Knowledge-Based Systems 238 (2022) 107942

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Diagnosis of Alzheimer's disease via an attention-based multi-scale convolutional neural network

Zhenbing Liu, Haoxiang Lu, Xipeng Pan, Mingchang Xu, Rushi Lan *, Xiaonan Luo

Guangxi Key Laboratory of Intelligent Processing Computer Image and Graphics, Guilin University of Electronic Technology, Guilin 541004, China

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 15 November 2021
Accepted 10 December 2021
Available online 17 December 2021

Keywords:
Alzheimer's disease
Multi-scale CNN
White matter
Gray matter
Attention mechanism

ABSTRACT

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases. Accurate diagnosis of mild cognitive impairment (MCI) in the prodromal stage of AD can delay onset. Therefore, the early diagnosis of AD is particularly essential. The convolutional neural network (CNN) extracts feature of image layer-by-layer, and the observed features are obtained by setting different receptive fields. However, the brain structure is very complicated, and the etiology of AD is unknown, in addition, most of the existing methods do not consider the details and overall structure of the image. To address this issue, we propose a novel multi-scale convolutional neural network (MSCNet) to enhance the model's feature representation ability. A channel attention mechanism is introduced to improve the interdependence between channels and adaptively recalibrate the channel direction's characteristic response. To verify the effectiveness of our method, we segment the original MRI data to obtain white matter (WM) and gray matter (GM) datasets and train the model. Extensive experiments show that our method obtains the state-of-the-art performance with fewer parameters and lower computational complexity.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative sickness, accompanied by cognitive impairment and behavioral defects. It can be classified as Mild Cognitive Impairment (MCI), Normal Control (NC) and AD according to clinical symptoms [1,2]. The number of people with dementias has witnessed a sharp increased with the aging of the global population. Research showed that approximately 50 million people suffered from dementia, and 60% to 70% of the population had AD in 2018 [3–5]. Therefore, early diagnosis and intervention (such as treatment) can delay the onset of AD.

Medical imaging technology has made unprecedented progress, and neuroimaging has gradually become the primary application of computer-aided diagnosis. Among them, magnetic resonance imaging (MRI) has been widely employed in the diagnosis of AD due to its high-resolution, non-invasive, and multi-directional imaging. It also can distinguish soft tissues and clearly display normal anatomical structures of the brain, such as white matter (WM) and gray matter (GM) [6,7]. GM in the brain is mainly composed of neuron cells and plays a dominant role in the nerve center. The thinking signals in human brain are generated in GM, and the nerve fibers that constitute WM are responsible for the transmission of nerve impulses. Researchers found that the brain MRI of patients diagnosed with AD is accompanied by varying degrees of WM-damage and GM-atrophy, which is also an important indicator in the clinic [8,9]. In order to explore the effectiveness of GM and WM in the diagnosis of AD, corresponding experiments are carried out in this work.

In the past few decades, machine learning has been widely used in AD diagnosis, which extracts the required features from AD data to fit the model and forecast unknown data categories. Generally, AD is diagnosed by manually extracting the characteristics of the region of interest (ROI), such as temporal lobe, entorhinal cortex volume, bilateral hippocampus, GM, WM, and thickness structure changes, etc. [10,11]. In addition, AD researchers have also made great efforts in the multi-modal multi-task learning. For example, Sheng et al. [12] used the combination of multimodal segmentation of the joint human connective project and logistic regression-recursive feature elimination to accurately identify the different stages of AD. Zeng et al. [13] proposed a multi-task learning algorithm based on deep confidence network (DBN). Through feature selection method of multi-task, the intrinsic correlation between multiple related tasks is considered and feature sets related to all tasks are selected. Dropout technology and zero masking strategy are used to overcome the problem of model overfitting. Lodewijk et al. [14] proposed a joint multimodal longitudinal regression and classification method for early diagnosis of AD. Prakash et al. [15] trained single-mode and multi-mode regression models based on baseline data including demographic, neuroimaging, cerebrospinal fluid markers, and genetic factors to predict future ADAS-COG scores. The process is

* Corresponding author.
E-mail address: rslan2016@163.com (R. Lan).

https://doi.org/10.1016/j.knosys.2021.107942
0950-7051/

relatively cumbersome, and the classification performance needs to be improved. If the method can automatically generate extracted features, it will solve the inaccuracy caused by manual feature extraction and significantly improve work efficiency.

Recently, deep learning has performed well in natural image classification, recognition, and detection [16–18]. It can automatically extract features without manual preprocessing. Researchers found that convolutional neural network (CNN) can also be used to extract medical image features for analysis [19]. Although CNN-based methods have been gradually used in AD diagnosis, there are still some limitations that need to be solved urgently. (1) The brain MRI structure is relatively complicated, and the cause of AD is not fully understood, most of the existing CNN-based methods are single scale representative features, and it is impossible to analyze the image information on the deep structure. (2) Most of the deformation of CNN is only stretched in the depth and width of the network structure, resulting in the model parameters that are too large to be used in practical applications.

In this paper, a novel multi-scale convolutional neural network (MSCNet) model for AD diagnosis was put forward. Unlike most existing CNN models, to avoid single-level of feature extraction and take full advantage of the efficiency of CNN, we provide a multi-scale structure and introduce an attention mechanism to double-weight the learned weight of neural network. The experimental results on WM and GM datasets show the effectiveness of our method, and verify that WM is more effective than GM in AD diagnosis. The main contributions are as follows.

(1) We propose a discriminative feature representation method for MRI of AD. A novel multi-scale residual block (MSRB), which does not rely on multiple convolution kernels of different sizes, and achieves the effectiveness of changing the size of the convolution kernel by setting different dilation rates, is used to extract multi-scale features and fuse the features between channels to obtain more comprehensive information.

(2) To improve the classification performance, a channel attention mechanism with global average pooling and global maximum pooling was introduced to learn the dependency relationship between each channel and assign corresponding weight coefficients.

(3) A unite network based on multi-scale convolutional neural network (MSCNet-U) is employed to solve the label expansion problem caused by slicing on the 3D volume.

The rest of this paper is organized as follows. In Section 2, we briefly review some related works of CNN in AD diagnosis and the multi-scale feature structure. In Section 3, the data preprocessing and the proposed method are described in detail. Section 4 shows the comparative results. Finally, the conclusion of this paper is drawn in Section 5.

2. Related works

2.1. CNN-based diagnosis of Alzheimer's disease

In recent years, AD diagnosis based on CNN has achieved good performance [20,21], and end-to-end training saves a large number of feature extraction operations. Sarraf et al. [22] used LeNet-5, one of the earliest convolutional neural networks, to train the AD/NC on the functional magnetic resonance imaging (fMRI) dataset, and the classification results was 96.86%. To relieve overfitting, Billones et al. [23] added a dropout after each pooling layer of VGG-16 to classify the AD/MCI/NC, and the accuracy was 91.85%. Farooq et al. [24] trained GoogLeNet, ResNet-18, and ResNet-152 on GM datasets, and all ran well. Sathiyamoorthi et al. [25] first preprocessed the image by adaptive histogram adjustment and 2D adaptive bilateral filter (2D-ABF), and then extracted the region of interest using the adaptive mean shift modified expectation maximization (AMS-MEM) algorithm. Finally, the 2-Dimensional Gray Level Co-Occurrence Matrix (2D-GLCM) was used to calculate the features and the classification decision was made by deep convolutional neural network. Almadhoun et al. [26] passed the entire brain image by the transmission of Xception learning architectures and then CNN, which is constructed by separable convolutional layer, was used to learn the general features of the image, and classifies them. Abdulazeem et al. [27] proposed an end-to-end framework for AD-classification based on CNN, realizing multi-classification at different stages of AD. Bringas et al. [28] employed CNN to analyze data from 35 patients with AD collected by smartphone, and found that mobility data can be a valuable resource for the treatment of patients with AD as well as to study the progress of the disease. Bae et al. [29] developed a convolutional neural network (CNN)-based AD classification algorithm using magnetic resonance imaging (MRI) scans from AD patients and age/gender-matched cognitively normal controls from two populations that differ in ethnicity and education level. Basheera et al. [9] used hybrid enhanced independent component analysis to collect segmented gray matter mri image, and then differentiation MCI and CN more accurately by CNN to escalate early diagnosis of AD.

