Pattern Recognition 152 (2024) 110450

Contents lists available at ScienceDirect

Pattern Recognition

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

# Interpretable medical deep framework by logits-constraint attention guiding graph-based multi-scale fusion for Alzheimer's disease analysis

Jinghao Xu [a], Chenxi Yuan [b], Xiaochuan Ma [a], Huifang Shang [c,*], Xiaoshuang Shi [a], Xiaofeng Zhu [a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[b] Department of Biostatistics, Epidemiology and Informatics Perelman School of Medicine, University of Pennsylvania, United States of America
[c] Department of Neurology, Laboratory of Neurodegenerative Disorders, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, Sichuan, China

ARTICLE INFO

*Keywords:*
Alzheimer's disease
Attention
Graph neural networks
Multi-scale feature fusion
Structural MRI

ABSTRACT

Deep learning using structural MRI has been widely applied to early diagnosis study of Alzheimer's disease. Among existing methods, attention-based 3D subject-level methods can not only provide diagnosis results but also interpret the significant brain regions, thereby attracting considerable attention. However, the performance of previous attention-based methods might be still restricted by: (i) the gap between attention scores and semantic significant regions; (ii) using only single-scale features or simply fusing multi-scale information by addition or concatenation for classification decision-making. To overcome these two issues, we propose an innovative dual-branch model called LA-GMF, which consists of two major modules: logits-constraint attention (LA) and graph-based multi-scale fusion (GMF). The LA module is designed to guide the model to focus on key areas to enhance the diagnostic performance of local lesions, by reducing the inconsistency between attention scores and class prediction probabilities. Meanwhile, by combining the graph neural network and the self-attention mechanism, the GMF module not only introduces the interaction between patches, but also explores the correlation and complementarity between features at different scales, thereby extracting feature representations more comprehensively. Experiments on the popular ADNI and AIBL datasets validate the potential of our model in boosting early AD diagnosis accuracy. Additionally, our interpretation experiments demonstrate the superior interpretability performance of the proposed method over recent state-of-the-art attention-based methods. Our source codes are released at: https://github.com/nollexu/LA-GMF.

## 1. Introduction

Alzheimer's disease (AD) is one common neurodegenerative disease characterized by progressive cognitive decline, memory loss, and impairment of daily life functions [1]. Currently, because there is no fundamental medical treatment to cure AD, early diagnosis and intervention are crucial to alleviate symptoms, delay disease progression, and improve life quality of patients. Structural magnetic resonance imaging (sMRI), which is a non-invasive neuroimaging technique and can generate 3-dimensional (3D) images to provide detailed information about brain structure and morphology, is one very important tool for early screening of AD [2]. However, manually examining sMRI is laborious, time consuming and even error prone. Hence, various computer-aided diagnostic methods have been proposed for early AD diagnosis, so as to reduce the workload of neurologists and boost their diagnosis accuracy and efficiency.

Traditional machine learning methods for AD diagnosis can be roughly divided into two categories: (i) Voxel-based methods [3,4], which utilize high-dimensional vectors as features and usually contain a large amount of redundancy and noise, resulting in high computational cost and poor diagnostic performance, and (ii) ROI-based methods [5,6], which manually extract features, such as volume, shape, and cortical thickness, from brain regions, thereby easily leading to information loss, i.e., rough and difficult to reflect the small changes related to brain diseases. Additionally, this manual feature extraction process is often subjective and time-consuming.

In recent years, deep learning methods have been widely used in computer-aid diagnosis (CAD), because of their powerful capability on automatically learning and extracting effective features from data. According to the input type of the network, existing deep learning models for AD diagnosis based on sMRI can be roughly divided into four categories [7]: 2D slice-level, ROI-based, 3D patch-level, and 3D subject-level. Specifically, 2D slice-level methods extract a set of 2D slices from 3D sMRI as the model input according to some general or

* Corresponding author.
*E-mail addresses:* hfshang2002@126.com (H. Shang), xsshi2013@gmail.com (X. Shi).

https://doi.org/10.1016/j.patcog.2024.110450
Received 31 October 2023; Received in revised form 7 March 2024; Accepted 20 March 2024
Available online 22 March 2024

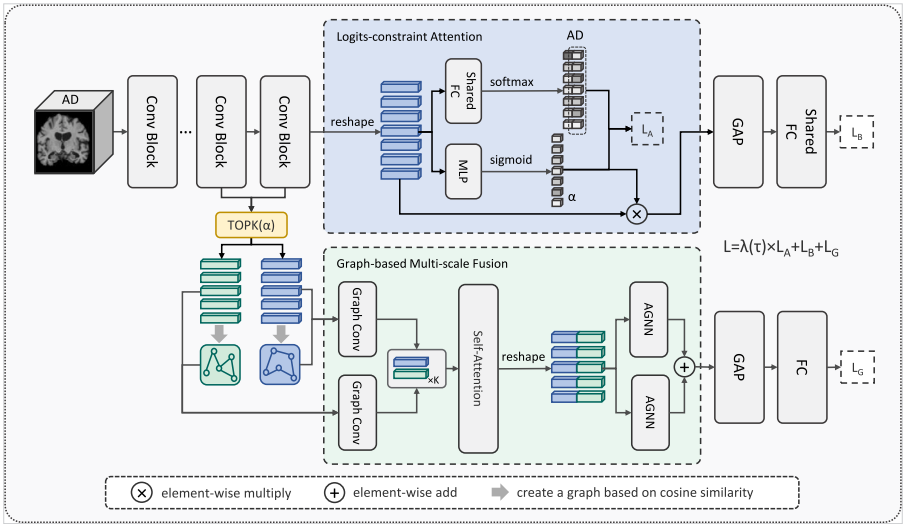


**Fig. 1.** The overall architecture of the proposed LA-GMF. It mainly consists of two modules, LA and GMF. $\alpha$ represents the attention score vector, and TOPK($\alpha$) represents the selection of $K$ important patches from different feature maps based on attention scores. $L_A$ represents the attention loss. $L_B$ and $L_G$ are the cross-entropy losses in the backbone network and graph branch network, respectively. AGNN, Attention-based Graph Neural Network, is a graph neural network architecture proposed in [25].

customized strategies (e.g., using all slices in a certain direction [8], or selecting the most informative slices based on image entropy [9]). These methods have fewer network parameters and can easily borrow existing successful models through transfer learning. However, analyzing 3D sMRI in a 2D manner can easily lose spatial information. ROI-based methods [10,11] usually require to segment disease-related regions under the guidance of prior medical knowledge, and then feed the segmented regions into the model to extract features and classify diseases. However, disease-related changes often span multiple brain regions, and these methods might not fully cover the relevant disease areas. 3D patch-level methods [12–16] usually utilize a data-driven method or based on medical prior knowledge to select local patches with large information content or related to diseases as model inputs. However, there methods usually divide patch localization and classification into two independent stages, and neglect the patch interactions, thereby often leading to the sub-optimal learning performance. 3D subject-level methods [17–19] adopt the entire sMRI as the model input and are able to utilize complete spatial information, but they are usually difficult to capture the local nature of AD pathology, thereby possibly causing poor diagnosis performance.

In order to overcome aforementioned limitations, various attention-based 3D subject-level methods [20–24] have been proposed to enable the model to focus more on local positions while maintaining a global perspective. These methods have achieved some significant improvements, but still face some challenges and drawbacks. Firstly, the number of voxels in the entire sMRI is usually large, while the number of subjects related to AD is relatively small. This makes it insufficient to guide the mining of discriminative regions solely based on the classification loss. Secondly, attention mechanisms do not always accurately reflect actual semantic information, i.e., some semantically important regions often obtain small the attention weights, this is the notorious attention-semantic gap, which often reduces the model interpretability and representation learning capability. Thirdly, most existing methods are based on single-scale features for decision-making, or rely solely on relatively basic methods (like addition or concatenation) for multi-scale information fusion, which might not be able to effectively capture the relevant and complementary information between scales, and thus possibly restricting the model capability.

Based on above observations, we propose an innovative dual-branch framework, namely LA-GMF, which mainly consists of two modules: logits-constrained attention (LA) and graph-based multi-scale fusion (GMF). For clarity, we show its overall architecture in Fig. 1. The proposed framework directly utilizes the whole 3D sMRI as the input to explore the complete spatial information, the LA module to mine discriminative patches for better capturing local pathology, and the GMF module to leverage multi-scale patches for learning more powerful feature representations.

