Computers in Biology and Medicine 157 (2023) 106790

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

MPS-FFA: A multiplane and multiscale feature fusion attention network for Alzheimer's disease prediction with structural MRI

Fei Liu [a,b], Huabin Wang [a,b,*], Shiuan-Ni Liang [c], Zhe Jin [a], Shicheng Wei [a], Xuejun Li [a,b], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging, Anhui University, Hefei, China
[b] School of Computer Science and Technology, Anhui University, Hefei, China
[c] School of Engineering, Monash University Malaysia, Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:
Alzheimer's disease
sMRI
Multiplane and multiscale
Attention fusion
Convolutional neural network

ABSTRACT

Structural magnetic resonance imaging (sMRI) is a popular technique that is widely applied in Alzheimer's disease (AD) diagnosis. However, only a few structural atrophy areas in sMRI scans are highly associated with AD. The degree of atrophy in patients' brain tissues and the distribution of lesion areas differ among patients. Therefore, a key challenge in sMRI-based AD diagnosis is identifying discriminating atrophy features. Hence, we propose a multiplane and multiscale feature-level fusion attention (MPS-FFA) model. The model has three components, (1) A feature encoder uses a multiscale feature extractor with hybrid attention layers to simultaneously capture and fuse multiple pathological features in the sagittal, coronal, and axial planes. (2) A global attention classifier combines clinical scores and two global attention layers to evaluate the feature impact scores and balance the relative contributions of different feature blocks. (3) A feature similarity discriminator minimizes the feature similarities among heterogeneous labels to enhance the ability of the network to discriminate atrophy features. The MPS-FFA model provides improved interpretability for identifying discriminating features using feature visualization. The experimental results on the baseline sMRI scans from two databases confirm the effectiveness (e.g., accuracy and generalizability) of our method in locating pathological locations. The source code is available at https://github.com/LiuFei-AHU/MPSFFA.

1. Introduction

People with Alzheimer's disease (AD) experience gradual and irreversible decline in their memory and cognitive abilities, eventually progressing to full dementia. The prodromal stage of AD can be further categorized into progressive mild cognitive impairment (pMCI) and stable MCI (sMCI). Over time, pMCI evolves into AD, whereas sMCI remains stable or is followed by only mild cognitive decline. It is important to estimate the likelihood of MCI conversion as accurately as possible in this early stage to provide appropriate treatment. Although no treatment has been proven to effectively prevent or reverse the process of neurodegeneration [1], early diagnosis still has important clinical value in delaying the onset of cognitive symptoms [2]. Previous studies have shown that the development of AD is correlated with the degree of cerebral cortex atrophy, and structural magnetic resonance imaging (sMRI) has been used as a biomarker in AD research [3–7] because sMRI is sensitive to the brain morphology changes caused by atrophy.

The volume reduction exhibited by the cerebral cortex and the changes in the voxel values in sMRI images [8–10] can be used as biomarkers in AD diagnosis. Researchers have focused on assessing the progression of AD based on the degree of change in brain tissues. To better extract the structural changes presented by locally abnormal brain regions associated with the disease, the corresponding MR images are divided into multiple regions based on different criteria. Voxel-based methods [7,11–15] extract tissue information from MR images to form feature vectors and construct classifiers after performing feature selection and dimensionality reduction. Region-based methods [16–26] first select specific regions of interest (ROIs) first and then extract feature information. Slice-based methods [27–32] directly use 2D image classification-based models for migration training to reduce the

* Corresponding author at: Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging, Anhui University, Hefei, China.
E-mail address: wanghuabin@ahu.edu.cn (H. Wang).

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

https://doi.org/10.1016/j.compbiomed.2023.106790
Received 5 November 2022; Received in revised form 13 February 2023; Accepted 11 March 2023
Available online 15 March 2023

complexity of the model training process. Compared with the above methods, methods [33–36] that directly use whole-brain sMRI data can prevent information loss. However, not all structural changes in the brain can be used as discriminating features for identifying AD. Thus, we design a feature discriminator to enhance the ability of the model to identify prominent abnormal atrophy.

Machine learning-based methods can be applied to analyze data by learning complex patterns in medical images. For example, in AD-related tasks, traditional machine learning-based approaches [11–15] first extract texture and morphometric features from sMRI data and then build models based on these features to identify AD and MCI. However, traditional machine learning-based approaches rely strongly on feature selection, and the amount of sMRI data is much less than the feature dimensionality. Deep learning-based methods can adaptively learn high-level semantic features from data without feature selection and make important features more prominent [37–46]. As a result, deep learning-based models can achieve better results than those produced by traditional machine learning algorithms [47–52].

To learn a better feature representation, many convolutional neural network (CNN)-based methods that integrate multiple attention mechanisms and multiscale features have been proposed, such as those in [53–57]. Though the aforementioned methods used various usage of conventional attention mechanisms, which can obtain decent performance, the AD classification task has unique characteristics. For instance, clinical assessment scores are available, which provides additional information for AD classification. Hence, our approach that considers these clinical assessment scores in the design of a new attention model is natural and reasonable. The experimental results indicate that our attention design is valid and superior to designs in aforementioned methods for AD classification.

In existing CNN-based AD diagnosis methods, to the best of our knowledge, features are generally extracted from a single plane (e.g., the coronal plane). However, the cerebral cortex atrophies in different planes have distinct characteristics, and multiplane feature fusion may yield a better diagnosis results. Furthermore, an attention module can be constructed using clinical data and combined with the general attention module to realize better AD diagnosis. Hence, we develop a model with fusion attention layers to detect multiscale discriminating cerebral cortex atrophy features in the sagittal, coronal, and axial planes. The model employs a feature encoder with multiple attention mechanisms to extract and integrate features distributed across the whole brain. Then, the obtained clinical scores are combined to select the most discriminating features. Furthermore, a feature discriminator is embedded in the model to enhance the features differences between categories. Finally, we demonstrate the interpretability of the predicted results by visualizing the features and showing the potential pathological locations.

The remainder of this paper is organized as follows. First, Section Two introduces recent progress in AD-related research. Then, the proposed method is described in detail in the Section Three. The experimental results are presented in Section Four. The effectiveness of the proposed method and the contribution of each component are discussed in Section Five and Six. The final section summarizes the work presented in this paper.

2. Related work

In the following subsections, the related works, such as those concerning AD diagnosis, multiscale features analysis, and attention mechanisms, are reviewed.

2.1. sMRI-based AD diagnosis

Many methods [7,11–15] use voxel-based morphological (VBM) measurement approaches to extract high-dimensional features (e.g., cortical thickness and gray matter intensity). Then the patterns associated with AD are analyzed based on these features. In one previous study [14], gray matter volumes and cortical thickness values were first extracted based on an automatic anatomic labeling (AAL) atlas [58] to represent the global structural features. In the study of [13], the proposed high-order feature correlation detection (HOFCD) method obtained the original features through texture analysis and VBM analysis. Then, the optimal feature subset was selected via clustering and sorting methods. Moradi et al. [12] combined MRI biomarkers with subjects' ages and cognitive measurements through a random forest classifier to construct polymeric biomarkers for predicting MCI conversion. However, the high dimensional features used in voxel-based methods tend to cause overfitting because the number of training samples is limited.

Region-based methods [16–26] segment ROIs in the whole brain. For instance, a multitask model was constructed to extract 3D patches from the hippocampus, and then the patches were input into a deep learning model to learn the patterns related to AD [17]. In a similar study [19], the left and right portions of the hippocampus were first segmented, and then the structural features derived from these two tissues were combined into a unified feature vector for global optimization through a neural network. Patch-based features were extracted from each landmark location to diagnose AD in [22,23]. Notably, region-based methods focus more on the information within regions, and how to represent the correlations between different regions is remains a challenge.

Slice-based methods [27–32] select slices from the original MRI images according to certain standards and then train models based on these slices. For example, multiple selected slices were sampled at fixed intervals along the coronal planes in MR images to train a classifier in [27]. Pan et al. [28] trained various basic classifiers by selecting multiple slices in the sagittal, coronal, and axial planes. Then, these basic classifiers were combined to construct a global classifier. In addition, the researchers not only extracted features from MRI slices but also combined these features with voxel-based methods to obtain better performance in [29]. Although slice-based methods can directly solve the problem of the small amount of available MRI data by using transfer learning and dataset expansion techniques, they have a major defect regarding to selecting disease-related slices while reducing possible data leakage during training [59,60].

In contrast to methods that use only partial data, whole-brain-based methods [33–36] train models at the subject level and their greatest advantage lies in the use of complete brain information. However, the different degrees of cortical atrophy caused by dementia are distributed in various brain areas. Thus, accurately locating disease-related discriminating information is the key to the successful identification of AD.

2.2. Multiscale feature analysis

Convolutional neural networks (CNNs) generally use fixed kernel sizes when extracting features, but this leads to information loss for small targets in the resulting feature maps. By introducing a multiscale kernel in the convolutional layer, more details of different-sized targets can be retained in the feature map. In the field of computer vision, researchers have proposed several models for representing multiscale features. For instance, a feature pyramid network (FPN) [61] obtains multiscale features by integrating the feature maps output by multiple convolutional layers, while multiscale convolutional kernels enable dilated convolution [62] to process targets with different scales. In medical neuroimaging analysis, the sizes of the possible lesion regions vary in different individuals, so it is more reasonable to fuse multiscale features.

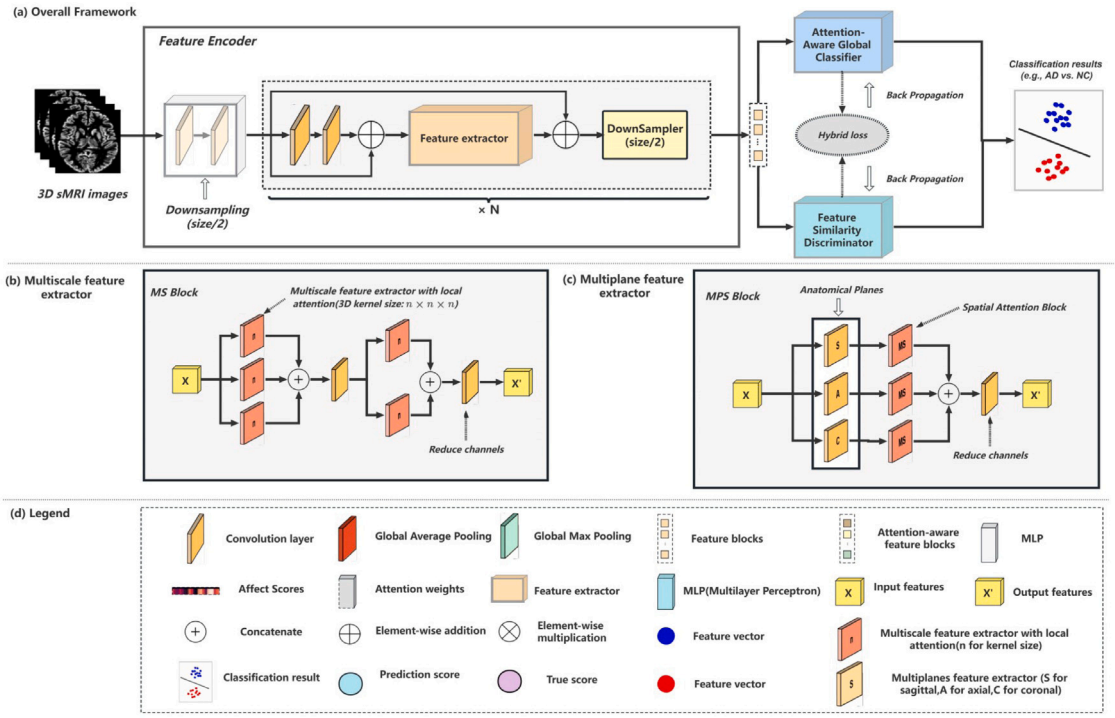

Fig. 1. The overall framework of our MPS-FFA model. The figure shows the general architecture, including an encoder, a classifier, and a feature similarity discriminator. The symbols shown in Fig. 1(d) are used to interpret the meaning of Figs. 1–3. Note: The model is constructed with a 3D CNN.