Researchers [5,30] have found that the information of various modalities can complement each other and obtain better results. Khvostikov et al. [31] proposed an improved 3D-CNN based on twin network which inputs n features of ROI of structural magnetic resonance imaging (sMRI) and diffusion tensor imaging (DTI) images into the pipeline in parallel. Then the features obtained by the n pipelines were passed through the flattened layer to obtain n vectors and spliced. Finally, the result was output through the fully connected layer (FC), dropout, and Softmax layer. Like multi-modal data information complementation, integrated learning combines multiple weakly supervised models to obtain a better and more comprehensive strongly supervised model. Kumar et al. [32] fine-tuned AlexNet and GoogLeNet to integrate with the AD dataset, and the final classification results were better than those of the single network. Feng et al. [33] combined three-dimensional convolutional neural networks (3D-CNNs) with MRI to execute binary and ternary disease classification models. Xia et al. [34] exploited a 6-layer 3D CNN to learn informative features of sMRI, then 3D CLSTM is leveraged to further extract the channel-wise higher-level information to realize AD diagnosis. Parmar et al. [35] presented one such synergy of fMRI and deep learning, where using a modified 3D-CNN to resting-state fMRI data for feature extraction and classification of AD. Folego et al. [36] proposed an end-to-end deep 3D CNN for the multiclass AD biomarker identification task, using the whole image volume as input. Shmulev [37] and Senanayake [38] built 3D-ResNet and 3D-DenseNet, respectively, but their experimental results were not satisfactory. However, the small number of MRI data make the 3DCNN-based network difficult to fit completely. At the same time, the network has larger number of parameters and longer training time. Therefore, we propose a lightweight network for end-to-end training in this paper.

2.2. Multi-scale network

Multi-scale networks can extract different feature information of ROI at different scales, including the information of both local and global. It is proven that improves the performance of the models in many areas [39,40], such as image classification [41], target detection [42–44] and segmentation [45,46].

Before the emergence of Inception [41] networks, most CNNs [47,48] achieved better performance by stacking convolutional layers and deepening the number of network layers, but this will inevitably lead to a considerable increase in the number of calculations. Due to the different positions of the image information,

the choice of the size of the convolution kernel is more challenging. If the image information distribution is global, it is suitable for the case where the convolution kernel is large. Conversely, it is suitable for smaller convolution kernels. The inception module uses 1×1, 3×3, and 5×5 convolution kernels and pooling in the same layer network to obtain different sizes receptive fields of the image. Finally, the channels are superimposed so that the features of different scales are fused and passed to the next layer. Gao et al. [49] proposed a new deep learning multi-scale structure Res2Net. It reconstructs the bottleneck block in ResNet, replacing the $1-3-1$ residual distribution with a hierarchical structure. In this module, the central convolution is divided into multiple groups. A group of filters first extracts features from a set of input feature maps. Then the output feature maps of the previous group are sent to the next set of filters together with another set of input feature maps. This process is repeated several times until all input feature maps have been processed. Finally, the feature maps of all groups are concatenated and sent to another 1×1 filter to thoroughly merge the information. These multi-scale structures perform well, but the network structure is more complex and has more parameters. Liu et al. [50] proposed a multi-scale multi-view re-sampling and color projection method for nodules, and applied it to process CT images. Zhu et al. [51] used filters of the same size to carry out multi-layer convolution and combined extended convolution with common convolution at different rates to design multi-scale modules to enrich features extracted from multi-layer convolution. In addition, multi-layer convolution and residual block cascade are used for image denoising. Ibtissam et al. [52] employed gaussian pyramid representation for multi-scale analysis to distinguish between normal and abnormal breast tissues. In this paper, we propose a simple, efficient, and lightweight multi-scale structure for diagnosing brain medical images.

3. Material and method

In this section, we will introduce the experimental materials and preprocess method, then analyze the framework of our method in detail, and finally the activation function and loss function are presented.

3.1. Material

Statistical parametric mapping (SPM), a medical statistical analysis software [53], is used to preprocess the MRI brain image data. The pipeline of data preprocess is shown in Fig. 1.

First, slice timing is employed on the original data to correct the difference in acquisition time between layers in a volume. Then, the head movement correction is used to eliminate slight head movement during the test, so that each frame of the experimental sequence is aligned with the first frame of the sequence. Next, to unify the data input size during training, the data need to be resampled, and the data size is adjusted to size during training, the data need to be resampled, and the data size is adjusted to $192 \times 192 \times 160$. The above is the medical professional processing flow for the AD dataset, which is recorded as stage I. Then, the tissue structure of the resampled 3D MRI was segmented, and WM and GM datasets were selected from the divided parts. Finally, Python and tools for Nifti and analyzing images were used to slice the GM and WM, which were recorded as stage II.

Alzheimer's Disease Neuroimaging Initiative (ADNI) [54] was established in 2003 to collect brain imaging data from multiple modalities, such as structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), and positron emission computed tomography (PET). There are many types of data in the ADNI, such as AD, MCI, NC, etc. Furthermore, ADNI will regularly track and update the dataset of the pre-diagnosed population. The initial MRI data of AD that used in the experiment including 160 AD, 200 MCI, and 160 NC were all from the ADNI. The data was preprocessed by SPM, and the clearest brain structures were taken as the experimental WM and GM datasets. The AD dataset is shown in Fig. 2. The ratio of training subsets to test subsets in our experiment is 7:3. When training and testing, we strictly divided the training subsets and test subsets according to the slices under the same volume.

Table 1
The accuracy and running time obtained in the MCI/NC of WM for slices 8, 12, 16, 20, 24, 28, 32 and 36.

Slice	8	12	16	20	24	28	32	36
Acc (%)	86.89	88.94	90.28	90.83	90.85	90.87	90.85	90.86
Time (ms)	44.69	46.38	47.69	49.25	55.98	59.64	63.39	70.04

Table 2
The accuracy in the MCI/NC of WM for randomly select 20 slices and continuously select the middle 20 slices.

Slice	20r	20c
ACC (%)	72.31	90.83

One piece of 3D data can be sliced into 160 pieces when we sliced, and we found that only the middle slices had relatively complete structural information. In order to determine the optimal number of slices, intermediate continuous slices of 8,12,16,20,24,28,32, and 36 were selected for comparison. The results of 8, 12, 16, 20, 24, 28, 32 and 36 slice are shown in Table 1, We can easily find that when the number of slices exceeds 20, the classification accuracy of the model is similar, but the training time is greatly increased. When the number of slices ranged from 8 to 20, both the classification accuracy and training accuracy increased but the running time increased slightly. Therefore, we select 20 consecutive and intermediate slice images in each 3D MRI and there are a total of 10,400 images for all classes.

In addition, we also randomly selected 20 images for experiments to determine the rationality of our selection method. The accuracy in the MCI/NC of WM is shown in Table 2. Among them, 20r represents randomly select 20 slices and 20c represents continuously select the middle 20 slices. Due to the selection of continuous middle 20 slices contained relatively complete structural information, so our method is more efficient and more reasonable.

3.2. Network architecture

Based on the multi-scale residual block integrated attention mechanism (MSAM), we proposed a novel AD diagnosis model, denoted as the multi-scale convolutional neural network (MSCNet) which can divide into five main convolutional blocks. We use ResNet-50 as the backbone network structure of MSCNet, the structure of MSCNet is shown in Fig. 3. The first convolution raises the number of channels of the network from 3 to 64, and the remaining convolution blocks are composed of MSAM modules. The number of channels is 64, 128, 256, and 512 in sequence, and the corresponding number of MSAM channels is 3, 4, 6, and 3.

Additionally, in order to solve the problem of label expansion caused by slicing on the 3D volume, we propose a multi-slice feature unite network based on MSCNet, denoted as MSCNet-U. The structure of MSCNet-U is shown in Fig. 4, there are 20 pipelines in MSCNet-U, and each pipeline has the same structure and parameters. We use the 3D volume obtained by putting the 20 slices into a separate pipeline, merge the feature vectors in the last layer of convolution operation.

