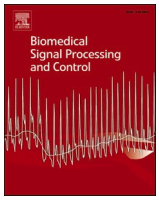
Biomedical Signal Processing and Control 86 (2023) 105189

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI

Jiaming Xin [a], Ancong Wang [a], Rui Guo [b], Weifeng Liu [a,*], Xiaoying Tang [a,*]

[a] School of Life Science, Beijing Institute of Technology, Beijing 100081, China
[b] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Alzheimer's disease
Image classification
Computational complexity
Magnetic resonance imaging
Deep learning

ABSTRACT

Alzheimer's disease (AD) is a primary cause of dementia. Its early diagnosis is crucial to delay the progression of the disease. So far, many computer aided diagnosis (CAD) methods that combined deep learning algorithms and structural MRI have achieved encouraging results. To improve the AD diagnosis performance, more and more models are based on 3D algorithms, which make the training and deployment of these methods unaffordable. In this study, a CNN and swin-transformer based efficient model, Efficient Conv-Swin Net (ECSnet), was developed. In this model: (1) a 2.5D-subject method and two-stream structure are used to help the model to encode 3D information to 2D feature maps; (2) convolution blocks are applied in the early stages of the transformer-based backbone network to improve the generalization ability; (3) a series of lightweight approaches are applied to reduce the parameters and computational cost of the model to enable the model to train and infer efficiently. Due to the lack of multi-center data and the differences between test sets, it is difficult to make a fair comparison between the previous methods. Our model was trained on the ADNI dataset and evaluated on an independent test set from AIBL. After being lightened, our proposed method showed no performance degradation on both ADNI and AIBL compared to models such as swin-T tiny. The ECSnet achieved 92.8% balance accuracy and 91.1% sensitivity on the AIBL, which are better than those of previous works while the model is more efficient than those 3D methods.

1. Introduction

Alzheimer's disease (AD) is one of the most prevalent progressive neurological diseases and a primary cause of dementia [1]. About 44 million people worldwide are diagnosed with AD each year, and this number is expected to grow to 131.5 million by 2050[2]. The progression of AD will result in the gradual deterioration, impairment of memory and cognitive functions, eventually leading to irreversible neuron damage in the brain and impairment of the ability of daily living [3]. As a result, AD has become a public health issue worldwide. Yet there is still no treatment proven to be effective in preventing the progression of AD, early diagnosis of AD is considered necessary to delay the progression of cognitive impairment and improve the quality of life of patients [4]. Computer aided diagnostic (CAD) can automatically diagnose diseases through developed algorithms with various medical data [5–7], and reduce doctors' involvement in the diagnosis pipeline, making the diagnoses more efficient. With the increase of AD patients, the need for effective CAD is also increasing.

Imaging techniques that can be used to observe the progression of AD in patients' brain include structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), positron emission tomography (PET) and diffusion tensor imaging (DTI) etc. Although fMRI, PET and DTI can detect changes in neural activity or lesions in patient's brain effectively [8–13], sMRI remains the most prevalent imaging technique due to the higher cost, complex detection process and longer scanning time of those techniques. sMRI are sensitive to morphological changes caused by brain atrophy in gray matter (GM) and white matter (WM) structures. With the development of machine learning techniques, many AD diagnosis algorithms with sMRI have been proposed for CAD. Traditional machine learning-based methods split brain sMRI images into different regions of interest (ROIs) by structural template and apply specific algorithms to extract structural variations or volume changes within regions considered relevant to the progression of AD [14,15], then a classification model will be trained by feeding these features into an ML model such as random forest or support vector machine.

Deep learning (DL) methods are gradually becoming mainstream in

* Corresponding authors.
E-mail addresses: breeze@bit.edu.cn (W. Liu), xiaoying@bit.edu.cn (X. Tang).

https://doi.org/10.1016/j.bspc.2023.105189
Received 7 January 2023; Received in revised form 26 May 2023; Accepted 21 June 2023
Available online 30 June 2023

the development of CAD, so that the classification algorithms represented by convolutional neural network (CNN) have been widely studied in the AD diagnosis task [16–18]. Compared with traditional machine learning methods, CNN and other deep learning models are mostly data-driven models that can automatically mine the latent structural features in the images [19], and the relatively large amount of sMRI data serves as a good data support for deep learning algorithms. Transformer-based models such as ViT [17] have been gradually applied in the CAD studies such as AD diagnosis for their excellent performance, where such models apply the self-attention mechanism to emphasize the dependencies between long sequences instead of learning a global representation. However, the main disadvantage of visional transformer models is that they require a large amount of data for training.

The use of deep learning algorithms instead of feature selection methods based on prior knowledge to extract sMRI image features is considered to better avoid the loss of key features due to manual selection. But in medical images field, deep learning models often face the problem that the size of the available dataset is similar to or even smaller than the number of features extracted by the model, also known as the "curse of dimensionality"[21,22]. When the dataset is relatively small and the task is difficult, the deep learning models are prone to overfitting rapidly during training, resulting in a poor performance on the unseen data. In order to achieve better AD diagnosis performance, many works in recent years have chosen to train an end-to-end 3D model to extract features or develop a method which combines a risk region identification model with a classification model. Those large models need to be trained and deployed on high-performance servers, which is inconvenient for research and clinical application, and the larger model capacity also leads to impairment of generalization ability.

Although deep learning methods have achieved many encouraging results, the lack of additional multicenter test data in many previous studies makes it difficult to effectively assess the model performance; and the differences in test sets due to various data screening criteria make the comparisons between different methods inconvenient and unfair. The problems limit the clinical application of deep learning methods, so it is necessary to compare models by using multi-center data as well as similar test sets.

To solve the above problems, an efficient deep learning network, Efficient Conv-Swin Net (ECSNet) based on CNN and swin-transformer [35] is proposed in this paper and the performance of the model was evaluated on an independent test set. Firstly, the model applies an early-stage CNN method and two-stream network structure to encode the 3D input images into 2D feature maps in a 2.5D-subject approach, allowing the model to encode as much information as possible to help the following swin-transformer to further extract features. We then apply a series of lightweight approaches in the convolution blocks and swin-transformer blocks to reduce parameters and computational cost, which also reduce the model capacity and make it efficient in training and inference phases while alleviating the overfitting caused by the lack of data.

2. Related work

In this section, we will briefly introduce the previous automatic AD diagnosis methods based on sMRI and deep learning in recent years, and review the lightweight techniques for convolution and attention mechanisms.

2.1. Deep learning models to diagnosis Alzheimer's disease with sMRI

Following the review of the deep learning methods for AD diagnosis [23], we divide previous deep learning methods into four categories: 2D slice-level, 2D subject-level, 3D patch-level (3D ROI-level) and 3D subject-level. The 2D slice-level, 2D subject-level and 3D subject-level models are trained on relatively complete or partially cropped sMRI images, and the 3D patch-level methods select the ROI that is considered effective for the task based on prior knowledge or an identification algorithm from whole-brain sMRI to perform the diagnosis.

For 2D slice-level and 2D subject-level methods, many of those previous works train their models on large-size natural image datasets such as ImageNet and transfer the models to the target medical datasets for fine-tuning, expecting the cross-domain knowledge can alleviate the overfitting; or train a self-supervised model like Auto Encoder or Generative Adversarial Networks (GAN) on target medical image datasets, then transferring the encoder or discriminator to the classification task [24,25]. Kang et al. [26] input all coronal-plane 2D slices obtained from 3D sMRI scans into DCGAN for unsupervised training, then transfer the discriminator to the AD/NC (normal control) task and select 2D slices at specific positions to feed into the network for fine-tuning, finally the model achieved accuracy of 90.36% and AUC of 0.897.

Because 3D sMRI images contain more information about brain structural changes, models that extract features via 3D algorithm are considered better in identifying AD-related lesions or atrophy, so most studies in recent years have chosen to use 3D patch-level or 3D subject-level images with 3D DL algorithms. Zhu et al. [27] split the original 3D sMRI into non-overlapping 3D patches, and made a group comparison on AD group and NC group of patch-level features at one patch location respectively using a *t*-test, considered that the patches with larger t-values between the two groups are more likely to include AD-related brain changes. Therefore, some patches with larger t-values were selected as training data, the final AD/NC classification accuracy on the ADNI reached 92.4%. Lian et al. [28] trained an FCN to generate disease risk maps representing the AD-related regions concerned by the model, and selected 36 high-risk 3D patches for the classification model, which achieved accuracy of 91.9% and 89.8% on ADNI-2 and AIBL respectively. Hedayati et al. [29] trained an unsupervised convolutional auto encoder to extract features from 3D images registered by different templates, and then got the classification results by using another one CNN. Finally, the method achieved accuracy of 95% on ADNI.

While the transfer learning, self-supervised learning and 3D models have shown potential performance, their common disadvantage is that the computational overhead is huge and the training often takes a long time. At the same time, there are also some problems such as models pre-trained on natural images may not match classification tasks on medical images, and 3D models are prone to overfit due to the large model capacity [18]. These problems limit the development and application of the models, making it necessary to design an efficient end-to-end model.

