Biomedical Signal Processing and Control 77 (2022) 103828

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

3D convolutional neural networks with hybrid attention mechanism for early diagnosis of Alzheimer's disease

Zhiwei Qin [a], Zhao Liu [b,*], Qihao Guo [c], Ping Zhu [a,*]

[a] State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] School of Design, Shanghai Jiao Tong University, Shanghai 200240, China
[c] Department of Gerontology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

ARTICLE INFO

Keywords:
3D convolutional neural networks
Hybrid attention mechanism
Alzheimer's disease
Early diagnosis
Magnetic resonance image

ABSTRACT

As a non-invasive and radiation-free imaging technique, magnetic resonance imaging (MRI) can intuitively display the three-dimensional tissues and structures of human brain, showing the great prospect in the early screening and diagnosis of Alzheimer's disease (AD). MR image processing on the basis of deep learning methods has aroused increasing attention, and the core of this type of method is to construct an efficient model to recognize and extract the key features of the images. In this article, a 3D Residual U-Net model incorporating hybrid attention mechanism (3D HA-ResUNet) is proposed for the auxiliary diagnosis of AD using 3D MR images. The backbone classification model consists of an up-sampling branch network, a down-sampling branch network, and intermediate connection residual blocks. The hybrid attention mechanism exploits the advantages of both channel and spatial attention, and is merged with the skip connection of the backbone classification model. In the binary classification task of AD vs. normal cohort (NC) on the ADNI dataset, the addition of the hybrid attention module helps improve accuracy, sensitivity, precision, F1 score and G-mean by 4.88%, 10.52%, 0.94%, 6.17% and 5.60%, respectively. Furthermore, the proposed method demonstrates superior generalization ability compared with other state-of-the-art methods. The 3D HA-ResUNet was further tested in the mild cognitive impairment (MCI) subtype classification task on the local dataset and achieved 100% of accuracy. In addition, an attribution-based visual interpretability method is employed to reveal the regions and features that the proposed model focuses on for classification. The visual interpretations combined with domain knowledge are capable of providing a valuable reference for physicians' clinical decision-making.

1. Introduction

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disease, and the leading source of dementia worldwide. Clinically, AD is characterized by comprehensive dementia manifestations such as memory impairment, language disorder, and executive dysfunction [1,2]. The pathogenesis of AD is still under study and no consistent conclusion has been reached yet, and no effective drugs and treatments have been found to cure AD diagnosed patients. Thus, early screening, assessment and intervention are critical to enhancing patients' quality of life and prognoses. With the rapid development of medical imaging technology and equipment, neuroimaging has become one of the most intuitive and reliable methods for the detection of AD. Among them, magnetic resonance imaging (MRI) is a non-invasive and radiation-free imaging technique that can display the three-dimensional images of brain tissues, and intuitively provide the information of tissue lesions. MRI has become an effective approach for clinical detection of biomarkers and recognition of brain atrophy patterns in the progression of AD [3]. In order to optimize the traditional manual film-reading process, avoiding a lot of time consumption and over-reliance on the expert's personal knowledge and experience, MRI scans processing methods based on general machine learning and deep learning have been heavily studied in recent years.

As one of the most prevalent research directions in machine learning, deep learning has the advantages of learning the inherent laws and representation levels of data, which makes it closer to the goal of artificial intelligence. At the same time, deep learning models can be trained end-to-end, which enables such techniques to be quickly deployed to specific applications in various fields. Given the difficulty of early AD diagnosis using 3D MR images, incorporating attention mechanism into

* Corresponding authors.
E-mail addresses: hotlz@sjtu.edu.cn (Z. Liu), pzhu@sjtu.edu.cn (P. Zhu).

https://doi.org/10.1016/j.bspc.2022.103828
Received 29 January 2022; Received in revised form 30 April 2022; Accepted 16 May 2022
Available online 20 May 2022

deep learning technique can be an effective approach. Introducing attention mechanism into deep learning model is not only a very interesting idea, but also a work that needs further exploration and improvement. In this work, we proposed a 3D pyramidal hierarchical deep learning model incorporating hybrid attention mechanism (3D HA-ResUNet) for the early diagnosis of AD based on 3D MR images.

The rest of this article is organized as follows: Section 2 summarizes the related work on AD diagnosis based on MR images, as well as the motivation and contributions of this work. Section 3 elucidates the attention mechanism and our proposed 3D HA-ResUNet model. The following Section 4 introduces the datasets, experimental settings and the classification results of different methods. Section 5 delivers a visual interpretation of the proposed model and a further comparison with some state-of-the-art studies. Finally, Section 6 provides a conclusion of this article along with an outlook for future work.

2. Related work

In this section, general machine learning methods for MR image feature extraction and classification are first reviewed, followed by a summary of deep learning architectures. Furthermore, the setbacks of existing approaches, together with the motivation and contributions of the proposed 3D HA-ResUNet are presented.

2.1. Machine learning methods

Machine learning provides automatic classification techniques based on data science, which can be easily combined with the image feature processing methods in traditional medicine to form a complete analysis process. In the classification studies of AD stages using MR images, researchers have used methods such as voxel-based morphometry to extract the image features (gray matter volume, white matter volume, cerebral cortical thickness, etc.), and input these numerical features into the classifier based on machine learning methods for further classification [4]. Klöppel et al. [3] extracted the voxel value of gray matter segments from T1-weighted MR scans and selected support vector machine (SVM) as the classifier, which was robust to MR images produced by different scanning devices. Nayak et al. [5] used two-dimension discrete wavelet transformation (2D DWT) to extract MR image features and probabilistic principal component analysis (PPCA) to select the significant features. The random forests approach was implemented to distinguish normal and diseased brains based on the feature set.

Cerebral cortical atrophy, hippocampal atrophy and ventricular enlargement are the main manifestations of early AD [6], so the whole brain can be divided into regions of interest (ROI) according to medical prior knowledge, and the image features of the ROIs can be extracted for analysis. Zhang et al. [7] extracted multi-biomarker features (e.g., volume features) from the 93 ROIs of multimodal brain neuroimages and classified the features using the multi-kernel SVM classifier. Uysal and Ozturk [8] proposed that the volume atrophy of hippocampus is the most important indicator of AD, therefore they built a machine learning model based on the volume features of left and right hippocampus as well as the age and gender information to predict the diagnostic type of subjects. Although machine learning methods can perform automatic classification of numerical features, the operations required from image data to numerical data, such as feature extraction, feature selection and feature fusion, still need additional manual design of algorithms in most cases.

2.2. Deep learning methods

Deep learning techniques have become mainstream research tools in computer vision due to their advantages in high-dimensional data representation [9–12], and have achieved satisfactory results in medical image analysis [13,14]. Compared to traditional image processing methods, deep learning has unique advantages: (1) as a data-driven feature learning algorithm, deep learning model can perform automatic feature extraction through end-to-end training, greatly reducing manual workload; (2) the deep structure of deep neural networks can capture the interactions between abstract features; (3) in addition to feature extraction, feature selection and classification can also be fulfilled in network training. As a representative technique of deep learning, convolutional neural networks (CNN) have been widely studied.

2.2.1. CNN architectures and their developments

The architecture of CNN greatly affects the feature extraction and classification performance of the model. Among various CNN architectures, ResNet [12] provides a solution to the degradation problem of deep CNNs. Korolev et al. [15] proposed two 3D CNN architectures for brain MR image classification, i.e., plain CNN and residual CNN, which revealed the potential of deep learning models for end-to-end analysis of complex MRI data. Karasawa et al. [16] built a 3D CNN model with 36 convolutional layers based on ResNet architecture. Liu et al. [17] established a multi-task CNN model for segmentation and feature extraction of brain MRI hippocampus, then learned the key features of the segmented hippocampus by building 3D DenseNet [18], and classified the AD status in combination with the two groups of features. This multi-model deep learning framework achieved better results than the single-model approach.

U-Net [19] is another prevalent CNN architecture, which enables feature extraction and fusion at different scales through an encoder with down-sampling and a decoder with up-sampling. U-Net was originally proposed for segmentation of medical images and has been customized with different characteristics for different medical image processing tasks [20,21]. Fan et al. [22] proposed a U-Net style model for AD diagnosis, applied the vanilla U-Net to the classification of 3D MR images, and achieved good results in both binary and multi-class classification. U-Net can be used as a classification method and combined with other image enhancement and feature extraction techniques. Ragupathy and Karunakaran [23] used fuzzy logic for brain MR image enhancement and dual tree-complex wavelet transform for feature extraction, the features were input into U-Net CNN for classification. Maqsood et al. [24] also used fuzzy logic for edge detection and U-Net model for classification. From a broader perspective, U-Net architecture can be easily integrated with other meticulous designs to make the model more comprehensive. For example, different activation function in the network [25], Bayesian approach for handling probabilistic graphical model [26]. However, the application of these techniques in MR image-based AD diagnosis needs further research.

2.2.2. Introducing attention mechanism into CNNs

CNNs are designed to simulate the image reception and processing of the brain. Inspired by this biology, adding an attention mechanism similar to human visual system to CNN facilitates the identification of valuable information in the region of interest and improving the accuracy and efficiency of information processing. Therefore, research has been accruing to incorporate attention mechanisms in classification models to improve the models' recognition capability of biomarker-related features in MR images.

Jin et al. [27] proposed a 3D attention-based ResNet for AD diagnosis, which embedded the lightweight attention module into the original ResNet architecture. Apart from improving the classification performance of the model, their work also made a beneficial exploration for visualizing the decision-making process of the deep learning model. Also based on ResNet architecture, Yu et al. [28] proposed a 3D spatial attention module which fused multi-scale spatial information of MRI brain scans from multiple branches. Zhang et al. [29] constructed a task-driven hierarchical attention network for AD classification. This framework consisted of a patch-based information sub-network generating a disease-related information map and an attention-based hierarchical sub-network extracting discriminative features with the help of

visual attention module and semantic attention module.

In terms of U-Net architecture, R. Karthik et al. [30] introduced attention mechanism into 2D fully convolutional network (FCN) for segmentation of brain MR images. The attention block helped the model concentrate on salient features of the lesion region, resulting in a good improvement in segmentation. Hashemi et al. [31] made modifications in the loss function of the 2D CNN models, and compared the U-Net and attention U-Net in segmentation of multiple sclerosis lesion. The experiments revealed that the attention gate in U-Net was helpful for the model to find the edges of the lesion area. Although U-Net approach combined with attention mechanism has achieved many successful cases in medical image segmentation, its direct application in the early diagnosis of AD based on MR images will still be a new attempt.

2.3. Research gaps and motivation

Our literature review triggered the conclusion that researchers have done extensive work in AD diagnosis based on MR image classification, and the related work has its distinctive characteristics. The current research work and areas for improvement are summarized as follows:

(1) General machine learning methods are able to deliver automatic classification of numerical features, but the operations of feature extraction, selection and fusion still require additional manual design of algorithms. Besides, machine learning methods have difficulty handling 3D data such as MR images.
(2) Deep learning methods provide the option of end-to-end learning, and transform the framework from 2D to 3D for handling 3D MR images. Nevertheless, a more comprehensive CNN architecture with pertinency designs needs to be further applied to the early diagnosis of AD. In addition, due to the complex characteristics of the brain structure and the difficulty of detecting early AD biomarkers, adding attention mechanism to the 3D CNN framework can be a good choice. However, many of the existing deep learning approaches that incorporate attention mechanisms take advantage of spatial attention, which focuses on the spatial relationships of features. Other types of attention, such as channel attention, have been less frequently introduced into relevant studies. Making full use of different attentions can help improve the performance of the model on brain MR image classification.
(3) Although deep learning methods and attention mechanism have advantages in processing 3D MR images, their models are black boxes and the lack of interpretability limits their further application in medical diagnosis.

