IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2024, accepted 10 July 2024, date of publication 18 July 2024, date of current version 29 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430325

RESEARCH ARTICLE

Alzheimer's Disease and Mild Cognitive Impairment Detection Using sMRI With Efficient Receptive Field and Enhanced Multi-Axis Attention Fusion

UTTAM KHATRI, JUN-HYUNG KIM, AND GOO-RAK KWON, (Senior Member, IEEE)
Department of Information and Communication Engineering, Chosun University, Gwangju 61452, South Korea

Corresponding author: Goo-Rak Kwon (grkwon@chosun.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2021R1I1A3050703, in part by the BrainKorea21Four Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 4299990114316, in part by the Following are results of a study on the ''Leaders in Industry-University Cooperation 3.0'' Project, in part by the Ministry of Education and National Research Foundation of Korea, in part by the Alzheimer's Disease Neuroimaging Initiative (ADNI) funded by the National Institutes of Health under Grant U01 AG024904, in part by the Department of Defense ADNI under Award W81XWH-12-2-0012, and in part by the ADNI's Clinical Centers in Canada funded by The Canadian Institutes of Health Research.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Chosun Institutional Review Board under Application No. 2-1041055-AB-N-01-2021-34.

ABSTRACT Deep neural networks have shown promising results in the analysis of structural magnetic resonance imaging (sMRI) data for the diagnosis of dementia, particularly Alzheimer's disease (AD). Different regions of the brain have diverse structures that are linked to specific functions, which could account for the variability in disease-related changes observed in sMRI scans of these areas. Understanding the overall characteristics of sMRI data is important since current popular convolutional neural networks (CNN) for deep learning do not consider the interconnection of voxels. Vision transformers have shown effectiveness in identifying long-distance connections in the brain, which has led to their success in applications such as disease detection. However, the image noise and limited scalability of self-attention mechanisms in relation to image size has hindered their widespread use in advanced Alzheimer's analysis. To enhance information retention and reduce network complexity, this study presents a novel adaptable efficient receptive field feature extraction network. Moreover, an advanced attention mechanism with both local grid attention block and dilated global attention module has been incorporated to highlight the characteristics of AD. Next, a more improved hierarchical inverted residual feed forward network in place of multi-layer perceptron is suggested to enhance the characterization of features through the integration of information from both lower and higher layers. Finally, the global average pooling and 1 × 1 convolution are used to reduce dimensionality, enhance non-linearity, and allow channel interactions in feature maps before being input into the classification head. The network achieved high performance in various scenarios, with average accuracies of 97.29% for AD vs. HC and 94.79% for MCI vs. HC classification using ADNI as experimental datasets.

INDEX TERMS Alzheimer's disease, CNN, sMRI, MBConv, partial convolution, vision transformer, block attention, grid attention.

I. INTRODUCTION

Neurodegenerative disorders refer to a set of ailments marked by the gradual decline of brain nerve cells, resulting in cognitive impairment, decreased mental functioning, and loss of

The associate editor coordinating the review of this manuscript and approving it for publication was Hasan S. Mir.

motor control [1], [2]. The prevalence of these illnesses is increasing worldwide [3]. As one of the most complicated and challenging neurodegenerative diseases, Alzheimer's disease emphasizes the significance of accurate disease identification and inclusive data collection for prompt detection [4]. The cognitive deterioration associated with Alzheimer's can be classified into three phases: mild demented, non-demented, and demented. These stages correspond to the disease's course [5]. In this field, there are distinct stages such as cognitively healthy (HC), mild cognitive impairment (MCI), and Alzheimer's disease (AD) [6]. To gather information about a patient, it is necessary to review their medical background, consider environmental influences, analyze genetic information, and perform a range of medical examinations using diagnostic imaging methods to understand the patient's anatomy and functionality [7]. Neuroimaging is an important area of study that centers on examining the structure of the brain and identifying biomarkers using various imaging techniques. Techniques such as sMRI and positron emission tomography (PET) scans are effective tools for characterizing diseases and assisting in precise diagnosis [8]. These methods provide a large amount of complex and varied healthcare data, which highlights how crucial it is to gather insightful information and develop specific disease profiles from this data [9]. sMRI is a non-invasive and effective technique that can be used to investigate and evaluate the changes in brain structure caused by Alzheimer's. These changes are crucial in clinical settings and play a major role in understanding the progression of disease [10].

CNN have demonstrated impressive achievements in deep learning (DL) techniques and have displayed exceptional performance in diagnosing brain diseases such as AD [11], [12]. CNNs are commonly utilized for the classification, detection, and prediction of AD based on brain imaging modalities like sMRI scans [13]. DL is utilized to analyze brain images in both 2D and 3D, aiming to detect important biomarkers linked to AD, such as amyloid-beta plaques and tau protein tangles [14], [15]. Additionally, DL can pinpoint particular brain areas that are connected to AD. Multiple studies have highlighted the successful utilization of DL in AD research, encompassing tasks like image classification and disease progression prognosis [16], [17]. Due to the ambiguity of available data, which makes it difficult to establish which attributes are vital for solving the problem, which greatly influences the medical imaging field. Nonetheless, CNN relies on specific details in the images, with the primary difficulty lying in effectively modeling imaging data on a global scale [18]. The complexity of brain modalities is further compounded by significant differences in quality stemming from factors like the method of image acquisition, type of scanner used, and patient motion during scanning [17]. These discrepancies pose a challenge for CNNs in discerning valuable features from the images. Vision transformers (ViT) have recently emerged as a potential new approach for deep learning applications in computer vision, indicating exciting possibilities for the future [19], [20]. ViT has demonstrated encouraging outcomes in the field of medical image interpretation [21]. The attention mechanism is the fundamental technique employed in ViT [22]. One major benefit of transformers compared to CNNs is their ability to understand intricate connections and distant dependencies within image features through self-attention [23].

Nevertheless, it has been noted that ViT struggles in AD recognition tasks without extensive pre-training. This is due to ViT having a strong model capacity with less inductive bias, leading to underfitting or overfitting [24]. To address this issue and enhance scalability, various research efforts have focused on sparse Transformer models designed for AD prognosis tasks, including local attention [25], [26]. These approaches typically reintroduce basic ViT architectures to make up for the lack of non-locality in ViT for AD recognition. One successful example of such modifications is the convolution-Swinformer [27], which applies self-attention on shifted non-overlapping windows. The utilization of this method resulted in outperforming for the first time on the Alzheimer's dataset benchmark using a ViT. Window-based attention has demonstrated low model capacity due to the lack of non-locality and struggles to scale well with tiny imaging datasets like AD, even though it offers more flexibility and generality than full attention in ViT [28], [29]. Nevertheless, early or high-resolution phases of a hierarchical network including global interactions through complete attention necessitate a computationally demanding attention operation due to its quadratic complexity. In order to balance model capacity and generalizability within a certain computational budget, the effective integration of both global and local interactions continues to be a difficulty. In addition to pure ViT, there is a growing interest in creating efficient ViTs and multilayer perceptron (MLPs). For instance, [30] have been developed to decrease computational complexity by combining double normalization methods with a modified attention mechanism. However, these models continue to face challenges requiring specialized training support for the modified attention mechanism. Furthermore, the integration of complex normalization and activation layers may restrict their AD prediction capability to a certain level. In this article, we introduce a novel Transformer module with adaptable efficient receptive field by integrating MBConv with block attention and Partial convolution (PConv) with grid attention in alternative fashion, which effectively functions as a fundamental building block for AD diagnosis. Within each block, local and global spatial interactions can be performed using the suggested method. In comparison to full self-attention, it offers increased flexibility and efficiency, making it naturally adaptable to sMRI brain images with linear complexity. Unlike shifted window or local attention mechanisms [27], [31], proposed methods enable a larger model capacity by introducing an efficient local and global receptive field. Furthermore, proposed model can be utilized as a general standalone attention module with only linear

TABLE 1. Information on the characteristics of demographics.