2.2. Efficient convolution & attention mechanism

The CNN models extract latent information through the stacked convolutional layers. Stacking a large number of convolution layers to make the model deeper is an effective way to improve the performance, but that means more parameters and computational overhead. In order to improve the training and inference efficiency with less impairment of the performance, many lightweight methods have been proposed. AlexNet proposed the group convolution [30], which reduces the computational complexity by grouping feature maps and convolution kernels. The depth-wise separable convolution based on group convolution proposed by Howard et al. [31] further reduces parameters and computational complexity by combining deep-wise convolution and point-wise convolution while maintaining good performance. DenseNet and GhostNet [32,33] consider that deep learning models extract many redundant feature maps in forward propagation. For this reason, DenseNet reduces feature maps obtained in convolution and introduces feature reuse which combines the feature maps from the early layer with those of the deep layers; GhostNet also reduces the feature maps obtained by convolution operation, and linearly projects the feature maps to get the "redundant" parts.

ViT proposed by Dosovitskiy et al. [20] splits the images into patches and embeds them into token vectors, thus introducing the transformer structure from NLP into the CV task. Different from traditional CNN

Table 1
The demographic information, including datasets, groups, gender, age, MMSE scores and ApoE4.

Dataset	Research group	Gender (Male/ Female)	Age (Mean ± std)	MMSE (Mean ± std)	ApoE4 positive(%)
ADNI	AD(n = 336)	184/152	75.12 ± 8.09	23.15 ± 2.09	68.75
	NC(n = 529)	230/299	73.64 ± 6.50	29.16 ± 1.07	29.49
AIBL	AD(n = 79)	33/46	73.34 ± 7.77	20.42 ± 5.46	68.35
	NC(n = 449)	184/265	72.47 ± 6.21	28.73 ± 7.21	26.16

models, visional transformer models extract image features without convolution operations, but apply self-attention mechanisms to encode connections within different patches [34]. The standard self-attention mechanism in ViT focuses on the regions that are crucial for the task by weighting the feature maps, for which three feature matrices are obtained by linearly projection and are performed matrix multiplication between each other to get the attention maps. The matrix multiplications make the computational complexity $O((hw)^2)$ where $h \times w$ is the image size. To reduce the computational overhead of the self-attention mechanism, some methods apply a local attention approach or an attention mechanism with less computational complexity to replace the original global attention[35,36]. Swin-transformer[35] applies a hierarchical structure to reduce the size of the feature map and converge the information, and the local self-attention within no-overlapped local window makes the computational complexity linear to the image size of $h \times w$; separable self-attention[37] replaces the matrix query in attention mechanism by a vector, and simplifies matrix multiplications as vector element-wise multiplication operation, also resulting in a linear complexity of the image size of $h \times w$.

3. Method

Firstly, this section presents the sources of the data used in this study, the preprocessing pipeline and the proposed 2.5D method; then we will describe the details of the proposed ECSnet, including the overall structure and the lightweight approaches for convolution and swin-transformer blocks.

3.1. Materials

All data used in this study were obtained from Alzheimer's Disease Neuroimaging Initiative [38] (ADNI, https://adni.loni.usc.edu) and the Australian Imaging Biomarkers & Lifestyle Flagship Study of Aging [39] (AIBL, https://aibl.csiro.au).

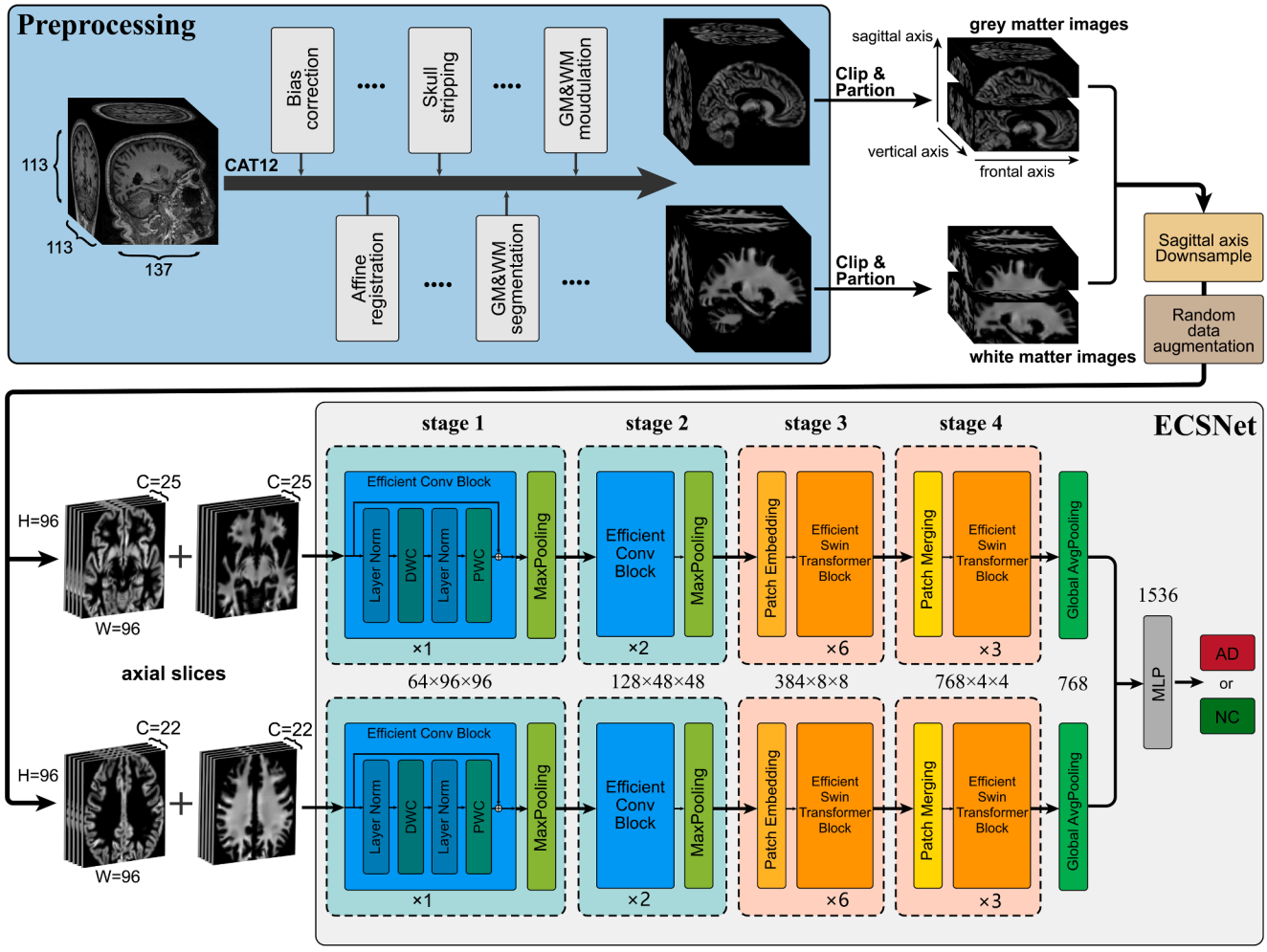


Fig. 1. Schematic diagram of the proposed AD diagnosis pipeline. The upper left part of the figure shows main operations of the sMRI preprocessing pipeline; and the 2.5D method and random data augmentation are shown in upper right and lower left part; Our ECSnet (lower right) consists of 2 backbone networks, each backbone contains two CNN stages and two swin-transformer stages, and the feature vectors extracted by the two-stream model are combined and input into the single-layer MLP to get the diagnosis result.

ADNI provides a public AD dataset for researchers worldwide to explore the early diagnosis methods and corresponding biomarkers of AD. There is a large amount of data from over 2000 subjects, including longitudinal sMRI scans, neuropsychological testing and biomarkers, etc. Most previous studies extracted and screened their own subsets from origin ADNI dataset with different criteria and do not provide reproducibility details, making it difficult to fairly compare the performance of different methods. When filtering the data, we excluded subjects whose diagnosis results repeatedly converted between AD and NC and the selected individuals can be divided into 529 NC and 336 CE subjects. Finally, we obtained a total of 865 1.5 T/3T 3D T1-weighted sMRI scans of different subjects from ADNI-1, ADNI-2 or ADNI-3 at their first visit.

AIBL was launched in 2006 and is the largest such study in Australia. The dataset contains longitudinal sMRI scans, neuropsychological testing, genetic information and lifestyle information of more than 800 subjects. In AIBL, we also excluded subjects with repeated AD/NC conversion, and the sMRI scans at selected subjects' first visit were obtained. Finally, there are 79 AD subjects and 449 NC subjects from AIBL. The demographic details of these subjects from the ADNI and AIBL datasets is shown in Table 1.

We didn't use no-imaging information such as demographic information or scores of neuropsychological scales in our study, which were used in some previous research [40,41], but only based on sMRI scans. Some previous studies used scores obtained from neuropsychological scales such as Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) as classification features in their models. We consider that applying these scores that relate to the golden standard (group label) in developments of the methods is contrary to the intention of CAD and prone to result in information leakage which is unfair for comparison between different methods.

3.2. Data preprocessing

The sMRI scans acquired from ADNI and AIBL are 3D T1-weighted images in Neuroimaging Informatics Technology Initiative (NIfTI) format. In order to remove non-brain tissue and align the brain to a uniform standard space, the computational anatomy toolbox CAT12 (available at https://www.neuro.uni-jena.de/cat/) was used to preprocess the image data. The main operations in the preprocessing pipeline are: 1. bias correction; 2. alignment with MNI152 template; 3. non-brain tissues removal; 4. segmentation of gray matter (GM), white matter (WM) and cerebrospinal fluid; 5. modulation of GM & WM; 6. Gaussian smoothing, etc. Finally, we obtained standardized 3D GM and WM images with the size of $113 \times 137 \times 113$ voxels and spatial resolution of $1.5 \times 1.5 \times 1.5\ mm^3$.