Many deep learning networks, such as ResNet [12] mentioned before, compress the feature maps to a small size in their deep layers, which is not suitable for visualizing the activation features, that is, not easily integrated with visual interpretability methods. To solve this issue, we designed a pyramidal hierarchical network with both down-sampling and up-sampling processes according to the U-Net architecture [19], which can expand the feature map to the size of the original input image in deep layers of the network, while ensuring excellent classification performance. Considering the convolution of CNN inherently models only the spatial information of the image but does not model the information between channels. The channel attention characterizes the relationship between channels, which can enhance the feature detector useful for the current task and suppress the feature detector of little use according to the importance of the channels. Therefore, the combination of channel attention and spatial attention will be more conducive to the identification and localization of decision features. This article introduced efficient channel attention and spatial attention to construct a lightweight hybrid attention module that can be fused with the aforementioned classification model to further improve its performance. Moreover, an attribution-based visual interpretability method is employed to explain the inference process of the proposed model and make the black box more transparent. In general, the proposed 3D Residual U-Net model incorporating with hybrid attention mechanism (3D HA-ResUNet) exhibits superior capabilities in brain MR image classification and a greater potential for model interpretation.

2.4. Research contributions

The main contributions of this study are as follows:

(1) A 3D pyramidal hierarchical convolutional neural network consisting of a down-sampling branch, an up-sampling branch and intermediate connection residual blocks is proposed to process 3D brain MR images. The proposed model is capable of generating feature maps of the same size as the input image in its deep layers and achieving excellent classification performance.
(2) A hybrid attention mechanism that fuses efficient channel attention and spatial attention is proposed in this article. The hybrid attention combines the advantages of both types of attention to obtain better feature recognition and location, which can be integrated with the basic classification model to further improve its performance.
(3) A visual interpretability method based on attribution and semantic explanations is applied to the inference process of the proposed model, making the deep learning model more transparent and easier to be popularized in clinic.

3. Methodology

The framework of the proposed 3D HA-ResUNet is illustrated in Fig. 1, which includes down-sampling branch network, up-sampling branch network, intermediate connection residual blocks and hybrid attention modules. The proposed model receives the pre-processed 3D MR image as input and outputs the diagnostic type of the sample. Furthermore, an attribution-based visual interpretability method is used to reveal the regions and features that the proposed model focuses on for classification.

3.1. Attention mechanism

The human visual system is considerably complex. As an important part of this system, the human eye, although considered as a refined sense, has mostly low resolution and unclear imaging except for a small area called the fovea located in the inner part of the back of the eyeball. Thus, when the human brain receives external visual information, it controls several eye movements, called saccades, to glimpse the most conspicuous or task-relevant parts of the scene. The attention mechanism in deep learning models is derived from this dynamic. Therefore, adding similar visual attention mechanisms to CNN is conducive to highlighting important information in the regions of interest and improving the accuracy and efficiency of information processing.

Attention mechanism in deep learning has developed into many different types. According to the differentiability of attention, it can be classified into hard attention and soft attention [32]. Hard attention is a non-differentiable attention and the training process is usually accomplished through reinforcement learning. Soft attention can be differentiated, and the weights of attention can be obtained by neural networks counting the gradients and learning from backward propagation. According to the attention domain, it can be divided into channel domain attention, spatial domain attention, layer domain attention, hybrid domain attention and temporal domain attention. The channel domain generally refers to the different channels of the neural networks and is concerned with the distribution of attention weights among different feature maps. The spatial domain emphasizes the distribution of attention weights within each feature map. The layer domain is commonly used for attention interactions between hierarchical feature maps of pyramidal network structure. The hybrid domain is a mixture of

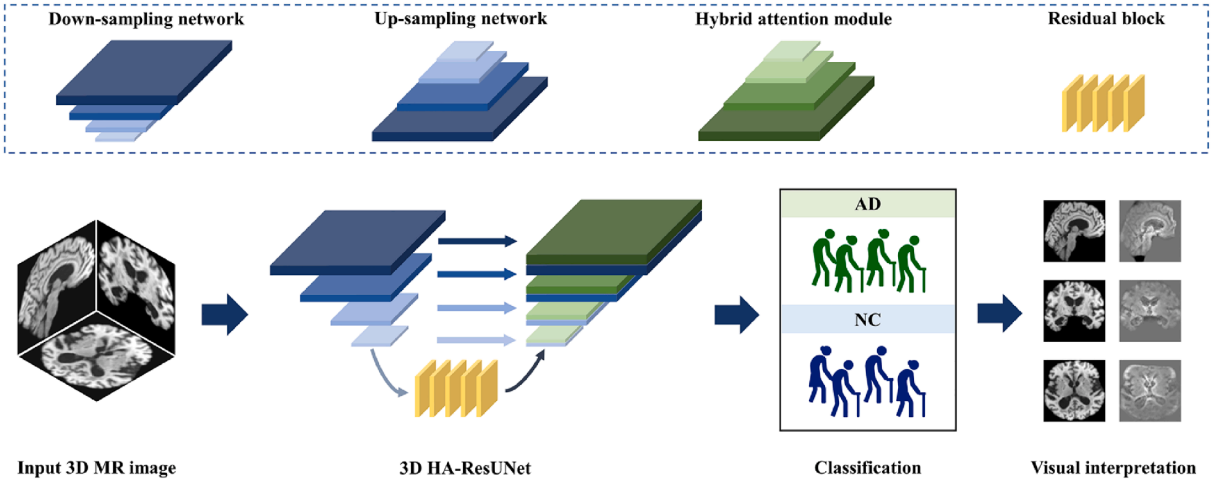


Fig. 1. Pipeline of the 3D HA-ResUNet.

different attention domains, and usually combines channel domain and spatial domain. The temporal domain attention can be considered as a special implementation of hard attention, which generally accompanies recurrent neural network (RNN) models. This study focuses on the channel and spatial domain attention mechanisms.

3.1.1. Channel domain attention

The channel domain attention mechanism is the weight assigned to the feature map of each channel in the network layer, which can be expressed by Equation 1:

$$F_c^{out} = w_c F_c^{in}, \qquad c = 1, 2, \cdots, C \tag{1}$$

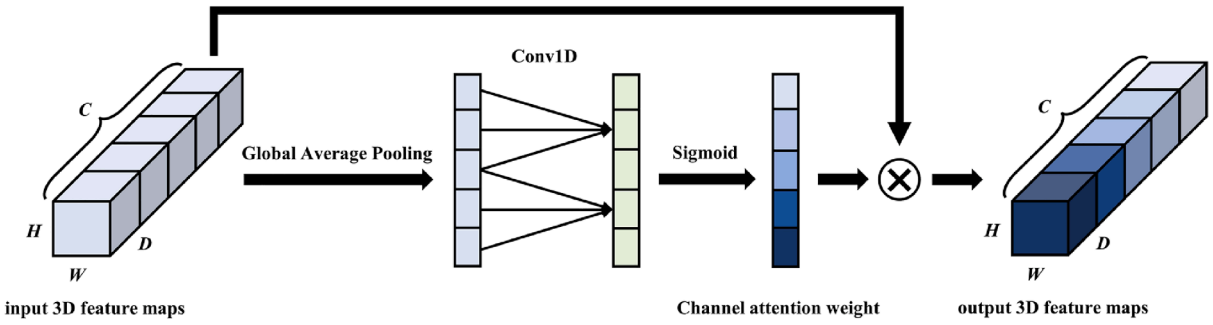
where c denotes the c-th channel, w_c is the attention weight of the c-th channel feature map, F_c^{in} and F_c^{out} are the c-th channel input and output feature maps respectively.

3.1.2. Spatial domain attention

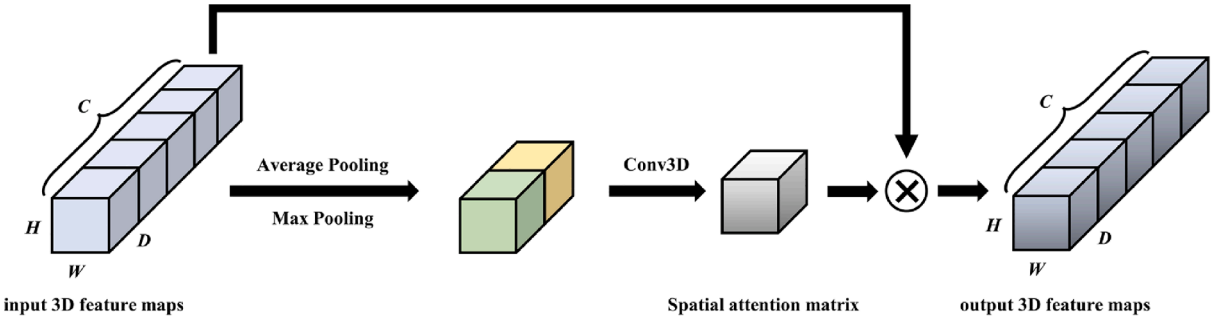
Spatial domain attention mechanism describes how much attention the feature maps receive in different regions of space, and all feature maps share the same attention weight matrix. It can be expressed as:

$$F_c^{out} = W^{\circ} F_c^{in}, \qquad c = 1, 2, \cdots, C \tag{2}$$

where c denotes the c-th channel, F_c^{in} and F_c^{out} are input and output



(a) 3D channel attention block



(b) 3D spatial attention block

Fig. 2. Diagram of each attention block. (a) 3D channel attention block generates a 1D channel attention weight which has C elements. (b) 3D spatial attention block generates a 3D spatial attention matrix with a size of $H \times W \times D$.

feature maps of the c-th channel, W is the spatial attention weight matrix of the feature map whose dimension is consistent with F_c^{in}. And $\circ$ represents the Hadamard product of two matrices of the same order.

3.1.3. The proposed hybrid domain attention

The hybrid domain attention mechanism in this study combines channel domain attention and spatial domain attention. Since each channel of the network is regarded as a feature detector, channel attention facilitates solving the issue of "what" features to focus on for the input image. While spatial attention focuses on distinguishing the spatial location of features ("where"), which is a supplement to channel attention [33]. Therefore, combining the attention of the two domains is expected to achieve a better feature recognition and localization effect. According to the data characteristics of this study, we proposed a hybrid attention module for 3D brain MR images, which contains a 3D channel attention block and a 3D spatial attention block to generate a 1D channel attention vector and a 3D spatial attention map, respectively. Fig. 2 illustrates the calculation diagram of each attention sub-block.

The 3D channel attention block was constructed based on the classical architecture of Squeeze-and-Excitation Networks (SENet) [34], which consists of squeeze, excitation and scale operations. The input 3D feature maps were squeezed into a $1 \times 1 \times C$ feature vector through global average pooling (GAP). Different from the original SENet, which utilizes two fully-connected layers to perform transformations of the feature vector, we employed 1D convolutional layer and sigmoid function to generate channel attention weight according to the design of ECA-Net [35]. Finally, the original input feature maps were multiplied with the channel attention vector to obtain new feature maps, which were assigned with per-channel weights. The calculation process of 3D channel attention can be expressed by the following equations:

$$W_{channel} = \sigma(Conv1D(GAP(F^{in}))) \tag{3}$$

$$F^{out} = W_{channel} F^{in} \tag{4}$$

where $W_{channel}$ is $1 \times 1 \times C$ channel attention weight. GAP denotes global average pooling, $Conv1D$ is 1D convolutional layer and σ represents sigmoid activation function.