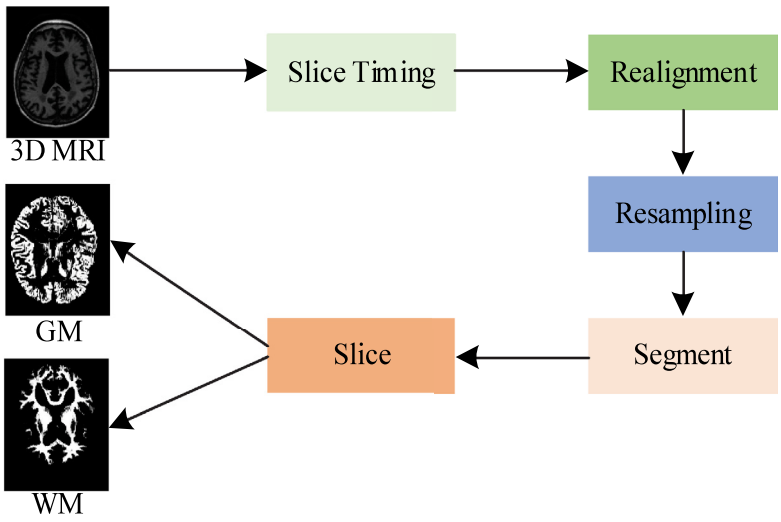

Fig. 1. The pipeline of data preprocess.

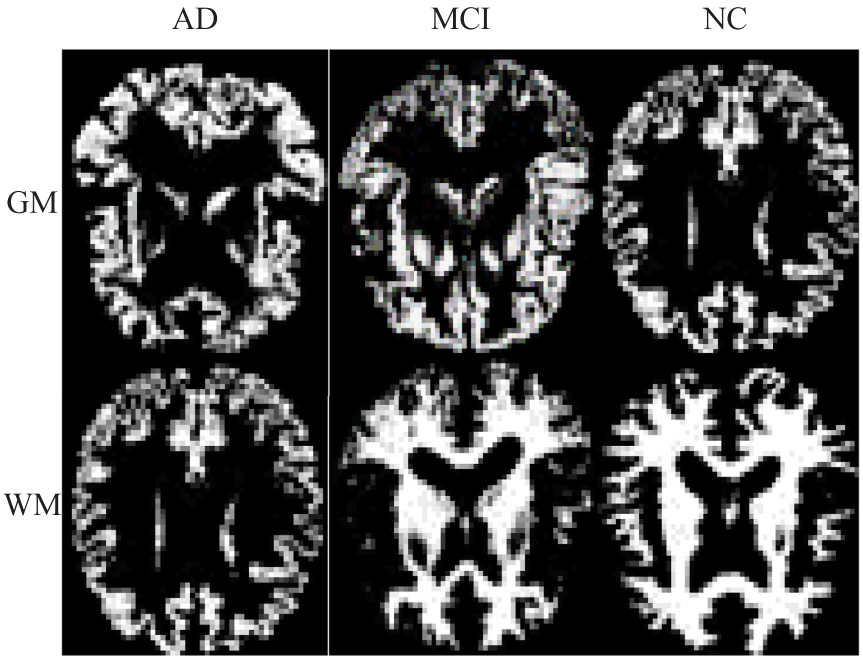

Fig. 2. The left, middle and right columns are AD, MCI, and NC, respectively.

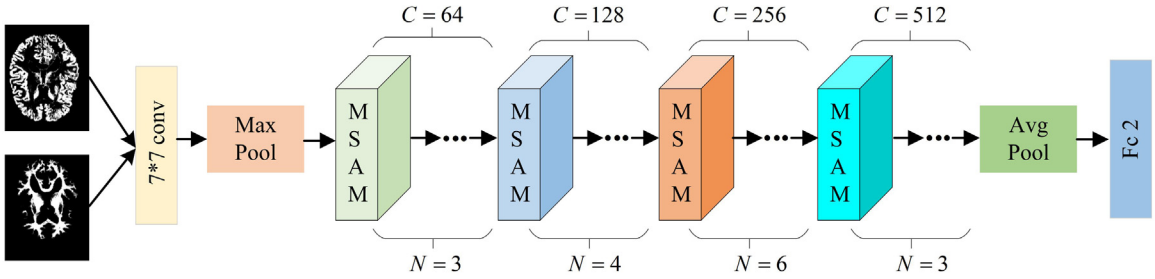

Fig. 3. The structure of MSCNet.

1) Residual neural network

ResNet is a convolutional neural network structure proposed by He et al. [55] in the 2015 ILSVRC competition and won first place in the competition. The emergence of ResNet solved the problem of deepening the number of layers of the network. Under the effect of the chain rule, the gradients are continuously multiplied, and eventually the problems of gradient disappearance, gradient explosion and network degradation. The residual block of ResNet-50, which is the bottleneck block, is shown in Fig. 5. The residual network uses the "shortcut connecting" method to

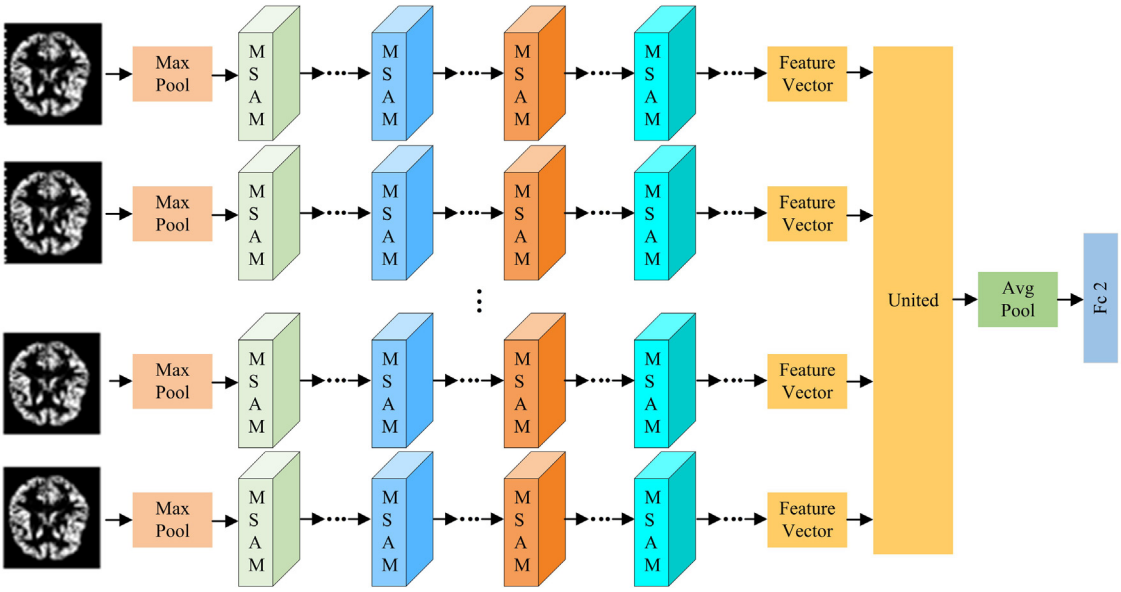

Fig. 4. Multi-slice feature unite network based on MSCNet (MSCNet-U).

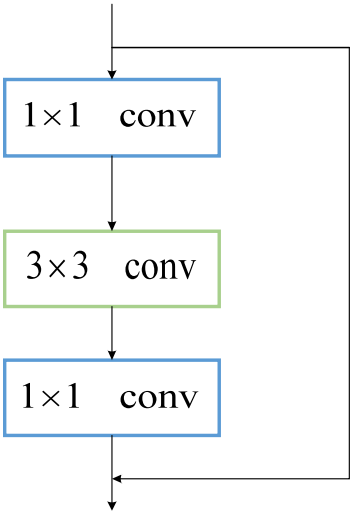

Fig. 5. The bottleneck block of ResNet-50.

send the current output to the next layer of the network, skipping this layer of operations without increasing the parameters. In the process of backpropagation, the gradient of the next layer of network is transferred to the previous network, which solves the problem of the disappearance of the gradient of the deep network. However, ResNet cannot characterize a specific aspect of the data, such as multi-scale and weights between channels. Therefore, it is particularly important to improve ResNet to make it suitable for actual data processing.