Patients with AD typically have moderate cortical atrophy in GM, and the reduction of tracts due to lesions in the WM such as myelin sheath injury will reduce the WM volume [42]. Therefore, in this study, we summed the standardized 3D GM and WM images as input images.

3.3. 2.5D subject-level method

As mentioned in Section 2, sMRI-based AD diagnosis models can extract features from 2D image slices or 3D images by different types of algorithms. Models which use 2D slices are prone to lose some task-relevant features due to the manually selected slice positions. Models based on 3D images can extract 3D spatial structure information, but the larger model capacity brought by the large number of parameters in 3D convolution or other 3D algorithms make them prone to overfit on small-size datasets and take a long time to train.

In order to inherit the relative efficiency of 2D models and retain the 3D spatial information, our proposed 2.5D subject-level method inspired by visual transformer models stacks multiple 2D slices into the 2.5D images and partitions them into patches, then the tokens obtained by embedding the patches are input into transformer blocks to extract features. By encoding 3D images through 2D algorithm, during the training process, the model is able to automatically select task-relevant 3D features to encode into 2D feature maps. Specifically, as shown in Fig. 1, in our 2.5D method, when a 3D image (size $= h \times w \times d$) is sliced into 2D slices, the slice plane is selected as the horizontal plane ($h \times w$), and then the 2D slices are stacked along the vertical axis. Because the number of the 2D feature maps extracted by the networks will be limited (e.g., 64 in the first conv block of our backbone network) when we try to ensure the computational efficiency, inputting all axial slices into a single backbone network would restrict the model's feature extraction capability due to the insufficient number of channels. So, in our method we split each GM and WM image into two parts along vertical axis and embed the two parts respectively. In addition, the slices in each part are downsampled along the vertical axis with a 50% sampling ratio.

According to previous pathology studies, lesions in the brains of AD patients are mostly concentrated in the gray matter, such as the gyrus of parietal, frontal and temporal lobe, hippocampus and amygdala, as well as in the white matter such as the corpus callosum [42]. To reduce the irrelevant regions in the images input into the model, and meanwhile, reduce the computation overhead of the model, we cropped the black edges and part of the cerebellar in the input images. Finally, the sizes of two parts of the 2.5D images after downsampling are $h \times w \times c = 96 \times 96 \times 25$ and $h \times w \times c = 96 \times 96 \times 22$.

3.4. Data augmentation

To alleviate overfitting due to the small-size training set, we applied data augmentation for the training set in the training phase. The data augmentation methods used include: 1. Flipping along sagittal plane; 2. Gaussian blur/sharpening; 3. contrast adjustment; 4. adding gaussian noise; 5. brightness adjustment. The above augmentation methods were randomly applied to each 2.5D image before inputting to the model.

3.5. Overall architecture

In recent years, ViT and other transformer-based models have shown great results in CV tasks such as classification and semantic segmentation. But so far, there are few works applying transformer in AD automatic diagnosis. We apply swin-transformer (swin-T) structure in our backbone network to improve the performance by introducing the self-attention mechanism. Since there is almost no convolution layer in the transformer-based models, the lack of inductive bias of convolution layer and relatively large model capacity make such type of models need to be trained on large datasets to get a good performance, and are more sensitive to hyperparameters such as weight decay. Xiao et al. [43] proposed that applying CNN in the early phase of ViT to encode low-level feature maps is beneficial for the above problems and can help the model converge faster and train more stably.

For the above reasons, as shown in Fig. 1, our proposed ECSnet applies convolution blocks to encode low-level features of input images before patch embedding block of swin-transformer. Since the self-attention operations in transformer are relatively costly compared to CNN, to reduce the size of the feature maps input into transformer layers, we replace the first two of the four stages of the original swin-transformer with convolution blocks. After being encoded and downsampled in CNN stages, the 256-dimensional feature maps with 1/4 size of the original input images are embedded into tokens and fed into swin-transformer blocks for further features extraction. Considered finer granularity helps the model understand the features better, different from the original setting of 4×4, we set the patch size $M^2 = m \times m$ to 3×3, and the window size in window attention is set to 4×4.

In order to get a more efficient model, we lightweighted the convolution blocks and swin-transformer blocks in our model, the modified blocks are called Efficient Conv Block (ECB) and Efficient swin-transformer block (ES-TB), both of which reduce the computational complexity and model capacity.

As with the swin-transformer, we divide the backbone network into

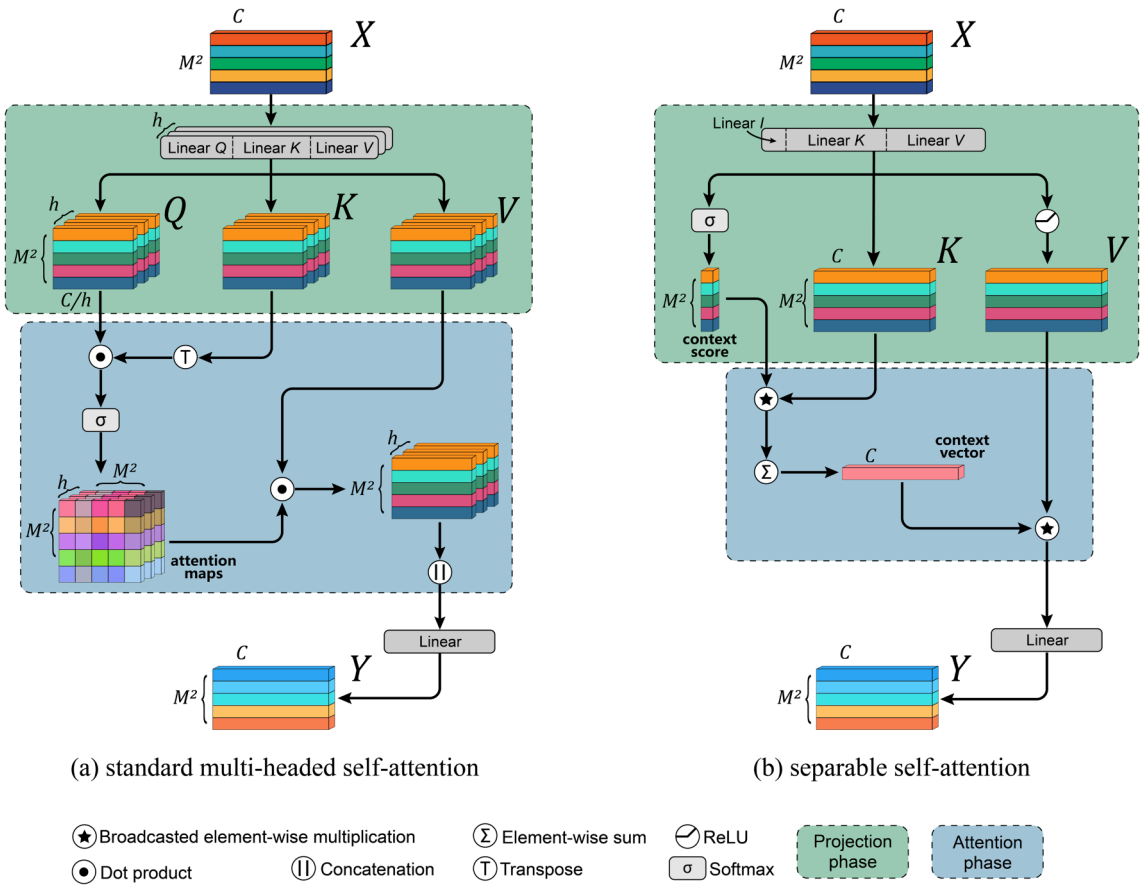


Fig. 2. (a) Standard multi-headed self-attention and (b) separable self-attention. The attention phase of standard multi-head self-attention contains two matrix multiplications which are replaced by element-wise multiplications in separable self-attention. This method makes the computational complexity of self-attention linear to the resolution of input images h × w. Fig. 2 is adapted from Fig. 3 from [37].

four stages (Fig. 1), each with a series of ECBs or ES-TBs, and the number of blocks in the four stages is set to [1,2,6,3]. Maxpooling is used to downsample the feature maps at the end of stage 1 and stage 2, and the same patch merging approach as in the original swin-transformer is used for downsampling before feature maps fed into stage 4.

Finally, we obtain a backbone network which combines CNN and swin-transformer. To correspond to the two parts of the input images based on the 2.5D method mentioned in Section 3.3, the model combines two backbone networks to form a two-stream structure. The features extracted by the two-stream network are obtained as two feature vectors of length 768 after global AveragePooling, and finally concatenated as one feature vector to feed into a single-layer MLP for classification. Softmax is finally performed on the output vector of length 2 to get the probabilities of group AD and NC.

3.6. *Efficient Swin-transformer block*

To reduce the computational complexity of the swin-transformer stages, standard multi-headed self-attention (MSA) is replaced with a separable self-attention (SSA) of linear complexity.

