Computer Methods and Programs in Biomedicine 254 (2024) 108259

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

MACFNet: Detection of Alzheimer's disease via multiscale attention and cross-enhancement fusion network

Chaosheng Tang [a], Mengbo Xi [a], Junding Sun [a,*], Shuihua Wang [b,*], Yudong Zhang [a,c,d,*], Alzheimer's Disease Neuroimaging Initiative [1]

[a] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, PR China
[b] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China
[c] School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, UK
[d] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Keywords:
Multi-scale attention
Cross-enhanced fusion
Multimodal image
Alzheimer's disease

ABSTRACT

Background and Objective: Alzheimer's disease (AD) is a dreaded degenerative disease that results in a profound decline in human cognition and memory. Due to its intricate pathogenesis and the lack of effective therapeutic interventions, early diagnosis plays a paramount role in AD. Recent research based on neuroimaging has shown that the application of deep learning methods by multimodal neural images can effectively detect AD. However, these methods only concatenate and fuse the high-level features extracted from different modalities, ignoring the fusion and interaction of low-level features across modalities. It consequently leads to unsatisfactory classification performance.

Method: In this paper, we propose a novel multi-scale attention and cross-enhanced fusion network, MACFNet, which enables the interaction of multi-stage low-level features between inputs to learn shared feature representations. We first construct a novel Cross-Enhanced Fusion Module (CEFM), which fuses low-level features from different modalities through a multi-stage cross-structure. In addition, an Efficient Spatial Channel Attention (ECSA) module is proposed, which is able to focus on important AD-related features in images more efficiently and achieve feature enhancement from different modalities through two-stage residual concatenation. Finally, we also propose a multiscale attention guiding block (MSAG) based on dilated convolution, which can obtain rich receptive fields without increasing model parameters and computation, and effectively improve the efficiency of multiscale feature extraction.

Results: Experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrate that our MACFNet has better classification performance than existing multimodal methods, with classification accuracies of 99.59 %, 98.85 %, 99.61 %, and 98.23 % for AD vs. CN, AD vs. MCI, CN vs. MCI and AD vs. CN vs. MCI, respectively, and specificity of 98.92 %, 97.07 %, 99.58 % and 99.04 %, and sensitivity of 99.91 %, 99.89 %, 99.63 % and 97.75 %, respectively.

Conclusions: The proposed MACFNet is a high-accuracy multimodal AD diagnostic framework. Through the cross mechanism and efficient attention, MACFNet can make full use of the low-level features of different modal medical images and effectively pay attention to the local and global information of the images. This work provides a valuable reference for multi-mode AD diagnosis.

1. Introduction

Alzheimer's disease (AD) is an irreversible degenerative brain disease, which is a serious disease to society [1]. As the global population continues to age, the number of dementia patients is increasing dramatically. Research suggests that about 50 million people were

* Corresponding author.
E-mail addresses: sunjd@hpu.edu.cn (J. Sun), shuihuawang@ieee.org (S. Wang), yudongzhang@ieee.org (Y. Zhang).

[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.cmpb.2024.108259
Received 20 November 2023; Received in revised form 10 May 2024; Accepted 29 May 2024
Available online 6 June 2024

affected by dementia in 2018, and this number is estimated to reach 152 million by 2050. The cost of the disease is currently estimated at $1 trillion per year and is expected to double by 2030 [2]. Due to the complexity of the pathogenesis, no effective drug or method has been developed to cure AD. Hence, accurate early detection and treatment of AD is of great significance.

According to the clinical symptoms of the patients, the subjects were divided into Controlled Normal (CN), Mild Cognitive Impairment (MCI), and AD [3,4]. Currently, the clinical examination methods for Alzheimer's disease primarily include electroencephalogram (EEG) examination, neuropsychological assessments, and neuroimaging examination [5]. Commonly used neuropsychological examinations include the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating (CDR) [6]. Neuropsychological examination is only an adjunctive diagnostic method in clinical practice. With advancements in technology, neuroimaging diagnosis has become the most crucial and intuitive method for diagnosing Alzheimer's disease. In neuroimaging diagnostics, the most commonly used are magnetic resonance images (MRI) and positron emission computed tomography (PET) images [7]. MRI is capable of demonstrating brain tissue with high-resolution imaging while clearly distinguishing between grey and white matter areas of the brain.PET can track and show the distribution of abnormalities in the brain using imaging agents.

In recent years, many researchers have used deep learning methods for AD diagnosis [8,9]. Convolutional neural networks (CNNs) can extract high-level features by stacking layers to capture subtle lesion sites [10]. However, unimodal AD images only contain partial abnormal information about the brain, making classification often difficult to achieve the desired results. For instance, MR images provide details of soft tissues and high-resolution anatomical information that reflect changes in brain structure, while PET images only provide functional information regarding blood flow and metabolic changes [11]. Considering the ability of multimodal images to provide rich and complementary information, the study of multimodal AD classification models has become an important approach in order to enhance classification accuracy [12,13].

There are three approaches to multimodal AD classification. (1) feature fusion method. This method feeds the original multimodal neuroimaging or clinical data into a multi-input network for feature extraction and then fuses the learned high-level features to improve AD classification accuracy [14,15]. The method utilizes the complementary information of different modes and effectively mitigates the problem of noise interference encountered in unimodal data. However, most studies focus only on fusing high-level features, thus ignoring the heterogeneity of different modality images and the interaction of low-level features. Additionally, simple feature fusion of low-dimensional clinical biomarker information with high-dimensional neuroimaging often leads to dimension mismatch [16]. (2) Image fusion method. This method integrates the complementary and related features of multiple images into a single fusion image, which can effectively improve the diagnosis and treatment effect [17]. However, due to factors such as image quality or availability, multimodal images may be incomplete, and this method is severely limited in practice. (3) Data generation method. This method primarily utilizes the Generative Adversarial Network (GAN) algorithm to directly generate missing data from available modal data [18]. However, due to the complexity of medical images, significant differences exist in semantics, resolution, and organization of edge information between synthetic and real images.

In addition, recent studies have found that despite multimodal approaches have achieved good results in AD diagnosis, there are still some challenges. For example, images from different modalities have different resolutions and feature expressions, and direct fusion of features from these images may lead to information loss or redundancy. Furthermore, due to the complexity of brain structure, the abnormalities associated with AD are distributed across multiple regions of the brain at different scales. Although CNNs enhance the ability to capture local information through convolutional operations, the resulting limitation is that CNNs are more concerned with local regional features of the brain, thus limiting their ability to model distant features between brain regions and global features. To solve this problem, multiscale methods have been proposed to capture detailed information on images at different scales [19]. The multi-scale method enhances image classification performance by constructing image pyramids with different scales and applying convolutional kernels of multiple scales to extract various features of different scale regions of interest (ROIs), including local and global information [20].

In summary, although CNNs have been extensively studied in multimodal AD diagnosis, some urgent problems still need to be solved. (1) The low-level features of different modalities contain more local and detailed information related to AD. Nevertheless, the existing multimodal feature fusion methods simply splice the high-level features, thus ignoring the cross-modal interactions of the low-level features, which limits the shared representation of the model [21]. Although some studies have employed self-attention to address the cross-modal interaction problem, there are still problems such as low fusion efficiency and large size of parameters. (2) Abnormalities associated with AD are distributed across multiple regions of the brain at different scales [22,23]. However, most CNN-based multimodal methods focus only on a single scale and cannot effectively extract global features across brain regions. In this context, the powerful extraction capability of multiscale methods for local and global features of images under different receptive fields becomes more and more prominent [24]. For example, Lu et al. [25] extracted multiscale features by manually segmenting multimodal images into patches of different sizes, and experimentally demonstrated the effectiveness of multiscale feature extraction in AD diagnosis; however, manually segmenting the patches resulted in the loss of details of AD-related lesions. Unlike Lu, Gao et al. [26] extracted multiscale information for AD based on multiscale pyramid convolution. However, the computational complexity of pyramid convolution is high, and features extracted at large scales may overlap with those extracted at small scales, leading to information redundancy.

In this paper, we design a multiscale attention and cross-enhancement fusion network, MACFNet, targeting the problem that existing feature fusion methods ignore the low-level feature interactions, the interaction and fusion of multilevel low-level features between different modalities are realised through a two-branch crossover structure, and then the proposed efficient spatial channel attention mechanism is used in the residual structure, forcing the network to pay attention to the important features related to AD, to achieve structural feature enhancement in MRI and functional feature enhancement in PET. In addition, a multi-scale attention guidance module based on inflationary convolution is proposed to address the redundant information and noise problems brought by existing multi-scale methods. In the multiscale feature extraction stage, compared with multiscale pyramid convolution, the dilated convolution can obtain rich receptive fields without increasing model parameters and computation, which helps to improve the efficiency of multiscale feature extraction. In the multiscale feature fusion stage, we propose the hard attention mechanism. Unlike existing multiscale methods, this hard attention is used to suppress redundant information from different scales before fusion to reduce the negative impact caused by redundant information. Finally, the effectiveness of the proposed MACFNet is verified in AD, CN and MCI classification tasks. Overall, our contributions are summarised as follows:

(1) A dual-branch fusion CNN based on a crossover mechanism, MACFNet, to localize the discriminative regions related to AD without prior knowledge, effectively improving AD classification accuracy.
(2) A cross-enhanced fusion module is proposed, which enables the fusion and interaction of MRI and PET low-level features through the cross-over mechanism. In addition, the ECSA attention mechanism is designed to effectively focus on important

information related to AD for MRI structural feature enhancement and PET functional feature enhancement.

(3) A concise multi-scale attention guidance block is proposed by setting different dilation rates to obtain different receptive fields. It can also get discriminative information related to AD via hard attention with different scales.

The rest of the paper is arranged as follows: Section 2 introduces the related work, Section 3 introduces the proposed methodology, Section 4 introduces the experimental results, Section 5 the discussion, and Section 6 is the conclusion.

2. Related works

In recent years, CNNs have performed well in various visual tasks, such as classification and semantic segmentation. It can automatically extract features from multimodal images like MRI and PET, reducing manual feature extraction's complexity. This section introduces methods related to AD diagnosis from two aspects. Firstly, deep learning-based multimodal AD diagnosis methods are introduced, including (i) feature fusion-based methods, (ii) image fusion-based methods, (iii) data generation-based methods, (iv) cross-modal interaction-based methods. In addition, since AD affects multiple regions of the brain, multiscale methods are considered to address the problem of traditional CNNs focusing only on localized information. Therefore, this section also introduces several multi-scale methods for AD diagnosis.

2.1. Multimodal-based AD diagnosis

2.1.1. Feature fusion methods

The method based on feature fusion is to put neuroimaging or clinical data of different modalities into the multi-branch neural network separately, then gradually learn the potential feature representation of different modal data, and finally achieve AD classification [27]. Zhu et al. [28] proposed a deep multimodal discriminative network DMDIN. Firstly, features of different modalities were reconstructed in the common space using MLP, then shared expression coefficients were used to embed inter and intra-class structural information of different modalities, and finally, generalised typical correlation analysis (GCCA) was used to generate the discriminative common space. With this approach, it is possible to aggregate the same type of features and separate different types of features. Their method resulted in a classification accuracy of 96.75 % for AD and CN. Unfortunately, their approach solely focused on binary classification and neglected the multi-class classification scenarios. Xing et al. [29] used the vision transformer (VIT) instead of CNNs to improve AD classification accuracy. Considering the high computational cost of 3D images, they projected two 3D scans of PET (PET-AV45 and PETFDG) into a 2D fused image, then put the fused images into VIT for feature extraction separately, and finally fused features from different modalities for AD classification. However, their method only fuses high-level features and ignores the interaction of low-level features of different modalities. Abde et al. [30] integrated the original neuroimaging features and the ROI in the brain into a CNN and concatenated the high-level features learned from both modalities for AD diagnosis. However, this method neglected the interaction of low-level features. In addition, the process of synthesizing ROI images is unstable, which is not conducive to model training.