In summary, our main contributions are listed as follows:

1. We propose a novel interpretable dual-branch framework, which can simultaneously mine significant patches from the whole 3D sMRI and classify it for AD diagnosis in an end-to-end manner.
2. We propose a new attention mechanism with an additional loss, which connects attention scores of patches with their corresponding class prediction probabilities, to reduce the attention-semantic gap.
3. We design a novel graph-based multi-scale fusion module, which employs the graph neural network and self-attention to extract multi-scale features of key patches mined by the LA module, to enhance the patch interaction and effectively fuse cross-scale information for boosting the model diagnosis performance.
4. Extensive experimental results demonstrate that the proposed framework outperforms recent state-of-the-art methods on AD diagnosis, with better interpretation capability.

The rest of this paper is organized as follows. Section 2 briefly reviews some related sMRI-based methods for AD diagnosis. Section 3 presents the proposed framework. Section 4 introduces the datasets and preprocessing methods used in our experiments, shows and analyzes experimental results. Finally, Section 5 concludes this paper and points out the future work.

## 2. Related work

Based on the input type, the proposed framework is one 3D subject-level method, however, it is also inspired by the idea of 3D patch-level methods. Hence, in this section, we briefly review some popular 3D patch-level and 3D subject-level methods for AD diagnosis.

### 2.1. 3D patch-level methods for AD diagnosis

3D patch-level methods extract patches from sMRI as the model input. They regard that brain atrophy usually occurs locally, i.e., only a few areas in sMRI are highly correlated with pathological features. For instances, Liu et al. [13] propose a multi-model deep learning

framework, which can simultaneously perform automatic hippocampal segmentation and AD classification, by using a patch containing the hippocampus as the input. Liu et al. [12] construct a CNN-based multiple instance learning (MIL) model, which utilizes pre-identified local image patches based on anatomical landmarks as the input, for AD classification and mild cognitive impairment (MCI) conversion prediction. Zhou et al. [14] design a dual attention multi-instance architecture, which can not only capture the relative importance of each patch, but also recognize the discriminative features within the patch, thereby obtaining better classification performance. Wang et al. [16] propose an innovative patch-based MIL framework for AD diagnosis. They employ the Relief algorithm [26] to select patches as the model input, and make the network capture local and global information by integrating inter-patch local attention blocks and outer-patch global attention blocks.

These methods demonstrate that mining the significant 3D patches can boost model classification performance. However, they usually fail to consider patch interactions and require to initialize patch extraction locations based on prior knowledge or through an independent localization stage, thereby restricting their model performance. In contrast to the 3D patch-level methods, the proposed framework abolishes the task-independent patch localization phase and introduce interactions between patches, so as to facilitate the feature representation learning.

### 2.2. 3D subject-level methods for AD diagnosis

3D subject-level methods utilize the entire 3D sMRI image as the model input, so that they can fully integrate and utilize spatial information. For example, Korolev et al. [17] adopt VGG- and ResNet-like 3D networks for AD classification. This is the first study to diagnose AD based on the whole brain sMRI. Fan et al. [18] introduce a UNet style model for AD diagnosis tasks, and experimentally demonstrate that skipping connections and deep supervision can achieve better classification model performance. Li et al. [19] propose a multi-channel contrastive learning strategy for AD diagnosis, which further improves the model classification accuracy and generalization ability by combining supervised classification loss with unsupervised contrastive loss.

Above methods can achieve promising performance on multiple datasets, however, features from disease-related regions often play an important role in disease identification. Therefore, various attention-based 3D subject-level methods have been proposed to enable models to focus more on local positions while maintaining a global perspective. For example, Jin et al. [27] insert a sparse attention module consisting of a convolutional layer and a rectified linear unit (ReLU) layer into 3D ResNet for AD and normal control (NC) recognition. Zhang et al. [20] propose ResAttNet based on residual connections and self-attention mechanisms, so as to simultaneously capture local and global information of sMRI for boosting diagnostic performance. Wu et al. [22] capture possible brain atrophy by adding attention modules after custom multi-scale ensemble blocks. Pei et al. [23] employ images from multiple scales as model inputs, they combine attention mechanisms and custom global context blocks to fuse features of different scales for the classification of AD and MCI. Zhang et al. [24] propose a UNet framework for progressive MCI (pMCI) and stable MCI (sMCI) recognition, by utilizing fine-grained spatial attention maps to highlight disease-related semantic features, and coarse-grained semantic attention maps to emphasize disease-related detailed features, so as to achieve disease diagnosis through hierarchical fusion of multi-scale features.

Relative to 3D subject-level methods, our framework using the attention mechanism focuses more on local feature details and enhances model interpretability. Compared with attention-based 3D subject-level methods, we introduce an additional loss function to narrow the gap between attention scores and semantically salient regions. Moreover, we further explore the correlation and complementarity between multi-scale features based on the graph neural network and self-attention mechanism.

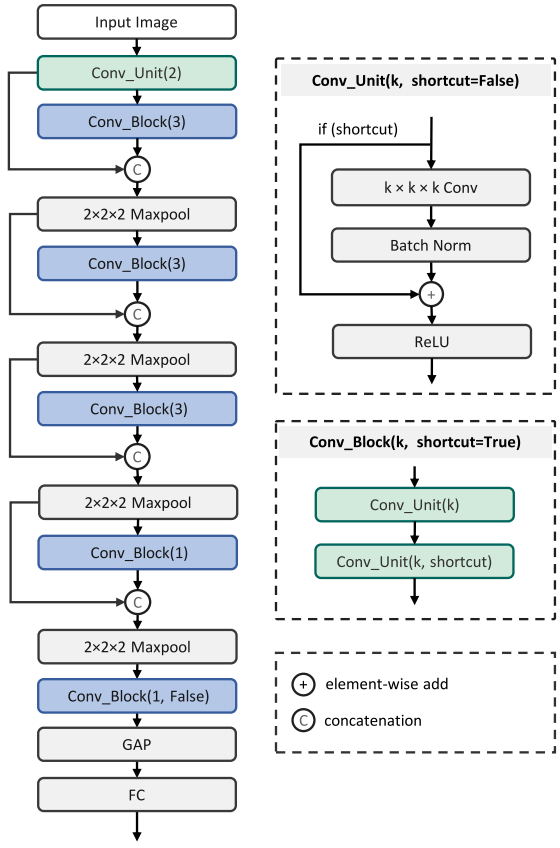


**Fig. 2.** The structure of the backbone network, which consists of 5 stages, with each one containing a Conv_Block. The number of output channels in each Conv_Block is 16, 32, 64, 128, and 256, respectively.

## 3. Method

In this section, we introduce three major components in the proposed framework, including the backbone network, LA and GMF modules.

### 3.1. Backbone network

Because AD datasets usually contain a small number of subjects, i.e., hundreds of individuals, we design a relatively small-sized 3D convolutional neural network as our backbone, which consists of 5 stages, including 11 convolutional layers, 4 max-pooling layers, a global average pooling (GAP) layer, and a classification layer. For clarity, we present the architecture of our backbone network in Fig. 2.

In Fig. 2, The Conv_Unit(k, shortcut) represents a convolutional unit that consists of a 3D convolutional layer, a Batch Normalization layer, and a ReLU activation function. By default, the shortcut parameter is set to False, i.e. when the shortcut parameter is not specified, there is no residual connection added. The parameter $k$ specifies the size of the convolutional kernel. Conv_Block denotes a convolutional block consisting of two Conv_Unit instances. Specifically, Conv_Unit(2) used in the first layer of the network, which has an output channel number of 16, stride of 2, and padding of 0, implements the downsampling and channel expansion operations. In Conv_Block(k), the convolution operation does not change the shape of the input, i.e., if $k = 3$, the padding of the convolution operation is set to 1 and stride is set to 1. If $k = 1$, no padding is added, and stride is set to 1. $2 \times 2 \times 2$ Maxpool represents a max-pooling operation with a kernel size of $2 \times 2 \times 2$, a stride of 2, and zero padding.

### 3.2. Logits-constraint attention

Given the feature map $\mathbf{F} \in \mathbb{R}^{d\times\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}}$ obtained from the last convolutional layer, with a downsampling factor of 32, where $d$ is the number of channels in the feature map, $H$, $W$ and $D$ represent the height, width, and depth of the input sMRI, respectively. Using a reshape operation, we transform it into a feature matrix $\mathbf{T} = [\mathbf{t}_0, \mathbf{t}_1, \ldots, \mathbf{t}_N] \in \mathbb{R}^{N\times d}$ ($N = \frac{H}{32} \times \frac{W}{32} \times \frac{D}{32}$ denotes the number of 3D patches), where $\mathbf{t}_i \in \mathbb{R}^d$ denotes the feature embedding of the $i$th 3D patch. Subsequently, we apply a two-layer MLP combined with the sigmoid function to calculate attention scores, as illustrated below:

$$\alpha_i = \frac{1}{1+\exp\{-\mathbf{w}^\top \mathrm{relu}(\mathbf{\Phi}\mathbf{t}_i^\top)\}}, \tag{1}$$

where $\alpha_i \in (0,1)$ represents the attention score of the $i$th patch embedding, $\mathbf{w} \in \mathbb{R}^{l\times 1}$ and $\mathbf{\Phi} \in \mathbb{R}^{l\times d}$ represent the parameters of the two fully connected layers, with $l$ being the dimension of the hidden layer.