In the MSA mechanism, to get the attention matrices, three matrices called Query, Key and Value (Q, K, V) are generated by three linear projection layers (shown in Fig. 2.a). In window multi-headed self-attention (W-MSA), Q, K, V$\in\mathbb{R}^{W\times M^2\times C}$, where W is the number of windows (windowed image size $= w \times w$, W=w^2). MSA and W-MSA compute self-attention in h groups by slicing Q, K, V into h heads, which allows the attention mechanism to converge knowledge from different subspaces and improve the understanding of models by computing multiple attention maps respectively. Since the self-attention mechanism can't understand the relative position information in the feature maps, according to [44,45], the relative position bias B, which contains the patch relative position information, is added into the attention computation. The standard W-MSA can be described as follows:

$$\begin{cases} Q = W_q X \\ K = W_k X \\ V = W_v X \end{cases} \qquad W_q, W_k, W_v \in \mathbb{R}^{C\times C} \tag{1}$$

$$W\text{-}MSA(Q,K,V)_{head} = SoftMax\left(\frac{QK^T}{\sqrt{C}} + B\right)VW_o \tag{2}$$

V will be linearly projected by $W_o \in \mathbb{R}^{C\times C}$ after being weighted by the attention maps. Supposing the size of feature map is C × h × w, the computational complexities of MSA and W-MSA are:

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C \tag{3}$$

$$\Omega(\mathrm{W-MSA}) = 4hwC^2 + 2M^2hwC \tag{4}$$

Although the local attention within the windows reduces the computational complexity from $O((hw)^2)$ to $O(hw)$, its large-batch matrix multiplication is still a very costly operation. So, we introduced the separable self-attention (SSA) into our swin-transformer block. The SSA replace Q with a vector Input (I $\in \mathbb{R}^{W\times M^2}$), which is projected by a linear layer with weight $W_i \in \mathbb{R}^C$, and then transform the I into context scores $C_s \in \mathbb{R}^{W\times M^2}$ by applying a softmax (shown in Fig. 2.b). Different from the attention scores computed for each token with respect to all M^2 tokens of

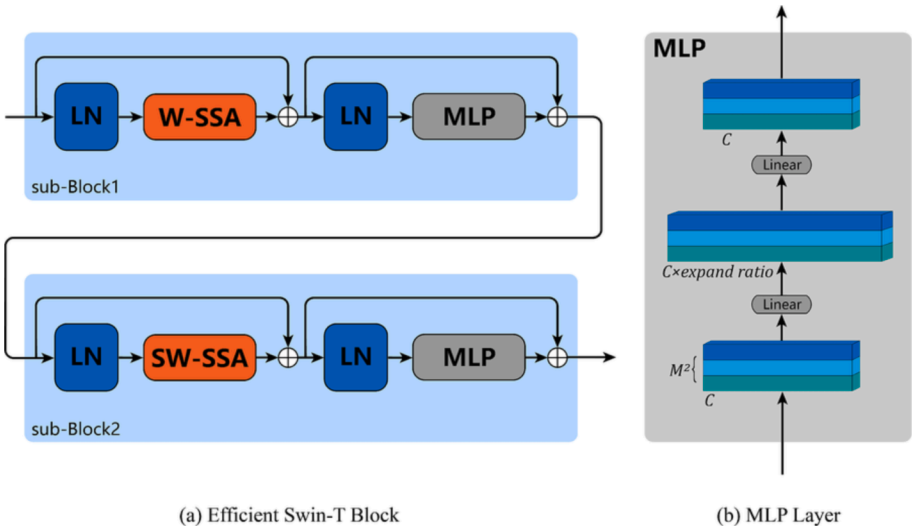


Fig. 3. Schematic diagram of the efficient swin-transformer block (ES-TB). (a) The structure of the proposed ES-TB is similar to the original swin-T block, but in order to get a more efficient block, we replace the W-MHA and SW-MHA by W-SSA and SW-SSA respectively. (b) The three-layer MLP in ES-TB. The expand ratio (r_{expand}) in the MLP layers is set to 1 in the ES-TB, instead of the 4 in the original swin-T block. (a) is adapted from Fig. 3(b) from [35] by Ze Liu et al., used under CC BY 4.0.

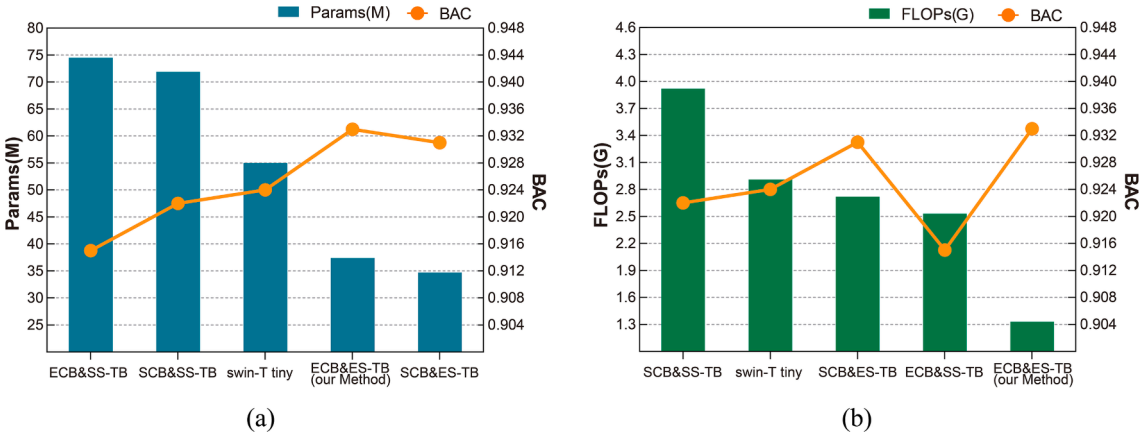


Fig. 4. Changes of BAC with respect to the parameters and FLOPs of swin-T tiny and the models with different block setting. (a) BAC and the number of parameters, (b) BAC and FLOPs.

K in W-MSA, separable attention only calculates the context scores with respect to a latent token L, which is implicitly represented as learnable weight W_i in the SSA. The context vector $C_v \in \mathbb{R}^{1\times C}$ is then computed by using C_s to weight K:

$$C_v = \sum_1^{M^2} C_s(i)K(i) \tag{5}$$

C_v is a cheap analogue of the attention map in W-MSA, where the encoded contextual information is shared by all tokens in each window. With C_s and C_v, the matrix computation in standard MSA and W-MSA can be simplified to two cheaper element-wise multiplication operations Window attention approach also can be applied in SSA. Window separable self-attention (W-SSA) can be described as:

$$\begin{cases} I = W_i X \\ K = W_k X \\ V = W_v X \end{cases} \qquad W_i \in \mathbb{R}^{C\times 1};\ W_k, W_v \in \mathbb{R}^{C\times C} \tag{6}$$

$$W\text{-}SSA(I,K,V) = \sum(\mathrm{SoftMax}(I)\cdot K)\cdot \mathrm{ReLU}(V)W_o \tag{7}$$

As with W-MSA, the output matrix will be finally linearly projected once more by W_o, and the computational complexity is:

$$\Omega(\text{W-SSA}) = hw\left(3C^2 + C\right) + 2hwC \tag{8}$$

The computational complexity of W-SSA is still $O(hw)$, but the computational overhead of both the projection part and the attention part have reduced. Although the computational complexity of W-SSA is independent of the window size M × M, we found in our experiments that using window attention instead of global attention on our task and model leads to a more stable convergence in the late period of training.

The structure of the proposed Efficient Swin-transformer Block is shown in Fig. 3. Each sub-block of ES-TB contains two sub-layers, which are two layer-normalization followed by W-SSA and MLP respectively. And the MLP contains two fully connected layers, the output dimension of input token is expanded to $r_{expand} \times C$ by the first layer, and then compressed back to C through the second one. The original setting of r_{expand} in swin-T is 4, but we found the quadruple expansion in the MLP brings a large amount of computational overhead to swin-T block. Therefore, we set the r_{expand} to 1 in the MLP layer. Moreover, window attention restricts the attention computation to the local parts of the feature maps, so that it cannot focus on long-range information as global attention does. In order to introduce connections across adjacent windows, we apply the same shifted window partitioning approach of swin transformer to W-SSA in the second sub-block of ES-TB.

Table 2
Classification performance on ADNI, number of parameters and computational overhead of the two-stream models using different block settings.

Block setting		ACC	BAC	SEN	SPC	AUC	Params	FLOPs
Efficient Conv Block	Efficient swin-T Block							
		0.926	0.922	0.904	0.939	0.962	71.9 M	3.92G
	✓	0.933	0.931	**0.926**	0.937	0.962	**34.7 M**	2.72G
✓		0.921	0.915	0.891	0.939	0.953	74.5 M	2.53G
✓	✓	**0.939**	**0.936**	0.925	**0.947**	**0.964**	37.4 M	**1.33G**

3.7. Efficient Conv block

The standard convolution operation consists of C_{output} kernels in each convolution layer, and the size of each convolution kernel is $C_{input} \times k \times k$, where $k \times k$ is the receptive field of the convolution kernel. Output feature maps are obtained by large amount of convolution multiplication within dense connections between C_{output} kernels and the C_{input} feature maps. Supposing the size of output feature maps is $h \times w$, the computational complexity of standard convolution is:

$$\Omega(\text{standard conv}) = C_{output} C_{input} hwk^2 \tag{9}$$

Obviously, the large amount of multi-add operations brings a high computational overhead, so the standard convolution blocks in the first two CNN stages are replaced with depth-wise separable convolution (DSC) in our method. A DSC consists of a deep-wise convolution (DWC) and a point-wise convolution (PWC). DWC is modified from group convolution, and the number of groups is set to be the same as the channels of the input feature maps. So that each convolution kernel of DWC contains only one filter and only multiplies with the corresponding one feature map, thereby getting a sparse convolution operation. And PWC is a standard convolution layer with a kernel size of 1×1, which introduces information exchange between different feature maps in PWC, and can linearly project the feature maps obtained from DWC into different output dimension. The computational complexity of DSC can be expressed as:

$$\Omega(\text{DSC}) = C_{input} hwk^2 + C_{output} C_{input} hw \tag{10}$$

We also introduce the residual connection of ResNet [47] and the pre-normalization structure of swin-transformer into the Efficient conv block (ECB), and the activation function and normalization method are kept same with swin-transformer by using GELU and layer-normalization. In order to get a larger receptive field, the kernel size of the DWC is set to 5×5.

