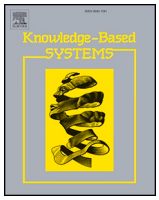
Knowledge-Based Systems 281 (2023) 111064

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



An end-to-end multimodal 3D CNN framework with multi-level features for the prediction of mild cognitive impairment

Yanteng Zhang [a], Xiaohai He [a,*], Yixin Liu [b], Charlene Zhi Lin Ong [c], Yan Liu [d], Qizhi Teng [a]

[a] College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
[b] National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu 610041, China
[c] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
[d] Department of Neurology, The Third People's Hospital of Chengdu, Chengdu 610031, China

ARTICLE INFO

Keywords:
Brain image
Multimodal framework
3D CNN
Multi-level features
AD diagnosis

ABSTRACT

In recent years, deep learning methods based on brain image have been used for the diagnosis of cognitive impairment-related disorders. With the development of neuroimaging techniques, multi-modality image such as structural magnetic resonance imaging (sMRI) and positron emission tomography (PET) reflect structural and functional information of the brain respectively, and provide more techniques for the diagnosis of cognitive impairment diseases. Combining these complementary image features can lead to more accurate diagnostic assessments compared to using a single modality. Therefore, how to effectively combine multi-modality image features to realize the diagnosis of cognitive impairment disease needs to be further explored. In this work, we propose an end-to-end multimodal 3D CNN framework based on ResNet architecture, which integrates multi-level features obtained under the role of attention mechanisms to better capture subtle differences among brain images, and achieves remarkable diagnostic performance through spatial pyramid pooling strategy and effective fusion of multi-modality features. In this process, we demonstrate that the multimodal framework is more effective by means of non-shared parameters for multi-modality features learning. Moreover, the visualized attention maps show that our model can focus on important brain regions relevant to disease diagnosis. The experimental results demonstrated that our method improved the diagnostic performance in AD diagnosis and MCI conversion prediction by 6.37 % and 3.51 % compared to the single modality, and it also outperformed some recent state-of-the-art multimodal methods. Especially in AD diagnosis achieved an average accuracy of 94.61 %, which provides a more feasible technology for diagnostic assessment of patients with AD.

1. Introduction

As global aging continues to accelerate, the number of people with cognitive impairment-related disorders is increasing every year. Alzheimer's disease (AD), as the most common symptom of brain disorders, is a degenerative disease of the central nervous system characterized by memory, language, cognitive and even emotional impairment, which has a significant impact on the daily life of patients [1]. In the early stages of AD, it is known as mild cognitive impairment (MCI). MCI has not yet reached the severity of dementia and will not affect the daily lives of patients, but more than one-third of mild cognitive impairments slowly progress to AD within 5 years [2]. However, at present, there is no specific drug or treatment option that can cure people with AD. If appropriate medical intervention is provided in the early stages of AD or even MCI, it can alleviate the continuous deterioration of the disease and improve the quality of life of patients [3]. Therefore, achieving early prediction of AD has become particularly important, but it also remains a very challenging issue in the clinical practice. In recent years, with the development of artificial intelligence techniques, some novel computer-aided methods have been proposed to diagnosis the cognitive disorders from brain images [4].

Currently, the clinical diagnosis of AD requires a comprehensive assessment based on neuropsychological assessment, cognitive-behavioral assessment, medical imaging and other ancillary tests [5]. Imaging is a relatively convenient and reliable diagnostic aid that plays an important role in detecting and confirming Alzheimer's disease, and the results of imaging examinations can assist physicians in improving the accuracy of Alzheimer's disease diagnosis to a certain extent [6].

* Corresponding author.
E-mail address: hxh@scu.edu.cn (X. He).

https://doi.org/10.1016/j.knosys.2023.111064
Received 17 April 2023; Received in revised form 14 September 2023; Accepted 5 October 2023
Available online 6 October 2023

With the development of medical imaging technology, a variety of neuroimaging techniques have emerged, including structural imaging (sMRI, DTI, etc.) and functional imaging (PET, fMRI, etc.). sMRI, as a non-invasive and radiation-free structural imaging, is currently the most widely used imaging tool in clinical practice. Doctors can observe some relatively obvious structural changes in the brain through sMRI to assess AD, and these changes mainly focus on cerebral cortex thinning, ventricular enlargement, and hippocampal atrophy. In functional imaging, FDG PET is a molecular examination modality to visualize glucose metabolism by injecting radioactive elements into the body to analyze the metabolic function of the brain, and its imaging mainly shows reduced glucose metabolism in the posterior cingulate gyrus and parietal temporal lobe cortical regions [7]. PET reflects functional metabolic changes in the brain different from structural imaging, and functional imaging is more advantageous in reflecting of some pathophysiological mechanisms and even in the accurate assessment [8]. Both of the above brain imaging techniques provide more basis for screening and early diagnosis of AD.

The accumulation of medical image data and the development of artificial intelligence technology provide opportunities and challenges for computer-aided diagnosis of AD. With the support of deep learning methods, through learning general features by 3D CNNs acting directly on brain images can realize the tasks of AD screening and assisted diagnosis [9,10,11]. Like CNN-based approach can provide better nonlinear representation than traditional machine learning methods and also eliminates the complex and time-consuming process of manual feature extraction. Some of these studies [12,13] combined with attention mechanisms and hybrid strategies based on CNN-backbone to achieve improved diagnostic performance. From the above and currently researched studies [14], the current methods for image-based AD diagnosis mainly utilize sMRI, but the information provided by single modality image is limited due to factors such as neuroimage manifestation of cognitively impaired patients with non-specific lesions and relatively subtle structural differences among subjects. Therefore, in the case of some patients with low specificity of neuroimage, single modality-based AD assessment remains difficult to realize.

Compared with single modality information, multi-modality brain images provide richer and more comprehensive complementary information. In neuroimaging, both sMRI and PET belong to the 3D volumetric image. PET simulates the metabolic changes in brain regions, while sMRI reflects the morphological changes of brain regions. For this reason, complementary and more comprehensive image features can be obtained by combining two different pathological manifestations, which can help to achieve more accurate AD screening and diagnosis. Huang et al. [15] a 3D CNN with VGG as the backbone to integrate multimodal MRI and PET information for AD diagnosis and prognosis. Lin et al. [16] utilized a reversible GAN to solve the problem of missing PET data and proposed a multimodal 3D CNN architecture for AD diagnosis, which achieved better diagnostic performance than using single modality. Aviles-Rivero et al. [17] proposed a semi-supervised hypergraph learning framework combining multi-modality MRI and PET with gene data for AD diagnosis, which considers higher order relations among multimodal data. These studies have shown that an improved diagnostic performance can be achieved by combining multi-modality sMRI and PET, which has gradually become the focus of current research in computer-aided diagnosis of AD. The study [18] also indicated that multimodal methods and biomarkers combined will lead to more accurate diagnosis on deep learning-based methods in the future work.

