Computers in Biology and Medicine 164 (2023) 107304

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification

Zhentao Hu [a], Yanyang Li [a], Zheng Wang [a,*], Shuo Zhang [a], Wei Hou [b], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] School of Artificial Intelligence, Henan University, Zhengzhou, 450046, China
[b] College of Computer and Information Engineering, Henan University, Kaifeng, 475004, China

ARTICLE INFO

Keywords:
Alzheimer's disease (AD)
Magnetic resonance imaging (MRI)
Deep learning (DL)
Transformer
Convolutional neural network (CNN)

ABSTRACT

Deep learning (DL) algorithms based on brain MRI images have achieved great success in the prediction of Alzheimer's disease (AD), with classification accuracy exceeding even that of the most experienced clinical experts. As a novel feature fusion method, Transformer has achieved excellent performance in many computer vision tasks, which also greatly promotes the application of Transformer in medical images. However, when Transformer is used for 3D MRI image feature fusion, existing DL models treat the input local features equally, which is inconsistent with the fact that adjacent voxels have stronger semantic connections than spatially distant voxels. In addition, due to the relatively small size of the dataset for medical images, it is difficult to capture local lesion features in limited iterative training by treating all input features equally. This paper proposes a deep learning model Conv-Swinformer that focuses on extracting and integrating local fine-grained features. Conv-Swinformer consists of a CNN module and a Transformer encoder module. The CNN module summarizes the planar features of the MRI slices, and the Transformer module establishes semantic connections in 3D space for these planar features. By introducing the shift window attention mechanism in the Transformer encoder, the attention is focused on a small spatial area of the MRI image, which effectively reduces unnecessary background semantic information and enables the model to capture local features more accurately. In addition, the layer-by-layer enlarged attention window can further integrate local fine-grained features, thus enhancing the model's attention ability. Compared with DL algorithms that indiscriminately fuse local features of MRI images, Conv-Swinformer can fine-grained extract local lesion features, thus achieving better classification results.

1. Introduction

The aging of the population has brought about various social problems, such as Alzheimer's disease (AD). According to statistics, the incidence of Alzheimer's disease among the elderly aged between 65 and 74 is about 3%; among the elderly aged between 75 and 84, the incidence of Alzheimer's disease is about 19%; among the elderly over the age of 84, the incidence of Alzheimer's disease is about 47% [1,2]. Elderly people with Alzheimer's disease have a serious decline in memory and cognitive ability, which seriously affects their quality of life. According to the report, the U.S. population with AD is expected to increase from 6.5 million to 13.8 million under today's medical conditions by 2060 [3]. Alzheimer's disease has caused a huge economic burden to society. In the United States alone, the annual treatment cost of Alzheimer's disease and other dementia patients exceeds 230 billion U.S. dollars [4].

The pathogenesis of Alzheimer's disease is still unclear. Some literature believes that Alzheimer's disease is caused by the accumulation of Aβ and tau proteins in the brain, leading to neuronal dysfunction and death, and causing brain atrophy and inflammatory responses in patients [5]. AD patients show heterogeneity in cognition, imaging, and other aspects. Among them, imaging heterogeneity is due to variations in the local regional distribution of brain atrophy [6]. Some clinical studies reveal that AD patients can be classified into different subtypes, and patients with different subtypes have different patterns of brain degeneration [7–10]. As a state of early AD, mild cognitive impairment (MCI) MCI is an intermediate stage between normal and

* Corresponding author.
E-mail address: wangzheng@henu.edu.cn (Z. Wang).

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete list of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.compbiomed.2023.107304
Received 1 March 2023; Received in revised form 14 July 2023; Accepted 28 July 2023
Available online 31 July 2023

AD, but without treatment and intervention, MCI will deteriorate to AD type with high probability [11,12], however, if MCI can be accurately identified, a range of pharmacological and behavioral treatments can be prescribed to reduce the rate or progression of disease symptoms.

Magnetic resonance imaging (MRI) can effectively establish structure images of the subject's brain with high resolution, allowing the observation of anatomical and functional neural changes associated with AD [13]. Deep learning (DL) algorithms based on brain MRI images of AD patients have achieved great success in the prediction of AD, with prediction accuracy exceeding even the most experienced clinical experts [14]. Compared with the traditional machine learning method that manually screens features from MRI images, the DL model can automatically obtain discriminative features from MRI images through iterative training, so it has better generalization [15]. To enable the DL model to automatically establish the dependency relationship between text or image features, Recurrent Neural Network (RNN) was proposed. RNN uses neurons to save current state information and historical information at each time point, so RNN is widely used in tasks with progressive relationships, such as speech recognition, text generation, etc [16], however, RNN applications are inherently flawed in computer vision tasks because it cannot process image features that do not have asymptotic relationships in parallel.

Recently, Vaswani et al. [17] proposed a new feature fusion model for machine translation tasks, called Transformer. Transformer uses a multi-head self-attention mechanism to model the entire input text. Unlike RNNs that operate in a recursive-like fashion, Transformer leverages a multi-head self-attention mechanism to model all input word vectors. Recently, researchers have found that Transformer is also suitable for modeling image features and achieved breakthrough results [18–21]. On this basis, Liu et al. [22] proposed a hierarchical Transformer model based on shifted window self-attention mechanism, and achieved state-of-the-art performance in image classification, object detection, and image semantic segmentation. In this model, the input natural image is divided into multiple local patches, each local patch is represented as a visual token, and the Transformer encoder based on the shift window mechanism performs self-attention in the visual tokens inside each window. Different from processing the input local image features equally, Liu et al. [22] limited the fusion of local image features to a specific image space by dividing the attention window to strengthen the local feature extraction and the adjacent attention windows were connected by window shift so that the boundary information of adjacent windows is saved.

Considering that Alzheimer's patients show heterogeneity in the distribution of local regions of brain atrophy, and inspired by the study of Liu et al. [22], we propose a deep learning model Conv-Swinformer that focuses on extracting and integrating local fine-grained features for AD classification. Conv-Swinformer integrates convolutional neural network (CNN) advantages in extracting low-level detail features and Transformer in modeling global context features. By introducing the shift window attention mechanism in the Transformer encoder, the model attention is focused on a small spatial area of the MRI image, which effectively reduces unnecessary background semantic information and enables the model to capture local features more accurately, and the second partition of the attention window can preserve the boundary information between adjacent local spaces. For small MRI datasets, Conv-Swinformer strengthens the extraction of local spatial features with strong semantic correlation, thus better utilizing local lesion features for AD classification.

The main contributions of this work are summarized as follows:

- A deep learning model integrating CNN and Transformer is proposed to better extract local features of brain atrophy in AD patients. The effectiveness of the proposed model is also trained and validated on the ADNI database, and the generalizability of the proposed model is evaluated on the OASIS dataset.
- The introduction of a shift window attention mechanism in the Transformer is proposed to enhance the extraction of MRI-localized features.
- The performance of several mainstream CNN architectures in extracting MRI slice features is compared and a final model architecture is obtained. The performance of the model on MRI slices in three directions is also compared and the most suitable slice sequence for AD classification is identified.
- The performance of the model under different data enhancement methods was verified. Meanwhile, the features extracted by the model in three classification tasks were visualized and analyzed, and the results showed that the model can focus on important brain regions closely related to AD, which is more needed in clinical diagnosis.

2. Related works

As an important biomarker of AD, MRI provides a key diagnostic reference for clinicians. Clinicians assess the extent of brain atrophy in AD patients based on MRI scans to determine the stage of disease progression. However, it is difficult for clinicians to delicately process massive and complex MRI images. To improve the diagnostic efficiency of clinicians, research on computer-aided AD diagnosis based on MRI images has become a hot spot [23]. In the past two decades, traditional machine learning techniques that manually extract features have been widely used in the diagnosis of Alzheimer's disease, and its diagnostic efficiency even exceeds that of the most experienced clinicians. According to different feature extraction regions, these machine learning methods can be roughly divided into three categories: Voxel-based, Region of Interest (ROI)-based, and local patch-based. The voxel-based machine learning method abstracts the inherent features of medical images through feature extraction, such as hippocampal volume and subcortical volume [24], and then feeds them into the classifier. F. Previtali et al. [25] proposed a technique for extracting features from MRI scans of patients' brains. This technique first uses prior knowledge to determine the locations of brain regions that are strongly associated with Alzheimer's disease, then extracts these spatial locations and their distribution information around the patient's brain, and combines them into new features, which are used as input to the classifier Support Vector Machine (SVM). In the ROI-based method, the MRI image is segmented into different regions of interest, and the features based on the region of interest are used to describe the MRI image. Li et al. [26] predicted mild cognitive impairment progression by segmenting the hippocampus in MRI. In local block-based methods, MRI is simply segmented into local blocks, such as Zhang et al. [27] utilize MRI slice texture features for Alzheimer's disease classification.