4. Experiment & results

In this section, we first present the implementation details of our experiments. Secondly, several ablation studies were performed to evaluate the effectiveness of the approaches used in our proposed ECSnet; then we compared the performance of the two-stream models formed by different backbone networks to validate the advantages of our backbone over classical 2D models. Finally, we compare the performance of our model with that of previous works which evaluated on similar subsets of AIBL.

4.1. Implementation and experimental settings

The training and testing tasks were implemented on Python 3.8.8 and Pytorch 1.10.0 with an Intel Core i5-11400H with 16 GB of RAM and an NVIDIA GeForce RTX 3060 GPU 6 GB.

In our experiments, all training data were allocated from the 865 subjects of ADNI dataset. When evaluating the performance on the ADNI, the models were evaluated by using 5-fold cross-validation to make the results more robust. 20% non-overlapping data were allocated each time as the validation set, and the average results that obtained from the five validation sets were used to evaluate the performance. When using the data from AIBL as the test set, all the data from ADNI was used as training data, and the models were trained on the training set for 55 epochs.

The feature dimension of the 2.5D input image is expanded to 64 through the first convolution block, and then expanded twice as large after each maxpooling layer. After the first two stages, 384-dimensional tokens are obtained by patch embedding layer.

In the training stage, the proposed ECSnet was trained by using AdamW optimizer with weight decay, the initial learning rate was set as $1 \times 10^{-5}$with an 85% decay per 20 epochs, and the weight decay was set as 0.04. A weighted cross entropy function was used to measure the classification loss to reduce the impact of class-imbalance, whose weight was calculated based on the ratio of AD and NC subjects in the training set, so that the loss function will be more biased towards the group with fewer subjects. The batch size and drop-rate in output layer are set to 32 and 0.3 respectively.

We evaluated the model performance from multiple perspectives by using metrics including accuracy, sensitivity(recall), specificity, balance accuracy (BAC), ROC curve and corresponding AUC [46]. The accuracy, sensitivity and specificity can be defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

Given the class-imbalance problem in the evaluation data, BAC is a better metric than ACC because the accuracies of both positive and negative groups are considered, while standard ACC will be more biased towards the group with more subjects. BAC can be expressed as:

$$BAC = \frac{Sensitivity + Specificity}{2} \tag{14}$$

4.2. Ablation study

In the proposed ECSnet, we lightened the convolutional blocks and the swin-transformer blocks by applying DSC and W-SSA. In order to evaluate the impact on performance and efficiency of these approaches, we performed an ablation study by applying different block settings on our model, listing the classification results on ADNI, number of parameters and computational overhead for each model (Table 2). When ECBs or ES-TBs are not applied in the backbone, we use standard convolutional block (SCB) and standard swin-T block (SS-TB) to replace them. The SCB consists of a layer-normalization layer followed by a standard convolution layer, and the kernel size is set to 3×3. The SS-TB has the same structure as the original swin-T block, and the r_{expand} of the MLP is also set to 4.

As shown in Table 2, the ES-TB reduces the computational overhead while also significantly reducing the number of parameters; however, the ECB contains one more layer normalization operation, so the parameters is more than that of the SCB, but it also significantly reduces the calculation overhead. While the number of parameters is reduced by

Table 3
Classification performance of models with or without CNN structure on ADNI and AIBL.

Method	Dataset	ACC	BAC	SEN	SPC	AUC
swin-T tiny with ES-TB	ADNI	0.927	0.926	0.922	0.930	0.962
Our model		**0.939**	**0.936**	**0.925**	**0.947**	**0.964**
swin-T tiny with ES-TB	AIBL	0.896	0.882	0.861	0.902	0.940
Our Method		**0.939**	**0.928**	**0.911**	**0.944**	**0.963**

adding the lightweight approaches into the model, the performance of the model does not degenerate. The sensitivity of 92.5% achieved by the model with ECB and ES-TB (our ECSnet) is slightly lower than 92.6% of the model applied SCB and ES-TB, but ECSnet has better performance in all other metrics. On the ADNI dataset, our ECSnet achieved 93.9% accuracy and 0.964 AUC, while the FLOPs and number of parameters were only 33.9% and 52.0% of those of the model with SCB and SS-TB.

We apply the CNN structure in the first two stages of the backbone network to enhance the model's ability by introducing the inductive bias, and we evaluated the effectiveness of this method on the ADNI and AIBL datasets. When not replacing the first two stages of the model with CNN in the experiments, the backbone network keeps the same structure as swin-T tiny with the number of blocks in the four stages [2,2,6,2], but still replace the W-MSA by W-SSA. The results (Table 3) show that on both ADNI and AIBL datasets, our proposed model achieves better performance in all metrics, with significantly higher BAC (92.8%) and SEN (91.1%) on AIBL.

To validate the effectiveness of our 2.5D method, we also compared the performance of different input methods (Fig. 5). Specifically, there are five input methods: 1. Extract two 2D slices as input images from the middle of the vertical axis direction of our two parts 2.5D image; 2. Feed the input images into a one-stream network without dividing them into two parts, and do not downsample them along the vertical axis; 3. Feed the input images into a one-stream network without dividing it into two parts, but downsample them along the vertical axis; 4. Divide the image into two parts but without downsampling; 5. Divide the images into two parts and perform downsampling.

With results reported in Fig. 5, the 2.5D method significantly improved the models' performance. And together with the use of downsampling method, the model performs more balanced in the classification of positive and negative subjects. In addition to the improvement in classification performance, the downsampling method can also reduce the computation on CPU during random data augmentation, making the training more efficient.

4.3. Methods comparison

4.3.1. Comparison of different hyperparameter settings

In ECSnet, we divide the backbone network into four stages, and the feature maps are downsampled while being fed forward through different stages. We evaluated the classification performance with different stage strategies, while keeping the depth of CNN stages and swin-T stages same as that in ECSnet. Through the downsampling, the receptive field is expanded. The results (Table 4) show that the 4-stage strategy is beneficial to the model performance.

The hyperparameters play an important role in the model training. We evaluated the classification performance of our ECSnet with different hyperparameter settings (Table 5 and Fig. 6), specifically, the weight decay in optimizer Adam, the expand ration of MLP in the ES-TB and the number of feature maps (channels).

4.3.2. Comparison of different methods

Then we compared the performance of models using previous solid 2D natural image classification backbone networks to form our proposed two-stream network. In the experiment, our proposed backbone network was replaced with swin-T tiny, ResNet34, ResNet18, DenseNet121 and SE_ResNet18 respectively and performed five-fold validations on ADNI and AIBL. We also calculated the number of parameters and the computational overhead of the models. All the models are compared under the same training strategy, data augmentation strategy and optimizer as our ECSnet. The learning rate and weight decay in AdamW optimizer are set to 1×10^{-5}, 0.01 for the CNN backbones and 1×10^{-5}, 0.05 for the swin-T tiny.

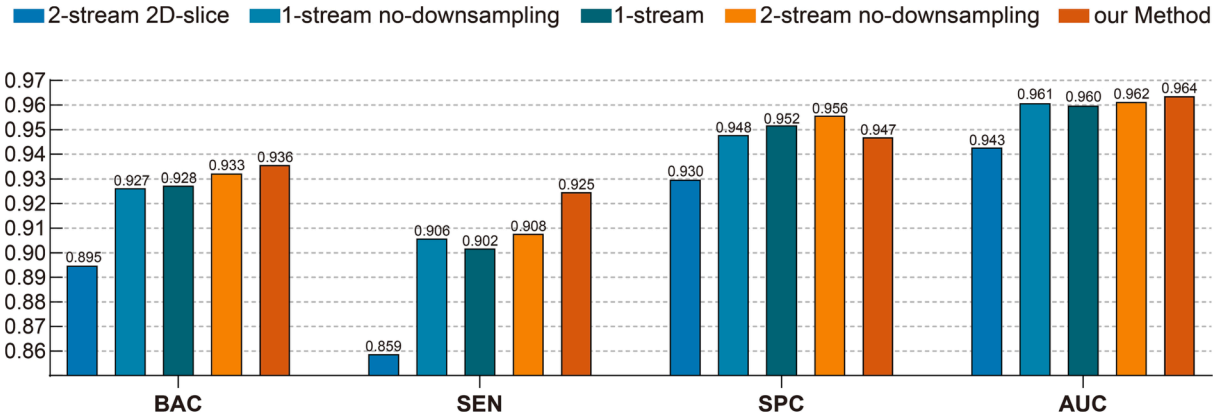


Fig. 5. Results of the models with different input methods on ADNI. The methods are: 1. Extract two 2D slices as input images from the middle of the vertical axis direction of our two parts 2.5D image (two-stream 2D-slice); 2. Feed the input images into a one-stream network without dividing them into two parts, and do not downsample them along the vertical axis (1-steam no-downsampling); 3. Feed the input images into a one-stream network without dividing it into two parts, but downsample them along the vertical axis (1-stream); 4. Divide the image into two parts but without downsampling (two-stream no-downsampling); 5. Divide the image into two parts and perform downsampling (our method).