As can be seen from the summary and review of the research in this area [19], most of the studies so far are still dominated by single modality methods. Although some multimodal methods are exemplified in the previous narrative, there are still some issues that need to be further solved and explored in the trend of multimodal methods. First, most of the CNN-backbone models are trained on natural images, not modeled for the feature representations of brain images [20,21]. For example, adopting VGG or ResNet backbone to extract features of brain image directly for AD diagnosis in fusing multimodal methods, which is not yet able to obtain specific features for the characteristics of the brain image itself. Secondly, brain atrophic and metabolic changes in cognitively impaired patients are often brain-wide without specific focal regions respectively, it makes a challenge to the effectiveness of two modality combinations and the performance improvement of some multimodal methods limited. In multimodal methods, many studies have achieved diagnosis only by concatenating extracted multimodal features [15,20,22], these approaches have not been able to fully utilize the advantages of complementary multi-modality image information. Narazani et al. [23] further explored multimodal methods through feature combination and image-level fusion, they found that the diagnostic performance of multimodal methods may not yet outperform that of PET methods. And the improvement of some multimodal methods is not obvious relative to PET modality, which is also the issues in current multimodal methods. To sum up, how to effectively utilize modality images for multimodal methods to achieve better AD diagnostic performance is the key of this area.

To this end, for two modality sMRI and PET, this study will construct a more effective diagnosis network and explore a combination method based on multi-modality image to achieve more accurate AD diagnostic performance. In clinical knowledge, structural and metabolic imaging features during the progression of cognitive normal to AD dementia will change differently. Structural changes primarily include the atrophy of brain gray matter (GM) [24], which is generally brain-wide and not limited to the hippocampus and amygdala. Such whole-brain changes are also present in metabolic features, including reduced glucose metabolism in the posterior cingulate gyrus and right parietal lobe [8,9], this is one of the reasons why cannot diagnose AD through specific brain regions alone. Based on the properties of neuroimages, in this work, we design an end-to-end 3D CNN framework that integrates 3D attention mechanisms and multi-layer feature fusion strategies for realizing AD diagnosis and MCI prediction tasks based on multi-modality brain images. Specifically, we implemented AD diagnosis using sMRI GM and PET as multimodal images and compared them with single modality method, feature combination approach, decision fusion approach, and some current studies, respectively, which demonstrated the effectiveness of our proposed multimodal method and achieved remarkable improvement in both AD diagnosis and MCI conversion prediction. In addition, we compare the AD diagnosis results of two modality images and analyze the effect of shared parameter training in multimodal approaches.

In summary, the main contributions of our work are as follows. (1) Considering the characteristics of brain images, to reduce the loss of effective information in convolutional feature extraction, we integrated multi-level features in the fully connected layer under the attention mechanism. (2) We implemented feature extraction by a twin-based network based non-shared parameters training for multi-modality images separately. (3) We adopted the strategy of spatial pyramid pooling to achieve down-sampling for fused features, that making features more robustness. (4) We integrated the strategy of feature combination and decision fusion for final diagnostic prediction. (5) We visualized attention maps show that our model can focus on important brain regions relevant to disease diagnosis and analyzed the advantages of our multimodal method. (6) We obtained improved results in the challenging prediction task of MCI conversion, which is more clinically relevant for the early screening of AD.

2. Dataset and materials

The public neuroimaging data used in this work were all from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1] database [25]. This study used T1-weighted MPRAGE structural MRI and

[1] http://adni.loni.usc.edu

18-fluorodeoxyglucose PET (FDG-PET) image (six 5-min frames 30–60 min post injection) data from the ADNI baseline for AD assessment, acquiring paired multimodal images from the same subject and from the closest acquisition date. MRI and PET acquisition was according to the ADNI acquisition protocol [26], more detailed information can be found at www.adni-info.org. In total, our dataset has 850 subjects including 215 AD, 246 NC (Normal Contral) and 389 MCI. We further divided the acquired MCI category into 151 pMCI (progressive MCI) and 238 sMCI (stable MCI), MCI subjects who developed AD within 3 years were labeled as pMCI and those who did not convert to AD were labeled as sMCI. The subjects included male and female, ranging in age from 61 to 87 years. Table 1 shows the demographic statistics for each category of subjects in our study. We divided the dataset by subject-level split way according to the number of subjects in the ratio of 70 %, 15 % and 15 %, namely the first n-1 numbered subjects used for training, and half of the subjects after n used for validation and the other half for test. Fig. 1 shows two ways of splitting the dataset. All subjects in our study are not from the same person, which avoids inability to accurately test diagnostic performance due to data leakage like the way of slice-level split [27,28].

This work used conventional procedures for brain image preprocessing, correction, affine registration. Specifically, all sMRI images were preprocessed by anterior commissure-posterior commissure correction and affine alignment by the SPM.[2] Where the N4 algorithm [29] was applied to correct the non-uniform tissue intensities, then the sMRI images were performed affine alignment to MNI152 space [30] with the normalized template. Final we used the SPM CAT12 to extract GM tissue from the preprocessed sMRI. For FDG-PET images, they were co-registered according to their corresponding N4 bias-corrected sMRI images on Clinica platform [31,32]. The resolution of both preprocessed brain images was 121 × 145 × 121. Fig. 2 shows the three sectional views of brain images preprocessing.

3. Methods

In this study, we propose an end-to-end multimodal 3D CNN framework for early diagnosis of AD. Firstly, to address the characteristics of brain images, we designed a feature extraction sub-network based on the ResNet [33] architecture, which integrates multiple layers of features under the role of attention mechanism to better capture the weak changes in brain images and further improve the performance of feature extraction. Secondly, we use twin-network to feature extract and learn features by the way of non-shared parameters for sMRI and PET images, respectively, and then use the strategy of spatial pyramid pooling to achieve dimensionality reduction for fused features. Finally, AD diagnosis based on multi-modality images is achieved by feature combination and decision fusion. Our proposed multimodal 3D CNN framework is shown in Fig. 3.

3.1. 3D ResNet architecture

Different from traditional methods of manually extracting features such as cerebral cortex thickness and brain volume, 3D CNN directly acts on brain images to learn general features, and eliminating the complex process of traditional manual feature extraction. In order to effectively encode the spatial information in brain images, we utilize 3D ResNet18 as the network architecture, which has achieved remarkable success in medical image tasks [34,35]. The network consists of a 7 × 7 × 7 down-sampling layer, four convolutional layers, pooling layer, fully connected layer, and softmax layer. In the convolutional layers, the size of the kernel is 3 × 3 × 3. The number of filters in the convolution layers is 64, 128, 256, and 512.

Table 1
The demographic information of dataset used in this study.