Although traditional machine learning-based methods for Alzheimer's diagnosis have high interpretability, they require specialized software operations and a large amount of expert knowledge for feature extraction [28], which can be time-consuming and the expert knowledge is likely to be subject to the subjective individual influence, thus making reproducibility an issue [29]. As a branch of machine learning, deep learning avoids manual feature extraction and is widely used for image feature extraction by constructing end-to-end models. Since Krizhevsky et al. [30] proposed AlexNet and succeeded in the recognition of the ImageNet-2012 dataset, many versions of the CNN architecture have been proposed in recent years [31–33]. Compared with machine learning methods, using CNN for AD classification can provide better flexibility and generalization by fine-tuning the network architecture to adapt to medical datasets [28], which is also verified by the study of Thayumanasamy et al. [34]. Thayumanasamy et al. [34] compared and discussed the traditional machine learning methods such as K-Nearest Neighbors (KNN), Gaussian Naive Bayes, Neural Networks, Decision Tree, and SVM, as well as deep learning methods including AlexNet, VGGNet, ResNet, DenseNet, DarkNet and EfficientNet with their subversions based on MRI data from ADNI, for the task of AD

classification, and the results showed that the deep learning model DenseNet achieved the best overall performance. Inspired by the human biological vision attention mechanism, Qin et al. [35] constructed a 3D U-shaped deep network, HA-ResUNet, with ResNet as the main network, introducing channel and spatial attention mechanisms. HA-ResUNet can automatically allocate attention by model training, which enables the model to focus on more valuable information within different levels of features. Due to the extensive use of medical devices over the past two decades, it is now possible to perform brain imaging scans in different modalities, and some research is being conducted based on multi-modal brain imaging data. Zhang et al. [36] designed two CNN architectures based on 2D slices of MRI and PET scans, respectively. The authors fused the outputs of CNN with MMSE and CDR scores as a decision-level feature for Cognitively normal (CN)/MCI/AD classification. However, using only one slice of an MRI or PET scan is not sufficient to characterize the entire scan, as the affected tissue or region is not expected to be present in all 2D slices. Aderghal et al. [37] improved the performance of network classification through the transfer learning of multi-modal brain scan data, whereas Abdelaziz et al. [38] proposed a multi-modal neural imaging and genetic data fusion approach based on a convolutional neural network that combines the high-level features learned from each modality to jointly recognize AD. To extract and integrate MRI local and global features, Altay et al. [39] proposed a DL architecture combining 3D CNN and RNN to predict preclinical AD, in which 3D CNN extracted local structural features of MRI images, and RNN established connections among these local features. However, modeling 3D MRI images with RNN has two limitations: first, due to the inherent progressiveness, RNN cannot process all local spatial features in parallel [17,40]; second, due to the limited storage capacity of LSTM units, it is very difficult for RNN to model local features with long distances in MRI image space [41]. In addition, although the feature extraction of MRI images using 3D CNN can extract 3D brain atrophy information more intuitively, due to the small data set, the traditional method based on 3D CNN may lead to overfitting.

Transformer provides a new solution for modeling local features of MRI images in parallel. Some studies classified AD based on pre-trained Transformer [42–44]. Kushol et al. [42] and Lyu et al. [43] used a pre-trained Transformer to extract features of 3D MRI slices and assign each slice a predicted value. Finally, a voting mechanism based on multiple predicted values determines the prediction result. The disadvantage of this approach is that feature extraction between slices is independent, which will lose the connection between slices at different positions in 3D MRI. Xing et al. [44] first projected 3D multi-modal PET images into 2D space, then adopted Transformer to extract multi-modal Positron Emission Tomography (PET) image features, then the extracted features are concatenated and sent to classification block to output prediction result. Due to CNN's ability to capture local information, some literature proposed DL models combining CNN and Transformer trained from scratch [39,45]. In [45], 3D CNN represents the original MRI image as a high-level 3D semantic feature map and then uses a pre-trained Transformer to model local spatial features of the 3D feature map. Altay et al. [39] proposed a new DL model integrating 2D CNN and Transformer. The model uses MRI slices at different locations as local spatial features and uses Transformer to fuse the local spatial features extracted by 2D CNN. Nevertheless, the above DL algorithms treat input features equally, which is different from our common sense that spatially adjacent voxels have a stronger semantic correlation than spatially distant voxels. In addition, for limited medical image datasets, feature fusion of input features indiscriminately makes it difficult to integrate local MRI image information, because the model may require a large number of iterative training to make local adjacent features generate a strong contact.

In this paper, we use MRI slices as local spatial features and compare the performance of mainstream CNN architectures in extracting slice planar features, and finally determine our model architecture Conv-Swinformer. We use the 2D CNN+Transformer architecture for AD recognition, which requires fewer parameters for training than using 3D CNN. Conv-Swinformer uses the MRI axial slice group as input, and VGGNet-16 represents the input axial slices group as a set of one-dimensional vectors, which are used as the input of the Transformer. By introducing the shift window mechanism in Transformer, feature fusion is limited to adjacent MRI slices, and the boundary information of adjacent local space is preserved through the second division of the shifted window. Conv-Swinformer can perform fine-grained modeling on the features of each local space of MRI images so that the model can pay more attention to local lesion features through training on small MRI dataset.

3. Proposed method

Fig. 1 depicts the automatic AD classification pipeline based on MRI images proposed in this paper. In the preprocessing pipeline, space normalization, skull dissection, and bias field correction are performed. Then the preprocessed MRI images are sliced, and the N 2D slices are obtained as the input of Conv-Swinformer. Conv-Swinformer can be roughly divided into two modules: CNN module and Transformer encoder module. The CNN module based on VGGNet-16 extracts the planar features of each axial slice and expresses it as a set of vectors (visual tokens), which are used as the input of the Transformer encoder. The Transformer encoder performs a series of weighted fusions of tokens, and the tokens after feature fusion are averaged to obtain the final classification vector. The MLP layer further combines features from different slices and maps the features to the sample category space to obtain the network prediction output.

Algorithm 1 MSA and masked-MSA algorithm

Input: X_W
Output: $\hat{X}_W$
$Q_1, Q_2, ..., Q_h \leftarrow split(X_W W^Q)$ // Generate query vector for each head
$K_1, K_2, ..., K_h \leftarrow split(X_W W^K)$ // Generate key vector for each head
$V_1, V_2, ..., V_h \leftarrow split(X_W W^V)$ // Generate value vector for each head
if MSA **then**
  for each $i \in \{1, 2, ..., h\}$ **do**
    $\hat{X}_W^i \leftarrow soft\max(\frac{Q_i K_i^T}{\sqrt{C}})V_i$
  end for
end if
if masked-MSA **then**
  $MASK \leftarrow zeros(C, C)$ // Generate the all-zero matrix
  $MASK[0 : w/2, w/2 : -1] \leftarrow -\infty$
  $MASK[w/2 : -1, 0 : w/2] \leftarrow -\infty$
  for each $i \in \{1, 2, ..., h\}$ **do**
    $\hat{Z}_W^i \leftarrow soft\max(\frac{Q_i K_i^T}{\sqrt{C}} + MASK)V_i$
  end for
end if
$\hat{X}_W \leftarrow concat(\hat{X}_W^1, \hat{X}_W^2, .., \hat{X}_W^h)W^O$
return $\hat{X}_W$

3.1. VGGNet-16 based CNN module

For each MRI image, take N slices as the input of Conv-Swinformer, and the size of each slice is $182 \times 218 \times 1$. VGGNet-16 based CNN module has 13 convolution layers, see Fig. 1(b). To match the input feature map size of VGGNet-16, we first scale each slice to $112 \times 112 \times 1$ using the bilinear interpolation algorithm. In this paper, VGGNet-16 (excluding 3 fully connected layers) accepts $112 \times 112 \times 3$ (height, weight, and channel) images, so it is necessary to expand the dimensionality of the input grayscale image. Add a convolutional layer with in-channel=1

