Computers in Biology and Medicine 171 (2024) 108116

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Diagnosis of Alzheimer's disease via optimized lightweight convolution-attention and structural MRI

Uttam Khatri [a], Goo-Rak Kwon [a,*]

[a] Dept. of Information and Communication Engineering, Chosun University, 309 Pilmun-Daero, Dong-Gu, Gwangju, 61452, Republic of Korea

ARTICLE INFO

Keywords:
Alzheimer's disease
Magnetic resonance imaging
Convolution neural network
Vision Transformer
Gradient centralization

ABSTRACT

Alzheimer's disease (AD) poses a substantial public health challenge, demanding accurate screening and diagnosis. Identifying AD in its early stages, including mild cognitive impairment (MCI) and healthy control (HC), is crucial given the global aging population. Structural magnetic resonance imaging (sMRI) is essential for understanding the brain's structural changes due to atrophy. While current deep learning networks overlook voxel long-term dependencies, vision transformers (ViT) excel at recognizing such dependencies in images, making them valuable in AD diagnosis. Our proposed method integrates convolution-attention mechanisms in transformer-based classifiers for AD brain datasets, enhancing performance without excessive computing resources. Replacing multi-head attention with lightweight multi-head self-attention (LMHSA), employing inverted residual (IRU) blocks, and introducing local feed-forward networks (LFFN) yields exceptional results. Training on AD datasets with a gradient-centralized optimizer and Adam achieves an impressive accuracy rate of 94.31% for multi-class classification, rising to 95.37% for binary classification (AD vs. HC) and 92.15% for HC vs. MCI. These outcomes surpass existing AD diagnosis approaches, showcasing the model's efficacy. Identifying key brain regions aids future clinical solutions for AD and neurodegenerative diseases. However, this study focused exclusively on the AD Neuroimaging Initiative (ADNI) cohort, emphasizing the need for a more robust, generalizable approach incorporating diverse databases beyond ADNI in future research.

1. Introduction

Alzheimer's disease (AD) is one of the more prevalent kinds of dementia, accounting for an estimated 60–70% of all cases [1]. According to data from the World Alzheimer's Survey, 78 million cases of Alzheimer's disease are expected to occur by 2030, with an estimated 55 million individuals suffering from the medical condition [2]. In addition to the terrible human effects AD has on those who have it and those who care for them, the disease carries enormous financial expenses. By 2023, it is anticipated that AD-related costs would reach $345 billion in the USA alone [3]. However, this is only the very tip of the iceberg. Our current situation has been identified as an AD epidemic [4], and as a result of the population's age distribution being skewed towards a higher number of elderly individuals, associated expenditures are anticipated to exceed threefold by the year 2050 [3]. The condition can be identified by a subtly progressive deterioration in cognitive, behavioral, and visuospatial abilities that are brought on by neurodegenerative disorders [5]. Monitoring symptoms aids in an initial AD diagnosis, yet efforts are underway to develop a technique for identifying precise biomarkers in cerebrospinal fluid (CSF) to enhance diagnostic accuracy [6]. However, this method is intrusive and poses potential harm to the patient. Moreover, modern imaging methods like positron emission tomography (PET) and structural magnetic resonance imaging (sMRI) can be utilized to detect molecular and structural AD-related biomarkers [7]. The structural changes to the brain caused by AD can be understood and evaluated using sMRI, which is a non-invasive and effective technology. They are considered essential in clinical practice and contribute significantly to the diagnosis of AD pathology [8–10]. Focus has lately been drawn to neural networks and deep learning techniques to automatically identify AD and other brain disorders using sMRI data [11]. Convolutional neural networks (CNNs), which have demonstrated outstanding ability in visual recognition, have been employed in the past to identify AD using sMRI [12–15]. Although convolutional processes improve their capacity for local knowledge transfer, they are not well adapted to simulate long-distance correlations. In the realm of natural language processing (NLP), a model by the name of Attention gained popularity as CNNs were being developed [16]. Utilizing the self-attention approach replicates the global context better than stacking hierarchical

* Corresponding author.
E-mail address: grkwon@chosun.ac.kr (G.-R. Kwon).

https://doi.org/10.1016/j.compbiomed.2024.108116
Received 30 September 2023; Received in revised form 28 January 2024; Accepted 4 February 2024
Available online 8 February 2024

convolution layers [8]. Several Attention-based Transformer techniques have been presented and have outperformed CNN-based techniques in a range of vision tasks including image classification [17–19], object recognition [20], and semantic segmentation [21]. Swin-Transformer [22] extends vision transformer (ViT) with a local self-attention mechanism to simplify calculations, while Transformer-iN-Transformer (TNT) [23] incorporates both local and global aspects of an image to increase classification efficiency. However, ViT application in the automatic classification of brain disorders is limited and using it directly on sMRI data would result in a significant computing overhead due to the intricate computations [8,24]. In the research [25]; Transformer was combined with CNN-based architectures to boost the performance accuracy. Transformer has been challenging to employ directly since most brain sMRI data sources are small in comparison to established natural image data sources. Altay et al. [26] employed CNNs and transformers to successfully find preclinical AD. Using a 2D CNN, they first retrieved attributes using numerous 2D slices of an sMRI scan before combining data from all the 2D image features with a Transformer for disease identification. Jun et al. [27] split 3D sMRI directly into 2D slices across the three distinct sections to feed deep models consisting of a transformer and a CNN encoder.

With the use of T1-weighted sMRI data and inspiration from convolutional attention [25], the goal of this research is to examine the Transformer's versatility in AD diagnosis tasks. To increase its effectiveness and versatility by applying spatial linkages, we present optimized convolution ViT (CViT) by stacking inverted residual block and sandglass local feed-forward networks inside the proposed architecture, an efficient design that takes advantage of CNNs and Transformers topologies for AD diagnosis. First, we accurately extracted 2D-slices from 3D sMRI images and a feature extraction method is developed to convert 2D sMRI data into a feature matrix using convolutional stem while considering 2D-tensor data. Then, four layers of LMHSA are used in the suggested framework to improve operational efficiency while lowering the quadratic complexity of the original self-attention mechanism [25]. To provide a comprehensive feature representation, the method incorporates depth wise convolution inside attention block which exhibits characteristics of high informativeness. Similarly, we redefined the input channel of the proposed architecture which helped to reduce computing complexity and simultaneous assurance of high performance. Later, we utilized the sandglass LFFN layer with more depth-wise convolution, a module that makes the process faster and more stable, and the AdamW optimizer with Gradient Centralization (AdamWGC) [28] for a more efficient and stable training process. Lastly, the framework's output was sent to the classifier for disease classification. We used the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to validate the suggested framework, which showed that the method was superior in terms of algorithm performance and a variety of medical criteria like accuracy, specificity, sensitivity, and precision. Suggested methods drastically reduce the number of parameters and processing expenses for brain sMRI. The following is a summary of the major contributions made by this study.

1. The strategy aims to improve the learning of features and better combine local and global information aspects of sMRI by using an IRU and sandglass LFFN with attention model.
2. Alzheimer's detection is used as the problem statement. The Gradient centralization technique was introduced into the AdamW optimizer to train the proposed model faster and consistently.
3. The model is compact and efficient which delivers state-of-the-art performance with accuracy of 94.31% for multi-class classification and 95.37% for AD/HC binary classification on the ADNI dataset while using fewer FLOPs and parameters.

2. Related works

2.1. Self-attention and vision Transformer

sMRI serves as a significant biomarker in the context of Alzheimer's disease, offering a pivotal diagnostic tool for healthcare professionals. Healthcare professionals utilize sMRI scans to evaluate the degree of brain atrophy in patients with AD, thus enabling the determination of the disease's progression stage. However, healthcare professionals encounter challenges in processing the extensive and intricate sMRI images with precision. Consequently, there has been a surge of interest in the exploration of computer-assisted AD diagnosis based on sMRI images, aiming to enhance the diagnostic efficacy of healthcare professionals. Over the past few decades, conventional machine learning methods [29,30] which involve manual extraction of features primarily using support vector machines, have been extensively employed in the diagnosis of Alzheimer's, and have even exhibited diagnostic efficacy surpassing that of the most healthcare professionals. CNNs [31] also have made a significant contribution to computer vision during the past few years due to their capacity to extract highly distinctive features. CNNs have shown promise in classifying healthy and AD brains, leading to their widespread use in predicting different stages of AD [30,32,33]. Various CNN-based algorithms, such as RNN [34], GCN [35], and Transfer Learning, have also been applied [36]. However, these methods are only focused on locale information of brain images. With the rise of deep learning recent studies on AD shifted towards complex deep learning algorithms focusing on accurate predictions for early AD diagnosis to assist physicians. Recently, transformer-based network structures, such as the ViT model, have been introduced and shown outstanding performance in computer vision tasks [17,18,22]. However, the direct use of ViT on brain images possesses the challenge of computational expense primarily due to hardware limitations. To integrate benefits of ViT by reducing the computational burden some researchers have experimented with transformer-based networks in medical image analysis, including AD diagnosis, and found that they can achieve equal or better performance than CNN models like Resnet [8,37,38]. However, the use of large datasets may not accurately represent the performance of transformers on small datasets in AD classification. Furthermore, the advancement of deep learning has led to significant advancements in merging convolution and attention mechanism in computer vision tasks [25,39]. These models were a new architecture that mixes the mechanism of self-attention with that of convolution and introduced some performance optimizations with lesser FLOPs than original ViTs. Few ViT networks have been proposed and have achieved impressive results in AD diagnosis tasks [8,38,40,41]. For instance, J. Zhu et al. [8] introduced efficient self-attention with structural distilling mechanism using sagittal sMRI in 2023, and later another study Hoang et al. [42] proposed transfer-learning ViT method using sagittal sMRI in 2023. Recently, Manzari et al. proposed MedViT [43] for medical images combing convolution and attention. In other studies, conducted in 2022 Kushol et al. [44] effectively employed the utilization of frequency an image domain features extracted from coronal 2D slices within a ViT architecture to achieve state-of-art performance in AD classification. Their work demonstrates the capacity of two transformers, one from the image domain and the other from the frequency domain, to capture global and local context as well as spatial features. These contributions have greatly influenced the field of medical image analysis and have contributed to the progress of ViT learning. While these architectures have exhibited remarkable performance, they still lack addressing computational burden of ViT. However, for the analysis of medical images, which may involve significantly larger inputs, a more computationally efficient approach is imperative. Additionally, convolution-attention offers a solution to the limited data problem in the medical imaging field, providing new inspiration for further advancements. In this paper, instead of just comparing the adversarial robustness of CNNs and ViTs without considering architecture design,

we go a step further and propose a robust optimized hybrid architecture. We incorporate a novel optimization technique into the Transformer architecture to enhance the resilience of Transformer models with reduced computing resources and auxiliary training for AD diagnosis.

2.2. Depth-wise convolution

The process of regular convolution and depth-wise convolution is illustrated in the diagram presented in Fig. 1 above. Unlike traditional convolution, which operates on all input channels independently, depth-wise convolution performs separate convolutions for each input channel. A single filter is used in depth-wise convolution to process each input channel separately, producing an output channel set that is equal to the number of input channels. This approach reduces the computational cost significantly compared to standard convolution, as it reduces the number of parameters and operations [12]. After performing depth-wise convolution, it is common to proceed with point-wise convolution, where output of the depth-wise convolution is processed using 1x1 filters. The point-wise convolution helps to capture channel interactions and mix information across channels. Overall, depth-wise convolution is an effective technique for reducing computational complexity while retaining information flow in convolutional neural networks, making it suitable for models with limited resources or when efficiency is a priority. Howard et al. [45], introduced the MobileNet architecture, which incorporates depth-wise separable convolutions, in their paper. This network has excellent mobile device compatibility and high computational efficiency. Since then, a few additional experiments have used depth-wise convolutions to build effective networks [46–48]. This motivates us to think about enhancing the localization of the network and ensuring its efficacy by including more depth-wise convolutions in the suggested model.

3. The proposed method

3.1. Overall architecture

In this paper, we propose a hybrid transformer design that adds extra convolution processes to the transformer's core portions to capture more locality, which improves classification efficiency and precision on small datasets. Fig. 2 presents an overview of the proposed model for Alzheimer's diagnostic task. In this work, a multi-stage architecture is used, where each stage utilizes a similar architecture made up of a convolution-ViT(CViT) block with inverted residual unit (IRU); Light-weight Multi-Heat Self-Attention (LMHSA) and Local Feed-forward Network (LFFN). Instead of employing a conventional tokenization approach such as ViT, which involves linearly projecting each image patch into visual tokens by dividing it into non-overlapping, equally sized patches, we utilize a convolution token embedding block. This block consists of 3x3 convolutions, followed by ReLU [49] activation and a batch normalization layer. This approach allows for efficient extraction of local information, taking inspiration from the research referenced in Ref. [25], and [50]. By incorporating the described convolution token embedding block, the model's capacity is enhanced in terms of capturing low-dimensional local information and preserving the patch edge data. This block helps prevent the loss of important details and enables the model to better retain and utilize fine-grained information at a local level. The tokens are then run through a new transformer block made of IRU; LMHSA; LFFN and extended channels blocks without the position embedding unlike in original ViT. Convolution IRU, LMHSA, and LFFN blocks efficiently capture both neighborhood knowledge as well as long-range relationships while reducing the computational cost and increasing versatility of the transformer architecture. The last layers of the model involve batch normalization, followed by global average pooling, and a classification layer that utilizes SoftMax activation.