	Numbers	Gender(M/F)	Age(yrs)
AD	215	126/89	74.9 ± 7.7
NC	246	125/121	74.1 ± 5.8
sMCI	238	135/103	72.5 ± 7.4
pMCI	151	89/62	74.4 ± 7.1

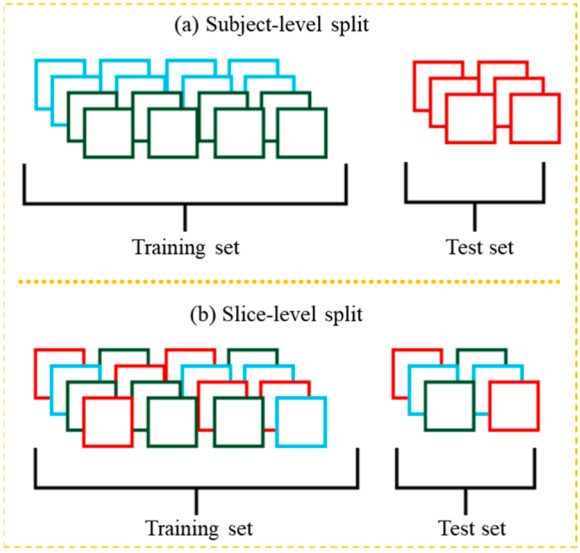


Fig. 1. Two ways of splitting the dataset, where the same color represents the image data from the same subjects.

3.1.1. Attention block

In order to pay more attention to the brain area in feature extraction, we designed the attention block to be applied in our network for brain image. In the neuroimaging of AD, there is general no lesion region, but rather morphological or metabolic changes in multiple brain regions. The spatial attention [36] is focused on a local region in space. For this reason, we designed the attention block to adapt to 3D brain images. The feature maps generated from the convolutional layer are fed to the attention block. The input feature maps are expressed as $M = [M_1, \cdots, M_C]$, where $M_i \in R^{H\times W\times D}$ $(i = 1, 2, \ldots, C)$ represents the feature map of the ith channel, and C represents the number of channels. Then, we perform cross-channel average pooling and max pooling on M to generate two feature channels, M_{avg} and M_{max} respectively. The M_{avg} and M_{max} share the same multi-layer perceptron to learn the dependency between channels. Finally, the weight coefficients are obtained by the sigmoid function as the nonlinear activation, and calculate the final attention maps A. Then the attention computation can be expressed as $A = \sigma(W_1(W_0M_{\text{avg}}) + W_1(W_0M_{\text{max}}))$.

The architecture of the attention block is shown in Fig. 4. Compared with the previous SE-Net attention [37], our attention block enhances the information interaction of channels and improves the overall feature representation. The attention block can be seamlessly integrated into our 3D CNN for end-to-end training. As mentioned above, the essence of our attention mechanism lies in modeling the importance between each feature, and once the weights of each feature channel are obtained, the weights are applied to each of the raw feature channels. In this way, some key feature maps of brain image in A are attentioned under the effect of weights when network training in AD diagnosis task.

3.1.2. Multi-level features concatenation

Different from other diseases, such as cognitive impairment, there is no specific target regions in brain. Moreover, the spatial information of

[2] http://www.fil.ion.ucl.ac.uk/spm

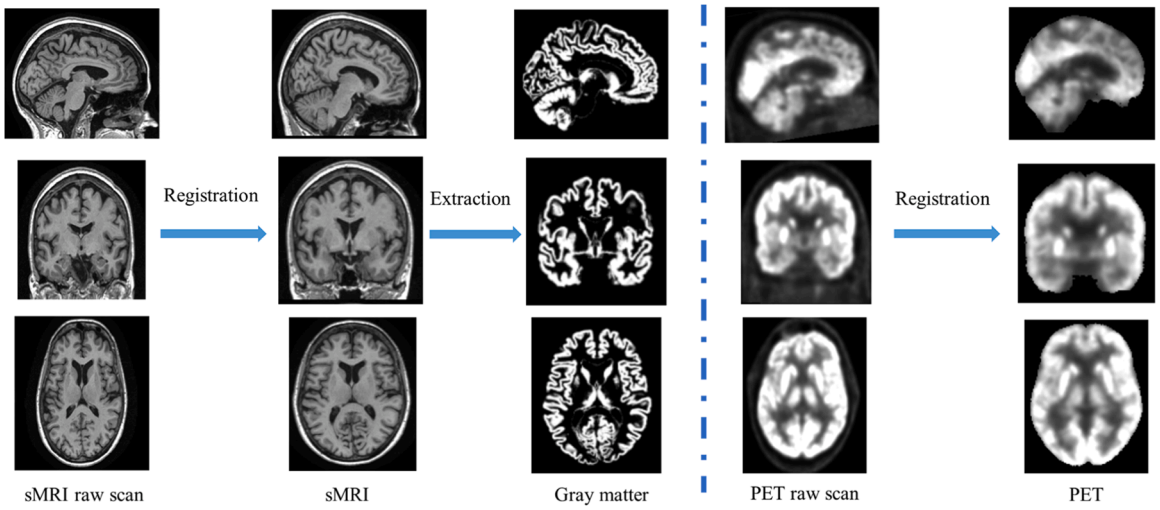


Fig. 2. The raw brain images were aligned to MNI152 space. The left and right sides show the three sectional views of sMRI and PET images respectively.

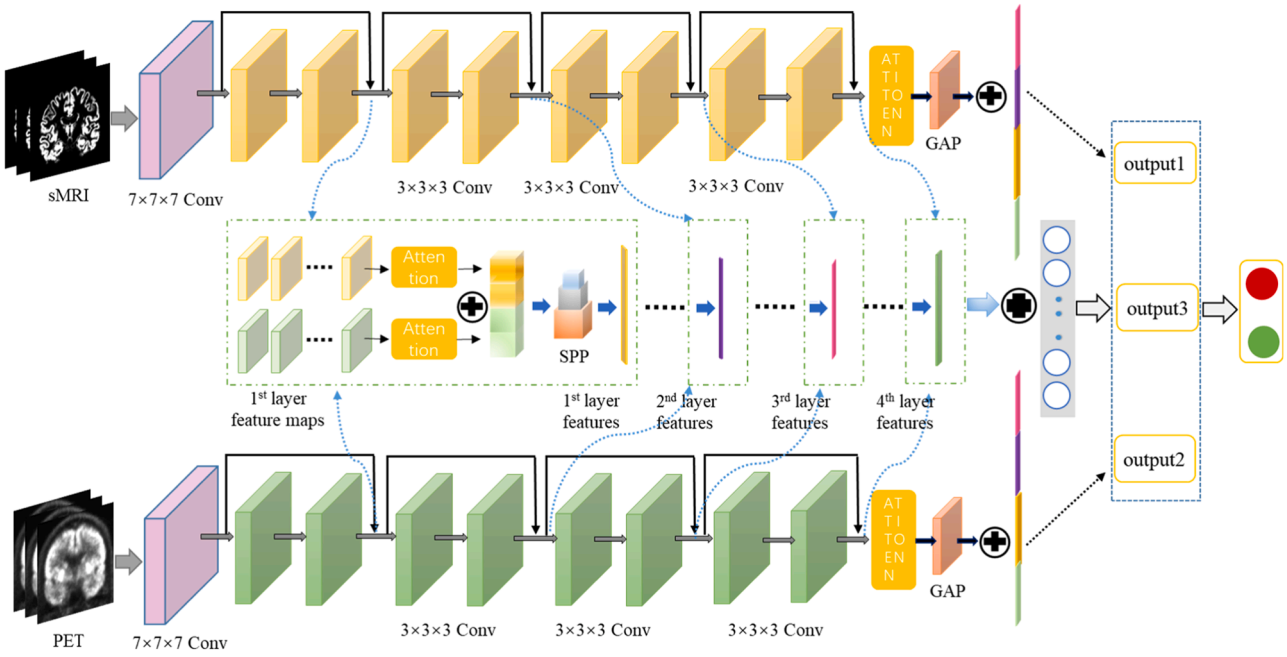


Fig. 3. Our end-to-end multimodal 3D CNN framework based on sMRI and PET images. The inputs of the network are sMRI GM and PET images respectively, and the trained multimodal network directly outputs the final diagnostic results.