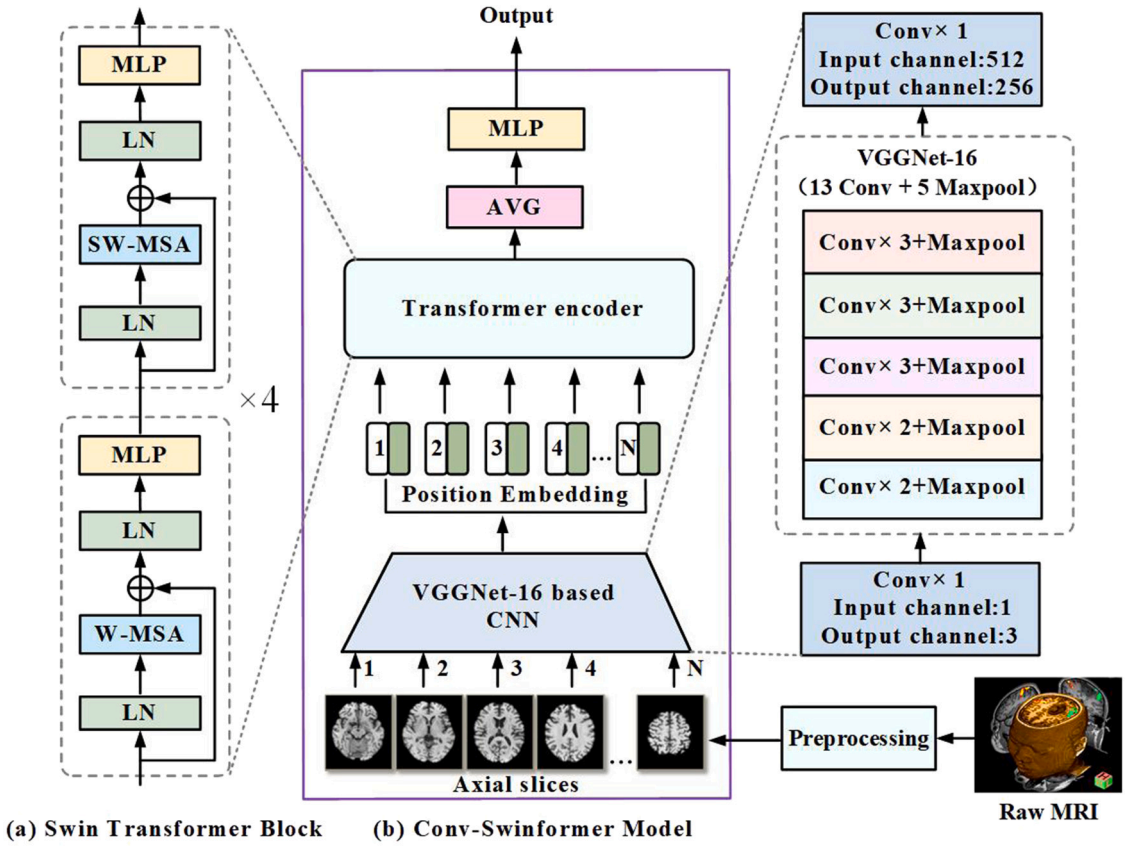


Fig. 1. Proposed MRI-based automatic classification pipeline. (a) Swin Transformer block; (b) the architecture of Conv-Swinformer.

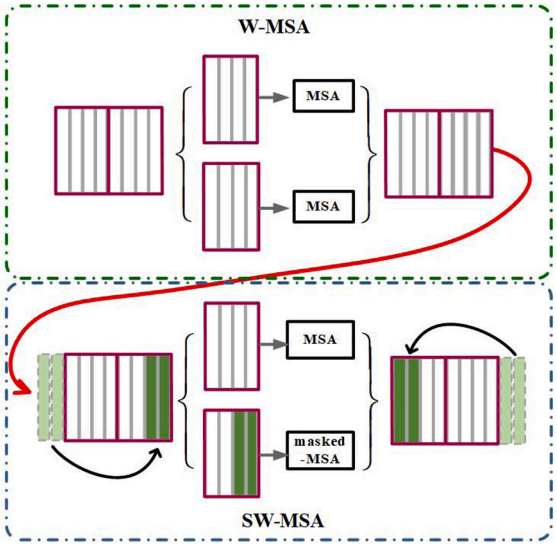


Fig. 2. Window division of W-MSA and SW-MSA modules.

and out-channel=3 for dimension expansion before VGGNet-16. Each convolutional layer contains multiple feature filters of different sizes for extracting features such as texture and shape in MRI slices. After 13 convolutions and activations and 5 pooling operations in VGGNet-16, the output feature map is $3 \times 3 \times 512$. Finally, a convolutional layer is added to the tail of VGGNet-16 for feature map to feature representation mapping, and this convolutional layer is set as in-channel = 512, out-channel = 256. After the last convolutional layer, N 256-dimensional tokens are obtained, representing N 2D slices in each MRI sample. Before N tokens are sent to the Swin Transformer encoder, position embedding is required to mark the position of each token corresponding to the 2D slice in the original 3D MRI image. This study refers to [17] to embed the spatial position within the token sequence. Represent N 256-dimensional tokens as $X \in \mathbb{R}^{N\times256}$, and perform positional embedding on X to get X_{PE}:

$$X_{PE} = X + PE \tag{1}$$

and $PE \in \mathbb{R}^{N\times256}$ is calculated as:

$$\begin{cases} PE_{(pos,2i)} = \sin(pos/10000^{2i/256}) \\ PE_{(pos,2i+1)} = \cos(pos/10000^{2i/256}) \end{cases} \tag{2}$$

where pos represents the position in the token sequence, i represents the ith dimension of a token.

3.2. Transformer

The original Transformer was designed for language translation, and its structure is an encoder–decoder model. The key idea of Transformer is that the self-attention mechanism can be used as the only mechanism to derive dependencies between input and output. In the Transformer for language translation, the input sentence is embedded into a token sequence, the encoder accepts the input token sequence and encodes it into a fixed-length vector, and the decoder takes the vector and decodes it into an output token sequence. In the Transformer-based image classification task, the encoder plays the role of weighted fusion. The image is segmented into patches, and each image patch is represented as a visual token. These visual tokens are sent to the encoder for mutual feature fusion to obtain an output, and the output is directly sent to the classification head for classification. Compared with translation tasks, Transformer does not require a decoder for image classification tasks. Therefore, this paper only uses the encoder in Transformer for research.

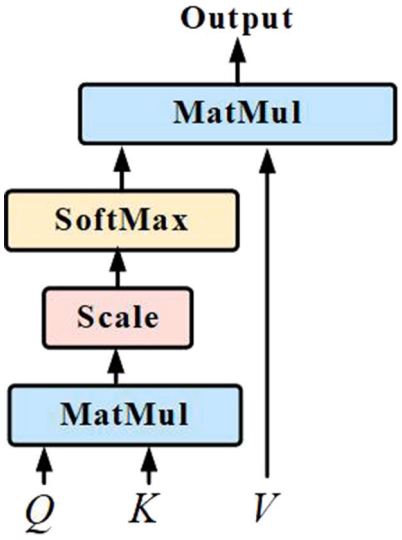


Fig. 3. Self-attention calculation process.

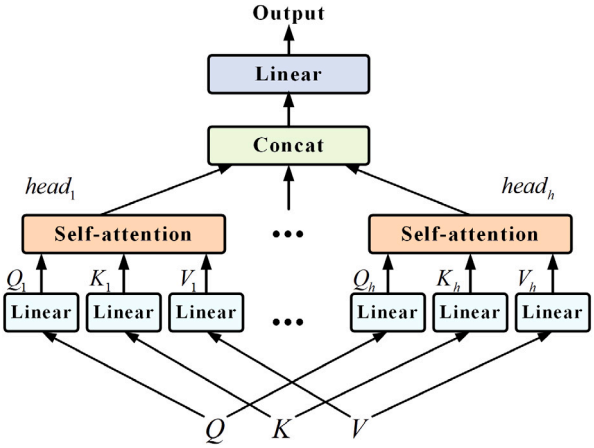


Fig. 4. Multi-head self-attention calculation process.

3.2.1. Self-attention

The self-attention (SA) mechanism is the core of the Transformer, and its essence is to perform a weighted fusion of all input tokens. As shown in Fig. 3, the calculation process of self-attention is as follows: First, a linear map is performed on each token in the input token sequence to obtain a set of query vectors (Q), key vectors (K), and value vectors (V). For each token, the query vector is used to query (multiply) the key vectors of other tokens to obtain the correlation value of this token with other tokens in the sequence. Then, multiply this correlation value with the value vectors of other tokens to obtain the weighted fusion of the token to all tokens in the sequence. SA is expressed by the formula as follows:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{3}$$

where d_K is the embedding dimension of K. When the dimension of K is very large, the result of the dot product will increase in magnitude, thus tending to the saturation area of the softmax activation function, resulting in a very small gradient, therefore, the result after the dot product of Q and K is divided by $\sqrt{d\_K}$ to prevent the gradient disappearing.