2) Multi-Scale Residual Block

The multi-scale feature representation of CNN performs well in computer vision tasks, such as target detection, semantic segmentation, and image classification. In addition, the MRI brain data used for the diagnosis of AD are difficult to identify due to their complex structure and numerous tissues. Moreover, the etiology of AD has not been known up to now, and it is impossible to extract the disease characteristics accurately. Multi-scale feature can extract more comprehensive information of image, and the extracted features are more advanced, which can effectively enhance the performance of the model. Most CNN-based AD models are simple invoking, without corresponding improvements to MRI data, and without considering multi-scale features.

In this paper, we propose a simple but efficient multi-scale structure to extract features of MRI images. Unlike the 3×3 convolution in the bottleneck block $1-3-1$ structure to extract features, we tend to divide the filters in the 3×3 convolution into smaller filter groups, and at the same time, a residual hybrid connection between groups is made to enhance the relationship between features. After 1×1 convolution, we divide the 3×3 convolution kernels with n channels into s groups of parallel smaller convolution kernel groups, denoted by X_i, where $i \in \{1, 2, \ldots, s\}$. Each feature subset X_i has $\frac{n}{s}$ channels. For the case where s is equal to 4, each X_i through a 3×3 dilated convolution [56], and the dilation rate is 1, 2, 3, and 5, denoted by Dconv $3 * 3(X_i)$. Each group X_i has a corresponding dilated convolution, and the output feature map is recorded as S_i. The dilation rate in dilated convolutions is the distance between convolution kernels. If the dilation rate is 1, it is the same as the ordinary convolution. Increasing the dilation rate can increase the receptive field and reduce the amount of calculation. We use dilated convolutions with different dilation rates in the same layer network to obtain receptive fields of different sizes and capture multi-scale context information better. Additionally, to fuse different scale information of the same layer network and comprehensively improve the feature hierarchy, we will mix and superimpose each X_i and denote by Y_i.

$$S_i = Dconv_{3*3}(X_i), \ 1 \le i \le s \tag{1}$$

$$Y_i = \sum_{i=1}^{s} S_i, \ 1 \le i \le s \tag{2}$$

After the concatenate operation on the feature map of the s group, the resulting feature map is sent into the 1×1 convolution and then sent to the Squeeze-and-Excitation (SE) [57] module to calculate the weight relationship between the channels. In Section 4, ablation study was conducted to verify the effectiveness of multi-scale structure in our model.

3) Multi-Scale Residual Block combined with Attention Mechanism (MSAM)

The attention mechanism is a simulation model of the attention of the human brain, and a weighted change strategy is

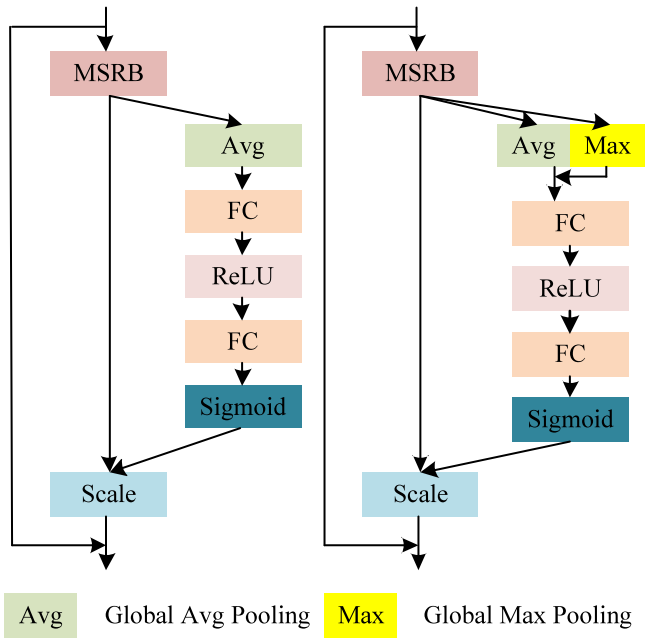

Fig. 6. The original module (left) and our improved module (right).

applied to the data to be processed. It shows good adaptability and gains in computer vision tasks [58,59]. In the SE module, the first operation is the squeeze. It performs feature compression from the spatial dimension, and pools the information of each channel into real numbers through global average pooling. As shown in Fig. 6, we changed the global average pooling value to the sum of the global maximum pooling value to consider the background and texture information of the image. This value has a global receptive field. The global information of the cth feature map of this layer is recorded as z_c.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) + \max_{i,j \in H \times W} u_c(i,j) \tag{3}$$

Among them, the spatial dimension of the input feature map is recorded as $H \times W$, and $\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j)$ is the global average pooling, and it means summing up all pixel values $u_c(i,j)$ in the spatial dimension $H \times W$ and then solving the average value, which can preserve the background information of the image. $\max_{i,j \in H \times W} u_c(i,j)$ is the global maximum pooling, and it means solving the max value of all pixel values $u_c(i,j)$ in the spatial dimension $H \times W$, which can extract texture information from image.

The second operation is the excitation, which generates weights for each feature channels through w, where w is learned to explicitly model the correlation between feature channels.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{4}$$

Finally, the reweight operation takes the weight of the excitation output as the importance of each feature channel, and then multiplies the channel-by-channel weighting to the previous feature to complete the re-calibration of the original feature in the channel dimension. Among them, s_c is the weight value obtained in the excitation operation, and u_c is a two-dimensional matrix representing the information of the cth channel.

$$\widetilde{X}_c = F_{scale}(u_c, s_c) = u_c \cdot s_c \tag{5}$$

To solve the problem of enormous resource consumption for network training caused by a large number of ResNet parameters, we introduce the channel split strategy in the residual network layer, dividing the channel into small convolution kernel groups of the same size for different convolution operations. The SE module can explicitly model the interdependence between channels, and automatically acquire important features between each channel by learning. It follows this idea to promote useful features while suppressing features that are less useful for the task. So, we use the channel attention mechanism to determine the weight ratio between channels. The structure of MSAM is shown in Fig. 7.

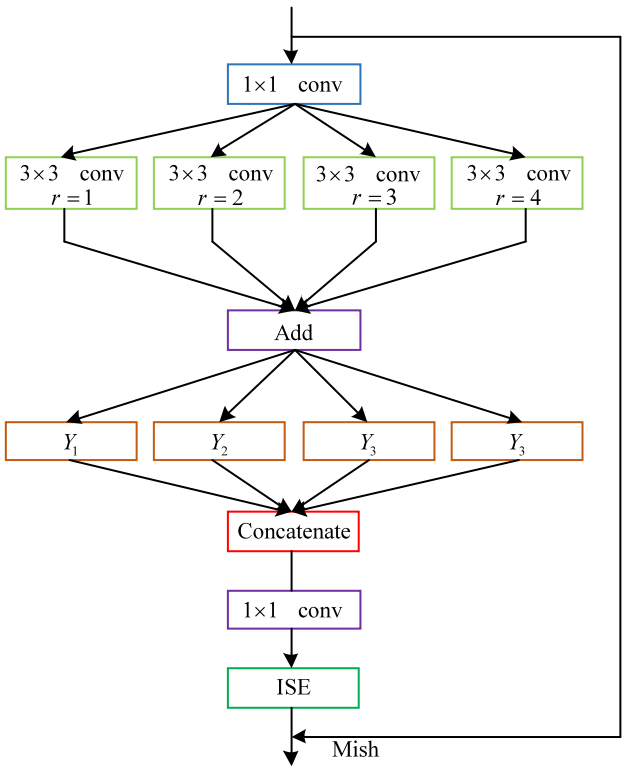

Fig. 7. The structure of MSAM.

3.3. Activation function

The activation function maps the input of the neuron to the output, increasing the nonlinear changeability of the neural network model. Mish is a self-regularized non-monotonic neural activation function [60]. Its formula is defined as:

$$Mish = x * \tanh\left(\ln\left(1 + e^x\right)\right) \tag{6}$$

The curves of ReLU and Mish are shown in Fig. 8. We can find that compared with ReLU at the negative zero boundaries, Mish allows a slightly negative value to obtain a better gradient flow. A smoother activation function can make the information go deeper into the neural network and improve generalization and accuracy, so we replace ReLU with Mish in MSCNet. It can benefit from Mish, although it is limited. We demonstrate this in the ablation study in Section 4.