Groups	Gender (M/F)	Education	Age (Years)	MMSE	CDR	APOEε4	FAQ
AD	160/155 #	15.47 ± 3.08	74.07 ± 7.5 *	24.54 ± 2.25 *	3.29± 1.7 *	0.88 ± 0.70	10.30± 7.05 *
MCI	200/170	15.52 ± 3.17	73.53 ± 7.6	27.42 ± 1.68	1.33 ± 0.74	0.65 ± 0.64	2.98 ± 3.50
HC	250/140	16.26 ± 3	74.76 ± 4.3	29.16 ± 0.95	0.04 ± 0.16	0.24 ± 0.46	0.12 ± 0.67

Values are means or numbers ± standard deviations. AD: Alzheimer's disease; HC: Healthy Control; MCI: Mild cognitive Impairment; MMSE: Mini Mental state Examination; CDR: Clinical Dementia Rate; FAQ: Functional Activities Questionnaires

*#Group-level chi-square tests are conducted for gender; *Group-level two-sample t-tests are conducted for age, education, MMSE, FAQ, and CDR.*

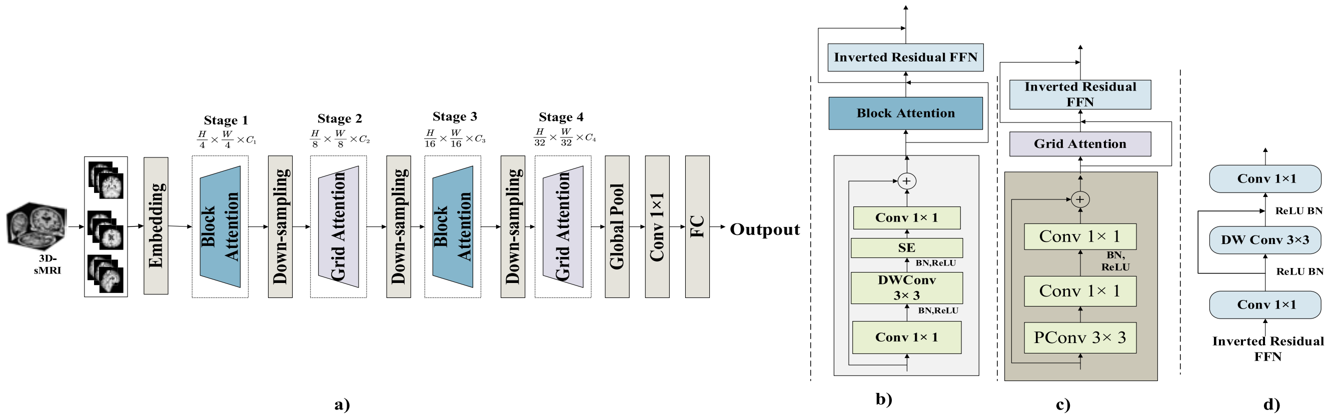


FIGURE 1. The suggested pipeline for AD diagnosis utilizing sMRI: a) Overall framework; b) MBCov with block attention and IRFFN; c) PConv with grid attention and IRFFN; d) inverted residual feed forward networks.

complexity, allowing it to be incorporated into any brain related conditions, including AD with high efficacy. Below are the main contributions of the research study:

1. Investigates Transformers' efficacy in AD diagnosis using T1-weighted sMRI data with efficient receptive field using MBConv and PConv module.
2. Proposes an attention fusion ViT architecture integrating block attention, and grid attention in alternate fashion for comprehensive feature representation and reduced processing costs with inverted residual feed forward network.
3. Validates superior performance on the ADNI dataset, highlighting improved accuracy, specificity, and sensitivity advancing brain sMRI analysis for AD diagnosis.

II. MATERIALS

A. DATASET

The information used in this study was obtained from the ADNI database, which can be accessed online at (http://adni.loni.usc.edu). The ADNI provides researchers with global access to an openly accessible database for the purpose of early detection of Alzheimer's Disease and the exploration of biological markers associated with the condition. For this research, we assessed 315 healthy control, 370 mild cognitive impairment, and 390 Alzheimer's disease samples. The MRI scans were conducted at a resolution of 3T and produced T1-weighted images using magnetization-prepared rapid-acquisition gradient-echo sequences. Each voxel in the acquired images measured 1 × 1 × 1 mm^3, and the resolution was 182 × 218 × 182. The individuals' MMSE scores, gender, age, and clinical dementia rate (CDR) are displayed in Table 1 together with other demographic and clinical data. Each of the three subject combinations had an equal distribution of ages. While there was noticeably larger variability in the results of the other two groups, the MMSE scores of the HC group only displayed minor differences.

B. DATA PREPROCESSING

Initially, the midpoint of the line connecting the anterior commissure (AC) and posterior commissure (PC) was selected as the new reference point for all original structural magnetic resonance imaging (sMRI) data. Next, we employed the computational anatomy toolbox (CAT12) for our study. This toolbox, available at the website http://www.neuro.uni-jena.de/cat/, consists of various morphometry methods such

TABLE 2. Specifications and parameters for architecture in detail.

Stage	Output Size	Parameters
Stem	112×112	3×3, 32, Stride 2 3×3, 32, Stride 1
Block1	56×56	2×2, 64, stride 2
		[MBConv, 64, E=4, R=4, Rel-MSA, P 8×8, H=4]×2
Block2	28×28	2×2, 128, stride 2
		[PConv, 128, R=4, r=1/4, Rel-MSA, G 8×8, H=4]×2
Block3	14×14	2×2, 256, stride 2
		[MBConv, 256, E=4, R=4, Rel-MSA, P 8×8, H=4]×5
Block4	7×7	2×2, 512, stride 2
		[PConv, 512, R=4, r=1/4, Rel-MSA, G 8×8, H=4]×2
	1×1	1×1,512
		2
		18.3M
		2.23B

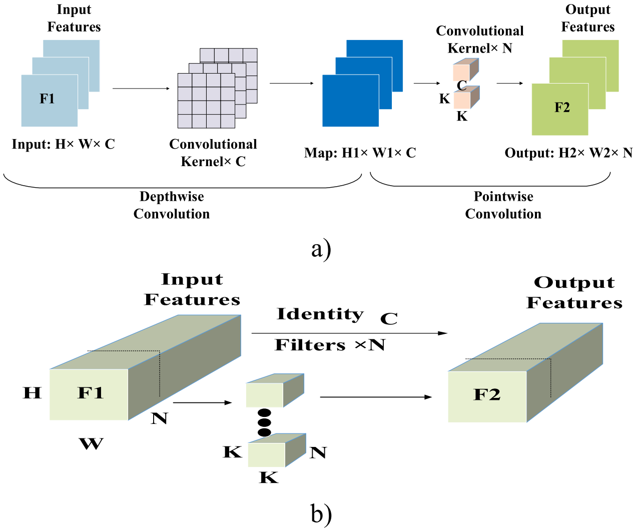


FIGURE 2. Illustration of a) Depthwise separable convolution and b) PConv.

as surface-based morphometry (SBM) and voxel-based morphometry (VBM) [32]. We implemented the following steps during preprocessing: elimination of non-brain tissue like the skull and neck. The process also involved normalization to the EPI template, modulation, and spatial smoothing with an 8 mm FWHM Gaussian filter.

III. METHODS

In this research, we propose a hybrid transformer design which integrate block and grid attention that improves the classification accuracy and efficiency on small sMRI datasets by adding MBConv and PConv processes to the main parts of the transformer and capturing more local information. Figure 1 and Table 2 present a summary of the suggested efficient convolution-transformer model for the Alzheimer's diagnostic task. This work employs a multi-stage architecture, wherein each block with a Rel-MSA and inverted residual feed forward network (IRFFN) comprise a comparable structure at each level. To enhance the receptive field of sMRI features, we alternate between inserting a MBConv with block attention and a PConv with grid attention block. This method helps retain important local and global features from each input image.

Incorporating two-convolution blocks enhances the model's capacity to preserve global information and capture local details. This prevents the loss of crucial information and enhances the utilization of specific, detailed data at a local level. Next, the features bypass the position embedding present in the original ViT and go via a block attention and grid attention made up of Rel-MSA, IRFFN, and classifier head. These blocks lower computing costs and increase transformer flexibility by effectively capturing both local and long-range interactions. The model's last layers consist of global average pooling, 1 × 1 convolution to perform a linear combination of input channels, enabling dimensionality reduction, feature transformation, and efficient parameter management, along classification layer with SoftMax activation.

A. MOBILE CONVOLUTION (MBCONV)

We utilize the MBConv block as the primary convolution operator with block attention. A pre-activation structure is also employed to enhance consistency between MBConv and Transformer blocks. More specifically, if X represents the input feature, the MBConv block without downsampling is expressed as:

$$X = X + Proj\left(SE\left(DWConv\left(Conv(X)\right)\right)\right) \quad (1)$$

In this architecture, BatchNorm is used for normalization, followed by a Conv1 × 1 expansion and ReLU activation, which is a common choice for Transformer-based models. The Depthwise Conv3 × 3 is utilized for DWConv, then BatchNorm and ReLU activation are applied. The Squeeze-Excitation layer is denoted as SE, while Proj refers to shrinking the number of channels with a Conv1 × 1 down projection. Down sampling is achieved in each step by using a stride-2 Depth wise Conv3 × 3 in the first MBConv block. Additionally, pooling and channel projection must be completed by the shortcut branch.

$$X = Proj\left(Pool\left(X\right)\right) + Proj(SE\left(DWConv\left(Conv(X)\right)\right)) \quad (2)$$

B. PARTIAL CONVOLUTION (PCONV)

The traditional approach to Convolution involves applying it only to a specific part of the input channel in order to capture spatial features, without altering the remaining parts of the channel [33]. The FLOPs of PConv are calculated as $H \times W \times K^2 \times N^2$. PWConv is added to the partial volume (PConv) in order to efficiently use data from all channels.