The 3D spatial attention block worked to generate a 3D attention matrix to emphasize different informative regions in a 3D feature map. The input 3D feature maps were transformed into two new feature maps through channel-wise average pooling and max pooling. The two feature maps were concatenated along the channel axis. A 3D convolutional layer was then applied on the concatenated feature map to generate 3D spatial attention matrix which contained the spatial attention weights assigned to the 3D feature maps. And the output feature map was the Hadamard product of each original input feature map and this spatial attention matrix. The 3D spatial attention block can be expressed as:

$$W_{spatial} = Conv3D([AvgPool(F^{in}); MaxPool(F^{in})]) \tag{5}$$

$$F_c^{out} = W_{spatial} \circ F_c^{in}, \qquad c = 1, 2, \cdots, C \tag{6}$$

where $W_{spatial}$ is 3D spatial attention matrix. $AvgPool$ and $MaxPool$ refer to channel-wise average pooling and max pooling, and [;] represents concatenation. $Conv3D$ is 3D convolutional layer. $\circ$ represents the Hadamard product of two matrices of the same order.

The channel attention block and spatial attention block were combined in a sequential manner. Thus, the hybrid attention mechanism with channel-first order can be achieved by Eq. (7):

$$F_c^{out} = W_{spatial} \circ (W_{channel} F^{in})_c, \qquad c = 1, 2, \cdots, C \tag{7}$$

3.2. Architecture of 3D HA-ResUNet

Two main aspects were considered in the design of the classification network:

(1) In order to improve the performance of the model on MR image classification. The low-level features after multiple down-samplings can provide the semantic information of the recognition target in the whole image, which is helpful for the category judgment of the target. The high-level features directly passed from the encoder to the same scale of the decoder through the skip connection can provide more refined features for classification. The brain MR images used in our study have relatively fixed structure and clear semantics, and low-level features can provide this information. However, the boundaries of brain tissues in the images are blurred, so more high-resolution information is required for identifying.

(2) The up-sampling network of the decoder can upscale the feature map to the size of the original input image, providing a basis for subsequent visual interpretation. For some classical deep neural networks, such as ResNet [12], their deep layers, especially the last convolutional layer, usually compress the feature map to a small size. When visualizing the feature maps output from these deep networks, many details will be lost if they are scaled up to the size of the original input image. Therefore, we designed a pyramidal hierarchical network according to the "down sampling-up sampling" quasi-symmetric architecture of U-Net [19], so that the feature map output from the deep layers can also maintain the size of the shallow layers.

Fig. 3 demonstrates the architecture of the proposed model, which contains a U-Net backbone, residual blocks and hybrid attention modules. Relevant network structures were extended to 3D because of the need to process 3D data.

3.2.1. 3D U-Net-like networks

A U-Net-like network structure and pyramid hierarchical structure are the major structure characteristics of the proposed model. The left half of the networks works like a decoder, receiving input images and performing feature extraction at different scales by down-sampling. The left half includes an initial convolutional block and four down-sampling blocks. The initial convolutional block receives 3D MR image input and contains a 3D convolutional layer with 64 $7 \times 7 \times 7$ filters, batch normalization (BN) layer [36] and ReLU activation function. The down-sampling block includes two 3D convolutional layers with filter size of $3 \times 3 \times 3$, and a 3D average pooling layer which reduces the size of the feature map to one-half of the original size. The $64 \times 64 \times 64$ input image becomes $4 \times 4 \times 4$ feature maps after 4 down-sampling blocks. While the right half of the networks uses up-sampling to restore the feature maps to the size of the original input image. The right half consists of four corresponding up-sampling blocks and a final convolutional block. The up-sampling block contains a 3D up-sampling layer and a 3D convolutional layer with filter size of $3 \times 3 \times 3$. Each up-sampling block is followed by the proposed hybrid attention module and a residual block. The final convolutional block contains a 3D convolutional layer with filter size of $7 \times 7 \times 7$ and a 3D global average pooling which compresses the 3D feature maps into a 1D feature vector. The predicted categories of the networks are finally output by the last fully-connected layer.

3.2.2. Residual block

Residual block acts as an intermediate connection bridge between the left and right halves of the networks. In order to further process the features output from down-sampling blocks, and provide more decision-informed features for up-sampling blocks, we implemented 7 residual blocks, each containing a 3D convolutional layer with filter size of $3 \times 3 \times 3$, a BN layer and ReLU activation. The features are fused by the residual connection to deal with the vanishing gradient problem in deep neural networks [12].

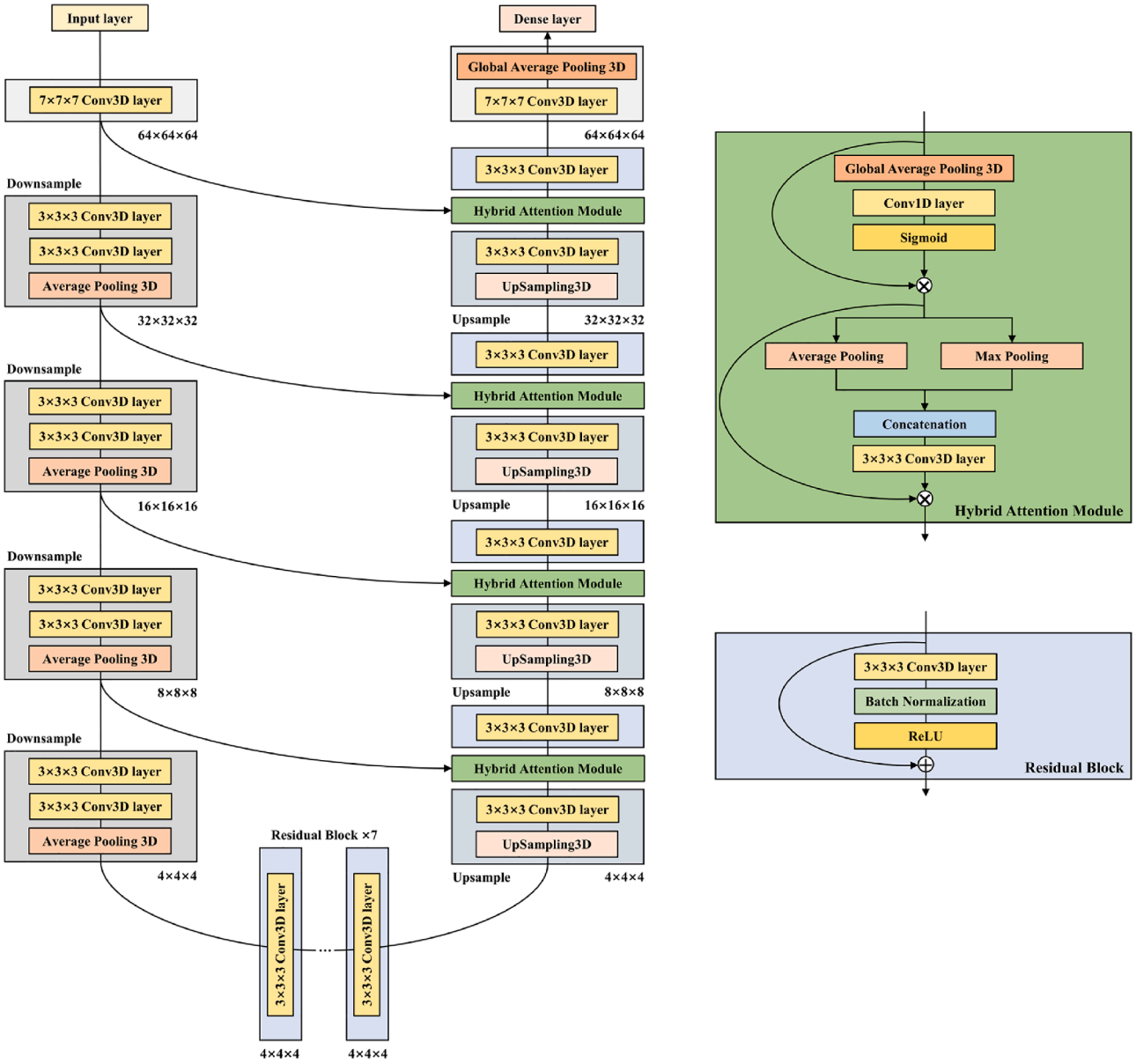


Fig. 3. Architecture of the proposed 3D Residual U-Net with hybrid attention mechanism. 3D U-net-like networks consist of left branch, right branch and the residual blocks which acts as the connecting bridge. The left branch is a down-sampling process, while the right branch is the corresponding up-sampling process, the middle connection works on feature processing. The feature maps of different levels are fused after hybrid attention weighting. The specific structures of hybrid attention module and residual block are demonstrated in the right half of the figure.

3.2.3. *Hybrid attention module*

The proposed hybrid attention module can be well incorporated with the skip connection in U-Net. We applied the hybrid attention to both low-level and high-level features and fused them by means of the skip connection. In other words, the attention weights that are conducive to the work of the shallow network and deep network can be obtained separately. And concatenating these two types of features not only integrates low-level and high-level features, but also allocates attention in line with the network characteristics, so that the model can focus on more valuable information for different levels of features. Section 3.1 describes the hybrid attention module, which includes a 3D global average pooling, a 1D convolutional layer and sigmoid activation function for the channel attention part, as well as the parallel average pooling and max pooling, concatenation operation and a 3D convolutional layer for the spatial attention part.

3.3. *Visual interpretability method*

Although deep learning models have played an outstanding role in many tasks, the lack of interpretability limits their further application in some scenarios such as medical diagnosis. Interpretability emphasizes the capability to furnish human with logical rules or at least some critical elements of the rules related to the domain knowledge [37,38]. In this section, we utilized attribution-based interpretability method to explain the decision-making process of the proposed 3D Residual U-Net with hybrid attention mechanism (3D HA-ResUNet), aiming to make the black box more transparent. Here, we employed and extended the Gradient-weighted Class Activation Mapping combined with guided backpropagation (Guided Grad-CAM) [39] approach to fit our 3D CNN model.

The gradient information flowing to the last convolutional layer in CNN model is able to reveal the importance of each filter for category recognition. For category n, the weight of the lth filter in the last

convolutional layer is calculated as follows:

$$\alpha_l^n = \frac{1}{Z}\sum_i\sum_j\sum_k \frac{\partial y^n}{\partial A_{ijk}^l} \tag{8}$$

where y^n denotes the gradient of the score for class n, A^l is feature map activation and $\frac{1}{Z}\sum_i\sum_j\sum_k$ represents 3D global average pooling.

Subsequently, ReLU activation is performed on the weighted feature map to eliminate the influence of negative values to obtain the classification location map of class n:

$$L_{Grad-CAM}^n = ReLU\left(\sum_l \alpha_l^n A^l\right) \tag{9}$$

Finally, the $L_{Grad-CAM}^n$ is element-wise multiplied with guided backpropagation to acquire Guided Grad-CAM visualizations. The Guided Grad-CAM is not only class-discriminative, but also fine-grain. The application of the Guided Grad-CAM to the proposed 3D HA-ResUNet model will provide visual explanations for the MRI-based classification process.

4. Experiments and results

This section introduces the datasets used in this research and experimental settings. We also provide a description of the comparison methods and the classification results on ADNI dataset. The proposed method was further compared with some state-of-the-art studies reported in the literature. The local MCI dataset was eventually applied to validate the ability of the proposed method for clinical application. All the experiments were implemented on 3.50 GHz CPU with 192 GB RAM, and NVIDIA Quadro P4000 8 GB GPU.