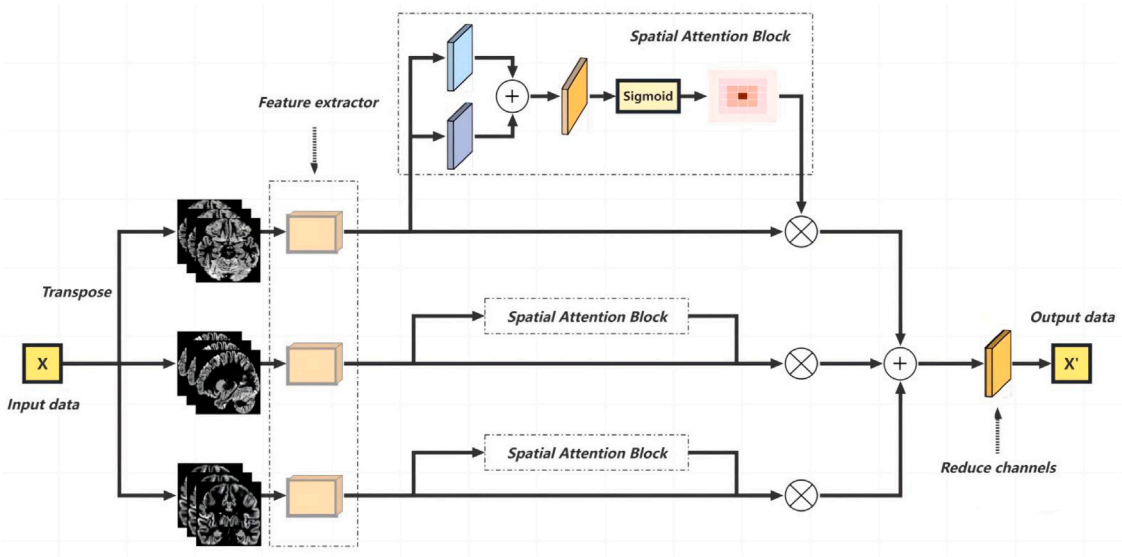

Fig. 2. Diagram of the multiplane feature extractor. Note: The meanings of the symbols used in this figure are described in Fig. 1(d).

Multiscale feature extraction methods have achieved good performance in computer-assisted medical neuroimaging analysis [59,63–71]. Existing AD diagnosis methods using multiscale features can be simply divided into four categories: (1) methods that combine whole-brain features (macro) and local ROIs (micro) to obtain multiscale information; (2) approaches that extract patches of different sizes from segmented ROIs to represent multiscale features; (3) techniques that extract multiscale features by using different size convolution kernels; and (4) methods that utilize wavelet frames to represent multiscale information. For example, a histogram-based directional gradient (HOG) descriptor quantifies the spatial gradient and calculates multiple small-scale features (SSHs) and large-scale features (LSHs) by dividing 18F-fluorodeoxyglucose (18F-FDG) positron emission tomography (PET) images of subjects' brains into different ROIs [64]. A standard voxel-based morphometry method was used in [65] to segment the brain, and then a multiscale image representation method based on wavelet frames was used to extract multiscale features in different directions from gray matter images. A joint learning-multiscale representation (JL-MSR) framework [66] that uses dilated convolutions with different expansion rates to construct multiscale feature representations has also been proposed. A new deep 3D-based multiscale CNN (3DM-SCNN) [67] was proposed, in which the features within the same region and different regions are fused. The hybrid fully convolutional network (H-FCN) [59] uses a patch-based subnetwork to generate patch-level feature representations and merges these features according to their regions.

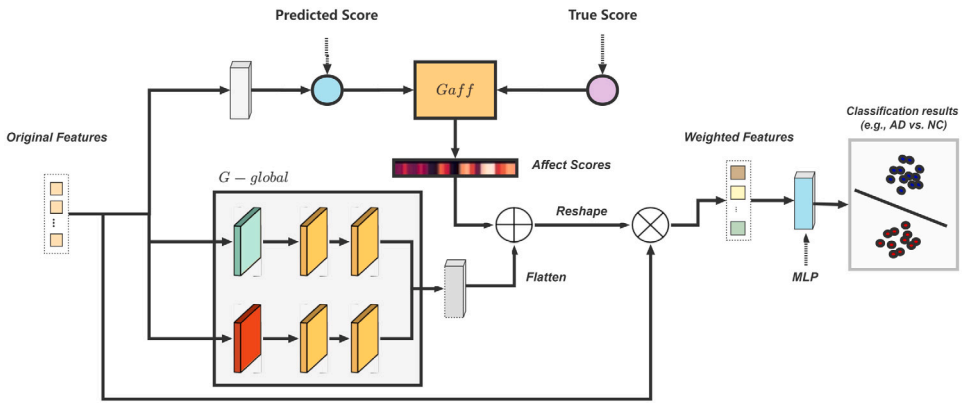

Fig. 3. Diagram of the attention-aware global classifier. Note: The meanings of the symbols used in this figure are described in Fig. 1(d).

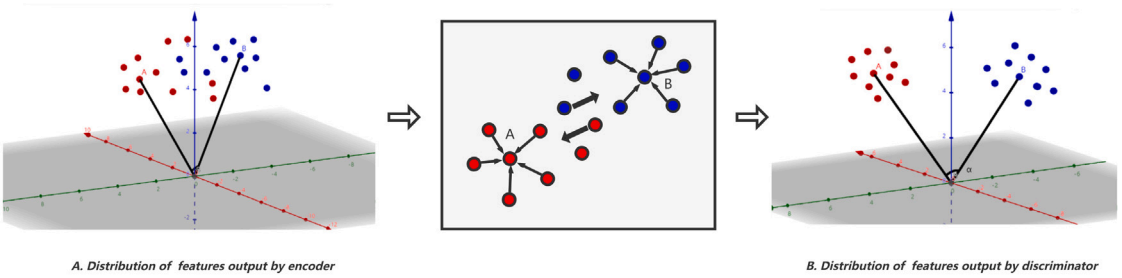

Fig. 4. Diagram of the feature similarity discriminator.

2.3. Attention mechanism

An attention mechanism can adaptively select task-related discriminating features by calculating the weight distribution of the input information to enhance local attention. An attention layer improves the feature discrimination ability of deep learning methods by enhancing important information while suppressing useless information.

In recent studies, the main role of attention layers in medical image analysis has been weighting features to enhance the contributions of important features [37–46,72,73]. For example, in the FFA-diffusion MRI (DMRI) framework proposed in [72], an attention layer was used to denoise MR images, and the final effect was superior to that obtained by traditional methods based on domain transformation and filtering. An attention-based network (ADVIAN) was proposed in which a convolutional block attention module [74] (CBAM) was added to enhance the network's feature representation ability [37]. An attention-based network in which attention blocks were used to automatically identify subject-specific discriminating features was proposed in [44]. A task-driven hierarchical attention network (THAN), which uses two attention modules to extract shallow visual features and deep semantic features, was also proposed, and the fusion of the two types of hierarchical features facilitated AD diagnosis [41].

Remarkably, researchers often combine the advantages of multiscale feature encoders and attention mechanisms to achieve better performance. For example, an attention layer was incorporated in a patch-based method to identify the important information in each region, and then the advanced features in all regions were weighted and balanced to construct a classifier [75]. Three multiscale methods were integrated into a unified framework in [68], and local anomaly representations were fused with the global anomaly information.

2.4. Research gaps and contributions

The proposed work addresses the following research gaps in Alzheimer's disease prediction approaches based on structural MRI.

(1) Existing methods, such as voxel-based, patch-based and slice-based methods [17,19,23,28,29], utilize partial sMRI images to realize reduced computational complexity, which inevitably leads to information loss.

(2) Attention modules, e.g., channel and spatial attention, have been widely applied in CNNs to realize good AD diagnosis [43,45,46]. However, existing attention modules no longer meet the increasing demands for state-of-the-art (SOTA) AD diagnosis. New mechanisms must be investigated in addition to traditional attention mechanisms.

(3) Insufficient sMRI data available for training the model limits the achievable diagnosis accuracy, and the resulting models have poor generalizability for different types of sMRI data.

Our contributions toward addressing these gaps can be summarized as follows.

(1) We propose a deep learning model, namely, MPS-FFA, that uses 3D sMRI data to extract pathological features from the entire brain, which ensures that no information is lost; in contrast, the 2D slice selection approach does suffer from information loss. MPS-FFA achieves a 97.7% accuracy for AD diagnosis, which outperforms existing SOTA methods. Considering the increased computational complexity when 3D sMRI data are used, a downsampling convolutional layer is applied along with a dense residual connection to efficiently propagate the gradients, which ensures that the computational complexity of the model remains at an acceptable level.

(2) We design a multiplane and multiscale feature fusion encoder, which is based on multiple parallel convolutions, that simultaneously integrates multiscale and multiplane features. The proposed encoder can extract more features at multiple scales from multiple planes. A visualization (Fig. 14) of the multiscale and multiplane approach is demonstrated to confirm the effectiveness of the AD pathological feature extraction.

(3) We introduce a novel attention mechanism that incorporates the clinical assessment score, which has yet to be studied. In principle, with this attention mechanism, the feature weights produced by the encoder are updated based on the clinical assessment score. This attention

Table 1
Demographic information of the subjects included in this study. Ages, years of education, and clinical scores are reported as the mean ± standard deviation.

Dataset	Category	Gender	Age	Education	CDR[a]	MMSE[b]
ADNI-1	AD	88/82	75.37 ± 7.48	14.61 ± 3.18	0.74 ± 0.24	23.22 ± 2.03
	pMCI	94/62	74.57 ± 7.11	15.74 ± 2.90	0.5 ± 0.0	26.53 ± 1.70
	sMCI	131/71	74.55 ± 7.59	15.54 ± 3.11	0.49 ± 0.03	27.37 ± 1.76
	NC	103/103	75.85 ± 5.10	15.92 ± 2.86	0.0 ± 0.0	29.14 ± 0.98
ADNI-2	AD	58/44	74.44 ± 7.89	15.99 ± 2.51	0.77 ± 0.27	22.99 ± 2.16
	pMCI	55/46	72.54 ± 6.96	15.99 ± 2.58	0.50 ± 0.04	27.55 ± 1.78
	sMCI	174/155	71.11 ± 7.51	16.16 ± 2.67	0.49 ± 0.02	28.18 ± 1.64
	NC	72/75	73.72 ± 6.39	16.68 ± 2.42	0.0 ± 0.0	29.06 ± 1.21
ADNI-3	AD	27/18	74.41 ± 8.93	15.57 ± 2.31	0.76 ± 0.30	22.37 ± 2.88
	MCI	97/80	72.08 ± 7.67	16.12 ± 2.55	0.5 ± 0.05	27.72 ± 2.02
	NC	149/228	70.95 ± 6.50	16.65 ± 2.29	0.002 ± 0.03	29.04 ± 1.24
AIBL	AD	30/44	73.35 ± 7.93	–	0.93 ± 0.55	20.18 ± 5.44
	pMCI	7/4	74.90 ± 5.97	–	0.5 ± 0.0	26.27 ± 1.60
	sMCI	33/36	75.36 ± 7.54	–	0.47 ± 0.13	27.04 ± 2.13
	NC	30/55	75.52 ± 6.63	–	0.029 ± 0.117	28.71 ± 1.35

[a]Clinical Dementia Rating.
[b]Minimum Mental State Examination.

mechanism can help the AD diagnosis, as clinical assessment scores are available widely.