In practical diagnosis, features of different modalities possess varying dimensional information [16], and simply splicing and fusing these features will lead to dimensional mismatch. Therefore, Shi et al. [31] proposed a novel adaptive similarity-based multimodal feature selection (ASMFS) method. This method solves the challenge of dealing with high-dimensional features and effectively captures the intrinsic similarity of various modalities. Unfortunately, their method neglects the multi-class classification case. Chen et al. [32] proposed an attention-based approach for multimodal AD diagnosis. They combined neuroimaging, clinical data, and genetic information to extract features from different dimensions of data. First, they use convolutional blocks to extract high-dimensional features from neuroimages while using embedding techniques to transform preprocessed clinical and genetic data into feature vectors as well. Then, attention blocks are used to process feature vectors from different modalities. Their approach resulted in a classification accuracy of 97.90 % for AD and CN. However, this method does not take into account the dimensional differences in the data from different modalities, nor does it consider the interaction of low-level features. Tu et al. [16] proposed a cutting-edge model for Alzheimer's disease (AD) diagnosis, focusing on feature transformation using multimodal data. Initially, they enhanced the subjects' low-dimensional clinical and biological features by employing base-geometry algebra, thereby converting them into high-dimensional features. Subsequently, they introduced a feature filtering algorithm to exclude irrelevant features that lack significant information for AD diagnosis. Ultimately, they integrated the transformed features with MRI data. Compared to unimodal data, the multimodal feature fusion strategy is indeed effective in improving the classification accuracy of AD [33]. However, deep learning methods based on feature fusion tend to have high computational complexity due to the heterogeneity of different modal data. Furthermore, the majority of multimodal methods applied in neuroimaging focus only on fusing high-level features and neglecting the interaction of low-level features. As a result, they fail to effectively integrate functional and structural information from multimodal neuroimaging.

2.1.2. Image fusion methods

Unlike feature fusion methods, multimodal image fusion methods integrate complementary and relevant information from different modal images into the fused image [34]. It reduces computational complexity and transcends the limitations of low-level feature interactions. Ismail et al. [35] proposed an integrated learning architecture (UltiAz-Net) based on multimodal image fusion. They simultaneously used several different CNNs such as (AlexNet, Inception -V3, and ResNet) to extract high-level features from the fused images. In addition, they used a multi-objective optimisation algorithm to optimise each layer in the network for automatic AD classification. However, this approach of integrating multiple CNNs can cause the model to have high computational complexity. Kang et al. [36] conducted preprocessing on MRI and diffusion tensor imaging (DTI) by FreeSurfer software. Then, the two-dimensional slices were fused into a single RGB image according to corresponding indexes. Finally, the RGB image was fed into a pre-trained VGG16 network for classification. However, they only focused on the classification of MCI and CN without considering other classification scenarios. To reduce the noise and irrelevant information in the fused images, Song et al. [37] integrated the gray matter (GM) tissue information from MRI and PET images into a novel "GM-PET" image, selectively retaining the GM regions highly associated with the AD diagnosis, which effectively reduced the interruption of noisy information. Compared with other image fusion methods, their approach significantly reduced the number of parameters, achieving a classification accuracy of 94.11 % for AD and CN. However, the preprocessing step of this method is highly time-consuming. Although multimodal fusion methods can achieve better classification performance, in clinical practice, it is difficult to obtain both MRI and PET images of the same subject due to their availability and high economic costs. Therefore, AD diagnosis based on image fusion methods is often difficult to achieve.

2.1.3. Data generation methods

To address the problem of missing unimodal data, some studies have generated missing data from existing images by the GAN. Lin et al. [38] proposed a reversible generative adversarial network (RevGAN) to reconstruct PET images from MR images and then put the multimodal images into a CNN classification model for AD diagnosis. Their method achieved an accuracy of 89.05 % in the classification of AD and CN.

Table 1
Detailed information about the subjects.

Category	Number	Age	Gender(F/M)	CDR	MMSE
AD	214	74.1 ± 7.8	96/118	0.9 ± 0.6	22.2 ± 4.3
CN	326	76.3 ± 6.2	162/164	0 ± 0	28.6 ± 1.3
MCI	226	76.2 ± 7.3	81/145	0.6 ± 0.2	25.8 ± 4.4

Table 2
Imaging parameters of the scanner.

Imaging Parameter	Manufacturer		
	SIEMENS	Philips Medical Systems	GE Medical Systems
repetition time [TR]/ms	3000	6.8005	7.332
echo time [TE]/ms	3.5	3.116	3.036
inversion time [TI]/ms	1000	0	400
flip angle/°	8	9	11
thickness/mm	1.2	1.2	1.2
matrix size/voxel	192 × 192 × 160	256 × 256 × 170	256 × 256 × 196
field strength/T	3.0	3.0	3.0

Unlike the direct generation of missing images, Ye et al. [39] proposed a feature generation-based GAN, where features are first extracted from the complete MRI. Then the GAN is used to generate missing PET features. Additionally, linear attention is employed to effectively preserve salient features related to the disease. However, their approach focuses solely on binary classification and disregards the issue of multi-classification.

In general, although GAN methods can generate missing images for multimodal AD diagnosis, there are still some shortcomings. Firstly, the issue of uncertainty in generating results has not been settled. Due to the GAN's stochastic nature, the GAN's missing data may vary each time, making it impossible to have complete control over the accuracy and consistency of the generated images. Secondly, the training process is unstable. To achieve high-quality generated images, it requires prolonged training and fine-tuning. Otherwise, it may lead to an unstable training process and low-quality generated images. Finally, there is an issue concerning the visual quality of generated images. Although the discriminator in GAN can regulate the distribution of images and improve visual quality, the synthesized images fail to preserve crucial disease-related features due to the intricate spatial structure of medical images.

2.1.4. Cross-modal low-level feature interaction

Although many multimodal AD diagnosis methods are based on feature fusion, they all ignore the key issue of cross-modal interaction in multimodal learning [40,41]. Most methods focus only on high-level features, ignoring the interaction of low-level features, which limits the shared representation ability of the model [42].

Some studies have presented related suggestions and solutions for implementation. For example, Golovanevsky et al. [43] employ self-attention and cross-modal attention to integrate MR images, clinical data, and genetic information. For each modality, self-attention is first used to learn the most important features in the unimodal data, and then cross-modal attention is used to acquire features from other modalities to enhance its features. Finally, the output of the feature from the cross-modal attention layer is fused for AD diagnosis. However, the differences in the dimensions of image features and clinical features may cause dimension mismatch to rely solely on attention in the process of cross-modal interaction. Pan et al. [44] combined Transformer and GAN to propose a new cross-modal network to fuse MRI and DTI images. They proposed a two-way attention mechanism, which can extract fMRI functional features and DTI structural features by CNNs and Graph Convolutional Networks (GCNs), respectively. The features of different modalities are fused layer by layer and then fed into the Transformer model to realize AD diagnosis. However, their approach implements the self-attention mechanism by computing a set of query matrix Q, key matrix K, and value matrix V by linear projection, which has a high time complexity. In addition, their method transforms functional information and structural information into each other to achieve the fusion of complementary information, which may lead to the loss of important features of individual modalities. Unlike them, Leng et al. [45] proposed a simple cross-enhanced fusion network to diagnose AD, in which they divided the original image into non-overlapping chunks for input into the network. Specifically, they first proposed a multiscale remote receiver module to extract multiscale information using deep convolution with different kernel sizes on four branches. In addition, they propose two spatial enhancement modules and a channel enhancement module with a crossover structure, which are implemented to realize the interaction of different modal information through two different modality convolution blocks and a residual connection. This approach further enhances the cross-modal fusion capability of the network through the crossover mechanism, but they only consider the fusion of high-level features and ignore the importance of low-level features. In addition, they only considered binary classification and did not consider the multiclassification case.

2.2. Multi-scale-based AD diagnosis

The lesions associated with AD occur in multiple regions at different scales of the brain, and many studies use the multiscale approach in order to AD diagnosis [46]. Lu et al. [25] segmented MRI and PET scans into patches of different sizes and then extracted multiscale features using six deep neural networks. These features were then fed into another DNN for fusion for AD diagnosis. It is worth noting that their approach involved manual segmentation of patches at different scales, which could potentially result in the loss of intricate details relevant to AD. Song et al. [37] argued that by combining multi-scale methods, shallow detailed information as well as deep semantic information can be extracted from images. They proposed a 3D multiscale CNN architecture based on U-Net that utilizes skip connections to combine features of different scales in the fused image, while applying dropout layers to prevent overfitting, and finally fuses the different scale features to a classifier for AD classification. However, their approach uses different stages of downsampling to obtain multi-scale information, which may lose useful features. For multimodal AD diagnosis, Gao et al. [26] introduced a multiscale model in the GAN. They designed two pyramidal convolution blocks in the generator to process input images from different scales. Pyramidal convolution can capture different levels of image detail through different receptive fields. Their method achieves 92.7 % accuracy in the classification of AD and CN. Unfortunately, pyramidal convolution has high computational complexity, and features extracted on large scales may overlap with those on small scales, which may generate redundancy of information. To solve this problem, Liu et al. [47] proposed a dilated convolution-based model (MSCNet) to extract multi-scale features by different dilation rates and receptive fields. To learn the dependence among each channel, they present a double-weight network based on an improved channel attention mechanism. However, this method contains sum and cascade operations, which will generate redundant noise. Furthermore, they focused only on grey matter (GM) regions and white matter (WM) regions of the brain in unimodal MRI and did not consider the multimodal situation.

3. Methods

3.1. Data acquisition and preprocessing

3.1.1. Data acquisition

The dataset used in this article is from the ADNI database (https://

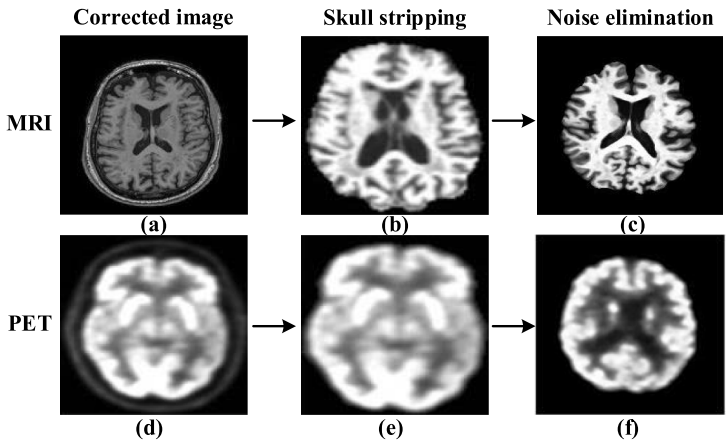

Fig. 1. Comparison of the corrected and preprocessed images. (a-c) show the preprocessing process of MRI axial image, and (d-f) show the preprocessing process of PET axial image.

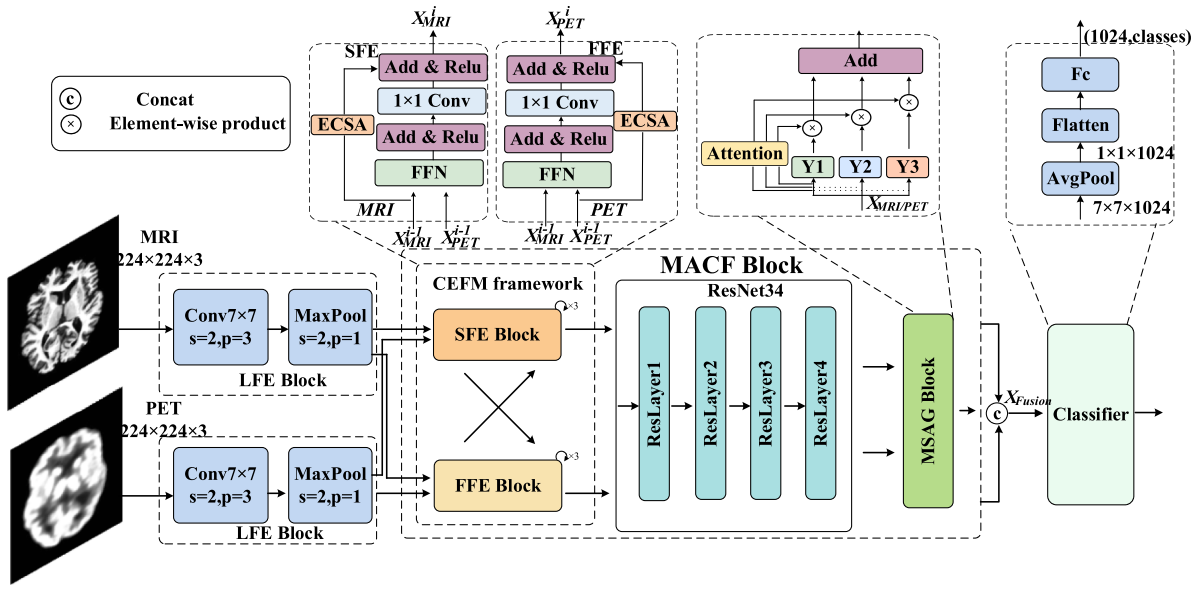

Fig. 2. The overall network architecture of MACFNet.

adni.loni.usc.edu/). ADNI is a multicenter longitudinal study designed to assist physicians in researching and developing the most effective clinical diagnostic and therapeutic protocols for AD. The database presents four studies (ADNI-1, ADNI-2, ADNI-GO, and ADNI-3). Following the methodology described in [43], we selected 766 subjects from the ADNI-GO and ADNI-2 phases who obtained MRI and PET images at baseline (10 months). These included 214 CE subjects, 326 CN subjects, and 226 MCI subjects. Each subject had one T1-weighted MRI image in NIFTI file format and one PET (FDG-PET) image. Table 1 shows the detailed information of the selected subjects.