The suggested model is more suited for tackling image classification problems using limited datasets since it has a good ability to capture local and long-range information with fewer parameters. Additionally, positional embedding is not needed for this model's training purposes. In this study, we start by utilizing the CViT block as the main framework and showcase the application of a new projection method. Subsequently, we delved into a thorough analysis of the IRU, LMHSA, and LFFN block, focusing on their effective design to improve the overall performance of the network.

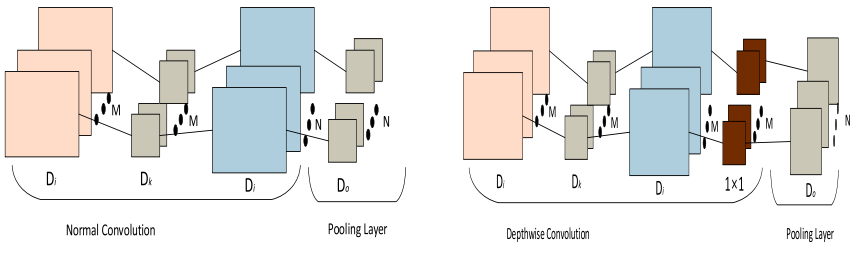
3.2. Inverted residual unit

The Inverted Residual Block, also referred to as the MBConv Block, is a specific type of residual block employed in vision models to improve efficiency. It was initially introduced in the MobileNet [45] CNN architecture and has been subsequently employed in various mobile-optimized CNNs. In contrast to the conventional Residual Block, which typically follows a wide-narrow-wide structure with the number of channels, the Inverted Residual Block adopts a narrow-wide-narrow configuration. It starts with a 1x1 convolution to widen the input, followed by a 3x3 depth wise convolution equipped with skip connection ensures efficacy while reducing computational burden. This promoted feature reuse and decreased the model's parameter count by ensuring that uniform weights distribution across multiple clusters of pixels in an image. Finally, a 1x1 convolution is used to decrease the number of channels so that the input and output can be added. It gets over the drawbacks of traditional positional embedding and the limits of traditional Vision Transformers in capturing local relationships and structured data that conventional CNNs capture inside individual patches. The residual block and the recommended IRU network have a very similar appearance. The inverted residual block consists of a depth-wise convolution, a projection layer, and an expansion layer. The IRU, however, features a unique shortcut connection location that improves its performance. Mathematically, it is represented as:

$$IRU(X) = Conv(\mathscr{F}(Conv(X))) \quad (1)$$

$$\mathscr{F}(X) = DWConv(X) + X \quad (2)$$

Let X be an input tensor of size $X \in \mathbb{R}^{H\times W\times d}$,where $H \times W$ represents the resolution of the current stage input and, d represents the dimension of the features. The function DWConv(.) represents the depth-wise



(a) (b)

Fig. 1. The visual representations of a) standard convolution and b) depth-wise convolution.

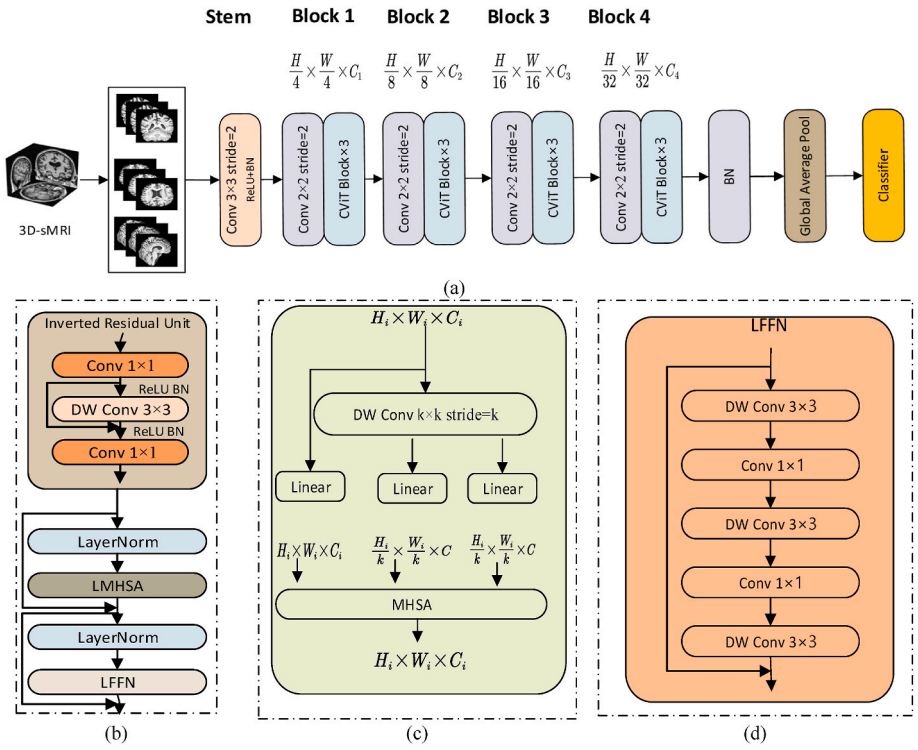


Fig. 2. Proposed optimized Convolution-ViT for Alzheimer's diagnosis, a) Overall architecture b) convolution transformer encoders with IRU, LMHSA and LFFN block, c) LMHSA block with additional depth-wise convolution and d) proposed sandglass LFFN block with skip connection.

convolution operation.

3.3. Lightweight multi-head self-attention

Fig. 3 depicted the various core blocks along with our proposed model. Our method consists of a k × k depth-wise convolution with a stride of k which minimizes the spatial size of the matrices K and V. The purpose of employing this technique is to minimize the computational burden involved in attention, thereby reducing the overall computational load. By utilizing fewer matrices produced by a convolution method, the quantity of self-attention computations is decreased. In original self-attention module, the input $X \in \mathbb{R}^{n\times d}$ is linearly transformed into query $Q \in \mathbb{R}^{n\times d_k}$, key $V \in \mathbb{R}^{n\times d_k}$ and value, $K \in \mathbb{R}^{n\times d_v}$, here, $n = H \times W$ represents the number of patches. For simplicity, we omit the reshape operation form $H \times W \times d$ to $n \times d$ tensors. The dimensions, d_k and d_v represent the sizes of the input key, query, and value, respectively. Afterward, the self-attention module is used as follows:

$$Attn = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Additionally, we introduce a relative location bias B inside every self-attention module, and the associated lightweight attention is described as:

$$Light\ Attn(Z_i) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{4}$$

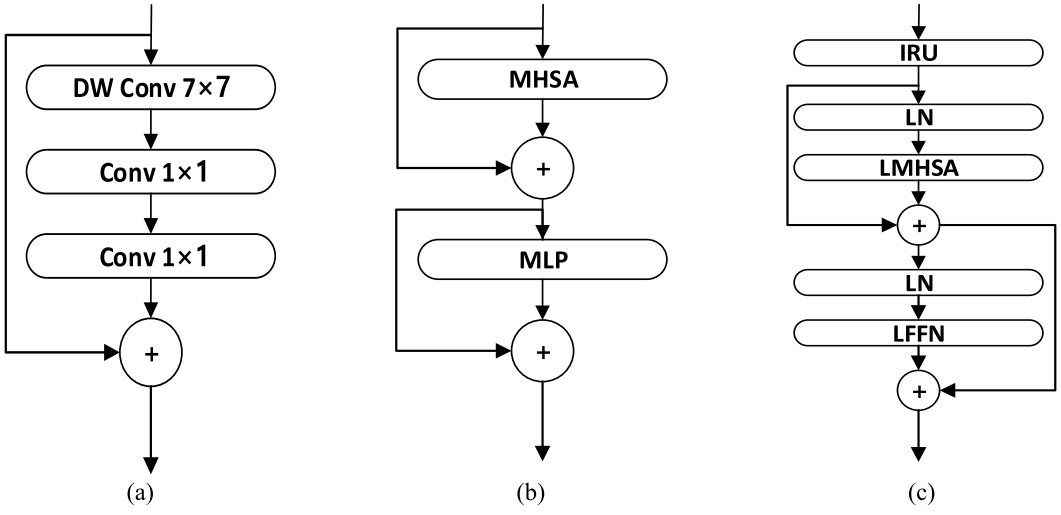


Fig. 3. A comparison between various core blocks a) ConvNext block b) ViT core block c) Ours (Convolution-Attention with IRU, LMHSA and LFFN) block.

3.4. Local feed-forward network

The final layer of each block substitutes an expansion layer for the traditional MLP in equation (5) of the Vision Transformers; Subsequently, a depth-wise convolution is applied, followed by a projection layer. Two linear layers are separated from one another in the original ViT by a GELU activation. Instead of GELU we employ ReLU [49] as the activation function because the more popular GELU [51] is frequently under supported by certain inference deployment platforms [52], as well as substantially slower than ReLU. Moreover, images dimension is multiplied by 4 in the first layer of LFFN and is decreased by the same number in the second layer:

$$FFN(X) = GELU(XW_1 + b_1)W_2 + b_2 \tag{5}$$

In this context, $W_1 \in \mathbb{R}^{d\times 4d}$ represents the weight matrix of first linear layer, and $W_2 \in \mathbb{R}^{4d\times d}$ represents the weight matrix of the second linear layer. Additionally, b_1 and b_2 denote the bias terms associated with the respective layers. To enhance the transformers' capacity to achieve locality in both higher and lower dimensions simultaneously, we suggest the LFFN block. This block incorporates a sandglass block with additional depth-wise convolutions into the original FFN of transformers.

The proposed LFFN resembles a residual block, which includes an expansion layer, followed by a depth-wise convolution, and finally a projection layer. The LFFN block carries out a series of actions. First rearranging the series of tokens into a 2D lattice will accurately reflect the feed-forward network, which is applied positionally to a sequence of tokens Z_i. The following equation (10) represent the final reshaped features:

$$LFFN\left(X_i^{d1}\right) = DWConv(z_i) \tag{6}$$

$$X_i^{l1} = Conv\left(X_i^{d1}\right) \tag{7}$$

$$X_i^{d2} = DWConv\left(X_i^{l1}\right) \tag{8}$$

$$X_i^{d2} = Conv\left(X_i^{d2}\right) \tag{9}$$

$$X_i^{d3} = DWConv\left(X_i^{d2}\right) + Z_i \tag{10}$$

Local information is extracted without almost any additional processing expense using the depth-wise convolution. The logic for the use of shortcuts is comparable to that of original residual networks, which can enhance the gradient's capacity to propagate between layers. In our research, we demonstrate that this shortcut aids the network in producing better outcomes. Overall, these components can be represented mathematically as:

$$X_i^{'} = IRU(X_{i-1}) \tag{11}$$

$$Z_i = LMHSA\left(LN\left(X_i^{'}\right)\right) + X_i^i \tag{12}$$

$$X_i^{d3} = DWConv\left(X_i^{d2}\right) + Z_i \tag{13}$$

where $X_i^{'}$ and Z_i denote the output features of the IRU and the LMHSA module for block i respectively. LN refers to Layer Normalization.

3.5. Gradient centralization

Fig. 4 above illustrates the gradient centralization (GC) operation. Gradient descent with a controlled loss function is a technique known as GC. An important factor in enhancing a Deep Neural Network's (DNN) performance is model optimization. Optimization can be done by Z-score standardization on the network's activations or using methods like Batch Normalization and Weight Standardization. For optimization in our approach, we employed a novel strategy called GC [28]. Unlike existing methods that primarily focus on activations or weights, GC directly operates on gradients by centralizing the gradient vectors to have zero mean. By adding a new restriction to the weight vector, it places such limits on the loss function. By controlling the weight space as well as the output feature, it enhances the DNNs' capacity for generalization. Additionally, it increases the gradient's and loss function's Lipshitzness, stabilizing the network's training process and boosting its effectiveness. The utilization of GC provides benefits in both the output feature space and the weight space regularization. This dual regularization contributes to enhancing the model's generalization performance while mitigating the risk of overfitting on the training data. As a result, the gradient of the weights becomes more predictable and stable enabling quick model training. Furthermore, it prevents gradient explosion, which stabilizes the model training procedure. It is simple to incorporate into existing gradient based DNN optimization techniques like Adam and SGDM.