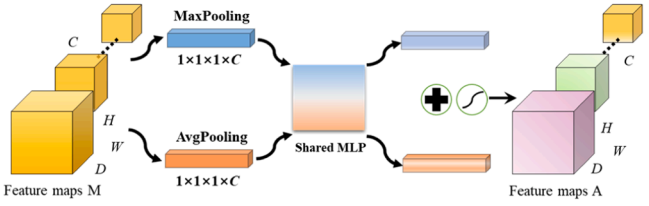


Fig. 4. The architecture of the attention block in our network.

brain image is complex, and the differences between individuals are not obvious, which makes the role of attention mechanism in brain images relatively limited. To consider these clinical characteristics of brain image, we integrate low-level features and high-level features of the convolutional layer in the last fully connected layer in order to reduce information loss in feature extraction. Fig. 5 shows the principle of integrated multi-level features in the sub-network of multimodal framework.

Throughout the sub-network, in order not to affect the acquisition of raw low-level features and raw high-level features of brain image, our attention block is not used directly in the convolutional layers, but acts independently on the brain feature maps of each convolutional layer, the raw feature maps in the network continue into the next convolutional layer. The global average pooling (GAP) is performed for the attention-based features obtained after four convolutional layers, and the multi-level features obtained after GAP are concatenated and then flattened into a 960 one-dimensional vector as the input of the fully connected layer. This sub-network has its backbone as 3D ResNet18 followed by a fully connected layer with softmax function to predict the disease categories in expression classification task.

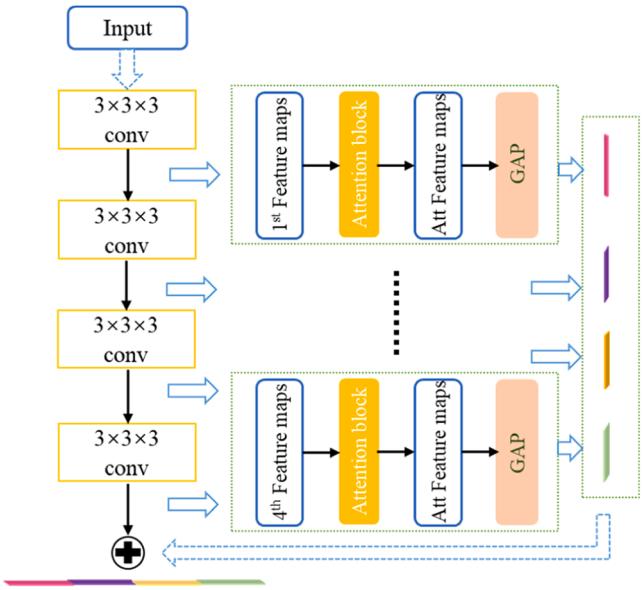


Fig. 5. The structure of integrated multi-level features in sub-network.

3.2. Multimodal network framework

Our multimodal network framework is a Twin-network structure, which contains of two same sub-network MAResNet18 but each network learns features through its own parameters. After multi-modality features acquisition and fusion, dimensionality reduction is achieved using spatial pyramid pooling strategy [38,39]. Finally, we combined with feature combination and decision fusion to achieve the final output prediction.

3.2.1. Twin-network for multimodal features learning

The performance of the feature acquisition is often enhanced by the learning approach of siamese-network with shared parameters in some multimodal studies, and the different presentation of cognitive disorders in sMRI and PET images makes the network shared parameters ineffective in learning different modality features. It has been shown that the diagnostic performance of sMRI and PET multimodal features is not effectively improved by the shared parameters approach [23]. In order to efficiently acquire the features of multimodal images, we acquire the features of each image separately in the form of twin-network (without shared parameters), which means the two sub-networks have the same structure but learn different image features by their respective network parameters. Fig. 6 shows two types of multimodal networks for features learning.

3.2.2. Integration with feature combination and decision fusion

Different from the traditional multimodal methods, our framework integrates the advantages of feature combining and decision fusion methods. First, the output feature maps are combined after every convolutional layer of the twin residual networks, which means that the feature maps of both modalities are combined with the corresponding features after four convolutional layers. Second, since our network architecture is containing of a twin-network, the output of the two sub-networks and the output of the combined features are fused in the final decision stage for the final prediction. In this way, our multimodal framework implements further decision making by learning the independent modality and the combination of different modalities.

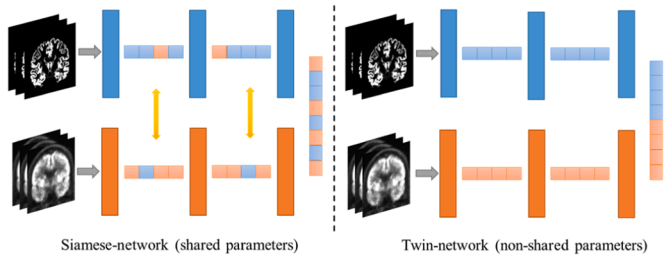


Fig. 6. Two types of multimodal networks for features learning.

3.2.3. Spatial pyramid pooling for combined features

Multi-modality images provide a diversity of complementary information with better specificity through feature combination. How to downscale and acquire multimodal features effectively is the key to improve the diagnostic performance. Spatial pyramid pooling (SPP) has a remarkable result in some medical imaging studies by reducing the dimensionality of multimodal features while retaining more information [39]. In both two sub-networks, the feature maps of each convolutional layer will generate new feature maps through the attention block, we perform 3D SPP after concatenating the two modality feature maps of each layer. The three scale feature sizes after SPP are $4 \times 4 \times 4$, $2 \times 2 \times 2$ and $1 \times 1 \times 1$, then the flattened features are down sampled by 1D convolution to the dimension of output feature maps in the corresponding convolutional layer, i.e., 128, 256, 512, and 1024, respectively (the concatenated features from two sub-networks). Fig. 7 illustrates the principle of 3D SPP block used for the combined feature maps. As can be seen through the principle of SPP action in Fig. 7, SPP turns one pooling into multiple scale pooling, and then integrates the multi-scale features obtained after pooling. Adopting different sizes of pooling windows to act on the feature map realizes the downsampling of the combined multimodal features while considering the multi-scale information in feature representation stage [40], which makes the concatenated pooling features more robust and makes the combination of the multi-modality image features with complementary properties more effective.

