parallel way. This module transforms the queries, keys, and values into linearly projected queries, keys, and values h times independently through multiplication with weight matrices

Table 1
Recent works in Alzheimer’s disease diagnosis using different deep learning methods.

Model	Architecture	Modality	Classes	Dataset	Accuracy
[6]	3D CNN	FDG PET	AD, CN, MCI	ADNI	75%
[9]	VGG16	MRI	nondemented, very mild to mild Alzheimer’s, nondemented, and nondemented	OASIS	71%
[10]	VGG16+attention mechanism	MRI	nondemented, very mild to mild Alzheimer’s, nondemented, and nondemented	OASIS	98%
[13]	multilayer perceptron+multihead ProbSparse self-attention +structural distilling+	MRI	AD, CN, MCI	ADNI	92.8%
[15]	3DCNN+ kernel attention mechanism +global attention	MRI	AD, CN, MCI	ADNI	97.28%
[12]	3D DenseNet+ wise attention mechanism	MRI	AD, CN, MCI	ADNI	87.28%
[8]	3Multi-Stream CNN	MRI	AD, CN, MCI	ADNI	85.96%
[7]	2 CNN	PET	AD, CN, MCI	ADNI	59.73%
[16]	Vision Transformer (ViT)	PET	AD, CN, MCI	ADNI	91%
[3]	Cascaded Modality Transformers architecture with cross-attention	MRI	AD, CN, MCI	ADNI+AIBL	94%
[24]	ResNet-50+SVM	PET+MRI	AD, CN, MCI	ADNIL	94%
[23]	ConvNeXt and ensemble of machine Learning classifiers	MRI	AD, CN, MCI	ADNIL	92%
[17]	Multi-layer perceptron+ ViT	MRI	AD, MCI, CN	ADNI	89.58%
[27]	multi-kernel learning of support vector machine	MEG+MRI	EMCI, LMCI, MCI	ADNI	77%
[20]	2D,3DCNN+ViT	MRI	AD, MCI, CN	ADNI	93.21
[21]	dual-branch vision transformer with cross-attention and graph pooling	MRI	AD, MCI, CN	ADNI	98%
[22]	DeiT+Instance learning	MRI	AD, MCI, CN	ADNI	93%
[30]	Resnet18	PET+MRI	EMCI, LMCI, MCI	ADNI	73.90%
[26]	3DCNN	PET+MRI	AD, CN, MCI	ADNI	71%

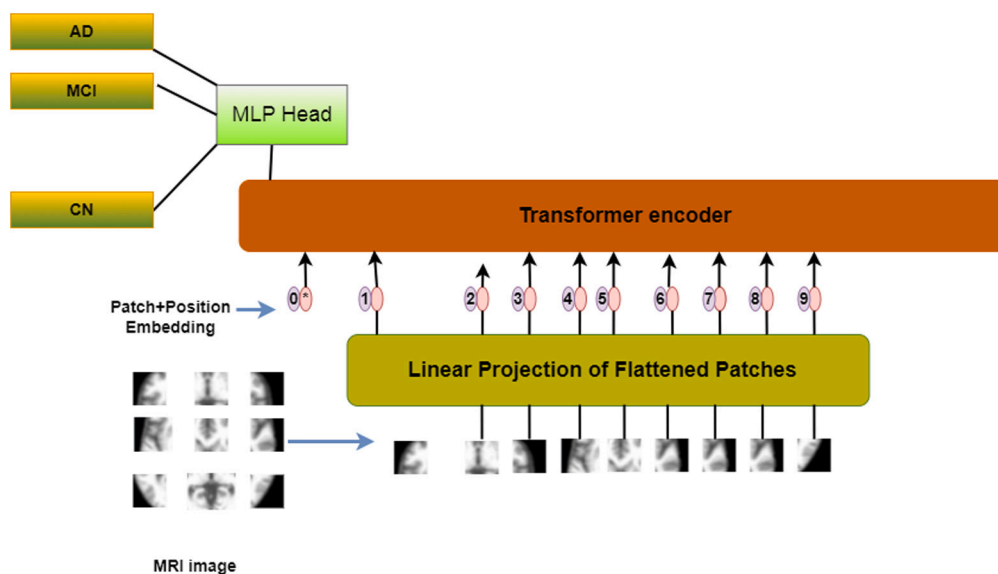


Fig. 1. ViT architecture.

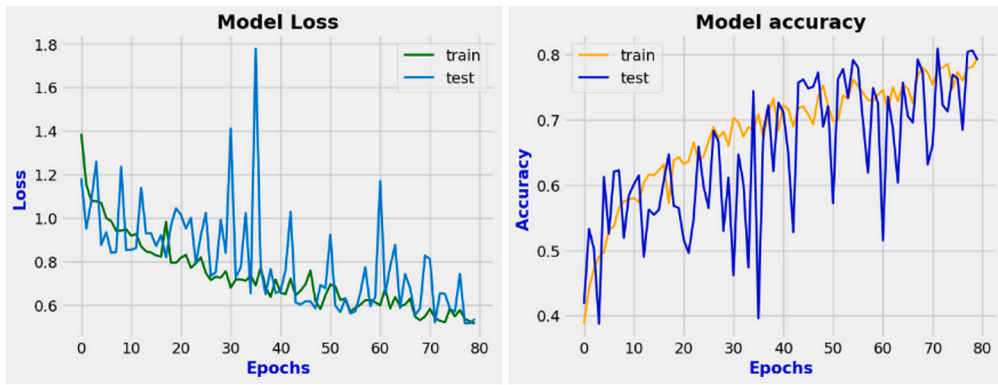


Fig. 2. The application of ViT for AD diagnosis.