(4) We design a loss function, namely, a feature similarity discriminator, that facilitates differentiating features of different classes (e.g., AD vs. normal control (NC)). With this loss function, the experimental results show a 3% improvement in the accuracy, which is satisfactory. The results confirm the necessity of the loss function.

3. Material and methods

In this section, we present the details of the proposed method in a stepwise manner. The overall framework of the method is introduced first, and then each component is analyzed in detail.

3.1. Subjects

The subjects were included in two databases: ADNI[2] and AIBL.[3] As is shown in Table 1, the subjects used in this study were categorized into three groups: AD, MCI, and NC. First, to ensure that all subjects were used only once, we removed duplicates from the two datasets. Then, the original sMRI images were preprocessed through a standard pipeline, namely, CAT12,[4] including skull dissection, intensity correction, and spatial registration.

The training dataset was specifically processed to prevent biased experimental results due to sample class imbalance. For example, in the AD classification task, the NC samples are divided into multiple subsets that contain the same number of samples as the AD group. Then, each NC subset is mixed with the AD group separately to ensure that the classes are balanced during training.

3.2. Problem setting

Let $D = \{X_i, Y_i\}_{i=1}^{N}$ represent the dataset used in this study, where X_i denotes the baseline sMRI scan of the ith sample, Y_i denotes the category label (e.g., AD, normal control (NC), and MCI), and N is the number of samples. In this study, the samples are randomly divided into two groups: $D_s = \{X_i^s, Y_i^s\}_{i=1}^{N^s}$ denotes the training data, which contains N^s samples, and $D_t = \{X_i^t, Y_i^t\}_{i=1}^{N^t}$ denotes the testing data. $D = D_t \cup D_s, (D_t \cap D_s = \emptyset)$ indicates all the data used in this study.

[2] Alzheimer's Disease Neuroimaging Initiative, https://adni.loni.usc.edu.
[3] Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing, https://aibl.csiro.au.
[4] Computational Anatomy Toolbox, https://neuro-jena.github.io/cat.

3.3. Overall framework

Our proposed MPS-FFA model (as shown in Fig. 1) consists of three main modules: a feature encoder, an attention-aware global classifier, and a feature similarity discriminator. The encoder ($F_{encoder}$) extracts features from the sagittal, coronal, and axial planes simultaneously and then obtains multiple degrees of atrophy features through multiscale feature extractors with local attention perception. Then, the features extracted by the encoder are weighted by the feature balancer (F_{score}) and output to the classifier ($\mathcal{L}_{cls}$). Next, the feature similarity discriminator ($\mathcal{L}_{diff}$) is used to calculate the feature differences between various categories before the features are input into the classifier. The proposed method is formulated in Eq. (1).

$$P(Y|X) = SoftMax\left(\mathcal{L}_{cls}\left(\mathcal{L}_{diff}\left(F_{score}\left(F_{encoder}(X)\right)\right)\right)\right) \quad (1)$$

3.3.1. Feature encoder

As shown in Fig. 1, the input images are first downsampled to reduce the image size. Then, the images are input into a parallel encoder module to extract multidimensional features. To retain more details, the downsampling operation is implemented by two pure convolution layers, which are composed of 32 channels (kernel: 4, stride: 2, padding: 1) and 64 channels (kernel: 3, stride: 1, padding: 1). Then, the images are input into a parallel 3D CNN structure, which is composed of multiplane feature extractors and multiscale feature extractors. Finally, the extracted features are combined to generate high-dimensional semantic information. The feature encoder outputs 64 feature blocks of size 4 × 5 × 4. Then, the feature blocks are weighted by the global feature balancer. Next, the feature similarity discriminator is used to calculate the feature differences. This step is performed to make the features with the same labels closer and the features with different labels farther apart and to enhance the ability of the network to extract discriminating features. Finally, the probabilities that the samples belong to different categories are output by the classifier. The following sections describe each part in detail. Notably, all the components in the feature encoder (e.g., down-sampling, MS block and MPS block) are realized by 3D-convolutional kernels.

The feature encoder constructs high-level task-related semantic information based on the original images. The feature encoder is serially connected by two structures: the multiplane feature extraction block (MPS-Block) (as shown in Figs. 2 and 1) and the multiscale feature extraction block (MS-Block) (as shown in Fig. 1). The MS-Block is composed of a group of convolution kernels with different receptive fields connected in parallel. Multiple multiscale feature extractors are connected in series to form a multiscale feature extraction subnetwork and serve as a branch of the main network, aiming at learning different

scale feature representations from the original images. The MPS-Block is composed of three feature extractors, which obtain richer feature representations by fusing multiscale features from different planes. The encoder contains four encoding layers with the same structure, as previously mentioned, and the depths of the different layers are 4, 6, 8, and 4. It is worth noting that the feature encoder is connected to a downsampling layer (kernel size: 3, stride: 2) to reduce the feature dimensionality. In addition, residual connections are used in the feature encoder to prevent gradient disappearance and explosion.

In particular, each MS-Block uses two fusion strategies: a large convolutional kernel first perceives information at a larger scale and captures interregional connections, and then a small convolutional kernel obtains local features. Each fusion strategy uses the same convolutional block structure (a base convolutional block and a multiscale convolutional block), and the depth of each layer is controlled by configuring the number of internal convolutional blocks. The base convolution block consists of a $1 \times 1 \times 1$ convolution and a $3 \times 3 \times 3$ convolution (padding: 1). The multiscale block extracts features in parallel through convolution layers with different kernel sizes and keeps the channels unchanged. The basic structure of the MS-Block is shown in Fig. 1, which consists of $3 \times 3 \times 3$, $5 \times 5 \times 5$ and $7 \times 7 \times 7$ convolutional layers or $1 \times 1 \times 1$ and $3 \times 3 \times 3$ convolutional layers connected in parallel. Then, a spatial attention layer is used to calculate the importance of each feature. Let the feature map output by the multiscale convolution layer in the MS-Block be $A = \{A^k\}$, where $A^k = [A_1^k, A_2^k, \ldots, A_c^k]$ denotes the feature map output by a convolution with a kernel size of k, $A_i^k \in \mathbb{R}^{D \times H \times W}$, $i \in [1, 2, \ldots, C]$ is the feature map in the ith output channel, and C indicates the number of channels. Corresponding to each A^k, a $1 \times 1 \times 1$ convolution is used to compute the position weights of the current map across the channels. The generated spatial weight distribution map S^k is then normalized by using the sigmoid function. This map reveals the parts that need to be emphasized or suppressed. Then, A^k is multiplied by the spatial attention map S^k in an elementwise manner to generate a feature representations with spatial attention awareness.

The proposed multiscale feature extractor with a spatial attention layer can be explained with the following mathematical formulation. Let f^w be a convolution with a kernel of size of $w \times w \times w$. The MS-Block is represented as $f_{ms} = f^1\left(\left[f^k \otimes \sigma\left(f^1\right)\right]\right)$, where k represents multiple convolution kernels, $\otimes$ denotes elementwise multiplication, and σ denotes sigmoid activation function.

The MPS-Block has the same structure as the MS-Block (a base convolution block and a multiplane convolution block). The MPS-Block first extracts features from multiple planes in parallel and then reduces the channel size with a $1 \times 1 \times 1$ convolution. The block separately extracts features along the axis and then reduces their dimensionality after combining these features along the channel dimension. Extracting features separately in three directions and then jointly locating and identifying local structures is equivalent to a multidimensional understanding of the information, which allows the model to obtain richer semantic information. Let f^s, f^c, and f^a denote the sagittal, coronal and axial feature maps, respectively; then, $f_{mps} = f^1([f^s, f^c, f^a])$ denotes the MPS-Block. Thus, the feature encoder can be expressed as Eq. (2), where DS indicates downsampling.

$$F_{encoder} = DS\left(\left[f_{mps}, f_{ms}\right]\right) \tag{2}$$

3.3.2. Attention-aware global classifier

As shown in Fig. 3, the attention-aware classifier balances the influences of different features and outputs the final classification results. The classifier adds attention weights to the features and selects the discriminating features.

Let $X = [X_1, X_2, \ldots, X_C]$ represent the features generated by the encoder. Each feature block $X_i \in \mathbb{R}^{D \times H \times W}$ $(i \in [1, 2, \ldots, C])$ is reshaped to $1 \times DWH$. G_{aff} calculates the impact score of each feature block using a shared parameter. The impact scores are then determined by combining the predicted and true values of the clinical scores (e.g., the minimum mental state examination (MMSE) scores). Additionally, G_{global} describes the global semantic information, and feature vectors with sizes of $2 \times C \times 1 \times 1 \times 1$ are output. We describe this process in Eqs. (3), (4) and (5). Moreover, $X_{predict}$ and X_{true} represent the predicted and true values of the clinical scores, respectively, σ denotes the sigmoid function, and δ denotes the softmax function.

$$F_{score} = \delta\left(G_{aff} + G_{global}\right) \tag{3}$$

$$G_{\text{aff}} = 1 - \sigma\left(\sqrt{\left(X_{predict} - X_{true}\right)^2}\right) \tag{4}$$

$$G_{global} = GMP\left(f^2\left(f^1(X)\right)\right) + GAP\left(f^2\left(f^1(X)\right)\right) \tag{5}$$

3.3.3. Feature similarity discriminator

We design a feature similarity loss function to enhance the encoder's ability to extract discriminating features by minimizing the feature similarities among features with different labels. We choose the cosine similarity function to calculate the feature similarity between the samples and average the obtained values to generate the final result. The following equations show the mathematical notations for describing the designed loss function, where l_i and l_j are the labels of different samples. By minimizing the loss $\mathcal{L}_{diff}$, the features with the same labels will become closer, and the features with different labels are easier to distinguish. We describe this process in Eqs. (6), (7) and (8).

$$\mathcal{L}_{diff} = \frac{1}{N}\sum_{i,\mathrm{j}}^{N} -1 * \left[\begin{array}{l} h\left(l_i, l_j\right)\log\left(similarity\right) + \\ g\left(l_i, l_j\right)\left(\log\left(similarity\right)\right)^{-1} \end{array}\right] \tag{6}$$

$$h\left(l_i, l_j\right) = \left\{\begin{array}{l} 1, \left(l_i = l_j\right) \\ 0, \left(l_i \neq l_j\right) \end{array}\right. \quad g\left(l_i, l_j\right) = \left\{\begin{array}{l} 1, \left(l_i \neq l_j\right) \\ 0, \left(l_i = l_j\right) \end{array}\right. \tag{7}$$

$$Similarity = \frac{\sum_{i,j}^{n} X_i \cdot X_j}{\sqrt{\sum_i^n \left(X_i\right)^2} \times \sqrt{\sum_j^n \left(X_j\right)^2}} \tag{8}$$

As shown in Fig. 4, the original features are difficult to distinguish in the vector space. However, the features in different categories become distinguishable after they are separated from each other.