3.2.2. Multi-head self-attention

The multi-head self-attention (MSA) mechanism is similar to the channel mechanism in convolution, which is essential for the model to extract multi-channel features. As shown in Fig. 4, based on SA, MSA remaps (or splits) Q, K, and V into multiple groups of Q', K', and V', and performs self-attention within each group of Q', K', and V'. Finally, the results of each group of self-attention operations are connected, and the output is obtained after linear mapping to the original token dimension. MSA is expressed by the formula as follows:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4}$$

$$MSA(Q,K,V) = Concat\left(head_1,\ldots,head_h\right)W^O \tag{5}$$

where $head_i$ is the calculation result of the ith head, W_i^Q, W_i^K, W_i^V are Q, K, and V correspond to the transformation matrix of the ith head, and W^O is used to transform the result of MSA to the original dimension of token.

3.3. Swin transformer encoder module

The brain is made up of several regions that are responsible for different functions, such as the hippocampus responsible for generating new memories, the hypothalamus responsible for regulating daily diet, the amygdaloid responsible for emotional experience and expression, etc. For AD patients, the degeneration of the brain does not always appear in all the above common areas, which requires an accurate grasp of the local lesion characteristics of AD patients for discrimination. In this study, Transformer is used to fuse the plane features of each position of 3D MRI extracted by CNN. For fine-grained feature extraction of local brain lesions, a shift window mechanism is introduced in the Swin Transformer encoder to limit the feature fusion operation to local windows.

As shown in Fig. 1(a), the Transformer encoder is composed of four Swin Transformer blocks, each block consists of two consecutive window MSA (W-MSA) and shifted window MSA (SW-MSA) modules. W-MSA and SW-MSA are shown in Fig. 2. In W-MSA, the token sequence is divided into multiple separate windows, and MSA is performed in each window. However, this partition method will lose window boundary information, that is, there is a lack of connection between adjacent windows. Therefore, in SW-MSA, the original token sequence is cyclically shifted, and the attention window is re-divided to perform in-window self-attention. In this way, each local space of MRI is connected to its adjacent local space. After the SW-MSA, performs a reverse loop shift on the token sequence to restore the original sequence. In special cases, in a Transformer block where the window size is equal to the length of the token sequence, then two W-MSA modules are used in that Transformer blocks instead of alternating W-MSA and SW-MSA modules.

Algorithm 1 is the pseudocode for MSA and masked-MSA calculation, where $X_W \in \mathbb{R}^{w\times C}$ is the token matrix in the window, w is the number of tokens in the window, C is the dimension of the token, $\hat{X}_W$ is the token matrix obtained after self-attention in the window, h is the number of heads in the multi-head attention, W^Q is the query transformation matrix, W^K is the key transformation matrix, W^V is the value transformation matrix, and $W^Q, W^K, W^V \in \mathbb{R}^{C\times C}$.

In this paper, the Transformer encoder includes four Swin Transformers blocks, the settings of four Transformer blocks are shown in Table 1. In the four Swin Transformer blocks, the size of the window is increased layer by layer. The size of the attention window affects the final classification performance. In Conv-Swinformer, the input tokens of the Transformer encoder are continuous (features of adjacent slices in space), so we start with a small window and gradually merge adjacent windows into larger attention windows to be able to fuse more global features. In addition, as the depth of the model increases, the number of attention heads in the window also changes, because as multiple local features are combined in one window, the information

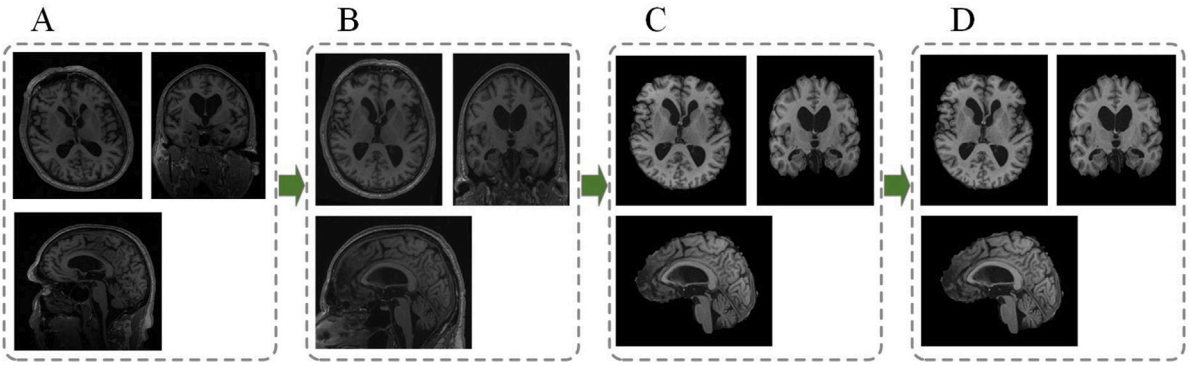


Fig. 5. MRI images preprocessing pipeline. A: Original brain MRI images. B: MRI images after space normalization. C: MRI images after skull dissection. D: MRI images after N4 bias field correction.

Table 1
Hyperparameter settings of 4 swin transformer blocks.

Block num	Module name	Window size	Attention heads
1	W-MSA	8	8
	SW-MSA	8	8
2	W-MSA	16	16
	SW-MSA	16	16
3	W-MSA	32	16
	SW-MSA	32	16
4	W-MSA	96	32
	SW-MSA	96	32

contained in each window is more diverse, so it is necessary to increase the number of attention heads to extract different semantic information to adapt to this change.

The N tokens after feature fusion by the Transformer encoder are averaged to obtain the final classification vector and sent to the final Multilayer Perceptron (MLP) layer. The MLP layer further combines the features from different slices and maps the features into the sample label space, which is the prediction output of the model.

4. Experiments

4.1. MRI dataset preprocessing

The MRI medical imaging data used in this study came from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI is a large clinical medical imaging database that provides all data and samples to scientists around the world [46,47]. Since 2003, ADNI has gone through three stages, ADNI1/GO, ADNI2, and ADNI3. ADNI provides researchers all over the world with a wealth of diagnostic, imaging, and genetic data, covering the pathological research of AD and the computer-aided diagnosis of AD.

Preprocessing of MRI images removes noise generated during data acquisition and helps the model focus on the differences in brain morphology of different categories of patients. The skull shapes of different subjects will have a noisy effect on the training of the model, which will lead to slow or even difficult convergence of the model. As shown in Fig. 5, we preprocess the brain MRI images to strip out brain regions. First, because the MRI images come from different stages of ADNI, the acquisition protocol will change, resulting in differences in the spatial resolution and slice thickness of the obtained MRI scans, so spatial normalization is required. The spatial normalization tool used is FMRIB Software Library (FSL), and the registration template is MNI152. Secondly, FSL is used to remove non-brain tissue voxels. Finally, due to the uneven brightness of the MRI image caused by the magnetic field strength during MRI acquisition, the Advanced Normalization Tools (ANTs) was used to correct the bias field. After preprocessing, all MRI images had a dimension of $182 \times 218 \times 182$ $(X \times Y \times Z)$ with a spatial resolution of $1 \times 1 \times 1$ mm^3 per voxel. After the conditional screening, 2890 MRI images are obtained, including 970 CN, 1412 MCI, and 508 AD. The subject demographics are shown in Table 2. The finally obtained MRI samples were sliced. Due to the limitation of the GPU and considering the effective pixels of the brain tissue contained in the slice, for each MRI, we selected 96 slices between the 43rd and 139th axial slices as the representatives of the MRI images.

Table 2
Demographics of subjects.

Class	Number	M/F	Age	MMSE	CDR
CN	970	460/510	76.72±5.59	28.98±1.27	0.05±0.15
MCI	1412	910/502	76.6±7.4	25.61±4.02	0.61±0.38
AD	508	278/230	76.18±7.75	21.45±4.16	0.94±0.47

4.2. Experimental setup

In this study, we conduct experiments on three binary classification tasks: AD vs. CN, AD vs. MCI, and MCI vs. CN. For each binary classification task, the training validation and test set account for 70%, 15%, and 15% of the total samples respectively, and all using random splits. Additionally, we randomly rotate the slices for data augmentation, and the range of rotation is (−15, 15).