4.1. Data and pre-processing

The data used in this study was collected from two sources: the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/) and the local dataset of subjects with mild cognitive impairment (MCI). As a longitudinal multicenter study, ADNI aims to detect and track AD at the earliest possible stage (pre-dementia) by means of clinical, imaging, genetic and biochemical biomarkers. In the ADNI dataset, T1-weighted structural MRI (sMRI) scans of AD and normal cohort (NC) were acquired. Since some subjects have longitudinal data, i.e., images from different time points, we only selected the first data of each subject to prevent possible similarity between these images. The ADNI dataset includes 98 AD samples and 114 NC samples in total.

In terms of the local dataset, MCI subjects were chosen as samples for their high probability of conversion to AD, which is critical for studying the early diagnosis of AD. The MCI subjects were further divided into three subtypes according to the results of neuropsychological assessment. The neuropsychological assessment mainly involves three cognitive domains: episodic memory domain, language domain and graphomotor speed/executive function domain [40–42]. The neuropsychological test methods used in each cognitive domain are specified in Table 1 and the detailed classification criteria for MCI subtypes are shown in Fig. 4 [40–42]. Subjects were diagnosed as aMCI for their impaired scores on both two measures within memory domain. The impaired score is defined as >1 standard deviation (SD) below the age-corrected normative mean. sMCI indicates the subjects suffered impaired scores on both two measures within language domain. The impaired score is defined as >1 SD below the education-corrected normative mean. And oMCI refers to impaired scores on both two measures within speed/executive function domain or one impaired score in each of the three cognitive domains. The impaired score of speed/executive function domain is defined as >1 SD below the age and education co-corrected normative mean. Based on the above criteria, a total of 43 aMCI samples, 46 sMCI samples and 5 oMCI samples were collected in the local MCI dataset. Due to the small sample size of oMCI, this study mainly focused on the classification of aMCI and sMCI.

Table 1
Cognitive function measures for MCI subtype classification.

Cognitive domain	Neuropsychological test method
Episodic memory domain	Rey Auditory Verbal Learning Test delayed recall
	Rey Auditory Verbal Learning Test delayed recognition
Language domain	Category Fluency Test ('Animals')
	Boston Naming Test
Graphomotor speed/executive function domain	Trail-Making Test Parts A
	Trail-Making Test Parts B

These six neuropsychological test methods were chosen because they are routinely used to assess early cognitive manifestations of AD.

The raw sMRI data were pre-processed before being fed into the proposed model. A general pre-processing procedure was conducted based on Python. Skull-stripping was performed on all data through a trained U-Net model. The 3D MR images were then cropped to remove the redundant background parts. And they were resized to $64 \times 64 \times 64$, which was the input data size of the proposed model. Finally, the 3D data were normalized by subtracting their mean and dividing by their standard deviation.

4.2. Experimental settings

The proposed model was experimented in the binary classification task of distinguishing AD and NC based on ADNI dataset. Additionally, the local MCI dataset was used to further verify the proposed method. For both the ADNI and MCI datasets, the samples were divided into training and test sets in the ratio of 8:2, where 10% of the training set was further divided into the validation set. Thus, there were 154 training samples, 17 validation samples and 41 test samples in ADNI dataset, 64 training samples, 7 validation samples and 18 test samples in MCI dataset. The experiments were conducted in the Keras environment with TensorFlow as the backend. After our numerical experiments and parameters optimization, Adam solver [43] was implemented to optimize network training with the initial learning rate of 0.0001. We used the validation error to monitor the training process. When the validation error of the model has not decreased for 3 consecutive epochs, the learning rate will be multiplied by a scaling factor of 0.5 until the minimum learning rate of 1e-5 is reached. The training epoch was set to 30 and the batch size was set to 2.

The performances of the classification models were evaluated by means of some common metrics like Accuracy (ACC), Sensitivity (SEN), Specificity (SPE) and Precision (PRE):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$SEN = \frac{TP}{TP + FN} \tag{11}$$

$$SPE = \frac{TN}{FP + TN} \tag{12}$$

$$PRE = \frac{TP}{TP + FP} \tag{13}$$

where TP represents true positive, TN represents true negative, FP means false positive and FN means false negative. Among these metrics, SEN is also known as true positive rate and SPE is known as true negative rate, they are crucial indicators in medicine.

Furthermore, to perform a more comprehensive evaluation of the model, F1 score and G-mean metrics are also considered in this work. F1 score is the criterion for integrating Precision (PRE) and Sensitivity

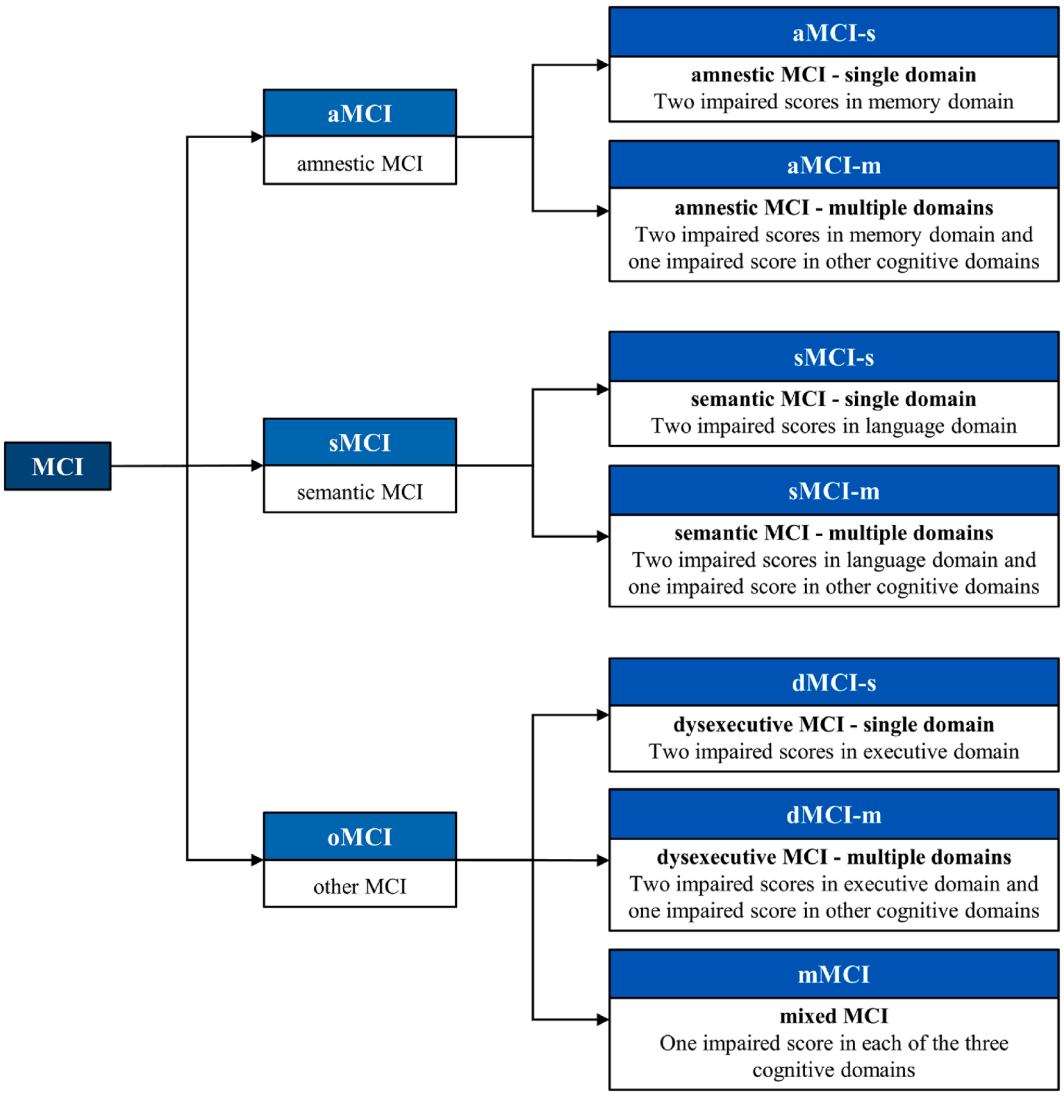


Fig. 4. Diagnostic criteria for MCI subtypes. MCI can be divided into amnestic MCI (aMCI), semantic MCI (sMCI) and other MCI (oMCI) based on the impaired scores of different cognitive domains.

(SEN), while G-mean combines Sensitivity (SEN) and Specificity (SPE). F1 score and G-mean can be calculated by the following equations:

$$F1 = \frac{2 \times PRE \times SEN}{PRE + SEN} \tag{14}$$

$$G - mean = \sqrt{SEN \times SPE} \tag{15}$$

4.3. Ablation study

This subsection aims to observe the effects of different attention mechanisms on the performance of the proposed 3D Residual U-Net model and to reveal the effectiveness of the proposed hybrid attention mechanism through ablation study. Both channel attention and spatial attention were introduced into this study and we have conducted separate experiments on each attention mechanism when acting alone and acting in combination.

The results displayed in Table 2 demonstrate that the proposed 3D Residual U-Net (3D ResUNet for short) has an overall good performance in the AD and NC binary classification of the ADNI dataset. The original 3D ResUNet framework is effective in solving this type of problem with the accuracy, specificity and precision of 87.80%, 95.45% and 93.75% respectively. Notably, the original 3D ResUNet achieved the tied highest specificity, but the lowest sensitivity, indicating that the model is more likely to predict the sample as NC. The introduction of attention mechanisms can improve this prediction tendency, the channel-based and spatial-based attention mechanisms had different effects on the

Table 2
Comparison of different attention mechanisms on the ADNI dataset (AD vs. NC).

Model	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	G-mean (%)
3D Residual U-Net	87.80	78.95	**95.45**	93.75	85.72	86.81
Cha + 3D Residual U-Net	90.24	**89.47**	90.91	89.47	89.47	90.19
Spa + 3D Residual U-Net	90.24	84.21	**95.45**	94.12	88.89	89.65
Cha-Spa + 3D Residual U-Net	**92.68**	**89.47**	**95.45**	**94.44**	**91.89**	**92.41**
Spa-Cha + 3D Residual U-Net	**92.68**	**89.47**	**95.45**	**94.44**	**91.89**	**92.41**

Cha represents channel attention mechanism and Spa represents spatial attention mechanism. Cha-Spa means the channel attention is performed before spatial attention while Spa-Cha is the opposite. Boldface indicates the best result in comparison.

performance of the model, although both made the accuracy of the original model increase to 90.24%.

The channel attention mechanism helped the 3D ResUNet model reach a certain degree of balance in the prediction of AD and NC, while the spatial attention mechanism retained the high specificity as the original model, but improved the sensitivity. The combination of the two attention mechanisms, that is, the hybrid attention mechanism proposed herein, can integrate the advantages of the two attention mechanisms and enable the model to obtain the highest evaluation scores. Compared with the original model, the hybrid attention mechanism helped increase accuracy, sensitivity, precision, F1 score and G-mean by 4.88%, 10.52%, 0.94%, 6.17% and 5.60%, respectively. In addition, the experimental study found that the execution order of the two attention mechanisms did not significantly affect the classification performance of the model. The same results were obtained whether channel attention or spatial attention was performed first. If not specified, the subsequent hybrid attention strategy means that channel attention is performed before spatial attention.

4.4. Results on ADNI dataset

4.4.1. Methods of comparison

In order to verify the effectiveness of the proposed 3D HA-ResUNet, we chose different representative and competitive methods for comparison: the machine learning classification models based on specific feature extraction methods, multilayer extreme learning machine methods and CNN-based methods, including the ResNet with different network depths.