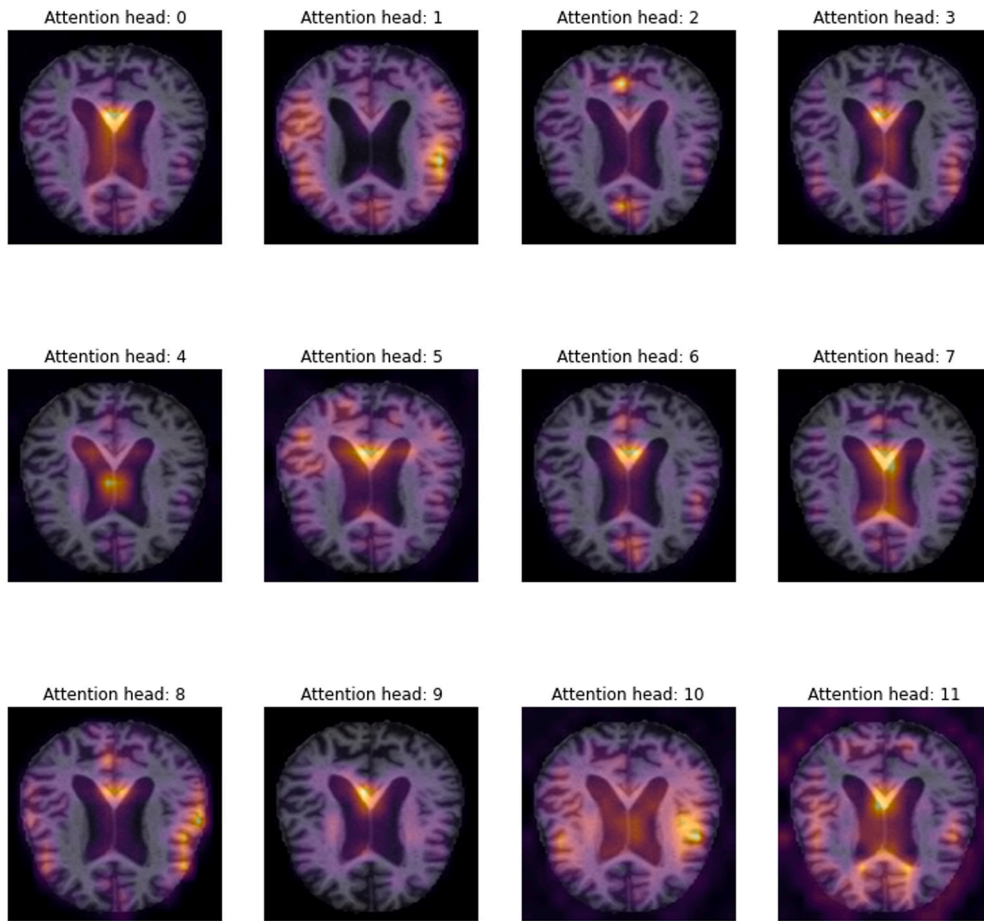


Fig. 3. Attention map visualization.

W_i^Q , W_i^K and W_i^V . Then a single attention mechanism is applied for each *head* projection in a parallel manner. The multi-head attention function can be defined as

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \dots, \text{head}_h] \mathbf{W}_0$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$. The final step is to concatenate the outputs of all the output of all the heads head_i , and apply the linear projection to the output with the weight matrix \mathbf{W}^O to obtain the final output.

The multihead attention enables the proposed model to learn and extract richer data representations. We also introduce the Depthwise Over-Parameterized (DO) module to the Efficient net B0 to improve the training speed of the network without adding parameters and to de-

crease computational complexity. An over-parameterized convolutional layer consists of a depthwise convolutional layer with a trainable kernel and a traditional convolutional with a trainable kernel.

3.1.2. Depthwise Over-Parameterized Convolution

Depthwise Over-Parameterized Convolution is introduced by [32]. It based on the combination of a depthwise convolution $D \in R^{(M*N)*D_{mul}*C_{in}}$ and conventional convolution kernels $W \in R^{C_{out}*D_{mul}*C_{in}}$ to speed up the model training and to boost its convergence. The over-parameterization refers to the combination of the two convolutions and adding learnable parameters to the model. The figure illustrates the flowchart of the DO-Conv. P represent the size of the input feature channel with spatial dimension $M * N$. C_{in} is the number of the input channels within the input feature maps. Whereas the C_{out} is the number

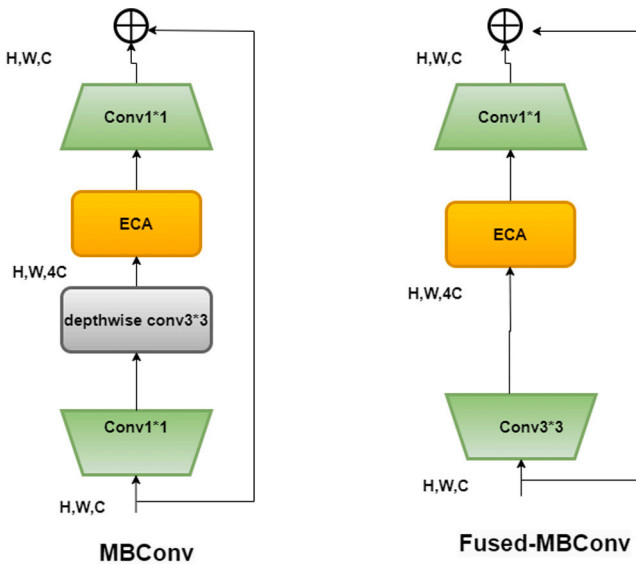


Fig. 5. MBConv and Fused MBConv blocks.

Given a feature map F , C is the number of input channels. H represents the height of the input feature map and W is the width of this input feature map. The depthwise convolution applied the convolution for each channel and a separate kernel to each channel. Here, the size of the kernel is $D_k \times D_k * 1$ with M filters the output has the size of $D_k \times D_k * M$.

The convolution operation requires a number of parameters proportional to the product of input channels, output channels, and kernel size. The key feature of the depthwise convolution is the reduction of the parameters since the number of these parameters is proportional to the number of input channels multiplied by the filter size which reduces the computational cost of the proposed model.

In addition, the application of the convolution to each channel separately allows the application of specific kernels to each input channel, which enable the network to capture and extract multi-scale features. The use of depthwise convolution enhances the network's ability to extract the local patterns in the input image.

Another key block of CoAtNet is the relative attention mechanism, which is a specific type of attention that takes into account the position of the image patches or tokens. The traditional attention mechanism is based on the dot product between the query and key vectors without computing the distance between the image patches and without considering its position in the calculation of attention. Whereas computing the distance between the tokens enables the model to better capture the long dependencies within the input image. The relative attention incorporates relative positional encodings to capture the relative relationship between image patches and the contextual relationships within the input image. This relative attention is coupled with the MBConv block to extract deep features and make the network more faster and generalize better. The MBConv block is a type of residual block that incorporates an inverted residual as depicted in Fig. 5. It consists of a depth-wise separable convolution with an inverse bottleneck and Squeeze-and-Excitation module. The module uses a 1×1 convolution flowed by an SE block to enhance the feature representation and uses a dynamic channel-wise feature recalibration. Then the module adopts a 3×3 depthwise convolution, which decreases the number of parameters. The output is passed again through a 1×1 which reduces the number of channels. The input is fed into the depth-wise convolution to reduce the number of the parameter. Then the output is fed into the convolution 1×1 layer to increase the number of channels and improve the feature representation to capture high level features.

The combination of depthwise convolution and relative self-attention within CoAtNet promotes it to encode and extract multi-scale features, including local and global features within the input brain modality. Fur-

thermore, this new technique decreases the computational cost and the complexity of the model. This enables it to handle the complex brain modalities of images and process various sizes and modalities.

The input image is passed through a stem stage (S0) consisting of convolution layers as depicted in 6. Then, the image is fed into two MBConv blocks that adopt depthwise convolution stages (S1 and S2) to reduce the dimensionality of the input.

The first MBConv blocks encode the spatial interaction between the image features. The second block compresses it before adding a residual.

The output of MBConv block stages is passed through transformer blocks with relative attention (S3 and S4). The network adopts pooling between stages. The final stage is to obtain the output class.