For AD vs. CN binary classification task, the model parameter learning gradient is set to 0.0001, and 100 epochs of training are performed. For AD vs. MCI and MCI vs. CN, the parameter learning gradient of the model is set to 0.00001, and 100 epochs of training are performed. The optimizers for the three training tasks all use SGD [48]. Due to GPU limitations, we take 80 slices per MRI image (N = 80) and set the batch size of the MRI sample to 2. Based on the above experimental settings, the validation loss and validation accuracy curves of Conv-Swinformer in the three binary classification tasks are shown in Figs. 6 and 7 respectively.

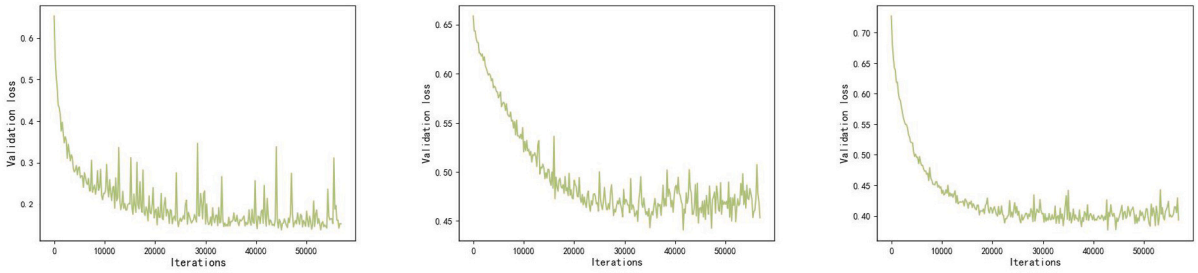
4.3. Results

4.3.1. Evaluation criteria

In this study, the performance metrics of the algorithm included accuracy, sensitivity, specificity, receiver operating curve (ROC), and area under the ROC curve (AUC). Generally, the real situation of all samples is divided into the positive class (severe case) and the negative class (mild case). For example, in AD vs. MCI, AD samples are in the positive class, and MCI samples are in the negative class. The calculation of the accuracy, sensitivity, and specificity are as follows:

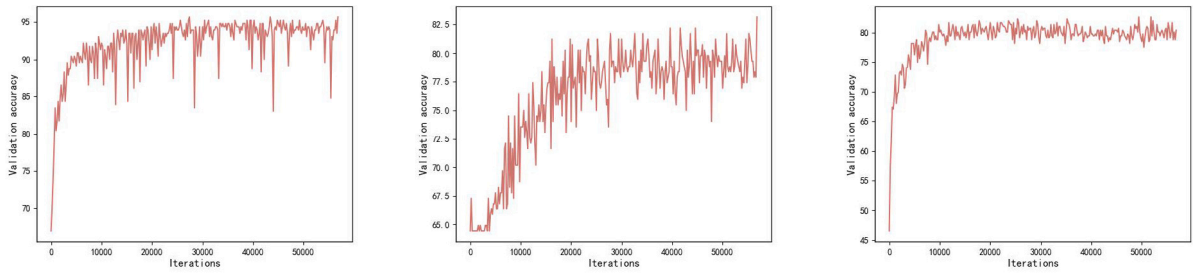
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$



(a) AD vs. CN task (b) AD vs. MCI task (c) MCI vs. CN task

Fig. 6. Validation loss curves in three classification tasks.



(a) AD vs. CN task (b) AD vs. MCI task (c) MCI vs. CN task

Fig. 7. Validation accuracy curves in three classification tasks.

Table 3
The comparative analysis of the various indicators using the proposed method with other representative algorithms.

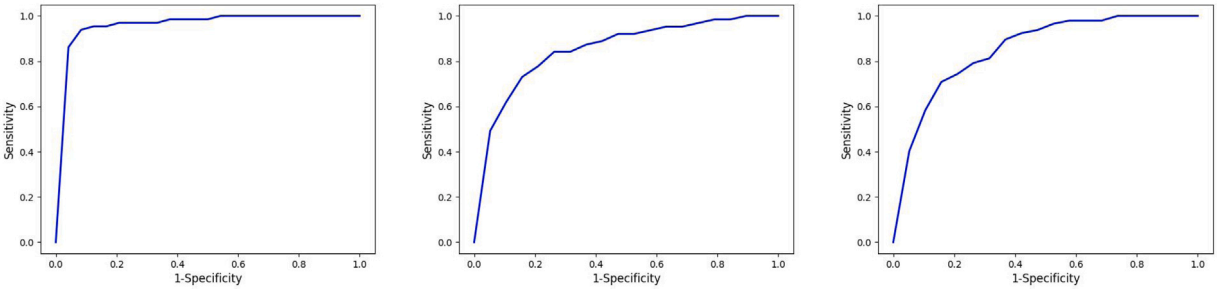
Tasks	Methods	Techniques	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
AD vs. CN	Korolev et al. [49]	3D CNN	87.00	–	–	0.8800
	Addformer [42]	Transformer	88.20	95.60	77.4	–
	Jia et al. [13]	3D CNN+SVM	92.00	**100**	80.00	0.9000
	Advit [44]	3D CNN+Transformer	91.34	–	–	0.9522
	3D CNN [39]	3D CNN	85.23	86.49	83.08	0.8857
	Recurrent Attention [39]	3D CNN+RNN	88.84	95.45	74.29	0.9331
	Attention Transformer [39]	2D CNN+Transformer	91.30	98.70	76.32	0.9503
	Conv-Swinformer	2D CNN+Transformer	**93.56**	93.81	**93.31**	**0.9749**
AD vs. MCI	Korolev et al. [49]	3D CNN	71.00	–	–	0.7700
	Choi et al. [50]	2D CNN	78.10	77.00	**79.30**	–
	3D CNN [39]	3D CNN	79.10	85.51	65.08	0.8127
	Recurrent Attention [39]	3D CNN+RNN	79.60	86.23	65.08	0.8490
	Attention Transformer [39]	2D CNN+Transformer	81.59	**92.03**	58.73	0.8424
	Conv-Swinformer	2D CNN+Transformer	**82.09**	86.96	71.43	**0.8569**
MCI vs. CN	Korolev et al. [49]	3D CNN	73.00	–	–	0.8000
	Li et al. [51]	DenseNets	73.80	**86.60**	51.50	0.7750
	3D CNN [39]	3D CNN	76.95	81.58	73.24	0.8340
	Recurrent Attention [39]	3D CNN+RNN	77.73	80.70	75.35	0.8408
	Attention Transformer [39]	2D CNN+Transformer	**79.07**	63.16	**91.67**	0.8410
	Conv-Swinformer	2D CNN+Transformer	**79.07**	79.82	78.17	**0.8580**

$$Specificity = \frac{TN}{FP + TN} \tag{8}$$

where TP represents the true positive, this is the total number of positive samples predicted to be in the positive class. FN represents the false negative, this is the total number of positive samples predicted to be in the negative class. FP represents the false positive, this is the total number of negative samples predicted as the positive class. TN represents the true negative, this is the total number of negative samples predicted as negative class. Sensitivity represents the detection rate of disease diagnosis, and specificity represents the exclusion rate of normal people. AUC represents the sorting ability of a classifier for samples. Since AUC is not sensitive to whether the number of samples of different categories is balanced, AUC can better reflect the classification performance of the model in unbalanced sample sets [52].

4.3.2. Experimental results

Figs. 6 and 7 show the training process of Conv-Swinformer in three classified tasks. It can be seen that the model converges under the condition of 100 rounds of training in the three tasks. ROC curves of Conv-Swinformer in three tasks are shown in Fig. 8. Altay et al. [39] propose three models based on mainstream deep learning architectures CNN, RNN, and Transformer. To verify the effectiveness of the proposed Conv-Swinformer, based on the data set of this study, we conducted comparative experiments with the three algorithms proposed in [39]. It can be seen that the performance of the 3D CNN model and the Recurrent Attention model is not as good as that of the Transformer-based method, which illustrates the advantage of the Transformer in modeling the global context. In the Transformer-based method, the proposed Conv-Swinformer performs better than the



(a) AD vs. CN task (b) AD vs. MCI task (c) MCI vs. CN task

Fig. 8. ROC curves of Conv-Swinformer in three classification tasks.

Table 4
Comparison of slice feature extractors of VGGNet-16, AlexNet, ResNet-18 and GoogLeNet.