Table 4
Classification performance of different stage strategies on ADNI.

Number of stages	ECBs in CNN stages	ES-TBs in Swin-T stages	ACC	BAC	SEN	SPC	AUC
2	[3]	[9]	0.910	0.909	0.911	0.906	0.947
3	[1,2]	[9]	0.925	0.919	0.894	0.944	0.954
4(ECSnet)	[1,2]	[6,3]	**0.939**	**0.936**	**0.925**	**0.947**	**0.964**

Table 5
Classification performance of different hyperparameter settings on AIBL. The default setting of the ECSnet is: Weight decay = 0.04, MLP expand ratio = 1.0, number of channels in each stage of the backbone networks = [64, 128, 384, 768].

hyperparameter	settings	ACC	BAC	SEN	SPC	AUC	Params	FLOPs
Weight decay	0.02	0.917	0.914	0.911	0.917	0.954	37.4 M	1.33G
	0.03	0.936	0.925	0.911	0.940	0.961		
	0.04	**0.939**	**0.928**	**0.911**	**0.944**	**0.963**		
	0.05	0.917	0.914	0.911	0.917	0.961		
MLP expand ratio	1.0	**0.939**	**0.928**	0.911	**0.944**	**0.963**	**37.4 M**	**1.33G**
	1.5	0.934	0.924	0.911	0.938	0.950	42.7 M	1.50G
	2.0	0.932	0.919	0.899	0.938	0.959	48.0 M	1.67G
	2.5	0.932	**0.928**	**0.924**	0.933	0.960	53.3 M	1.84G
	3.0	0.920	0.912	0.899	0.924	0.961	58.7 M	2.01G
Stage channels	[32,64,192,384]	0.922	0.907	0.886	0.928	0.953	**11.7 M**	**0.36G**
	[48,96,288,576]	0.924	0.903	0.873	0.933	0.953	22.1 M	0.77G
	[64,128,384,768]	**0.939**	0.928	0.911	**0.944**	**0.963**	37.4 M	1.33G
	[80,160,480,960]	0.936	**0.931**	**0.924**	0.938	0.960	55.9 M	2.04G

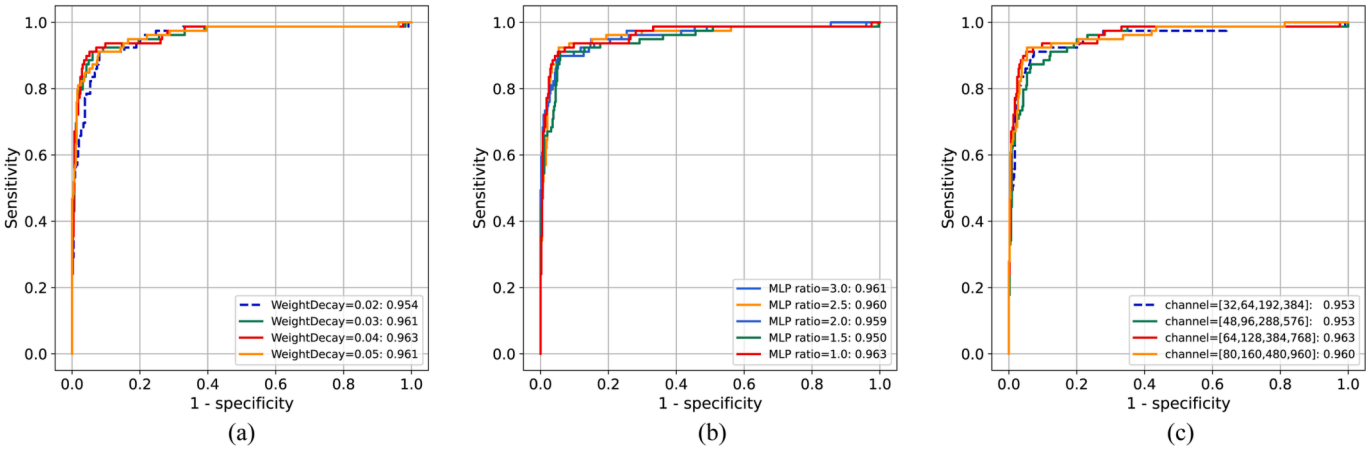


Fig. 6. ROC curves and AUC of different hyperparameter settings on AIBL. (a)weight decay in Adam, (b)expand ratio of MLP in ES-TB, (c) number of channels in each stage of the backbone networks.

Table 6
Performance on ADNI of the proposed two-stream networks with different 2D backbones.

Backbone	ACC	BAC	SEN	SPC	AUC	Params	FLOPs
swin-T tiny [35]	0.926	0.924	0.909	0.938	0.962	55.0 M	2.91G
ResNet34 [47]	0.905	0.898	0.866	0.930	0.947	42.7 M	1.64G
ResNet18 [47]	0.916	0.914	0.905	0.923	0.956	22.5 M	**0.96G**
SE_ResNet18 [48]	0.926	0.919	0.887	**0.951**	0.962	22.7 M	0.96G
DenseNet121 [32]	0.908	0.906	0.899	0.913	0.948	**14.0 M**	1.35G
our Backbone(ECSnet)	**0.939**	**0.936**	**0.925**	0.947	**0.964**	37.4 M	1.33G

As shown in Table 6, obviously, the parameters and computational cost of the models based on transformer are relatively large compared with the CNN models. Our proposed backbone network reduces the computational complexity to be close to that of the 2D CNN models while achieving the best results on ADNI in all evaluation metrics except SPC. On the AIBL (Fig. 7), it can be seen that the model based on our proposed backbone (ECSnet) achieved the best results in most of the metrics, and the ECSnet has obvious advantages in BAC and SEN (92.8% and 91.1%), and 93.9% of ACC and 0.963 of AUC are also higher than those of other models.

In order to compare with previous works more fairly, we used the data from AIBL as an independent test set and compared the results with SOTA models which were evaluated on similar subsets.

As shown in Table 7, these previous methods are all based on 3D networks such as 3D CNN to extract features, but our 2.5D method whose backbone network is only based on 2D methods allows our model to have a relatively small computational overhead. At the same time, our model achieved a great performance. The BAC (92.8%) and SEN (91.1%) of our ECSnet are significantly better than the 88.2 ~ 91.0% and 72.3% ~88.9% of other methods, other metrics are also close to the SOTA models. It is worth noting that the method in [41] included clinical information such as MMSE scores as input features, while the rest of the methods only used original or standardized sMRI images.

5. Discussion

5.1. Experimental discussion

Although many works have achieved encouraging results in computer aided AD diagnosis based on sMRI and deep learning, there are still many limitations in applying the methods to the clinic. In this study, in order to avoid the use of 3D algorithms which would greatly increase the model parameters and computational overhead in our model, our proposed ECSnet was built on a two-stream structure and applied a 2.5D subject-level method which can reduce the loss of 3D information while keeping efficient by using 2D algorithms. The ECSnet was also applied a

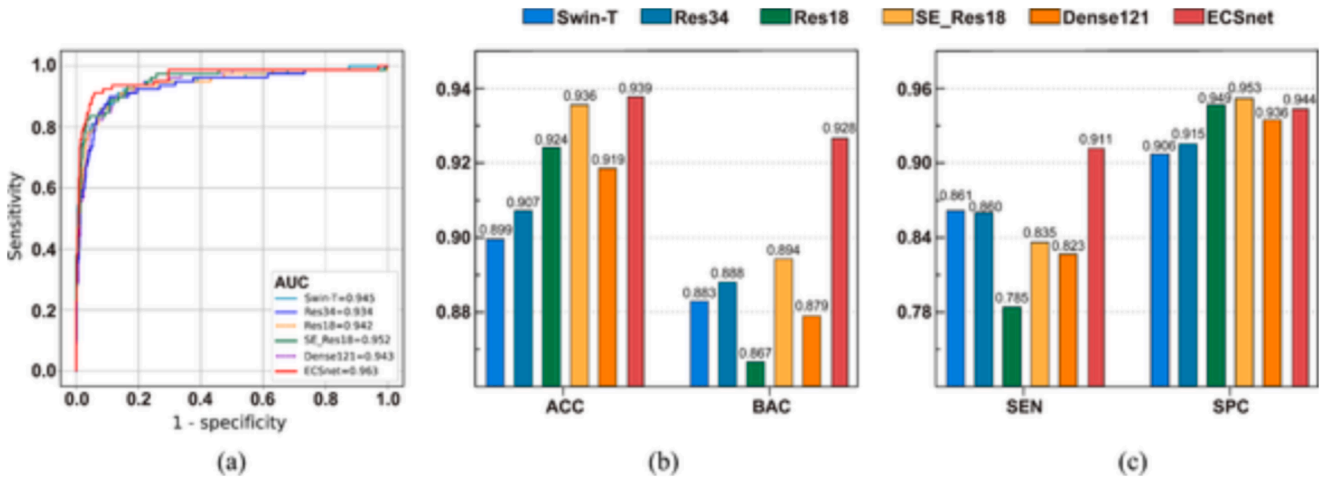


Fig. 7. Performance on AIBL of the proposed two-stream networks with different 2D backbones. (a) ROC curve of the models, (b) histogram of the ACC and BAC, (c) histogram of the SEN and SPC.