Feature extraction combined with machine learning: The classification models were constructed by specific feature extraction methods combined with machine learning classifier. According to the characteristics of brain MR image, we chose gray-based and texture-based methods to conduct feature extraction. As for gray-based method, gray histogram, gray mean, gray variance, gray contrast amplitude, gray energy, and gray entropy were selected as metrics. Local binary pattern (LBP) was chosen as the texture-based feature extraction method because of its advantages in capturing global texture changes and local gray changes. Conventional feature extraction methods are used to process 2D images. For 3D MR images, we extracted features from three slice direction, namely sagittal, coronal and axial plane, and integrated the features to better characterize the information in different dimensions of the data. Support vector machine (SVM) was chosen as the machine learning-based classifier for its generalization capability.

Multilayer extreme learning machine: Multilayer extreme learning machine (ML-ELM) was proposed to accelerate the computational process of deep learning with its non-iterative and random feature mapping mechanism [44,45]. Among the ML-ELMs, stacked ELM autoencoder (ELM-AE) has the similar hierarchical structure as the deep neural network without tedious training, so it can quickly process high-dimensional features and obtain comparable generalization performance. We chose ELM-AE as the classifier to combine with the above feature extraction methods. Furthermore, we also constructed ML-ELM for feature extraction and classification, which facilitates direct comparison with the end-to-end CNN-based methods.

ResNet: ResNet (Deep Residual Networks) [12] proposes the residual learning unit, which applies identity mapping to solve the degradation problem of deep network, so as to give full play to the advantages of deeper CNN. ResNet has structures of different depths, and we chose ResNet50 and ResNet18, i.e., ResNet with 49 and 17 convolutional layers. We further extended the network structures to 3D to process 3D MR images.

4.4.2. Results and comparison

The proposed 3D HA-ResUNet achieved good performance in the classification of AD and NC on the ADNI dataset. We also trained other competitive methods on the same dataset for further comparison and the results are listed in Table 3.

As illustrated in Table 3, the proposed method got the highest accuracy, specificity, precision, F1 and G-mean scores of 92.68%, 95.45%, 94.44%, 91.89% and 92.41% respectively among all the competing methods. The sensitivity score of the proposed method ranked the second, reaching 89.47%. ResNet, a frequent architecture in the literature, also showed strong competitiveness in this comparison. 3D ResNet18 reached 90.24%, 89.47%, 90.91%, 89.47%, 89.47% and 90.19% in accuracy, sensitivity, specificity, precision, F1 and G-mean metrics, respectively. However, 3D ResNet50, with deeper network structure (more convolutional layers), failed to demonstrate the classification performance matching its model scale. The reason is that the model has undergone severe overfitting on the ADNI dataset used in this study, i.e., the sample size in our experiment is still too small compared with the model parameters of 3D ResNet50, resulting in poor outcomes on the test set. The predictions of 3D ResNet50 show obvious bias with low sensitivity (68.42%), that is, the model tends to predict the sample as NC, resulting in a high false negative rate (FNR) of the result.

The same phenomenon also occurred in the method of using gray features in three slice directions combined with SVM classifier. It has very low sensitivity (52.63%) but high specificity (95.45%), which will lead to serious underdiagnosis in practice. In contrast, the use of texture-based features like LBP provided better results overall, with high sensitivity (94.74%) but relatively low specificity (77.27%), which will increase the possibility of misdiagnosis in practice. However, when ELM-AE was used as the classifier, the results obtained by the gray features are overall better than those obtained by the LBP features. Similarly, whether the classifier is SVM or ELM-AE, gray features have a great probability of underdiagnosis (highest specificity, low sensitivity), while LBP features are more likely to lead to misdiagnosis (highest sensitivity, low specificity).

The combination of gray-based and texture-based features seems to be a good option to improve the classification performance and alleviate the prediction bias. In particular, when ELM-AE combines the two types of features, it can achieve an overall performance comparable to 3D ResNet18 and second only to the proposed method. This indicates that machine learning techniques and ML-ELMs also have great potential in the classification of 3D structural MR images by extracting targeted features.

Table 3
Classification results of different methods on the ADNI dataset (AD vs. NC).

Method	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	G-mean (%)
Gray features + SVM	75.61	52.63	**95.45**	90.91	66.67	70.88
Gray features + ELM-AE	87.80	78.95	**95.45**	93.75	85.72	86.81
LBP features + SVM	85.37	**94.74**	77.27	78.26	85.72	85.56
LBP features + ELM-AE	82.93	**94.74**	72.73	75.00	83.72	83.01
Gray + LBP features + SVM	87.80	89.47	86.36	85.00	87.18	87.90
Gray + LBP features + ELM-AE	90.24	**94.74**	86.36	85.71	90.00	90.45
3D ML-ELM	87.80	89.47	86.36	85.00	87.18	87.90
3D ResNet18	90.24	89.47	90.91	89.47	89.47	90.19
3D ResNet50	82.93	68.42	**95.45**	92.86	78.79	80.81
3D HA-ResUNet	**92.68**	89.47	**95.45**	**94.44**	**91.89**	**92.41**

Gray features include gray histogram, gray mean, gray variance, gray contrast amplitude, gray energy, and gray entropy. LBP means local binary pattern. Support vector machine (SVM) is the machine learning classifier and ELM-AE is the classifier based on stacked extreme learning machine autoencoder. 3D ML-ELM is multilayer extreme learning machine for feature extraction and classification. 3D HA-ResUNet is the proposed 3D Residual U-Net with hybrid attention mechanism. Boldface indicates the best result.

In many cases, accuracy and speed are often difficult to balance. Although the proposed 3D HA-ResUNet performs best on the ADNI dataset, its model also consumed the longest training time. In this experiment, the proposed model was trained 30 epochs, and each epoch took 190 s on average. The overall training time of the proposed model is slightly longer than 3D ResNet50, indicating the computational complexity of this deep CNN-based methods, especially when dealing with 3D data. As a comparison, the ML-ELM method does not require training parameters, and direct calculations can be performed for the input data, so it is fast, reflecting the characteristics of this type of method. Among them, ML-ELM took 14 s, ELM-AE took 158 s and SVM took 143 s. If the model training time is not considered, there is no significant difference in the time consumed by all methods for the prediction of the test set. It is worth pointing out that the time consumption is relative, and the time required for computing varies on different computers. As long as the time consumption is within a reasonable range, considering the accuracy that the model can achieve, deep learning-based methods can still be a good option.

4.4.3. Results on imbalanced dataset

As a possible situation in practical application scenarios, class imbalance in datasets is a problem worthy of study. In the clinical application of AD diagnosis based on MR images, the number of diseased samples (AD) accumulated in hospitals is likely to be less than the number of normal samples (NC). We refer to the literature [46] and utilize the Imbalanced ration (IR) to measure the imbalance of the dataset:

$$IR = \frac{N_+}{N_-} \tag{16}$$

where N_+ denotes the number of positive samples in the dataset, which is the number of AD samples for the ADNI dataset in this study. And N_- denotes the number of negative samples, which is the number of NC samples.

We used 100%, 80%, 50% and 30% of the AD samples to generate four levels of imbalanced datasets, i.e., IRs are 0.86, 0.68, 0.43 and 0.18, respectively. And the top 3 algorithms in Table 3 were compared here. As can be seen from Table 4, the proposed 3D HA-ResUNet shows obvious advantages when the class imbalance is not particularly severe, and even achieve a slight performance improvement over the original dataset (IR = 0.86) at 80% of the AD samples (IR = 0.68). As the proportion of AD samples becomes smaller, the performance of the proposed method starts to decrease. When IR is 0.18, although the accuracy is still high, the algorithm cannot well identify the AD samples in the test set. This is because there are fewer AD samples for model training, and the model does not fully learn the features that can identify AD well in the limited data. However, 3D ResNet18 exhibits superior classification ability when the IR value is small. ELM-AE method shows good stability in this comparison and maintains a good level of accuracy. In practical application, for severely imbalanced datasets, we can consider setting some class-specific regular parameters as in [46] and adding them to the original method to improve the model performance.

4.5. Results on MCI dataset

Mild cognitive impairment (MCI) is a state between normal aging and AD, with a high probability of eventual conversion to AD. The identification of different subtypes of MCI is a crucial issue in the early diagnosis of AD. We validated the clinical feasibility of the proposed method using MCI subjects collected from a local hospital. The advantages of using the local MCI dataset are: First, local dataset is more in line with clinical scenario in terms of small sample size and data characteristics compared to a large database like ADNI. Furthermore, local MCI subjects can be subdivided into amnestic MCI (aMCI), semantic MCI (sMCI) and other types of MCI (oMCI) based on neuropsychological assessment, which is beneficial to accurately predicting the progression from MCI to dementia.

The proposed 3D HA-ResUNet performed well on MCI dataset. To avoid the contingency on small sample dataset, we set different test set proportions (15%, 20%, 25% and 30%) to observe the prediction performance of the model on different test sets (refer to Table 5 for specific results). 3D HA-ResUNet accurately predicted aMCI and sMCI on different proportions of the test set, indicating that the proposed method is good at addressing this MCI subtype classification with a small sample size and has a satisfactory application prospect in clinic. According to our previous experiments, 3D ResNet18 also had an impressive classification performance on the ADNI dataset. Thus, it was selected for comparison on this dataset and it achieved good prediction results as well. However, when the proportion of the test set is small, the model overfitted, resulting in misclassification, which is not as robust as the proposed method overall. As indicated by the metric scores in Table 5, 3D ResNet18 misclassified aMCI cases as sMCI at 15% and 20% of the test set ratio. Fig. 5 additionally provides an intuitive representation for comparing the performance of the two methods on different scale test sets: the proposed 3D HA-ResUNet maintained 100% prediction accuracy in all tests, while 3D ResNet18 produced fluctuations in prediction performance when the scale of test set was small.

5. Discussion

This section discusses the visual interpretability and broader performance comparison of our proposed model. Section 5.1 visualizes the basis of the model for MCI subtype classification and interprets it in the context of brain anatomy. Section 5.2 compares the classification results of the proposed method with other state-of-the-art methods in the ADNI database.

Table 4
Classification results of different imbalanced ratios on the ADNI dataset (AD vs. NC).

Imbalanced ratio	Model	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	G-mean (%)
0.86	ELM-AE	90.24	**94.74**	86.36	85.71	90.00	90.45
	3D ResNet18	90.24	89.47	90.91	89.47	89.47	90.19
	3D HA-ResUNet	**92.68**	89.47	**95.45**	**94.44**	**91.89**	**92.41**
0.68	ELM-AE	87.18	75.00	95.65	92.31	82.76	84.70
	3D ResNet18	92.31	**87.50**	95.65	93.33	90.31	91.48
	3D HA-ResUNet	**94.87**	**87.50**	**100.00**	**100.00**	**93.33**	**93.54**
0.43	ELM-AE	87.88	**80.00**	91.30	80.00	80.00	**85.46**
	3D ResNet18	**90.91**	70.00	**100.00**	**100.00**	**82.35**	83.67
	3D HA-ResUNet	**90.91**	70.00	**100.00**	**100.00**	**82.35**	83.67
0.18	ELM-AE	88.89	75.00	91.30	60.00	66.67	82.75
	3D ResNet18	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	3D HA-ResUNet	88.89	25.00	**100.00**	**100.00**	40.00	50.00

ELM-AE refers to the combination of ELM-AE classifier with Gray and LBP features. Boldface indicates the best result.

Table 5
Classification results of different proportions of test sets on the local MCI dataset (aMCI vs. sMCI).