3.3.4. Hybrid loss function

The learnable parameters of the encoder and classifier are jointly optimized by minimizing the hybrid loss function, which enables the model to efficiently learn disease-related discriminating features. Specifically, the hybrid loss function consists of a joint cross-entropy loss $\mathcal{L}_{cls}$ that evaluates the effectiveness of the classification process and a feature similarity loss $\mathcal{L}_{diff}$ that measures feature differences. In Eq. (10), α is a hyperparameter, which is set to 0.1 in our experiments.

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{diff} \tag{9}$$

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(P\left(Y_i | X_i\right)\right) \tag{10}$$

4. Results

In the following subsections, we first introduce the setup of the proposed method for conducting relevant experiments with two datasets. Then, we summarize the subjects used in this study and compare our experimental results with those reported in recent studies.

Table 2
The results of two tasks (i.e., AD vs. NC and pMCI vs. sMCI) performed with the ADNI database. The best and second-best results are **highlighted** and underlined, respectively.

Method	AD vs. NC				pMCI vs. sMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
VBM [15]	0.924	0.915	0.953	0.968	–	–	–	–
ROI [18]	0.855	0.808	0.898	0.902	0.678	0.646	0.700	0.682
PLM [26]	0.923	0.915	0.945	0.969	0.724	0.367	0.909	0.734
ATT [46]	0.919	0.887	0.945	0.965	0.827	0.579	0.866	0.793
MIL [75]	0.924	0.910	0.938	0.965	0.802	0.771	0.826	0.851
MSM [63]	0.935	0.941	0.929	0.962	0.833	**0.875**	0.810	**0.908**
Ours	**0.977**	**0.968**	**0.985**	**0.977**	**0.883**	0.840	**0.944**	0.892

Table 3
Comparison with related approaches conducted using the ADNI database. The best and second-best results are **highlighted** and underlined, respectively.

Reference	Method	AD vs. NC				pMCI vs. sMCI			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
VBM [12]	VBM+RF	–	–	–	–	0.820	**0.870**	0.740	**0.900**
ROI [17]	ROI+CNN	0.889	0.866	0.908	0.925	–	–	–	–
ROI [19]	ROI+CNN	0.925	0.882	0.949	**0.978**	–	–	–	–
PLM [23]	Patch+PCA	0.895	0.879	0.908	0.924	–	–	–	–
SLC [28]	Slice+CNN	0.840	–	–	0.920	0.620	–	–	0.590
SLC [29]	Slice+SVM	0.876	0.841	–	0.903	0.671	0.345	–	0.865
ATT [38]	Whole Brain+CNN	0.910	0.910	0.920	–	0.820	0.810	0.810	–
ATT [42]	Patch+CNN	0.937	0.889	0.980	0.951	0.800	0.650	0.900	0.745
MSM [65]	GM+Hip+Wavelet	0.841	0.824	0.855	0.900	0.766	0.718	0.822	0.790
MSM [59]	Patch+FCN	0.903	0.824	0.965	0.951	0.809	0.526	0.854	0.781
Ours	GM+CNN	**0.977**	**0.968**	**0.985**	0.977	**0.883**	0.840	**0.944**	0.892

Table 4
Comparison with attention-based methods. The best and second-best results are **highlighted** and underlined, respectively.

Model	Method	AD vs. NC				pMCI vs. sMCI			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
3DResAttNet [38]	Self Attention	0.910	0.910	0.920	–	0.820	0.810	0.810	–
TPA-GAN [40]	Self Attention	0.920	0.891	0.940	0.956	0.753	0.773	0.741	0.786
AD2A [43]	Spatial Attention	0.925	0.750	**0.957**	0.957	0.780	0.534	0.866	0.788
pABN [45]	Dual Attention[a]	0.872	0.890	0.856	0.927	0.793	0.546	0.841	0.776
HybNet [46]	Spatial Attention	0.919	0.887	0.945	**0.965**	0.827	0.579	0.866	0.793
MPS-DA (Ours)[b]	Dual Attention[a]	0.919	0.884	0.936	0.910	0.884	0.880	**0.889**	0.884
MPS-GA (Ours)[c]	Global Attention	**0.947**	**0.953**	0.942	0.948	**0.907**	**0.920**	**0.889**	**0.904**

[a]Spatial Attention + Channel Attention.
[b]MPS with dual attention.
[c]MPS with an attention mechanism that incorporates clinical assessment scores.

Table 5
Comparison with related approaches conducted using the AIBL database. The best and second-best results are **highlighted** and underlined, respectively.

Method	AD vs. NC			
	ACC	SEN	SPE	AUC
ATT [43]	0.903	0.873	0.908	**0.953**
ATT [46]	0.898	0.873	0.902	0.946
MIL [75]	0.902	0.848	0.915	0.939
MIL [76]	0.923	0.889	0.930	0.950
Ours	**0.949**	**0.929**	**0.972**	0.951

Table 6
Summary of the ablation studies.

Experiment	AD vs. NC		pMCI vs. sMCI	
	ACC	AUC	ACC	AUC
Multiscale feature encoder (MS)	0.887	0.887	0.884	0.884
MS with channel attention (MSA)	0.914	0.895	0.884	0.877
Multiplane feature encoder (MPS)	0.895	0.893	0.884	0.884
MPS with spatial attention (MPSA)	0.914	0.884	0.837	0.844
Feature encoder[a] (MSA + MPSA)	0.919	0.910	0.884	0.884
Attention-aware global classifier[b]	0.947	0.948	0.907	0.904
Feature similarity discriminator[c]	0.977	0.977	0.884	0.892

[a]Feature encoder with dual attention mechanisms.
[b]The proposed attention mechanism incorporates the clinical assessment scores.
[c]Feature similarity loss function and cross-entropy loss function.

4.1. Experimental settings

The MPS-FFA model is programmed with the PyTorch framework. We use batch normalization (BN) and an activation function (a rectified linear unit (ReLU)) to ensure a smoother training process and add more nonlinear constraints. We use an early stopping strategy (the process stops if the validation accuracy does not improve for 5 consecutive batches) to prevent overfitting. The model uses an Adam optimizer with an initial learning rate of 0.0001 to update the parameters. If the number of parameters is significantly larger than the number of samples (sMRI images), models tend to overfit. To alleviate this problem, we randomly augment the data during training. Specifically, the input images are augmented through three steps: (1) the images are randomly rotated by ±10° along the axial, sagittal or coronal plane; (2) the images are randomly flipped along the z-axis; and (3) the images are randomly masked. The random probability is set to 0.5 in each step.

To evaluate the experimental results, several metrics, such as the accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC), are used in this paper.

Furthermore, TP, TN, FP, and FN, represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. These evaluation metrics can be formulated as:

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

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

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

4.2. Competing methods

We compare the proposed MPS-FFA method with a variety of mainstream methods, including: VBM methods [15], ROI-based methods [18], patch-based methods (PLMs) [26], fused attention methods (ATT) [46], and multi-instance methods (MIL) [75]. In addition, we also compare our approach with various multiscale feature-based methods (MSM) [63]. Brief descriptions of these competing methods are provided below.

(1) VBM [15]: This approach first analyzes the significant gray matter volume differences between AD and NC patients as candidate regions, and the voxel intensity values extracted from the candidate regions are used as raw features. Then, the most discriminating feature subset is selected to construct the classifier.

(2) ROI [18]: A mapping of cortical thickness differences is constructed by a t-test, which is used to generate candidate ROIs, and then a set of discriminating feature vectors is generated to train the classifier after feature selection

(3) PLM [26]: A patch-based model (MM-DBM) is built by selecting distinct patches for each class through interclass significance tests, which are used to determine shared feature representations for pairs of patches.

(4) ATT [46]: This method utilizes an attention-based framework to obtain multilevel discriminating information from sMRI data for AD diagnosis.

(5) MIL [75]: This technique employs a dual-attentive network (DA-MIDL), which uses multi-instance learning [77,78] to balance the relative contributions of features.

(6) MSM [63]: A multiscale structural mapping procedure is used to obtain a wider range of regional differences by quantifying morphometric-based macrostructure and microstructure features in multiple cortical regions.

4.3. Experimental results obtained with the ADNI dataset

The values of the four metrics are listed in Table 2. The results show that in general, our proposed method obtains the best performance. For example, our method obtains the best classification accuracy on both tasks (0.977 and 0.883). Specifically, in the AD classification task, the sensitivity value of our method is 0.968, which is significantly higher than those of other competing methods. These results indicate that our method has high accuracy while ensuring a low misclassification rate. Furthermore, the results in Table 2 also show that accurately predicting future MCI conversion is more challenging than identifying AD. However, our method still achieves an overall better and more stable performance than the competing methods. This may be due to the use of multiscale and multiplane feature extractors, as these components detect small-scale structural changes related to AD.

In the voxel-based and region-based methods, feature locations first need to be discriminated by preselecting the brain regions associated with the disease. These VBM-based [15], ROI-based [18], and PLM-based [26] use feature selection techniques that allow the model to learn fewer features and therefore achieve better results, as shown in Table 2. However, these methods based on feature selection have two main limitations. First, they are all based on the assumption of brain location consistency and thus ignore individual variability; second, feature selection is often based on local features without fully considering global feature correlations. Therefore, compared with methods based on whole-brain features for localization and diagnosis (e.g., ATT [46], MIL [75], and MSM [63]), the methods based on local feature selection do not have advantages in terms of the overall performance. Furthermore, the method based on multiscale feature fusion (e.g., MSM [63]) obtains better performance on the second task. The potential reason for this result may be that the structural changes in the brains of MCI patients are not obvious and that large individual differences are present; thus, the single-scale features cannot characterize the subtle structural changes distributed throughout the brain.

Fig. 5 shows the convergence curves of the training loss and validation accuracy. Since we applied an early stopping strategy during model training, the training time is reduced. Fig. 5 displays that the loss curve stabilizes rapidly after several training epochs. The loss and accuracy plots are consistent for the two datasets, indicating that the model has good generalizability.

4.4. Comparison with related works

We compare the proposed method more extensively with the approaches developed in related studies in Table 3. Our method achieves better results than the comparison methods in both tasks. The results obtained by methods based on deep learning algorithms [19,38,42,59] are significantly better than those obtained by methods based on traditional machine learning algorithms [12,23,29] in both tasks. The possible reason for this is that deep learning algorithms can extract generic low-level features such as spatial information and structural information from the original images, thereby generating better high-level semantic features through multilayer encoders. Moreover, in contrast to region-based methods, our approach extracts structural features directly from the whole brain. However, our method still achieves competitive results, which indicates that the model can effectively identifies pathological locations after fusing the attention layer and multiscale features. Finally, we obtain better performance using gray matter information than the comparison methods using the whole brain [38]. Thus, changes in gray matter atrophy patterns may accurately reflect disease progression [3,8–10].

4.5. Comparison with attention-based methods

The results of our method are also compared with those of attention-based approaches.