After obtaining the attention score $\alpha$, we utilize it to weight the embedding vectors, and then after GAP operation we feed it into the classification layer to obtain the corresponding logits vector for the entire image. They can be described as:

$$\mathbf{t}_i \leftarrow \alpha_i \mathbf{t}_i, \tag{2}$$

$$\mathbf{z}^I = \frac{1}{N}\sum_{i=1}^{N} \mathbf{t}_i \theta, \tag{3}$$

where $\mathbf{z}^I \in \mathbb{R}^{1\times C}$ represents the image-level logits vector in the main branch, $C$ represents the total number of classes, $\theta \in \mathbb{R}^{d\times C}$ denotes the parameters of the classification layer. By sharing the classification layer, we can obtain the logits vector for each patch embedding in $\mathbf{T}$ as follows:

$$\mathbf{z}_i = \mathbf{t}_i \theta, \tag{4}$$

where $\mathbf{z}_i \in \mathbb{R}^{1\times C}$ represents the logits vector corresponding to the $i$th patch embedding.

Finally, we adopt the $\ell_2$-norm to calculate the distance between probabilistic $\mathbf{z}_{i,c}$ and $\alpha_i$ for the $i$th patch embedding, and employ it as an additional loss to supervise model training to reduce the gap between the attention score and semantic classification. The loss is:

$$L_A = \sqrt{\sum_{i=1}^{N}\left(\frac{\exp(\mathbf{z}_{i,c})}{\sum_{j=1}^{C}\exp(\mathbf{z}_{i,j})} - \alpha_i\right)^2}, \tag{5}$$

where $c$ denotes the true class of the corresponding image. More specifically, Eq. (5) aims to make the patches with high class prediction probabilities have large attention scores, thereby reducing the inconsistency between attention scores and semantic significant regions.

Considering the challenge of directly capturing local and subtle structural abnormalities from the whole brain sMRI, the attention map generated by Eq. (1) can provide key guidance for identifying and localizing subject specific whole brain atrophy caused by AD.

### 3.3. Graph-based multi-scale fusion module

3D patch-level methods have proven that mining the significant 3D patches can effectively improve the model classification performance. Inspired by them, we hope to employ attention mechanisms to identify important patches from the entire sMRI data and further classify them in an end-to-end manner. However, due to the use of CNN as the backbone network, the extracted patches usually only involve local information, ignoring the information exchange between patches, which might become a bottleneck for improving the model performance. Fortunately, as a powerful and flexible modeling framework, Graph Neural Network (GNN) can explore the relationships between patches and enhance their interaction. Additionally, compared with high-level feature maps that capture abstract and semantic information, low-level feature maps play an irreplaceable role in supplementing detailed information, such as textures and edges. In order to obtain more expressive feature representations, it is crucial to exploit both low- and high-level features. However, existing multi-scale methods usually only fuse information from different scales through simple concatenation or addition, which might ignore the correlation and complementarity between scales.

Based on the above analysis, we design a graph-based multiscale fusion (GMF) module. Specifically, we first extract key patches in feature maps of different scales based on the attention map as the input of the graph branch. Next, we construct multiple graphs based on the feature matrices of these key patches and utilize graph convolutional networks (GCN) [28] to model the interactions between features within scales. In addition, we also introduce a self-attention mechanism to capture the correlation between different scales. In order to achieve a more comprehensive and effective fusion, we further utilize AGNN to mine the complementary relationship between scales based on the scale-specific graph structure and the similarity between the fused features. We show the detailed introduction of the GMF module in the following.

Let $\{\mathbf{T}^{(m)}\}_{m=1}^{M}$ represent the inputs of GMF module, where $\mathbf{T}^{(m)} \in \mathbb{R}^{K\times d}$ denotes the feature matrix corresponding to the $m$th scale, $M$ represents the number of scales, $K$ is the number of patches with the highest attention scores that we selected based on Eq. (1). We initially employ cosine similarity along with $M$ thresholds $\{e_m\}_{m=1}^{M}$ to construct $M$ adjacency matrices $\{\mathbf{A}^{(m)}\}_{m=1}^{M}$, where $\mathbf{A}^{(m)} \in \mathbb{R}^{K\times K}$ denotes the graph structure of the $m$th graph, and it is obtained by:

$$\mathbf{A}_{i,j}^{(m)} = \begin{cases} 1 & \text{if } \cos(\mathbf{t}_{\mathrm{i}}^{(\mathrm{m})}, \mathbf{t}_{\mathrm{j}}^{(\mathrm{m})}) >= e_m, \\ 0 & \text{otherwise;} \end{cases} \tag{6}$$

where $\mathbf{t}_i^{(m)} \in \mathbb{R}^d$ and $\mathbf{t}_j^{(m)} \in \mathbb{R}^d$ represent the $i$th and $j$th rows in the feature matrix $\mathbf{T}^{(m)}$, respectively. Specifically, we fused features from two scales to make the final decision. We extract the feature vectors corresponding to $K$ voxel positions from the feature map output from the last convolutional block, forming the feature matrix $\mathbf{T}^{(1)}$. Meanwhile, in the feature map output from the previous stage of convolutional blocks, we treat the $2\times2\times2$ blocks as a whole, and then flatten them, next, reduce the dimensionality through a linear layer with an input dimension of 8 and an output dimension of 2 to obtain the feature matrix $\mathbf{T}^{(2)}$, which has the same size as $\mathbf{T}^{(1)}$.

Then, we employ the graph convolutional layer $\mathrm{g}^{(\mathrm{m})}(\cdot)$ to generate node representations $\mathbf{H}^{(m)} \in \mathbb{R}^{K\times d}$ based on the node features $\mathbf{T}^{(m)}$ and the graph structure $\mathbf{A}^{(m)}$ of each graph, i.e.,

$$\mathbf{H}^{(m)} = \mathrm{g}^{(\mathrm{m})}(\mathbf{T}^{(\mathrm{m})}, \mathbf{A}^{(\mathrm{m})}) = \sigma\left(\hat{\mathbf{D}}^{(m)-\frac{1}{2}}\hat{\mathbf{A}}^{(\mathrm{m})}\hat{\mathbf{D}}^{(m)-\frac{1}{2}}\mathbf{T}^{(\mathrm{m})}\Theta^{(\mathrm{m})}\right), \tag{7}$$

where $\hat{\mathbf{A}}^{(m)} = \mathbf{A}^{(m)} + \mathbf{I}_K$ denotes the adjacency matrix with self-loops, $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ is an identity matrix. Since our $\mathbf{A}^{(m)}$ is generated using cosine similarity, which already includes self-loops, $\hat{\mathbf{A}}^{(m)}$ is equal to $\mathbf{A}^{(m)}$. $\hat{\mathbf{D}}^{(m)}$ is a diagonal degree matrix, e.g., $\hat{\mathbf{D}}_{ii}^{(m)} = \sum_{j=1}^{K}\hat{\mathbf{A}}_{ij}^{(m)}$, $\sigma(\cdot)$ represents an activation function, and $\Theta^{(m)} \in \mathbb{R}^{d\times d}$ is the weight matrix of $\mathrm{g}^{(\mathrm{m})}(\cdot)$.