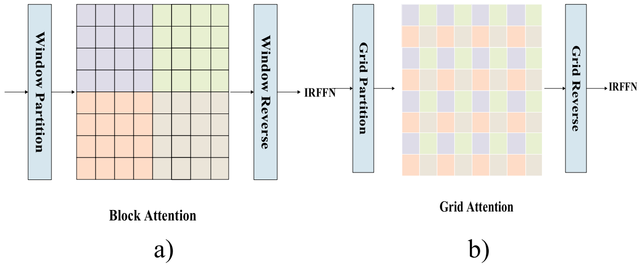


FIGURE 3. Illustration of a) Block attention and b) grid attention process.

Compared to the centrally concentrated Regular Convolution, the active sensing region on the input attribute network resembles a T-shaped Convolution. This T-shaped receptive field is justified by computing the position Frobenius norm, which quantifies the relative relevance of each position. If a position has a higher Frobenius norm than others, it is seen as more significant. Equation (3) is used to determine the Frobenius parameter for point i in the Conv filter $F \in R^{k^2 \times c}$.

$$\|F_i\| = Softmax\sqrt{\sum_{j=1}^{c} |f_{ij}|^2} i = 1, 2, 3, \ldots, k^2 \tag{3}$$

The results indicate that the most frequently observed position in the filter is the middle position. Accordingly, the central location is more important than the positions around it, which is consistent with the T-shaped calculation that is centered in the central position. Although T-shaped Convolution can be used directly for efficient calculation, it can also be divided into PConv and PWConv, as shown in Figure 2, to reduce filter redundancy and save FLOPs in the process. The FLOPs of a T-shaped Convolution can be determined as shown in Equation (4), which is larger than the FLOPs of PConv and PWConv as seen in Equation (5), for the same input $F_1 \in R^{c \times h \times w}$ and output $F_2 \in R^{c \times h \times w}$.

$$h \times w \times (k^2 \times N \times c + c \times (c - N) \tag{4}$$

$$h \times w \times \left(k^2 \times N^2 + c \times N\right) \tag{5}$$

Among them, $c > N, c - N > N$. Moreover, a two-step implementation can be effortlessly achieved with standard Conv. There are three stages in all, with two PWConv (or Conv1 × 1) layers after a PConv layer in each step.

C. ATTENTION MODULE

Several previous studies have investigated relative attention in both natural language processing (NLP) and computer vision. For the sake of simplicity, we will concentrate on our model in this study that uses only one head of the multi-head self-attention mechanism [34]. In practice, however, we always employ multi-head attention with the same head dimension. The following is a description of the relative attention concept:

$$RelAttn\,(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right) V \tag{6}$$

There is a hidden dimension d in the query, key, and value matrices Q, K and $V \in \mathbb{R}^{(H \times W) \times C}$, where H and W denote the matrices' height and width and C denotes the number of channels. Both a learnt static location-aware matrix B and the scaled input-adaptive attention $QK^T \big/ \sqrt{d}$ determine the attention weights. The relative position bias B is represented by a matrix $\hat{B} \in \mathbb{R}^{(2H-1)(2w-1)}$) to account for changes in 2D coordinates. Bilinear interpolation is used to transfer the relative positional bias from $\mathbb{R}^{(2H-1)(2W-1)}$to $\mathbb{R}^{(2H'-1)(2W'-1)}$ when fine-tuning at a higher resolution, such as $H' \times W'$. For 2D vision tasks, this relative attention is better than standard self-attention because it makes use of input-adaptivity, translation equivariance, and global interactions. Equation 6 defines this relative attention, which is the default setting for all attention operators in our model.

Since L and C stand for sequence length and channels, it is assumed that the relative attention operator in Equation 6 treats the second last dimension of an input $(\ldots, L, C)$ as the spatial axis in all cases. The self-attention operation does not require any changes in order to employ the suggested Multi-Axis Attention. First, the input image/feature$X \in \mathbb{R}^{H \times W \times C}$ is divided into non-overlapping blocks with a block size of $P \times P$ using the Block $(\cdot)$ operator with parameter P as shown in Figure 3a. The block dimensions are integrated onto the spatial dimension (i.e., -2 axis) following window division, which is very crucial to remember.

$$Block\,(H, W, C) = \left(\frac{H}{P} \times P, \frac{W}{P} \times P, C\right) = \frac{HW}{p^2}, p^2, C) \tag{7}$$

The block partition process previously described is the opposite of the operation known as Unblock $(\cdot)$. Similar to this, the operation Grid $(\cdot)$ is described using the parameter G as follows: the input feature is split into a uniform grid measuring $G \times G$, with a size of $H\big/G \times W\big/G$ for each lattice as shown in Figure 3b. The grid dimension needs to be adjusted to the anticipated spatial axis (i.e., -2 axis) using a transpose, as opposed to a grid operator. The inverse procedure Ungrid $(\cdot)$ can be used to return the original gridded input to the regular 2D feature space.

$$Grid\,(H, W, C) = \left(G \times \frac{H}{G}, G \times \frac{W}{G}, C\right) = G^2\frac{HW}{G^2}, \\ C = \frac{HW}{G^2}, G^2, C \tag{8}$$

We will now talk about the local Block Attention to help make the concept of the multi-axis attention module clearer. If X is an input tensor that lies within $X \in \mathbb{R}^{H \times W \times C}$, then the local Block Attention can be expressed as follows:

$$X = X + Unblock(RelAttn(Block\,(BN\,(X))) \tag{9}$$

$$X = X + IRU\,(BN\,(X)) \tag{10}$$

The extended global Grid Attention module is described as follows:

$$X = X + Ungrid(RelAttn(Grid\,(BN\,(X))) \tag{11}$$

$$X = X + IRU\left(BN\left(X\right)\right) \tag{12}$$

We simplify the RelAttn operation by not including the QKV input format. Layer Normalization is represented by LN; we used IRU, a standard inverted residual network with DWConv, skip connection, and point-wise convolution, in place of MLP.

D. INVERTED RESIDUAL FEED FORWARD NETWORK

To increase efficiency, a particular kind of residual block used in this vision models is the Inverted Residual Block, also called the MBConv Block. It was first included in the CNN design of MobileNet [35] and has since been used in a number of CNNs that have been tailored for mobile devices. The configuration of the Inverted Residual Block is narrow-wide-narrow, in contrast to the wide-narrow-wide structure that the traditional Residual Block usually follows with the number of channels. First, a 1 × 1 convolution is used to expand the input. Next, a 3 × 3 depth-wise convolution is used to drastically cut down on the number of parameters. Ultimately, the number of channels is reduced using a 1 × 1 convolution, allowing the input and output to be added. It overcomes the shortcomings of conventional positional embedding and the constraints of standard Vision Transformers to collect structured data and local relationships that are captured inside individual patches by conventional CNNs. The preferred network IRU, looks a lot like the residual block. The depth-wise convolution, projection layer, and expansion layer make up the inverted residual block. Nonetheless, the IRU has a special shortcut connection point that enhances its functionality. In mathematical notation, it is expressed as:

$$IRU\left(X\right) = Conv\left(\mathcal{F}\left(Conv(X)\right)\right) \tag{13}$$

$$\mathcal{F}\left(X\right) = DWConv\left(X\right) + X \tag{14}$$

Consider X to be an input tensor of size $X \in \mathbb{R}^{H\times W\times d}$, where d denotes the feature dimension and $H \times W$ denotes the resolution of the current stage input. The depth-wise convolution process is represented by the function DWConv(.).

IV. EXPERIMENTS

A. EXPERIMENTAL DETAILS

The public ADNI database provided the datasets that we used in our investigations. To improve the training data, we made random shifts of up to 2 pixels in three dimensions and randomly flipped images horizontally. To assist with network training, we initially trained the effective receptive field multi-axis attention network on the 2D brain sMRI scans. This study included two binary classification tasks to evaluate the effectiveness of the model: differentiating between HC and individuals with AD, as well as between HC and individuals MCI. In order to guarantee the accuracy of the findings, all models were trained using 10-Fold cross-validation. The data was split into ten parts, with nine segments utilized for training and one for testing during each cycle. For the ultimate assessment, the average and variability of every statistic were computed following ten training rounds. We look at the receiver operating characteristic's area under the curve, sensitivity, specificity, and classification accuracy to assess the system's performance. We compare the results by considering AD/HC subjects and MCI/HC subjects as positive and negative samples in the performance calculation.