The output of MBConv block stages is passed through a transformer block with relative attention (S3 and S4). The network uses pooling between stages. The output class is obtained in the final stage.

Inspired by [34] we modify the structure of the MBConv by replacing the SE block with an ECA module within the mbconv as depicted in 7. Efficient Channel Attention (ECA) is a lightweight, efficient channel attention module that encodes and enhances the cross channel interaction within the input image by considering every channel and its neighbors. It is based on the squeeze excitation module. The Fig. 8 illustrates the Modified CoAtNet structure.

The key feature of the ECA module is the reduction of model complexity without dimensionality reduction. We also include fused inverted residuals as early blocks in the model. Fused inverted residuals modify the structure of the MBConv by fusing the depthwise conv 3×3 and the conv 1×1 into a single regular conv 3×3 to make the model faster as depicted in 7. Hence, the input MRI passes through Fused EMBCConv and is then fed into EMBCConv blocks. The output from the first stage is then passed through the transformer blocks. Then the output from the stage 2 is fed into global average pooling and soft max layers for classification.

3.3. Data collection

In this study, we selected different AD stages based on the MRI and PET images from the OASIS-1 and OASIS-3 datasets. The Open Access Series of Imaging Studies dataset (OASIS) is a public multi-modal dataset consisting of various longitudinal multimodal neuroimaging data, clinical, cognitive data, and brain modalities. The main objective of the dataset is to make the neuroimaging data freely available to enhance the Alzheimer's diagnosis. It involves demographic information from structural MRI, functional MRI, and PET scans.

The OASIS dataset is publicly available for research purposes, and can be accessed through the OASIS website or various data repositories, such as the OpenNeuro platform. Access to the dataset requires a registration process and acceptance of a data usage agreement.

The OASIS-1 dataset consists of cross-sectional data and MRI data about 416 subjects with 434 sessions aged 18 to 96 of middle-aged, nondemented and demented. The OASIS-2 is a longitudinal collection of 150 subjects about nondemented and Demented Older Adults. OASIS-3 is multimodal neuroimaging, clinical, and cognitive dataset that contains different brain modalities such as MRI, PET and CT. This dataset contains 2842 MR sessions with different types: T1w, T2w, FLAIR, ASL. It contains also 2157 PET imaging includes various tracers, PIB, AV45, and FDG. OASIS-3 includes 1472 CT sessions.

We also evaluated our proposed models based on the Alzheimer's Disease Neuroimaging Initiative (ADNI). We selected different stages of AD disease (AD, CN, MCI) and different brain modalities (MRI, PET and CT). This dataset is a well-known multimodal dataset created in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, and private pharmaceutical companies to promote the early diagnosis of Alzheimer's disease.

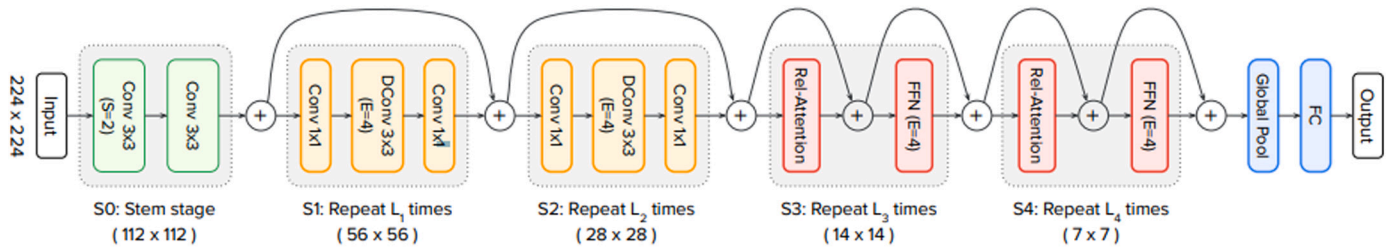


Fig. 6. CoAtNet architecture adapted from [33].

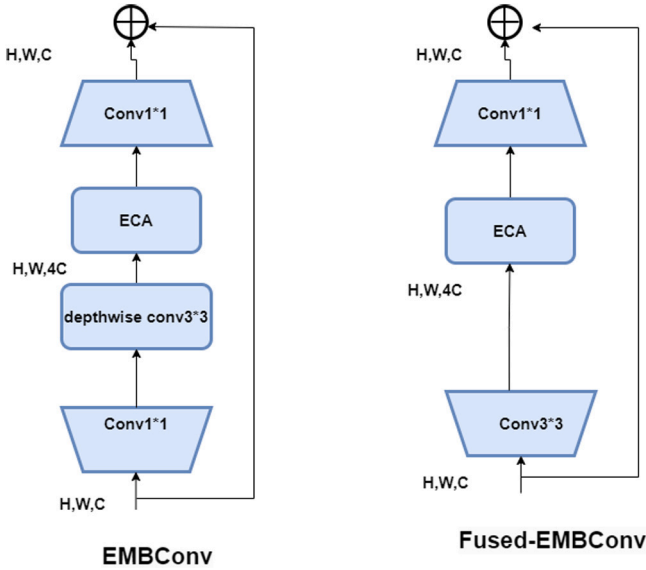


Fig. 7. EMBConv and FusedEMBConv.

It involves various biomarkers such as brain modalities and heterogeneous data (genetics, cognitive tests) about subjects with different disease stages (Alzheimer’s disease, mild cognitive impairment, and healthy controls) to study the disease progression. The dataset. In addition, it encompasses data about over 1,500 participants who were followed longitudinally for up to 10 years. Each participant is characterized by different data, such as clinical and cognitive assessments, brain modalities, and genetic data.

3.4. Data pre-processing

The vital step for effective feature extraction is image pre-processing. We applied skull stripping to remove irrelevant information from the MRI and PET images, such as the non-brain tissues. This technique reduces the complexity by focusing only on the relevant regions associated with the disease. We applied bias field correction to the MRI images using the N4ITK method to correct the intensity. The image resizing is also a crucial step to prepare the MRI image and adjust its size according to the proposed model. We enhance the input image by removing background noise from these images. We adopted the FreeSurfer software to apply motion correction, intensity normalization, and skull stripping for the brain image modalities.

3.5. Data augmentation

Data augmentation is a technique to increase the size of the training data. It consists of applying a set of transformations such as cropping, rotation, scaling, flipping, zooming, and translation to the existing dataset to generate new examples. This technique improves the model generalization to capture more deep features from the images.

3.6. Multimodal methods

In the first part, we developed a unimodal method based on the MRI modality. Multi-modal image fusion consists of combining different brain modalities for AD detection, such as MRI and PET. MRI is a non-invasive imaging tool that detects changes in the brain’s structure related to the disease such as the shrinkage of the hippocampus. This technique also measures the volume of relevant brain regions that can be affected by the disease and illustrates the atrophy of these regions.