Table 7
Performance comparison with state-of-the-art methods on AIBL.

Methods	Groups (AD/ NC)	Data	ACC	BAC	SEN	SPC	AUC	Approach
Lian et al. (2020)[28]	72/447	skull-stripped sMRI	0.898	0.887	0.873	0.902	0.946	3D patch-level
Qiu et al. (2020)[41]	62/320	sMRI+clinical data	0.932	0.910	0.877	0.943	**0.974**	3D subject-level
Zhu et al. (2021)[27]	79/307	skull-stripped sMRI	0.902	0.882	0.848	0.915	0.939	3D patch-level
Han et al. (2022) [49]	72/359	skull-stripped sMRI	0.923	0.910	0.889	0.930	0.950	3D subject-level
Li et al. (2022)[50]	79/448	sMRI GM	**0.939**	0.851	0.723	**0.978**	0.957	3D subject-level
Our Method	79/449	sMRI GM&WM	**0.939**	**0.928**	**0.911**	0.944	0.963	2.5D subject-level

series of lightweight approaches that make it easy to train and deploy the model on computers without high performance.

We applied a series of lightweight methods to simplify our backbone network and deal with the overfitting by applying data augmentation, two-stream structure and 2.5D method. With a 5-fold validation, our model achieved an accuracy of 93.9% and an AUC of 0.964 on the ADNI dataset, which are close to the previous works. We also performed an ablation study to evaluate the impact of the lightweight methods on model performance, parameters, and computational overhead (Fig. 4.a). The results show that the BAC on ADNI dataset subsequently increases as ECB and ES-TB are applied in the backbone; and as shown in Fig. 4.b, the lightweight methods significantly reduce the computational overhead, while the model performance is not impaired. The number of parameters and FLOPs are only 68% and 45.7% of those of the swin-transformer tiny respectively while the performance is even better. With the results reported in Table 3, the performance degraded slightly on ADNI but sharply on AIBL when the early-stage CNN method is not applied to the backbone. That validates the generalization ability of the model on unseen data, which is brought by the inductive bias of convolution layers. In brief, our 2.5D method combined with early-stage CNN method allows the model to achieve a good performance while being lightweight.

We also validate the effectiveness of the methods on the independent test set allocated from AIBL (Table 7). Our model achieved BAC of 92.8% and SEN of 91.1%, which are significantly higher than those of the compared state-of-the-art models, and other metrics are also close to the SOTA models. The results also indicate that the 2.5D method, two-stream structure and the CNN used in the early stages can improve the model's ability of encoding features and the generalization performance. Moreover, compared to the 3D subject-level and 3D patch-level models, the training of our lightweight 2.5D models requires less computing resource and takes less time (about 12 mins each training in our experiments). In recent years, some works [27,41] applied risky region identification algorithms that were usually developed on 3D models like fully convolution net (FCN) in the AD diagnosis. With the identification algorithms, only high-risk regions were kept, so that the redundant data fed into the following classification models could be reduced. However, the approaches also made the training of these methods more costly. In contrast, our ECSnet can achieve close or even better and more balanced performance with only a one-time end-to-end training.

Table 8
Classification performance of the proposed model on MCI diagnosis task.

Groups(pMCI/ sMCI)	Task	ACC	SEN	SPC	AUC
138/218	(AD & pMCI) vs sMCI	0.771	0.820	0.682	0.807
	pMCI vs NC	0.932	0.877	0.947	0.957
	sMCI vs NC	0.842	0.779	0.863	0.897
	(sMCI & pMCI) vs NC	0.859	0.863	0.857	0.930

5.2. Future works

As the early stage of AD, the diagnosis of MCI is also important to the early disease intervention. We performed 5-fold validations to evaluate the model performance on MCI diagnosis tasks on ADNI. Subjects from the MCI group who converted to AD within 36 months after the first scan were classified as pMCI and the rest as sMCI. As shown in Table 8, our model achieved potential results. In the future, we can try to extend our

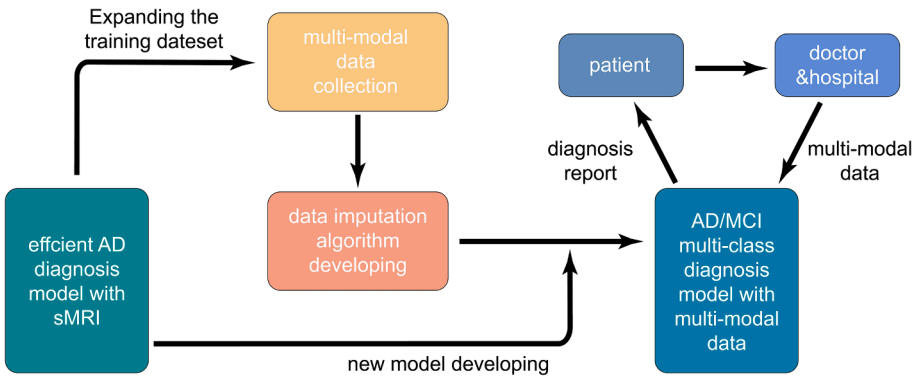


Fig. 8. Block diagram of possible future works and implementation [51].

work to improve the performance of MCI classification.

Except for image data such as sMRI and PET, several types of non-imaging data are often used for the diagnosis or study of AD, such as subjects' gender, age, ApoE4, Aβ42 and neuropsychological scales. However, the genetic information, biomarker information and images like PET are only available for part of the subjects in most datasets, and the small amount of data makes their application in deep learning methods more challenging. For the above reasons, we did not include these modalities in this study. In possible future works, we can try to collect more multi-modal data or develop an effective data imputation algorithm (e.g., GAN), improving the model performance by utilizing the multi-modal model. The Fig. 8 depicts the block diagram of possible future works and implementation of AD/MCI diagnosis model.

6. Conclusion

We proposed a backbone network that combines CNN and swin-transformer to form an efficient two-stream model named ECSnet for automatic AD diagnosis with sMRI. Firstly, we applied CNN structure at the early stage of our transformer-based backbone to improve the generalization ability by introducing the inductive bias. Secondly, we built the model on a two-stream structure and applied a 2.5D method to the model to enhance the ability of encoding 3D features with 2D algorithms. Finally, we applied a series of lightweight methods in our model to make the parameters and computational overhead less than recent state-of-the-art models which are mostly based on costly 3D algorithms. Compared to the SOTA models, our ECSnet is a 2D end-to-end model that is efficient in training and inference, and can achieve similar or even better performance, demonstrating the effectiveness of our proposed methods and the trade-off between computational overhead and model performance.

CRediT authorship contribution statement

Jiaming Xin: Methodology, Software, Investigation, Writing – original draft. **Ancong Wang:** Investigation, Validation, Conceptualization. **Rui Guo:** Investigation, Validation. **Weifeng Liu:** Conceptualization, Writing – review & editing. **Xiaoying Tang:** Supervision, Conceptualization, Project administration.

Declaration of Competing Interest

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

Data availability

The data that has been used is from the public dataset, and the links to the data are mentioned in the article

Acknowledgements

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: 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 affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The Australian Imaging Biomakers and Lifestyle (AIBL) study was supported by funding from the Science and Industry Endowment Fund, the Dementia Collaborative Research Centres, the McCusker Alzheimer's Research Foundation, the National Health and Medical Research Council (AUS), and the Yulgilbar Foundation, plus numerous commercial interactions supporting data collection. Details of the AIBL consortium can be found at www.AIBL.csiro.au and a list of the researchers of AIBL is provided at http://aibl.csiro.au/.

References

[1] E.D. Roberson, L. Mucke, 100 years and counting: prospects for defeating Alzheimer's disease, Science 314 (5800) (2006) 781–784, https://doi.org/10.1126/science.1132813.
[2] Alzheimer's disease facts and figures. (2020). *Alzheimers Dement*. https://doi.org/10.1002/alz.12068.
[3] W. Jagust, Vulnerable neural systems and the borderland of brain aging and neurodegeneration, Neuron 77 (2) (2013) 219–234, https://doi.org/10.1016/j.neuron.2013.01.002.
[4] A. Atri, Current and future treatments in Alzheimer's disease, Semin. Neurol. 39 (02) (2019) 227–240, https://doi.org/10.1055/s-0039-1678581.
[5] S.I. Khan, R.B. Pachori. Automated Eye Movement Classification Based on EMG of EOM Signals Using FBSE-EWT Technique. Ieee Transactions on Human-Machine Systems. https://doi.org/10.1109/thms.2023.3238113.
[6] S.I. Khan, S.M. Qaisar, R.B. Pachori, Automated classification of valvular heart diseases using FBSE-EWT and PSR based geometrical features, Biomed. Signal Process. Control 73 (2022), 103445, https://doi.org/10.1016/j.bspc.2021.103445.

[7] S.I. Khan, R.B. Pachori. Empirical Wavelet Transform-Based Framework for Diagnosis of Epilepsy Using EEG Signals. (2022). In R. K. Chaurasiya, D. Agrawal, & R. B. Pachori (Eds.), AI-Enabled Smart Healthcare Using Biomedical Signals (pp. 217-239). IGI Global. https://doi.org/10.4018/978-1-6684-3947-0.ch012.