4. Experimental setting

4.1. Studied dataset

The dataset utilized in this investigation was retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, available online at (http://adni.loni.usc.edu). To aid in the early detection of AD and the investigation of biological markers for the disorder, the ADNI

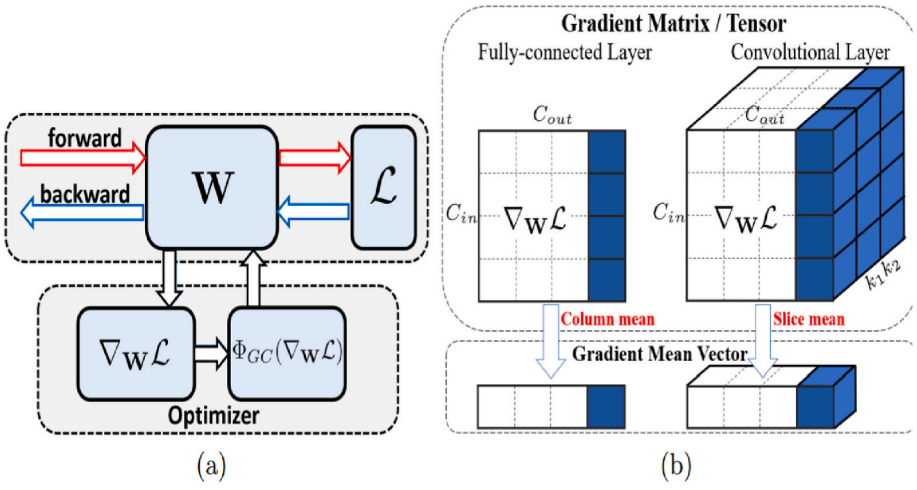


Fig. 4. a) Drawing of a gradient centralization (GC) map b) examples of the GC operation on the weight of gradient matrices. The gradient column mean of the gradient matrix is calculated using GC, and each column/slice is centralized to have a mean of zero [28].

offers researchers worldwide access to a publicly available AD database [53]. We evaluated 315 HC, 370 MCI, and 390 AD samples for this study. All sMRI scans were performed at a 3T resolution, and the resulting images were T1-weighted which is comprised of magnetization-prepared rapid-acquisition gradient-echo sequences. The obtained images had a spatial resolution of 182 × 218 × 182 with a voxel size of 1 × 1 × 1 mm^3. Finally, for the purpose of subsequent model training, the process involved the division of each 3D sMR image into 2D images by means of slicing and tiling, thereby achieving a dimension of 224 × 224 pixels. Table 1 presents the demographic and clinical information of the participants, including their gender, age, MMSE results, and clinical dementia rate (CDR) scores. The ages of the three subject combinations were evenly distributed. The MMSE scores of the HC group showed slight differences, in contrast to the considerably greater variations observed in the scores of the remaining two group.

4.2. Dataset preprocessing

Fig. 5 above shows the sample slice used in this study. First, the centre of the anterior commissure (AC) - posterior commissure (PC) line was chosen as the new location for all raw sMRI data. Then, for SPM [54], we utilized the computational anatomy toolbox (CAT12, accessible at http://www.neuro.uni-jena.de/cat/), which includes several morphometry techniques, including surface-based morphometry (SBM) and voxel-based morphometry (VBM). Our preprocessing procedures included the following steps: removal of non-brain tissue, such as the skull and neck, etc. Normalization to the EPI template, modulation, and spatial smoothing using an 8 mm full-width at half-maximum (FWHM) Gaussian filter.

4.3. Training setup

To evaluate the efficacy of the proposed optimized model, we compared it with other baseline CNN and Transformer-based networks, including ResNet-50 [31], DensNet121 [55], and various versions of CMT [25]. This comparison aimed to assess the performance and capabilities of the proposed model in relation to these established architectures. To identify Alzheimer's brain sMRI data, CViT hybrid model was utilized. We randomly rearranged the images and conducted all experiments by dividing the data into 10% for testing and 90% for training. Additionally, 20% of the training set was set aside as a validation set. The initial learning rate was set to 0.003 and 200 epochs to train the model and it was intended to become zero after a single cycle of the cosine. CViT models were trained using the AdamW and AdamWGC [28] optimizer with a weight decay of 0.03 and a batch size of 32. The optimal value for the batch size of models was determined by identifying the ideal point between the batch sizes of 8 and 128, utilizing increments of the power of 2. As we move from larger batch sizes to smaller ones, we notice that for batch size 32 have the least error rate. According to this research, large-batch approaches tend to converge to sharp minima of the testing and training functions, and sharp minima result in less effective generalization. Small-batch techniques, on the other hand, always lead to flat minima. The cross-entropy loss makes sure that the network training goes smoothly. The implementation uses ADNI Dataset

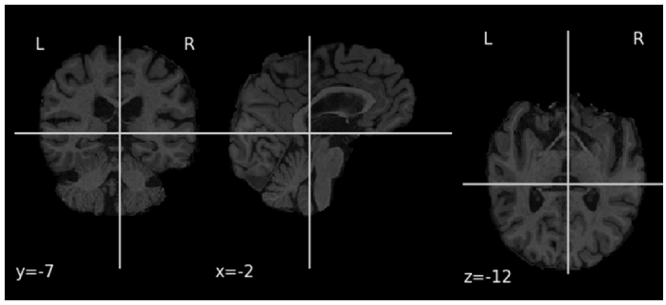


Fig. 5. Illustration of Coronal, Sagittal and Axial slice of sMRI ADNI dataset used in the experiment.

[56] and deals with Alzheimer's diagnosis as multi-class classification problem statement. We used 3D structural sMRI scans of 1075 individuals (390 AD, 370 MCI, and 315 HC) to construct our 2D model with balance dataset. Data augmentation is used to increase amount of data during training phase to make the model to gain more information about the Alzheimer's atrophy present in the sMRI images. We specifically excluded a colour jitter, Gaussian blur, and solarization image augmentations and opted instead for random horizontal flip, vertical flip, height shift, and random zoom augmentations. Utilizing the same training methodology as [25], the suggested approach for classification was executed in Python 3.9.13 using the Keras library, which relies on Tensorflow 2.11.0. The process was then conducted on a computer equipped with an NVIDIA RTX3090 GPU and tested in the Ubuntu 20.04.6-x64 operating system.

4.4. Network architecture

Following the fundamental CMT settings, we design the suggested model architectures. The detailed architecture of the proposed model is presented in Table 2. Firstly, the convolutional layer with a kernel size of 3 × 3 and a stride of 2 were utilized as convolution stem, which produced 32 enhanced channels. Additionally, a Batch Normalization layer is introduced for stable training for convolutional stem. The next step is to utilize an inverted residual unit inside CViT blocks with a depth wise convolution of kernel size of 3 × 3. The number of LMHSA blocks are individually set to 3 for each block respectively. In the LMHSA block, which is used for convolution projection, the size of the convolution kernel is set to 3, and the number of heads is set to 1,2,4,8 with reduction rates 8,4,2,1 respectively while setting the expansion ratio as 4.

4.5. Evaluation

The anticipated results for the diagnostic tasks are represented by the abbreviations TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). A positive sample, as defined by TP, is one that was accurately projected to be positive. A sample that was appropriately identified as negative is referred to as TN. A negative sample that was mistakenly labeled as a positive sample is indicated by the symbol FP. The abbreviation FN denotes inaccurate prediction of a positive sample as a negative sample. We employ the commonly used

Table 1
Information on the collected individuals' demographics.

Groups	Gender (M/F)	Education	Age (Years)	MMSE	CDR	APOEε4	FAQ
AD	160/155[a]	15.47 ± 3.08	74.07 ± 7.5[b]	24.54 ± 2.25[b]	3.29 ± 1.7[b]	0.88 ± 0.70	10.30 ± 7.05[b]
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; MCI: Mild cognitive Impairment; CN: Normal Control; CDR: Clinical Dementia Rate; MMSE: Mini Mental state Examination. FAQ: Functional Activities Questionnaires.

[a] Group-level two-sample t-tests are conducted for age, education, MMSE, FAQ, and CDR.

[b] group-level chi-square tests are conducted for gender.

Table 2
Overall architecture of proposed model for Alzheimer's classification, with the output size matching the input resolution of 224 × 224. Convolutional layers and optimized ViT blocks are indicated within brackets along with the number of stacked blocks.

Stage	Output Size	Layer Name	CViT
Stem	112 × 112	Convolutional layer	3 × 3, 32, Stride 2
Block1	56 × 56	Patch Embedding	2 × 2, 64, stride 2
		CViT Block	[3 × 3, 64, H1 = 1, k1 = 8, R1 = 4] × 3
Block2	28 × 28	Patch Embedding	2 × 2, 128, stride 2
		CViT Block	[3 × 3, 128, H1 = 2, k1 = 4, R1 = 4] × 3
Block3	14 × 14	Patch Embedding	2 × 2, 256, stride 2
		CViT Block	[3 × 3, 256, H1 = 4, k1 = 2, R1 = 4] × 3
Block4	7 × 7	Patch Embedding	2 × 2, 512, stride 2
		CViT Block	[3 × 3, 512, H1 = 8, k1 = 1, R1 = 4] × 3
	1 × 1	Global Average Pooling	1 × 1512
		FC	3
		#Params	**15.9M**
		#FLOPs	**2.2 B**

metrics of accuracy, specificity, sensitivity, precision, F1 score, and receiver operating characteristic curve (ROC curve) to assess the performance of our diagnostic model. Accuracy is the percentage of correctly diagnosed test samples among all test samples, as demonstrated in relation (14).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

The proportion of correctly classified samples, as shown in equation (15), represents specificity.

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

The definition of sensitivity, as depicted in equation (16), measures the ability of a model to accurately identify AD patients among all the positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

Precision, as expressed in equation (17), quantifies the proportion of correctly predicted positive observations out of all the expected positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

The F1 score, as demonstrated in equation (18), is computed by taking the weighted average of precision and recall. It combines both measures to provide an overall evaluation of the model's performance.

$$F1\ score = 2\left(\frac{Precision \times Sensitivity}{Precision + Sensitivity}\right) \tag{18}$$

An ROC curve graphically represents the performance of a classification model across various classification thresholds. The ROC curve illustrates the relationship between two variables: The first parameter, True Positive Rate (TPR), is also commonly referred to as recall and is represented by equation (19). which quantifies the proportion of correctly predicted positive instances (true positives) out of all the actual positive instances.

$$TPR = \frac{TP}{TP + FN} \tag{19}$$

The second parameter, False Positive Rate (FPR), is defined in equation (20). which measures the proportion of incorrectly predicted negative instances (false positives) out of all the actual negative instances. It represents the rate at which the model incorrectly identifies negative cases as positive.

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

5. Experimental results and analysis

Using limited ADNI Alzheimer's datasets, we assess the proposed optimized model in this section. Additionally, to validate the design of the suggested architecture, we also conducted a comparative and investigative study on existing baseline CNN, ViT and CMT model.

5.1. Demographics analysis

In the comparison between AD/HC/MCI, MCI/HC, and the comparison between MCI/AD, there were no statistically significant differences in age between the groups. However, in combinations of all groups, there was a significant variation in Mini-Mental State Examination (MMSE) ($P \leq 0.05$) and Clinical Dementia Rating (CDR) ($P \leq 0.05$). AD exhibits a higher percentage of males, in contrast to HC which shows a similar percentage of females' population. On the other hand, MCI have a slight variation of females and males' population. The prevalence of males in AD is 64.10%, while the prevalence of females in MCI is 45.94%. In contrast, the prevalence of male and females in HC is similar with 50.79% and 49.20%, respectively. All numerical and clinical data, including the APOEƐ4 positive rate, were expressed as the mean value plus or minus the standard deviation. Variables included education, MMSE, and CDR. Group-level two-sample t-tests were performed for age, education, MMSE, and CDR, while group-level chi-square tests were conducted for gender. These variables have been comprehensively described and analyzed in Table 1 and Fig. 6 below respectively. From the figure we can say that education level and APOE4 greatly influence the clinical dementia rate on the patients. Similarly, person education level linearly related to MMSE score; while individuals age increases their chance of AD is increase based on the presence of APOE4 gene in the individuals.

5.2. Performance comparison of the proposed method with different models on ADNI dataset

The proposed method was evaluated using the ADNI database, specifically for a multi-class classification task involving AD, MCI, and HC individuals. The results of the proposed method were compared with other models trained on the same dataset, namely ResNet50, VGG16, VGG19, DenseNet121, baseline ViT and CMT varieties. The evaluation was performed using metrics: accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE) and area under receiver operating characteristic (AUROC).