To enhance the model's ability to capture discriminative features, most of the existing proposals [43,45,46] introduced the spatial or channel attention layers, while other works utilized the self-attention mechanism [38,40]. We integrate parallel dual attention layers (spatial and channel attention) in the feature encoders (MPS-DA). Specifically, in order to consider the cross-channel feature weights, we introduce a channel attention layer in the multiscale feature encoder. Moreover, to effectively consider lesions in different planes, we add a spatial attention layer to the multiplane feature encoder. As shown in Table 4, our model achieved the best pMCI and sMCI classification results when the dual attention layer is used. In addition, we combine the predicted clinical scores with a global attention layer (MPS-GA) to balance the relative contributions of the features output by the encoder, resulting in significant performance improvements in both tasks.

4.6. Generalization on the AIBL dataset

We also evaluate our method with the AIBL dataset, which was not used in the training process. As shown in Table 5, our method obtains better performance than the comparison approaches in terms of accuracy, sensitivity and specificity. In addition, the performance metrics of our method all remain stable without significant changes. This result indicates that our method is robust and can effectively identify AD-related discriminating features.

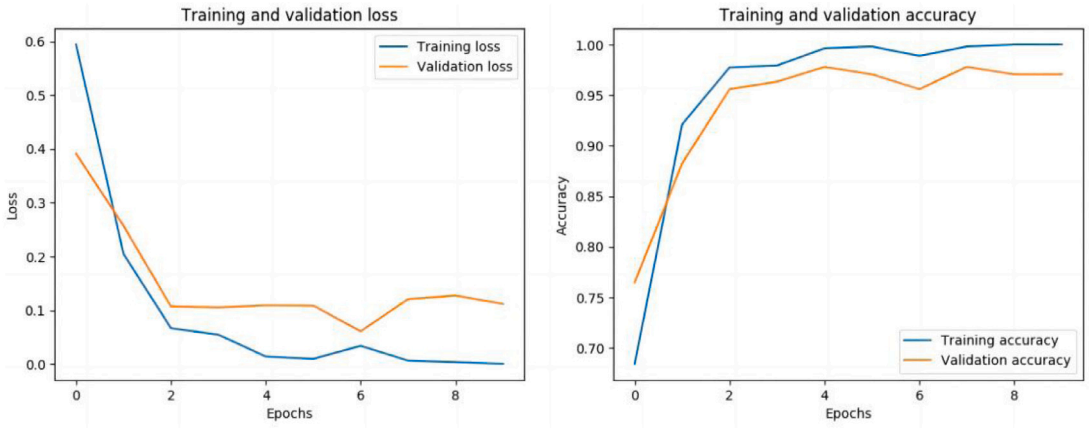

Fig. 5. Convergence curves of the loss and accuracy with the proposed MPS-FFA architecture.

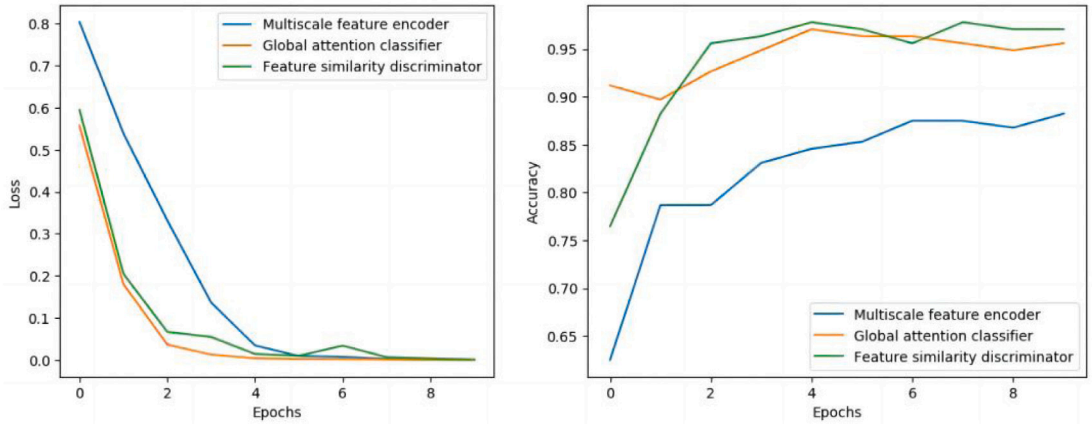

Fig. 6. Convergence curves of the loss and accuracy after different modules are integrated into the model.

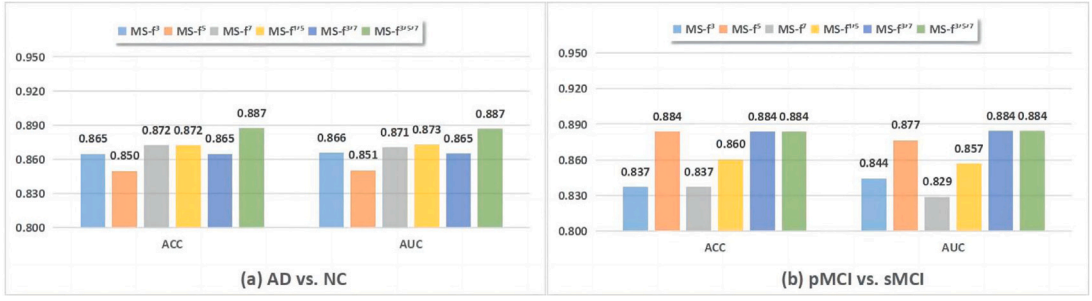

Fig. 7. The impact of the encoder using a multiscale feature extractor. The left panel shows the performance achieved on the AD classification task, and the right panel shows the performance achieved on the MCI conversion prediction task. The impacts of different convolutional kernel sizes (e.g. $f^3, f^5, f^7, f^{1,5}, f^{3,7}, f^{3,5,7}$) on the results are indicated in the figure legend.

Fig. 8. The impact of the encoder using a multiplane feature extractor. The impacts of extracting features from the axial, sagittal, and coronal planes are shown in the plots.

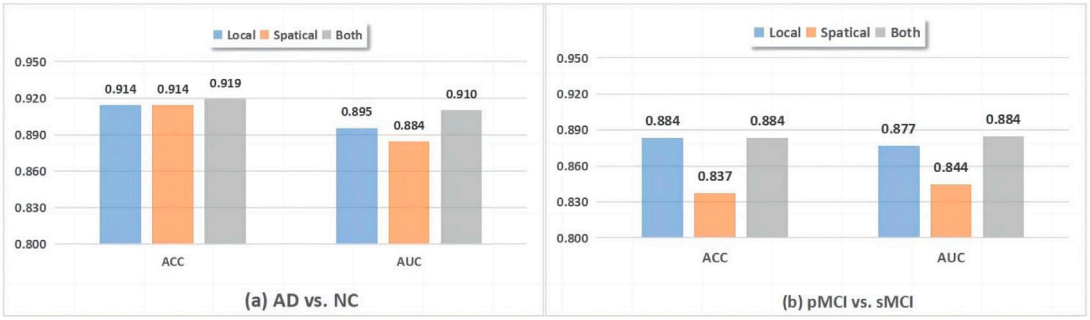

Fig. 9. The impact of the encoder after adding attention layers. "Local" means adding only a local attention layer to the multiscale encoder, while "Spatial" means adding only a spatial attention layer to the multiplane encoder, and "Both" indicates adding both layers.

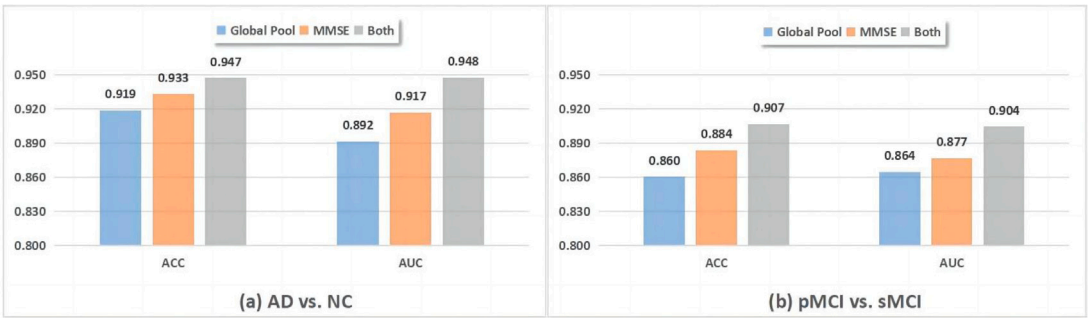

Fig. 10. Results of the ablation experiments conducted on the attention module, where "Global Pool" means that only the global pooling layer is used, "MMSE" means that only the clinical scores are used, and "Both" represents that both methods are combined.

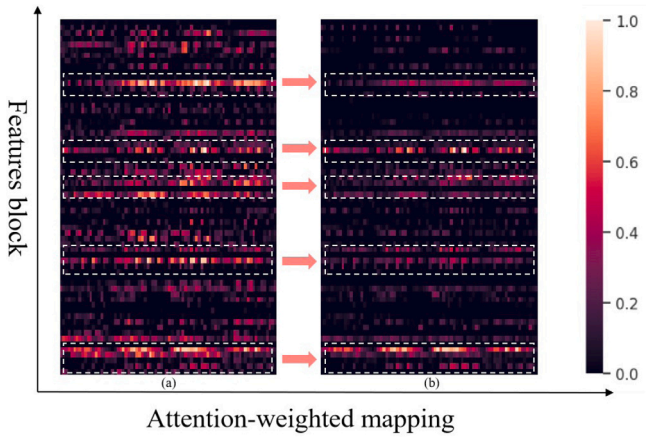

Fig. 11. The proposed fusion attention balancer strengthens the important discriminating features while suppressing useless features. The left panel (a) shows the original feature matrix and the right panel (b) shows the feature matrix after attention-weighted mapping. Note: The white dashed boxes in the figure represent the features with strong discriminative power, which are retained by the model.

5. Ablation studies

In this section, ablation studies are carried out based on different component in the proposed model. Each of the components is analyzed independently, and then the components are aggregated to construct the final model. First, the multiscale feature extractor (MS) and a multiplane feature extractor (MPS) are independently validated. Then, these modules are combined in parallel to extract features from the sMRI images. The encoded features are input into the latter modules to evaluate the performance of the attention-aware global classifier and the feature similarity discriminator. The impacts of each individual component and their combinations on the model performance are shown in Table 6 and Fig. 6. For the classification of AD and NC, the accuracy is significantly improved by approximately 2% after dual attention layers are integrated into the encoder. In addition, the integration of the feature similarity discriminator improves the classification efficiency by about 5%. For the classification of pMCI and sMCI, all indicators increased after the inclusion of clinical scores and the two global attention layers were incorporated. The results of the ablation experiments are analyzed in the following sections.

5.1. Feature encoder performance analysis

The encoder extracts features with different scales in various directions to obtain richer feature representations. To further validate the effectiveness of the encoder, we conduct separate experiments and evaluate its performance.

The performance achieved by the feature extractor with different scales is shown in Fig. 7. The figure shows that the best results are obtained by using the multiscale feature extractor. In particular, the performance improvement in the AD classification task obtained by fusing multiscale features is more pronounced than that achieved in the MCI conversion prediction task. The potential reason for this result may be that the degrees of brain atrophy differ significantly between AD and NC patients, and thus, better performance can be obtained when multiscale features are considered than when single-scale features are applied. Although pMCI and sMCI are difficult to distinguish, the use of multiscale features still realizes better performance than the use of single-scale features.