PET scans have also shown promising results in the detection of Alzheimer’s disease. This technique illustrates critical hallmark features of AD disease, such as the existence of beta-amyloid plaques and tau tangles in the brain. Furthermore, the PET modality provides valuable information about the metabolic activity in the brain, which can be useful for diagnosing early-stage AD. Several studies applied deep learning to the MRI modalities to track and analyze structural changes related to the disease, such as brain volume, cortical thickness. Deep learning methods have also been applied to the PET modality to analyze the abnormal metabolic activity and the abnormal protein accumulation related to the disease. In addition, deep learning methods are used based on this modality to track and predict disease progression by detecting the structural changes that occur in the brain. The key step toward an effective diagnosis of Alzheimer’s disease is building a comprehensive and complete picture of the biomarkers related to the disease. Several studies have noticed that combining brain modalities provides an accurate and comprehensive view for the diagnosis of Alzheimer’s disease.

Each modality provides information related to the disease, and extracting and combining this information enhances the disease diagnosis. In this section, we propose two different multi fusion modalities for Alzheimer diagnosis. The first method consists of combining the MRI and PET modalities. The second method is based on the fusion of the MRI and CT modalities.

1. The first step is data preprocessing, which is a vital step that consists of preparing the images to have the same resolution and be aligned spatially. This includes image registration, image resizing, intensity normalization.
2. The second step is to feed each modality into the proposed model for feature extraction. The proposed model extracts features from the MRI and the PET modality, and then we fuse the obtained features.
3. The fusion step consists of fusing the vector features of each modality into a single vector using a fusion technique such as the late or early fusion technique. The fusion stage consists of fusing the features extracted from the MRI and PET modality and feeding the result into a single classifier, as depicted in Fig. 9.
4. The last step is the classification, which consists of passing the obtained fused features for classification and obtaining the output class of the disease.

We proposed two multi-modal methods, the first combining the MRI and the PET modality. Each modality is fed into the network in parallel way and we fuse the extracted features from each modality. The second method is based on the fusion of MRI and CT.

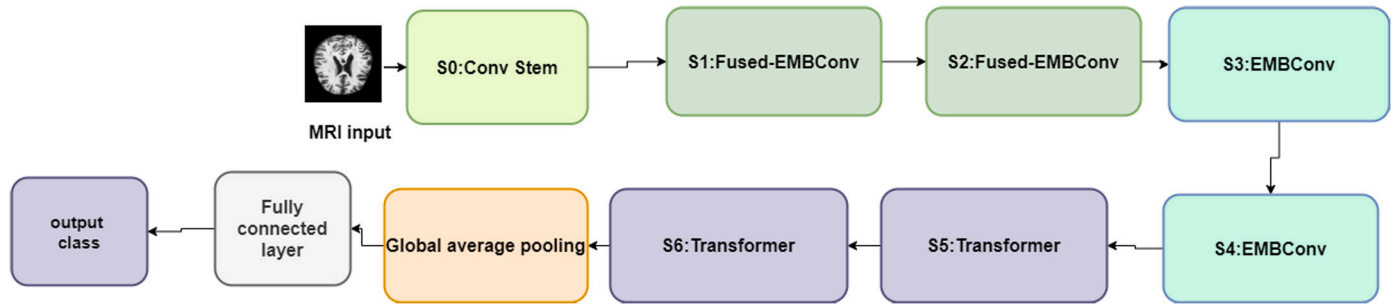


Fig. 8. Modified CoAtNet.

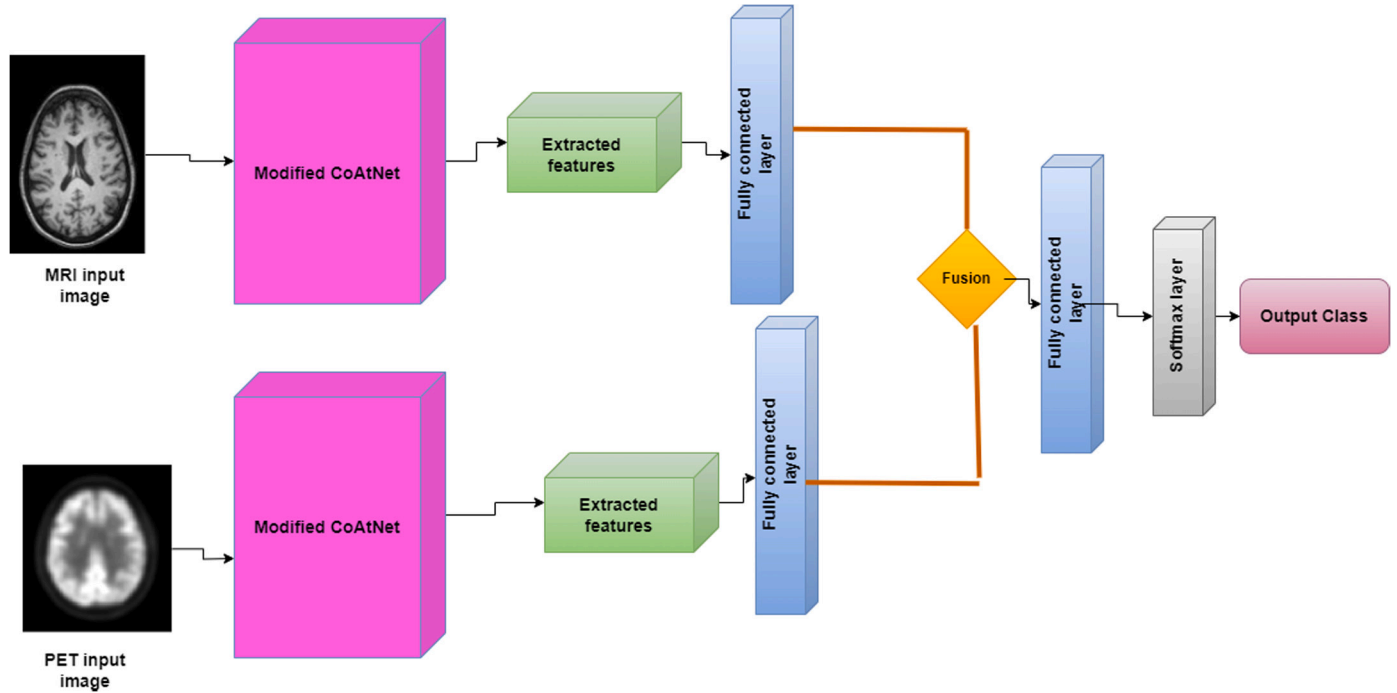


Fig. 9. Fusion modality.

4. Experimental results and discussion

4.1. Evaluation metrics

The metrics used to evaluate our proposed methods are accuracy, precision, recall and F1-score. Where the accuracy is defined as the ratio of the number of correct predictions and the total number of samples. Recall is a metric that represents the ratio between the actual correct prediction (true positives) and the total prediction made by the proposed network (true positives + false positives). The precision is the ratio of true positives over the sum of false positives and true negatives. F1 score is a weighted average of precision and recall.