Proportion of test set	Model	ACC (%)	SEN (%)	SPE (%)	PRE (%)	F1 (%)	G-mean (%)
15%	3D ResNet18	92.86	85.71	100.00	100.00	92.31	92.58
	3D HA-ResUNet	100.00	100.00	100.00	100.00	100.00	100.00
20%	3D ResNet18	88.89	77.78	100.00	100.00	87.50	88.19
	3D HA-ResUNet	100.00	100.00	100.00	100.00	100.00	100.00
25%	3D ResNet18	100.00	100.00	100.00	100.00	100.00	100.00
	3D HA-ResUNet	100.00	100.00	100.00	100.00	100.00	100.00
30%	3D ResNet18	100.00	100.00	100.00	100.00	100.00	100.00
	3D HA-ResUNet	100.00	100.00	100.00	100.00	100.00	100.00

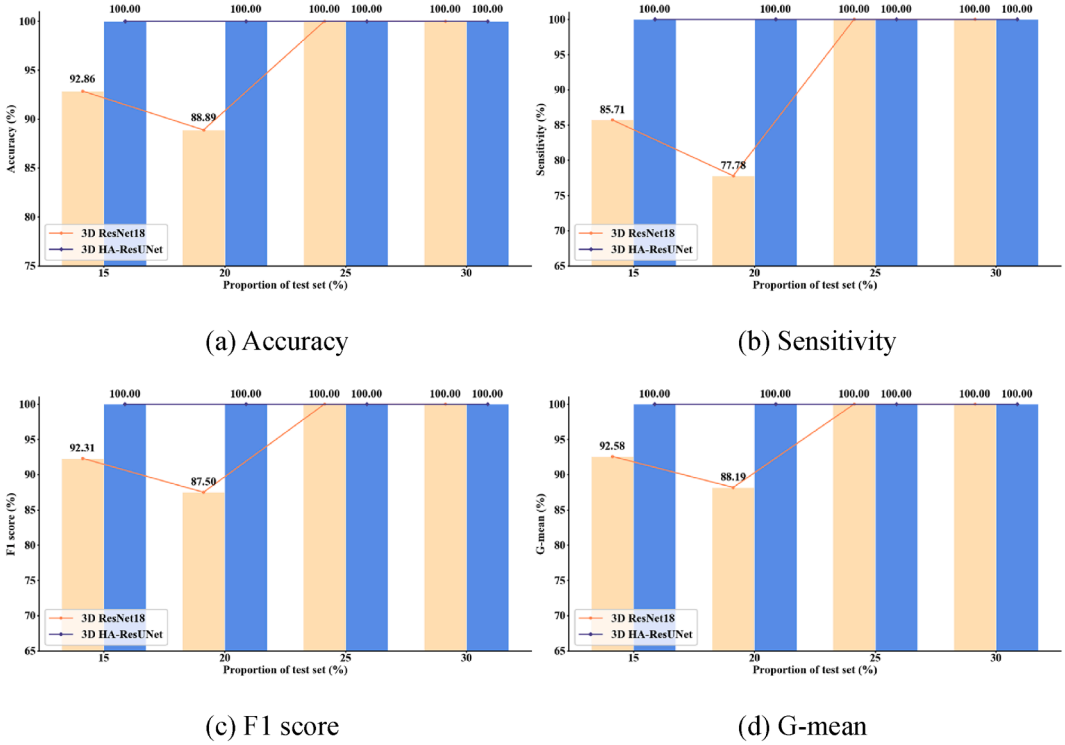


Fig. 5. Test accuracy, sensitivity, F1 score and G-mean of the model with different test set proportions on MCI dataset. The proposed 3D HA-ResUNet remained 100% for the metrics while 3D ResNet18's performance fluctuated when the proportion was small.

5.1. Visual interpretation

There are two pivotal issues to be studied in the clinical application of AD auxiliary diagnosis based on deep learning methods: one is the demand for high accuracy of the deep learning model; the other is the interpretability of the model, providing an interaction with domain knowledge. Unlike natural images, brain MR images depict the complex tissues and structures of the brain and therefore require visual interpretability methods to be both category-discriminative and capable of exhibiting fine-grained brain details. Attribution-based interpretability is able to present visual interpretation of inferences from deep learning networks, which can be well integrated with MRI-based classification process and meet the needs of most clinical diagnosis scenarios. In Section 4.5, the proposed 3D HA-ResUNet shows superior performance on the local MCI dataset, and we have made an attempt to observe the features of concern for the model to make decisions by means of the visual interpretability method mentioned in Section 3.3. For a random sample input of the MCI dataset, the importance of each voxel in the input 3D image to the prediction can be deduced according to the prediction category of the proposed model: the saliency map can be calculated from the gradient information of the last convolutional layer of the networks. Fig. 6 and Fig. 7 display the regions of attention of the proposed model's last convolutional layer for the inputs of aMCI and sMCI respectively.

In order to deliver a better visualization of the generated 3D saliency maps, we displayed slices from coronal, sagittal and axial planes separately. The significant regions and corresponding features of the data are activated with high pixel values in the saliency map, so they are white contours and textures on the slices. Fig. 6 is a visualization of aMCI data input, from Fig. 6 (a) we can see that the upper contours of the coronal plane are the key activation regions, and from the detailed magnified image we further know that these regions mainly include the frontal lobe and parietal lobe. It should be noted that the highlighted white or black areas on the slices are artifacts rather than valid activation features. On the sagittal slices, parietal lobe and occipital lobe are the key regions for the model's decision making and corpus callosum as well as its surrounding areas was also activated in specific slices. And the activated region presented on the axial slices is an external circle of contours, mainly concentrated in the regions such as cuneus, superior parietal lobule, angular gyrus and so on. We can conclude from the respective display of slices from three directions that the proposed 3D HA-ResUNet focused on the parietal lobe, occipital lobe and part of frontal lobe of aMCI samples, so as to extract key features for classification. According to the study of Li et al. [47], aMCI is significantly correlated with white matter volumes in the areas of frontal, parietal and occipital lobes.

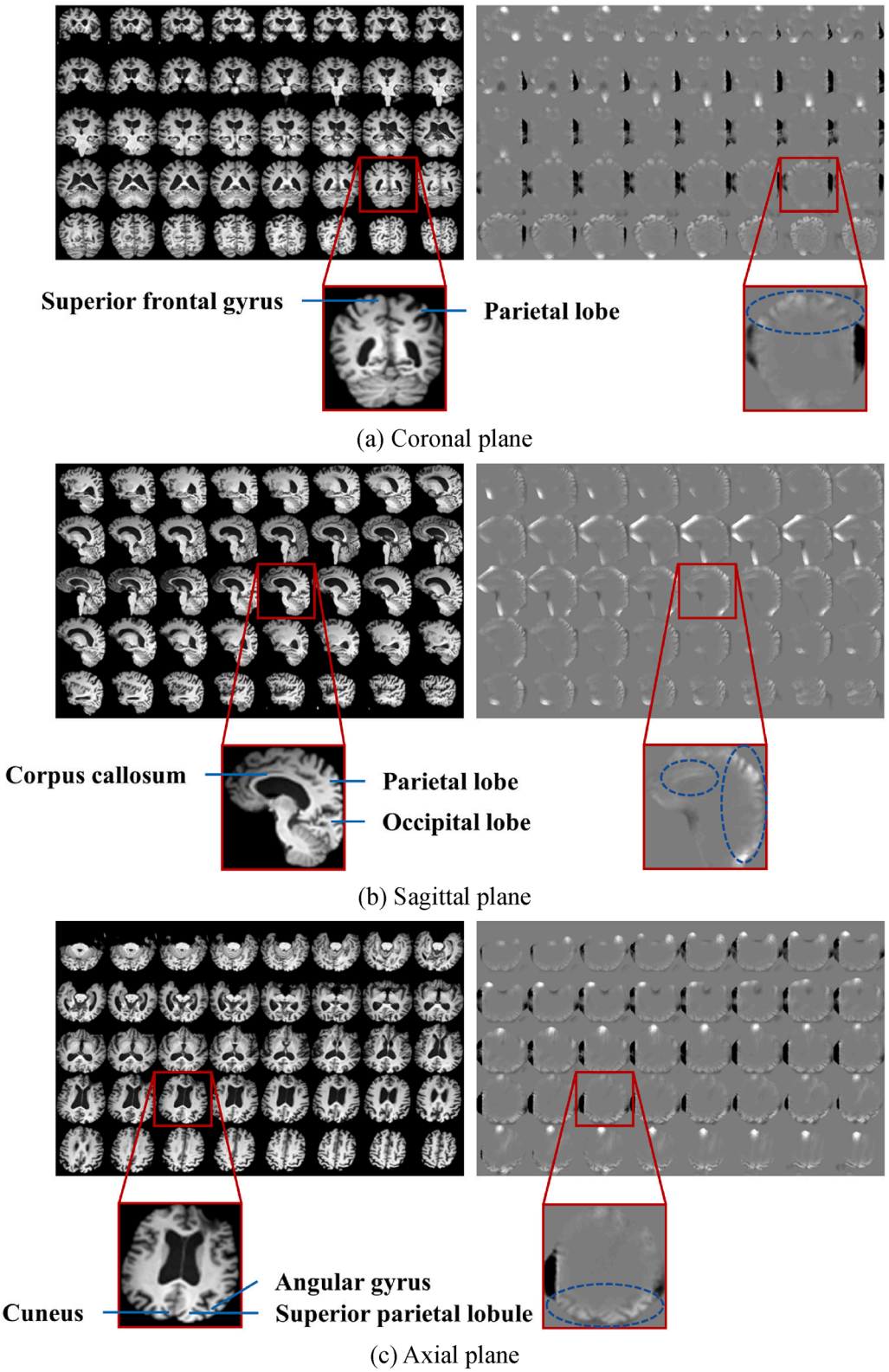
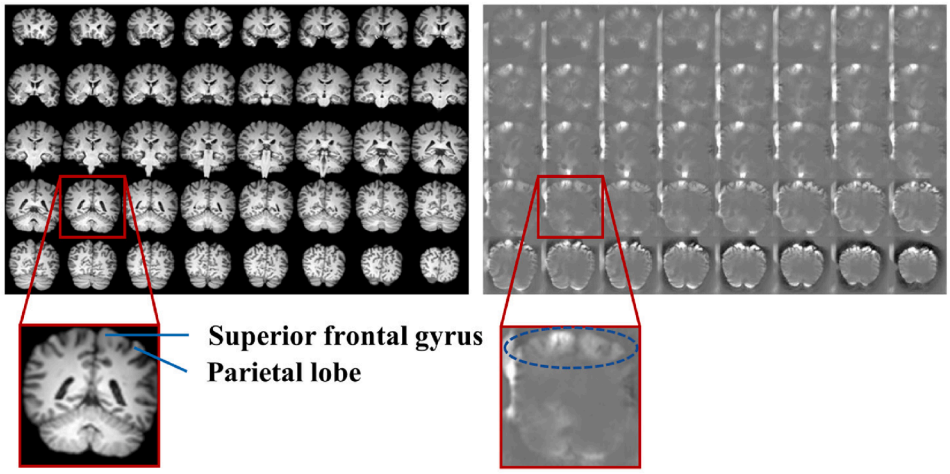
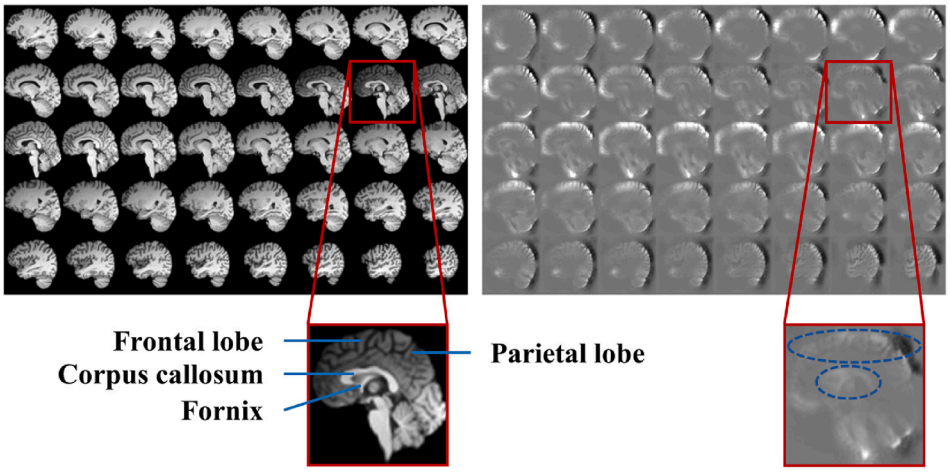