The encoder with a multiplane extractor can realize better performance than the comparison encoders. In the AD classification task, the encoder that extracts features from only the sagittal plane obtains better results than the encoders that extract features in the axial and coronal plane. This result is likely due to the better perception of the structure of the hippocampus and the surrounding cortex in the sagittal plane. In contrast, in the MCI conversion prediction task, the encoder that extracts features from only the sagittal plane does not perform well.

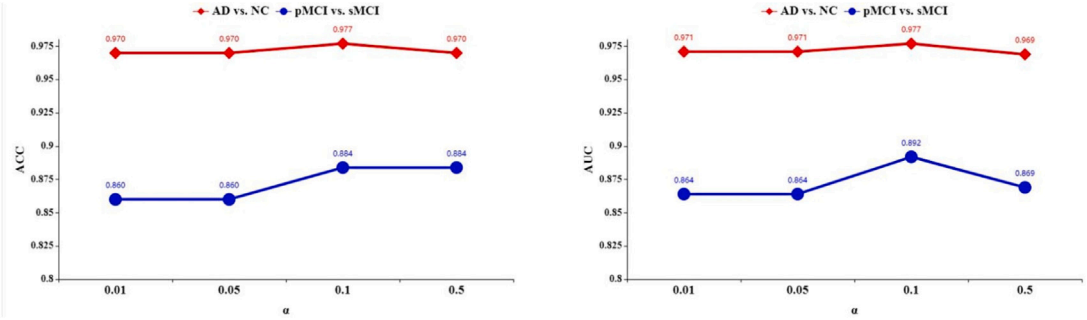

Fig. 12. Effects of different hyperparameter values on the loss function. The horizontal axis indicates different parameter values, and the vertical axis indicates the results in terms of the relevant performance assessment metrics (the ACC and AUC).

This result indicates that the structural changes in the brain vary greatly between AD and MCI patients, and features from multiple planes must be extracted to ensure better results. The performance of the multiplane extractor is displayed in Fig. 8.

We also add an attention layer to the feature extractor to ensure that the model focuses on the most important regions. As shown in Fig. 9, large performance gains are obtained after adding attention layers to both feature extractors alone. This result indicates that the attention layer enhances the ability of the encoder to identify important regions. In addition, better performance is obtained by using both attention layers in the feature encoder.

5.2. Effectiveness of the attention layer

To assess the effectiveness of the feature balancer, four further experiments are conducted with alternate models, including encoders with only a fully-connected layer (Baseline), global attention weighting (AFF_Global), clinical score-based prediction weighting (AFF_Mmse), and a combination of both weighting methods (AFF_All). We evaluate the performance of these four methods in both tasks.

As shown in Fig. 10, our proposed global feature balancer generally improves the classification performance in general. For example, the approach with dual-feature balancing obtains higher accuracy (94.7% and 90.6%) in the two tasks than the method using only one weighting method. The method using MMSE prediction scores obtains better performance than the approach using global pooling, suggesting that the use of clinical scores are advantageous in guiding the prediction process. Moreover, better overall performance gains are obtained by using both approaches, indicating that weighting different characteristics is beneficial for AD-related tasks.

A comparison between the original features and the features weighted by the global feature balancer is shown in Fig. 11. The global feature balancer strengthens the important discriminating features while suppressing the influence of useless features. Moreover, the feature balancer helps to reduce the risk of overfitting.

5.3. Effect of the hyperparameter on the loss function

The impacts of different parameter values in the loss function are summarized in Fig. 12. The parameter value is restricted to the range [0.01, 0.5]. As shown in Fig. 12, the performance indicators increase significantly with increasing parameter values when the parameters are restricted to the range of [0.05, 0.1], while the performance significantly decreases when the parameter value exceeds 0.1. Furthermore, the effects of the parameter values are more pronounced in the MCI conversion prediction task than in the AD classification task. Specifically, when the parameter value is changed from 0.05 to 0.1, the ACC value increases by 2.4% (from 86% to 88.4%) and the AUC value increases by 4.4% (from 86.4% to 89.2%) in the MCI conversion prediction task, while the values of these two metrics increase by only 0.7% (from 97% to 97.7%) and 0.6% (from 97.1% to 97.7%) in the AD classification task, respectively. This result indicates that MCI is more sensitive to the parameter value. The potential reason for this may be that MCI is difficult to distinguish, while the feature similarity loss function is effective in increasing the differences among features, thus improving the overall performance. However, this impact is not obvious when the parameter is set to a small value. For example, when the parameter value is less than 0.05, the two performance metrics do not change because when the parameter value is very small, the similarity loss function cannot effectively guide the model parameter update process. Therefore, it is reasonable for us to set the parameter to 0.1 in our experiments.

6. Discussion

The proposed model uses fusion attention layers to identify abnormal voxels that can be used to distinguish AD/NC and pMCI/sMCI. We visualize these voxels through spatial heat maps, and find that they are highly consistent with the pathological locations of AD and MCI respectively.

As an AD diagnosis aid, our model can accurately identify possible pathological locations in sMRI images, thereby helping physicians focus on the areas that are most relevant to the development of AD. In addition, the model can accommodate the variability in the locations of pathologies across subjects, and accurately identify changes in the local microstructures that are widely distributed in the brain.

We visualize and analyze the results generated by the model with the ADNI dataset. The highlighted areas in the upper panel of Fig. 13 show the most discriminating voxel distributions localized by the model, while the lower panel shows the relevant regions. The model localizes its focal areas to the sulcus gyrus and the edge of the gray matter, including the hippocampus, amygdala, occipital lobe, temporal lobe, and central sulcus. Notably, the regions labeled by the model agree with the results of previous studies [3,9]. These regions are highly correlated with cognition and long-term memory and are considered neuroanatomical markers for AD diagnosis. Thus, our approach is effective in identifying local structural changes caused by brain atrophy.

To further illustrate the brain ROIs identified by the model, we provide 3D visualizations [79] of the important cortices according to the model. As shown in Fig. 14, the spatial heatmaps demonstrate the different levels of voxels identified by the model. As a result, abnormal voxels can more easily be distinguished. From a global perspective, the distributions of relevant discriminating voxels in the brains of AD patients are more widespread than those in the brains of MCI patients, and this result is consistent with the results presented in Fig. 13. From a local perspective, the voxels of interest to the model are mainly distributed in the parahippocampal gyrus, visual cortex, frontal middle gyrus, cuneus, and motor areas. These voxels are mainly concentrated in areas of the brain that are related to memory and visual and motor control; this finding is consistent with existing studies [3,9,80].

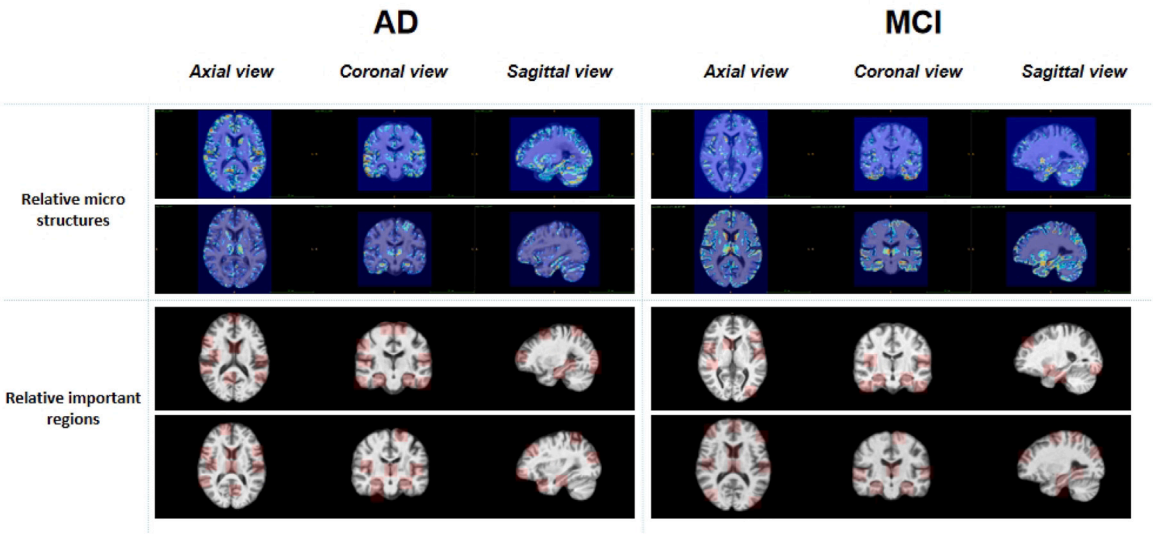

Fig. 13. Discriminating pathological locations identified by the model in the AD classification (left panel) and MCI conversion prediction (right panel) tasks. In the upper panel, the brain regions identified by the model after weighting the features by their impact scores and the attention layer are shown. The lower panel shows the marked significant regions. The subjects were randomly selected from the AD/MCI groups, and the figure displays the pathological locations identified by the model.

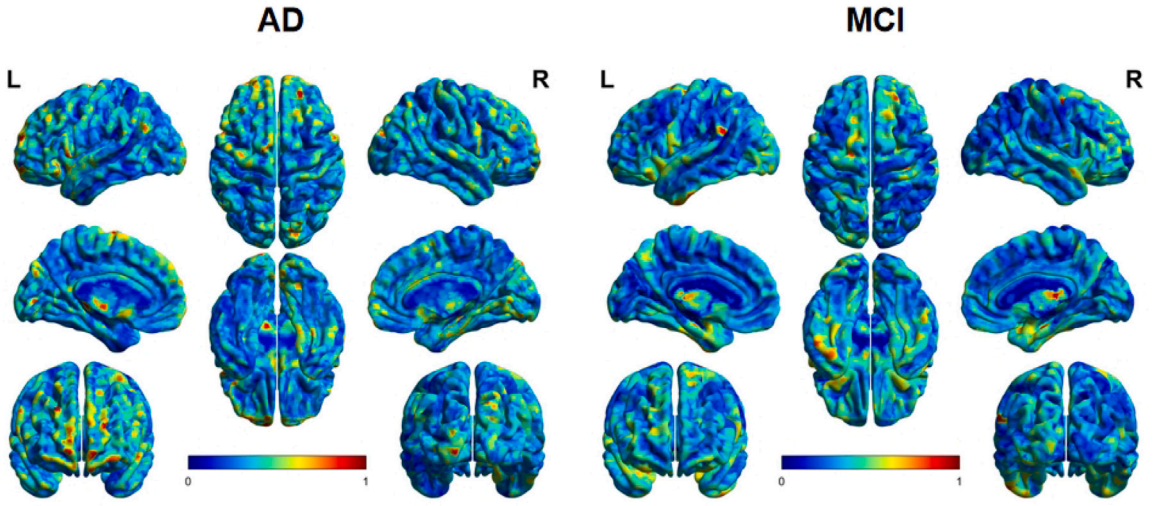

Fig. 14. The spatial heatmaps show the ROIs identified by the model. From left to right, the outer views, such as the left hemisphere, the top view, and the right hemisphere, are shown in the first row. The inner views of these regions are displayed in the middle row. The third row illustrates the views of the front and back regions of the brain. The subjects were randomly selected from the AD/MCI groups, and the figure displays the pathological locations identified by the model.

The overlaps among the pathological locations identified by the model suggest a correlation between AD and MCI. For example, in both tasks, the discriminating regions identified by the model contain the hippocampus, amygdala, and temporal lobe, which are highly correlated with memory. However, AD patients have more severe brain atrophy and wider regional distributions than MCI patients, which is consistent with medical observations.