Tasks	CNN Framework	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
AD vs. CN	VGGNet-16	**93.56**	93.81	**93.31**	**0.9749**
	pre-trained VGGNet-16	91.57	95.58	84.62	0.9740
	AlexNet	88.20	95.58	75.38	0.9686
	pre-trined AlexNet	89.89	95.58	80.00	0.9654
	ResNet-18	80.34	92.04	60.00	0.8610
	pre-trined ResNet-18	82.02	**96.46**	56.92	0.8726
	GoogLeNet	85.96	94.69	70.77	0.9191
	pretrained GoogLeNet	84.83	88.50	78.46	0.8952
AD vs. MCI	VGGNet-16	**82.09**	86.96	71.43	0.8569
	pre-trained VGGNet-16	81.09	88.41	65.08	0.8502
	AlexNet	78.11	**91.30**	49.21	0.8320
	pre-trined AlexNet	78.61	88.41	57.42	0.8353
	ResNet-18	79.10	81.59	74.60	**0.8711**
	pre-trined ResNet-18	80.10	81.16	**77.78**	0.8569
	GoogLeNet	74.63	75.36	73.02	0.8348
	pretrained GoogLeNet	77.61	78.99	74.60	0.8432
MCI vs. CN	VGGNet-16	**79.07**	**79.82**	78.17	**0.8580**
	pre-trained VGGNet-16	76.74	71.05	81.25	0.8562
	AlexNet	74.42	70.18	77.78	0.8353
	pre-trined AlexNet	76.74	74.56	78.47	0.8437
	ResNet-18	74.03	64.04	81.94	0.8250
	pre-trined ResNet-18	76.36	70.18	81.25	0.8381
	GoogLeNet	75.97	66.67	83.33	0.8436
	pretrained GoogLeNet	77.13	62.28	**88.89**	0.8249

Table 5
Performance comparison of 4 versions of Conv-Swinformer.

Versions	Settings	AD vs. CN		AD vs. MCI		MCI vs. CN	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
V1	WINs=[2,4,8,16]+HEADs=[4,8,8,16]	0.9101	0.9698	0.7811	0.8290	0.7752	0.8537
V2	WINs=[4,8,16,32]+ HEADs=[8,8,16,16]	0.9213	**0.9763**	0.8109	0.8306	0.7636	0.8491
V3	WINs=[8,16,32,96]+ HEADs=[8,16,16,32]	**0.9356**	0.9749	**0.8209**	**0.8569**	**0.7907**	**0.8580**
V4	WINs=[96,96,96,96]+ HEADs=[32,32,32,32]	0.8989	0.9745	0.7662	0.8376	0.7597	0.8475

Attention Transformer, which proves that the shifted window attention mechanism introduced in the Transformer encoder is more suitable for the classification of AD under a small MRI dataset.

In addition, compared with some ADNI-based studies [13,42,44,49–51], the proposed automated diagnostic pipeline has achieved competitive diagnostic results. We compare the overall effect of a series of methods including experimental setup, data preprocessing, data augmentation, and classification models. As shown in Table 3, for AD vs. CN, the proposed method achieves better performance on accuracy, specificity, and AUC; for AD vs. MCI, the proposed method achieves better performance on accuracy and AUC; for MCI vs. CN, the proposed model achieves better performance on accuracy and AUC. Notably, the proposed algorithm outperforms the comparison algorithms on AUC in all three classification tasks.

4.4. Model performance analysis

4.4.1. Comparison of different CNN architectures

Conv-Swinformer can be divided into two modules: CNN-based slice planar feature extraction module and Transformer based spatial feature fusion module. To investigate which CNN architecture is more suitable for extracting planar features from MRI slices, we compare the performance of AlexNet [30], VGGNet-16 [31], GoogLeNet [32] and ResNet [33] and their pre-trained versions as MRI slice feature extractors. As shown in Table 4, for AD vs. CN classification task, it can be seen that VGGNet-16 outperforms other CNN models in accuracy, specificity, and AUC, and the pre-trained version of VGGNet-16 does not help improve model performance. For the AD vs. MCI classification task, VGGNet-16 achieves the highest accuracy, and AlexNet achieves the highest sensitivity. ResNet-18 with pre-trained parameters and

Table 6
Comparison of Conv-Swinformer performance based on axial, coronal and sagittal plane slices.

Tasks	Metrics	Axial	Coronal	Sagittal
AD vs. CN	Accuracy	**0.9356**	0.9101	0.9213
	AUC	**0.9749**	0.9720	0.9526
AD vs. MCI	Accuracy	**0.8209**	0.7562	0.7962
	AUC	**0.8569**	0.8124	0.8414
MCI vs. CN	Accuracy	**0.7907**	0.7670	0.7614
	AUC	0.8580	**0.8582**	0.8508

original ResNet-18 achieve the highest specificity and AUC, respectively. For MCI vs. CN classification task, VGGNet-16 outperforms other models in accuracy, sensitivity, and AUC metrics. Based on the above analysis, the model has the best overall performance when using the original VGGNet-16 as the MRI slice feature extractor.

4.4.2. Transformer encoder hyperparameter settings

For the Transformer encoder module, we experimentally determine the settings of four Swin Transformer blocks. As shown in Table 5, the elements in WINs represent the size of the attention window in four Transformer blocks, and the elements in HEADs represent the number of self-attention heads within the window in each of the four Transformer blocks. When WINs = [96, 96, 96, 96], the model performs an indiscriminate self-attention operation on all tokens. It can be seen that the model performs best when WINs = [8, 16, 32, 96]. We speculate that this is because when WINs = [2, 4, 8, 16] and WINs = [4, 8, 16, 32], small windows fail to establish long-distance spatial connections in MRI images, while when WINs = [96, 96, 96, 96], larger windows make it difficult for the model to focus on local lesion features, resulting in poor classification performance.

4.4.3. Comparison of three plane slices of MRI

MRI has three plane views: axial, coronal, and sagittal. This work investigates the performance of Conv-Swinformer trained on these three plane slices. As shown in Table 6, for the AD vs. CN and AD vs. MCI classification task, the model trained on axial plane slices achieves the best performance on accuracy and AUC. For the MCI vs. CN, the model trained based on axial plane slices leads on the accuracy. In general, the overall performance of the model is the best when using axial slices, we speculate that this is because AD is accompanied by atrophy of the temporal lobe and hippocampus, and axial slices can clearly show the atrophy of the hippocampus, especially the temporal lobe. However, the AUC on task MCI vs. CN is lower than coronal slices. Therefore, the analysis using only the axial slice group does not fully utilize the spatial information of the MRI images, and how to fuse the features of these three plane slice groups while avoiding model overfitting is a problem that we need to solve in future work.

4.4.4. Ablation experiments

To demonstrate the effectiveness of the shifted window attention mechanism introduced in Conv-Swinformer, two ablation experiments are performed on the Conv-Swinformer model. In the ablation experiment 1, only the W-MSA part in each Swin Transformer block in the Transformer encoder is retained, and the SW-MSA is no longer performed; in the ablation experiment 2, the SW-MSA in each Swin Transformer block in the Transformer encoder is replaced with W-MSA, that is, W-MSA is performed twice in each Swin Transformer block. The experimental results are shown in Fig. 9. In ablation experiment 1 and ablation experiment 2, it can be seen that due to the lack of boundary information fusion between adjacent windows, the classification performance of the model has declined to varying degrees, which proves that the introduced shift window attention helps to improve the classification performance of the model.

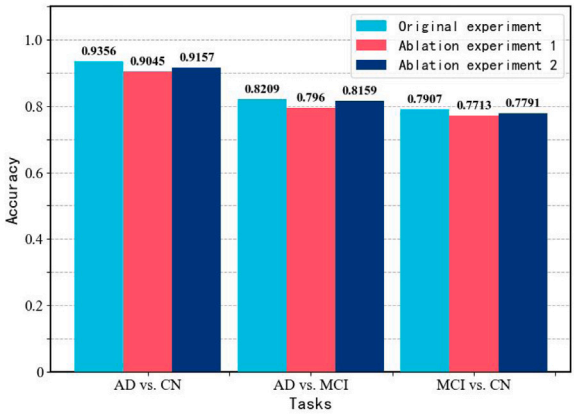


Fig. 9. Comparison of the accuracy of the original experiment and the ablation experiments.