3.4. Loss function

Cross entropy is a concept in Shannon's information theory, which mainly measures the information difference between two probability distributions. The smaller the value of cross-entropy is, the closer the two probability distributions are. Compared with kullback leibler divergence, cross entropy has simpler operations

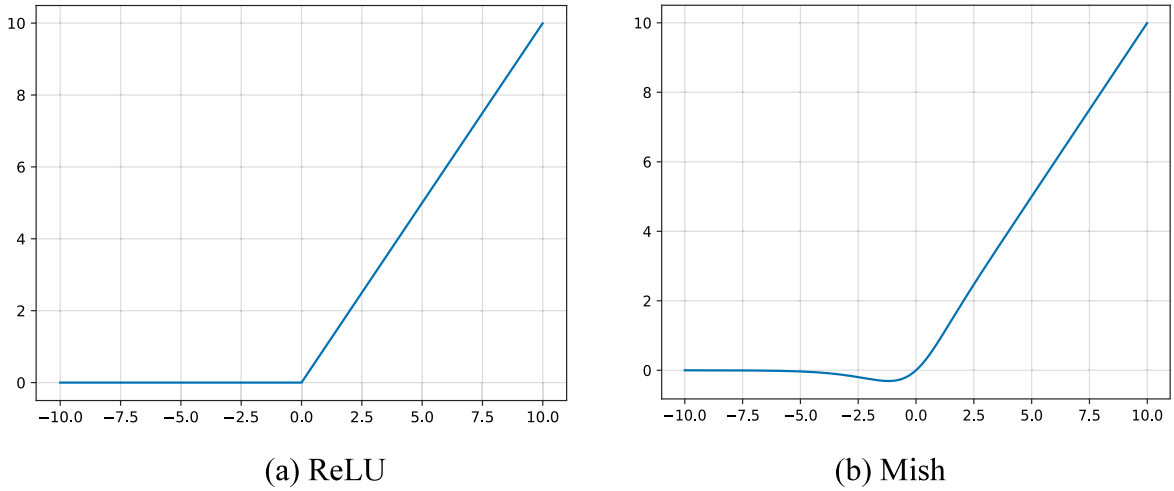

Fig. 8. The curves of ReLU and Mish.

and faster operations. Another benefit is that the use of the sigmoid function can avoid the problem of decreasing the learning rate of the mean square error loss function during gradient descent due to the learning rate can be controlled by the output error. So, the cross-entropy loss function was used in our model. Assuming that p is the probability of the expected output, q is the probability of the actual output, and $H(p, q)$ is the cross-entropy, then:

$$H(p, q) = -\sum_{x} p(x) \log q(x) \tag{7}$$

4. Experiments and discussion

In this section, experimental setting and assessment criteria are introduced briefly, and then we will present some ablation experiments to illustrate the effectiveness of each module in our method, finally our method is compared with the previous approaches to evaluate its performance.

4.1. Experimental setting

The experiments in this paper were conducted in a server environment configured with an NVIDIA DXG-1 based on the Ubuntu 16.04 and a Tesla P100 GPU. SPM12 were used for the medical processing of data. Our method was implemented based on PyTorch and Python3.5. The parameters of our model were set as follows: batch size was set to 16, the optimizer of the training network was Adam, the decay rate was 1e-5, the number of epochs was 60, and the loss function as Cross-Entropy Loss. The learning rate was initialized to 2.5e−4, and reduced by half every 15 epochs.

4.2. Assessment criteria

We evaluated the performance of our model by classification accuracy (ACC), specificity (SPE), sensitivity (SEN), and area under the ROC curve (AUC). Among them, TP, FN, TN, and FP represent true positive, false negative, true negative, and false positive, respectively. Our final evaluation indicators are calculated based on each volume. We took the average of the slice labels obtained from the test as the final single volume label.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$SEN = \frac{TP}{TP + FN} \tag{9}$$

Table 3
ACC, SEN, SPE and AUC of various classification methods for AD, MCI, and NC in GM.

	Models	ACC (%)	SPE (%)	SEN (%)	AUC
AD *vs* NC	ResNet-50	94.80	99.03	96.52	0.96
	ResNet-50+MSRB	96.87	99.56	97.36	0.97
	MSCNet	**97.91**	**99.89**	**98.99**	**0.99**
	MSCNet-U	97.91	99.58	97.29	0.96
AD *vs* MCI	ResNet-50	90.74	92.36	90.47	0.94
	ResNet-50+MSRB	92.60	95.32	92.31	0.96
	MSCNet	**94.44**	**98.52**	**94.91**	**0.97**
	MSCNet-U	93.52	96.82	92.98	0.96
MCI *vs* NC	ResNet-50	87.96	90.25	84.62	0.92
	ResNet-50+MSRB	89.81	91.32	86.18	0.94
	MSCNet	**90.74**	**93.65**	**89.36**	**0.95**
	MSCNet-U	89.81	92.14	88.69	0.94

$$SPN = \frac{TN}{TN + FP} \tag{10}$$

4.3. Ablation study

To illustrate the effectiveness of the multi-scale resnet block, the attention mechanism, the slices on 3D volume labels and Mish in our method, ablation studies were conducted in this part and all methods took the same testing subjects for fair comparison.

1) The effectiveness of MSRB. MSRB represents the network model where the bottleneck module is improved into a multi-scale module. In Tables 3 and 5, we can see from the table that the MSRB has improved performance over the baseline ResNet-50. In the MCI/NC identification experiments of GM and WM, the multi-scale structure improved the classification accuracy of ResNet-50 by 1.85%.

To further verified the effectiveness of MSRB, a new_MSRB is designed by convolution kernels of different sizes. The structure of new_MSRB is shown in Fig. 9, experiments are also conducted. The result is shown in Table 4 and MSCNet is made of MSRB, MnSCNet is made of new_MSRB. From Table 4, we can easily find that MSCNet and MnSCNet have the similar accuracy in the MCI/NC of WM data, but MnSCNet need more time than MSCNet. And compared with MnSCNet, MSCNet has fewer Params and FLOPs.

2) The effectiveness of the attention mechanism. In Tables 3 and 5, MSCNet is an MSRB network model that integrates the

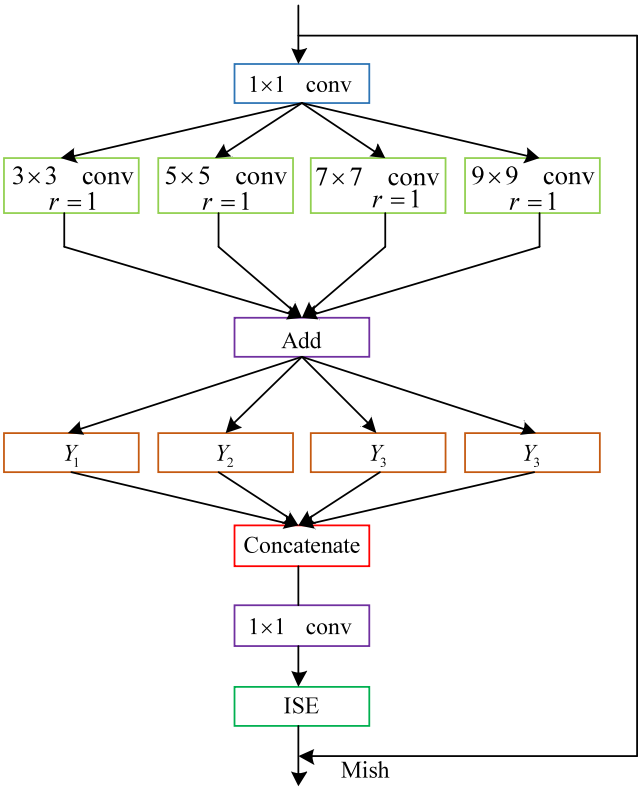

Fig. 9. The structure of new_MSRB.

Table 4
The Params, FLOPs, Acc and running time of MnSCNet and MSCNet in the MCI/NC of WM data.