7. Conclusions

In this study, the proposed MPS-FFA approach can be applied to AD-related tasks and provides interpretable feature visualizations. Notably, the MPS-FFA model includes all of its components into one module and achieves high performance. Finally, the experimental results with two independent datasets indicate that our method can accurately locate pathological locations.

Some areas are worth studying in future work. First, the feature extractors used in the encoder are relatively independent, so further integration is possible. The use of clinical cognitive data in this study effectively improves the ability of the network to extract discriminating features based on sMRI data. Thus, a joint analysis of multiple data types can be considered in future studies. Finally, since there is redundant information in sMRI data, introducing feature sparsity in the model may improve the computational performance.

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

This work were supported by the National Natural Science Foundation of China (Grant No. 61972001), the Natural Science Foundation of Anhui Province, China (Grant No. 1908085MF209), and the Natural

Science Foundation for the Higher Education Institutions of Anhui Province (Grant No. 2022AH050091).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health, United States Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, United States, the National Institute of Biomedical Imaging and Bioengineering, United States, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] J. Cummings, G. Lee, P. Nahed, M.E.Z.N. Kambar, K. Zhong, J. Fonseca, K. Taghva, Alzheimer's disease drug development pipeline: 2022, Alzheimer's Dementia Transl. Res. Clin. Interv. 8 (1) (2022) e12295, http://dx.doi.org/10.1002/trc2.12295.

[2] K.G. Yiannopoulou, S.G. Papageorgiou, Current and future treatments in Alzheimer disease: An update, J. Cent. Nerv. Syst. Dis. 12 (2020) 1179573520907397, http://dx.doi.org/10.1177/1179573520907397.

[3] C. Möller, H. Vrenken, L. Jiskoot, A. Versteeg, F. Barkhof, P. Scheltens, W.M. van der Flier, Different patterns of gray matter atrophy in early- and late-onset Alzheimer's disease, Neurobiol. Aging 34 (8) (2013) 2014–2022, http://dx.doi.org/10.1016/j.neurobiolaging.2013.02.013.

[4] F. Farina, D. Emek-Savaş, L. Rueda-Delgado, R. Boyle, H. Kiiski, G. Yener, R. Whelan, A comparison of resting state EEG and structural MRI for classifying Alzheimer's disease and mild cognitive impairment, NeuroImage 215 (2020) 116795, http://dx.doi.org/10.1016/j.neuroimage.2020.116795.

[5] R. Wurtman, Biomarkers in the diagnosis and management of Alzheimer's disease, Metabolism 64 (3) (2015) S47–S50, http://dx.doi.org/10.1016/j.metabol.2014.10.034.

[6] M. Ewers, R.A. Sperling, W.E. Klunk, M. Weiner, H. Hampel, Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia, Trends Neurosci. 34 (8) (2011) 430–442, http://dx.doi.org/10.1016/j.tins.2011.05.005.

[7] C. Salvatore, A. Cerasa, P. Battista, M.C. Gilardi, A. Quattrone, I. Castiglioni, Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach, Front. Neurosci. 9 (2015) 307, http://dx.doi.org/10.3389/fnins.2015.00307.

[8] A.M. Brickman, L.B. Zahodne, V.A. Guzman, A. Narkhede, I.B. Meier, E.Y. Griffith, F.A. Provenzano, N. Schupf, J.J. Manly, Y. Stern, J.A. Luchsinger, R. Mayeux, Reconsidering harbingers of dementia: Progression of parietal lobe white matter hyperintensities predicts Alzheimer's disease incidence, Neurobiol. Aging 36 (1) (2015) 27–32, http://dx.doi.org/10.1016/j.neurobiolaging.2014.07.019.

[9] J. Yang, P. Pan, W. Song, R. Huang, J. Li, K. Chen, Q. Gong, J. Zhong, H. Shi, H. Shang, Voxelwise meta-analysis of gray matter anomalies in Alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation, J. Neurol. Sci. 316 (1) (2012) 21–29, http://dx.doi.org/10.1016/j.jns.2012.02.010.

[10] J. Izzo, O.A. Andreassen, L.T. Westlye, D. van der Meer, The association between hippocampal subfield volumes in mild cognitive impairment and conversion to Alzheimer's disease, Brain Res. 1728 (2020) 146591, http://dx.doi.org/10.1016/j.brainres.2019.146591.

[11] H.T. Shen, X. Zhu, Z. Zhang, S.H. Wang, Y. Chen, X. Xu, J. Shao, Heterogeneous data fusion for predicting mild cognitive impairment conversion, Inform. Fusion 66 (2021) 54–63, http://dx.doi.org/10.1016/j.inffus.2020.08.023.

[12] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, NeuroImage 104 (2015) 398–412, http://dx.doi.org/10.1016/j.neuroimage.2014.10.002.

[13] Y. Ding, C. Luo, C. Li, T. Lan, Z. Qin, High-order correlation detecting in features for diagnosis of Alzheimer's disease and mild cognitive impairment, Biomed. Signal Proces. 53 (2019) 101564, http://dx.doi.org/10.1016/j.bspc.2019.101564.

[14] J. Liu, Y. Pan, F.X. Wu, J. Wang, Enhancing the feature representation of multi-modal MRI data by combining multi-view information for MCI classification, Neurocomputing 400 (2020) 322–332, http://dx.doi.org/10.1016/j.neucom.2020.03.006.

[15] I. Beheshti, H. Demirel, F. Farokhian, C. Yang, H. Matsuda, Structural MRI-based detection of Alzheimer's disease using feature ranking and classification error, Comput. Methods Programs Biomed. 137 (2016) 177–193, http://dx.doi.org/10.1016/j.cmpb.2016.09.019.

[16] O. Valenzuela, X. Jiang, A. Carrillo, I. Rojas, Multi-objective genetic algorithms to find most relevant volumes of the brain related to Alzheimer's disease and mild cognitive impairment, Int. J. Neural Syst. 28 (9) (2018) 1850022, http://dx.doi.org/10.1142/S0129065718500223.

[17] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, NeuroImage 208 (2020) 116459, http://dx.doi.org/10.1016/j.neuroimage.2019.116459.

[18] S.F. Eskildsen, P. Coupé, D.G. a Lorenzo, V. Fonov, J.C. Pruessner, D.L. Collins, Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning, NeuroImage 65 (2013) 511–521, http://dx.doi.org/10.1016/j.neuroimage.2012.09.058.

[19] S. Katabathula, Q. Wang, R. Xu, Predict alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations, Alzheimers Res. Theory 13 (1) (2021) 104, http://dx.doi.org/10.1186/s13195-021-00837-0.

[20] J. Lötjönen, R. Wolz, J. Koikkalainen, V. Julkunen, L. Thurfjell, R. Lundqvist, G. Waldemar, H. Soininen, D. Rueckert, Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease, NeuroImage 56 (1) (2011) 185–196, http://dx.doi.org/10.1016/j.neuroimage.2011.01.062.

[21] V. Gonuguntla, E. Yang, Y. Guan, B.B. Koo, J.H. Kim, Brain signatures based on structural MRI: Classification for MCI, PMCI, and AD, Human Brain Mapp. 43 (9) (2022) 2845–2860, http://dx.doi.org/10.1002/hbm.25820.

[22] M. Liu, J. Zhang, D. Nie, P.T. Yap, D. Shen, Anatomical landmark based deep feature representation for MR images in brain disease diagnosis, IEEE J. Biomed. Health 22 (5) (2018) 1476–1485, http://dx.doi.org/10.1109/JBHI.2018.2791863.

[23] F. Li, M. Liu, Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks, Comput. Med. Imag. Grap. 70 (2018) 101–110, http://dx.doi.org/10.1016/j.compmedimag.2018.09.009.

[24] S. Qiu, P.S. Joshi, M.I. Miller, C. Xue, X. Zhou, C. Karjadi, G.H. Chang, A.S. Joshi, B. Dwyer, S. Zhu, M. Kaku, Y. Zhou, Y.J. Alderazi, A. Swaminathan, S. Kedar, M.-H. Saint Hilaire, S.H. Auerbach, J. Yuan, E.A. Sartor, R. Au, V.B. Kolachalama, Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, Brain 143 (6) (2020) 1920–1933, http://dx.doi.org/10.1093/brain/awaa137.

[25] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, The Alzheimer's Disease Neuroimaging Initiative, Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment, Front. Neurosci. 12 (NOV) (2018) 777, http://dx.doi.org/10.3389/fnins.2018.00777.

[26] H.I. Suk, S.W. Lee, D. Shen, Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, NeuroImage 101 (2014) 569–582, http://dx.doi.org/10.1016/j.neuroimage.2014.06.077.

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

[28] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, X. Song, Early detection of Alzheimer's disease using magnetic resonance imaging: A novel approach combining convolutional neural networks and ensemble learning, Front. Neurosci. 14 (2020) 259, http://dx.doi.org/10.3389/fnins.2020.00259.

[29] L. Nanni, S. Brahnam, C. Salvatore, I. Castiglioni, Texture descriptors and voxels for the early diagnosis of Alzheimer's disease, Artif. Intell. Med. 97 (2019) 19–26, http://dx.doi.org/10.1016/j.artmed.2019.05.003.

[30] R. Mendoza-Léon, J. Puentes, L.F. Uriza, M. Hernández Hoyos, Single-slice Alzheimer's disease classification and disease regional analysis with supervised switching autoencoders, Comput. Biol. Med. 116 (2020) 103527, http://dx.doi.org/10.1016/j.compbiomed.2019.103527.

[31] F. Ren, C. Yang, Q. Qiu, N. Zeng, C. Cai, C. Hou, Q. Zou, Exploiting discriminative regions of brain slices based on 2D CNNs for Alzheimer's disease classification, IEEE Access 7 (2019) 181423–181433, http://dx.doi.org/10.1109/ACCESS.2019.2920241.

[32] H.W. Kim, H.E. Lee, S. Lee, K.T. Oh, M. Yun, S.K. Yoo, Slice-selective learning for Alzheimer's disease classification using a generative adversarial network: a feasibility study of external validation, Eur. J. Nucl. Med. Mol. Imaging. 47 (9) (2020) 2197–2206, http://dx.doi.org/10.1007/s00259-019-04676-y.

[33] M. Dyrba, M. Hanzig, S. Altenstein, S. Bader, T. Ballarini, F. Brosseron, K. Buerger, D. Cantré, P. Dechent, L. Dobisch, E. Duzel, M. Ewers, K. Fliessbach, W. Glanz, J.D. Haynes, M. Heneka, D. Janowitz, D. Keles, I. Kilimann, S. Teipel, Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease, Alzheimers Res. Theory 13 (1) (2021) 191, http://dx.doi.org/10.1186/s13195-021-00924-2.

[34] S. Dwivedi, T. Goel, R. Sharma, R. Murugan, Structural MRI based Alzheimer's disease prognosis using 3D convolutional neural network and support vector machine, in: 2021 Advanced Communication Technologies and Signal Processing, ACTS, 2021, pp. 1–4, http://dx.doi.org/10.1109/ACTS53447.2021.9708107.

[35] P. Lu, L. Hu, N. Zhang, H. Liang, T. Tian, L. Lu, A two-stage model for predicting mild cognitive impairment to Alzheimer's disease conversion, Front. Aging Neurosci. 14 (2022) 826622, http://dx.doi.org/10.3389/fnagi.2022.826622.