The proposed model can be partitioned into two components: series of convolution layers (3 × 3 filters) with stride 2 for a patch embedding from input sMRI slices and Transformer module to preserve global information's. The classification performance of each model was compared for the multiclass classification tasks setting same training parameters for fair comparison, and the results were presented in Table 3, Fig. 7, and Fig. 8. Figs. 7 and 8 represent confusion matrix, the ROC curve, Precision-Recall curve, loss, and accuracy plot for two optimizers respectively. The effectiveness of the proposed model was demonstrated in both binary classification scenarios AD/HC; MCI/HC and multi-class classifications. According to the evaluation, by incorporating IRU, LMHSA and LFFN in our method achieved the highest accuracy rate of 94.31%, outperforming CNN, baseline ViT and CMT variant in terms of accuracy while reducing the computational complexity and model parameters. VGG16 achieved an accuracy rate of 90.15%, which was slightly higher than ResNet50 accuracy rate 89.24%.

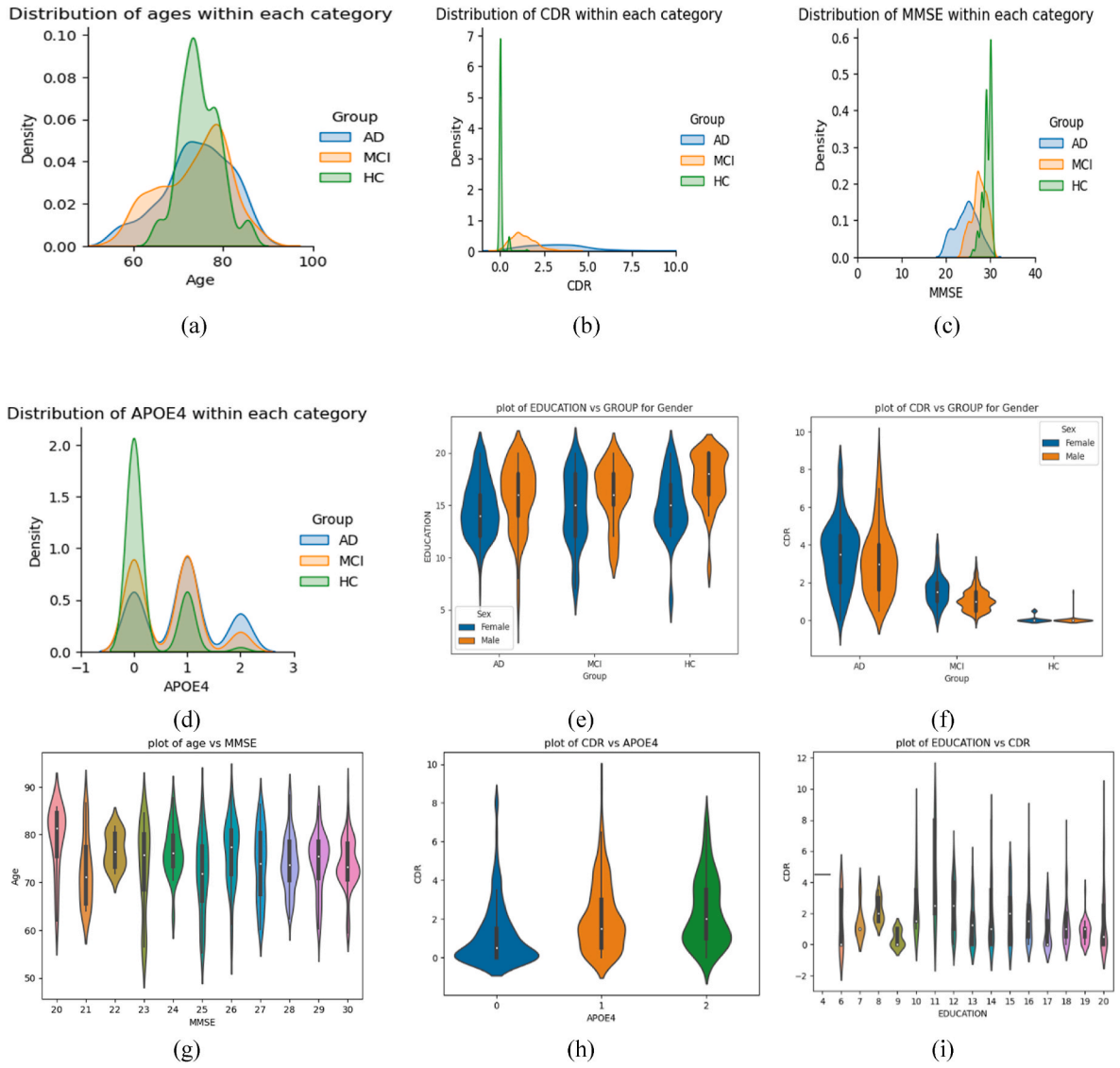


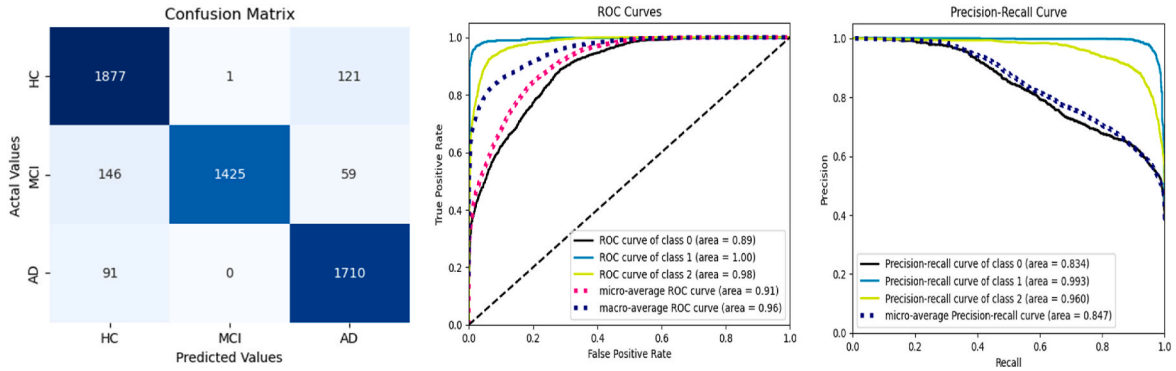
Fig. 6. Visualizing demographic and clinical density maps: Distributions of age, CDR MMSE scores, and APOE4 scores are shown in (a), (b), (c) and (d). To show the relationship between AD stages and various indicators in the clinical AD we depicted the violin plot: (e), (f), (g), (h) and (i).

Table 3
Multiclass classification results (AD/MCI/HC) for different models on ADNI dataset.

Models	Image Size	Param (M)	FLOPs (B)	ACC	SEN	SPE	Precision	Recall	F1-Score
VGG-16 [58]	224X224	134.2	15.4	90.15	89.74	94.23	92.01	90.25	90.86
VGG-19 [58]	224X224	139.5	19.6	86.25	93.10	87.51	89.04	88.07	91.02
DensNet-121 [55]	224X224	7.03	2.8	87.15	87.17	93.18	89.25	88.41	88.19
Resnet-50 [31]	224X224	23.5	3.8	89.24	88.07	93.13	90.37	89.11	89.20
ViT-S [17]	224X224	50	22	78.47	82.56	74.06	78.55	78.47	78.42
ViT-B [17]	224X224	86	35.1	81.09	72.87	89.96	82.31	81.09	80.99
ViT-L [17]	224X224	303	122.9	83.90	88.76	78.66	84.12	83.9	83.84
CMT-Ti [25]	160X160	8.2	0.65	74.87	87.14	74.38	83.01	74.33	85.02
CMT-XS [25]	192X192	14.09	1.57	85.41	91.03	84.71	87.02	85.35	88.97
CMT-S [25]	224X224	25.1	4.08	86.05	92.17	85.31	88.03	86.43	90.05
CMT-B [25]	256X256	44.6	9.42	88.01	81.41	95.43	89.12	88.54	85.09
Ours (AdamW)	224X224	15.9	2.2	92.07	95.43	92.21	93.08	92.54	94.24
Ours (AdamWGC)	224X224	15.9	2.2	94.31	97.14	94.11	95.02	94.55	96.06

FLOPs: Floating point operations; ACC: Accuracy, SEN: Sensitivity; SPE: Specificity.

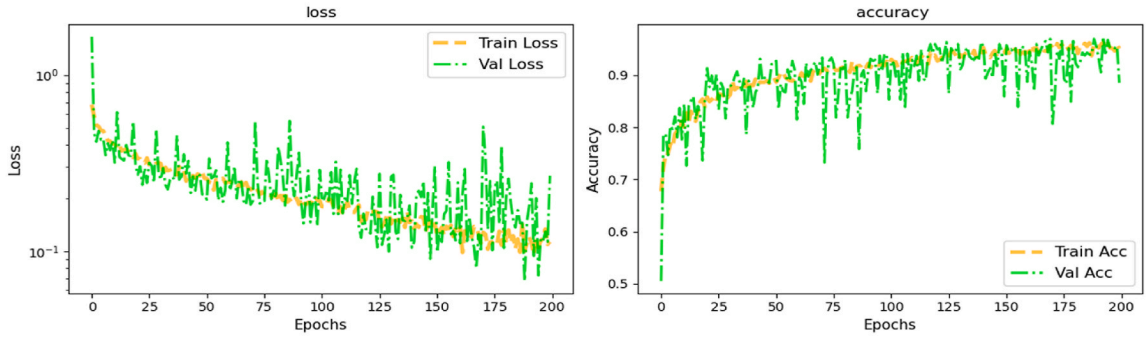
Similarly, CMT-S performs better as compared to the other variants with accuracy rate of 88.01%. However, the proposed method surpassed them in terms of accuracy rate and other evaluation criteria as present in results Table 3. The proposed method demonstrated higher sensitivity and precision; for example, it achieved a sensitivity of 97.14% and precision of 95.02%, which were both higher than the baseline CNN, ViT



(a)AdamW optimizer's Confusion matrix for multiclass classification

(b)AdamW optimizer's Receiver Operating Characteristics (ROC) curve for multiclass classification

(c)AdamW optimizer's Precision-Recall curve for multiclass classification



(d) AdamW optimizer's loss and accuracy graph on training/validation dataset for multi-class classification.

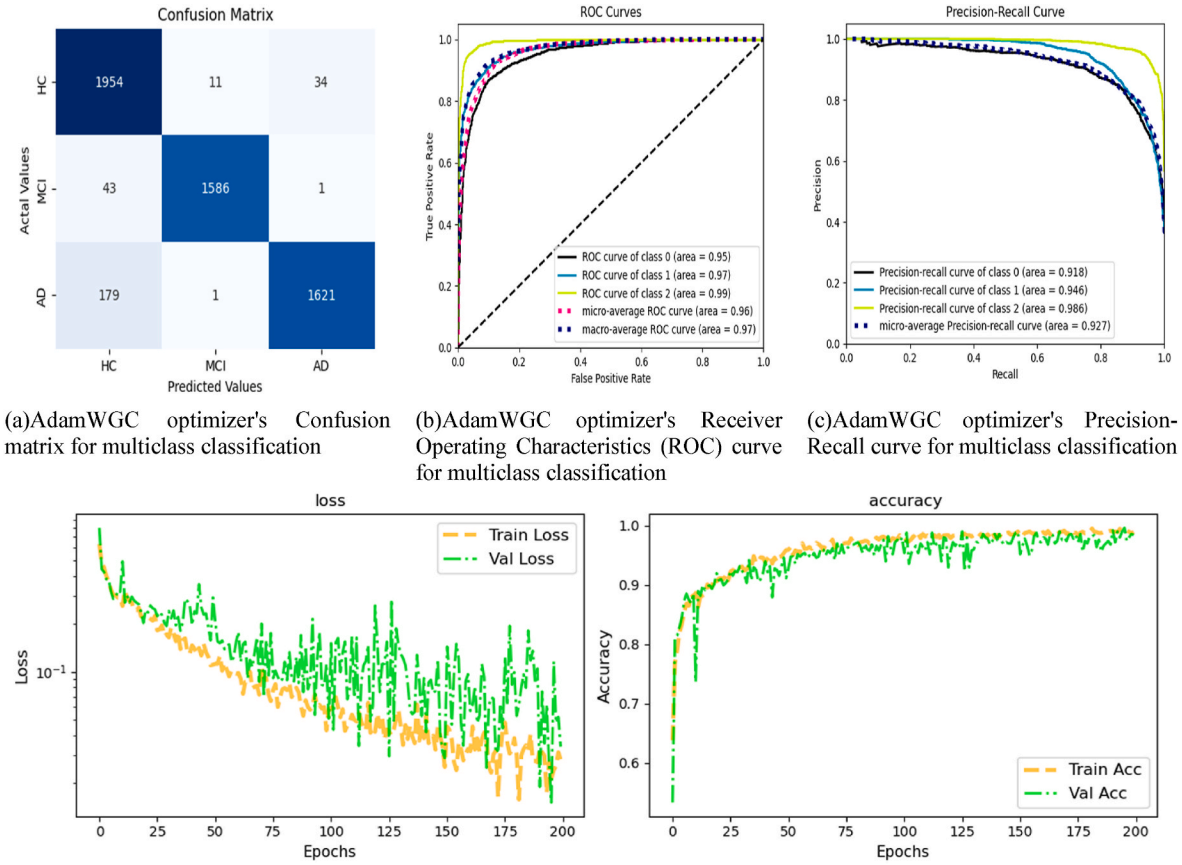
Fig. 7. Visualizing model performance on AdamW optimizer: a) confusion matrix for multi-class classification of Alzheimer's disease diagnosis, b) ROC curves for multiclass classification tasks. To evaluate the overall effectiveness of the model, we utilize the area under the curve (AUC). Better performance is represented by a larger area, c) Precision-Recall plot for evaluating the model. The larger area indicates better performance, d) loss and accuracy graph on training/validation dataset.

and CMT models. However, CMT-B and VGG16 showed a slightly better specificity score rate of 95.43% and 94.23% respectively. We trained our model with AdamWGC optimizers which demonstrated better and smooth training process and achieved excellent performance, especially training and validation loss reduced greatly as compared to AdamW optimizer (Fig. 8(d)). Specifically, GC operates directly on gradients and removed the mean from the gradient vectors and centralized them to have zero mean. Which improves the loss function with a constraint on weight vectors, and regularizes both weight space and output feature space and helps to train the model smoothly [28].