Due to the ADNI volunteers being sourced from various countries and regions, the imaging equipment used for scanning differs across all subjects. For the convenience of the study, we selected the three most widely used imaging devices in current practice. MR images of all subjects were obtained using three magnetic resonance scanners. The imaging parameters of each scanner are shown in Table 2.

3.1.2. Data preprocessing

Due to variations in imaging techniques, significant differences and noise interference exist between different brain images, making it difficult to extract effective features from them. The preprocessing operation removes redundant information from the image irrelevant to AD diagnosis and lays the foundation for subsequent analysis. The preprocessing of our method is divided into three steps:

(1) Image correction. The raw MR images were first processed for image correction to remove phenomena such as motion artifacts and noise from the images. Image correction includes head motion correction and bias field correction. In this paper, head motion correction is performed on MR images based on anterior perineum (AC) and posterior perineum (PC) localization criteria. Then, the bias field correction is performed by the N4BiasFieldCorrection.sh module integrated into the ANTs tool.[2] The image size was set to 3, 3, and 3, and the scaling factors were set to 8, 4, and 2. In addition, the robustfov tool in the FMRIB Software Library[3] (FSL) was used to remove the neck region from the MR images.

(2) Image registration. We aligned MRI and PET images to address the spatial geometric inconsistencies of the different modality images. First, the PET images were aligned to the structural space

[2] Available at https://github.com/ANTsX/ANTs
[3] Available at https://fsl.fmrib.ox.ac.uk/fsl/fslwiki

Table 3
Architectural specification of the proposed MACFNet.

Module name	Layer	Output size (c, w, h)	Hyper-parameters
Input	/	3 × 224×224	/
LFE Block	Conv	64×112×112	$k=7, p=3, s=2$
	MaxPool	64×56×56	$k=3, p=1, s=2$
SFE Block	Conv × 4	64×56×56	$k=3, p=1, s=1$
	ECSA	64×56×56	/
	Conv	64×56×56	$k=1, p=0, s=1$
FFE Block	Conv × 4	64×56×56	$k=3, p=1, s=1$
	ECSA	64×56×56	/
	Conv	64×56×56	$k=1, p=0, s=1$
MSAG Block	Conv	32×7 × 7	$k=1, p=0, s=1$
	Scale1	512×7 × 7	$k=3, p=1, d=1$
	Scale2	512×7 × 7	$k=3, p=2, d=2$
	Scale3	512×7 × 7	$k=3, p=3, d=3$
	Attention	1 × 7 × 7	$k=3$
Output	AvgPool	1024×1 × 1	/

(* k is the kernel size, p is the padding size, s is the step size, and d is the dilation rate).

of the corresponding MR images of each subject using the FMRIB linear image alignment tool (FLIRT). The similarity measurement function was set to Normal Mutual Information (NMI), the image interpolation was set to B-spline interpolation, and the degree of freedom (DOF) was set to 6. Then, the MR images were aligned to the MNI standard space using the FMRIB Nonlinear Image Alignment Tool (FNIRT). Finally, PET images in the MRI structure space were registered to the MNI structure space according to the MRI transformation matrix.

(3) Tissue segmentation. Atrophy and lesions in grey matter, white matter, and regions are a focus of Alzheimer's research as they are closely associated with disease progression and symptoms. However, aligned MRI and PET images contain regions of the skull, cerebellum, etc., that are not directly relevant to the diagnosis of AD, and these regions may increase the computational burden and interfere with the diagnosis. Therefore, tissue segmentation of images is required. In this paper, we use the CAT12 toolbox in SPM[4] software to remove the cranium and cerebellum from MRI and PET images to reduce the computational burden and exclude the interference of irrelevant regions. Finally, to improve the image quality further, the images were smoothed using a Gaussian kernel function to suppress the noise in the functional images and to make the grey and white matter changes in the images clearer.

As is shown in Fig. 1, the MRI and PET images were processed in the axial view by the med2image tool.[5] Fig. 1(a)-Fig. 1(c) represents the axial image of MRI, while Fig. 1(d)-Fig. 1(f) represents the axial image of PET. Fig. 1(a) is the origin corrected image after removing the neck, Fig. 1(b) is the sliced image after skull stripping based on the corrected image, and Fig. 1(c) is the final image obtained after further noise elimination on the skull-stripped image. The same processing was applied to PET images.

3.2. Overview of proposed MACFNet

We propose the MACFNet, which is mainly composed of the cross-enhanced fusion (CEFM) module and the multi-scale attention guidance (MSAG) module. Fig. 2 shows the overall architecture of MACFNet. It consists of two dual-branch structures with different computational complexity.

[4] Available at https://neuro-jena.github.io/cat//index.html
[5] Available at https://github.com/FNNDSC/med2image

Firstly, MRI and PET images are simultaneously input into a proposed low-level feature extraction (LFE) block, which is a combination of 7 × 7 convolution followed by max pooling.

$$f_{LFE}(X) = MaxPool_{3\times3}[Conv(X_{MRI||PET}, k=7)], \tag{1}$$

where $X_{MRI||PET}$ denotes the MRI and PET images and f_{LFE} denotes feature maps output from the LFE block.

Secondly, MRI and PET images are fed into the SFE block and FFE block in CEFM, respectively, which is mainly composed of the feed-forward network (FFN) and efficient channel spatial attention (ECSA). A cross-connection is set between the SFE and FFE block in both branches to fuse the complementary information of different modalities fully.

$$\begin{cases} f_{SFE}(X) = CEFM_{SFE}\{Attention_{ECSA}(X_{MRI}) + concat[FFN(X_{MRI}), FFN(X_{PET})]\} \\ f_{FFE}(X) = CEFM_{FFE}\{concat[FFN(X_{MRI}), FFN(X_{PET})] + Attention_{ECSA}(X_{PET})\} \end{cases}, \tag{2}$$

where $f_{SFE}(X)$ and $f_{FFE}(X)$ denote the feature maps output from the SFE and FFE blocks, respectively, $Attention_{ECSA}$ denotes the ECSA attention mechanism, *concat*denotes the concatenation operation.

Subsequently, the fusion features obtained from the CEFM block are further extracted by ResNet34 to extract high-level features. These features that contain information from different modalities are fed into the MSAG block to extract detailed information at different scales. Finally, the outputs of the two branches are subjected to concat fusion operation and then fed into the classifier for classification.

Table 3 shows the detailed parameter settings of MACFNet, where we provide the output feature map sizes of the LFE Block, SFE Block, FFE Block, and MSAG Block modules, as well as the hyper-parameter values corresponding to each module.

3.3. CEFM module

The proposed CEFM module consists of three SFE modules and three FFE modules. The cross-structuring of the attention mechanism is used in order to enhance the structural features of MRI and the functional features of PET, while realizing the fusion of multi-level low-level features of different modalities.

3.3.1. SFE block and FFE block

(1) SFE block

The network architecture of the SFE block is shown in Fig. 3. The SFE block consists of a dual-branch residual architecture to fuse low-level features among different modalities and enhance the structural features of MRI.

The input feature map of SFE is represented by input1 and input2. where input1 represents $X_{MRI}^{i-1}(i=1,2,3)$ and input2 represents $X_{PET}^{i-1}(i=1,2,3)$. When $i=1$, the outputs of the LFE block on both branches are used as inputs to the SFE. When $i>1$, the outputs of the previous SFE and FFE blocks in the CEFM are used as inputs to the SFE block. To fully extract structural information from MR images and enhance the interaction of low-level features of different modalities, SFE employs a two-level residual architecture.

Specifically, the feature map $X_{MRI}^{i-1} \in \mathbb{R}^{H\times W\times C}$ is fed into the first-stage residual architecture ResBlock1 to extract structural features. Where *H*, *W*, and *C* denote the height, width, and number of channels of the feature map, respectively. To train the network better, We use LeakyRelu as the activation function, which is mathematically represented as follows:

$$\delta(x) = \max(0,x) + \alpha_i \times \min(0,x), \tag{3}$$

where δ denotes the LeakyRelu activation function, x denotes the input, and max and min denote maximum and minimum values, respectively. It can be seen that LeakyRelu assigns a non-zero slope α_i to all negative

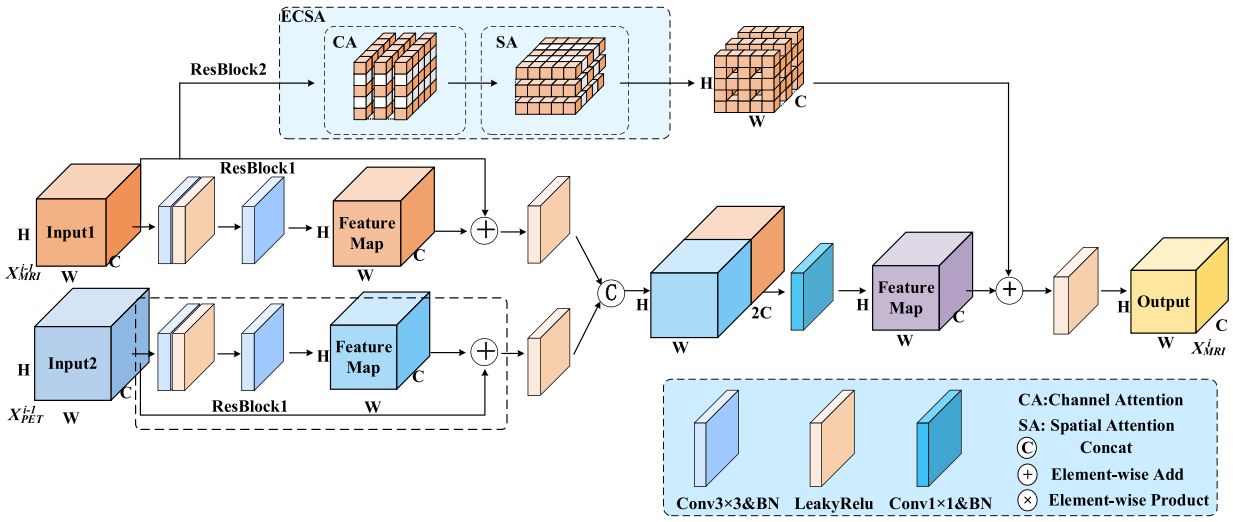

Fig. 3. Structural Feature Enhancement Block (SFE). X_{MRI}^{i-1} and X_{PET}^{i-1} denote the MRI and PET feature maps of the input of stage $i-1$, and X_{MRI}^{i} denotes the enhanced MRI structural feature map of the output of stage $i-1$, which is one of the inputs of the SEF Block in stage i.

Algorithm 1
The procedure of SFE.

Variable Definition:
$f(X)$: Output feature map; $concat$: Concatenation operation; +: Element-wise add;
SA: Spatial Attention; CA: Channel Attention; s: Residual structures stage;

Input: The Images X_{MRI}^{i-1} and X_{PET}^{i-1} of size H×W×C, residual structures stage s, convolution $Conv$
Output: An enhanced structural feature map $f_{fusion}(X_{MRI}^{i})$ of size H×W×C

for s in the range [1,3) do
  % Stage1: Achieving interaction of different modal low-level features
  if $s = 1$ do
    % Step 1: Using stage1 residual structure ResBlock1 to extract structural feature map
    $$f_s(X_{MRI}^{i-1}) = X_{MRI}^{i-1} + Conv_{3\times3}(X_{MRI}^{i-1})$$
    % Step 2: Using stage1 residual structure ResBlock1 to extract functional feature map
    $$f_s(X_{PET}^{i-1}) = X_{PET}^{i-1} + Conv_{3\times3}(X_{PET}^{i-1})$$
    % Step 3: Concatenate feature map
    $$Z_{fusion} = concat([f_s(X_{PET}^{i-1}), f_s(X_{MRI}^{i-1})]), Z_{fusion} \in \mathbb{R}^{H\times W\times 2C}$$
    % Step 4: Calculate the fusion feature map of ResBlock1
    $$U_{fusion} = Conv_{1\times1}(Z_{fusion}), U_{fusion} \in \mathbb{R}^{H\times W\times C}$$
  end if
  % Stage2: Enhancing MRI structural features
  if $s = 2$ do
    % Step 5: Using ECSA block to extract structural feature map
    $$f_s(X_{MRI}^{i-1}) = ECSA(X_{MRI}^{i-1}) = SA(CA(X_{MRI}^{i-1}))$$
  end if
end for
Calculate the enhanced structural feature map by ResBlock2
$$f_{fusion}(X_{PET}^{i}) = U_{fusion} + f_2(X_{MRI}^{i-1})$$

values, and during backpropagation, the gradient can be computed for the portion of the input that is less than zero, which allows the neurons in the network to maintain a certain update ability to better train the network.