We merge $\{\mathbf{H}^{(m)}\}_{m=1}^{M}$ and convert them into a 3D matrix $\mathbf{J} = [\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_K] \in \mathbb{R}^{K\times M\times d}$, which is then fed into a self-attention module to fuse multi-scale information. In the self-attention module, we first utilize three projection matrices (i.e., $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$) to translate $\mathbf{J}$ to $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$. Then, we compute the inter-scale attention to update $\mathbf{J}$ as follows:

$$\mathbf{J} = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \tag{8}$$

By concatenating features of different scales at the same location, we transform $\mathbf{J}$ into $\widetilde{\mathbf{J}} \in \mathbb{R}^{K\times Md}$. Then, in order to better fuse multi-scale information, we utilize $M$ parallel single-layer AGNN to promote the fused vector to retain scale-specific relevant information, based on

the original graph structure and the cosine similarity between the fused feature vectors. A single-layer AGNN consists of a linear layer and an attention-guided propagation layer, as shown below:

$$\mathbf{E}^{(m)} = \sigma(\widetilde{\mathbf{J}}\mathbf{W}^{(m)}), \tag{9}$$

$$\widetilde{\mathbf{E}}^{(m)} = \mathbf{P}^{(m)}\mathbf{E}^{(m)}, \tag{10}$$

where $\mathbf{E}^{(m)} \in \mathbb{R}^{K\times d}$ is a matrix after dimension reduction by a projection matrix $\mathbf{W}^{(m)} \in \mathbb{R}^{Md\times d}$, $\sigma(\cdot)$ represents the activation function. $\mathbf{P}^{(m)} \in \mathbb{R}^{K\times K}$ is the propagation matrix, e.g., $\mathbf{P}_i^{(m)} = \mathrm{softmax}\left([\beta^{(m)}\cos(\mathbf{E}_i^{(m)}, \mathbf{E}_j^{(m)})]_{j\in\mathcal{N}(i)\cup\{i\}}\right)$, $\beta^{(m)} \in \mathbb{R}$ is the parameter of the propagation layer, $\mathcal{N}(i)$ represents the neighborhood of node $i$, and $\widetilde{\mathbf{E}}^{(m)}$ is the output of the $m$th AGNN.

By adding the outputs of $M$ AGNN, the final output of the GMF module $\mathbf{U} = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_K] \in \mathbb{R}^{K\times d}$ can be obtained as:

$$\mathbf{U} = \frac{1}{M}\sum_{m}^{M}\widetilde{\mathbf{E}}^{(m)}. \tag{11}$$

### 3.4. Loss function

In addition to the attention loss $L_A$ described in Eq. (5), the proposed framework also contains two classification losses, $L_B$ and $L_G$, where $L_B$ is used in the backbone network for model training and $L_G$ is to guide the optimization of the GMF module, while collaboratively optimizing the convolutional parameters of the backbone network to guarantee that the network can accurately learn classification decisions. Specifically, by feeding one whole sMRI into the backbone network, its final output is $\mathbf{z}^I$, $L_B$ is represented by:

$$L_B = -\sum_{i=1}^{C} y_i \cdot \log\left(\frac{e^{\mathbf{z}_i^I}}{\sum_{j=1}^{C} e^{\mathbf{z}_j^I}}\right), \tag{12}$$

where $\mathbf{y} \in \mathbb{R}^C$ is the image label vector and $y_i \in \{0, 1\}$.

Given the fused feature matrix $\mathbf{U}$ in the graph branch, we utilize the same method as Eq. (3) to obtain its logits vector $\mathbf{z}^G$ by combining a linear layer with a GAP layer. Then, we can obtain the classification loss of the graph branch. The specific process is:

$$\mathbf{z}^G = \frac{1}{K}\sum_{i=1}^{K}\mathbf{u}_i\theta^G, \tag{13}$$

$$L_G = -\sum_{i=1}^{C} y_i \cdot \log\left(\frac{e^{\mathbf{z}_i^G}}{\sum_{j=1}^{C} e^{\mathbf{z}_j^G}}\right). \tag{14}$$

Based on the aforementioned three losses, we can obtain the final loss as follows:

$$L = \lambda(\tau) \times L_A + L_B + L_G \tag{15}$$

where $\lambda(\tau)$ is a weighted function changing with the number of epochs to adjust the term $L_A$.

For clarity, we present the detailed training procedure in Algorithm 1.

**Algorithm 1** LA-GMF

**Input:** Training data $\{\mathbf{X}_i, \mathbf{y}_i\}_{i=1}^N$, where $\mathbf{X}_i$ is a whole sMRI and $\mathbf{y}_i$ is its label vector, the number of training epochs T.
**Output:** Well trained model LA-GMF.

**for** $\tau \in [1, T]$ **do**
  **for** $\mathbf{X}_i, \mathbf{y}_i$ in $\{\mathbf{X}_i, \mathbf{y}_i\}_{i=1}^N$ **do**
    $\mathbf{T} \leftarrow \mathrm{reshape}(f_{CNN}^{-1}(\mathbf{X}_i), (N, -1))$ ▷ Obtain and reshape the feature map of the last convolutional layer.
    $\alpha, \mathbf{z}^I, \{\mathbf{z}_i\}_{i=1}^N \leftarrow$ Eqs. (1)-(4)
    $\widetilde{\alpha} \leftarrow \mathrm{topk}(\alpha)$ ▷ Obtain the $K$ positions with the highest attention scores.
    $\{\mathbf{T}^{(m)}\}_{m=1}^M \leftarrow \mathrm{getPatches}(f_{CNN}(\mathbf{X}_i), \widetilde{\alpha})$ ▷ Extract multi-scale patches using top-k indices from the backbone.
    $\{\mathbf{A}^{(m)}\}_{m=1}^M, \{\mathbf{H}^{(m)}\}_{m=1}^M \leftarrow$ Eqs. (6)-(7) ▷ Create $M$ graphs and perform graph convolution operations.
    $\widetilde{\mathbf{J}} \leftarrow$ Eq. (8) ▷ Perform self-attention and reshape operations.
    $\{\widetilde{\mathbf{E}}^{(m)}\}_{m=1}^M \leftarrow$ Eqs. (9)-(10) ▷ Calculate the output of $M$ AGNN
    $\mathbf{U} \leftarrow$ Eq. (11) ▷ Obtain the fused output
    $L \leftarrow$ Eq. (15) ▷ Calculate the final loss
  **end for**
  Back-propagate $L$ to update model parameters;
**end for**

## 4. Experiments

### 4.1. Datasets

In our experiments, we evaluate the proposed framework on two popular datasets: the AD Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu), the Australian Imaging, Aging Biomarkers and Lifestyle Flagship Study (AIBL, https://aibl.csiro.au). **ADNI** consists of a total of 1335 1.5T/3T T1-weighted baseline sMRI scans spanning two ADNI stages (i.e., ADNI-1, ADNI-2). These subjects are categorized into three groups: AD, MCI and NC, comprising 324 AD subjects, 518 MCI subjects and 493 NC subjects. **AIBL** is composed of baseline sMRI scans from 531 different subjects, including 75 AD and 456 NC subjects. The demographic information of the subjects in these datasets is summarized in Table 1.

### 4.2. Image preprocessing

We preprocess all sMRI scans using a standard pipeline. First, we convert the DCM files to NIFTI format based on the SimpleITK [29] library (only required when processing the AIBL dataset). Then, we adopt the fslreorient2std tool to reorient the image to match the orientation of the standard template image. Next, we employ the robustfov tool to crop the sMRI image to effectively remove the neck and lower head areas. This preprocessing step facilitates subsequent registration and skull stripping operations. The FLIRT tool [30] is used to register all sMRI images into a Colin27 template [31], remove global linear differences, and resample all images to the same resolution (i.e. $1 \times 1 \times 1$ mm$^3$) and size (i.e. $191 \times 217 \times 191$). Finally, we apply the BET tool [32] to remove the skull and dura mater. It is worth noting that fslreorient2std, robustfov, FLIRT and BET are all integrated tools in the FSL package [33].

### 4.3. Comparison methods

We employ five 3D CNN models as the comparison methods, which utilize the whole sMRI as the model input and train the model in an end-to-end manner. We briefly describe them as follows:

- **VoxCNN** [17]: which extends the VGG model to a 3D CNN structure for AD diagnosis.
- **VoxResNet** [17]: which extends the ResNet model to a 3D version for AD diagnosis.
- **Attention ResNet** [27]: which implements a simple attention via 3D convolution with the ReLU function, and embed the attention module into a 3D ResNet network, so as to provide interpretability analysis.
- **pABN** [21]: which designs a parallel attention-augmented bilinear network for AD diagnosis, aiming to extract fine-grained representations with small parameter overhead.
- **AMSNet** [22]: which integrates the multi-scale fusion module of dilated convolution and improves the attention mechanism in [27].