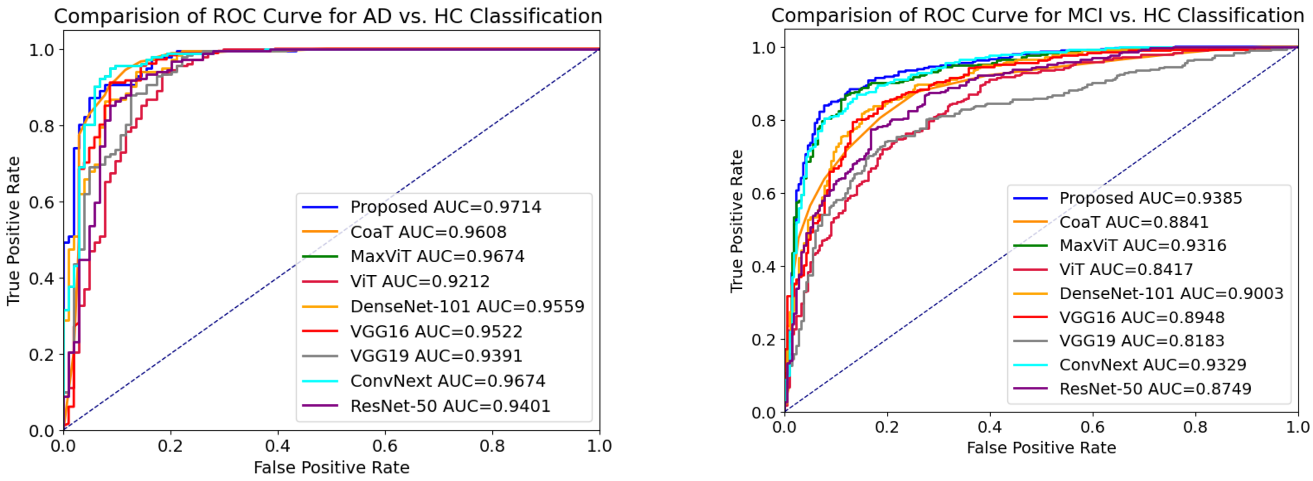
Our suggested approach was carried out utilizing Python 3.9.13 using the Keras library, which relies on Tensorflow 2.11.0. We employed AdamW as the optimizer, with a learning rate set at 0.0001 and weight decay of 0.01, using a batch size of 64. The ideal batch size for our models was determined by achieving a harmonious balance between batch sizes ranging from 8 to 128, incrementing in multiples of 2. When transitioning from larger to smaller batch sizes, it was observed that a batch size of 64 yielded the lowest error rate. Research suggests that using large batches often leads to convergence at sharp minima for both testing and training functions, which in turn hinders effective generalization. Conversely, small-batch methods tend to result in flatter minima. The utilization of cross-entropy loss helps ensure smooth network training and is aimed at addressing sample imbalance among various classes. At first, our research concentrated on evaluating how well the attention module worked in classifying diseases. Afterwards, we carried out a study to analyze the essential elements of our method. Finally, we conducted a comparison of our approach with previous research that utilized baseline sMRI scans from the ADNI database for the detection of AD-related disorders.

B. EFFECTIVENESS OF ATTENTION MODULE

The main goal of the initial trial is to assess how well the attention module works in the multi-axis attention convolution network in relation to classification accuracy. This attention module comprises MBConv with block attention and PConv with grid attention branches for the aggregation of multi-axis sMRI features. In this experiment, the outcomes are compared between models with baseline CNN and those with latest attentions-based models in Table 3. Specifically, the implementation of MBConv with block attentions and PConv with grids attentions involves retaining the respective field while eliminating the less informative sMRI features information within the attention module. Table 3 illustrates a comparison of ACC, SEN, SPE, and AUC findings for AD vs. HC and MCI vs. HC categorizations. Similarly, Figure 4, and Figure 5 present a comparison of AUC and Accuracy plot for baseline and hybridViT model with our finding respectively. The findings indicate that multi-axis attention with MBConv and PConv outperforms existing hybrid CNN-ViT attention in terms of Alzheimer's classification performance due to its ability to aggregate multi-scale features by integrating efficient receptive field within attention module in alternate fashion in different stages. Nonetheless, the combination of efficient receptive field, block and grid attentions in our multi-axis mixed attention can further enhance classification performance. To demonstrate, attention maps are created using the Rel-MSA layer of the fully connected block, showing the weights of multi-axis convolution kernels in various

TABLE 3. Results of binary classification (AD vs. HC and MCI vs. HC) using the ADNI dataset for multiple models.

Model			AD vs. HC				MCI vs. HC			
	Param#(M)	Flops(G)	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
VGG16[36]	134.2	15.4	88.75±5.96	93.03±3.61	82.97±2.98	95.22±5.98	85.55±4.80	69.17±3.05	81.51±2.83	89.48±3.84
VGG19[36]	139.5	19.6	88.71±4.63	94.05±7.22	84.89±1.50	93.91±6.87	86.99±2.57	91.61±4.35	91.19±1.05	81.83±3.40
DenseNet-101[37]	7.03	2.8	92.90±2.28	85.53±5.74	80.15±7.61	95.59±2.33	91.48±1.73	83.54±5.54	79.34±4.69	90.03±1.69
ResNet50[38]	23.5	3.8	89.14±6.39	92.44±4.26	76.81±3.24	94.01±1.46	87.87±2.60	90.60±7.51	95.10±7.15	87.49±2.36
ConvNext[39]	28.6	4.5	91.17±1.34	92.27±2.72	89.93±2.62	96.74±1.32	90.04±5.84	93.50±8.68	96.49±3.22	93.29±6.05
ViT[20]	50	22	86.51±3.02	84.88±5.77	87.15±5.18	92.12±2.93	81.68±5.25	83.06±2.26	91.56±4.24	84.17±4.03
CoaT[40]	5.5	4.33	93.62±3.09	94.41±3.72	89.75±7.04	96.08±3.53	91.60±1.37	87.21±2.57	94.57±4.35	88.41±1.48
MaxViT[34]	31	5.6	93.34±0.09	96.72±0.31	89.52±1.94	96.74±0.96	90.02±0.75	87.10±7.38	93.33±7.54	93.16±0.64
Proposed	18.3	2.23	97.29±1.44	95.96±1.21	96.15±3.12	97.14±1.54	94.79±2.86	93.57±6.75	97.85±2.22	93.85±5.34



a) b)

FIGURE 4. ROC curve for different classification model with our proposed model. A larger area denotes better performance: a) ROC curve for AD vs. HC and b) MCI vs. HC classification task respectively.

brain regions. The SoftMax function ensures that the total weight at each position is 1, with efficient receptive field created using MBConv and PConv kernels contributing more to feature aggregation. The attention maps are resized to fit the input image size using bilinear interpolation. To improve visualization, an average attention map for all test subjects is calculated and overlaid on the original sMRI image, as shown in Figure 6. It is evident that our methodology sharply focuses on atrophic brain regions in sMRI images and gives them priority over background noise. It has demonstrated exceptional performance in identifying these areas reliably in various slices. Our approach correctly locates and aligns atrophic regions in AD brains, whether viewed in coronal, sagittal, or axial perspectives. This demonstrates its ability to integrate scale-invariance and intrinsic localization from efficient convolution receptive field to Rel-MSA (block and grid attention) and capture high-quality sMRI features. Our strategy improves efficacy by significantly reducing background bias as compared to previous methods.

Efficient receptive field and multi-axis attention hold more weight around the brain regions with coarse structures like the parietal lobe, frontal lobe, temporal lobe, occipital lobe, and ventricle regions. However, baseline CNN and other hybrid-ViT show greater weights in the areas of the brain that have finer structures, like the amygdala and hippocampal regions. The recalibrated attention maps provide a fundamental illustration of the reasoning behind the attention mechanism in our proposed method.

C. EFFECTIVENESS OF EFFICIENT RECEPTIVE FIELD AND ENHANCED MULTI-AXIS ATTENTION

The second experiment aims to assess the efficiency of the arrangement of efficient receptive field and two attention module blocks in enhancing classification performance.