Overall, the self-attention-based models exhibited the best classification performance, followed by the CNN-based models. Fig. 8 shows the application of Grad-CAM (Gradient-weighted Class Activation Mapping) [57] on the AD identification task. In general, the utilization of GC in the AdamW optimizer resulted in an increased accuracy. Comparing the performance of the two optimizers, the AdamWGC achieved a higher accuracy rate of 94.31% compared to the accuracy rate of 92.07% achieved by an AdamW. This suggests that incorporating GC into the AdamW optimizer had a positive impact on the model's ability to classify AD brain images correctly, leading to improved accuracy by 2.24%. Similarly, notable improvements can be observed in terms of micro average AUROC by 5 % (Fig. 7(b) and Fig. 8(b)) and the area under precision-recall by 8% (AUPR; Fig. 7(c) and Fig. 8(c) curve. A higher AUROC and AUPR indicates better discrimination and overall performance of the model in distinguishing between positive and negative instances. Overall, the utilization of GC in optimizer likely contributes to improved classification performance and increased separability of the predicted AD classes. Furthermore, the precision-recall curve, which depicts the trade-off between precision and recall, shows substantial improvement when GC is incorporated into optimizer while training the proposed convolution-attention model. The precision-recall curve is especially informative in situations where data imbalance exists or when the focus is on positive instances. The enhancement observed in the precision-recall curve suggests that the model achieves higher precision for a given level of recall, indicating improved performance which correctly identifying positive instances while minimizing false positives rate. In summary, applying GC to the AdamW optimizer yields significant improvements in both AUC and the precision-recall curve, highlighting the enhanced discrimination and classification performance of the model for Alzheimer's diagnosis task using ADNI database.

5.3. Ablation study

To examine the impact of design choices on the multiclass classification of AD using the proposed model, we conducted an ablation experiment. This experiment aimed to investigate the effects of different block sizes on model performance and computational complexity. For fair comparison, how number of blocks affect the performance we keep all parameters same, and we only change the number of blocks. We tested three versions of our method: as shown in Table 4. All models were trained using the same training setup with AdamWGC optimizer. The results are presented in Table 4. It can be observed that the accuracy of classification is comparable between block size 3 and 4 but the computational complexity of the model increased largely. More specifically, as the total number of attention blocks goes from 12 to 16, the classification accuracy of model decreases from 94.31% to 91.74%, resulting in a decrease of 2.57% overall accuracy with 8.3% increase in computation cost. Nevertheless, the trend of enhancing classification performance also decreases while reducing the number of blocks into 8.



(a)AdamWGC optimizer's Confusion matrix for multiclass classification

(b)AdamWGC optimizer's Receiver Operating Characteristics (ROC) curve for multiclass classification

(c)AdamWGC optimizer's Precision-Recall curve for multiclass classification

(d) AdamWGC optimizer's loss and accuracy graph on training/validation dataset for multi-class classification.

Fig. 8. Visualizing model performance on AdamWGC optimizer: a) confusion matrix for multi-class classification of Alzheimer's disease diagnosis, b) ROC curves for multiclass classification tasks. To evaluate the overall effectiveness of the model, we utilize the area under the curve (AUC). Better performance is represented by a larger area, c) Precision-Recall plot for evaluating the model. The larger area indicates better performance, d) loss and accuracy graph on training/validation dataset.

Table 4
Multiclass classification results (AD/MCI/HC) on different block sizes for ADNI dataset.

No. of blocks	Image Size	#Parameters	#FLOPs	ACC	SEN	SPE	Precision	Recall	F1-Score
2,2,2,2	224 × 224	11.8 M	1.4 B	88.53	86.05	91.21	88.70	88.53	88.53
3,3,3,3	224 × 224	15.9 M	2.2B	94.31	97.14	94.11	95.02	94.55	96.06
4,4,4,4	224 × 224	21.8 M	2.6B	91.74	85.27	96.65	91.39	90.74	90.73

For instance, while the total number of blocks reduced to 8, the accuracy was decreased to 88.53%. Consequently, considering the overall computation costs and benefits, we set attention blocks size as 3 in each step in our experiments. Our findings indicate that the small version with three blocks in each attention-head achieved the highest classification performance with less parameters and computational complexity for AD data set as compared to others. We believe that a block size of three captures the most effective and informative features of sMRI images by extracting brain regions with corresponding AD atrophy. Larger block sizes result in overly generalized information, leading to computationally complex and loss of details. Conversely, small blocks sizes can compromise the semantic information of the sMRI scan even though it is computationally efficient.

5.4. Performance comparison with existing state-of-art-machine learning, CNN methods

Several recent literary works have examined the neuroimaging technique for discriminative classification of AD, focusing on patients with MCI and the identification of individuals with Alzheimer's from healthy controls. Nevertheless, making a direct comparison with the current state-of-the-art methods is challenging due to the utilization of different datasets and classification techniques in most literary works, both of which have had a significant impact on performance accuracy. By employing various classifiers architectures for the discrimination between AD and HC, previous literary works have reported different accuracy ranges as depicted in Table 5. These literary works have employed the ADNI database to evaluate their proposed methods, and it is evident that the classification accuracy has been influenced by the number of subjects. For a fair evaluation we conducted a comparison of our model's classification results with previous studies using the ADNI database. Initially, we conducted a comparison between our proposed model and traditional machine learning methods. In the study by Liu et al. [59] introduced a whole-brain hierarchical network that extracted brain features from regions of interest (ROI) and employed the multiple kernels boosting (MKBoost) algorithm for classification. Using a single structural MRI modality dataset, they achieved accuracy rates of 94.65% for distinguishing AD/HC, and 85.79% for differentiating MCI/HC. By proposing an SVM-based method that integrated spatial-anatomical information and employed a group lasso penalty to induce sparsity Sun

Table 5
Performance comparison with state-of-the-art machine learning and CNN methods for AD/MCI/HC for different models on Alzheimer's dataset.

Reference	Methods	Modality	Subjects (AD/MCI/HC)	AD/MCI/HC			AD/HC			HC/MCI		
				ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
Liu et al. [59]	MKBoost,SVM	MRI	200/280/230	–	–	–	94.65	95.03	91.76	84.79	88.91	80.34
Sun et al. [60]	SVM	MRI	137/210/162				95.10	93.8	83.80	70.80	72.10	69.10
Lian et al. [61]	Hierarchical CNN	MRI	429/-/358	–	–	–	90.30	82.40	96.50	–	–	–
Kang et al. [64]	2D CNN	MRI	229/382/187	–	–	–	90.04	93.9	83.80	72.40	74.70	84.80
Li et al. [62]	3D CNN	MRI	330/299/299	–	–	–	93.20	95	89.80	80.40	83.20	78.60
Liu.J et al. [46]	Multi-layer NN	MRI	90/136/266	78.02	83.21	75.32	–	–	–	84.65	82.35	79.50
Liu.M et al. [63]	3D DenseNet	MRI	97/233/119	88.90	86.60	90.80	–	–	–	76.20	79.50	69.80
Proposed Method	CViT(AdamW)	MRI	315/370/390	92.07	95.43	92.21	–	–	–	–	–	–
	CViT(AdamWGC)			94.31	97.14	94.11	95.37	91.09	1.00	92.15	89.92	94.56

et al. [60] achieved comparable outcomes. Their approach yielded accuracy rates of 95.1% for distinguishing AD/HC and 70.80% for differentiating MCI/HC.

Furthermore, we conducted a comparison between our model and existing deep learning architectures. Lian et al. [61] proposed a hierarchical CNN based method that captured anatomical atrophy localization in structural MRI scans of the brain. They achieved accuracies of 90.30% for AD vs HC with 82.40% sensitivity, and 96.50% specificity for binary classification tasks. A 3D deep learning system was created by Li et al. [62] based on structural MRI images for the detection of Alzheimer's disease in individuals. For AD over HC and MCI over HC, their model had accuracy rates of 93.20% and 80.40%, respectively. Liu.J et al. [46] designed a CNN-based architecture using the OASIS dataset and obtained accuracy rates of 78.02% for multiclass classification, 84.65% for distinguishing MCI/HC, and 75.32% for distinguishing AD/MCI when applied to the ADNI dataset. They made further enhancements to their approach by employing a deep separable convolution model, which helped reduce the number of parameters. As a result, they achieved an accuracy of 77.79%. In order to emphasize the features extracted specifically from the segmented region of the hippocampus, Liu.M et al. [63] devised an architecture that integrated 3D DenseNet and a multi-task CNN. Their model achieved accuracy rates of 88.90% for distinguishing multi-class AD/MCI/HC and 76.20% for differentiating MCI/HC.

In comparison to these previous CNN base studies, our proposed model achieved competitive or improved classification accuracies, demonstrating the efficacy of our approach in AD diagnosis. Table 5 provides a summary comparison of different studies, including the proposed depth wise convolution method by Liu. et al. [46], in terms of classification performance. The results indicate that our proposed method tends to achieve high discrimination accuracy while utilizing a reduced number of parameters. The reduced number of parameters suggests that the proposed method can achieve competitive classification accuracy while being computationally efficient for clinical diagnostic of Alzheimer's.

5.5. Performance comparison with existing state-of-art-transformer methods

In this section, we set out to evaluate the effectiveness of our proposed framework by comparing its performance with recent ViT based methods in the classification and diagnosis of AD. Due to the limited availability of ViT-based AD research and the absence of multi-class classification reports in the literature, our comparison focuses on binary classification. The results are detailed in Table 6, showcasing our findings alongside those reported in existing studies. Each investigation is accompanied by information on the methodology employed and the recorded performance measures.

Specifically, we compare our results with attention-based methodologies introduced by Xin et al. [65], BranInf proposed by Zhu et al. [66], Conv-Swin by Hu et al. [67], Addaformer by Kushol [44], Zhu et al. [8] and Zhang and Kalavati et al. [40], as they utilized sMRI images from the ADNI database in their experiments. Hu et al. [67] employed a Conv-Swin transformer model, combining VGG16 for convolutional feature extraction and Swin transformer for feature fusion, achieving an accuracy rate of 93.56% for AD/HC and 79.07% for HC/MCI. Zhu et al. [8] introduced a BranInf model with ProbSpares attention as the main ViT backbone, focusing on enhancing the efficacy of attention mechanism for AD classification and obtained accuracy rate of 97.97% with higher specificity 98.17% for AD/HC. However, their outcomes are inferior in terms of both accuracy and specificity indicators for HC/MCI classification task. Kushol et al. [44] proposed an Addformer, combining frequency domain and extracted features of sMRI neuroimages for AD classification using ViT as the primary architecture. However, many state-of-the-art methodologies, including those mentioned above, rely on basic ViT backbone approaches for AD diagnostic classification, impacting the computational complexity of the model. In contrast, our approach incorporates fully automated convolutional attention in deep learning to identify AD stages with reduced computational burden, enabling timely AD identification without human intervention. Our sMRI-based convolution vision transformers yield promising results for

Table 6
Performance comparison with state-of-the-art transformer methods for AD/MCI/HC for different models on Alzheimer's dataset.