The ResBlock1 consists of FFN and residual connections, where the FFN includes convolutional layers, Rectified Linear Unit layers, and batch normalization layers, and the mathematical representation of ResBlock1 is shown below:

$$FFN(X) = BN\{Conv_{C_{in}\to C_{out}}[\delta(BN(Conv_{C_{in}\to C_{out}}(X_{MRI}^{i-1}, k=3)))]\}, \quad (4)$$

$$ResBlock1(X_{MRI}^{i-1}) = X_{MRI}^{i-1} + FFN(X_{MRI}^{i-1}), \quad (5)$$

$$f(X_{MRI}^{i-1}) = ResBlock1(X_{MRI}^{i-1}), \quad (6)$$

where $Conv$ denotes the convolution layers, k denotes the kernel size, C_{in}

Algorithm 2
The procedure of FFE.

Variable Definition:
$f(X)$: Output feature map; $concat$: Concatenation operation; $+$: Element-wise add;
SA: Spatial Attention; CA: Channel Attention; s: Residual structures stage;

Input: The Images X_{MRI}^{i-1} and X_{PET}^{i-1} of size H×W×C, residual structures stage s, convolution $Conv$
Output: An enhanced functional feature map $f_{fusion}(X_{PET}^{i})$ of size H×W×C

for s in the range [1,3) do
  % Stage1: Achieving interaction of different modal low-level features
  if $s = 1$ do
    % Step 1: Using stage1 residual structure ResBlock1 to extract functional feature map
    $$f_s(X_{PET}^{i-1}) = X_{PET}^{i-1} + Conv_{3\times3}(X_{PET}^{i-1})$$
    % Step 2: Using stage1 residual structure ResBlock1 to extract structural feature map
    $$f_s(X_{MRI}^{i-1}) = X_{MRI}^{i-1} + Conv_{3\times3}(X_{MRI}^{i-1})$$
    % Step 3: Concatenate feature map
    $$Z_{fusion} = concat([f_s(X_{PET}^{i-1}), f_s(X_{MRI}^{i-1})]), Z_{fusion} \in \mathbb{R}^{H\times W\times 2C}$$
    % Step 4: Calculate the fusion feature map of ResBlock1
    $$U_{fusion} = Conv_{1\times1}(Z_{fusion}), U_{fusion} \in \mathbb{R}^{H\times W\times C}$$
  end if
  % Stage2: Enhancing PET functional features
  if $s = 2$ do
    % Step 5: Using ECSA block to extract functional feature map
    $$f_s(X_{PET}^{i-1}) = ECSA(X_{PET}^{i-1}) = SA(CA(X_{PET}^{i-1}))$$
  end if
end for
Calculate the enhanced functional feature map by ResBlock2
$$f_{fusion}(X_{PET}^{i}) = U_{fusion} + f_2(X_{PET}^{i-1})$$

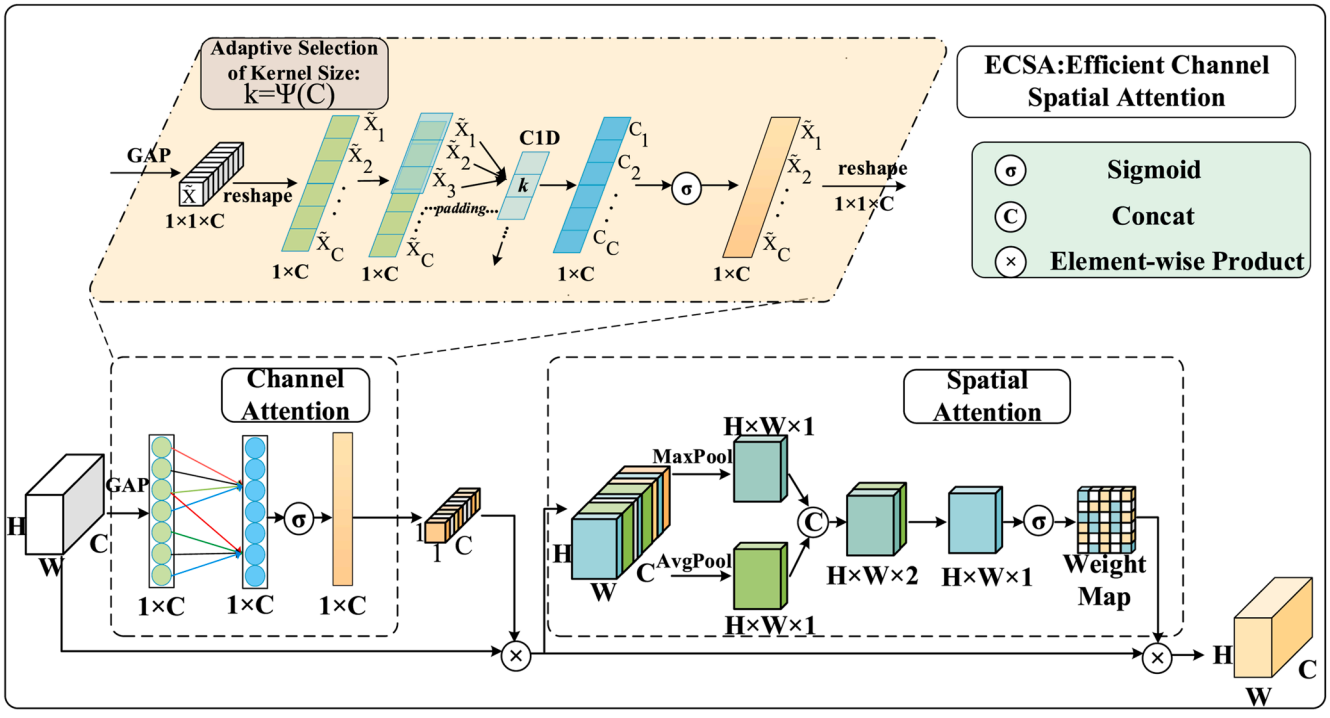

Fig. 4. Efficient channel spatial attention (ECSA).

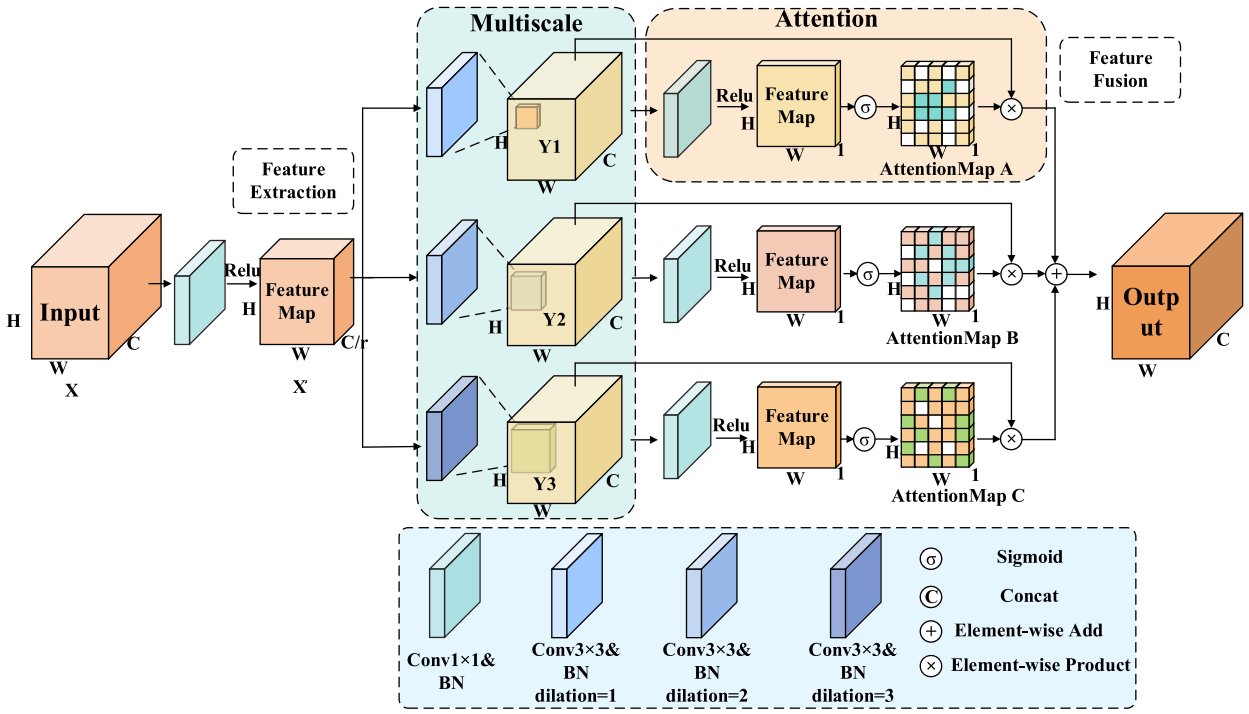

Fig. 5. Multi-scale attention guided block (MSAG).

Algorithm 3
The procedure of MSAG.

Variable Definition:
$A, B,$ and C: Attention map; $f(X)$: Output feature map $Conv$: Convolution convolution; $+$: Element-wise add; $\otimes$: Element-wise product;

Input: The feature map X_{MRI} or X_{PET} of size H×W×C, the dilation rate d, the scale i
Output: A multi-scale stacked receptive fields feature map $f_{fusion}(X_{MRI||PET})$ of size H×W×C
for i in the range [1, 3] do
  % Stage1: Calculate feature map in different scale receptive field
  if $i = 1$ do
    Calculate the feature map of the receptive field for scale 1
    $f(X^i_{MRI}) = Conv_{3\times3}(X_{MRI}, d = i) \parallel f(X^i_{PET}) = Conv_{3\times3}(X_{PET}, d = i)$
    Calculate the attention map
    $$A = \sigma(Conv_{1\times1}(X^i_{MRI\parallel PET})), A \in \mathbb{R}^{H\times W\times1}$$
  elif $i = 2$ do
    Calculate the feature map of the receptive field for scale 2
    $f(X^i_{MRI}) = Conv_{3\times3}(X_{MRI}, d = i) \parallel f(X^i_{PET}) = Conv_{3\times3}(X_{PET}, d = i)$
    Calculate the attention map
    $$B = \sigma(Conv_{1\times1}(X^i_{MRI\parallel PET})), B \in \mathbb{R}^{H\times W\times1}$$
  else do
    Calculate the feature map of the receptive field for scale 3
    $f(X^i_{MRI}) = Conv_{3\times3}(X_{MRI}, d = i) \parallel f(X^i_{PET}) = Conv_{3\times3}(X_{PET}, d = i)$
    Calculate the attention map
    $$C = \sigma(Conv_{1\times1}(X^i_{MRI\parallel PET})), C \in \mathbb{R}^{H\times W\times1}$$
end for
% Stage2: Calculate Multiscale Integration Feature Maps
$$f_{fusion}(X_{MRI||PET}) = A \otimes f(X^1_{MRI\parallel PET}) + B \otimes f(X^2_{MRI\parallel PET}) + C \otimes f(X^3_{MRI\parallel PET})$$

Table 4
Ablation experiments of ECSA in CEFM.

Auxiliary Diagnosis	SEN(%)	SPE(%)	ACC(%)	AUC(%)
AD/CN	**99.75**	**98.33**	**99.28**	**99.72**
w/o ECSA	98.83	98.14	98.91	99.69
AD/MCI	**99.76**	**96.72**	**98.47**	**99.89**
w/o ECSA	98.33	96.54	98.39	99.84
CN/MCI	**99.61**	**99.31**	**99.47**	**99.97**
w/o ECSA	99.07	99.02	99.07	99.93
AD/CN/MCI	**96.12**	**98.5**	**97.34**	**99.68**
w/o ECSA	96.04	98.18	96.55	99.26

(* SEN: sensitivity; SPE: specificity; ACC: accuracy, 'w/o' means without).

$\rightarrow C_{out}$ denotes the change in the number of channels of the feature map before and after the convolution operation. Here $C_{out} = C_{in} = C$, *BN* denotes the batch normalization layer. Where the 3×3 convolution uses padding to ensure that the size of the output feature map remains unchanged, and then the X_{MRI}^{i-1} is summed with the FFN output to get the ResBlock1 output feature map $f(X_{MRI}^{i-1})$.

To extract functional features from PET images, the feature map $X_{PET}^{i-1} \in \mathbb{R}^{H\times W\times C}$ has also been processed by the same residual structure. The outputs of these two residual structures are concatenated in the channel dimension to initially fuse structural and functional features, and the fusion process can be described as below:

$$f_Z(X_{fusion}) = concat\left[f(X_{MRI}^{i-1}), f(X_{PET}^{i-1})\right], f_Z(X_{fusion}) \in \mathbb{R}^{H\times W\times 2C}, \tag{7}$$

where $f_Z(X_{fusion})$ represents the fused feature map, $f(X_{MRI}^{i-1})$ and $f(X_{PET}^{i-1})$ represent the output of ResBlock1 on the respective corresponding branch.