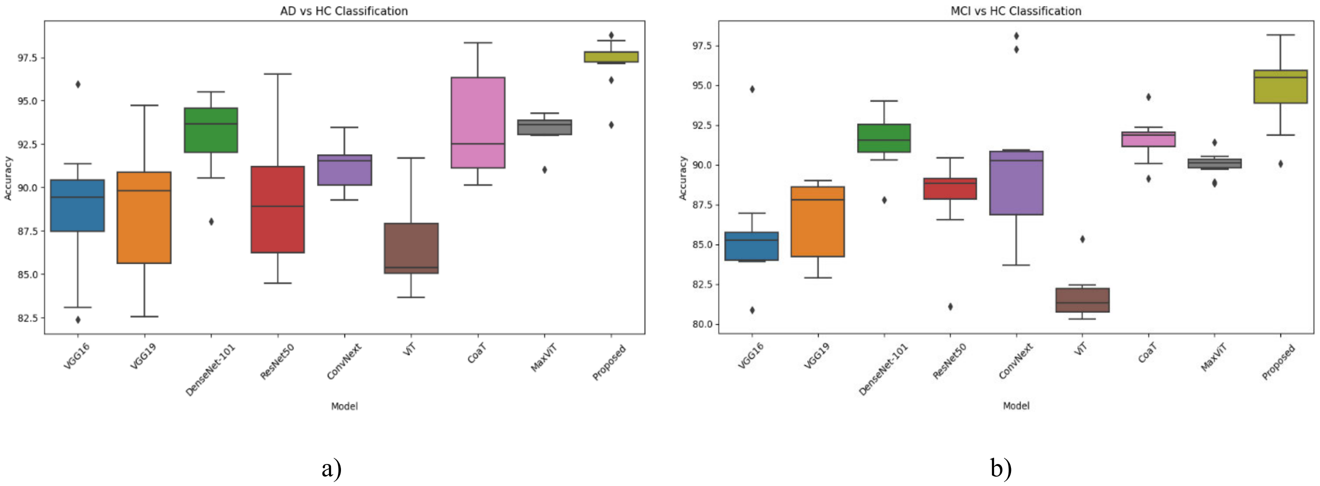


FIGURE 5. Box plot showing the average classification accuracy for 10-fold cross-validation results: a) AD vs. HC and b) MCI vs. HC classification accuracy respectively.

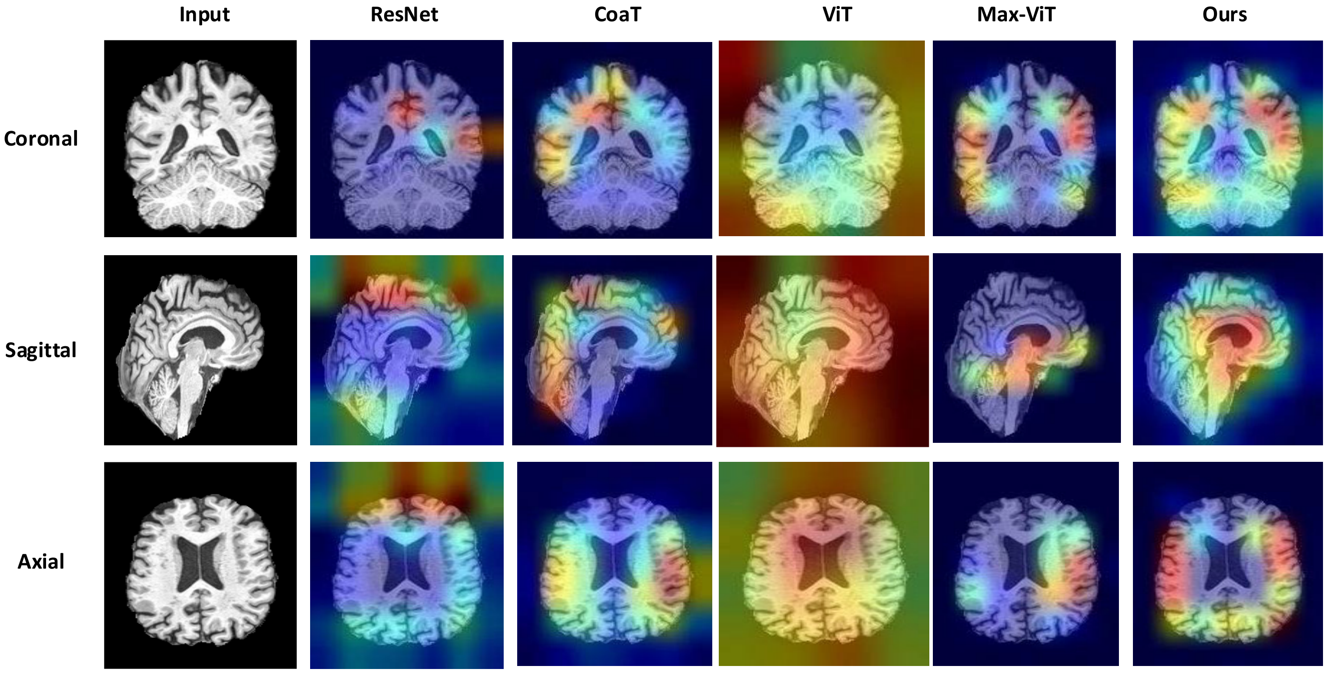


FIGURE 6. Brain maps showing the important brain region for Alzheimer's disease prediction using Grad-CAM technique with efficient receptive field and multi-axis attention modules.

The results of experiments conducted using block attention with MBConv and grid attention with PConv were compared with those obtained when MBConv and PConv were incorporated in both attentions within each block. Additionally, comparisons were made with a single attention module for each receptive field module separately. Table 4 shows the labels assigned to the proposed experimental multi-axis attention models: "MBConv + BlockAttn+GridAttn(S1,S2,S3,S4)", "PConv + BlockAttn + GridAttn(S1,S2,S3,S4)," "MBConv + BlockAttn(S1,S2,S3,S4)," "PConv + BlockAttn(S1,S2,S3,S4)," "MBConv + GridAttn(S1,S2,S3,S4)," "PConv + GridAttn (S1,S2,S3,S4)" and "Proposed(S1 = MBConv + BlockAttn, S2 = PConv + GridAttn, S3 = MBConv + BlockAttn, S4 = PConv + Gird Attn)" respectively. To show the effectiveness of direct utilization of image features learned by the multi-axis attention network for AD classification, we visualize the t-SNE features maps for last layer of each model in Figure 7. The outcomes presented in Table 4 and Figure 7 and Figure 8 demonstrate that the inclusion of efficient convolution receptive field, block and grid attention can marginally enhance classification results, while our suggested multi-axis attention with efficient receptive filed approach outperforms the others approaches. Therefore, capturing correlations

TABLE 4. Results of binary classification (AD vs. HC and MCI vs. HC) on different block arrangement.

Architecture Design	Param# (M)	Flops (G)	AD vs. HC				MCI vs. HC			
			ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
MBConv+Block Attn +Grid Attn(S1,S2,S3,S4)	30.86	5.6	93.79±0.03	95.78±0.45	91.71±1.71	97.41±0.87	90.71±0.89	90.03±6.41	94.12±6.11	93.95±0.87
PConv+Block Attn +Grid Attn(S1,S2,S3,S4)	27.03	17.85	95.12±2.75	94.27±5.72	95.94±3.57	95.1±2.8	91.76±1.65	86.95±2.77	94.09±1.9	93.1±1.82
MBConv+ Block Attn(S1,S2,S3,S4)	21.49	1.6	96.54±5.23	94.29±4.04	88.75±4.59	95.84±3.97	93.77±4.46	86.32±2.04	93.45±4.64	93.18±3.4
PConv+Block Attn(S1,S2,S3,S4)	16.46	10.82	94.3±3.4	93.24±5.29	96.78±1.78	93.51±2.13	91.65±0.69	92.27±2.81	90.95±2.56	92.61±0.64
MBConv+Grid Attn(S1,S2,S3,S4)	21.49	1.6	94.75±2.06	94.85±5.35	94.67±7.67	94.76±3.72	93.27±0.98	93.69±2.51	92.8±2.68	93.25±0.99
PConv+Grid Attn(S1,S2,S3,S4)	16.46	10.82	94.52±1.36	88.64±9.47	80.22±8.68	94.43±1.54	93.45±1.04	94.27±2.84	92.53±3.89	93.4±1.12
Proposed (S1=MBConv+Block Attn, S2=PConv+Grid Attn, S3=MBConv+Block Attn, S4=PConv+Gird Attn)	18.3	2.23	97.29±1.44	95.96±1.21	96.15±3.12	97.14±1.54	94.79±2.86	93.57±6.75	97.85 ± 2.22	93.85±5.34

among sMRI images at varying Alzheimer's stage with suggested approach proves beneficial for diagnostic tasks. Furthermore, integration of MBConv, PConv, block and grid attention in efficient setting plays a crucial role in Alzheimer's detection. In comparison to baseline CNN and hybridViT, the models are lighter and faster due to the large reduction in model parameters and computational cost achieved by alternating the integration of MBConv and PConv inside two attention modules. To enhance comprehension, we conducted a comparison of t-SNE features space between the feature vectors of AD vs. HC group at last layers of each block setting, as illustrated in Figure 7. Features matrices are created by arranging the feature vectors in each group test sample. It is evident that the integration of efficient receptive field, block attention and grid attention not only enhance feature disparity among subjects of two classification group but also boosts the ability to differentiate features between different classes with computationally efficient way. Furthermore, Figures 8 illustrate the superiority of our methods in terms of accuracy compared to different block settings.