**Table 1**
Baseline demographic information of subjects included in three public datasets (i.e., ADNI-1, ADNI-2, and AIBL). Gender is presented in the form of male/female. Age, years of education, and MMSE scores are presented as mean ±standard deviation (SD).

| Dataset | Group type | Gender (Male/Female) | Age (Mean ± SD) | Education (Mean ± SD) | MMSE (Mean ± SD) |
|---|---|---|---|---|---|
| ADNI-1 | AD | 93/91 | 75.30 ± 7.44 | 14.59 ± 3.12 | 23.40 ± 2.04 |
| | MCI | 176/110 | 74.51 ± 7.16 | 15.60 ± 2.93 | 27.04 ± 1.76 |
| | NC | 107/107 | 75.89 ± 4.99 | 16.07 ± 2.86 | 29.13 ± 0.99 |
| ADNI-2 | AD | 81/59 | 74.59 ± 8.24 | 15.77 ± 2.63 | 23.09 ± 2.06 |
| | MCI | 136/96 | 71.47 ± 7.21 | 16.39 ± 2.60 | 27.88 ± 1.75 |
| | NC | 127/152 | 73.05 ± 5.98 | 16.62 ± 2.52 | 29.03 ± 1.23 |
| AIBL | AD | 29/46 | 73.56 ± 7.52 | – | 20.31 ± 5.47 |
| | NC | 191/265 | 72.36 ± 6.12 | – | 28.71 ± 1.22 |

### 4.4. Experiment setup

We conducted our experiments by using the framework PyTorch on a single GPU (i.e. NVIDIA GeForce 3090 24 GB).

All images used in our experiments are cropped to $160 \times 192 \times 160$ by removing uninformative background parts. Additionally, we augment training images by using translation and the mirror operation to enlarge the diversity of training data, where translation is to translate the image by one voxel in one of six directions (up, down, front, back, left, right) with equal probability, and the mirror operation is to mirror the left and right brains of the input image with a probability of 0.5.

We adopt a five-fold cross-validation strategy to comprehensively evaluate the performance of the proposed framework. Specifically, we design four sets of experiments to examine the model performance on different tasks, including AD-NC, AD-MCI, MCI-NC and AD-MCI-NC. For the binary classification task of AD-NC, We randomly divide the ADNI dataset into 5 subsets, with four of these subsets (including 80% subjects of the dataset) dedicated to model training and the remaining one used for testing. For the AIBL dataset, we utilize it as an independent testing set to assess the model's generalization ability. For the other three tasks, we report the average results of the model using five-fold cross-validation on the ADNI dataset. During the model training process, we adopt Stochastic Gradient Descent (SGD) as the optimizer, initialize the learning rate to 0.001 and decay it by 0.5 for every 20 epochs. Additionally, we totally train the model 100 epochs, with setting the batch size as 3.

We utilize three crucial metrics – $F_1$-score, Area Under the Curve (AUC), and Accuracy – to rigorously assess the model during the testing phase, ensuring a comprehensive evaluation.

In the proposed framework, there is one essential hyperparameter $\lambda(\tau)$ to weight the loss function $L_A$, where $\tau$ is the number of current training epochs. For the binary classification task of AD-NC, we recommend to set $\lambda(\tau)$ as 0 ($\tau \leq 20$) and 0.2 ($\tau > 20$). When processing AD-MCI, MCI-NC and AD-MCI-NC tasks, we recommend to set $\lambda(\tau)$ as 0 ($\tau \leq 40$) and 0.2 ($\tau > 40$). This is because the classification performance is low in the early stage of network training, imposing attention constraints based on logits can easily lead to trivial solutions. Therefore, for these three relatively difficult new tasks, we choose to postpone the application of the $L_A$ constraint. Additionally, considering that high-level features typically carry more semantic information, to ensure our graph towards homogeneity, we set the thresholds used for graph construction to $e_1 = 0.4$ and $e_2 = 0$, where $e_1$ represents the threshold employed for high-level features.

### 4.5. Classification results

Table 2 shows the classification results of all methods on the ADNI and AIBL datasets for AD-NC classification. As we can see, the proposed LA-GMF achieves better performance than the other comparison methods on the two datasets, in terms of all the three metrics, $F_1$-score, AUC and Accuracy. Specifically, compared to the best competitor, LA-GMF obtains the improvements of 1.06%, 0.07%, and 0.89% in terms of $F_1$-score, AUC, and accuracy, respectively, on the ADNI dataset. On the AIBL dataset, LA-GMF obtains 1.5% and 0.97% higher performance than the best competitors, in terms of $F_1$-score and Accuracy, respectively.

We believe that the outstanding performance of LA-GMF can be attributed to two main factors: Firstly, compared with methods that ignore the attention mechanism or rely only on image classification loss to guide attention mining, LA-GMF constrains attention scores through probability logic vectors. This effectively bridges the gap between attention scores and semantic categories, significantly enhancing the model's representation learning capabilities. Secondly, the graph branch network of LA-GMF not only models the complex relationships between related patches, but also combines self-attention mechanisms to deeply explore the complementarity and correlation between features at different scales, providing richer information for the final classification decision.

In order to comprehensively evaluate the generalization performance of LA-GMF, we further compare LA-GMF with five comparison methods on two binary classification tasks (AD-MCI and MCI-NC) and one ternary classification task (AD-MCI-NC). The detailed results are shown in Table 3, which suggests that LA-GMF still have superior classification performance over the comparison methods on these three more challenging tasks. For example, on the AD-MCI classification task, LA-GMF improves AUC and accuracy by 1.35% and 1.18%, respectively, compared to the best competitor. On the MCI-NC classification task, LA-GMF improves $F_1$-score, AUC and Accuracy by 0.42%, 1.89% and 0.69%, respectively. Interestingly, we note that existing attention-based 3D subject-level methods perform worse than methods without attention on these more challenging tasks. The reason might be that the attention mechanism is too simple or inflexible, so that it fails to effectively identify and focus on the key relationships on more difficult tasks, thereby introducing noise or misdirection which can degrade the model performance.

### 4.6. Ablation study

In this subsection, we conduct a detailed evaluation of two key components in the proposed framework, namely the LA module and the GMF module, based on the AD-NC binary classification task on the ADNI dataset. In order to comprehensively verify the effectiveness of these two modules, we design four methods, including Baseline, Baseline+LA, Baseline+LA+GCN and LA-GMF. Baseline indicates that only the backbone network is used for diagnosis. Baseline+LA embeds the LA module into the backbone network. Baseline+LA+GCN adds a graph branch based on single-scale features to Baseline+LA, and uses a two-layer GCN for feature processing. LA-GMF is a complete model using two-scale information. During this experiment, we adopt the same experimental setting as that in Section 4.5.

Table 4 shows the classification results of the four methods on the AD-NC task. As we can see, compared to Baseline, LA can boost $F_1$-score from 88.04% to 88.31%. Additionally, the graph branch network using single-scale modeling further improves $F_1$-score to 89.45%. By

**Table 2**
Comparison of five-fold cross-validation results obtained by different methods on ADNI and AIBL datasets for AD-NC binary classification. We bold and underline the best and second-best results at each setting, respectively.

| Model | ADNI | | | AIBL | | |
|---|---|---|---|---|---|---|
| | $F_1$-score | AUC | Accuracy | $F_1$-score | AUC | Accuracy |
| VoxCNN | 86.71 ± 2.23 | 91.79 ± 1.52 | 89.84 ± 1.45 | 68.70 ± 7.72 | 92.91 ± 1.20 | 88.66 ± 4.89 |
| VoxResNet | 87.12 ± 1.23 | 93.16 ± 1.32 | 90.21 ± 0.76 | 68.84 ± 2.39 | **93.41 ± 0.72** | 89.30 ± 1.48 |
| Attention ResNet | 87.16 ± 2.41 | 93.92 ± 0.74 | 89.97 ± 1.74 | 66.66 ± 5.98 | 92.24 ± 0.43 | 88.44 ± 3.98 |
| pABN | 79.70 ± 4.03 | 88.57 ± 2.21 | 84.83 ± 2.25 | 57.05 ± 6.71 | 85.64 ± 1.94 | 86.78 ± 2.56 |
| AMSNet | 89.94 ± 1.33 | 94.86 ± 1.10 | 92.13 ± 0.99 | 72.72 ± 4.25 | 93.16 ± 1.56 | 91.27 ± 2.03 |
| LA-GMF | **91.00 ± 2.11** | **94.93 ± 1.74** | **93.02 ± 1.57** | **74.22 ± 1.10** | **93.41 ± 0.72** | **92.24 ± 0.28** |

**Table 3**
Comparison of the five-fold cross-validation results of different methods on the three tasks (AD-MCI, MCI-NC and AD-MCI-NC) within the ADNI dataset. For the AD-MCI-NC task, we report macro-averaged results. We bold and underline the best and second-best results at each setting, respectively.