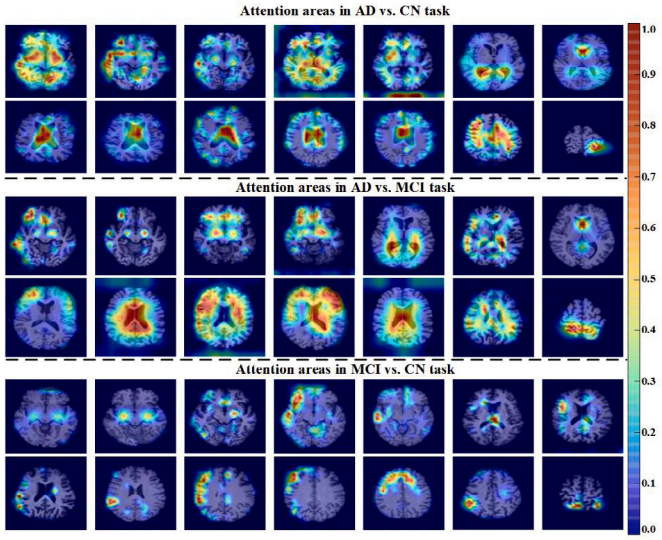


Fig. 10. Conv-Swinformer attention areas on three classification tasks.

4.4.5. Model attention visualization

Attention visualization can improve the interpretability of the model. For clinicians, the attention heat map can mark the attention areas of the model, to quickly find important information and provide strong reference evidence for doctors to make decisions. To show the effectiveness of Conv-Swinformer in feature extraction, this paper uses Grad-CAM [53] to visualize the features extracted by the model. As shown in Fig. 10, compared with the classification task of MCI vs. CN, the attention area of the model is wider in AD vs. CN and AD vs. MCI classification tasks. In the AD vs. CN classification task, the model's attention involves various areas of the cerebral cortex, the most concerned area is the ventricle, and individual samples involve the hippocampus. In the AD vs. MCI classification task, the focus regions of the model are essentially the same as the focus regions of AD vs. CN. However, in the task of MCI vs. CN, the brain regions that the model focused on changed significantly. First, the regions were relatively small and scattered, which also proved that the brain atrophy in the early stage of Alzheimer's disease showed obvious localization. In addition, the brain regions the model focused on involved the hippocampus, amygdala, and parts of the frontal and parietal lobes, suggesting that these regions are more affected in the early stages of Alzheimer's disease.

Table 7
Comparison of model performance under different data enhancement strategies.

Strategy	AD vs. CN		AD vs. MCI		MCI vs. CN	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
No data augmentation	0.9270	0.9730	0.7861	0.8267	0.7713	0.8400
Random rotation	**0.9356**	0.9749	**0.8209**	**0.8569**	**0.7907**	**0.8580**
Mirror inversion	0.9213	**0.9751**	0.7861	0.8062	0.7674	0.8333
Random rotation + Mirror inversion	0.8708	0.9623	0.7512	0.7907	0.7558	0.8229

Table 8
Comparison of model performance on ADNI and OASIS datasets.

Tasks	ADNI		OASIS	
	Accuracy	AUC	Accuracy	AUC
AD vs. CN	**0.9356**	**0.9749**	0.9231	0.9417
AD vs. MCI	0.8209	**0.8569**	**0.8333**	0.8393
MCI vs. CN	**0.7907**	**0.8580**	0.7353	0.7924

5. Discussion

Overfitting can be easily encountered when training a deep learning model on a medical imaging dataset with limited training samples. Data augmentation can enrich the original samples and help the model to better capture the features of the images, increasing the robustness and accuracy of the deep learning model. Based on MRI axial slices, we compared the performance of models using random rotation and mirror reflection data augmentation strategies, as shown in Table 7. It can be seen that when the model adopts the random rotation strategy, the three tasks' test performance improved. When using mirror reflection, there is a small improvement in the AD vs. CN task, and no noticeable improvement in the AD vs. MCI, MCI vs. CN tasks. When both random rotation and mirror reflection strategies are adopted, there is no improvement in the test performance of the three tasks. This is because the improper data augmentation may bring noise to the model due to a lack of valid information and context, resulting in instability during the training process, thus affecting the accuracy of prediction.

In computer-aided diagnosis, the generalizability of the classifier is an important evaluation criterion. The higher generalizability on different datasets, the more practical the classifier is to make good predictions on new data. We conducted experiments based on the cross-sectional MRI dataset (OASIS-1) in the OASIS dataset to verify the generalization of the proposed model[2]. The OASIS-1 dataset followed 416 subjects, including 218 aged between 18 and 59 and 198 aged between 60 and 96, to evaluate their dementia severity through the CDR scale. Since there were no dementia patients in the younger subjects in the dataset, this study used 198 MRI images from OASIS-1, dividing them into 98 normal (CDR = 0), 70 mild dementia (CDR = 0.5), and 30 severe dementia (CDR > 0.5) according to the CDR scores. Additionally, due to the OASIS-1 dataset also provides preprocessing methods such as skull stripping and bias field correction, the proposed model was tested on the preprocessed 198 images. The test results on the OASIS-1 dataset are shown in Table 8, it can be seen that the model achieved good generalization in the classification tasks of AD vs. CN and AD vs. MCI. In the AD vs. MCI classification task, the test accuracy of the model on the OASIS dataset even exceeded that on the ADNI dataset.

From the above experimental results, it can be inferred that MCI recognition is more difficult than the classification between AD and CN due to the subtle differences between MCI and the other two patient types in brain structures. In this study, to more accurately capture these subtle differences, we introduced the shifted window attention mechanism in the Transformer encoder, limiting the model's attention to the local spatial region of the MRI image, thus enhancing the extraction of local fine-grained features and achieving better classification results. This paper directly transplanted existing mainstream CNN architectures without exploring the relationship between the number of CNN layers and the model's classification performance. In the next step, we can try to adjust the number of CNN layers on the existing basis to reduce the total parameters of the model, hoping to further alleviate overfitting and improve model performance. We study the classification task of Alzheimer's disease based on deep learning technology, considering the heterogeneity of the local area distribution of brain atrophy in patients, and focusing more on the extraction of local fine-grained features. However, this study does not distinguish between Alzheimer's disease with different patterns of brain degeneration, including natural aging. In addition, the performance of the model proposed in this paper is obtained through experiments on existing datasets and has not been verified by real clinical MRI data. Therefore, cooperating with neuroscientists to verify the clinical value of the results of this paper and focusing on personalized diagnosis and treatment of Alzheimer's disease is also the direction to further improve this research.

6. Conclusion

This paper introduces a Convn-swinformer model combining CNN and Transformer for pairwise classification of CN, MCI, and AD. In Conv-Swinformer, VGGNet-16 based CNN extracts low-level planar features from 2D MRI slices, and Transformer performs further feature fusion on the extracted features. To better integrate local planar features, We introduced a shift window mechanism in Transformer to limit the fusion of planar features to the divided attention window, in other words, the local space. Through the cyclic displacement of planar features and the second division of the attention window, adjacent local spaces preserve the boundary information of each other. Experiments have proved that compared with treating the features of each plane of MRI equally, Conv-Swinformer can fine-grainedly fuse the features of nearer planes in MRI 3D images, so that the model can focus on local tiny brain atrophy for AD discrimination.

This study uses 2D CNN to extract features from MRI slices and uses Transformer to re-establish spatial connections, which avoids the shortcomings of 3D CNN in extracting MRI global spatial information to a certain extent, but 2D CNN still has this problem when extracting MRI slice information. In the next step, we plan to introduce the attention mechanism into the 2D slice, hoping to have a finer-grained extraction of MRI features. MCI is a pre-AD stage, and the brain atrophy of MCI patients is the most rapid in the course of the disease. Therefore, longitudinal research on MCI patients can hopefully predict the development trend of patients, which is of great clinical significance. We will explore longitudinal studies of AD in the future in the hope of better classification or prediction performance.

Funding

This research is supported by the National Natural Science Foundation of China(Grant No. 61976080), the Academic Degrees and Graduate Education Reform Project of Henan Province (Grant No. 2021SJGLX195Y), the Key Project on Research and Practice of Henan University Graduate Education and Teaching Reform (Grant No. YJSJG2023XJ006), and the Innovation and Quality Improvement Project for Graduate Education of Henan University (Grant No. SYLKC2023016, No. SYLYC2023191, No. SYLYC2022191, No. SYLYC2022192).

[2] The OASIS data repository can be accessed via http://www.oasisbrains.org.

Declaration of competing interest

The authors declared no conflict of interest.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI), United States (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81X-WH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, United States, 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.

References

[1] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, Dementia prevention, intervention, and care: 2020 report of the lancet commission, Lancet 396 (2020) 413–446.