	MnSCNet	MSCNet
Params	98.5×10^6	20.3×10^6
FLOPs	12.4×10^9	2.8×10^9
ACC (%)	92.61	92.59
Time (ms)	96.3	59.2

Table 5
ACC, SEN, SPE and AUC of various classification methods for AD, MCI, and NC in WM.

	Models	ACC (%)	SPE (%)	SEN (%)	AUC
AD *vs* NC	ResNet-50	95.83	98.63	95.02	0.96
	ResNet-50+MSRB	97.91	99.06	96.66	0.98
	MSCNet	**98.96**	**99.21**	**98.29**	**0.99**
	MSCNet-U	97.91	98.36	96.25	0.98
AD *vs* MCI	ResNet-50	91.67	90.34	90.01	0.95
	ResNet-50+MSRB	93.51	93.62	91.85	0.97
	MSCNet	**95.37**	**97.19**	**93.95**	**0.98**
	MSCNet-U	93.51	94.54	92.19	0.97
MCI *vs* NC	ResNet-50	88.89	88.85	83.82	0.93
	ResNet-50+MSRB	90.74	90.65	86.08	0.95
	MSCNet	**92.59**	**92.45**	**88.96**	**0.96**
	MSCNet-U	91.67	91.86	87.31	0.95

channel attention mechanism. In MCI/NC identification experiments of GM, the introduction of channel attention mechanism improved the final classification accuracy by 0.93%. In the MCI/NC identification experiments of GM, the final classification accuracy was improved by 1.86%. The experimental results proves that the channel attention mechanism we added makes the model benefit from it.

3) The effectiveness of slices on 3D volume labels. As shown in Tables 3 and 5, the results of MSCNet-U on WM and GM are very close to MSCNet. Moreover, the main structures of the two network models are the same. The experimental results show that it is reasonable to select 20 continuous slices on the 3D volume in the experimental preprocess stage. As shown in Fig. 10(a) and Fig. 10(b), ROC curves of ResNet50, MSRB, and MSCNet in the MCI/NC control group on the GM and WM datasets are plotted. It can be seen from Fig. 9 that the ROC curve corresponding to the MSCNet model proposed in this paper is closest to the left and upper boundaries and has the best performance.

Table 6
The MCI/NC of WM data, ACC and model test time comparison when the activation function is ReLU and Mish.

	MSCNet-ReLU	MSCNet-Mish
ACC (%)	91.66	92.60
Time (ms)	48.60	59.20

Table 7
The Params, FLOPs and memory of ResNet-50 and our Method.

	ResNet-50	MSCNet
Params	25.5×10^6	20.3×10^6
FLOPs	4.1×10^9	2.8×10^9
Memory	109.15M	78.03M

Table 8
The accuracy and standard deviation of Our method and ResNet-50 in the MCI/NC of WM and GM.

Data	Model	Acc	Std
GM	MSCNet	**90.74**	**0.0071**
	ResNet-50	87.96	0.2807
WM	MSCNet	**92.59**	**0.0018**
	ResNet-50	88.85	0.3394

4) The effectiveness of Mish. To evaluate the effectiveness of Mish in MSCNet, we performed two sets of experiments in MCI/NC in WM: one set used ReLU, and the other set used Mish. The results are shown in Table 6, The replacement of Mish improves the classification accuracy by 0.94%. However, Mish consumes a large amount of calculation time.

4.4. Result analysis

To evaluate the performance of our method, our method was compared with ResNet-50 and the state-of-the-art methods. We compared the number of params, FLOPs and memory of the MSCNet and ResNet-50. The results are shown in Table 7. From Table 7, we can see that the MSCNet proposed by ResNet-50 as the backbone network has different degrees of reduction in params and FLOPS and memory, which meets the definition of lightweight.

The performance of our method and ResNet-50 was verified by five-fold cross-validation. The experimental results are shown in Table 8, we can find that the accuracy of our method is higher than ResNet-50 and the standard deviation of our method is lower than ResNet-50 in the MCI/NC group of WM and GM. Furthermore, we use the method of randomly extracting data to conduct 20 experiments independently, in which ResNet-50 and our method use the same test set and training set for each experiment. Then we calculated the p values of the two methods by T-test, and its value was 0.03. So, there is statically significant difference between ResNet-50 and our method.

The experimental results of our method and the state-of-art methods are shown in Table 9, we also listed the size of the dataset and the characteristic forms used by these methods. It can be seen that some researchers extracted ROIs such as the hippocampus in the image, and some researchers send the entire

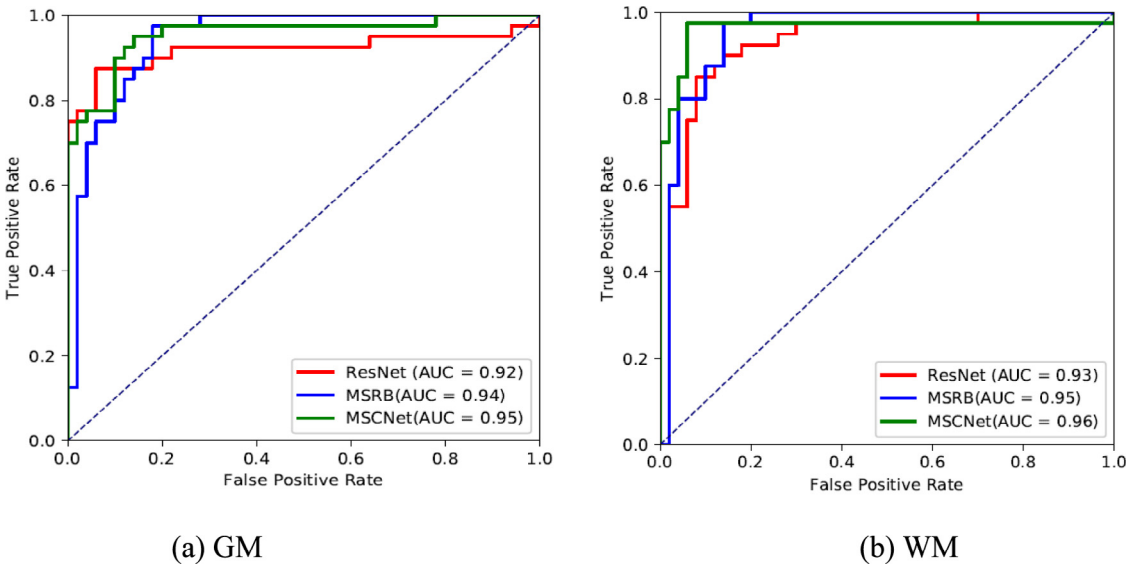

Fig. 10. The ROC curves for the classification between MCI and NC in GM (a) and WM (b).

Table 9
The results of each method (%) in the AD/NC, AD/MCI and MCI/NC.

Articles	Content	Data size	AD *vs* NC	AD *vs* MCI	MCI *vs* NC
Altaf et al. [61]	GM+WM+CSF	92AD+105MCI+90NC	97.80	85.3	**91.8**
Suk et al. [17]	93ROIs	51AD+99MCI+52NC	98.80	83.7	90.7
Shi et al. [62]	Full brain	51AD+99MCI+52NC	97.13	–	87.24
Sarraf et al. [63]	Full brain	28AD+15NC	96.86	–	74.37
Liu et al. [18]	ROI	93AD+204MCI+100NC	93.26	–	–
Fang et al. [30]	Full brain	93AD+204MCI+100NC	**99.27**	92.57	90.35
Our method (GM)	GM	160AD+200MCI+160NC	97.91	**94.44**	90.74
Our method (WM)	WM	160AD+200MCI+160NC	98.96	**95.37**	**92.60**

image to CNNs for training. In our method, we segmented GM and WM from MRI images and conducted experiments separately. The total number of datasets we use is far greater than of other researchers. Moreover, in the AD/NC group, the identification of the AD and NC brain images is relatively high, so the ACC is relatively high. In the AD/MCI group, the difference of brain image structure was further reduced, and the ACC began to decrease, but the ACC of our model on GM and WM was 94.44% and 95.37% respectively, which were both higher than other methods in the table. In the MCI/NC group, the brain image structure is very close, and this stage is of great significance for the clinical diagnosis of AD. The ACC of our method on WM reaches 92.60%, which is improved by 0.8% on the method proposed by Altaf et al. [61]. Based on these results, we believe that the MSCNet model has a better performance than other methods, verifying the effectiveness of MSCNet for AD diagnosis.