3.2.4. Decision output of the network

There are three softmax functions at the final output of the network, corresponding to the output of the fully connected layer of the two sub-networks and the output of the fully connected layer after feature combination. The softmax function for N class probabilities of the output layer is as follows:

$$\mathrm{softmax}(z_j) = \frac{\exp(z_j)}{\sum_{j=1}^{N}(\exp(z_j))} \tag{1}$$

where z_j in the above (13) represents the jth value in the last output $N \times 1$ vector of the network. N is the number of categories, the calculated softmax(z_j) value is between (0, 1).

The three output probabilities are given weights to achieve decision fusion and the final output is expressed as:

$$\mathrm{softmax}(z)_{final} = a \times \mathrm{softmax}_{\mathrm{m}} + b \times \mathrm{softmax}_{\mathrm{p}} + c \times \mathrm{softmax}_{\mathrm{c}} \tag{2}$$

where $\mathrm{softmax}_{\mathrm{m}}$ represents the output probability from the MRI network

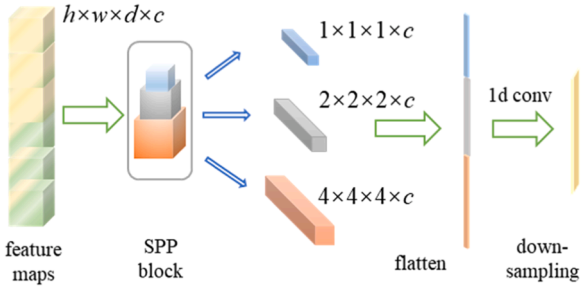


Fig. 7. The principle of SPP block used for the fused feature maps.

branch, softmax_p represents the output probability from the PET network branch, and softmax_c represents the output probability from the combined features branch. *a, b, c* are the weighting coefficients and $a + b + c = 1$ ($a = b$), which we set to 0.2, 0.2, 0.6 respectively in our multimodal network.

4. Experimental results and prospect

4.1. Experimental settings

Our experiments were implemented in PyTorch platform and ran on a Windows x86–64 computer equipped with a GPU NVIDIA P100 16 G. We followed the same data split strategy to construct three different subsets with different test sets, the average of three classification accuracies is used as the final result. The purpose of this is to avoid some subjects with distinct imaging features appearing in a particular test set, since no fixed subjects are formulated for the test set in the ADNI database. Further, we also select a subset to evaluate the performance of model by a five-fold cross validation strategy (the ratio of validation set is nearly 15 %).

The hyperparameters tuning in network training are set as follows. We use the Adam algorithm as the optimizer, and the cross-entropy loss function is used to optimize the parameters. The learning rate of the optimizer is initially set to 1e−4, after 30 epochs, the learning rate is halved every 10 epochs until to 5e−6. The network model is trained for 60 epochs. During the network training, the batch size is set to 6, a 0.5 dropout layer is used before the fully connected layer. To compare all methods in this work, the training hyperparameters are same in each model. In addition to the usefulness of the AD vs. NC classification for disease exclusion, the prediction task of MCI conversion (sMCI vs. pMCI) is of great importance for the early treatment of AD patients. The convolution kernel is initialized randomly in AD vs. NC task and then we use the network parameters learned to initialize the training network for sMCI vs. pMCI task. We evaluated diagnostic performance based on accuracy (ACC), sensitivity (SEN), specificity (SPE) and the area under curve (AUC).

4.2. Experimental results and discussion

In this section, first we performed ablation experiments on two types of single modality respectively: sMRI and PET images. In the single modality-based approaches, we used the baseline 3D ResNet18, attention-based 3D ResNet18 (which is named AResNet18), AResNet18 with multi-level features (which is named MAResNet18) for comparison, and showed the visual effects of the attention mechanism on AD diagnosis. Second, ablation experiments of multimodal diagnosis methods based on our feature extraction method are conducted, including the effects of shared parameters in multimodal methods on AD diagnosis, as well as the comparison and analysis of feature combination, decision fusion, and our proposed multimodal method. Finally, our proposed multimodal network is compared with other state-of-the-art methods, which demonstrates the effectiveness and superiority of our method.

Tables 2 and 3 summarize the results of the two single modality sMRI and PET based classification tasks performed in our work, respectively. When using only the 3D ResNet baseline model, the sMRI-based classification of AD vs. NC achieved an accuracy of 85.29 % and the PET-based achieved 88.24 %. The PET-based performance outperformed the sMRI, as well as in the sMCI vs. pMCI prediction. Then we combined the attention block after the last convolution layer of ResNet18 architecture as the AResNet18. From the results, it is seen that the diagnostic performance of the model based on the attention mechanism is improved. Due to the complex properties of brain image, the application of attention mechanism in brain image needs more research and exploration. Although some studies have demonstrated changes in the hippocampus and amygdala in sMRI and in the cingulate gyrus and parietal lobe in PET [7], their changes are not significant for the whole brain. Moreover, there is no clinical gold standard for the diagnosis of AD based on brain image by far [41]. Thus, the attention mechanisms also have a limited effectiveness in the diagnosis of cognitive impairment. The diagnostic results based on our single-modality method (MAResNet18) are further improved. In the classification of AD vs. NC, the accuracy based on sMRI reached 88.24 %, while the accuracy based on PET reached 90.69 %. For the classification of sMCI vs. pMCI, the accuracy based on sMRI and PET was 73.68 % and 75.44 %, respectively. With our method, we integrated the high-level and low-level features acquired under the role of attentional blocks, making the subtle differences in brain image can be captured to achieve better diagnostic performance. It can also be found in our above results that the accuracy based on PET is higher than that of sMRI, the result that is also consistent with established clinical knowledge. Metabolic changes in PET imaging responses can detect functional brain changes and specific lesions in AD earlier than sMRI, which also provides a basis for future studies on the application of functional imaging in the early diagnosis of cognitive impairment.

Then we analyzed the ablation experiments for the multimodal methods, which contains two same MAResNet18 as sub-networks in our multimodal framework to achieve feature extraction for brain images of both modalities separately, and the results are shown in Table 4. First, for shared parameters approach we performed ablation experiments in methods with feature combination (FC) and decision fusion (DC). From the results, we can see that the diagnostic performance of learning by shared parameters is not improved, and even decreased relative to that of PET, which is consistent with the research results [23]. The reason we inferred that the representation of AD features in sMRI and PET responses are different, such as those in sMRI mainly include atrophy of hippocampus and temporal lobe, while those in PET are mainly metabolic changes in cingulate gyrus and parietal lobe, and these different changes are reflected in different brain regions, so the different features of each modality cannot be learned through sharing parameters. From the comparison results, it is clear that the twin-based network (non-shared parameters) works better due to the fact that each network learns the features of the corresponding modality features individually, thus allowing the feature fusion takes advantage of modality complementarity. In addition, from the standard deviation, it can be seen that the FC method is more stable. Although the DC method achieves a higher accuracy on a certain subset of tests, it differs from the results of the validation set during model training. Fig. 8 shows the classification accuracy over three folds based on two different training types of multimodal framework. To sum up, in order to capture the features of each modality effectively, our multimodal approaches are performed by the way of twin-network.

Under the twin-network approach, we conducted further ablation experiments with the SPP-based feature-combined method (SPFC) and our multimodal method (SPDFC), the results are shown in Table 5.