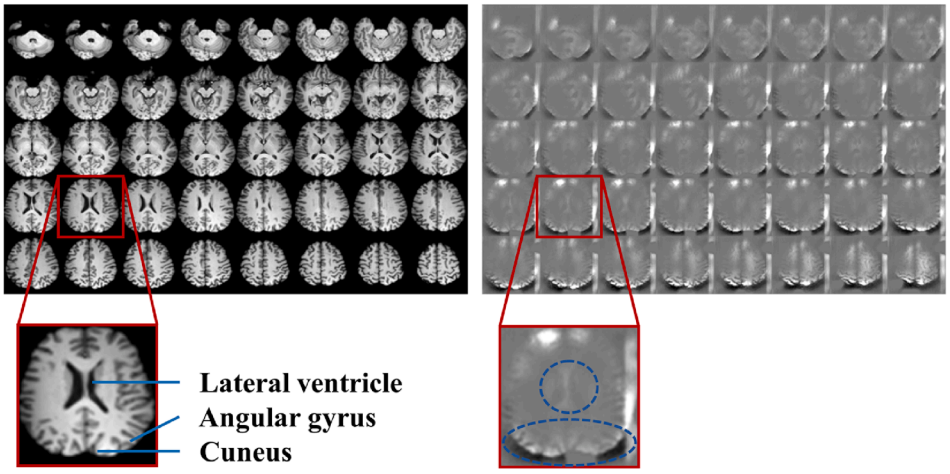
Fig. 6. Visualization of model decision features under aMCI sample input. A random sample from aMCI was fed into the trained 3D HA-ResUNet model. The regions of attention in the 3D saliency maps of the model's last convolutional layer were generated by the 3D Guided Grad-CAM method. Left side demonstrates the original slices from different directions and the right side shows the corresponding Guided Grad-CAM visualizations. For the Guided Grad-CAM images, the activated features have high pixel values, so they are white contours and textures on the images. (a) Slices and Guided Grad-CAM images of 3D MR images on coronal plane. Superior frontal gyrus and parietal lobe were significantly activated. (b) Slices and Guided Grad-CAM images of 3D MR images on sagittal plane. Parietal lobe and occipital lobe were significantly activated and corpus callosum also appeared in some slices. (c) Slices and Guided Grad-CAM images of 3D MR images on axial plane. Activation regions included cuneus, superior parietal lobule, angular gyrus and so on.



(a) Coronal plane



(b) Sagittal plane



(c) Axial plane

Fig. 7. Visualization of model decision features under sMCI sample input. A random sample from sMCI was fed into the trained 3D HA-ResUNet model. The regions of attention in the 3D saliency maps of the model's last convolutional layer were generated by the 3D Guided Grad-CAM method. (a) Slices and Guided Grad-CAM images of 3D MR images on coronal plane. Superior frontal gyrus and parietal lobe were significantly activated. (b) Slices and Guided Grad-CAM images of 3D MR images on sagittal plane. Frontal lobe and parietal lobe were significantly activated, corpus callosum and fornix were also activated in certain slices. (c) Slices and Guided Grad-CAM images of 3D MR images on axial plane. Activation regions included cuneus, angular gyrus, lateral ventricle and so on.

As for a random sample from sMCI subtype, Fig. 7 depicts the region of interest for the model to make a prediction. On the whole, the regions of attention of the model for sMCI sample are similar to those of aMCI, including partial frontal lobe and parietal lobe on coronal slices, parietal lobe and corpus callosum on sagittal slices, cuneus and angular gyrus on axial slices, etc. The difference is that the saliency maps of sMCI appear to contain richer activations, such as lateral ventricle and its surrounding areas (shown on both coronal and axial slices), and more parts like supramarginal gyrus, posterior central gyrus, anterior central gyrus are activated on axial slices. In addition, more frontal areas are activated on sagittal slices for the sMCI sample, while more occipital areas for the aMCI sample. For the same activated regions of aMCI and sMCI, we suppose that the model is able to extract discriminative features (like tissue volume, cortical thickness, depth of sulcus and gyrus, etc.) from these regions. For example, angular gyrus, known as the visual language center, is highly correlated with semantic processing [48,49], extracting features from this region may be conducive to distinguishing sMCI. And the different regions presented on sMCI can be considered as some supplementary information captured by the model, such as the lateral ventricle, central sulcus and regions around them, which may provide additional decision-making features.

The above visual interpretation process manifests the regions and features that the deep learning model focuses on for classification, making the model more transparent to a certain extent. This visual interpretation can be well integrated with the professional knowledge of medical experts because of its semantic and attribute explanations. It can not only provide a valuable reference for physicians' clinical decision-making, but also inspires physicians to further conduct multi-modal and multi-analytic research for specific ROIs.

5.2. Comparison with state-of-the-art methods

The proposed 3D HA-ResUNet is compared with other state-of-the-art methods using the ADNI database in recent literature to acquire a broad perspective on the level of the proposed method in the research field of AD computer-aided diagnosis based on MR images. These SOTA methods can be divided into two categories: with or without attention mechanism. The methods without attention mechanism screened in the literature include Marginal Fisher Analysis based on multi-kernel learning (MKMFA) [50], ensemble of multiple CNN models (Multi-CNNs) [51], 3D plain and residual CNNs (VoxResNet) [15], multi-modality cascaded CNNs (Cascaded CNNs) [52], hierarchical fully convolutional network (H-FCN) [53] and multi-model deep learning framework (multi-model CNNs) [17]. The methods combined with attention mechanism include 3D ResNet with spatial attention (3D ResNet) [27], task-driven hierarchical attention network (THAN) [29] dual attention multi-instance deep learning network (DA-MIDL) [54]. Details are listed in Table 6.

It should be emphasized that, although all methods were trained and tested on the ADNI database, the exact dataset parameters such as imaging equipment, scanning parameters, sample size, etc., and the training process of the models used in each literature are different. Therefore, the results are only used for the comparison of relative levels between methods, and the numbers do not represent the absolute superiority or inferiority. Overall, the proposed 3D HA-ResUNet is very competitive in comparison with the SOTA results recorded in the literature. Only from the value of each metric, the proposed 3D HA-ResUNet ranks second in terms of accuracy and specificity among all listed methods. In fact, 3D Cascaded CNNs [52], which ranks first in accuracy, utilized two modality neuroimaging data (MRI and PET) for training. Its accuracy is less than 93.26% when using MR images only, which means that our method achieves a high level of accuracy compared with other SOTA methods. Throughout the trends revealed in the literature studies, the use of model ensemble and multimodal data is instrumental in improving the performance of early diagnosis of AD based on neuroimaging [51–53,17]. Additionally, the attention mechanism shows great potential, and the algorithms in Table 6 that incorporate the attention mechanism have all achieved superior results. Jin et al. [27] introduced spatial attention to 3D ResNet architecture and Zhang et al. [29] devised visual and semantic attention modules in the hierarchical attention sub-network. Zhu et al. [54] constructed a dual attention framework, including spatial attention block and attention-based multi-instance learning pooling operation. As a comparison, we also proposed a hybrid attention strategy, combining channel and spatial attentions, and experimentally verified that the integration of channel and spatial attentions can fully exploit the advantages of both.

Table 6
Comparison with state-of-the-art methods using ADNI database.

Reference	Method	Attention	Number of samples	AD vs. NC (%)		
				ACC	SEN	SPE
Cao et al., 2017 [50]	MKMFA	No	AD-192, NC-229	88.60	85.70	90.40
Cheng et al., 2017 [51]	3D Multi-CNNs	No	AD-199, NC-229	87.15	86.36	85.93
Korolev et al., 2017 [15]	3D VoxResNet	No	AD-50, NC-61	80.00	–	–
Liu et al., 2018 [52]	3D Cascaded CNNs	No	AD-93, NC-100	93.26	92.55	93.94
Lian et al., 2018 [53]	H-FCN	No	AD-358, NC-429	90.30	82.40	96.50
Liu et al., 2020 [17]	Multi-model CNNs	No	AD-97, NC-119	88.90	86.60	90.80
Jin et al., 2019 [27]	3D ResNet	Spatial	AD-227, NC-305	92.10	89.00	94.40
Zhang et al., 2021 [29]	THAN	Hierarchical	AD-327, NC-416	92.00	90.30	93.10
Zhu et al., 2021 [54]	DA-MIDL	Dual	AD-398, NC-400	92.40	91.00	93.80
Proposed	3D HA-ResUNet	Hybrid	AD-98, NC-114	92.68	89.47	95.45

According to Korolev et al. [21], SEN and SPE values were not reported in their article.

6. Conclusion

In this study, a 3D Residual U-Net model incorporating hybrid attention mechanism (3D HA-ResUNet) is proposed for early diagnosis of AD based on 3D MR images. The characteristics and innovations of this study are summarized as follows:

(1) The backbone 3D CNN model contains up-sampling and down-sampling branch networks and intermediate connection residual blocks, which has both excellent classification performance and a good foundation for model interpretability.
(2) The hybrid attention mechanism combines the advantages of efficient channel attention and spatial attention, and can be applied to the skip connection of the proposed classification model to further improve the performance.
(3) A visual interpretability method based on attribution and semantic explanations is employed to reveal the regions and features that the proposed model focuses on for classification.

Compared with different representative methods, the proposed 3D

HA-ResUNet demonstrates superior generalization ability in both tasks of AD vs. NC on ADNI dataset and MCI subtype classification on local dataset. From the related research of AD computer-aided diagnosis based on MR images, the proposed method is also very competitive with other state-of-the-art methods. Furthermore, the visual interpretations can be well integrated with the domain knowledge of medical experts, making the proposed method easier to be comprehended and clinically popularized.

The visual interpretability approach used in this article is post-hoc, that is, retrospecting the saliency maps of the trained model based on the input. Future work will focus on more comprehensive interpretability techniques, which are able to actively participate in the inference process of different components of the deep learning model, and integrate more closely with the biomarker research of AD diagnosis.

CRediT authorship contribution statement

Zhiwei Qin: Conceptualization, Methodology, Software, Writing – original draft, Visualization. **Zhao Liu:** Validation, Investigation, Writing – review & editing. **Qihao Guo:** Resources, Investigation. **Ping Zhu:** Resources, Supervision, Project administration.

Declaration of Competing Interest

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

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 82171198). Data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and the local dataset of subjects with mild cognitive impairment (MCI). The contributors within the ADNI did not participate in analysis or writing of this work. A complete listing of ADNI contributors can be found online (http://adni.loni.usc.edu/).