| Task | Metrics | VoxCNN | VoxResNet | Attention ResNet | pABN | AMSNet | LA-GMF |
|---|---|---|---|---|---|---|---|
| AD-MCI | F1-score | 59.48 ± 4.81 | **61.08 ± 4.37** | 59.53 ± 6.53 | 44.71 ± 17.14 | 55.49 ± 7.80 | 60.76 ± 3.71 |
| | AUC | 73.24 ± 4.23 | 73.77 ± 4.72 | 72.66 ± 4.06 | 65.33 ± 6.76 | 70.94 ± 4.11 | **75.12 ± 3.38** |
| | Accuracy | 72.57 ± 1.47 | 72.57 ± 2.78 | 70.79 ± 1.93 | 67.58 ± 4.78 | 70.67 ± 2.26 | **73.75 ± 1.89** |
| MCI-NC | F1-score | 67.22 ± 2.79 | 69.73 ± 2.23 | 68.48 ± 4.10 | 65.03 ± 3.82 | 69.38 ± 3.27 | **70.15 ± 3.04** |
| | AUC | 73.26 ± 2.74 | 73.35 ± 2.77 | 70.38 ± 2.77 | 68.87 ± 6.56 | 72.15 ± 3.45 | **75.24 ± 2.59** |
| | Accuracy | 70.33 ± 1.76 | 69.34 ± 2.26 | 67.46 ± 1.90 | 66.97 ± 4.48 | 69.05 ± 3.12 | **71.02 ± 2.27** |
| AD-MCI-NC | F1-score | 59.07 ± 2.22 | 58.53 ± 2.59 | 56.59 ± 1.99 | 55.19 ± 2.33 | 56.87 ± 1.23 | **61.35 ± 2.53** |
| | AUC | 76.29 ± 1.37 | 74.40 ± 3.08 | 74.78 ± 1.14 | 72.23 ± 2.29 | 73.20 ± 2.91 | **76.60 ± 1.86** |
| | Accuracy | 60.00 ± 2.00 | 58.88 ± 2.93 | 57.90 ± 1.13 | 56.78 ± 1.83 | 57.68 ± 1.79 | **61.42 ± 2.39** |

**Table 4**
Classification results obtained by Baseline, Baseline+LA, Baseline+LA+GCN and LA-GMF on the AD-NC (ADNI) task. We bold and underline the best and second-best results at each setting, respectively.

| Metrics | Baseline | Baseline+LA | Baseline+LA+GCN | LA-GMF |
|---|---|---|---|---|
| $F_1$-score | 88.04 ± 1.17 | 88.31 ± 2.67 | 89.45 ± 2.29 | **91.00 ± 2.11** |
| AUC | 94.22 ± 1.06 | 94.67 ± 2.05 | 94.63 ± 1.24 | **94.93 ± 1.74** |
| Accuracy | 90.82 ± 1.22 | 91.19 ± 1.63 | 91.92 ± 1.51 | **93.02 ± 1.57** |

introducing dual scale features and combining them with the fusion module, the model can achieve 91.00% $F_1$-score, which is better than that only using single-scale features. Similar findings can be observed on other metrics and tasks. These results indicate that: (1) The proposed LA module, which aims to locate significant discriminative regions, can guide the network to learn better feature representations. (2) The interaction between patches introduced by our graph branching network is very beneficial to diagnostic tasks. (3) Our proposed fusion module using dual-scale features can be greatly contributed to the improvement of model performance.

In addition to the overall evaluation of the module, we also conduct ablation experiments on the internal design of the LA module to evaluate the effectiveness of different implementations. Specifically, our experimental settings are shown as follows:

1. (FC & MLP): Utilizing FC for patch classification and MLP to calculate attention weight. Note that the FC for patch classification has different parameters from that for image classification.
2. (Shared FC & FC): Using shared FC for patch classification, while using additional FC to calculate attention weights. Note that the additional FC has different parameters from that for patch and image classification.
3. (All Shared FC): Shared FC is employed for both patch classification and attention calculation.
4. (All MLP): Using two independent MLPs for patch classification and attention calculation, respectively.
5. (MLP & Shared FC): MLP is employed for patch classification, but shared FC is used for attention calculation.
6. (Shared FC & MLP): Shared FC is employed for patch classification, and MLP is used for attention calculation. This is our choice in the proposed framework.

When using shared FC to obtain the attention weight, it requires to modify the calculation way of the attention mechanism to match the dimensions of logits and attention. Here, we adopt the similar calculation way in loss-based attention [34]. Specifically, the patch-level logits are first obtained by shared FC, and then the $L_2$ norm of the logits is calculated and processed by sigmoid to obtain the attention score. We display the experimental results of above six cases in Table 5, which demonstrates that using shared FC for patch classification and MLP for attention calculation can obtain better performance than the other five cases. This phenomenon might be because shared FC can simultaneously classify patch- and image-level features to better maintain their semantic consistency. Additionally, MLP introduces non-linear layers, which can learn better learn attention weights.

### 4.7. Interpretation experiments

In addition to classification experiments, we also evaluate the interpretability of the proposed framework from both ROI and patch aspects based on the AD-NC classification task. We select two attention based methods, i.e. Attention ResNet and AMSNet, for comparative analysis. Due to the existence of two parallel attention modules in pABN, it is difficult to effectively extract patches, so we exclude it here. Additionally, in order to better evaluate the interpretation performance of the LA module, we only replace LA with one popular attention mechanism, i.e., loss-based attention [34], in our framework and maintain the remaining modules unchanged, in which loss-based attention connects attention mechanism with loss function to mine significant patches. We refer to this modified model as Loss-Attention.

**ROI-based Interpretability Evaluation**. As shown in Fig. 3, we upsample the attention maps and overlay them on the original images to provide visualization-based explanations, where dark red indicates larger weights and dark blue indicates smaller weights. This provides preliminary evidence that our method can accurately localize important brain regions. However, this visualization method is unstable and can only provide a qualitative explanation of the model's positioning

**Table 5**
Classification results of ablation experiments within the LA module on the AD-NC (ADNI) task. We bold and underline the best and second-best results at each setting, respectively.

| Metrics | Baseline+LA (FC & MLP) | Baseline+LA (Shared FC & FC) | Baseline+LA (All Shared FC) | Baseline+LA (All MLP) | Baseline+LA (MLP & Shared FC) | Baseline+LA (Shared FC & MLP) |
|---|---|---|---|---|---|---|
| F1-score | 87.16 ± 1.56 | 87.55 ± 2.22 | 86.80 ± 2.55 | 87.84 ± 1.73 | 87.56 ± 2.86 | **88.31 ± 2.67** |
| AUC | 94.01 ± 1.21 | 93.97 ± 2.51 | 93.60 ± 1.22 | 94.45 ± 1.05 | 93.43 ± 1.72 | **94.67 ± 2.05** |
| Accuracy | 90.09 ± 0.98 | 90.33 ± 1.53 | 90.08 ± 1.68 | 90.21 ± 1.37 | 90.57 ± 1.54 | **91.19 ± 1.63** |