In order to demonstrate the generalizability of our method, additional experiments were performed on a larger dataset containing only MRI images, called the multi-atlas label propagation with expectation–maximization (MALPEM) dataset, which available on the website https://biomedia.doc.ic.ac.uk/software/malpem/. The initial MRI data of AD that used in the experiment including 1355 AD, and 1506 NC were all from the MALPEM. The data was also preprocessed by SPM, and the clearest brain structures were taken as the experimental WM and GM datasets. The ratio of training subsets to test subsets in our experiment is 7:3. When training and testing, we strictly divided the training subsets and test subsets according to the slices under the same volume. The results of our method and ResNet-50 are shown in Table 10, we can easily find that ACC, SPE and SEN of our method are higher than those of ResNet-50 in both GM and WM. So, compared with ResNet-50, our method has better generalization ability.

Table 10
ACC, SEN and SPE of ResNet-50 and Our method for AD and NC in WM and GM.

Data	Model	ACC (%)	SPE (%)	SEN (%)
WM	ResNet-50	96.01	98.95	96.62
	Our method	**98.85**	**99.39**	**98.71**
GM	ResNet-50	95.88	98.95	97.05
	Our method	**98.11**	**99.91**	**97.95**

5. Conclusion and future work

We propose a model named MSCNet for AD diagnosis. Compared with the previous simple extraction of ROIs such as the hippocampus or sending raw AD data to CNNs for training, we first segment the MRI data into WM and GM. Then, a new and efficient network architecture with a multi-scale structure and channel attention mechanism is introduced to accurately classify images. Extensive experiments show that our method achieve a good performance in AD diagnosis, and its model size is satisfactory. In addition, experiments prove that WM is more effective in the diagnosis of AD. But the data preprocessing and model training stages in our method are carried out separately. Therefore, we will build a fully-automated system that includes data preprocessing and lightweight models to achieve the effect of assisting medical diagnosis.

CRediT authorship contribution statement

Zhenbing Liu: Writing – review & editing, Supervision, Funding acquisition. **Haoxiang Lu:** Conceptualization, Methodology, Writing – original draft. **Xipeng Pan:** Validation, Writing – review & editing, Supervision, Investigation. **Mingchang Xu:** Visualization, Data curation, Resources, Software. **Rushi Lan:** Writing

– review & editing, Project administration, Funding acquisition. **Xiaonan Luo:** Funding acquisition.

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.

Acknowledgments

The experimental data used in this paper was downloaded from Alzheimer's disease neuroimaging initiative (ADNI) and multi-atlas label propagation with expectation–maximization (MALPEM). This work was supported in part by the National Natural Science Foundation of China under grants (61866009, 62172120, 62002082), the Guangxi Key Research and Development Project (AB21220037), the Guangxi Key Laboratory of Cognitive Radio Image and Graphics (GIIP2001) and the Guanxi Natural Science Foundation under grants (2019GXNSFFA245014, 2020GXNSFBA238014, 2020GXNSFAA297061).

References

[1] B. Lei, S. Yu, X. Zhao, A.F. Frangi, S. Wang, Diagnosis of early alzheimer's disease based on dynamic high order networks, Brain Imag. Behav. 15 (3) (2021) 276–287, 15(3).
[2] J.W. Feng, S.W. Zhang, L.N. Chen, J. Xia, Alzheimer's disease classification using features extracted from nonsubsampled contourlet sub band-based individual networks, Neurocomputing 421 (2021) 260–272.
[3] B.Y. Lei, M.Y. Yang, P. Yang, F. Zhou, W. Hou, W.B. Zou, X. Li, T.F. Wang, X.H. Xiao, S.Q. Wang, Deep and joint learning of longitudinal data for alzheimer's disease prediction, Pattern Recognit. 102 (C) (2020) 107247.
[4] J.L. Peng, X.F. Zhu, Y. Wang, L. An, D.G. Shen, Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis, Pattern Recognit. 88 (2019) 370–382.
[5] A. Mounim, E. Yacoubi, S.G. Salicetti, C. Kahindo, A.S. Rigaud, V.C. Lacroix, From aging to early-stage alzheimer's: uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning, Pattern Recognit. 86 (2019) 112–133.
[6] X. Bi, X. Hu, H. Wu, Y. Wang, Multimodal data analysis of alzheimer's disease based on clustering evolutionary random forest, IEEE J. Biomed. Health. 24 (10) (2020) 2973–2983.
[7] X.Z. Liu, W. Chen, H.T. Hou, X.L. Chen, J.T. Zhang, J. Liu, Z.W. Guo, G.H. Bai, Decreased functional connectivity between the dorsal anterior cingulate cortex and lingual gyrus in alzheimer's disease patients with depression, Behav. Brain. Res. 326 (2017) 132–138.
[8] C. Luo, M. Li, R. Qin, H. Chen, H. Zhao, White matter microstructural damage as an early sign of subjective cognitive decline, Front. Aging Neurosci. 11 (2020) 378–389.
[9] S. Basheera, M. Ram, A novel cnn based alzheimer's disease classification using hybrid enhanced ica segmented gray matter of mri, Comput. Med. Imag. Grap. 81 (2020) 101713.
[10] B.S. Ángela, V.M.T. José, L.C. Susana, M.L. Pedro, F.J. Eduardo, MRI evidence of brain atrophy, white matter damage, and functional adaptive changes in patients with cervical spondylosis and prolonged spinal cord compression, Eur. Radiol. 30 (2020) 357–369.
[11] M.F. Núria, M.F. Marc, F.P. Laia, Z.H. Daniel, B.G.M. Geisa, G.F. Esteban, P.C. Maribel, G.O. Joan, Network change point detection in resting-state functional connectivity dynamics of mild cognitive impairment patients, Int. J. Clin. Hlth. Psyc. 20 (3) (2020) 200–212.
[12] J. Sheng, M. Shao, Q. Zhang, R. Zhou, Y. Xin. Alzheimer's disease, mild.cognitive. impairment, And normal aging distinguished by multi-modal parcellation and machine learning, Sci. Rep. 10 (1) (2020) 5475.
[13] N. Zeng, H. Li, Y. Peng, A new deep belief network-based multi-task learning for diagnosis of alzheimer's disease, Neural Comput. and Appl. 12 (1) (2021) 1–12.
[14] B. Lodewijk, N. Kai, W. Hua, S. Li, H. Heng, Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction, IEEE T. Med. Imaging 39 (6) (2020) 1845–1855.
[15] F. Haghighi, M. Taher, Z. Zhou, M.B. Gotway, J. Liang, Transferable visual words: exploiting the semantics of anatomical patterns for self-supervised learning, IEEE T. Med. Imaging (2021) http://dx.doi.org/10.1109/TMI.2021.3060634, in press.
[16] X. Zhang, L. Wang, Y. Su, Visual place recognition: a survey from deep learning perspective, Pattern Recognit. 113 (2020) 107760.
[17] D.G. Shen, G.R. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (1) (2017) 221–248.
[18] M.H. Liu, D.N. Cheng, K.D. Wang, Y.P. Wang, Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis, Neuroinformatics 16 (3) (2018) 295–308.
[19] X. Wang, W. Wang, Adaptive weights integrated convolutional neural network for alzheimer's disease diagnosis, J. Med. Imag. Health in. 10 (12) (2020) 2893–2900.
[20] X. Zhang, L.X. Han, W.Y. Zhu, L. Sun, D.Q. Zhang, An explainable 3D residual self-attention deep neural network for joint atrophy localization and alzheimer's disease diagnosis using structural MRI, IEEE J. Biomed. Health. (2021) http://dx.doi.org/10.1109/JBHI.2021.3066832, in press.
[21] M. Dua, D. Makhija, P. Manasa, P. Mishra, A cnn–rnn–lstm based amalgamation for alzheimer's disease detection, J. Med. Biol. Eng. 40 (1) (2020) 688–706.
[22] S. Sarraf, G. Tofighi, Deep learning-based pipeline to recognize alzheimer's disease using fmri data, in: FTC, San Francisco, CA, USA, 2016, pp. 816-820.
[23] C.D. Billones, O.J.L.D. Demetria, D.E.D. Hostallero, P.C. Naval, Demnet: a convolutional neural network for the detection of alzheimer's disease and mild cognitive impairment, in: TENCON, Marina Bay Sands, Singapore, 2016, pp. 3724–3727.
[24] A. Farooq, S.M. Anwar, M. Awais, S. Rehman, A deep cnn based multi-class classification of alzheimer's disease using mri, in: IST, Beijing, China, 2017, pp. 1–6.
[25] V. Sathiyamoorthi, A.K. Ilavarasi, K. Murugeswari, S.T. Ahmed, M. Kalipindi, A deep convolutional neural network based computer aided diagnosis system for the prediction of alzheimer's disease in mri images, Measurement 171 (2021) 108838.
[26] H.R. Almadhoun, S.S.A. Naser, Classification of alzheimer's disease using traditional classifiers with pre-trained cnn, J. Acad. Health Med. Res. 5 (4) (2021) 17–21.
[27] Y. Abdulazeem, W.M. Bahgat, M. Badawy, A CNN based framework for classification of alzheimer's disease, Neural Comput. and Appl. 33 (16) (2021) 10415–10428.
[28] S. Bringas, S. Salomón, R. Duque, C. Lage, J.L. Montaa, Alzheimer's disease stage identification using deep learning models, J. Biomed. Inform. 109 (2020) 103514.
[29] J.B. Bae, S. Lee, W. Jung, S. Park, K.W. Kim, Identification of alzheimer's disease using a convolutional neural network model based on t1-weighted magnetic resonance imaging, Sci. Rep. 10 (1) (2020) 22252.
[30] X. Fang, Z. Liu, M. Xu, Ensemble of deep convolutional neural networks based multi-modality images for alzheimer's disease diagnosis, IET Image Process. 14 (2) (2020) 318–326.
[31] A. Khvostikov, K. Aderghal, J.B. Pineau, S.K. Andrey, G. Catheline, 3D CNN-based classification using smri and md-dti images for Alzheimer's disease studies, in: CVRR, Salt Lake City, UT, USA, 2018, pp. 1–14.
[32] A. Kumar, J. Kim, L. David, F. Michael, D.G. Feng, An ensemble of fine-tuned convolutional neural networks for medical image classification, IEEE J. Biomed. Health. 21 (1) (2017) 31–40.
[33] W. Feng, L.N.V. Halm, H. Tang, A. Mecum, X. Guo, Automated mri-based deep learning model for detection of alzheimer's disease process, Int. J. Neural Syst. 30 (6) (2020) 2050032.
[34] Z. Xia, G. Yue, Y. Xu, C. Feng, M. Yang, T. Wang, B. Lei, A novel end-to-end hybrid network for alzheimer's disease detection using 3D cnn and 3D clstm, in: ISBI, Iowa City, IA, USA, 2020, pp. 1–9.
[35] H. Parmar, B. Nutter, R. Long, S. Antani, S. Mitra, Spatiotemporal feature extraction and classification of alzheimer's disease using deep learning 3D-cnn for fmri data, J. Med. Imaging 7 (5) (2020) 056001.
[36] G. Folego, M. Weiler, R.F. Casseb, R. Pires, A. Rocha, Alzheimer's disease detection through whole-brain 3d-cnn mri, Front. Bioeng. Biotechnol. 8 (2020) 534592.
[37] Y. Shmulev, M. Belyaev, Predicting conversion of mild cognitive impairments to alzheimer's disease and exploring impact of neuroimaging, in: MIC, Hobart, TAS, Australia, 2018, pp. 83–91.
[38] U. Senanayake, A. Sowmya, L. Dawes, Deep fusion pipeline for mild cognitive impairment diagnosis, in: ISBI, Washington, DC, USA, 2018, pp. 1394–1997.
[39] H.L. Li, C.M. Pun, F. Xu, L.S. Pan, H.M. Lu, Hybrid feature selection algorithm based on discrete artificial bee colony for parkinson diagnosis, ACM T. Internet Techn. 21 (3) (2021) 1–22.
[40] R.S. Lan, L. Sun, Z.B. Liu, H.M. Lu, X.N. Luo, Madnet: A fast and lightweight network for single-image super resolution, IEEE T. Cybern. 51 (3) (2021) 1443–1453.
[41] S. Christian, I. Sergey, V. Vincent, A.A. Alexander, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI, San Francisco, CA, USA, 2017, pp. 4278–4284.
[42] W. Zou, Z. Zhang, Y. Peng, C. Xiang, L. Zhang, Sc-rpn: a strong correlation learning framework for region proposal, IEEE T. Image Process. 30 (2021) 4084–4098.