Finally, we use 1×1 convolution to reduce the dimensionality of the fused feature map $f_Z(X_{fusion})$, mathematically represented as follows:

$$f_U(X_{fusion}) = Conv_{2C\rightarrow C}(f_Z(X_{fusion}), k=1), f_U(X_{fusion}) \in \mathbb{R}^{H\times W\times C}, \tag{8}$$

where $f_U(X_{fusion})$ represents the output fused feature map after dimensionality reduction.

In the second residual architecture, ResBlock2 is used to process the feature map $X_{MRI}^{i-1} \in \mathbb{R}^{H\times W\times C}$, which employs the ECSA attention mechanism, where the interdependencies between channels can be efficiently modeled through channel attention. In addition, ECSA utilizes spatial attention to extract the background and texture information of the image. We perform element-wise addition of ECSA attention-processed feature maps with 1×1 convolution-processed fusion feature maps $f_U(X_{fusion})$, and enhance MRI features by LeakyRelu. The procedure is as follows:

$$ResBlock2(X_{MRI}^{i-1}) = \delta\left(ECSA(X_{MRI}^{i-1}) + f_U(X_{fusion})\right), \tag{9}$$

$$f(X_{MRI}^{i}) = ResBlock2(X_{MRI}^{i-1}). \tag{10}$$

where $f(X_{MRI}^{i})$ denotes the enhanced structural feature map output by the SFE Block. *ECSA* represents the ECSA attention mechanism. The pseudo-code is shown in Algorithm 1:

(2) FFE block

Similar to SFE, the FFE module also consists of a two-branch residual architecture for fusing the low-level features of different modalities and enhancing the functional features of PET. The difference is that input1 of FFE denotes $X_{PET}^{i-1}(i = 1, 2, 3)$ while input2 denotes $X_{MRI}^{i-1}(i = 1, 2, 3)$. The rest of the architecture is the same as SFE's and will not be repeated. The pseudo-code is shown in Algorithm 2:

3.3.2. ECSA block

Research has shown that using attention to capture the dependencies of all channels in a feature map is both inefficient and unnecessary [48]. To effectively capture local cross-channel interactions, we propose an efficient channel spatial attention (ECSA), which can filter unimportant channel information. As shown in Fig. 4, it consists of a concatenated structure of efficient channel attention and spatial attention. Channel attention includes a global average pooling layer, a 1D convolutional layer, and the corresponding activation function, effectively focusing on the channel information related to AD pathological regions. Spatial attention is composed of global average pooling and global max pooling. It can be used to get both background information and image texture information.

The first is the efficient channel attention mechanism. Specifically, for the input feature map, first use the global average pooling (GAP) to pool the information of each channel into a real number to obtain aggregated features:

$$\widetilde{X} = f_{GAP}(X) = \frac{1}{W\times H}\sum_{w=1}^{W}\sum_{h=1}^{H} X(w,h), \widetilde{X} \in \mathbb{R}^{1\times 1\times C}, \tag{11}$$

where $X \in \mathbb{R}^{H\times W\times C}$ represents the input feature map. Here W and H denote the width and height, respectively, w and h represent pixel values, f_{GAP} represents the global average pooling operation and $\widetilde{X}$ denotes the statistical information associated with the channel.

To achieve local cross-channel interaction, we use the band matrix W_k representing the output attention feature map to learn the channel attention. If the input is the MRI feature matrix X_{MRI}, the mathematical expression of W_k is as follows:

Table 5
Ablation experiments of CEFM and MSAG in different classification tasks.

Task	Method	RN	CEFM	MSAG	SEN(%)	SPE(%)	ACC(%)	AUC(%)
AD/CN	MACFNet	√	√	√	**99.91**	**98.92**	**99.59**	**99.94**
	w/o CEFM	√		√	99.72	96.09	98.54	99.90
	w/o MSAG	√	√		99.75	98.33	99.28	99.72
	w/o CEFM&MSAG	√			98.39	94.93	98.21	99.80
AD/MCI	MACFNet	√	√	√	**99.89**	**97.17**	**98.85**	**99.91**
	w/o CEFM	√		√	99.88	96.05	98.60	99.94
	w/o MSAG	√	√		99.76	96.72	98.47	99.89
	w/o CEFM&MSAG	√			97.67	95.44	98.17	99.81
CN/MCI	MACFNet	√	√	√	**99.63**	**99.58**	**99.61**	**99.98**
	w/o CEFM	√		√	98.86	99.11	98.98	99.98
	w/o MSAG	√	√		99.61	99.31	99.47	99.97
	w/o CEFM&MSAG	√			98.16	98.78	98.83	99.91
AD/CN/MCI	MACFNet	√	√	√	**97.75**	**99.04**	**98.23**	**99.89**
	w/o CEFM	√		√	97.72	99.02	98.19	99.71
	w/o MSAG	√	√		96.12	98.5	97.34	99.68
	w/o CEFM&MSAG	√			96.21	98.46	97.19	99.68

* (RN means ResNet).

Table 6
Performance comparison with other methods.

Task	Method	SEN (%)	SPE (%)	ACC (%)	AUC (%)
AD/CN	Song et al.(2021) [37]	93.33	94.27	94.11	n/a
	Zhang et al.(2019) [54]	96.58	95.39	98.47	98.61
	Fang et al.(2020) [52]	95.89	98.72	99.27	n/a
	Gao et al.(2022) [26]	91.70	93.50	92.70	96.4
	Zhang et al.(2022) [53]	n/a	n/a	96.23	99.00
	Tu et al.(2022) [16]	97.40	93.00	96.20	98.60
	Abde et al.(2022) [30]	98.82	97.52	98.24	97.70
	Shi et al.(2022) [31]	96.10	97.47	96.76	97.03
	Leng et al.(2023) [45]	97.22	98.21	97.67	98.55
	Ismail et al.(2023) [35]	95.00	94.00	94.40	n/a
	MACFNet(ours)	**99.91**	**98.92**	**99.59**	**99.88**
AD/MCI	Song et al.(2021) [37]	71.19	85.94	80.80	n/a
	Zhang et al.(2019) [54]	97.43	84.31	88.20	88.01
	Fang et al.(2020) [52]	89.71	93.59	92.57	n/a
	Zhang et al.(2022) [53]	n/a	n/a	88.12	91.00
	Liu et al.(2022) [47]	94.91	**98.52**	94.44	97.00
	Ismail et al.(2023) [35]	89.20	93.30	90.00	n/a
	MACFNet(ours)	**99.89**	97.07	**98.85**	**99.90**
CN/MCI	Song et al.(2021) [37]	84.69	85.60	85.00	n/a
	Zhang et al.(2019) [54]	90.11	91.82	85.74	88.15
	Fang et al.(2020) [52]	88.36	92.56	90.35	n/a
	Zhang et al.(2022) [53]	n/a	n/a	87.45	95.0
	Abde et al.(2022) [30]	90.26	96.98	94.59	93.3
	Shi et al.(2022) [31]	85.98	70.90	80.73	78.75
	Ismail et al.(2023) [35]	96.00	89.20	93.20	n/a
	MACFNet(ours)	**99.63**	**99.58**	**99.61**	**99.98**
AD/CN/MCI	Song et al.(2021) [37]	55.67	83.40	71.52	n/a
	Zhang et al.(2022) [53]	n/a	n/a	80.34	95.00
	Golo et al.(2022) [43]	n/a	n/a	96.88	n/a
	Han et al.(2022) [55]	n/a	n/a	67.74	n/a
	Ismail et al.(2023) [35]	n/a	n/a	92.30	n/a
	MACFNet(ours)	**97.75**	**99.04**	**98.23**	**99.82**

(* Bold indicates the best value in terms of the evaluation indicator. 'n/a' means not available.).

$$W_k = \begin{bmatrix} \omega_{mri}^{1,1} & \cdots & \omega_{mri}^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \omega_{mri}^{2,2} & \cdots & \omega_{mri}^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \omega_{mri}^{C,C-k+1} & \cdots & \omega_{mri}^{C,C} \end{bmatrix}, \tag{12}$$

where k denotes the number of cross-channels, C denotes the number of channels of the input band matrix, and ω represents the learnable channel weights, the channel weights ω_i can be obtained by computing the interactions between $\widetilde{X}_i(i=1,2..C)$ and its k neighbors. ω_i is described as follows:

$$\omega_i = \sigma\left(\sum_{j=1}^{k} \omega_i^j \widetilde{X}_i^j\right), \widetilde{X}_i^j \in \Omega_i^k, \tag{13}$$

where Ω_i^k represents the set of k neighboring channels to $\widetilde{X}_i$. The above process can be implemented by a one-dimensional convolution:

$$W_k = \sigma(C1D_k(\widetilde{X})), \tag{14}$$

where $C1D_k$ represents a one-dimensional convolution of size k, and σ denotes the sigmoid function. In addition, the coverage range of the cross-channel is determined based on the size of the convolution kernel k, implying a potential mapping ψ between k and C. Since the number of input channels is usually a power of 2, the following mapping relationship is established:

$$k = \psi(C) = \left|\frac{b + \log_2 C}{\gamma}\right|, \tag{15}$$

where γ and b denote the hyperparameters of the mapping process, and the size of the k can be computed from the input number of channels C.

Finally, the matrix $W_k \in \mathbb{R}^{1\times1\times C}$ is element-wise multiplied with the input X to obtain the final output feature map of the CA.

$$f_{CA}(X) = X \otimes W_k, f_{CA}(X) \in \mathbb{R}^{H\times W\times C}, \tag{16}$$

where ⊗ denotes the element-wise product,$f_{CA}(X)$ represents the output feature map after channel attention.

The right side of Fig. 4 shows the spatial attention mechanism. For each channel of $f_{CA}(X)$, the maximum value and average value are computed separately. This preserves the background and texture information of the image. Subsequently, these two outcomes are stacked, and a 7 × 7 convolution operation with a single channel is performed to adjust the channel number. Finally, a sigmoid function is applied to generate the weight map of the input feature map:

$$f_{SA}(X) = \sigma[Conv_{2\to1}(concat(AvgPool(f_{CA}(X)), MaxPool(f_{CA}(X))), k=7], \tag{17}$$

where $f_{SA}(X) \in \mathbb{R}^{H\times W\times1}$ represents the output spatial attention weight map, which is then multiplied by the original input to obtain the ECSA output:

$$f_{ECSA}(X) = f_{CA}(X) \otimes f_{SA}(X), f_{ECSA}(X) \in \mathbb{R}^{H\times W\times C}, \tag{18}$$

where $f_{ECSA}(X)$ represents the output feature map of the ECSA.

3.4. MSAG block

The lesions associated with AD occur in regions of the brain at different scales. As illustrated in Fig. 5, to avoid the problem that feature extraction from a single scale leads to ignoring the details and overall structure of the image, the Multiscale Attention Guided (MSAG) framework is proposed, consisting of feature extraction and feature fusion.

In the feature extraction stage, dilated convolutions with different dilation rates are applied at multiple scales to identify brain atrophy. The utilization of dilated convolutions allows for a rich receptive field without increasing the model parameters or computational complexity. In the feature fusion stage, we use convolution and sigmoid as hard attention mechanisms to capture global contextual information at each scale. This approach helps mitigate the noise caused by redundant information at different scales. Finally, the feature maps extracted at different scales are element-wise added to obtain a fused feature map containing information at different scales.