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

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

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

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

4.2. Comparison between the classification results of the proposed methods and basic convolutional neural networks

We evaluate different CNN models such as ResNet50, VGG19, Densenet 169, and EfficientNet. The main advantage of the CNN models is that they have strong inductive biases that ensure an effective model generalization. CNN models have the reliable capability to extract deep features from the MRI and PET modalities. Furthermore, CNN adopted parameter sharing which reduced the number of parameters. It extracts patterns at different levels. The first layers capture low-level features, whereas the high-level layers identify deep features. However, CNN requires a large amount of data, which is a big challenge due to the limited labeled neuroimaging data. In addition, It relies on a limited receptive field due to the fixed size of the filters used in feature extraction, while expanding the CNN receptive field increases network complexity and computational cost. This limited receptive field makes the CNN focusing only on the local features. CNN has demonstrated its unique effectiveness as a local feature extractor. Nevertheless, The brain modalities are potential biomarkers that involve spatial information and capture the different regions that are affected by the disease. In addition, the brain modalities capture the changes that occur in the entire brain, making it crucial to capture global features to understand the

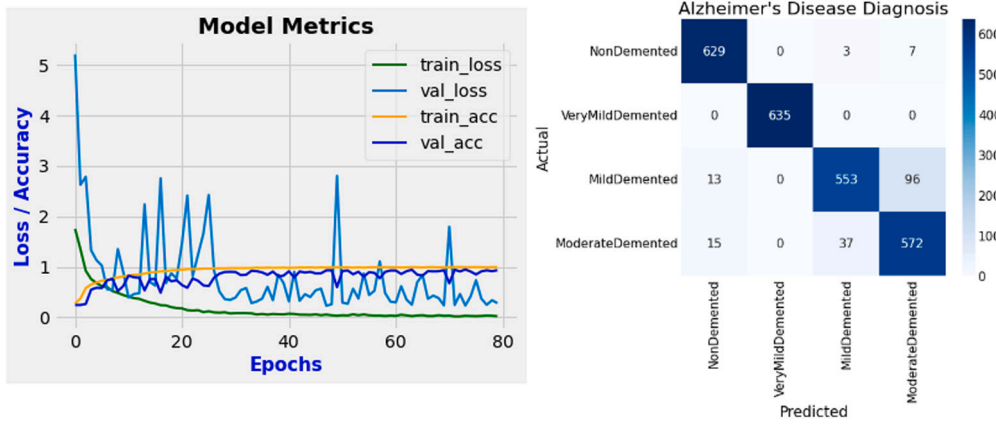


Fig. 10. The Output of the Swin transformer with an improved EfficientNet B0.

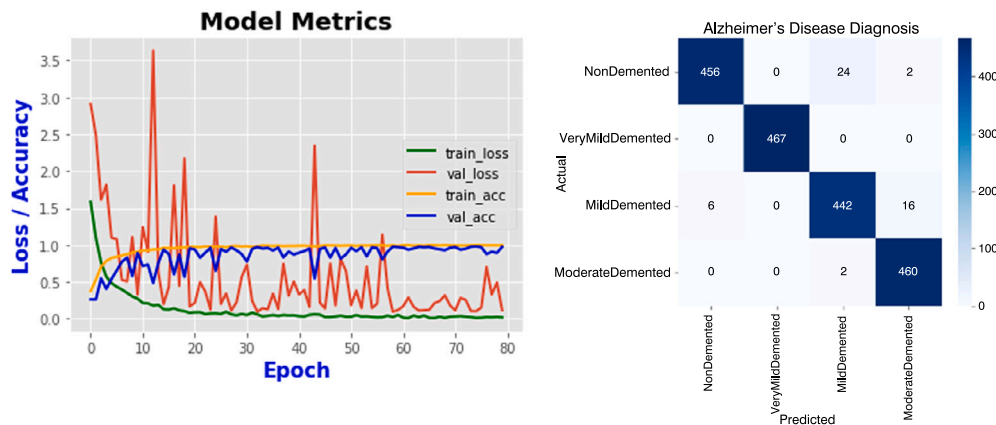


Fig. 11. Model metrics of Modified CoAtNet on OASIS dataset.

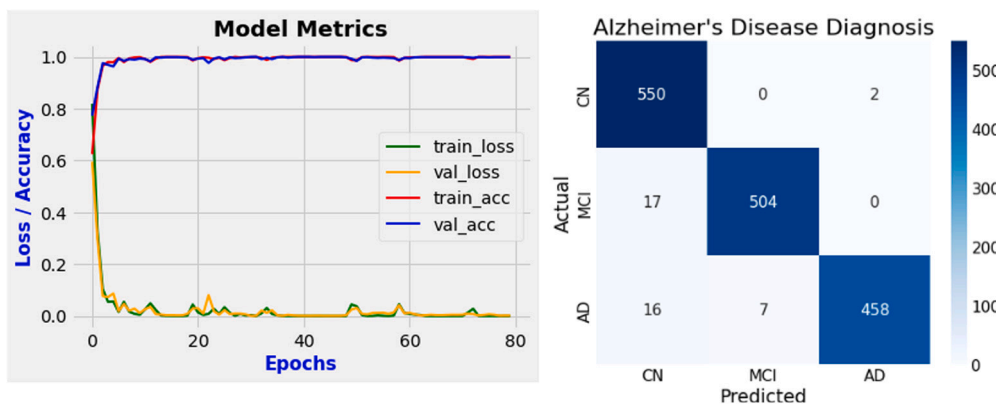


Fig. 12. Application of Modified CoAtNet on ADNI dataset.

overall brain structure and identify patterns associated with Alzheimer’s disease. Whereas CNN cannot capture long-range dependencies and the complex relations or interactions between the different modalities or image regions. Thus, CNN cannot extract the global features within the brain modalities and cannot encode the subtle changes depicted in the input brain modality. The proposed models extracted global and local features and outperformed the CNN models, as depicted in the Table 2. We compared our proposed models with some CNN models based on various metrics. VGG16 achieved 78%, ResNet50 85%, Densenet 169 89% and EfficientNetb0 87% as test accuracy. In addition, compared to the CNN models, our proposed models capture long-range dependencies between the image regions associated with the disease and extract global feature.

4.3. Comparison between the classification results of the proposed methods and the vision transformer

Vision transformer relies on a self attention mechanism that enables it to capture long-range dependencies between the MRI image regions. It handles a great potential in extracting global features from the image. Vision transformer require large-scale datasets, which is a big challenge. Furthermore, it lacks from the inductive bias, which makes it harder to train and has an high computational complexity. The transformers ensure better scalability than CNNs. The hyperparameters, the dataset, network depth, and the regularization methods are key factors in the vision transformer’s performance. The network is hard to optimize. Self-attention is a core component of vision transformer. However, the main

Table 2
Comparative table between our proposed models, CNN models and Vision Transformer.