Table 2
The classification results of ResNet architecture based on single-modality sMRI image.

Methods	AD vs. NC				sMCI vs. pMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
ResNet18	85.29	81.25	88.89	0.849	70.18	59.09	78.09	0.686
AResNet18	87.50	85.94	88.89	0.874	72.81	54.55	84.29	0.694
MAResNet18	88.24	85.42	90.74	0.881	73.68	65.91	79.05	0.721

Table 3
The classification results of ResNet architecture based on single-modality PET image.

Methods	AD vs. NC				sMCI vs. pMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
ResNet18	88.24	84.38	91.67	0.880	71.93	54.55	82.86	0.687
AResNet18	88.98	85.94	91.67	0.888	73.68	63.64	80.00	0.718
MAResNet18	90.69	87.5	93.52	0.905	75.44	63.64	85.71	0.724

Table 4
The classification results of two different training types in our multimodal methods.

Multimodal	AD vs. NC			
Siamese-network	ACC	SEN	SPE	AUC
DC(S)	88.73(1.70)	87.50	89.81	0.891
FC(S)	89.71(1.47)	86.46	92.59	0.895
Twin-network	ACC	SEN	SPE	AUC
DC(T)	91.67(3.70)	89.58	93.52	0.916
FC(T)	92.65(1.47)	89.58	95.37	0.925

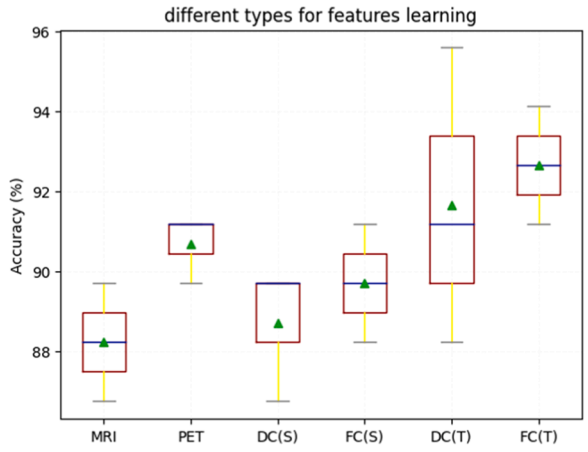


Fig. 8. The accuracy over three folds based on two different training types of siamese-network(S) and twin-network(T).

Incidentally, the diagnostic performance of the SPP-based (SPFC) is higher than that of the multimodal method without SPP block (FC(T) in Table 4). This shows that SPP plays a better pooling role for multimodal features. Under the influence of feature combination and decision fusion, our multimodal method has been further improved. The best accuracy was achieved in both prediction of AD vs. NC with 94.61 % and sMCI vs. pMCI with 77.19 % respectively, and there were significant improvements compared to previous single-modality methods. Meanwhile, it also has good performance in sensitivity and specificity. Sensitivity indicates the accuracy with which AD patients are diagnosed correctly, and specificity refers to the accuracy with which a healthy population can be correctly diagnosed. We also compared the weight parameters of the multimodal method, in the formula (2), c was assigned weighting values of 0.5 and 0.6, respectively. From the standard deviation, it can be seen that the model is more stable when the weight is 0.6. Overall, our proposed multimodal method produced the best results in terms of classification performance and stability. We conclude that the performance improvement depends on the following four aspects. First, our effective feature acquisition, which we have validated on single modality based on attention mechanism and multi-level features integrated approach. Secondly, we trained the network by means of twin-network so that each sub-network can effectively learn the respective features of different modality images. Thirdly, the strategy of SPP makes the down-sampled fused features more robust, effectively preserving multimodal feature information. Finally, we combine the feature fusion and decision fusion under the framework of multimodal twin-network. Fig. 9 summarizes the comparisons between single modality and multimodal approaches in a box plot showing classification accuracy over three experiments. Fig. 10 illustrates the plots of training process for single modality and multimodal methods. During network optimization, the training and validation curves of the multimodal method converge faster and smoother, and the accuracy of validation is obviously improved, which reflect the better performance of our proposed multimodal method. In addition, due to the small amount of brain image data used for AD diagnosis, the validation curves are accompanied by more vibrated on the single modality method, whereas the vibrate phenomenon is significantly less on the multimodal method.

3D Grad CAM can provide better interpretability of AD diagnostic models. We applied 3D Grad CAM to our proposed method to show some important regions of interest to the network. Fig. 11 shows the feature maps of low-level features and high-level features in our model under MRI GM of an AD subject (128S0740) and a NC subject (128S4599), the coronal view of MRI is more intuitive. As can be seen from the feature maps of both convolutional layers, our attention is focused on the hippocampal regions and part of the cerebral cortex. Fig. 12 shows the heat maps of attention to PET of an AD subject (128S0740) and a NC subject (128S4599), the sagittal and axial views of PET image are more intuitive, where include the posterior cingulate gyri, parietal lobule and precuneus marked as some important regions. The regions are also consistent with medical clinical variations [7,42]. While the higher-level features and the lower-level features in our method reflect detailed and global attention effects, respectively, which we further combine to obtain better diagnostic performance.

The datasets used for brain disorders diagnosis tend to be small, and in many cases the trained models are unstable. For example, the performance of the models trained on the different dataset varies. For this purpose, we performed a five-fold cross validation strategy on a fixed sub-dataset via our SPDFC (0.6) multimodal method. Fig. 13 shows these five test ACCs and AUCs of two diagnostic tasks. Specifically, we achieved average accuracy of 93.83 % and 77.19 % in AD diagnosis and MCI conversion prediction, with standard deviations of 1.23 and 1.24, respectively. In terms of AUC, 0.937 and 0.762 were obtained, with corresponding standard deviations of 1.18 and 2.38 respectively. The experimental results show that the ACCs and AUCs of our repeatedly trained models are similar in several tests, and the two standard deviations are relatively small, which also indicates the feasibility and

Table 5
The classification results of our multimodal methods based on sMRI and PET images.