In the feature extraction stage, for the input $X \in \mathbb{R}^{H\times W\times C}$, the dimensionality of the original channels is first reduced using 1 × 1 convolution to obtain $X' \in \mathbb{R}^{H\times W\times C'}$:

$$X' = Conv_{C\to C'}(X, k=1), \tag{19}$$

where $C' = C/r$ represents the number of channels in the output feature map, where r denotes the scaling factor to extract features at different scales. We use dilation convolution with size 3 × 3 and dilation rate 1, 2, and 3 on scales $F1$, $F2$, and $F3$, respectively. At the same time, in the process of convolution, padding is used to ensure that the outputs Y_1, Y_2, and Y_3 have the same size, and the output of each scale via the BN and ReLu activation functions:

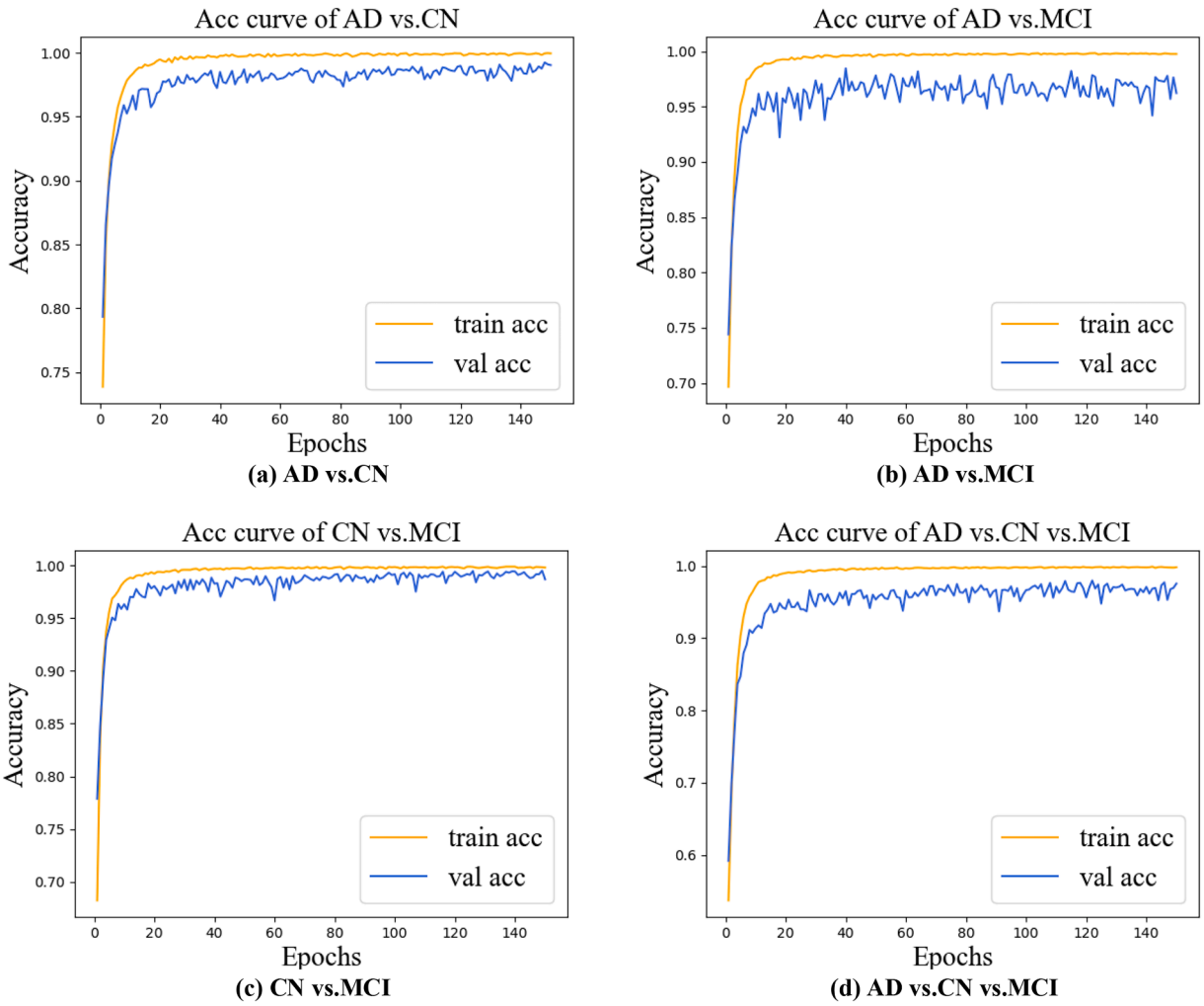

Fig. 6. MACFNet acc curves on training and validation sets.

$$Y_i = Relu(BN(Conv_{C'\to C}(X', k_i)))_{i=1,2,3}, \tag{20}$$

where i denotes different scales, $Y_i \in \mathbb{R}^{H\times W\times C}$ represents the extracted feature maps at different scales, and k_i represents the kernel size at different scales.

In the feature fusion stage, unlike Liu et al. [47] and Ge et al. [49] which proposed aggregating or concatenating features on all scales to fuse multi-scale information, we use the hard attention mechanism to suppress redundant information that comes from each scale. In this paper, 1×1 convolution and Sigmoid as hard attention. For feature map input at different scales, attention is used on $\mathrm{F1}_{\mathrm{scale}} - -\mathrm{F3}_{\mathrm{scale}}$ scales respectively:

$$\begin{cases} A = \sigma(Conv_{C\to 1}(Y_1, k=1)), \\ B = \sigma(Conv_{C\to 1}(Y_2, k=1)), \\ C = \sigma(Conv_{C\to 1}(Y_3, k=1)), \end{cases} \tag{21}$$

where $A, B, C \in \mathbb{R}^{H\times W\times 1}$ represent attention weight maps of different scales. They are element-wise multiplied with the original feature maps at each scale. Finally, multi-scale features are generated by element-wise addition operation:

$$Z = Y_1 \otimes A + Y_2 \otimes B + Y_3 \otimes C. \tag{22}$$

where $Z \in \mathbb{R}^{H\times W\times C}$ denotes the output multi-scale fusion feature map. Different from Liu et al.'s method [47], our method focuses on "important features under different scale receptive fields". Specifically, we employ attention to suppress redundant information before fusion effectively. The pseudo-code is shown in Algorithm 3:

4. Results

This section first describes the experimental environment, parameter settings as well as the evaluation metrics of the model. Secondly, the effectiveness of MACFNet on different classification tasks is analyzed, including ablation experiments and comparison analysis with other multimodal classification methods. Finally, the effectiveness of MACFNet is visualized directly.

4.1. Experimental setup

All experiments were conducted on a workstation equipped with four Nvidia A100 graphics processing units (GPUs), the operating system Ubuntu 20.04, and a total of 160 GB of video memory. We resized the input images to a size of 224×224 and then trained MACFNet on the Pytorch 1.11.0 framework. more detailed settings are as follows: (i) the optimizer uses the Adam optimizer; (ii) the batch size is set to 32; (iii) the loss function uses the CrossEntropy; and (ii) the learning rate is set to 1×10^{-5}.

To validate the effectiveness of the proposed model, drawing on Meng et al. [50] and Tang et al. [51], in this paper, the preprocessed dataset is randomly divided into a training set, a validation set, and a test set in the ratio of 6:2:2. Among them, the training set is used to train the model parameters so that the model can learn the patterns and features

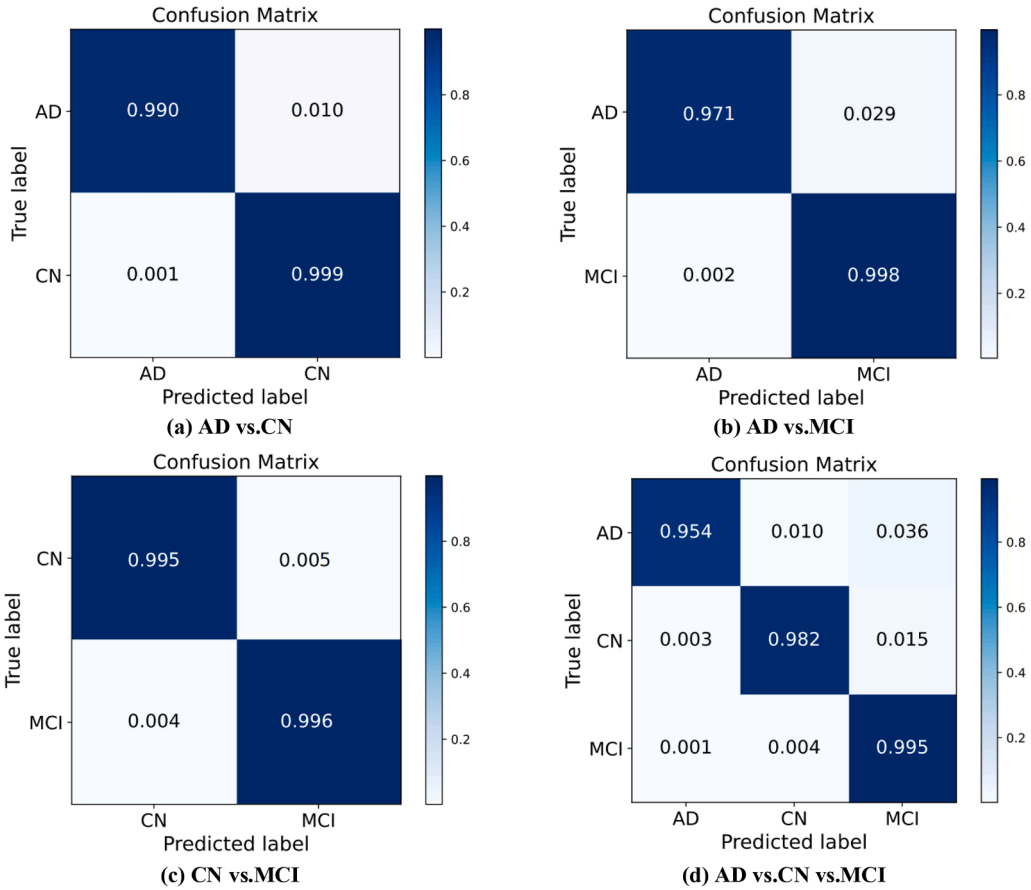

Fig. 7. Confusion matrixes of MACFNet for different classification tasks.

of the data. Secondly, the validation set is used to select the optimal model and tune the hyperparameters; finally, the performance of the model is evaluated using the test set.

4.2. Performance evaluation

To evaluate the effectiveness of MACFNet, several evaluation metrics were computed, including sensitivity, specificity, accuracy, and the area under the curve (AUC).

$$\begin{cases} \text{Accuracy} = \dfrac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{TN}+\text{FP}} \\ \text{Sensitivity} = \dfrac{\text{TP}}{\text{TP}+\text{FN}} \\ \text{Specificity} = \dfrac{\text{TN}}{\text{FP}+\text{TN}} \end{cases} \tag{23}$$

The terms "true positive," "true negative," "false positive," and "false negative" are represented as "TP," "TN," "FP," and "FN," respectively. The ROC (Receiver Operating Characteristic) curve is a graphical tool for evaluating the performance of a classification model. The calculation method for AUC involves integrating the area under the ROC curve, which is used to measure the classifier's ability to classify samples at different thresholds. A higher AUC value reflects better performance of the classifier.

4.3. Ablation experiments

The purpose of this section is to demonstrate the validity of the proposed components (ECSA, CEFM, and MSAG) through ablation experiments, all experiments using the same parameter settings to ensure fairness.

4.3.1. Performance of ECSA block

We use different strategies to verify the effectiveness of ECSA. As shown in Table 4, firstly, the CEFM module containing only SFE and FFE is added to the baseline model ResNet34, and the classification accuracies are 98.91 %, 98.39 %, 99.07 %, and 96.55 % for AD vs. CN, AD vs. MCI, CN vs. MCI and AD vs. CN vs. MCI, respectively. After adding the ECSA module to the CEFM, the classification accuracies are 99.28 %, 98.47 %, 99.47 %, and 97.34 %, respectively. It can be seen that the inclusion of the ECSA module increases the classification accuracy by 0.37 %, 0.08 %, 0.4 % and 0.79 %, respectively. This result proves the effectiveness of ECSA. Specifically, ECSA is able to effectively focus on important information related to AD through spatial attention and efficient channel attention mechanisms to enhance MRI structural features and PET functional features. This enabled the model to capture better the abnormal representation of Alzheimer's disease in brain images, such as grey and white matter regions.

In addition, it can be seen that among the four groups of classification tasks, the classification accuracies of the AD vs. CN, CN vs. MCI tasks are much higher than those of the AD vs. MCI and multiple classification tasks, which is because the AD-related lesion areas do not change significantly at the early stage. Meanwhile, it is difficult for MCI to distinguish AD from CN.