Reference	Methods	Modality	Subjects (AD/MCI/HC)	AD/HC			HC/MCI		
				ACC	SEN	SPE	ACC	SEN	SPE
BraInf [8]	distilling-ViT	MRI	313/319/324	97.97	97.94	98.17	91.89	90.66	93.01
Xin et al. [65]	Conv-Swin Net	MRI	336/-/529	0.939	0.925	0.947			
Conv-Swin [67]	Conv-Swinformer	MRI(Axial-Slice)	508/1412/970	0.9356	–	–	0.7907	–	–
Addformer [44]	Addaformer	MRI	159/-/229	0.882	–	–	–	–	–
Zhu et al. [66]	DA-MIDL	MRI	389/-/391	0.924	0.91	0.938			
Zhang and Khalvati [40]	VViT-tiny	MRI	180/-/214	0.72	–	–	–	–	–
	VViT-small			0.72	–	–	–	–	–
	VViT-Base			0.74	–	–	–	–	–
	CVVT-tiny			0.84	–	–	–	–	–
	CVVT-small			0.86	–	–	–	–	–
	CVVT-Base			0.84	–	–	–	–	–
Proposed Method	CViT(AdamW)	MRI	315/370/390	–	–	–	–	–	–
	CViT(AdamGC)			95.37	91.09	1.00	92.15	89.92	94.56

predicting AD progression in a fully automated manner, as presented in Table 3, Figs. 7, and Fig. 8. The proposed model, integrating an efficient version of vision transformers with a convolution block inside the main transformer, consistently outperforms previous studies in sensitivity, specificity, and accuracy.

In summary, our convolution self-attention-based model exhibits optimal overall classification performance, surpassing both deep learning models and ViT methods alone for AD diagnosis. Our proposed model achieves an accuracy rate of 94.31% for multi-class and 95.37% for AD/HC binary classification tasks, demonstrating improvements compared to previous ViT-based studies. Notably, our model outperforms Zhu et al. [8] in accuracy for HC/MCI classification. The confusion matrix of our model, illustrated in Fig. 8, showcases its superior performance with an AUC of 0.96.

These results affirm the efficiency of our proposed methodology, showcasing its effectiveness compared to the latest research utilizing ViT and neuroimaging for the predictive diagnosis of AD. Moreover, integration of IRU, LMHSA, and LFFN preserves both global and local information in sMRI with lower computational cost, making it a promising approach for diagnostic classification of Alzheimer's and its clinical application.

5.6. Attention sensitive pathological brain regions

Identifying the specific brain region that is closely associated with the predictions made by deep learning models is crucial in the context of computer-aided diagnosis. When it comes to the clinical diagnosis of AD, observing structural changes in the brain plays a significant role. In our study, we employ the Grad-CAM [57] technique to investigate convolution and attention maps which brain regions the conv-attention layers of our model focus on to classify Alzheimer's classes (Figs. 9 and 11). The depicted different slices highlight several locations that our proposed method identifies.

In Fig. 9(b), we compare the AD/HC classes within each marked location for convolution maps and Fig. 9(c) compare the LMHSA maps for AD/HC classes. Our findings reveal that main regions, namely the medial-occipital gyrus, superior frontal gyrus, third ventricle, putamen, thalamus, hippocampus, amygdala, medial frontal gyrus, superior temporal gyrus, frontal lobe, medial frontal, ventricular, and occipital areas, carry the most informative features for our model's predictions. Similarly, Fig. 11 provides an overview of the related brain attention score corresponding to the AD brain sMRI regions for different slices on each attention head. First, we generate an attention map by aggregating the attention scores from the multiple heads of the self-attention mechanism. This can be done by taking the mean or max attention scores across each attention head [68]. By examining the attention map or heatmap, we can gain insights into which areas of the Alzheimer's brain sMRI are considered important by the model for making accurate predictions. Which provide a better understanding of the model's decision-making process and potentially reveals regions of interest for further analysis and research. More importantly these identified key atrophic brain areas can help the doctor properly analyze and help to find the viable treatment for AD. These identified regions align with the findings of numerous previous studies on AD diagnosis [61,69,70], which further validate the reliability and effectiveness of our proposed model.

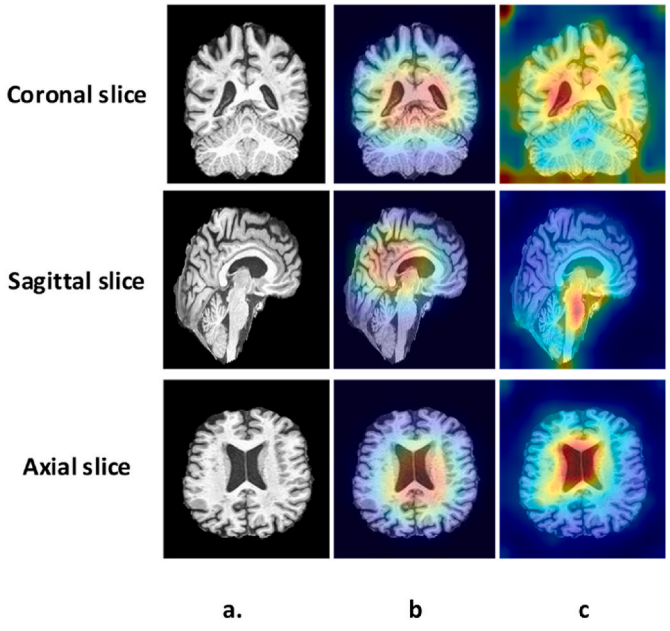


Fig. 9. To demonstrate the advantages of our suggested approach, attention and convolution maps are utilized to highlight a particular area of the different image slice that provides information significant to the diagnosis of AD. a) Input AD sMRI images. b) Convolution maps for AD sMRI images. c) Attention-maps obtained through LMHSA for AD sMRI images.

6. Discussion

For effective early support and treatment, precise diagnosis of Alzheimer's disease is essential. Researchers have explored computer-based systems for early detection, with CNN-based image recognition widely used in medical diagnosis. However, designing an effective deep learning model for desirable outcomes can be challenging.

In this study, our focus was on enhancing the accuracy of sMRI image classification for AD using convolution-attention features while minimizing parameters compared to the original ViT models. Previous models sought to improve classification performance by increasing image size, attention blocks, and network complexity. However, vision transformer models often faced challenges with huge parameters and computational costs. We proposed a redesigned CMT network to lower parameters and computational costs for brain Alzheimer's classifications. Our network comprises three-layer types: convolution, attention, and local feedforward neural networks. It applies the convolution transformer architecture in AD recognition, providing a new perspective. Experimental findings show our proposed model is more effective than well-known backbone networks in identifying Alzheimer's disorders.

The proposed architecture differs from the baseline ViT in several ways. Firstly, we introduce an improved variant of the transformer block with integrated IRU, LMHSA, and LFFN block to enhance local information on brain sMRI. Secondly, the features from the first stage have a higher resolution compared to ViT, maintaining a resolution of $H/4 \times W/4$.Thirdly, the method adopts a stage-wise architecture design, using four convolutional layers to gradually reduce resolution and increase dimensionality, allowing for the extraction of multi-scale features and reducing computational burden. To achieve this, we incorporated one standard convolutional layer with stride two as convolution stem and four CViT blocks with 4 output channels of size 64, 128, 256 and 512 respectively. Each block consists of a depth-wise convolution preceding a point-wise convolution of size 1x1 with groups of similar size to the number of channels. The ReLU activation function and Batch Normalization were applied after each convolution process in the block. A skip connection-inspired inverted residual convolutional element was also used in the model. The convolution filters' dimensions were set to 3x3, and the number of blocks was set fixed as 3 for each block in the model with stem width 32, number of heads as 1,2,4,8, and reduction rate 8,4,2,1 respectively along with expansion ratio as 4. This promoted feature reuse and decreased the model's parameter count by ensuring that uniform weights distribution across multiple clusters of pixels in an sMRI images. Lastly, the proposed model replaces the class token used in ViT with average pooling for better classification results and incorporates a simple scaling strategy to preserve important features in

sMRI data. These architectural differences outperform ViT in accuracy and computational efficiency, achieving a top accuracy of 94.31% on the ADNI dataset with fewer FLOPs compared to ViT-based models.

In the ablation study, we explored the impact of number of block size to gain a better understanding of the efficacy of our proposed method. We observed that an AdamW with GC allowed our method to train the model more smoothly and efficiently as shown in Fig. 8(d) and performance comparison in Fig. 10. To reduce the model complexity, we introduced the IRU, and LFFN with lesser number of attention block in our model to achieved better accuracy for Alzheimer's dataset. By introducing IRU and LFFN inside architecture we get excellent efficacy for AD diagnosis (with reduced parameters 15.9 M and 2.2 B FLOPs) as compared to previous models for the same 224 × 224 input size.

The main contribution of this article is the development of the hybrid convolution-attention model, an advanced deep learning architecture for efficient AD diagnosis using structural MRI data. The model incorporates representation inverted residual unit, lightweight multi-head self-attention, local feedforward neural network with additional depth-wise convolution and classifier modeling into a unified framework, addressing limitations of current computational and memory costs. We also incorporate the GC on optimizer for consistency and smooth training process to train the high dimensional sMRI data demonstrating the superiority of the proposed model in algorithm performance and various medical metrics.

The practical implication of this research is that this model provides an efficient tool for diagnosing AD using sMRI data. With high accuracy in classifying AD and MCI compared to other methods, the proposed model can potentially assist medical professionals in accurately identifying and monitoring AD patients. By leveraging the power of deep learning and advanced data analysis techniques, this model offers a valuable contribution to the field of neuroimaging and can aid in early detection and intervention for AD.

Additionally, we investigated the brain regions that significantly influenced the predictions of our suggested technique. We identified key regions with the highest attention scores as medial-occipital gyrus, superior frontal gyrus, third ventricle, putamen, thalamus, medial frontal gyrus, superior temporal gyrus, frontal lobe, hippocampus, amygdala, and occipital regions. Fig. 8 illustrates examples of sMRI scans for AD cases. The thalamus, known as the primary relay for sensorimotor information in the brain, is believed to play a crucial role in early-stage memory processing affected by AD [71]. Numerous cognitive processes, including attention, spatial awareness, and long-term memory, are mediated by the medial frontal area [72]. In AD, there is a volume loss in the occipital area, which controls visual perception of things like color, shape, and motion [42,73]. These results point to useful areas for future research, and which can be more valuable for identifying regions of interest in clinical application (Figs. 9 and 11). Our model can be valuable for analysis of different brain imaging modalities as well as other medical images analysis.

6.1. Limitations and future works

Although our proposed method demonstrates favorable results in diagnosing AD, there exist certain limitations that require further enhancement in future studies. In the following sections, we outline these limitations and propose potential solutions to address them. Firstly, our current method utilizes 2D scan which may result in missing global anatomical information from other brain regions, that could affect the accuracy of our predictions. To overcome this limitation, future work should focus on incorporating 3D architectures and segmentation techniques to accurately identify and include these regions in the analysis. Secondly, the model's performance needs to be validated on a wider range of datasets to assess its generalization capabilities. Currently, the evaluation is based on a specific ADNI dataset, and it is important to ensure that the model performs consistently well on different datasets with varying characteristics, such as imaging techniques, demographics, and disease prevalence. In contrast to using a single MRI modality, utilization of multimodal imaging data has the potential to provide a richer set of information to improved classification performance. Hence, future studies will focus on incorporating multimodal brain data, including functional MRI (fMRI), Positron Emission Tomography (PET), and other clinical features. By integrating multiple imaging modalities, researchers aim to enhance the discriminative power of the models and achieve even better performance and expected to provide a more comprehensive understanding in the classification of AD related brain conditions.