[8] S. Sarraf, J. Sun. Functional brain imaging: A comprehensive survey (2016). *arXiv preprint arXiv:1602.02225*. https://doi.org/10.48550/arXiv.1602.02225.

[9] P. Vemuri, D.T. Jones, C.R. Jack, Resting state functional MRI in Alzheimer's Disease, Alzheimers Res. Ther. 4 (1) (2012) 1–9, https://doi.org/10.1186/alzrt100.

[10] V. Camus, P. Payoux, L. Barré, B. Desgranges, T. Voisin, C. Tauber, R. La Joie, M. Tafani, C. Hommet, G. Chételat, Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment, Eur. J. Nucl. Med. Mol. Imag. 39 (4) (2012) 621–631, https://doi.org/10.1007/s00259-011-2021-8.

[11] C. Marcus, E. Mena, R.M. Subramaniam, Brain PET in the diagnosis of Alzheimer's disease, Clin. Nucl. Med. 39 (10) (2014) e413.

[12] T.M. Nir, N. Jahanshad, J.E. Villalon-Reina, A.W. Toga, C.R. Jack, M.W. Weiner, P. M. Thompson, A.S.D.N. Initiative, Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging, NeuroImage: clinical 3 (2013) 180–195, https://doi.org/10.1016/j.nicl.2013.07.006.

[13] T. Maggipinto, R. Bellotti, N. Amoroso, D. Diacono, G. Donvito, E. Lella, A. Monaco, M.A. Scelsi, S. Tangaro, A.S.D.N. Initiative, DTI measurements for Alzheimer's classification, Phys. Med. Biol. 62 (6) (2017) 2361, https://doi.org/10.1088/1361-6560/aa5dbe.

[14] S. Leandrou, S. Petroudi, P.A. Kyriacou, C.C. Reyes-Aldasoro, C.S. Pattichis, Quantitative MRI brain studies in mild cognitive impairment and Alzheimer's disease: A methodological review, IEEE Rev. Biomed. Eng. 11 (2018) 97–111, https://doi.org/10.1109/RBME.2018.2796598.

[15] S. Rathore, M. Habes, M.A. Iftikhar, A. Shacklett, C. Davatzikos, A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages, Neuroimage 155 (2017) 530–548, https://doi.org/10.1016/j.neuroimage.2017.03.057.

[16] E. Hosseini-Asl, R. Keynton, A. El-Baz, Alzheimer's disease diagnostics by adaptation of 3D convolutional network, IEEE international conference on image processing (ICIP) 2016 (2016) 126–130, https://doi.org/10.1109/ICIP.2016.7532332.

[17] X.W. Gao, R. Hui, Z. Tian, Classification of CT brain images based on deep learning networks, Comput. Methods Programs Biomed. 138 (2017) 49–56, https://doi.org/10.1016/j.cmpb.2016.10.007.

[18] M. Khojaste-Sarakhsi, S.S. Haghighi, S.F. Ghomi, E. Marchiori, Deep learning for Alzheimer's disease diagnosis: A survey, Artif. Intell. Med. 102332 (2022), https://doi.org/10.1016/j.artmed.2022.102332.

[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly. An image is worth 16x16 words: Transformers for image recognition at scale (2020). *arXiv preprint arXiv: 2010.11929*. https://doi.org/10.48550/arXiv.2010.11929.

[21] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, G. Varoquaux, Machine learning for neuroimaging with scikit-learn, Front. Neuroinf. 14 (2014), https://doi.org/10.3389/fninf.2014.00014.

[22] B. Mwangi, T.S. Tian, J.C. Soares, A review of feature reduction techniques in neuroimaging, Neuroinformatics 12 (2) (2014) 229–244, https://doi.org/10.1007/s12021-013-9204-3.

[23] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation, Med. Image Anal. 63 (2020), 101694, https://doi.org/10.1016/j.media.2020.101694.

[24] M. Hon, N.M. Khan, Towards Alzheimer's disease classification through transfer learning, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, https://doi.org/10.1109/BIBM.2017.8217822.

[25] R. Mendoza-Léon, J. Puentes, L.F. Uriza, M.H. Hoyos, Single-slice Alzheimer's disease classification and disease regional analysis with Supervised Switching Autoencoders, Comput. Biol. Med. 116 (2020), 103527, https://doi.org/10.1016/j.compbiomed.2019.103527.

[26] W. Kang, L. Lin, B. Zhang, X. Shen, S. Wu, A.S.D.N. Initiative, Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis, Comput. Biol. Med. 136 (2021), 104678, https://doi.org/10.1016/j.compbiomed.2021.104678.

[27] W. Zhu, L. Sun, J. Huang, L. Han, D. Zhang, Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI, IEEE Trans. Med. Imaging 40 (9) (2021) 2354–2366, https://doi.org/10.1109/TMI.2021.3077079.

[28] C. Lian, M. Liu, Y. Pan, D. Shen, Attention-guided hybrid network for dementia diagnosis with structural MR images, IEEE Trans. Cybern. (2020), https://doi.org/10.1109/tcyb.2020.3005859.

[29] R. Hedayati, M. Khedmati, M. Taghipour-Gorjikolaie, Deep feature extraction method based on ensemble of convolutional auto encoders: Application to Alzheimer's disease diagnosis, Biomed. Signal Process. Control 66 (2021), 102397, https://doi.org/10.1016/j.bspc.2020.102397.

[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90, https://doi.org/10.1145/3065386.

[31] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017). *arXiv preprint arXiv:1704.04861*. https://doi.org/10.48550/arXiv.1704.04861.

[32] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700-4708 (2017). https://doi.org/10.1109/CVPR.2017.243.

[33] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu. Ghostnet: More features from cheap operations. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020).

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, International conference on machine learning (2015) 2048–2057, https://doi.org/10.5555/3045118.3045336.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 10012-10022. https://doi.org/10.48550/arXiv.2103.14030.

[36] S. Wang, B.Z. Li, M. Khabsa, H. Fang, H. Ma. Linformer: Self-attention with linear complexity (2020). *arXiv preprint arXiv:2006.04768*. https://doi.org/10.48550/arXiv.2006.04768.

[37] S. Mehta, M. Rastegari. Separable Self-attention for Mobile Vision Transformers (2022). *arXiv preprint arXiv:2206.02680*. https://doi.org/10.48550/arXiv.2206.02680.

[38] C.R. Jack Jr, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, L.J. Whitwell, C. Ward, The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27 (4) (2008) 685–691, https://doi.org/10.1002/jmri.21049.

[39] K.A. Ellis, C.C. Rowe, V.L. Villemagne, R.N. Martins, C.L. Masters, O. Salvado, C. Szoeke, D. Ames, Group, A. R. Addressing population aging and Alzheimer's disease through the Australian Imaging Biomarkers and Lifestyle study: Collaboration with the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & dementia*, *6*(3),(2010), 291-296. https://doi.org/10.1016/j.jalz.2010.03.009.

[40] Q. Zhang, Q. Du, G. Liu, A whole-process interpretable and multi-modal deep reinforcement learning for diagnosis and analysis of Alzheimer's disease*, J. Neural Eng. 18 (6) (2021), 066032 https://doi.org/10.1088/1741-2552/ac37cc.

[41] S. Qiu, P.S. Joshi, M.I. Miller, C. Xue, X. Zhou, C. Karjadi, G.H. Chang, A.S. Joshi, B. Dwyer, S. Zhu, Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, Brain 143 (6) (2020) 1920–1933, https://doi.org/10.1093/brain/awaa137.

[42] M.A. DeTure, D.W. Dickson, The neuropathological diagnosis of Alzheimer's disease, Mol. Neurodegener. 14 (1) (2019) 1–18, https://doi.org/10.1186/s13024-019-0333-5.

[43] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better, Adv. Neural Inf. Proces. Syst. 34 (2021) 30392–30400.

[44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67, https://doi.org/10.5555/3455716.3455856.

[45] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei. Relation networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 3588–3597. https://doi.org/10.1109/CVPR.2018.00378.

[46] S.I. Khan, R.B. Pachori, Derived vectorcardiogram based automated detection of posterior myocardial infarction using FBSE-EWT technique [Article], Biomed. Signal Process. Control 70 (2021), 103051, https://doi.org/10.1016/j.bspc.2021.103051.

[47] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770-778.

[48] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 7132-7141. https://doi.org/10.1109/TPAMI.2019.2913372.

[49] K. Han, M. He, F. Yang, Y. Zhang, Multi-task multi-level feature adversarial network for joint Alzheimer's disease diagnosis and atrophy localization using sMRI, Phys. Med. Biol. 67 (8) (2022), 085002, https://doi.org/10.1088/1361-6560/ac5ed5.

[50] C. Li, Y. Cui, N. Luo, Y. Liu, P. Bourgeat, J. Fripp, T. Jiang. Trans-ResNet: Integrating Transformers and CNNs for Alzheimer's disease classification. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (2022), 1-5. https://doi.org/10.1109/ISBI52829.2022.9761549.

[51] S.I. Khan, R.B. Pachori, Automated classification of lung sound signals based on empirical mode decomposition, Expert Syst. Appl. 184 (2021), 115456, https://doi.org/10.1016/j.eswa.2021.115456.