[36] J. Li, Y. Wei, C. Wang, Q. Hu, Y. Liu, L. Xu, 3-D CNN-based multichannel contrastive learning for Alzheimer's disease automatic diagnosis, IEEE Trans. Instrum. Meas. 71 (2022) 1–11, http://dx.doi.org/10.1109/TIM.2022.3162265.

[37] S.H. Wang, Q. Zhou, M. Yang, Y.D. Zhang, ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation, Front. Aging Neurosci. 13 (2021) 687456, http://dx.doi.org/10.3389/fnagi.2021.687456.

[38] X. Zhang, L. Han, W. Zhu, L. Sun, D. 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 PP (2021) 1, http://dx.doi.org/10.1109/JBHI.2021.3066832.

[39] Y. Zhang, Q. Teng, Y. Liu, Y. Liu, X. He, Diagnosis of Alzheimer's disease based on regional attention with sMRI gray matter slices, J. Neurosci. Methods 365 (2022) 109376, http://dx.doi.org/10.1016/j.jneumeth.2021.109376.

[40] X. Gao, F. Shi, D. Shen, M. Liu, Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease, IEEE J. Biomed. Health 26 (1) (2022) 36–43, http://dx.doi.org/10.1109/JBHI.2021.3097721.

[41] Z. Zhang, L. Gao, G. Jin, L. Guo, Y. Yao, L. Dong, J. Han, the Alzheimer's Disease NeuroImaging Initiative, THAN: Task-driven hierarchical attention network for the diagnosis of mild cognitive impairment and Alzheimer's disease, Quant. Imag. Med. Surg. 11 (7) (2021) 3338–3354, http://dx.doi.org/10.21037/qims-21-91.

[42] K. Han, J. Luo, Q. Xiao, Z. Ning, Y. Zhang, Light-weight cross-view hierarchical fusion network for joint localization and identification in Alzheimer's disease with adaptive instance-declined pruning, Phys. Med. Biol. 66 (8) (2021) 085013, http://dx.doi.org/10.1088/1361-6560/abf200.

[43] H. Guan, Y. Liu, E. Yang, P.T. Yap, D. Shen, M. Liu, Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification, Med. Image Anal. 71 (2021) 102076, http://dx.doi.org/10.1016/j.media.2021.102076.

[44] C. Lian, M. Liu, L. Wang, D. Shen, Multi-task weakly-supervised attention network for dementia status estimation with structural MRI, IEEE Trans. Neur. Netw. Learn. 33 (8) (2021) 4056–4068, http://dx.doi.org/10.1109/TNNLS.2021.3055772.

[45] H. Guan, C. Wang, J. Cheng, J. Jing, T. Liu, A parallel attention-augmented bilinear network for early magnetic resonance imaging-based diagnosis of Alzheimer's disease, Human Brain Mapp. 43 (2) (2022) 760–772, http://dx.doi.org/10.1002/hbm.25685.

[46] C. Lian, M. Liu, Y. Pan, D. Shen, Attention-guided hybrid network for dementia diagnosis with structural MR images, IEEE Trans. Cybernetics 52 (4) (2022) 1992–2003, http://dx.doi.org/10.1109/TCYB.2020.3005859.

[47] X. Pan, T.L. Phan, M. Adel, C. Fossati, T. Gaidon, J. Wojak, E. Guedj, Multi-view separable pyramid network for AD prediction at MCI stage by 18F-FDG brain PET imaging, IEEE Trans. Med. Imaging 40 (1) (2021) 81–92, http://dx.doi.org/10.1109/TMI.2020.3022591.

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

[49] A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review, Comput. Methods Programs Biomed. 187 (2020) 105242, http://dx.doi.org/10.1016/j.cmpb.2019.105242.

[50] L. Nanni, M. Interlenghi, S. Brahnam, C. Salvatore, S. Papa, R. Nemni, I. Castiglioni, The Alzheimer's Disease Neuroimaging Initiative, Comparison of transfer learning and conventional machine learning applied to structural brain MRI for the early diagnosis and prognosis of Alzheimer's disease, Front. Neurol. 11 (2020) 576194, http://dx.doi.org/10.3389/fneur.2020.576194.

[51] G. Battineni, N. Chintalapudi, F. Amenta, E. Traini, Deep learning type convolution neural network architecture for multiclass classification of Alzheimer's disease, in: BIOIMAGING 2021 - 8th International Conference on Bioimaging, 2021, pp. 209–215, http://dx.doi.org/10.5220/0010378602090215.

[52] E.E. Bron, S. Klein, J.M. Papma, L.C. Jiskoot, V. Venkatraghavan, J. Linders, P. Aalten, P.P. De Deyn, G.J. Biessels, J.A. Claassen, H.A. Middelkoop, M. Smits, W.J. Niessen, J.C. van Swieten, W.M. van der Flier, I.H. Ramakers, A. van der Lugt, Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease, NeuroImage: Clinical 31 (2021) 102712, http://dx.doi.org/10.1016/j.nicl.2021.102712.

[53] R. Karthik, R. Menaka, M. Hariharan, D. Won, Ischemic lesion segmentation using ensemble of multi-scale region aligned CNN, Comput. Methods Programs Biomed. 200 (2021) 105831, http://dx.doi.org/10.1016/j.cmpb.2020.105831.

[54] L. Wang, J. Shen, E. Tang, S. Zheng, L. Xu, Multi-scale attention network for image super-resolution, J. Vis. Commun. Image Represent. 80 (2021) 103300, http://dx.doi.org/10.1016/j.jvcir.2021.103300.

[55] W. Yu, D. Pi, L. Xie, Y. Luo, Multiscale attentional residual neural network framework for remaining useful life prediction of bearings, Measurement 177 (2021) 109310, http://dx.doi.org/10.1016/j.measurement.2021.109310.

[56] R. Karthik, T.T. George, Y. Shah, P. Sasidhar, A novel multi-feature fusion method for classification of gastrointestinal diseases using endoscopy images, Diagnostics 12 (2022) 10, http://dx.doi.org/10.3390/diagnostics12102316.

[57] R. Karthik, R. Menaka, M. Siddharth, Classification of breast cancer from histopathology images using an ensemble of deep multiscale networks, Biocybern. Biomed. Eng. 42 (2022) 963–976, http://dx.doi.org/10.1016/j.bbe.2022.07.006.

[58] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, NeuroImage 15 (1) (2002) 273–289, http://dx.doi.org/10.1006/nimg.2001.0978.

[59] C. Lian, M. Liu, J. Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 880–893, http://dx.doi.org/10.1109/TPAMI.2018.2889096.

[60] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, X. Zhao, Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease, Neurocomputing 333 (2019) 145–156, http://dx.doi.org/10.1016/j.neucom.2018.12.018.

[61] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 936–944, http://dx.doi.org/10.1109/CVPR.2017.106.

[62] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, http://dx.doi.org/10.48550/arXiv.1706.05587, arXiv.

[63] I. Jang, B. Li, J.M. Riphagen, B.C. Dickerson, D.H. Salat, Multiscale structural mapping of Alzheimer's disease neurodegeneration, NeuroImage Clinical 33 (2022) 102948, http://dx.doi.org/10.1016/j.nicl.2022.102948.

[64] X. Pan, M. Adel, C. Fossati, T. Gaidon, J. Wojak, E. Guedj, Multiscale spatial gradient features for 18F-FDG PET image-guided diagnosis of Alzheimer's disease, Comput. Methods Programs Biomed. 180 (2019) 105027, http://dx.doi.org/10.1016/j.cmpb.2019.105027.

[65] K. Hu, Y. Wang, K. Chen, L. Hou, X. Zhang, Multi-scale features extraction from baseline structure MRI for MCI patient classification and AD early diagnosis, Neurocomputing 175 (2015) 132–145, http://dx.doi.org/10.1016/j.neucom.2015.10.043.

[66] H.D. Li, R. Guo, J. Li, J. Wang, Y. Pan, J. Liu, Joint learning of primary and secondary labels based on multi-scale representation for Alzheimer's disease diagnosis, in: Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, 2020, pp. 637–642, http://dx.doi.org/10.1109/BIBM49941.2020.9313422.

[67] C. Ge, Q. Qu, I.Y.H. Gu, A.S. Jakola, Multi-stream multi-scale deep convolutional networks for Alzheimer's disease detection using MR images, Neurocomputing 350 (2019) 60–69, http://dx.doi.org/10.1016/j.neucom.2019.04.023.

[68] E. Lee, J.S. Choi, M. Kim, H.I. Suk, Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning, NeuroImage 202 (2019) 116113, http://dx.doi.org/10.1016/j.neuroimage.2019.116113.

[69] W. Yan, V. Calhoun, M. Song, Y. Cui, H. Yan, S. Liu, L. Fan, N. Zuo, Z. Yang, K. Xu, J. Yan, L. Lv, J. Chen, Y. Chen, H. Guo, P. Li, L. Lu, P. Wan, H. Wang, H. Wang, Y. Yang, H. Zhang, D. Zhang, T. Jiang, J. Sui, Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data, EBioMedicine 47 (2019) 543–552, http://dx.doi.org/10.1016/j.ebiom.2019.08.023.

[70] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease, Med. Image Anal. 46 (2018) 26–34, http://dx.doi.org/10.1016/j.media.2018.02.002.

[71] K. Hett, V.T. Ta, I. Oguz, J.V. Manjón, P. Coupé, Multi-scale graph-based grading for Alzheimer's disease prediction, Med. Image Anal. 67 (2021) 101850, http://dx.doi.org/10.1016/j.media.2020.101850.

[72] D. Hong, C. Huang, C. Yang, J. Li, Y. Qian, C. Cai, FFA-DMRI: A network based on feature fusion and attention mechanism for brain MRI denoising, Front. Neurosci. 14 (2020) 577937, http://dx.doi.org/10.3389/fnins.2020.577937.

[73] H. Deng, Y. Zhang, R. Li, C. Hu, Z. Feng, H. Li, Combining residual attention mechanisms and generative adversarial networks for hippocampus segmentation, Tsinghua Sci. Technol. 27 (1) (2022) 68–78, http://dx.doi.org/10.26599/TST.2020.9010056.

[74] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: Computer Vision, ECCV 2018, 2018, pp. 3–19, http://dx.doi.org/10.1007/978-3-030-01234-2_1.

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

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

[77] M. Sun, T.X. Han, M.C. Liu, A. Khodayari-Rostamabad, Multiple instance learning convolutional neural networks for object recognition, in: Proceedings International Conference on Pattern Recognition, ICPR 2016, 2016, pp. 3270–3275, http://dx.doi.org/10.1109/ICPR.2016.7900139.

[78] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Artificial Intelligence 201 (2013) 81–105, http://dx.doi.org/10.1016/j.artint.2013.06.003.

[79] M. Xia, J. Wang, Y. He, BrainNet viewer: A network visualization tool for human brain connectomics, PLoS One 8 (7) (2013) e68910, http://dx.doi.org/10.1371/journal.pone.0068910.

[80] J. Fleming Beattie, R.C. Martin, R.K. Kana, H. Deshpande, S. Lee, J. Curé, L. Ver Hoef, Hippocampal dentation: Structural variation and its association with episodic memory in healthy adults, Neuropsychologia 101 (2017) 65–75, http://dx.doi.org/10.1016/j.neuropsychologia.2017.04.036.