Network	Accuracy (Test)	Recall	Precision	Training Accuracy
VGG16	78%	74.45%	77.65	79.95
ResNet50	85%	85.78%	86.72	88.87
Densenet 169	89%	89.9%	89.92	91.97
EfficientNetb0	87%	87.22%	79.07	89.49
ViT	79%	79.36%	79.59	81.87
Proposed model 1	93.23%	93.86%	93.52	99.92
Proposed model 2 (OASIS dataset)	97.33%	97.34%	97.35%	99.96%
Proposed model 2 (ADNI dataset)	98.87%	98.88%	98.89%	99.98%
Multi-modal (PET and MRI)	99.42%	99.42%	99.42%	99.99%
Multi-modal (CT and MRI)	94.55%	94.57%	94.59%	99.62%

Table 3
Comparative table between our proposed models and recent models based on the test accuracy.

Source	Model	Modality	Accuracy
[35]	ensemble CNN models (DenseNet196, VGG16 and ResNet50)	MRI	89%
[12]	3D DenseNet+ wise attention mechanism	MRI	87.28%
[8]	3Multi-Stream CNN	MRI	85.96%
[13]	Multilayer perceptron+multihead ProbSparse self-attention +structural distilling	MRI	92.8%
[17]	Multi-layer perceptron+ ViT	MRI	89.58%.
[23]	ConvNeXt and ensemble of machine Learning classifiers	MRI	92%
[3]	Cascaded Modality Transformers architecture with cross-attention	MRI	94%
[30]	Resnet18	PET+MRI	73.90%
[24]	ResNet-50+SVM	MRI+PET	94%
[26]	3DCNN	PET+MRI	71%
Proposed model 1 (Ours)	Enhanced EfficientNet and Swin transformer	MRI	93.23%
Proposed model 2 (OASIS dataset) (Ours)	Modified CoAtNet	MRI	97.33%
Proposed model 2 (Ours) (ADNI dataset)	Modified CoAtNet	MRI	98.87%
Multi-modal (Ours)	Modified CoAtNet	MRI+PET	99.42%
Multi-modal (Ours)	Modified CoAtNet	MRI+CT	94.55%

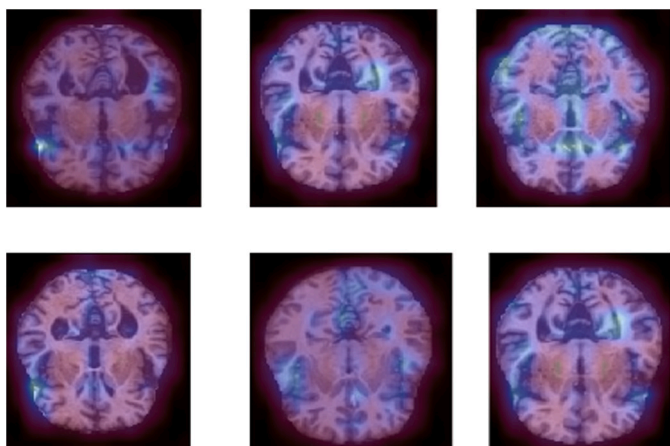


Fig. 13. Visualizations with Gradient-based Localization (Grad-CAM).

limitation of the self attention is the quadratic computational complexity with the size of the image, which that make the model hard to train. Furthermore, ViT cannot extract local features. We evaluated the ViT architecture and we obtained an accuracy of 79%. Our proposed methods address these issues. It extracts global and local information within the MRI data compared to the CNN models that capture only local features and the vision transformer that capture only global features. Our proposed methods combine the generalization capability of CNN with the robust feature representation of Transformer. The first model, which combines the EfficientNet and the Swin transformer achieves a 93.23% accuracy as depicted in Fig. 10.

We improve the EfficientNet by using multi-head attention to extract the relevant feature of the MRI image and the DO-Conv layer to boost the network performance without increasing the model complexity. The second method is based on an improved CoAtNet using the ECA module and fused mbconv to enhance the model's feature extraction and performance. The second method achieved 97.33% accuracy of classifi-

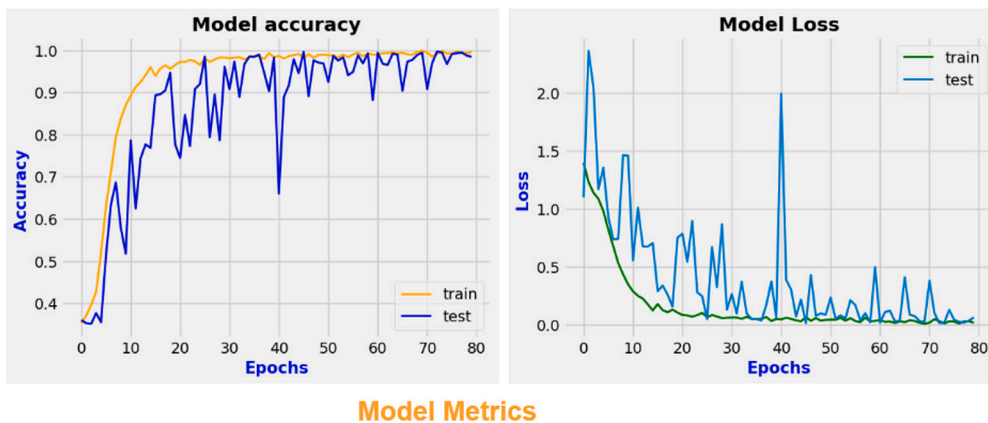


Fig. 14. Fusion modality PET and MRI based on Modified CoAtNet.

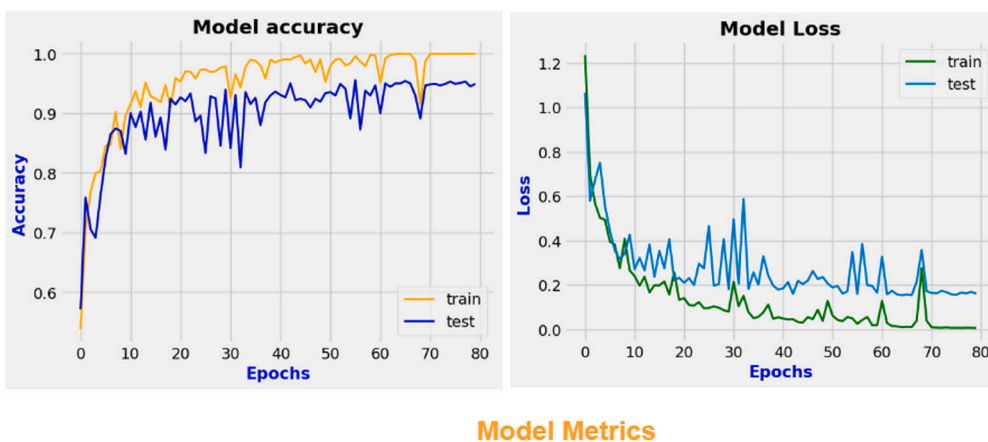


Fig. 15. Fusion modality MRI and CT based on Modified CoAtNet.

cation on the OASIS as depicted in Fig. 11 and 98.87% on ADNI dataset (Fig. 12). Another advantage of our proposed models is that we applied two different multi-fusion modalities. We combine the PET and MRI modalities to extract more accurate information from different brain modalities. See Fig. 13. This method achieved 99.42% accuracy, as illustrated in the Fig. 14. The second method combines the fusion of the MRI and CT modalities. This method has 94.55% accuracy, as depicted in Fig. 15.