References

[1] Alzheimer's Association, 2017 Alzheimer's disease facts and figures, Alzheimer's & Dementia. 13 (4) (2017) 325–373. 10.1016/j.jalz.2017.02.001.
[2] Alzheimer's Association, 2018 Alzheimer's disease facts and figures, Alzheimer's & Dementia. 14 (3) (2018) 367–429. 10.1016/j.jalz.2018.02.001.
[3] S. Klöppel, C.M. Stonnington, C. Chu, et al., Automatic classification of MR scans in Alzheimer's disease, Brain. 131 (3) (2008) 681–689, https://doi.org/10.1093/brain/awm319.
[4] R. Cuingnet, E. Gerardin, J. Tessieras, et al., Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database, NeuroImage. 56 (2) (2011) 766–781, https://doi.org/10.1016/j.neuroimage.2010.06.013.
[5] D.R. Nayak, R. Dash, B. Majhi, Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests, Neurocomputing. 177 (2016) 188–197, https://doi.org/10.1016/j.neucom.2015.11.034.
[6] H. Braak, E. Braak, Neuropathological staging of Alzheimer related changes, Acta Neuropathol. 82 (4) (1991) 239–259, https://doi.org/10.1007/BF00308809.
[7] D.Q. Zhang, Y.P. Wang, L.P. Zhou, H. Yuan, D.G. Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage. 55 (3) (2011) 856–867, https://doi.org/10.1016/j.neuroimage.2011.01.008.
[8] G. Uysal, M. Ozturk, Hippocampal atrophy based Alzheimer's disease diagnosis via machine learning methods, J. Neurosci. Methods 337 (2020), https://doi.org/10.1016/j.jneumeth.2020.108669.
[9] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90, https://doi.org/10.1145/3065386.
[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014. https://arxiv.org/abs/1409.1556.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, Z. Wojna, Rethinking the inception architecture for computer vision, in, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 (2016) 2818–2826, https://doi.org/10.1109/CVPR.2016.308.
[12] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, Deep residual learning for image recognition, in, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 (2016) 770–778, https://doi.org/10.1109/CVPR.2016.90.
[13] S.C. Lo, S.L. Lou, J.S. Lin, M.T. Freedman, M.V. Chien, S.K. Mun, Artificial convolution neural network techniques and applications for lung nodule detection, IEEE Trans. Med. Imaging 14 (4) (1995) 711–718, https://doi.org/10.1109/42.476112.
[14] G. Litjens, T. Kooi, B.E. Bejnordi, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005.
[15] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3D brain MRI classification, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI), 2017, pp. 835–838, https://doi.org/10.1109/ISBI.2017.7950647.
[16] H. Karasawa, C.-L. Liu, H. Ohwada, Deep 3D convolutional neural network architectures for Alzheimer's disease diagnosis, in: Proceedings of the Asian Conference on Intelligent Information and Database Systems (ACIIDS), 2018, pp. 287–296, https://doi.org/10.1007/978-3-319-75417-8_27.
[17] M.H. Liu, F. Li, H. Yan, K.D. Wang, Y.X. Ma, L. Shen, M.Q. Xu, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, NeuroImage. 208 (2020), https://doi.org/10.1016/j.neuroimage.2019.116459.
[18] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 (2017) 2261–2269, https://doi.org/10.1109/CVPR.2017.243.
[19] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.
[20] F.Q. Zhao, Z.W. Wu, L. Wang, et al., Spherical deformable U-Net: Application to cortical surface parcellation and development prediction, IEEE Trans. Med. Imaging 40 (4) (2021) 1217–1228, https://doi.org/10.1109/TMI.2021.3050072.
[21] R. Karthik, R. Menaka, M. Hariharan, D. Won, Ischemic lesion segmentation using ensemble of multi-scale region aligned CNN, Comput. Meth. Programs Biomed. 200 (2021), https://doi.org/10.1016/j.cmpb.2020.105831.
[22] Z.H. Fan, J. Li, L. Zhang, et al., U-net based analysis of MRI for Alzheimer's disease diagnosis, Neural Comput. Appl. 33 (20) (2021) 13587–13599, https://doi.org/10.1007/s00521-021-05983-y.
[23] B. Ragupathy, M. Karunakaran, A fuzzy logic-based meningioma tumor detection in magnetic resonance brain images using CANFIS and U-Net CNN classification, Int. J. Imaging Syst. Technol. 31 (1) (2020) 379–390, https://doi.org/10.1002/ima.22464.
[24] S. Maqsood, R. Damasevicius, F.M. Shah, An efficient approach for the detection of brain tumor using fuzzy logic and U-NET CNN classification, Computational Science and Its Applications (ICCSA) (2021) 105–118, https://doi.org/10.1007/978-3-030-86976-2_8.
[25] R. Karthik, U. Gupta, A. Jha, R. Rajalakshmi, R. Menaka, A deep supervised approach for ischemic lesion segmentation from multimodal MRI using fully convolutional network, Appl. Soft Comput. 84 (2019), https://doi.org/10.1016/j.asoc.2019.105685.
[26] H. Xiong, S.D. Liu, R.V. Sharan, E. Coiera, S. Berkovsky, Weak label based Bayesian U-Net for optic disc segmentation in fundus images, Artif. Intell. Med. 126 (2022), https://doi.org/10.1016/j.artmed.2022.102261.
[27] D. Jin, J. Xu, K. Zhao, F.Z. Hu, Y. Liu, Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), 2019, pp. 1047–1051, https://doi.org/10.1109/ISBI.2019.8759455.
[28] F. Yu, B.Q. Zhao, Q.Q. Ge, Z.J. Zhang, J.M. Sun, X.M. Li, A lightweight spatial attention module with adaptive receptive fields in 3D convolutional neural network for Alzheimer's disease classification, in: International Conference on Pattern Recognition (ICPR), 2021, pp. 575–586, https://doi.org/10.1007/978-3-030-68763-2_44.
[29] Z.H. Zhang, L.L. Gao, G. Jin, L.J. Guo, Y.D. Yao, L. Dong, J.M. Han, THAN: task-driven hierarchical attention network for the diagnosis of mild cognitive impairment and Alzheimer's disease, Quant. Imaging Med. Surg. 11 (7) (2021) 3338–3354. 10.21037/qims-21-91.
[30] R. Karthik, M. Radhakrishnan, R. Rajalakshmi, J. Raymann, Delineation of ischemic lesion from brain MRI using attention gated fully convolutional network, Biomed. Eng. Lett. 11 (2021) (2020) 3–13, https://doi.org/10.1007/s13534-020-00178-1.
[31] M. Hashemi, M. Akhbari, C. Jutten, Delve into Multiple Sclerosis (MS) lesion exploration: A modified attention U-Net for MS lesion segmentation in Brain MRI, Comput. Biol. Med. 145 (2022), https://doi.org/10.1016/j.compbiomed.2022.105402.
[32] B. Zhao, X. Wu, J.S. Feng, Q. Peng, S.C. Yan, Diversified visual attention networks for fine-grained object classification, IEEE Trans. Multimedia 19 (6) (2017) 1245–1256, https://doi.org/10.1109/TMM.2017.2648498.
[33] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, European Conference on Computer Vision (ECCV) (2018) 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.
[34] J. Hu, L. Shen, S. Albanie, G. Sun, E.H. Wu, Squeeze-and-Excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2017) 2011–2023, https://doi.org/10.1109/TPAMI.2019.2913372.

[35] Q.L. Wang, B.G. Wu, P.F. Zhu, P.H. Li, W.M. Zuo, Q.H. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, 2019. https://arxiv.org/abs/1910.03151v4.
[36] S. Ioffe, C. Szegedy, B. Normalization, Accelerating deep network training by reducing internal covariate shift, in: The 32nd International Conference on International Conference on Machine Learning (ICML), 2015, pp. 448–456.
[37] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. https://arxiv.org/abs/1702.08608v2.
[38] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, 2020. https://arxiv.org/abs/2012.14261.
[39] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep Networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359, https://doi.org/10.1007/s11263-019-01228-7.
[40] M.W. Bondi, A.J. Jak, L. Delano-Wood, M.W. Jacobson, D.C. Delis, D.P. Salmon, Neuropsychological contributions to the early identification of Alzheimer's disease, Neuropsychol. Rev. 18 (2008) 73–90, https://doi.org/10.1007/s11065-008-9054-1.
[41] A.J. Jak, M.W. Bondi, L. Delano-Wood, C. Wierenga, J. Corey-Bloom, D.P. Salmon, D.C. Delis, Quantification of five neuropsychological approaches to defining mild cognitive impairment, Am. J. Geriatr. Psychiatry. 17 (5) (2009) 368–375, https://doi.org/10.1097/JGP.0b013e31819431d5.
[42] M.W. Bondi, E.C. Edmonds, A.J. Jak, et al., Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates, J. Alzheimers Dis. 42 (1) (2014) 275–289, https://doi.org/10.3233/JAD-140276.
[43] D. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014. https://arxiv.org/abs/1412.6980v8.
[44] L.L.C. Kasun, H. Zhou, G.B. Huang, C.M. Vong, Representational learning with extreme learning machine for big data, IEEE Intell. Syst. 28 (6) (2013) 31–34.
[45] J. Zhang, Y.J. Li, W.D. Xiao, Z.Q. Zhang, Non-iterative and fast deep learning: multilayer extreme learning machines, J. Frankl. Inst.-Eng. Appl. Math. 357 (13) (2020) 8925–8955, https://doi.org/10.1016/j.jfranklin.2020.04.033.
[46] W.D. Xiao, J. Zhang, Y.J. Li, S. Zhang, W.D. Yang, Class-specific cost regulation extreme learning machine for imbalanced classification, Neurocomputing. 261 (25) (2017) 70–82, https://doi.org/10.1016/j.neucom.2016.09.120.
[47] Y.J. Li, L. Cong, T.T. Hou, et al., Characterizing global and regional brain structures in amnestic mild cognitive impairment among rural residents: a population-based study, J. Alzheimers Dis. 80 (4) (2021) 1429–1438, https://doi.org/10.3233/JAD-201372.
[48] J.R. Binder, R.H. Desai, W.W. Graves, L.L. Conant, Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies, Cereb. Cortex 19 (12) (2009) 2767–2796, https://doi.org/10.1093/cercor/bhp055.
[49] M.L. Seghier, The angular gyrus, Neuroscientist. 19 (1) (2013) 43–61, https://doi.org/10.1177/1073858412440596.
[50] P. Cao, X.L. Liu, J.Z. Yang, D.Z. Zhao, M. Huang, J. Zhang, O. Zaiane, Nonlinearity aware based dimensionality reduction and over-sampling for AD/MCI classification from MRI measures, Comput. Biol. Med. 91 (1) (2017) 21–37, https://doi.org/10.1016/j.compbiomed.2017.10.002.
[51] D.N. Cheng, M.H. Liu, J.L. Fu, Y.P. Wang, Classification of MR brain images by combination of multi-CNNs for AD diagnosis, in: 9th International Conference on Digital Image Processing (ICDIP), 2017, https://doi.org/10.1117/12.2281808.
[52] 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 (3–4) (2018) 295–308, https://doi.org/10.1007/s12021-018-9370-4.
[53] C.F. Lian, M.X. Liu, J. Zhang, D.G. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. 42 (4) (2018) 880–893, https://doi.org/10.1109/TPAMI.2018.2889096.
[54] W.Y. Zhu, L. Sun, J.S. Huang, L.X. Han, D.Q. Zhang, Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI, IEEE Trans. Med. Imaging 40 (9) (2021) 2354–2366, https://doi.org/10.1109/TMI.2021.3077079.