[2] J.M. Long, D.M. Holtzman, Alzheimer disease: An update on pathobiology and treatment strategies, Cell 179 (2019) 312–339.

[3] J. Gaugler, B. James, T. Johnson, J. Reimer, M. Solis, J. Weuve, R.F. Buckley, T.J. Hohman, Alzheimer's disease facts and figures, Alzheimers Dement 18 (2022) (2022) 700–789.

[4] M.G. Ulep, S.K. Saraon, S. McLea, Alzheimer disease, J. Nurse Pract. 14 (2018) 129–135.

[5] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C.E. Teunissen, J. Cummings, W.M. van der Flier, Alzheimer's disease, Lancet. 397 (2021) 1577–1590.

[6] B. Lam, M. Masellis, M. Freedman, D.T. Stuss, S.E. Black, Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome, Alzheimers. Res. Ther. 5 (2013) 1–14.

[7] M.E. Murray, N.R. Graff-Radford, O.A. Ross, R.C. Petersen, R. Duara, D.W. Dickson, Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: A retrospective study, Lancet Neurol. 10 (2011) 785–796.

[8] J.L. Whitwell, D.W. Dickson, M.E. Murray, S.D. Weigand, N. Tosakulwong, M.L. Senjem, D.S. Knopman, B.F. Boeve, J.E. Parisi, R.C. Petersen, Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: A case-control study, Lancet Neurol. 11 (2012) 868–877.

[9] K.A. Jellinger, Neuropathological subtypes of Alzheimer's disease, Acta Neuropathol. 123 (2012) 153–154.

[10] P. Chen, H. Yao, B.M. Tijms, P. Wang, D. Wang, C. Song, H. Yang, Z. Zhang, K. Zhao, Y. Qu, Four distinct subtypes of Alzheimer's disease based on resting-state connectivity biomarkers, Biol. Psychiatry. 93 (2023) 759–769.

[11] H.-I. Suk, S.-W. Lee, D. Shen, A.D.N. Initiative, Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, Neuroimage 101 (2014) 569–582.

[12] H.-I. Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification, in: Int. Conf. Med. Image Comput. Comput. Interv., Springer, 2013, pp. 583–590.

[13] H. Jia, H. Lao, Deep learning and multimodal feature fusion for the aided diagnosis of Alzheimer's disease, Neural Comput. Appl. 34 (2022) 19585–19598.

[14] S. Fathi, M. Ahmadi, A. Dehnad, Early diagnosis of Alzheimer's disease based on deep learning: A systematic review, Comput. Biol. Med. (2022) 105634.

[15] Y. AbdulAzeem, W.M. Bahgat, M. Badawy, A CNN based framework for classification of Alzheimer's disease, Neural Comput. Appl. 33 (2021) 10415–10428.

[16] M.A. Abdou, Literature review: Efficient deep neural networks techniques for medical image analysis, Neural Comput. Appl. (2022) 1–22.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Adv neural inf process syst, Neural Info. Process. Syst. 30 (2017).

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Eur. Conf. Comput. Vis., Springer, 2020, pp. 213–229.

[19] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H.S. Torr, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6881–6890.

[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 Prepr. arXiv:2010.11929.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Int. Conf. Mach. Learn., PMLR, 2021, pp. 10347–10357.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10012–10022.

[23] N. Mahendran, D.R.V. PM, A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease, Comput. Biol. Med. 141 (2022) 105056.

[24] A.L. Benedet, N.J. Ashton, T.A. Pascoal, A. Brinkmalm, J. Nilsson, H. Kvartsberg, S. Mathotaarachchi, M. Savard, J. Therriault, C. Tissot, SNAP25 reflects amyloid-and tau-related synaptic damage: Associations between PET, vbm and cerebrospinal fluid biomarkers of synaptic disfunction in the Alzheimer's disease spectrum: Neuroimaging: imaging the human synapse in AD, Alzheimer's Dement. 16 (2020) e046358.

[25] F. Previtali, P. Bertolazzi, G. Felici, E. Weitschek, A novel method and software for automatically classifying Alzheimer's disease patients by magnetic resonance imaging analysis, Comput. Methods Programs Biomed. 143 (2017) 89–95.

[26] H. Li, M. Habes, D.A. Wolk, Y. Fan, A.D.N. Initiative, A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data, Alzheimer's Dement. 15 (2019) 1059–1070.

[27] Y. Zhang, S. Wang, Y. Sui, M. Yang, B. Liu, H. Cheng, J. Sun, W. Jia, P. Phillips, J.M. Gorriz, Multivariate approach for Alzheimer's disease detection using stationary wavelet entropy and predator–prey particle swarm optimization, J. Alzheimer's Dis. 65 (2018) 855–869.

[28] T. Illakiya, R. Karthik, Automatic detection of Alzheimer's disease using deep learning models and neuro-imaging: Current trends and future perspectives, Neuroinformatics 21 (2023) 339–364.

[29] J. Samper-González, N. Burgos, S. Bottani, S. Fontanella, P. Lu, A. Marcoux, A. Routier, J. Guillon, M. Bacci, J. Wen, Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data, Neuroimage 183 (2018) 504–521.

[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM. 60 (2017) 84–90.

[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, ArXiv Prepr. arXiv:1409.1556.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[34] I. Thayumanasamy, K. Ramamurthy, Performance analysis of machine learning and deep learning models for classification of Alzheimer's disease from brain MRI, Trait. Du Signal. 39 (2022) 1961.

[35] Z. Qin, Z. Liu, Q. Guo, P. Zhu, 3D convolutional neural networks with hybrid attention mechanism for early diagnosis of Alzheimer's disease, Biomed. Signal Process. Control. 77 (2022) 103828.

[36] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, X. Zhang, Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease, Neurocomputing 361 (2019) 185–195.

[37] K. Aderghal, K. Afdel, J. Benois-Pineau, G. Catheline, Improving Alzheimer's stage categorization with convolutional neural network using transfer learning and different magnetic resonance imaging modalities, Heliyon 6 (2020).

[38] M. Abdelaziz, T. Wang, A. Elazab, Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks, J. Biomed. Inform. 121 (2021) 103863.

[39] 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, in: Proc. AAAI Conf. Artif. Intell., 2021, pp. 15088–15097.

[40] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin, Convolutional sequence to sequence learning, in: Int. Conf. Mach. Learn., PMLR, 2017, pp. 1243–1252.
[41] W. Zhang, T. Zhu, C. Yang, J. Xiao, H. Ning, Sensors-based human activity recognition with convolutional neural network and attention mechanism, in: 2020 IEEE 11th Int. Conf. Softw. Eng. Serv. Sci., IEEE, 2020, pp. 158–162.
[42] 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 Int. Symp. Biomed. Imaging, IEEE, 2022, pp. 1–5.
[43] Y. Lyu, X. Yu, D. Zhu, L. Zhang, Classification of Alzheimer's disease via vision transformer: Classification of Alzheimer's disease via vision transformer, in: Proc. 15th Int. Conf. PErvasive Technol. Relat. to Assist. Environ., 2022, pp. 463–468.
[44] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, N. Jacobs, Advit: Vision transformer on multi-modality pet images for Alzheimer disease diagnosis, in: 2022 IEEE 19th Int. Symp. Biomed. Imaging, IEEE, 2022, pp. 1–4.
[45] 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, in: 2022 IEEE 19th Int. Symp. Biomed. Imaging, IEEE, 2022, pp. 1–5.
[46] C.R. Jack Jr., M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging An Off. J. Int. Soc. Magn. Reson. Med. 27 (2008) 685–691.
[47] M. Liu, J. Zhang, P.-T. Yap, D. Shen, View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data, Med. Image Anal. 36 (2017) 123–134.
[48] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. (1951) 400–407.
[49] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3D brain MRI classification, in: 2017 IEEE 14th Int. Symp. Biomed. Imaging, ISBI 2017, IEEE, 2017, pp. 835–838.
[50] B.-K. Choi, N. Madusanka, H.-K. Choi, J.-H. So, C.-H. Kim, H.-G. Park, S. Bhattacharjee, D. Prakash, Convolutional neural network-based mr image analysis for Alzheimer's disease classification, Curr. Med. Imaging 16 (2020) 27–35.
[51] F. Li, M. Liu, A.D.N. Initiative, Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks, Comput. Med. Imaging Graph. 70 (2018) 101–110.
[52] C.X. Ling, J. Huang, H. Zhang, AUC: A better measure than accuracy in comparing learning algorithms, in: Conf. Can. Soc. Comput. Stud. Intell., Springer, 2003, pp. 329–341.
[53] 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: Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.