4.4. Comparison of the proposed models with recent methods

In this section, we compare our proposed models with recent methods based on CNN, vision transformer and hybrid methods that combine CNN and vision transformer. Our experiments were conducted based on two datasets (ADNI and OASIS). The Table 3 provide a comparison of the proposed models with recent methods.

As depicted in 3 our proposed methods outperformed various state-of-the-art Alzheimer’s disease diagnosis methods based on CNN, ViT, and hybrid methods. The main advantage of Our multi-modal method is among the first fusion modalities methods that applied recent architecture transformers and CNN for AD diagnosis because most multimodal studies are based on CNN. Thus, our methods combine the strengths of CNN and transformers and address the issues of the methods that are based on CNN and the ViT based methods. Our proposed methods extract global and local features. Further, it ensures good generalization, enhances feature representation, and extraction with reduced computation and memory requirements.

5. Conclusion

In this paper, we have proposed two new methods based on recent architectures. The first method consists of a Swin transformer and an improved EfficientNet using multihead attention and a DO-Conv layer. The second method is based on an improved CoAtNet. We modify the structure of the building blocks of this network, the MBConv, by replacing the SE module with the ECA module, and we add an improved fused MBConv as early layers to enhance the generalization capability and reduce the model complexity. Our proposed model shows astonishing results compared to CNN models. We improved the disease diagnosis by proposing two multi-modalities methods. The first method combines features from the PET and MRI modalities. Whereas the second method fuses the CT and MRI modalities. The fusion method based on PET and MRI outperforms the fusion method based on MRI and CT. The appropriate choice of brain modalities is critical for an affective disease diagnosis.

Ethical approval

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Funding

Not applicable.