[43] J. Leng, Y. Liu, Single-shot augmentation detector for object detection, Neural Comput. Appl. 33 (2) (2021) 3583–3596.
[44] R. Lan, Y. Zhou, Z. Liu, X. Luo, Prior knowledge-based probabilistic collaborative representation for visual recognition, IEEE T. Cybern. 50 (4) (2020) 1498–1508.
[45] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE T. Pattern Anal. 40 (4) (2018) 834–848.
[46] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE T. Pattern Anal. 39 (4) (2017) 640–651.
[47] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: ICML, Lille, France, 2015, pp. 448–456.
[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, Las Vegas, NV, USA, 2016, pp. 2818–2826.
[49] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, P.H.S. Torr, Res2net: A new multi-scale backbone architecture, IEEE T. Pattern Anal. 43 (2) (2021) 652–662.
[50] X. Liu, F. Hou, H. Qin, A. Hao, Multi-view multi-scale cnns for lung nodule type classification from ct images, Pattern Recognit. 77 (2018) 262–275.
[51] F. Zhu, H. Xia, H. Li, S. Song, X. Mou, The combination of multi-scale and residual learning in deep cnn for image denoising, IET Image Process. 14 (10) (2020) 2013–2019.
[52] B. Ibtissam, A. Karim, Multi-scale cnn based on region proposals for efficient breast abnormality recognition, Multimedia Tools Appl. 78 (2019) 12939–12960.
[53] A. John, Computational anatomy with the spm software, Magn. Reson. Imaging 27 (8) (2009) 1163–1174.
[54] R.J. Clifford, A.B. Matt, C.F. Nick, T. Paul, W.W. Michael, The alzheimer's disease neuroimaging initiative (adni): mri methods, J. Magn. Reson. Imaging. 27 (4) (2008) 685–691.
[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, Las Vegas, NV, USA, 2016, pp. 770–778.
[56] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: ICLR, Jiang Su, Zhe Jian, China, 2016, pp. 1–9.
[57] J. Hu, L. Shen, G. Sun, Squeeze and excitation networks, in: CVPR, Salt Lake City, UT, USA, 2018, pp. 7132-7141.
[58] A. He, T. Li, N. Li, K. Wang, H. Fu, Cabnet: category attention block for imbalanced diabetic retinopathy grading, IEEE T. Med. Imaging 40 (1) (2021) 143–153.
[59] X.F. Ding, Y.X. Peng, C.M. Shen, T.Y. Zeng, CAB U-net: An end-to-end category attention boosting algorithm for segmentation, Comput. Med. Imag. Grap. 84 (2020) 101764.
[60] D. Misra, Mish: A self-regularized non-monotonic neural activation function, 2019, CoRR, volume abs/1908.08681.
[61] T. Altaf, S.M. Anwar, N. Gul, M.N. Majeed, M. Majid, Multi-class alzheimer¨s disease classification using image and clinical features, Biomed. Signal Proces. 43 (2018) 64–74.
[62] J. Shi, X. Zheng, Y. Li, Q. Zhang, S.H. Ying, Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease, IEEE J. Biomed. Health. 22 (1) (2017) 173–183.
[63] S. Sarraf, G. Tofighi, Deep learning-based pipeline to recognize alzheimer's disease using fmri data, in: FTC, San Francisco, CA, USA, 2016, pp. 816–820.