7. Conclusion

In this research article, we introduced a highly efficient model that utilizes the convolution self-attention mechanism for classifying MRI data related to AD. By employing the convolution and self-attention mechanism, we were able to significantly reduce computational complexity, allowing for the application of self-attention to high-

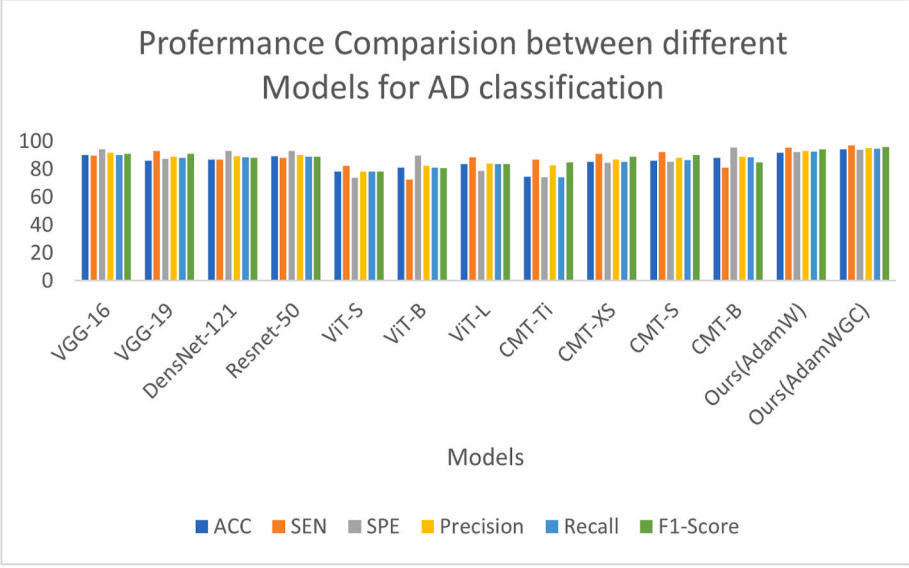
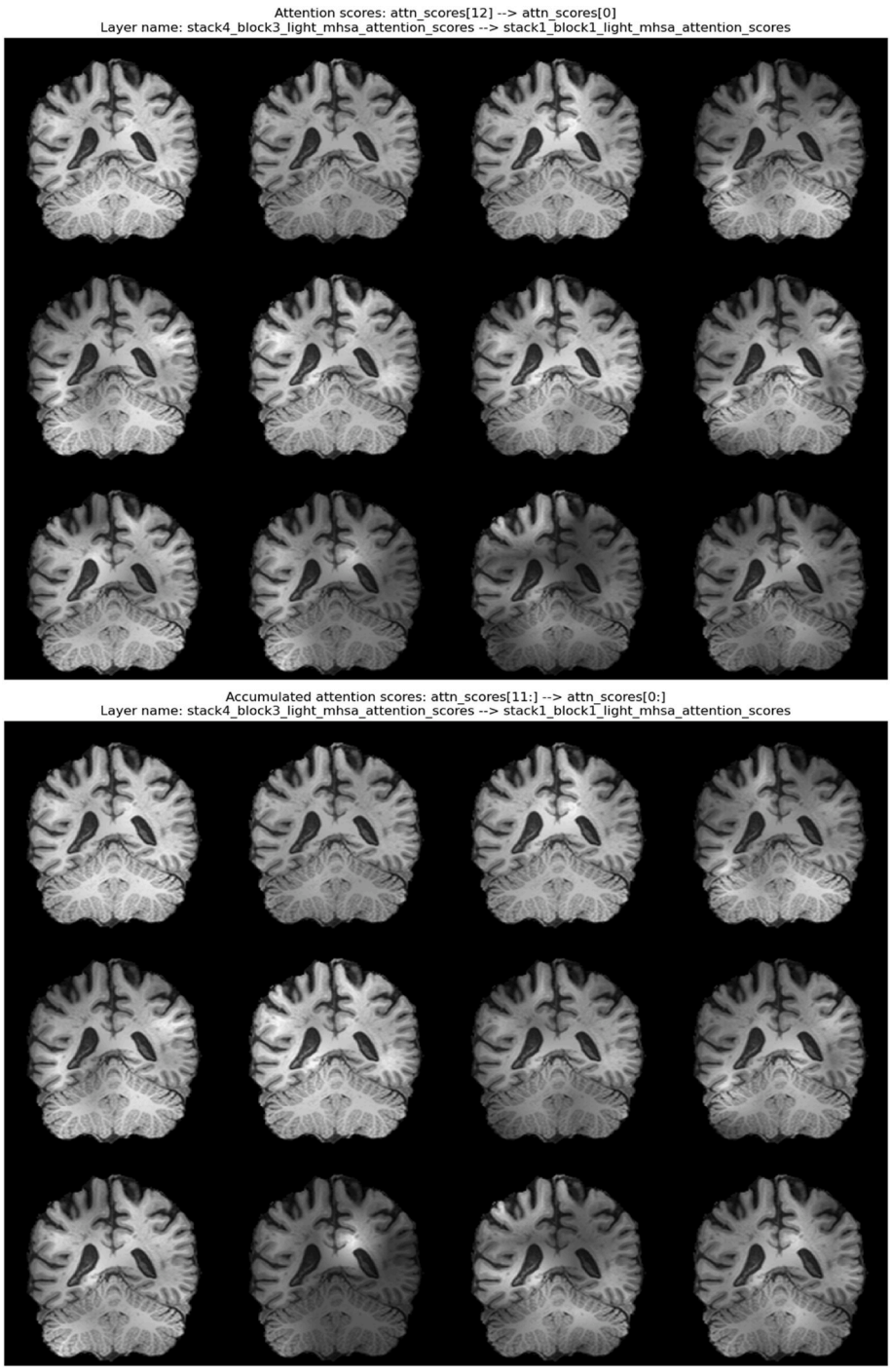


Fig. 10. Visualizing models comparison in bar graph including different classification parameters for AD/MCI/HC multi-class classification task.



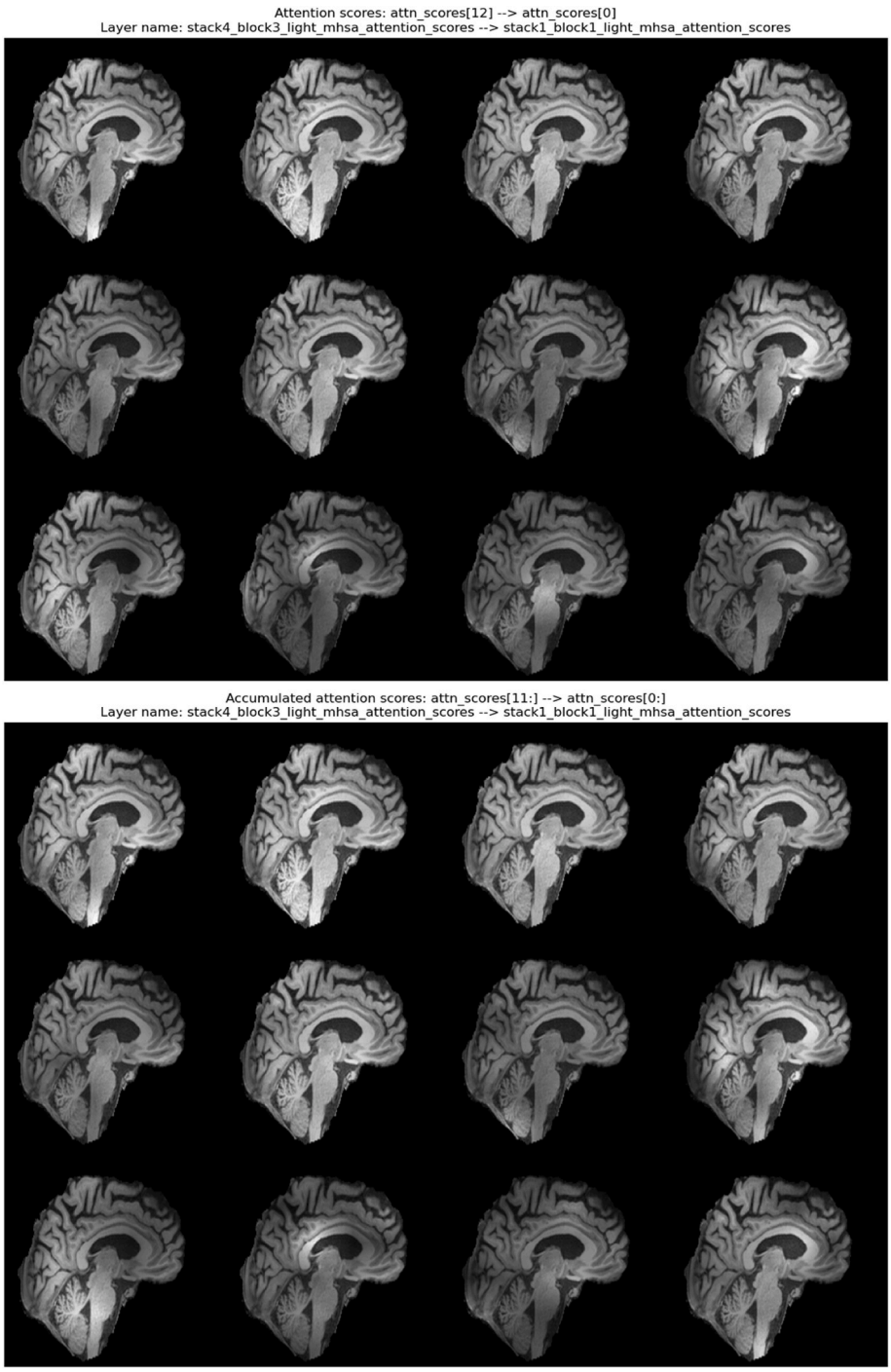
(a)

Fig. 11. Visualizing model attention score maps on each self-attention head for the Alzheimer's brain's sMRI. Highlight area of the brain has higher attention scores: (a) Coronal slice (b) Sagittal slice (c) Axial slice.

dimensional sMRI data. Additionally, our model incorporates an inverted residual unit layer that performs feature down-sampling and retaining important features while minimizing computational costs. When evaluated on the ADNI dataset, the proposed optimized architecture achieved impressive classification accuracies of 94.31% for multiclass classification, 95.37% for AD/HC and 92.15 % for MCI/HC classification, surpassing the performance of other state-of-the-art methods. A series of comparison studies on baseline models further demonstrated the efficient learning capabilities of the optimized architecture when applied to brain sMRI data. This research introduces new insights and methodologies for leveraging deep learning in the study of Alzheimer's diseases.

Funding

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2021R1I1A3050703, in part



(b)

Fig. 11. (*continued*).

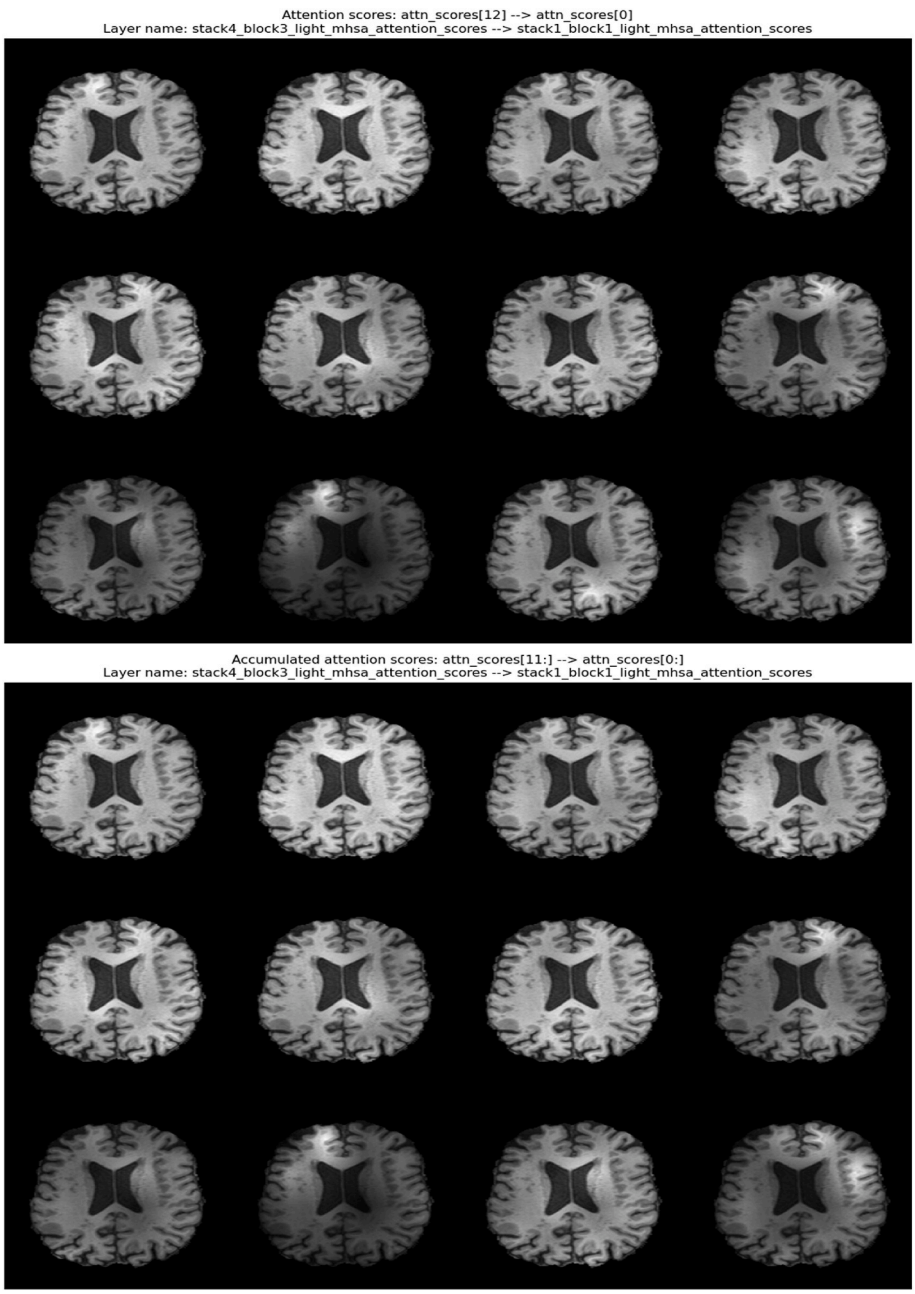
by the BrainKorea21Four Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 4299990114316, in part by the Alzheimer's Disease Neuroimaging Initiative (ADNI) funded by the National Institutes of Health under Grant U01 AG024904, and in part by the Department of Defense ADNI under Award W81XWH-12-2-0012.