References

- [1] X. Feng, F.A. Provenzano, S.A. Small, A deep learning MRI approach outperforms other biomarkers of prodromal Alzheimer's disease, *Alzheimer's Res. Ther.* 14 (1) (Mar. 2022), <https://doi.org/10.1186/s13195-022-00985-x>.
- [2] M. Dong, L. Xie, S.R. Das, J. Wang, L.E. Wisse, R. deFlores, D.A. Wolk, P.A. Yushkevich, DeepAtrophy: teaching a neural network to detect progressive changes in longitudinal MRI of the hippocampal region in Alzheimer's disease, *NeuroImage* 243 (2021) 118514, <https://doi.org/10.1016/j.neuroimage.2021.118514>.
- [3] R. Kushol, A. Masoumzadeh, D. Huo, S. Kalra, Y.-H. Yang, Addformer: Alzheimer's disease detection from structural mri using fusion transformer, in: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.
- [4] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, N. Jacobs, Advit: vision transformer on multi-modality pet images for Alzheimer disease diagnosis, in: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–4.
- [5] R. Kadri, B. Bouaziz, M. Tmar, F. Gargouri, CrossViT wide residual squeeze-and-excitation network for Alzheimer's disease classification with self attention ProGAN data augmentation, *Int. J. Hybrid Intell. Syst.* 17 (3–4) (2022) 163–177, <https://doi.org/10.3233/his-220002>.
- [6] S. Kim, P. Lee, K.T. Oh, M.S. Byun, D. Yi, J.H. Lee, Y.K. Kim, B.S. Ye, M.J. Yun, D.Y. Lee, Y. Jeong, Deep learning-based amyloid PET positivity classification model in the Alzheimer's disease continuum by using 2-[18F]FDG PET, *EJNMMI Res.* 11 (1) (Jun. 2021), <https://doi.org/10.1186/s13550-021-00798-3>.
- [7] A.B. Tufail, N. Anwar, M.T.B. Othman, I. Ullah, R.A. Khan, Y.-K. Ma, D. Adhikari, A.U. Rehman, M. Shafiq, H. Hamam, Early-stage Alzheimer's disease categorization using PET neuroimaging modality and convolutional neural networks in the 2d and 3d domains, *Sensors* 22 (12) (2022) 4609, <https://doi.org/10.3390/s22124609>.
- [8] M. Ashtari-Majlan, A. Seifi, M.M. Dehshibi, A multi-stream convolutional neural network for classification of progressive MCI in Alzheimer's disease using structural MRI images, *IEEE J. Biomed. Health Inform.* 26 (8) (2022) 3918–3926, <https://doi.org/10.1109/jbhi.2022.3155705>.
- [9] S. Sharma, K. Guleria, S. Tiwari, S. Kumar, A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer disease using MRI scans, *Meas. Sens.* 24 (2022) 100506, <https://doi.org/10.1016/j.measen.2022.100506>.
- [10] S.-H. Wang, Q. Zhou, M. Yang, Y.-D. Zhang, ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation, *Front. Aging Neurosci.* 13 (Jun. 2021), <https://doi.org/10.3389/fnagi.2021.687456>.
- [11] W. Zhu, L. Sun, J. Huang, L. Han, D. Zhang, Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI, *IEEE Trans. Med. Imaging* 40 (9) (2021) 2354–2366, <https://doi.org/10.1109/tmi.2021.3077079>.
- [12] J. Zhang, B. Zheng, A. Gao, X. Feng, D. Liang, X. Long, A 3d densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification, *Magn. Reson. Imaging* 78 (2021) 119–126, <https://doi.org/10.1016/j.mri.2021.02.001>.
- [13] J. Zhu, Y. Tan, R. Lin, J. Miao, X. Fan, Y. Zhu, P. Liang, J. Gong, H. He, Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis, *Comput. Biol. Med.* 147 (2022) 105737, <https://doi.org/10.1016/j.cmpbiomed.2022.105737>.
- [14] B. Yan, Y. Li, L. Li, X. Yang, T. qiang Li, G. Yang, M. Jiang, Quantifying the impact of pyramid squeeze attention mechanism and filtering approaches on Alzheimer's disease classification, *Comput. Biol. Med.* 148 (2022) 105944, <https://doi.org/10.1016/j.cmpbiomed.2022.105944>.
- [15] Z. Pei, Z. Wan, Y. Zhang, M. Wang, C. Leng, Y.-H. Yang, Multi-scale attention-based pseudo-3d convolution neural network for Alzheimer's disease diagnosis using structural MRI, *Pattern Recognit.* 131 (2022) 108825, <https://doi.org/10.1016/j.patcog.2022.108825>.
- [16] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, N. Jacobs, Advit: vision transformer on multi-modality pet images for Alzheimer disease diagnosis, in: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2022, pp. 1–4.
- [17] Y. Duan, R. Wang, Y. Li, Aux-vit: classification of Alzheimer's disease from mri based on vision transformer with auxiliary branch, in: *2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2023, pp. 382–386.
- [18] Q. Zhao, G. Huang, P. Xu, Z. Chen, W. Li, X. Yuan, G. Zhong, C.-M. Pun, Z. Huang, IDA-net: inheritable deformable attention network of structural MRI for Alzheimer's disease diagnosis, *Biomed. Signal Process. Control* 84 (2023) 104787, <https://doi.org/10.1016/j.bspc.2023.104787>.
- [19] Z. Hu, Z. Wang, Y. Jin, W. Hou, VGG-Tswinformer: transformer-based deep learning model for early Alzheimer's disease prediction, *Comput. Methods Programs Biomed.* 229 (2023) 107291, <https://doi.org/10.1016/j.cmpb.2022.107291>.
- [20] J. Jang, D. Hwang, M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20686–20697.
- [21] C. Tang, M. Wei, J. Sun, S. Wang, Y. Zhang, CsAGP: detecting Alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling, *J. King Saud Univ, Comput. Inf. Sci.* 35 (7) (2023) 101618, <https://doi.org/10.1016/j.jksuci.2023.101618>.
- [22] Y. Yin, W. Jin, J. Bai, R. Liu, H. Zhen, Smil-deit: multiple instance learning and self-supervised vision transformer network for early Alzheimer's disease classification, in: *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–6.
- [23] C. Techa, M. Ridouani, L. Hassouni, H. Anoun, Automated Alzheimer's disease classification from brain MRI scans using ConvNeXt and ensemble of machine learning classifiers, in: *Lecture Notes in Networks and Systems*, Springer Nature Switzerland, 2023, pp. 382–391.
- [24] S. Dwivedi, T. Goel, M. Tanveer, R. Murugan, R. Sharma, Multimodal fusion-based deep learning network for effective diagnosis of Alzheimer's disease, *IEEE Multimed.* 29 (2) (2022) 45–55, <https://doi.org/10.1109/mmul.2022.3156471>.
- [25] Z. Kong, M. Zhang, W. Zhu, Y. Yi, T. Wang, B. Zhang, Multi-modal data Alzheimer's disease detection based on 3d convolution, *Biomed. Signal Process. Control* 75 (2022) 103565, <https://doi.org/10.1016/j.bspc.2022.103565>.
- [26] J. Song, J. Zheng, P. Li, X. Lu, G. Zhu, P. Shen, An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis, *Front. Dig. Health* 3 (Feb. 2021), <https://doi.org/10.3389/fdgh.2021.637386>.
- [27] D. Vaghari, E. Kabir, R.N. Henson, Late combination shows that MEG adds to MRI in classifying MCI versus controls, *NeuroImage* 252 (2022) 119054, <https://doi.org/10.1016/j.neuroimage.2022.119054>.
- [28] A. Fedorov, L. Wu, T. Sylvain, M. Luck, T.P. DeRamus, D. Bleklov, S.M. Plis, V.D. Calhoun, On self-supervised multimodal representation learning: an application to Alzheimer's disease, in: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1548–1552.
- [29] P. Sherwani, P. Nandhakumar, P. Srivastava, J. Jagtap, V. Narvekar, R. Harikrishnan, Comparative analysis of Alzheimer's disease detection via mri scans using convolutional neural network and vision transformer, in: *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, 2023, pp. 1–9.
- [30] M. Odusami, R. Maskeliūnas, R. Damaševičius, S. Misra, Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images, *J. Med. Biol. Eng.* 43 (3) (2023) 291–302, <https://doi.org/10.1007/s40846-023-00801-3>.
- [31] J. Zhang, X. He, Y. Liu, Q. Cai, H. Chen, L. Qing, Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data, *Comput. Biol. Med.* 162 (2023) 107050, <https://doi.org/10.1016/j.cmpbiomed.2023.107050>.
- [32] J. Cao, Y. Li, M. Sun, Y. Chen, D. Lischinski, D. Cohen-Or, B. Chen, C. Tu, DO-conv: depthwise over-parameterized convolutional layer, *IEEE Trans. Image Process.* 31 (2022) 3726–3736, <https://doi.org/10.1109/tip.2022.3175432>.
- [33] Z. Dai, H. Liu, Q.V. Le, M. Tan, Coatnet: marrying convolution and attention for all data sizes, *Adv. Neural Inf. Process. Syst.* 34 (2021) 3965–3977.
- [34] R. Karthik, T.S. Vaichole, S.K. Kulkarni, O. Yadav, F. Khan, Eff2net: an efficient channel attention-based convolutional neural network for skin disease classification, *Biomed. Signal Process. Control* 73 (2022) 103406, <https://doi.org/10.1016/j.bspc.2021.103406>.
- [35] C. Techa, M. Ridouani, L. Hassouni, H. Anoun, Alzheimer's disease multi-class classification model based on CNN and StackNet using brain MRI data, in: *Proceedings of the 8th International Conference on Advanced Intelligent Systems and Informatics 2022*, Springer International Publishing, 2022, pp. 248–259.

Rahma Kadri received her Master Thesis degree in Computer Science from Higher Institute of Computer Science and Multimedia, Sfax University, Tunisia in 2018. She is currently a PhD student at Sfax University, Tunisia and a member of Multimedia, Information systems and Advanced Computing Laboratory (MIRACL). Her research areas include Image Processing, Machine Learning and Deep Learning.

Bassem Bouaziz Hold a PhD in computer science from University of Sfax. He is currently University council member. He is senior member of the Digital Research Center of Sfax (CRNS). He coordinates several multinational projects on biodiversity informatics and digital health technologies using Artificial Intelligence in partnership with industry and academia. His research areas include Multimedia document indexing and processing, Computer Vision for video analysis, Deep learning, biomedical signals processing.

Mohamed Tmar Currently, he is Associate Professor at the Department of Computer Science of the Higher Institute of Computer Science and Multimedia at the University of Sfax, Tunisia. He is a member of the Multimedia, Information systems and Advanced Computing Laboratory, University of Sfax. His main research areas are Multimedia document indexing and processing, Computer Vision for video analysis, Deep learning, Objects recognition.

Faiez Gargouri Professor of Computer Science at University of Sfax, he is a member of the Multimedia, Information systems and Advanced Computing Laboratory and

Vice-President of the University of Sfax. He was the head of the Higher Institute of Computer science and Multimedia from 2007 to 2011. He has got his maitrise diploma in computer management, faculty of economics and management of Sfax (1988), his master in computer science from the Paris 6 University (1990) and his PhD thesis, Paris 5 Uni-

versity (1995). He got his Habilitation Degree in Computer Science, Faculty of Sciences of Tunis (2002). His research interests include Business Information Systems, Business Intelligence, multimedia Information systems, Ontology, Deep learning.