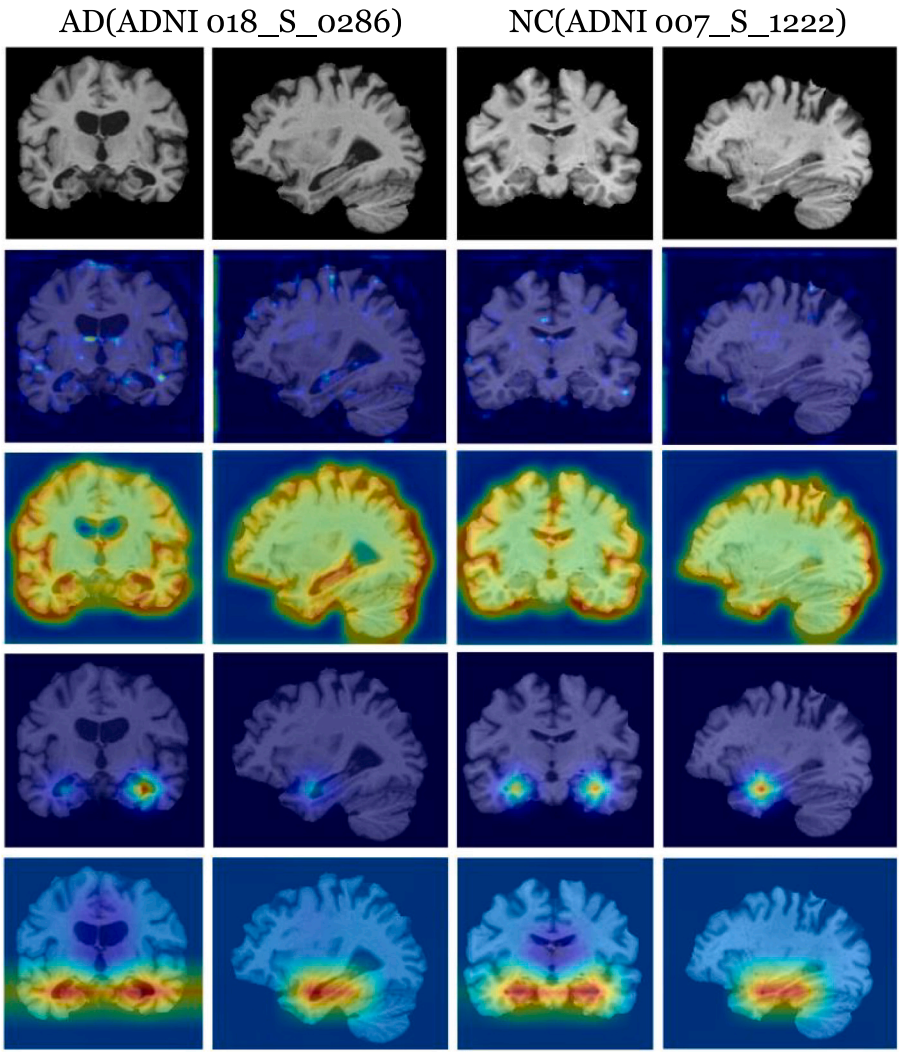


**Fig. 3.** Heatmaps generated by four different models. From the second to the fifth row, the heatmaps are generated by Attention ResNet, AMSNet, Loss-Attention and LA-GMF, respectively.

ability. Therefore, we designed a quantitative evaluation method based on ROI and average precision (AP).

We upsampled the attention map to the original image size and then divided it into multiple ROIs using the AAL template [35] (as both our image and AAL template are registered onto the Colin27 template, this approach is feasible). The top 20 ROIs with the highest scores in each image are counted as sample-level discriminative regions mined by the model. For each method, we saved 5 model weights (one for each fold). Among the discriminant regions of all samples mined by the five model weights, the top 20 ROIs with the highest frequency of occurrence are regarded as the population-level discriminant regions corresponding to this method.

Several literatures [36–38] have illustrated that the hippocampus, amygdala, fusiform, parahippocampal gyrus, entorhinal cortex, uncus and precuneus are the most relevant brain regions for Alzheimer's disease. Table 6 shows the corresponding indices of these brain regions in the AAL template.

**Table 6**
The corresponding number of brain regions in AAL.

| ROI | AAL |
|---|---|
| Hippocampal, uncus | 37/38 |
| Parahippocampal, entorhinal cortex | 39/40 |
| Amygdale | 41/42 |
| Fusiform | 55/56 |
| Precuneus | 67/68 |

We calculate the AP of the important brain regions in Table 6 and the discriminative brain regions mined by each method to evaluate the

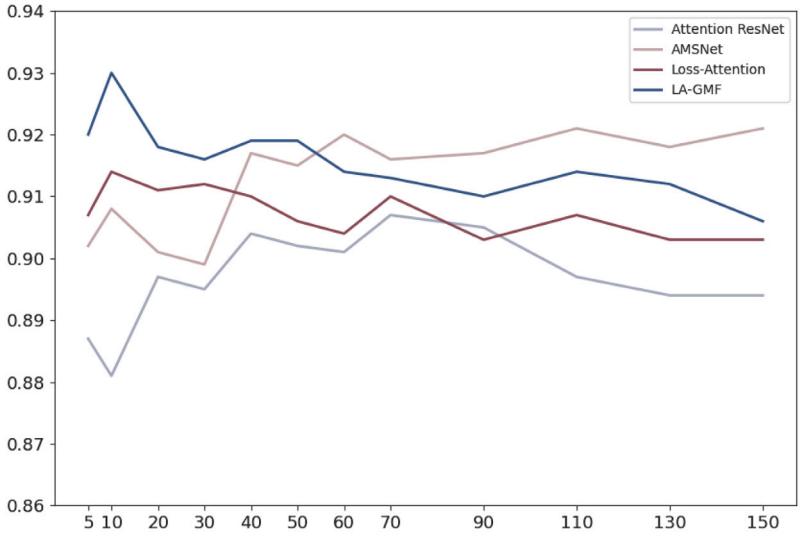


**Fig. 4.** Comparison of classification results using different numbers of patches on the AD-NC (ADNI) task.

**Table 7**
The top 20 brain regions with the highest scores among the four methods. The bold font represents the area related to Alzheimer's disease.

| Method | Index | AP |
|---|---|---|
| Attention ResNet | **41**, **38**, 84, 93, 110, **37**, 83, 3, 15, 34, 91, 88, 19, 78,7, 57, 79, 5, 26, 60 | 0.250 |
| AMSNet | **41**, 111, 47, 82, 18, **42**, 64, **40**, 112, 12, 43, 39, 86, 17, **37**, 31, 80, 48, 44, 115 | 0.238 |
| Loss-Attention | **42**, **41**, **39**, **40**, **37**, **38**, **55**, 83, 89, 87, **56**, 84, 88, 97, 76, 107, 75, 96, 98, 22 | 0.773 |
| LA-GMF | **41**, **42**, **40**, **39**, **38**, **37**, **55**, **56**, 96, 98, 89, 108, 107, 83, 90, 97, 87, 95, 84, 88 | **0.800** |

interpretability of the model. AP is defined as:

$$AP = \frac{\sum_{i=1}^{N_t} (P(i) \times e(i))}{N_s}, \tag{16}$$

where $N_s$ and $N_t$ represent the number of significant and total brain regions in the sequence, respectively; $e(i) \in \{0, 1\}$ represents whether the $i$th brain region is significant, and $P(i)$ represents the precision of the $i$th region containing significant brain regions, namely:

$$P(i) = \frac{m_i}{i}, \tag{17}$$

where $m_i$ is the number of searched significant brain regions in the first $i$ brain regions. The discriminative brain regions selected by the four methods and their AP scores are shown in Table 7. We can see that the proposed method achieves 55.0%, 56.2% and 2.7% higher AP than Attention ResNet, AMSNet, and Loss-Attention respectively. This demonstrates the superior interpretation performance of the proposed framework, which might be significantly beneficial to clinical diagnosis.

**Patch-based Interpretability Evaluation**. Literature [39] shows that powerful interpretation methods often achieve the best performance with only a small number of patches. Therefore, in addition to ROI-based interpretability, we also use classification accuracy as a metric to evaluate the model's ability to select important patches without using any medically relevant prior knowledge. Fig. 4 shows the classification accuracy of four models using different patch quantities. From this, we can see that LA-GMF can achieve the best performance using only 10 patches, indicating that the discriminative patches mined by LA have higher information content.

## 5. Conclusion

This paper proposes an attention-guided deep learning framework, namely LA-GMF, to mine the significant patches from the whole-brain sMRI for identifying discriminative locations related to AD and boosting its diagnosis performance. In the proposed framework, the logits-constraint attention utilizes the patch prediction class probability to constrain the attention score, so as to reduce the gap between the attention mechanism and semantic significant regions. Meanwhile, the graph-based multi-scale fusion module effectively integrates multi-scale information while retaining the characteristics of the scale itself to learn better feature representations. Experimental results on two public datasets demonstrate the superior classification and interpretation performance of the proposed framework, i.e., integrating feature extraction, discriminative localization and multi-scale feature fusion into an end-to-end deep learning framework is feasible and beneficial for automatic diagnosis of AD.

However, AD diagnosis often requires a comprehensive evaluation combining multiple modalities. Our current approach relies solely on structural MRI, in the future, we plan to integrate clinical diagnostic information with other modalities to further boost AD diagnosis accuracy. Additionally, our method does not take into account the timing characteristics of AD, in the future, we will consider longitudinal tracking data in the ADNI data set to improve AD diagnosis performance. Moreover, the proposed framework focuses on binary classification tasks in this paper, in the future, we will extend the proposed framework to more multi-classification tasks.

## CRediT authorship contribution statement

**Jinghao Xu:** Data curation, Methodology, Writing – original draft. **Chenxi Yuan:** Data curation, Investigation, Methodology. **Xiaochuan Ma:** Validation, Visualization. **Huifang Shang:** Supervision, Writing – review & editing. **Xiaoshuang Shi:** Methodology, Project administration, Supervision, Writing – review & editing. **Xiaofeng Zhu:** Supervision, Writing – review & editing.

## Declaration of competing interest

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

## Data availability

The authors do not have permission to share data.