Methods	AD vs. NC				sMCI vs. pMCI			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
SPFC	93.38(0.85)	89.06	97.22	0.931	76.61(1.76)	68.18	81.90	0.75
SPDFC (0.5)	94.12(2.08)	92.19	95.83	0.933	76.61(3.65)	60.61	86.67	0.736
SPDFC (0.6)	94.61(0.85)	92.19	97.22	0.947	77.19(1.01)	68.18	82.86	0.755

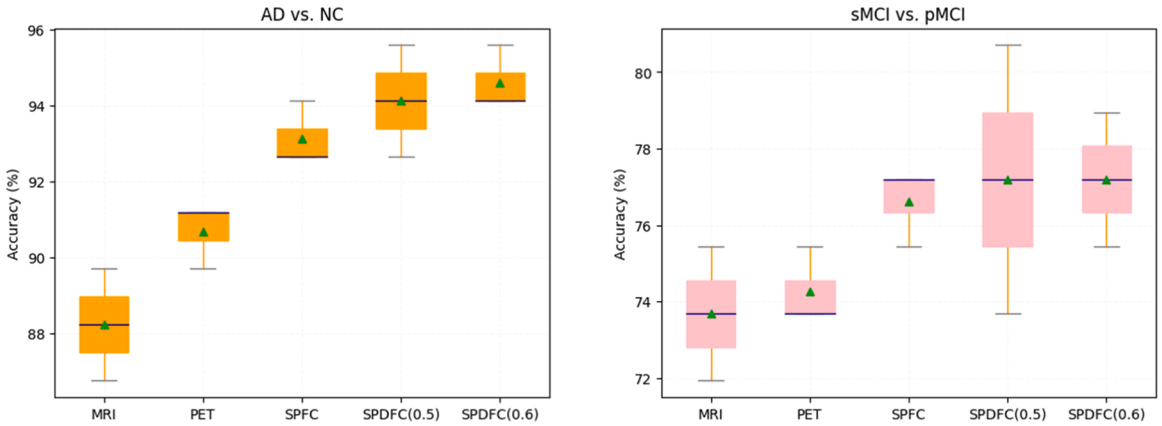


Fig. 9. The classification accuracy over three experiments based on single-modality and multimodal methods.

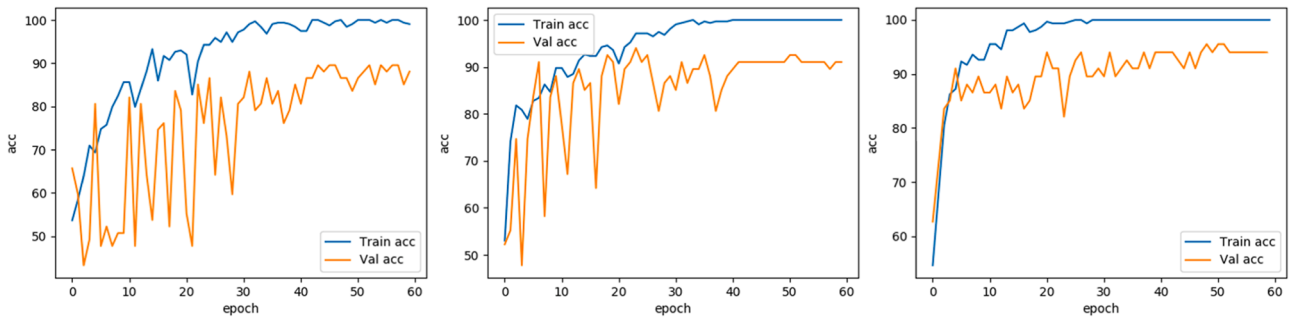


Fig. 10. The plots from left-to-right are the training and verification curve based on sMRI (MAResNet18), PET (MAResNet18) and proposed SPDFC multimodal method, respectively.

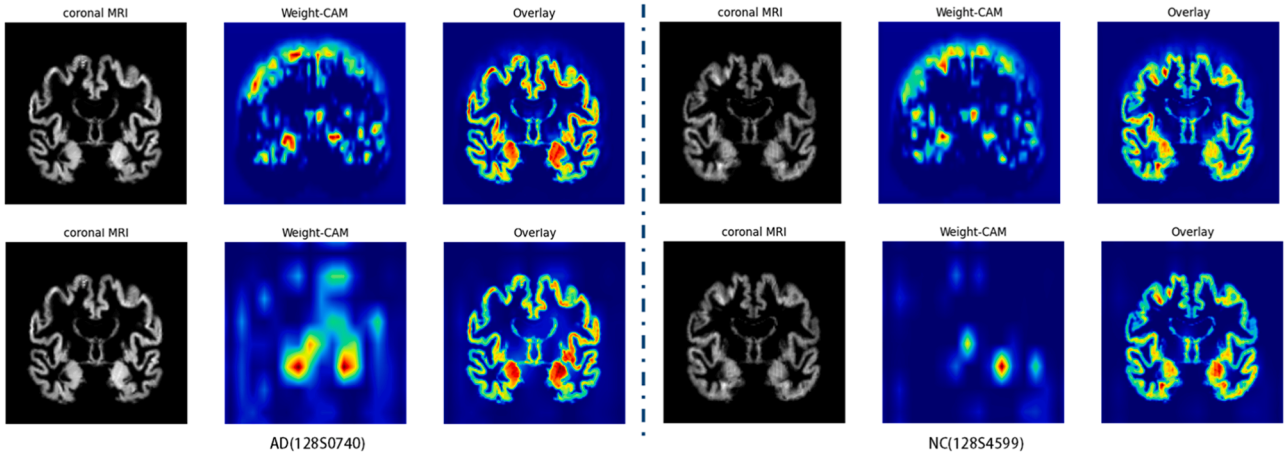


Fig. 11. The left and right panels shown for an AD subject and a NC subject, respectively. From left to right are the coronal view of MRI GM, CAM weights, and visual interpretation of the heat maps. We enumerate the application of Grad Cam to the first convolutional layer (top row) and the third convolutional layer (bottom row) to show the visualization of attention maps.

stability of our method.

4.3. Comparison to the related research and prospect

Furthermore, we compared our method with other deep learning-based studies that are based on the ADNI dataset. Although some 2D CNN-based studies achieve the ideal diagnostic accuracy, many of the results are caused by the data leakage with dataset split strategy, with disparities of 10 % or even more than 20 % points in rigorous test evaluation [27]. From the selected AD diagnosis methods listed in Tables 6 and 7, it can be seen that our proposed method achieves 94.61 % and 77.19 % accuracy in the classification of AD vs. NC and sMCI vs. pMCI, respectively, both showing superior diagnostic performance compared to some state-of-the-art multimodal methods. In addition to a better accuracy, there are advantages in terms of sensitivity, specificity and AUC. Sensitivity indicates the accuracy of patients being diagnosed correctly, i.e., a low rate of missed diagnoses, and specificity refers to the accuracy of being able to correctly diagnose a healthy population, all of which illustrate the effectiveness of our proposed multimodal method. At the same time, our single-modality