Data availability statement

The dataset used in this study were acquired from ADNI homepage, which is available freely for all researcher and scientist for experiments on Alzheimer's disease and can be easily downloaded from ADNI websites: http://adni.loni.usc.edu/about/contact-us/.

CRediT authorship contribution statement

Uttam Khatri: Writing – original draft, Visualization, Validation, Software, Conceptualization. **Goo-Rak Kwon:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation.



(c)

Fig. 11. (*continued*).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors declare that data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The patients/participant provided their written informed consent to participate in this study. As such, the funder, and the investigators within ADNI contributed to the data collection, but did not participate in analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

The design and implementation of ADNI (Alzheimer's Disease Neuroimaging Initiative) 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, as well as 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. and its affiliate Genentech Inc., Fujirebio, and GE Healthcare. ADNI's clinical centers in Canada are funded by The Canadian Institutes of Health Research. 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 at the College of Southern California. The ADNI data is disseminated by the Laboratory for Neuro Imaging at the College of Southern California. Correspondence should be addressed to GR-K, grkwon@chosun.ac.kr.

References

[1] L.-K. Huang, S.-P. Chao, C.-J. Hu, Clinical trials of new drugs for Alzheimer disease, J. Biomed. Sci. 27 (1) (Jan. 2020) 18.
[2] "World Alzheimer Report 2021: Journey through the Diagnosis of Dementia".
[3] alz.13016, 2023 Alzheimer's Disease Facts and Figures, Alzheimers Dement., Mar.2023.
[4] R.A. Sperling, et al., Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement 7 (3) (May 2011) 280–292.
[5] H. Braak, E. Braak, Neuropathological stageing of Alzheimer-related changes, Acta Neuropathol. 82 (4) (Sep. 1991) 239–259.
[6] S. Palmqvist, et al., Detailed comparison of amyloid PET and CSF biomarkers for identifying early Alzheimer disease, Neurology 85 (14) (Oct. 2015) 1240–1249.
[7] D.P. Veitch, et al., Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's Disease Neuroimaging Initiative, Alzheimers Dement 15 (1) (Jan. 2019) 106–152.
[8] J. Zhu, et al., Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis, Comput. Biol. Med. 147 (Aug. 2022) 105737.
[9] S. Lahmiri, A. Shmuel, Performance of machine learning methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer's disease, Biomed. Signal Process Control 52 (Jul. 2019) 414–419.
[10] E. Hosseini-Asl, R. Keynton, A. El-Baz, Alzheimer's disease diagnostics by adaptation of 3D convolutional network, in: 2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, pp. 126–130.
[11] M.I. Razzak, S. Naz, A. Zaib, Deep learning for medical image processing: overview, challenges and the future, in: N. Dey, A.S. Ashour, S. Borra (Eds.), Classification in BioApps: Automation of Decision Making, Lecture Notes in Computational Vision and Biomechanics, Springer International Publishing, Cham, 2018, pp. 323–350.
[12] F.U.R. Faisal, G.-R. Kwon, Automated detection of Alzheimer's disease and mild cognitive impairment using whole brain MRI, IEEE Access 10 (2022) 65055–65066.
[13] M. Liu, J. Zhang, E. Adeli, D. Shen, Landmark-based deep multi-instance learning for brain disease diagnosis, Med. Image Anal. 43 (Jan. 2018) 157–168.
[14] D. Jin, et al., Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Apr. 2019, pp. 1047–1051.
[15] X. Xing, et al., Dynamic image for 3D MRI image Alzheimer's disease classification, in: A. Bartoli, A. Fusiello (Eds.), Computer Vision – ECCV 2020 Workshops, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 355–364.
[16] A. Vaswani, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017 [Online]. (Accessed 24 April 2023).
[17] A. Dosovitskiy, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021 arXiv, Jun. 03.
[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 10347–10357.
[19] X. Chu, Z. Tian, B. Zhang, X. Wang, C. Shen, Conditional positional Encodings for vision transformers, arXiv (Feb. 12, 2023).
[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End object detection with transformers, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 213–229.
[21] S. Zheng, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890. (Accessed 24 April 2023).
[22] Z. Liu, et al., Swin transformer: hierarchical vision transformer using shifted windows, in: Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022. (Accessed 24 April 2023).
[23] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in transformer, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2021, pp. 15908–15919. (Accessed 24 April 2023).
[24] S. Wang, B.Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-Attention with Linear Complexity, arXiv, Jun. 14, 2020.
[25] J. Guo, et al., CMT: Convolutional Neural Networks Meet Vision Transformers, arXiv, Jun. 14, 2022.
[26] F. Altay, G.R. Sánchez, Y. James, S.V. Faraone, S. Velipasalar, A. Salekin, Preclinical stage Alzheimer's disease detection using magnetic resonance image scans, Proc. AAAI Conf. Artif. Intell. 35 (17) (May 2021) 17.
[27] E. Jun, S. Jeong, D.-W. Heo, H.-I. Suk, Medical Transformer: Universal Brain Encoder for 3D MRI Analysis, arXiv, Apr. 28, 2021.
[28] H. Yong, J. Huang, X. Hua, L. Zhang, Gradient Centralization: A New Optimization Technique for Deep Neural Networks, arXiv, Apr. 07, 2020.
[29] J. Dukart, et al., Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI, Psychiatry Res 212 (3) (Jun. 2013) 230–236.
[30] F.J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, D. Castillo-Barnes, Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders, IEEE J. Biomed. Health Inform. 24 (1) (Jan. 2020) 17–26.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv, Dec. 10, 2015.
[32] H. Li, M. Habes, D.A. Wolk, Y. Fan, A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data, Alzheimers Dement 15 (8) (Aug. 2019) 1059–1070.
[33] Z. Cui, Z. Gao, J. Leng, T. Zhang, P. Quan, W. Zhao, Alzheimer's disease diagnosis using enhanced inception network based on brain magnetic resonance image, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov. 2019, pp. 2324–2330.
[34] M. Nguyen, N. Sun, D.C. Alexander, J. Feng, B.T.T. Yeo, Modeling Alzheimer's disease progression using deep recurrent neural networks, in: 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), Jun. 2018, pp. 1–4.
[35] X. Zhao, F. Zhou, L. Ou-Yang, T. Wang, B. Lei, Graph convolutional network analysis for mild cognitive impairment prediction, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Apr. 2019, pp. 1598–1601.
[36] M. Hon, N.M. Khan, Towards Alzheimer's disease classification through transfer learning, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov. 2017, pp. 1166–1169.
[37] C. Matsoukas, J.F. Haslum, M. Söderberg, K. Smith, Is it Time to Replace CNNs with Transformers for Medical Images? arXiv, Aug. 20, 2021.
[38] S. Sarraf, A. Sarraf, D.D. DeSouza, J.A.E. Anderson, M. Kabia, The Alzheimer's disease neuroimaging initiative, "OViTAD: optimized vision transformer to predict various stages of Alzheimer's disease using resting-state fMRI and structural MRI data,", Brain Sci. 13 (2) (Feb. 2023) 2.
[39] J. Li, et al., Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios, arXiv, Aug. 16, 2022. (Accessed 29 July 2023).
[40] Z. Zhang, F. Khalvati, Introducing Vision Transformer for Alzheimer's Disease Classification Task with 3D Input, arXiv, Oct. 03, 2022.
[41] H. Shin, S. Jeon, Y. Seol, S. Kim, D. Kang, Vision transformer approach for classification of Alzheimer's disease using 18F-florbetaben brain images, Appl. Sci. 13 (6) (Jan. 2023) 6.
[42] G.M. Hoang, U.-H. Kim, J.G. Kim, Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI, Front. Aging Neurosci. 15 (2023). (Accessed 1 July 2023).
[43] O.N. Manzari, H. Ahmadabadi, H. Kashiani, S.B. Shokouhi, A. Ayatollahi, MedViT: a robust vision transformer for generalized medical image classification, Comput. Biol. Med. 157 (May 2023) 106791.
[44] 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), Mar. 2022, pp. 1–5.
[45] A.G. Howard, et al., MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv, Apr. 16, 2017.
[46] J. Liu, M. Li, Y. Luo, S. Yang, W. Li, Y. Bi, Alzheimer's disease detection using depthwise separable convolutional neural networks, Comput. Methods Programs Biomed. 203 (May 2021) 106032.
[47] R. Kadri, B. Bouaziz, M. Tmar, F. Gargouri, Depthwise separable convolution ResNet with attention mechanism for Alzheimer's detection, in: 2022 International Conference on Technology Innovations for Healthcare (ICTIH), Sep. 2022, pp. 47–52.
[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, arXiv (Mar. 21, 2019).
[49] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines"..
[50] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, L. Sagun, ConViT: improving vision transformers with soft convolutional inductive biases, J. Stat. Mech. Theory Exp. 2022 (11) (Nov. 2022) 114005.
[51] D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (GELUs), arXiv, Jun. 05, 2023.
[52] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer Normalization, arXiv, Jul. 21, 2016.
[53] C.R. Jack Jr., et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging 27 (4) (2008) 685–691.
[54] SPM - Statistical Parametric Mapping. Accessed: January. 11, 2023.https://www.fil.ion.ucl.ac.uk/spm/.
[55] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 2261–2269.
[56] ADNI | Alzheimer's Disease Neuroimaging Initiative Accessed: June. 15, 2023. https://adni.loni.usc.edu/.
[57] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-Cam, Visual explanations from deep networks via gradient-based localization, in: Presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626. (Accessed 28 June 2023).
[58] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv (Apr. 10, 2015).

[59] J. Liu, M. Li, W. Lan, F.-X. Wu, Y. Pan, J. Wang, Classification of Alzheimer's disease using whole brain hierarchical network, IEEE/ACM Trans. Comput. Biol. Bioinform. 15 (2) (Mar. 2018) 624–632.
[60] Z. Sun, Y. Qiao, B.P.F. Lelieveldt, M. Staring, Integrating spatial-anatomical regularization and structure sparsity into SVM: improving interpretation of Alzheimer's disease classification, Neuroimage 178 (Sep. 2018) 445–460.
[61] C. Lian, M. Liu, J. Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (Apr. 2020) 880–893.
[62] J. Li, Y. Wei, C. Wang, Q. Hu, Y. Liu, L. Xu, 3-D CNN-based multichannel contrastive learning for Alzheimer's disease automatic diagnosis, IEEE Trans. Instrum. Meas. 71 (1–11) (2022).
[63] M. Liu, et al., A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, Neuroimage 208 (Mar. 2020) 116459.
[64] W. Kang, L. Lin, B. Zhang, X. Shen, S. Wu, Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis, Comput. Biol. Med. 136 (Sep. 2021) 104678.
[65] J. Xin, A. Wang, R. Guo, W. Liu, X. Tang, CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI, Biomed. Signal Process Control 86 (Sep. 2023) 105189.
[66] 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) (Sep. 2021) 2354–2366.
[67] Z. Hu, Y. Li, Z. Wang, S. Zhang, W. Hou, Conv-Swinformer: integration of CNN and shift window attention for Alzheimer's disease classification, Comput. Biol. Med. 164 (Sep. 2023) 107304.
[68] leondgarse, Keras_cv_attention_models, Nov. 13, 2023. https://github.com/leondgarse/keras_cv_attention_models. (Accessed 14 November 2023).
[69] W. Shao, Y. Peng, C. Zu, M. Wang, D. Zhang, Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease, Comput. Med. Imaging Graph. 80 (Mar. 2020) 101663.
[70] E.E. Bron, et al., Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease, NeuroImage Clin 31 (Jan. 2021) 102712.
[71] J.P. Aggleton, A.J.D. Nelson, Why do lesions in the rodent anterior thalamic nuclei cause such severe spatial deficits? Neurosci. Biobehav. Rev. 54 (Jul. 2015) 131–144.
[72] Role of the Medial Prefrontal Cortex in Cognition, Ageing and Dementia | Brain Communications | Oxford Academic.".
[73] A. Brewer, B. Barton, Visual cortex in aging and Alzheimer's disease: changes in visual field maps and population receptive fields, Front. Psychol. 5 (2014).