## Acknowledgments

Xiaofeng Zhu and Xiaoshuang Shi were supported by the National Key Research and Development Program of China (No. 2022YFA1004100), Jinghao Xu and Huifang Shang were supported by Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China and West China Hospital (No. ZYGX2022YGRH009).

## References

[1] W. Jagust, Vulnerable neural systems and the borderland of brain aging and neurodegeneration, Neuron 77 (2) (2013) 219–234.
[2] G. Frisoni, N. Fox, C. Jack Jr., P. Scheltens, P. Thompson, The clinical use of structural MRI in Alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77.
[3] S. Klöppel, C. Stonnington, C. Chu, B. Draganski, et al., Automatic classification of MR scans in Alzheimer's disease, Brain 131 (3) (2008) 681–689.
[4] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, et al., Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset, Neuroimage 48 (1) (2009) 138–149.
[5] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, et al., Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI, Neuroradiology 51 (2009) 73–83.
[6] X. Tang, D. Holland, A. Dale, L. Younes, M. Miller, A.D.N. Initiative, Baseline shape diffeomorphometry patterns of subcortical and ventricular structures in predicting conversion of mild cognitive impairment to Alzheimer's disease, J. Alzheimer's Dis. 44 (2) (2015) 599–611.
[7] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, et al., Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation, Med. Image Anal. 63 (2020) 101694.
[8] A. Farooq, S. Anwar, M. Awais, S. Rehman, A deep CNN based multi-class classification of Alzheimer's disease using MRI, in: 2017 IEEE International Conference on Imaging Systems and Techniques, IST, IEEE, 2017, pp. 1–6.
[9] M. Hon, N. Khan, Towards Alzheimer's disease classification through transfer learning, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2017, pp. 1166–1169.
[10] K. Aderghal, M. Boissenin, J. Benois-Pineau, G. Catheline, K. Afdel, Classification of sMRI for AD diagnosis with convolutional neuronal networks: A pilot 2-D+ study on ADNI, in: Proceedings of the International Conference on Multimedia Modeling, Springer, 2016, pp. 690–701.
[11] R. Cui, M. Liu, Hippocampus analysis by combination of 3-D DenseNet and shapes for Alzheimer's disease diagnosis, IEEE J. Biomed. Health Inform. 23 (5) (2018) 2099–2107.
[12] M. Liu, J. Zhang, E. Adeli, D. Shen, Landmark-based deep multi-instance learning for brain disease diagnosis, Med. Image Anal. 43 (2018) 157–168.
[13] M. Liu, F. Li, H. Yan, K. Wang, et al., A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, Neuroimage 208 (2020) 116459.
[14] 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.
[15] Y. Chen, Y. Xia, Iterative sparse and deep learning for accurate diagnosis of Alzheimer's disease, Pattern Recognit. 116 (2021) 107944.
[16] T. Wang, Q. Dai, A patch distribution-based active learning method for multiple instance Alzheimer's disease diagnosis, Pattern Recognit. (2024) 110341.
[17] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3D brain MRI classification, in: Proceedings of the IEEE International Symposium on Biomedical Imaging, IEEE, 2017, pp. 835–838.
[18] Z. Fan, J. Li, L. Zhang, G. Zhu, et al., U-net based analysis of MRI for Alzheimer's disease diagnosis, Neural Comput. Appl. 33 (2021) 13587–13599.
[19] 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.
[20] 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 Inf. 26 (11) (2021) 5289–5297.
[21] 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, Hum. Brain Mapp. 43 (2) (2022) 760–772.
[22] Y. Wu, Y. Zhou, W. Zeng, Q. Qian, M. Song, An attention-based 3D CNN with multi-scale integration block for Alzheimer's disease classification, IEEE J. Biomed. Health Inf. 26 (11) (2022) 5665–5673.
[23] Z. Pei, Z. Wan, Y. Zhang, M. Wang, C. Leng, Y.-H. Yang, Multi-scale attention-based pseudo-3D convolution neural network for Alzheimer's disease diagnosis using structural MRI, Pattern Recognit. 131 (2022) 108825.
[24] Z. Zhang, L. Gao, P. Li, G. Jin, J. Wang, A.D.N. Initiative, et al., DAUF: A disease-related attentional UNet framework for progressive and stable mild cognitive impairment identification, Comput. Biol. Med. 165 (2023) 107401.
[25] K. Thekumparampil, C. Wang, S. Oh, L. Li, Attention-based graph neural network for semi-supervised learning, 2018, arXiv preprint arXiv:1803.03735.
[26] L. Sun, T. Yin, W. Ding, Y. Qian, J. Xu, Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems, Inform. Sci. 537 (2020) 401–424.
[27] D. Jin, J. Xu, K. Zhao, F. Hu, et al., Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration, in: Proceedings of the IEEE International Symposium on Biomedical Imaging, IEEE, 2019, pp. 1047–1051.
[28] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, ICLR, 2017.
[29] Z. Yaniv, B.C. Lowekamp, H.J. Johnson, R. Beare, SimpleITK image-analysis notebooks: A collaborative environment for education and reproducible research, J. Digit. Imaging 31 (3) (2018) 290–303.
[30] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, Neuroimage 17 (2) (2002) 825–841.
[31] C. Holmes, R. Hoge, L. Collins, R. Woods, A. Toga, A. Evans, Enhancement of MR images using registration for signal averaging, J. Comput. Assist. Tomogr. 22 (2) (1998) 324–333.
[32] S. Smith, Fast robust automated brain extraction, Hum. Brain Mapp. 17 (3) (2002) 143–155.
[33] M. Jenkinson, C. Beckmann, T. Behrens, M. Woolrich, S. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.
[34] X. Shi, F. Xing, K. Xu, P. Chen, et al., Loss-based attention for interpreting image-level prediction of convolutional neural networks, IEEE Trans. Image Process. 30 (2020) 1662–1675.
[35] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, et al., 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.
[36] A. Convit, M. De Leon, C. Tarshish, S. De Santi, et al., Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease, Neurobiol. Aging 18 (2) (1997) 131–138.
[37] G. Karas, P. Scheltens, S. Rombouts, R. Van Schijndel, et al., Precuneus atrophy in early-onset Alzheimer's disease: A morphometric structural MRI study, Neuroradiology 49 (2007) 967–976.
[38] C. Galton, K. Patterson, K. Graham, M. Lambon-Ralph, et al., Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia, Neurology 57 (2) (2001) 216–225.
[39] H. Xiang, J. Shen, Q. Yan, M. Xu, X. Shi, X. Zhu, Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis, Med. Image Anal. 89 (2023) 102890.

**Jinghao Xu** is a current graduate student in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China. He received his B.S. degree in Software Engineering from Qingdao University of Science and Technology in 2020. His research interests include deep learning and medical image analysis.

**Chenxi Yuan** is a postdoctoral researcher at the University of Pennsylvania Perelman School of Medicine. She earned her Ph.D. in Industrial Engineering from Northeastern University in 2022, where her work centered on the development and utilization of generative adversarial networks (GANs) and natural language processing (NLP) in intelligent manufacturing. At present, Dr. Yuan's methodological endeavors are directed toward the development of deep learning techniques that facilitate precision medicine and equitably improve health outcomes. She received multiple awards including the Outstanding Student Scholarship from Northwest University in 2015, the Achievement Award Scholarship from the University of Florida in 2016, the MIE Graduate Student Conference Award from NEU in 2020, and the Reviewer's Favorite Paper Award in ICED 2023.

**Xiaochuan Ma** is an undergraduate student in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China. His current main research directions are deep learning and graph learning.

**Huifang Shang** is a faculty member in the Department of Neurology at the West China Hospital of Sichuan University. She obtained her medical doctor degree (2003) from University of Bern, Master degree (1999) from West China University of Medical Sciences. Her major research interests include clinical, genetic and basic research in neurodegenerative disorders and movement disorders.

**Xiaoshuang Shi** is a faculty member in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China (UESTC). He obtained his Ph.D. degree (2019) from University of Florida, Master degree (2013) from Tsinghua University, and Bachelor degree (2009) from Northwestern Polytechnical University. Before joining UESTC, he worked as a Postdoctoral fellow at the National Institutes of Health (NIH) (2020.01–2021.04), and as a research assistant at Tsinghua University (2013.09–2015.04). His major research interests include large-scale data retrieval, deep learning, medical image analysis.

**Xiaofeng Zhu** is a faculty member of University of Electronic Science and Technology of China, Chengdu, China. His current research interests include large-scale multimedia retrieval, feature selection, sparse learning, data preprocess, and medical image analysis.