D. ABLATION STUDY

The third experiment aims to perform an ablation study on the window size of attention block. We experimented with finding the settings of four efficient receptive fields with four hierarchical multi-axis attention blocks in the transformer encoder module. Table 5 shows that Head denotes the number of self-attention heads in the window, and W is the size of the attention window in each Transformer block. The model applies Rel-MSA to both block and grid attention without discrimination for window. Because larger window widths, such as w = 16 have difficulty connecting grid features in sMRI images, the model performs best when w = 8 in our multi-axis attention setting in combination with efficient receptive field modules. On the other hand, wider windows, such as w = 32, make it more difficult for both attention blocks to concentrate on local lesion features, which produces poor classification outcomes.

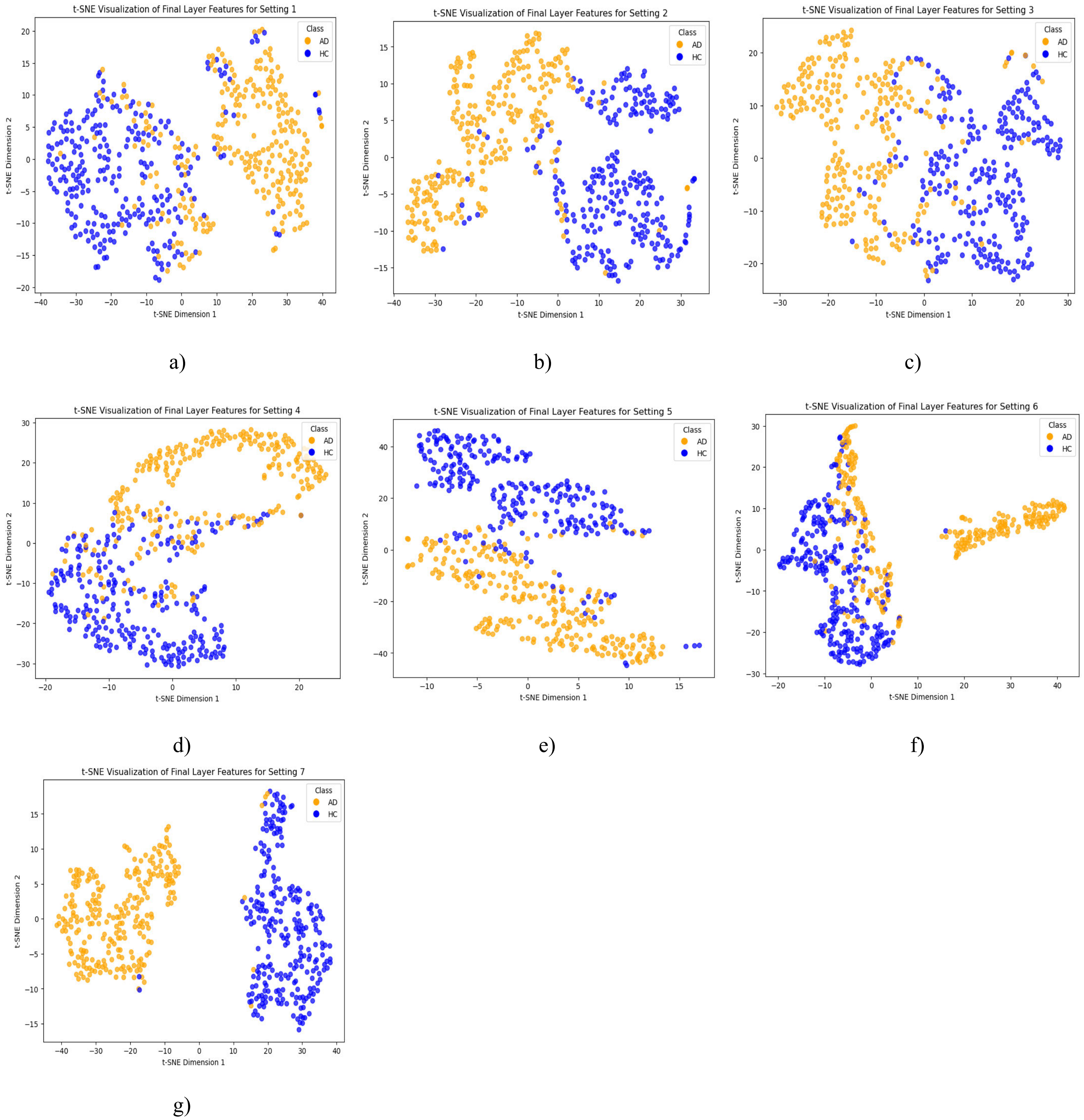
The findings reveal that window size 8 leads to a 2.47% and 1.32% increase in ACC and AUC for AD vs. HC classification, as well as more than a 1% and 2% enhancement in ACC and AUC for MCI vs. HC classification respectively. Furthermore, the results Table indicates that w = 8 performs better in ROC curve for classification than w = 32. It could be more advantageous to include this window size in the low-level sMRI feature map as well. Moreover, the proposed efficient receptive field multi-axis attention module demonstrates the most superior outcomes for AD and MCI diagnosis, highlighting the advantages of modeling correlations and discrimination between sMRI features vectors of different groups in AD-related diagnosis.

E. COMPARISON WITH RECENT STATE-OF-THE-ART METHODS

In this part, we are evaluating our suggested method against other modern techniques that have used deep neural networks and hybridViT model on the sMRI data from the ANDI database. One of these techniques involved implementing a 3D convolution network for diagnosing Alzheimer's disease and investigating biomarkers [41]. A 3D CNN+RNN mechanism was used in another technique for the diagnosis and interpretability analysis of AD [42].

A Conv-Swinformer network was proposed in another study to gradually learn and merge local and global features of sMRI data [27]. A comprehensible deep learning model was proposed to predict the probability of a class by utilizing randomly selected patches, followed by making predictions using a multilayer perceptron [43]. A deep network based on the attention framework name as Addformer was proposed for predicting AD using sMRI slices [44].

The outcomes demonstrate the superiority of the attention-based model in capturing global context, since the



a)

b)

c)

d)

e)

f)

g)

FIGURE 7. Performance illustration of different block setting using t-SNE features visualization for last layer of each model showing the better clustering of proposed setting. a) Setting 1 (MBConv+Block Attn +Grid Attn(S1,S2,S3,S4); b) Setting 2 (PConv+Block Attn +Grid Attn(S1,S2,S3,S4); c) Setting 3 MBConv+ Block Attn(S1,S2,S3,S4); d) Setting 4 PConv+Block Attn(S1,S2,S3,S4); e) Setting 5 (MBConv+Grid Attn(S1,S2,S3,S4); f) Setting 6 PConv+Grid Attn(S1,S2,S3,S4); g) Setting 7: proposed (S1=MBConv+Block Attn, S2=PConv+Grid Attn, S3=MBConv+Block Attn, S4=PConv+Gird Attn) methods.

3D CNN model and Recurrent Attention model perform less better than the Transformer-based approach. The proposed efficient receptive field with multi-axis attention performs better in the Transformer-based method than the pure Attention Transformer, suggesting that the block and grid attention with MBConv and PConv mechanism is superior for classifying AD for a small sMRI dataset. Furthermore, the proposed automated diagnostic process has produced competitive diagnostic outcomes when compared to some ADNI-based research. This study examines the collective impacts of various methods, such as experimental planning, data preprocessing, data expansion, and classification algorithms.

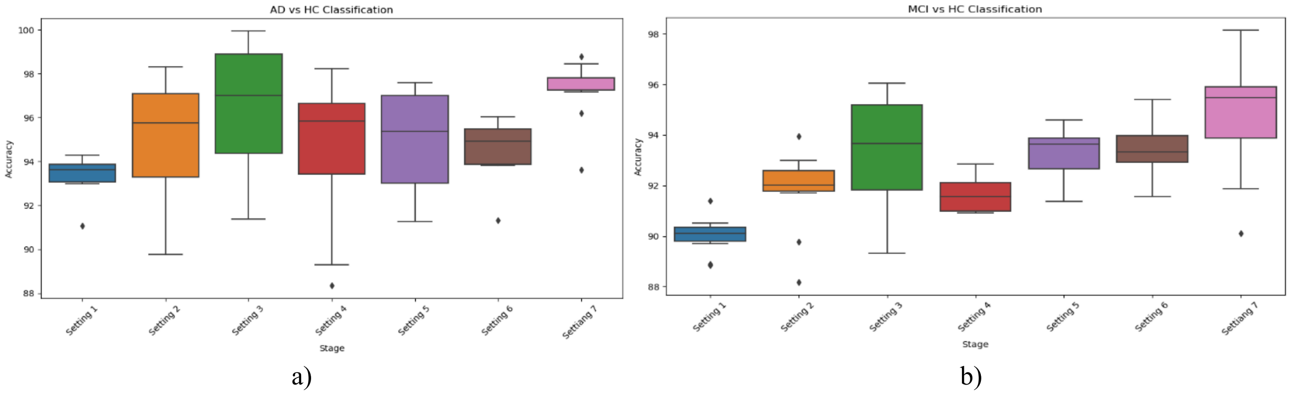


FIGURE 8. Box plot showing the average classification accuracy for 10-fold cross-validation results: a) AD vs. HC and b) MCI vs. HC classification accuracy for different block setting respectively.