4.3.2. Performance of CEFM module

To verify the effectiveness of CEFM, we added the CEFM module to the baseline model. As shown in Table 5, compared with the baseline model, with the addition of CEFM, the classification accuracy improved by 1.07 %, 0.3 %, and 0.64 % in the binary classification tasks of AD vs. CN, AD vs. MCI, and CN vs. MCI, respectively. In the multiple classification task, the classification accuracy improved by 0.15 %. These results suggest that fusing multilevel low-level features from different

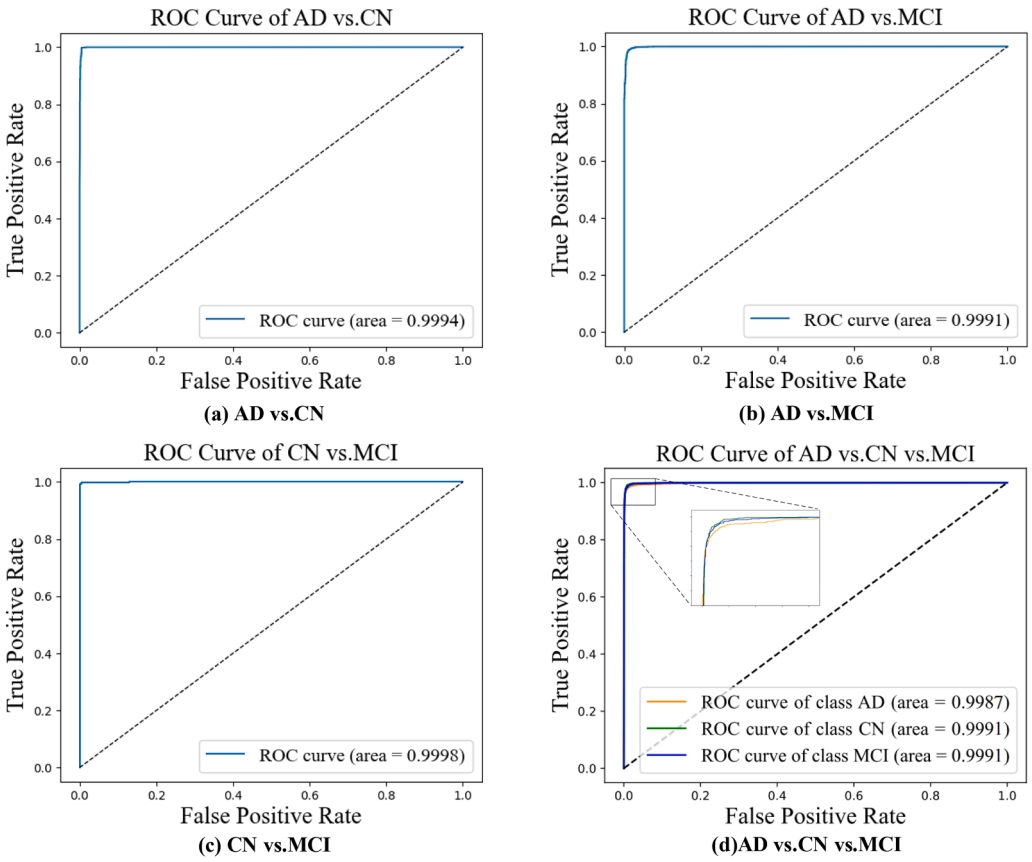

Fig. 8. ROC curves of MACFNet for different classification tasks.

modalities can further enhance the effectiveness of the model for AD diagnosis. The high-level features tend to contain more abstract and semantically rich information, which helps better capture the image's global contextual information. However, low-level features focus on image details and texture information, which is crucial for AD diagnosis. By fusing low-level features from different modalities, the complementary nature of low-level and high-level features can be fully utilised to provide a more comprehensive and enriched feature representation, enhancing the understanding and representation of the model. In addition, different modalities have differentiation, with MRI focusing more on structural information and PET focusing more on functional information. High-level features show less sensitivity to these differences, while low-level features are more sensitive. By integrating multilevel features, the effect of heterogeneity across modalities can be reduced, thus enhancing the generalisation and robustness of the model to multimodal images.

4.3.3. Performance of MSAG block

The MSAG module was added to the baseline model to validate the diagnostic performance of the module in different classification tasks. As shown in Table 5, the classification accuracies of AD vs. CN, AD vs. MCI, CN vs. MCI, and AD vs. CN vs. MCI were improved by 0.33 %, 0.43 %, 0.15 %, and 1.00 %, respectively, with the MSAG module. This shows that multiscale feature extraction can localize the different lesion regions associated with AD and improve classification accuracy. Furthermore, our MSAG module results in higher classification accuracy compared to the CEFM module in the multi-class classification task. This is because the pathological regions associated with AD subtypes may be distributed in multiple ROI regions, and the utilization of different scales of receptive fields can effectively focus on these regions.

4.4. Comparison with other methods

In this section, a comparison is made between the MACFNet and the multimodal methods based on the ADNI database. It includes feature fusion-based methods [16,30,31,52], image fusion-based methods [35,37], data generation-based methods [26,53], and cross-modal interaction-based methods [43,45]. In addition, multiscale-based methods [47] were also compared, as shown in Table 6.

In the AD vs. MCI classification task, MACFNet achieved superior results in the three metrics of sensitivity, accuracy, and AUC compared to the multiscale approach proposed by Liu et al. [47]. This is because they use a diagnostic approach based on unimodal data, whereas MACFNet extracts features from both MRI and PET modalities. It can be inferred that making full use of complementary information from different modalities is beneficial for AD diagnosis. However, Liu et al. [47] reported a specificity of 98.52 %, slightly higher than the MACFNet method. It shows that their ability to recognize negative samples is superior to ours. This is because they proposed a multiscale approach to extract features from ROI, including gray matter (GM) and white matter (WM) regions, reducing the negative impact of redundant information.

4.5. Performance analysis

To analyze the performance of our MACFNet model, we present the accuracy curves of MACFNet on the training and validation sets. In addition, confusion matrices and ROC curves demonstrate the model's classification effectiveness on the test set.

The accuracy convergence curves of MACFNet on the training and validation sets are shown in Fig. 6. It can be seen that MACFNet achieves excellent classification performance. In addition, MACFNet achieves the highest accuracy in the AD vs. CN and CN vs. MCI tasks, and slightly lower accuracy in the AD vs. MCI and multiple classification tasks. This

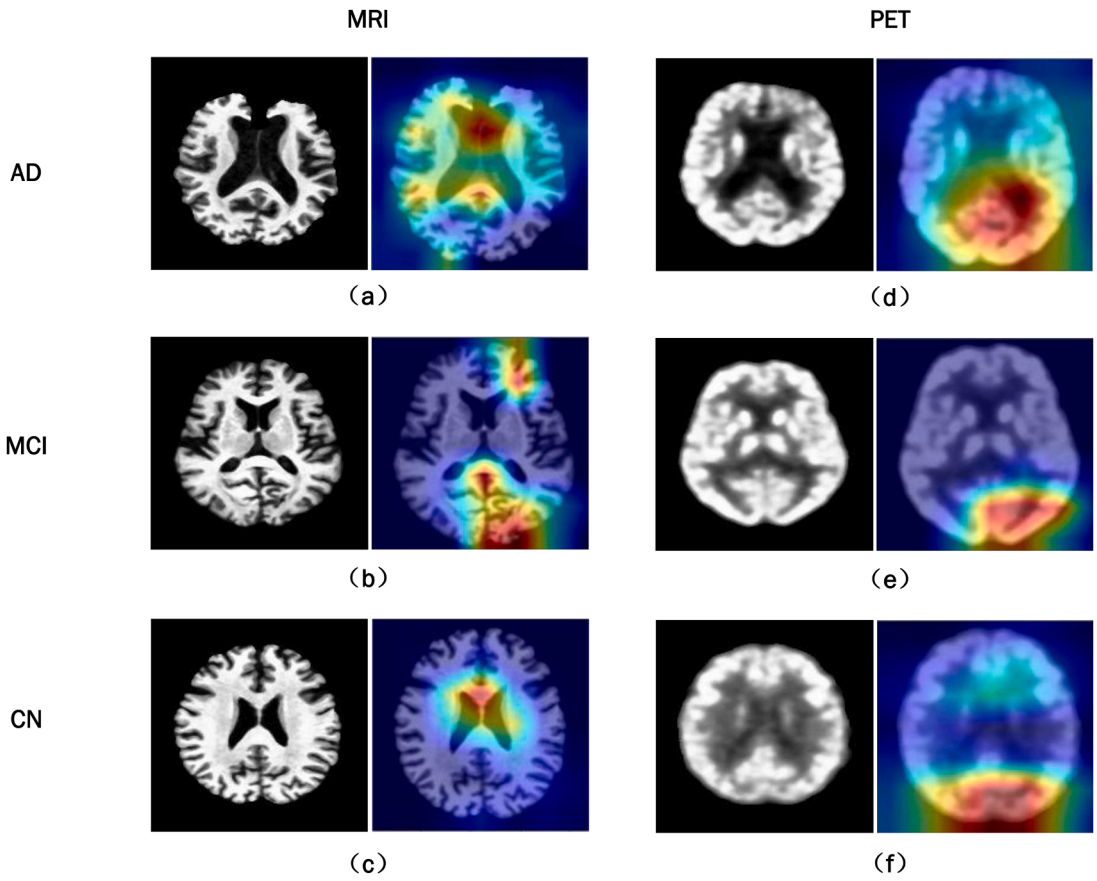

Fig. 9. The visualization of MACFNet using Grad-CAM.

is because MCI is an early stage of AD where brain changes are not obvious.

Fig. 7 illustrates the confusion matrix of MACFNet on the test dataset, revealing that MACFNet achieved the highest classification accuracy for CN, reaching a remarkable 99.9 %.

The ROC curves for different classification tasks are illustrated in Fig. 8. The results demonstrate the exceptional classification performance of our MACFNet across the four classification tasks. The classification performance of AD vs. CN and CN vs. MCI is superior to that of the other two classification tasks, which also indirectly reflects the consistency of MACFNet's classification results on the training and validation sets.

5. Discussion

5.1. Evaluation of MACFNet

As shown in Table 6, our MACFNet achieves the best classification performance in binary classification tasks such as AD and CN, CN and MCI, as well as multi-classification tasks. This can be attributed to several factors below. Firstly, MACFNet not only considers the fusion of different modal high-level features but also pays attention to the interaction between cross-modal low-level features. Secondly, in order to reduce the noise effects from irrelevant channel features, double weighting is performed using ECSA attention. This enables MACFNet to effectively capture functional and structural information related to AD. In addition, by utilizing multi-scale feature fusion in the multimodal approach, MACFNet is also able to capture both local and global contextual information of AD, thus further improving the performance of the model.

However, in the AD and MCI classification tasks, Liu et al.'s model was slightly higher in specificity than MACFNet. This is because brain atrophy in Alzheimer's disease affects ROI regions of the brain, such as grey and white matter. Although the relevant regions were also preprocessed in this study, the ROI regions were not analysed separately. Instead, Liu et al. diagnosed AD by secondary processing the images again through image analysis tools such as Python and Nifti, manually delineating the GM and WM regions, and fusing multi-scale grey and white matter features of the MRI images. San et al. [56] and Mancho et al. [57] similarly demonstrated the importance of ROI regions. These regions are often closely associated with disease progression, and by specifically targeting these regions for analysis, it may be possible to further improve the model's performance. Therefore, we will consider designing automated ROI extraction modules in the future and embedding them into an end-to-end image classification network. This could reduce the complexity of manually classifying ROIs and further improve the generalisation performance of the model.

5.2. Visualization of MACFNet

Among visualisation techniques, Grad-CAM has been widely used as an effective gradient visualisation method for explanatory studies of deep learning models. In diagnosing AD, many studies have used Grad-CAM for visualisation [37,58,59]. With this method, it is possible to directly show the image regions that the model focuses on when making predictions. Therefore, the same Grad-CAM technique is used in this paper. In particular, the work of Lian et al. [60] has been remarkably effective in the analysis of gradient visualisation, and their study provides valuable insight into understanding the model's decision making. Fig. 9 shows the results of the visualization of MACFNet on Grad-CAM. It can be found that MACFNet can accurately focus on the brain lesion areas. As shown in Fig. 9(a), the ROI regions are distributed throughout

the entire brain portion. This is because the entire brain of the subject undergoes atrophy at the AD stage.

As can be seen in Fig. 9(b), the ROI regions of MCI subjects are relatively concentrated. This is because MCI, as a precursor stage of AD, has no obvious brain changes, and brain atrophy occurs in localized areas.

In addition, it can be seen from Fig. 9(d)-(f) that the heat map of the PET image focuses on different areas from that of the MRI. That is mainly caused by the different imaging mechanisms. This result further demonstrates that integrating different modalities enables the acquisition of complementary information, which can help diagnose AD.

6. Conclusion

In this paper, we propose a multimodal CNN network, MACFNet, with a two-branch crossover mechanism for multimodal AD classification. The MACFNet employs a cross-enhanced fusion algorithm based on an efficient attentional mechanism to enhance the structural and functional information of neuroimages of different modalities, and to achieve the fusion and interactions of the multimodal low-level features. In addition, MACFNet can focus on the local and global information related to AD by adopting a multi-scale approach in order to obtain different scales of receptive fields. Experiments on the ADNI database show that our MACFNet achieves better classification performance than existing methods. The classification accuracies reach 99.59 %, 98.85 %, 99.61 %, and 98.23 % for AD vs. CN, AD vs. MCI, CN vs. MCI and AD vs. CN vs. MCI, respectively.

However, the proposed MACFNet only considers the use of multimodal neuroimaging for AD diagnosis, ignoring clinical and biological information. In addition, MACFNet performs feature extraction on the whole image, ignoring features in ROI regions such as GM or WM. In the future, we will consider extracting and analyzing ROI features from neuroimaging data while further optimizing the variants of MACFNet by combining clinical data to improve the model's generalization ability.

CRediT authorship contribution statement

Chaosheng Tang: Writing – original draft, Visualization, Formal analysis, Conceptualization. **Mengbo Xi:** Validation, Software, Resources, Investigation. **Junding Sun:** Writing – original draft, Visualization, Supervision, Software. **Shuihua Wang:** Writing – review & editing, Validation, Investigation, Funding acquisition. **Yudong Zhang:** Writing – review & editing, Validation, Software, Funding acquisition.

Declaration of competing interest

Tche 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 is supported by the National Natural Science Foundation of China(62276092); Key Science and Technology Program of Henan Province (212102310084); MRC (MC_PC_17171); Royal Society (RP202G0230); BHF (AA/18/3/34220); Hope Foundation for Cancer Research (RM60G0680); GCRF (P202PF11).; Sino-UK Industrial Fund (RP202G0289); LIAS (P202ED10, P202RE969); Key Scientific Research Projects of Colleges and Universities in Henan Province(22A520027); Data Science Enhancement Fund (P202RE237).; Fight for Sight (24NN201); Sino-UK Education Fund (OP202006); BBSRC (RM32G0178B8).

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12–2–0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering: AbbVie, Alzheimer's Drug Discovery Foundation; Araclon Biotech; Alzheimer's Association; 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.; Lumosity; Lundbeck; MerckCo., Inc.; Meso Scale Diagnostics, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; NeuroRx Research; Neurotrack Technologies; Takeda Pharmaceutical Company; Novartis Pharmaceuticals Corporation; Pfizer Inc.; and Transition Therapeutics. 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. 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). ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix

Table 7 shows the mathematical symbols used in this paper.

Table 7
Mathematical symbols.

Symbol	Meaning
$+$	Element-wise add
$\times$	Multiplication operation
$\otimes$	Element-wise product
concat	Concatenation operation
max	Maximum value
Min	Minimum value
X	Input feature map
$f(X)$	Output feature map
$\mathbb{R}$	Set of real numbers
H	The height of the feature map
W	The width of the feature map
C	The number of channels of the feature map
(w, h)	Coordinate of pixel

(*continued on next page*)

Table 7 (*continued*)

Symbol	Meaning
k	Kernel size
W_k	Learnable band matrix (k means the number of cross-channels)
ω	Learnable channel weight
ψ	Mapping relation
δ	LeakyRelu activation function
σ	Sigmoid activation function
r	Scaling factor
$C_{in} \rightarrow C_{out}$	Variation in the number of channels of the feature map

References

[1] Z. Hu, Z. Wang, Y. Jin, W. Hou, VGG-TSwinformer: transformer-based deep learning model for early Alzheimer's disease prediction, Comput. Methods Programs Biomed. 229 (2023) 107291.

[2] C. Patterson, "World Alzheimer report 2018," 2018.

[3] B. Lei, et al., Diagnosis of early Alzheimer's disease based on dynamic high order networks, Brain Imaging Behav. 15 (2021) 276–287.

[4] X. Tian, Y. Liu, L. Wang, X. Zeng, Y. Huang, Z. Wang, An extensible hierarchical graph convolutional network for early Alzheimer's disease identification, Comput. Methods Programs Biomed. 238 (2023) 107597.

[5] S. Tomassini, et al., Brain-on-cloud for automatic diagnosis of Alzheimer's disease from 3D structural magnetic resonance whole-brain scans, Comput. Methods Programs Biomed. 227 (2022) 107191.

[6] M. Khojaste-Sarakhsi, S.S. Haghighi, S.F. Ghomi, E. Marchiori, Deep learning for Alzheimer's disease diagnosis: a survey, Artif. Intell. Med. 130 (2022) 102332.

[7] Y. Zhang, X. He, Y.H. Chan, Q. Teng, J.C. Rajapakse, Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans, Comput. Biol. Med. 164 (2023) 107328.

[8] L. Liu, S. Liu, L. Zhang, X.V. To, F. Nasrallah, S.S. Chandra, Cascaded multi-modal mixing transformers for Alzheimer's disease classification with incomplete data, Neuroimage 277 (2023) 120267.

[9] N. Rahim, S. El-Sappagh, S. Ali, K. Muhammad, J. Del Ser, T. Abuhmed, Prediction of Alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data, Inf. Fusion 92 (2023) 363–388.

[10] S.M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, M.K. Khan, Medical image analysis using convolutional neural networks: a review, J. Med. Syst. 42 (2018) 1–13.

[11] Y. Liu, X. Chen, J. Cheng, H. Peng, A medical image fusion method based on convolutional neural networks, in: 2017 20th international conference on information fusion (Fusion), IEEE, 2017, pp. 1–7.

[12] M. Wang, W. Shao, S. Huang, D. Zhang, Hypergraph-regularized multimodal learning by graph diffusion for imaging genetics based Alzheimer's disease diagnosis, Med. Image Anal. 89 (2023) 102883.

[13] G. Martí-Juan, M. Lorenzi, G. Piella, A.s.D.N. Initiative, MC-RVAE: multi-channel recurrent variational autoencoder for multimodal Alzheimer's disease progression modelling, Neuroimage 268 (2023) 119892.

[14] T. Zhang, M. Shi, Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease, J. Neurosci. Methods 341 (2020) 108795.

[15] J. Zhang, X. He, Y. Liu, Q. Cai, H. Chen, L. Qing, Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data, Comput. Biol. Med. 162 (2023) 107050.

[16] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, P. Zhang, Alzheimer's disease diagnosis via multimodal feature fusion, Comput. Biol. Med. 148 (2022) 105901.

[17] B. Rajalingam, R. Priya, R. Bhavani, Multimodal medical image fusion using hybrid fusion techniques for neoplastic and Alzheimer's disease analysis, J. Comput. Theor. Nanosci. 16 (4) (2019) 1320–1331.

[18] Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, D. Shen, Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis, in: Medical Image Computing and Computer Assisted Intervention–MICCAI2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11, Springer, 2018, pp. 455–463.

[19] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern. Anal. Mach. Intell. 40 (4) (Apr 2018) 834–848.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI conference on artificial intelligence 31, 2017.

[21] A. Khan, A. Chefranov, H. Demirel, Image scene geometry recognition using low-level features fusion at multi-layer deep CNN, Neurocomputing. 440 (2021) 111–126.

[22] O. Camara, et al., Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal Alzheimer's disease images, Neuroimage 42 (2) (2008) 696–709.

[23] J.L. Whitwell, et al., Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study, Lancet Neurol. 11 (10) (2012) 868–877.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern. Anal. Mach. Intell. 40 (4) (2017) 834–848.

[25] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images, Sci. Rep. 8 (1) (2018) 5697.

[26] X.Y. Gao, F. Shi, D.G. Shen, M.H. Liu, Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease, IEEE J. Biomed. Health Inform. 26 (1) (Jan 2022) 36–43.

[27] M. Odusami, R. Maskeliūnas, R. Damaševičius, S. Misra, Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images, J. Med. Biol. Eng. (2023) 1–12.

[28] Q. Zhu, et al., Deep Multi-modal discriminative and interpretability network for Alzheimer's disease diagnosis, IEEE Trans. Med. ImAging (2022).

[29] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, N. Jacobs, Advit: vision transformer on multi-modality pet images for Alzheimer disease diagnosis, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE, 2022, pp. 1–4.

[30] M. Abdelaziz, T.F. Wang, A. Elazab, Fusing Multimodal and anatomical volumes of interest features using convolutional auto-encoder and convolutional neural networks for Alzheimer's disease diagnosis, Front. Aging Neurosci. 14 (Apr 2022). Art no. 812870.

[31] Y. Shi, et al., ASMFS: adaptive-similarity-based multi-modality feature selection for classification of Alzheimer's disease, Pattern. Recognit. 126 (Jun 2022). Art no. 108566.

[32] H. Chen, et al., Multimodal predictive classification of Alzheimer's disease based on attention-combined fusion network: integrated neuroimaging modalities and medical examination data, IET. Image Process. 17 (11) (2023) 3153–3164.

[33] Y. Dai, et al., DE-JANet: a unified network based on dual encoder and joint attention for Alzheimer's disease classification using multi-modal data, Comput. Biol. Med. 165 (2023) 107396.

[34] G. Zhang, et al., A multimodal fusion method for Alzheimer's disease based on DCT convolutional sparse representation, Front. Neurosci. 16 (2023) 1100812.

[35] W.N. Ismail, F.R. PP, M.A. Ali, A meta-heuristic multi-objective optimization method for Alzheimer's disease detection based on multi-modal data, Mathematics 11 (4) (2023) 957.

[36] L. Kang, J. Jiang, J. Huang, T. Zhang, Identifying early mild cognitive impairment by multi-modality MRI-based deep learning, Front. Aging Neurosci. 12 (2020) 206.

[37] J. Song, J. Zheng, P. Li, X. Lu, G. Zhu, P. Shen, An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis, Front. Digit. Health 3 (2021) 637386.

[38] W. Lin, et al., Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease, Front. Neurosci. 15 (2021) 646013.

[39] H. Ye, Q. Zhu, Y. Yao, Y. Jin, D. Zhang, Pairwise feature-based generative adversarial network for incomplete multi-modal Alzheimer's disease diagnosis, Vis. Comput. (2022) 1–10.

[40] C. Tang, M. Wei, J. Sun, S. Wang, Y. Zhang, A.s.D.N. Initiative, CsAGP: detecting Alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling, J. King Saud Univ.-Comput. Inf. Sci. (2023) 101618.

[41] X. Hao, et al., Multi-modal self-paced locality preserving learning for diagnosis of Alzheimer's disease, IEEE Trans. Cogn. Dev. Syst. (2022).

[42] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 689–696.

[43] M. Golovanevsky, C. Eickhoff, R. Singh, Multimodal attention-based deep learning for Alzheimer's disease diagnosis, J. Am. Med. Inform. Assoc. 29 (12) (2022) 2014–2022.

[44] J. Pan and S. Wang, "Cross-modal transformer GAN: a brain structure-function deep fusing framework for Alzheimer's disease," *arXiv preprint arXiv:2206.13393,* 2022.

[45] Y. Leng, et al., Multimodal cross enhanced fusion network for diagnosis of Alzheimer's disease and subjective memory complaints, Comput. Biol. Med. 157 (2023) 106788.

[46] S. Miao, et al., MMTFN: multi-modal multi-scale transformer fusion network for Alzheimer's disease diagnosis, Int. J. ImAging Syst. Technol 34 (1) e22970.

[47] Z.B. Liu, H.X. Lu, X.P. Pan, M.C. Xu, R.S. Lan, X.N. Luo, Diagnosis of Alzheimer's disease via an attention-based multi-scale convolutional neural network, Knowl. Based. Syst. 238 (2022). FebArt no. 107942.

[48] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.
[49] 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.
[50] X. Meng, et al., Multi-modal neuroimaging neural network-based feature detection for diagnosis of Alzheimer's disease, Front. Aging Neurosci. 14 (2022) 911220.
[51] C. Tang, M. Wei, J. Sun, S. Wang, Y. Zhang, A.s.D.N. Initiative, CsAGP: detecting Alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling, J. King Saud Univ.-Comput. Inf. Sci. 35 (7) (2023) 101618.
[52] X.S. Fang, Z.B. Liu, M.C. Xu, Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis, IET. Image Process. 14 (2) (Feb 2020) 318–326.
[53] G. Zheng, et al., A transformer-based multi-features fusion model for prediction of conversion in mild cognitive impairment, Methods 204 (2022) 241–248.
[54] F. Zhang, Z. Li, B.Y. Zhang, H.S. Dua, B.J. Wang, X.H. Zhang, Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease, Neurocomputing. 361 (Oct 2019) 185–195.
[55] P.H.HAN Kun, Z.H.A.N.G. Wei, B.I.A.N. Xiaofei, C.H.E.N. Chunling, H.E. Shuning, Alzheimer's disease classification method based on multi-modal medical images, J. Tsinghua Univ. 60 (8) (2020) 664–671, 682, 2020-08-15.
[56] Á. Bernabéu-Sanz, J.V. Mollá-Torró, S. López-Celada, P.Moreno López, E. Fernández-Jover, MRI evidence of brain atrophy, white matter damage, and functional adaptive changes in patients with cervical spondylosis and prolonged spinal cord compression, Eur. Radiol. 30 (2020) 357–369.
[57] N. Mancho-Fora, et al., Network change point detection in resting-state functional connectivity dynamics of mild cognitive impairment patients, Int. J. Clin. Health Psychol. 20 (3) (2020) 200–212.
[58] 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 Inform. 26 (11) (2022) 5665–5673.
[59] 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 Inform. 26 (11) (2021) 5289–5297.
[60] C. Lian, M. Liu, Y. Pan, D. Shen, Attention-guided hybrid network for dementia diagnosis with structural MR images, IEEE Trans. Cybern. 52 (4) (2020) 1992–2003.