To ensure a fair comparison, we compared our technique with ADNI base sMRI image with comparable numbers of data sample. The outcomes in terms of classification results such as ACC, SEN, and SPE for AD and MCI diagnosis are shown in Table 6. Our approach significantly outperformed the other methods in both classification tasks. In comparison to other techniques, our approach employs a highly effective receptive field and multi-axis attention transformer to capture flexible multi-scale characteristics and enhanced long-range features specific to groups.

V. DISCUSSION

In this part of the report, we will first examine the main differences between our proposed method and previous research. Our proposed deep learning technique can simultaneously acquire features and diagnose diseases without the need for complicated image processing like ROI segmentation and rigid registration, unlike traditional methods. Our approach differs in significant ways from existing deep learning methods. To begin with, unlike current CNN-based methods that employ the same convolution across the entire brain, we introduce a MBConv and PConv attention to combine feature maps learned using various convolution module, enabling the capture of diverse morphological changes in local brain regions by ensuring the efficient computational cost. Furthermore, we include a block and grid attention mechanism to capture long-distance connections in more advanced feature maps using Rel-attention, overcoming the limitations of convolution as a localized operation. Lastly, rather than using MLP as like original transformer we introduce IRFFN to ensure fine sMRI features representation, we propose IRFFN network to integrate image features and uncover latent relationships among different groups for more precise diagnostic outcomes.

Interpreting the deep network in addition to disease classification is crucial in brain image analysis. To aid in interpretation, Grad-CAM [46] technique is utilized for AD vs. HC patient in the testing set using a specific method. Grad-CAM maps are generated by calculating the gradients of classification scores to determine the significance of individual voxels in disease classification. These maps are normalized and do not consider negative gradients. The mean class activated map is computed to pinpoint areas of pathology, and it is superimposed on a reference image to emphasize the brain regions most linked to AD, demonstrated in Figure 6. The important areas of the brain that have been recognized include the amygdala, hippocampus, parahippocampal gyrus, superior temporal gyrus, and ventricle. These areas are recognized for their important involvement in memory and cognitive functions, and shrinking in these regions has been associated with Alzheimer's disease. Additionally, ventricular enlargement due to adjacent brain atrophy is a crucial biomarker for measuring AD progression. The atrophic brain regions identified for AD diagnosis are generally consistent, although there are some differences. One example is that the diagnosis of Alzheimer's is more accurate when there is atrophy in the hippocampal regions due to its important role that hippocampal atrophy plays in the progression of AD.

In our study, we trained efficient receptive field and multi-axis attention deep networks using ADNI dataset. The ADNI data was divided into 10 parts for the experiments, with 9 parts being utilized for training the deep network and 1 part for testing purposes. The classification results showed an accuracy of 97.29% and an AUC of 97.14% for AD and an accuracy of 94.79% and an AUC of 93.85% for MCI diagnosis respectively.

VI. LIMITATIONS

Even though the system we created has demonstrated encouraging outcomes in tasks pertaining to pathology localization and diagnosing AD, there is still room to improve the system's functionality and usefulness by tackling the following issues. Although variations in brain atrophy may manifest at different scales in different slices, in our current system, the size of diseased patches stays stable and uniform across

TABLE 5. Results of binary classification (AD vs. HC and MCI vs. HC) for different window size with attention head.

Window Size (W)	Head	AD vs. HC				MCI vs. HC			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
8	32	97.29±1.44	95.96±1.21	96.15±3.12	97.14±1.54	94.79±2.86	93.57±6.75	97.85±2.22	93.85±5.34
16	32	94.82±3.13	94.60±0.04	95.03±0.06	95.82±0.07	93.27±0.98	93.69±0.02	93.80±0.02	93.25±0.09
32	32	93.70±2.52	94.32±0.05	96.23±0.04	94.77±0.03	88.73±2.37	87.28±0.15	90.25±0.12	91.76±0.02

TABLE 6. Comparing the performance of binary classification on the ADNI dataset using the most recent CNN and hybridViT techniques.

Reference	Methods	Modality	Subjects (AD/MCI/HC)	AD/HC			HC/MCI		
				ACC	SEN	SPE	ACC	SEN	SPE
Jin et al.[41]	Attention+ResNet	MRI	227/-/305	90.6	89.4	91.5			
Altay et al. [42]	3D CNN	MRI	493/-/605	85.23	86.49	83.08	76.95	81.58	73.24
Zhu et al. [45]	distilling-ViT	MRI	313/319/324	97.97	97.94	98.17	91.89	90.66	93.01
Xin et al. [28]	Conv-Swin Net	MRI	336/-/529	93.9	92.5	94.7			
Hu et al. [27]	Conv-Swinformer	MRI(Axial-Slice)	508/1412/970	93.56	93.81	97.49	79.07	79.82	78.17
Kushol et al. [44]	Addaformer	MRI	159/-/229	88.2	95.6	77.4			
Proposed Method	Block+Grid Attn	MRI	315/370/390	97.29	95.96	96.15	94.79	93.57	97.85

multiple portions of sMRI images. Going forward, utilizing 3D sMRI images could be one way to make our architecture more flexible. The current technique, which uses 2D slices, may introduce false positive samples, contaminating the data representation and affecting diagnosis accuracy. Confirming the model's generalizability also requires evaluating its performance on a larger variety of datasets. Although our assessment is currently predicated on a particular ADNI dataset, it is imperative to confirm that the model operates uniformly across datasets with varying attributes. By offering a more complete collection of information, the incorporation of multimodal imaging data could improve classification performance as opposed to depending only on MRI data. Future studies will therefore investigate the integration of multimodal brain data, including fMRI, PET scans, and clinical characteristics.

VII. CONCLUSION

We presented a highly effective model in this study that utilizes the self-attention mechanism to classify MRI data associated with Alzheimer's disease. Using a combination of mobile convolution, partial convolution, and multi-axis (block and grid) attention mechanism in an alternating manner, we successfully decreased computational complexity while maintaining effectiveness. This enabled the application of attention to high-dimensional sMRI data. Moreover, our design includes a layer with an inverted residual unit that conducts feature down sampling, preserving critical features while reducing computational expenses. The optimized architecture proposed in this study achieved outstanding classification results on the ADNI dataset, outperforming other cutting-edge methods. Using brain sMRI data, a number of comparative investigations between the ViT model and the baseline CNN further illustrated the effective learning capabilities of the suggested architecture. This work offers fresh perspectives and methods for applying deep learning to the investigation of brain disorders. Despite encouraging results in pathology localization and AD diagnosis, our system can be enhanced by addressing issues such as the uniform size of diseased patches across sMRI slices, improving flexibility with 3D sMRI images, reducing false positives from 2D slices, ensuring model generalizability across diverse datasets, and integrating multimodal imaging data, including fMRI, PET scans, and clinical characteristics for better classification performance.

ACKNOWLEDGMENT

The design and implementation of Alzheimer's Disease Neuroimaging Initiative (ADNI) involved the contribution of ADNI researchers and the provision of data. However, they were not directly involved in the analysis or writing of this report. A comprehensive list of ADNI investigators can be accessed at (http://adni.loni.usc.edu/wp-content/uploads/how-to-apply/ADNI Acknowledgement List.pdf). ADNI receives support from various sources, including the National Institute on Aging, the National Institute of

Biomedical Imaging and Bioengineering, and several generous donors, such as AbbVie, Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Araclon Biotech, BioClinica Inc., Biogen, Bristol-Myers Squibb Company, CereSpir Inc., Cogstate, Eisai Inc., Elan Pharmaceuticals Inc., Eli Lilly and Company, EuroImmun, F. Hoffmann-La Roche Ltd., its affiliate Genentech Inc., Fujirebio, and GE Healthcare. The Foundation for the National Institutes of Health facilitates contributions from the private sector. The grantee of ADNI is the Northern California Institute for Research and Education and the study Coordinator is the Alzheimer's Therapeutic Research Institute, College of Southern California. The ADNI data is disseminated by the Laboratory for Neuro Imaging, College of Southern California.

REFERENCES

[1] *The Neuropathological Diagnosis of Alzheimer's Disease | Molecular Neurodegeneration*. Accessed: May 4, 2024. [Online]. Available: https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-019-0333-5

[2] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dementia*, vol. 7, no. 3, pp. 263–269, May 2011, doi: 10.1016/j.jalz.2011.03.005.

[3] "2023 Alzheimer's disease facts and figures," *Alzheimers Dement.*, Mar. 2023, Art. no. alz.13016, doi: 10.1002/alz.13016. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13016

[4] K. Blennow, M. J. de Leon, and H. Zetterberg, "Alzheimer's disease," *Lancet*, vol. 368, no. 9533, pp. 387–403, Jul. 2006, doi: 10.1016/s0140-6736(06)69113-7.

[5] E. J. Mufson, L. Binder, S. E. Counts, S. T. DeKosky, L. de Toledo-Morrell, S. D. Ginsberg, M. D. Ikonomovic, S. E. Perez, and S. W. Scheff, "Mild cognitive impairment: Pathology and mechanisms," *Acta Neuropathologica*, vol. 123, no. 1, pp. 13–30, Jan. 2012, doi: 10.1007/s00401-011-0884-1.

[6] *Mild Cognitive Impairment: Clinical Characterization and Outcome | Dementia and Cognitive Impairment | JAMA Neurology | JAMA Network*. Accessed: Mar. 28, 2024. [Online]. Available: https://jamanetwork.com/journals/jamaneurology/article-abstract/774828

[7] R. J. Bateman, C. Xiong, T. L. S. Benzinger, A. M. Fagan, A. Goate, N. C. Fox, D. S. Marcus, N. J. Cairns, X. Xie, T. M. Blazey, and D. M. Holtzman, "Clinical and biomarker changes in dominantly inherited Alzheimer's disease," *New England J. Med.*, vol. 367, no. 9, pp. 795–804, Aug. 2012, doi: 10.1056/nejmoa1202753.

[8] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, J. C. Morris, R. C. Petersen, A. J. Saykin, M. E. Schmidt, L. Shaw, L. Shen, J. A. Siuciak, H. Soares, A. W. Toga, and J. Q. Trojanowski, "The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception," *Alzheimer's Dementia*, vol. 9, no. 5, pp. 111–194, Sep. 2013, doi: 10.1016/j.jalz.2013.05.1769.

[9] A. Kumar and A. Singh, "A review on Alzheimer's disease pathophysiology and its management: An update," *Pharmacological Rep.*, vol. 67, no. 2, pp. 195–203, Apr. 2015, doi: 10.1016/j.pharep.2014.09.004.

[10] P. Vemuri and C. R. Jack, "Role of structural MRI in Alzheimer's disease," *Alzheimer's Res. Therapy*, vol. 2, no. 4, p. 23, Aug. 2010, doi: 10.1186/alzrt47.

[11] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, no. 11, p. 226, Oct. 2018, doi: 10.1007/s10916-018-1088-1.

[12] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[13] B. Lee, W. Ellahi, and J. Y. Choi, "Using deep CNN with data permutation scheme for classification of Alzheimer's disease in structural magnetic resonance imaging (sMRI)," *IEICE Trans. Inf. Syst.*, vol. 102, no. 7, pp. 1384–1395, Jul. 2019.

[14] U. Khatri and G.-R. Kwon, "Multi-biomarkers-base Alzheimer's disease classification," *J. Multimedia Inf. Syst.*, vol. 8, no. 4, pp. 233–242, Dec. 2021, doi: 10.33851/jmis.2021.8.4.233.

[15] S. W. Park, N. Y. Yeo, Y. Kim, G. Byeon, and J.-W. Jang, "Deep learning application for the classification of Alzheimer's disease using 18F-flortaucipir (AV-1451) tau positron emission tomography," *Sci. Rep.*, vol. 13, no. 1, May 2023, Art. no. 1, doi: 10.1038/s41598-023-35389-w.

[16] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

[17] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps: Automation of Decision Making* (Lecture Notes in Computational Vision and Biomechanics), N. Dey, A. S. Ashour, and S. Borra, Eds., Cham, Switzerland: Springer, 2018, pp. 323–350, doi: 10.1007/978-3-319-65981-7_12.

[18] Y. Kinoshita and H. Kiya, "Convolutional neural networks considering local and global features for image enhancement," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2110–2114, doi: 10.1109/ICIP.2019.8803194.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst* Curran Associates, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[21] R. Azad, A. Kazerouni, M. Heidari, E. Khodapanah Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," 2023, *arXiv:2301.03505*.

[22] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[23] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" 2021, *arXiv:2108.08810*.

[24] G. Hcini, I. Jdey, and H. Dhahri, "Investigating deep learning for early detection and decision-making in Alzheimer's disease: A comprehensive review," *Neural Process. Lett.*, vol. 56, no. 3, p. 153, Apr. 2024, doi: 10.1007/s11063-024-11600-5.

[25] 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 Aging Neurosci.*, vol. 15, Apr. 2023, Art. no. 1102869.

[26] H. Shin, S. Jeon, Y. Seol, S. Kim, and D. Kang, "Vision transformer approach for classification of Alzheimer's disease using 18F-florbetaben brain images," *Appl. Sci.*, vol. 13, no. 6, p. 3453, Mar. 2023, doi: 10.3390/app13063453.

[27] Z. Hu, Y. Li, Z. Wang, S. Zhang, and W. Hou, "Conv-swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification," *Comput. Biol. Med.*, vol. 164, Sep. 2023, Art. no. 107304, doi: 10.1016/j.compbiomed.2023.107304.

[28] J. Xin, A. Wang, R. Guo, W. Liu, and X. Tang, "CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105189, doi: 10.1016/j.bspc.2023.105189.

[29] C. Chen, H. Wang, Y. Chen, Z. Yin, X. Yang, H. Ning, Q. Zhang, W. Li, R. Xiao, and J. Zhao, "Understanding the brain with attention: A survey of transformers in brain sciences," *Brain-X*, vol. 1, no. 3, p. e29, Sep. 2023, doi: 10.1002/brx2.29.

[30] M. Jiang, B. Yan, Y. Li, J. Zhang, T. Li, and W. Ke, "Image classification of Alzheimer's disease based on external-attention mechanism and fully convolutional network," *Brain Sci.*, vol. 12, no. 3, p. 319, Feb. 2022, doi: 10.3390/brainsci12030319.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[32] *SPM—Statistical Parametric Mapping*. Accessed: Jan. 11, 2023. [Online]. Available: https://www.fil.ion.ucl.ac.uk/spm/

[33] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher FLOPS for faster neural networks," 2023, *arXiv:2303.03667*.

[34] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," 2022, *arXiv:2204.01697*.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.

[40] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," 2021, *arXiv:2104.06399*.

[41] D. Jin, J. Xu, K. Zhao, F. Hu, Z. Yang, B. Liu, T. Jiang, and Y. Liu, "Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1047–1051, doi: 10.1109/ISBI.2019.8759455.

[42] F. Altay, G. R. Sánchez, Y. James, S. V. Faraone, S. Velipasalar, and A. Salekin, "Preclinical stage Alzheimer's disease detection using magnetic resonance image scans," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 17, pp. 15088–15097, doi: 10.1609/aaai.v35i17.17772.

[43] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, and M. Kaku, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," *Brain*, vol. 143, no. 6, pp. 1920–1933, Jun. 2020, doi: 10.1093/brain/awaa137.

[44] R. Kushol, A. Masoumzadeh, D. Huo, S. Kalra, and Y.-H. Yang, "Addformer: Alzheimer's disease detection from structural mri using fusion transformer," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5, doi: 10.1109/isbi52829.2022.9761421.

[45] J. Zhu, Y. Tan, R. Lin, J. Miao, X. Fan, Y. Zhu, P. Liang, J. Gong, and H. He, "Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105737, doi: 10.1016/j.compbiomed.2022.105737.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

UTTAM KHATRI received the B.Eng. degree in electronics and communication engineering from Pokhara University (Nepal Engineering College), Nepal, in 2015. In 2015 and 2018, he was a Junior Professor with Nepal Engineering College. Currently, he is a Research Scholar with Chosun University, Gwangju, Republic of Korea. His research interests include artificial neural networks, artificial intelligence systems, and machine learning on image processing, especially in medical image processing.

JUN-HYUNG KIM received the B.S. and Ph.D. degrees in electronics engineering from Korea University, Seoul, Republic of Korea, in 2006 and 2012, respectively. He was with the Agency for Defense Development, Daejeon, Republic of Korea, from 2012 to 2021, as a Senior Researcher, working on developing algorithms for electro-optical/infrared sensors. Then, he spent a half year with the Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju, Republic of Korea, as a Postdoctoral Researcher. Since 2023, he has been an Assistant Professor with Chosun University. His research interests include image processing, deep learning, and their applications in target detection and image understanding.

GOO-RAK KWON (Senior Member, IEEE) received the Ph.D. degree from the Department of Mechatronic Engineering, Korea University, in 2007. He was the Chief Executive Officer and the Director of Dalitech Company Ltd., from 2004 to 2007. He joined the Department of Electronic Engineering, Korea University, from 2007 to 2008, where he was a Postdoctoral Researcher supporting the BK21 Information Technique Business. He has been a Professor with Chosun University, since 2017. His research interests include medical image analysis, A/V signal processing, and video communication and applications. He was a Life Member of the Signal Processing Society in the IEIE, KMMS, KIPS, KICS, KING, and KISM, as a Senior Member.