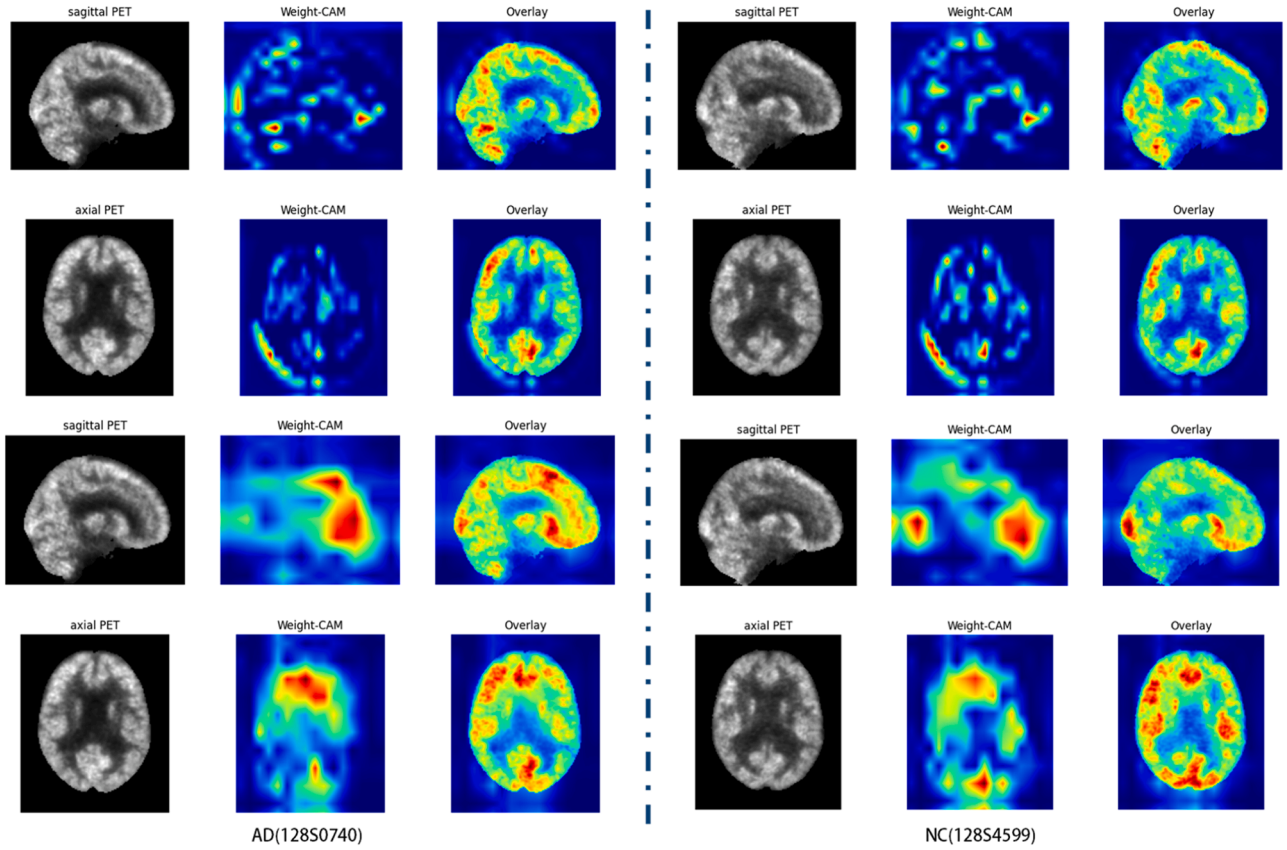


Fig. 12. The left and right panels shown for an AD subject and a NC subject, respectively. From left to right are the sagittal and axial views of PET image, CAM weights, and visual interpretation of the heat maps. We enumerate the application of Grad Cam to the second convolutional layer (top two rows) and the fourth convolutional layer (bottom two rows) to show the visualization of attention maps.

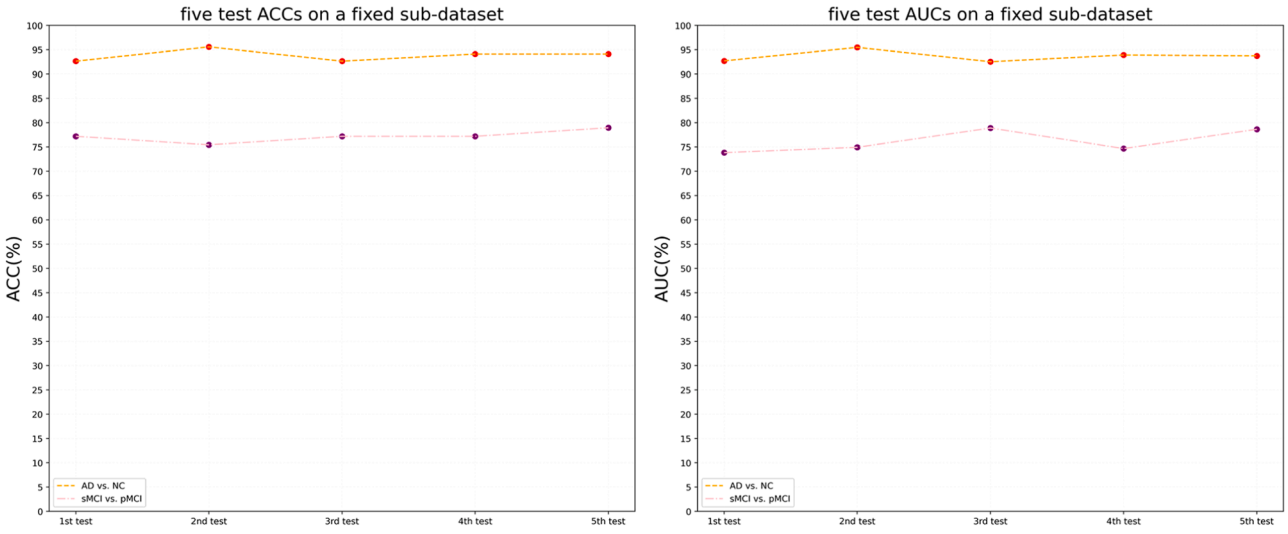


Fig. 13. Five test results over five-fold cross validation strategy on a fixed sub-dataset, the left and right sides show the ACC and AUC respectively.

method also has some advantages, such as a PET-based accuracy of 90.69 %, which outperforms some single-modality methods. In addition, our proposed method is an end-to-end AD diagnosis network framework that eliminates the need for manual feature extraction and other processes, which has feasible applications in AD computer-aided diagnosis.

Currently, some CNN-based multimodal approaches mainly focus on the structural improvement of network such as features concatenation at the fully connected layer, ignoring the importance of how to effectively combine multimodal features. In this work, we firstly extract features of both modalities effectively through the integration of attention mechanism and multi-lever features. Secondly, we make each sub-network learn the respective features of different modalities through the training of twin-network, and then use SPP for dimensionality reduction to make the fused feature more robust. Finally, our networks achieve more effective diagnostic prediction through feature combination and decision fusion. In addition, we show the heat map visualization of the

Table 6
Classification results of AD vs. NC in studies based on single-modality and multimodal methods.

Studies	ACC	SEN	SPE	AUC	Method	Subjects
Wen et al. [28]	89.00	–	–	–	ROI-based 3D CNN	336AD + 330NC
Duan et al. [43]	89.58	–	–	–	2D VIT	64AD + 376NC
Adeli et al. [44]	92.10	–	–	–	RFS-LDA	93AD + 101NC
Shao et al. [45]	92.68	89.47	95.45	–	Hypergraph based	160AD + 211NC
Gao et al. [46]	90.50	84.60	95.0	93.90	Mutil-modal based	352AD + 427NC
Xing et al. [47]	91.34	86.23	93.51	–	Mutil-modal CNN	167AD + 214NC
Gao et al. [48]	92.00	89.10	94.00	90.50	Mutil-modal CNN	196AD + 227NC
Huang et al. [15]	90.10	90.85	89.21	90.84	Mutil-modal CNN	465AD + 480NC
Lin et al. [16]	92.28	90.38	94.37	92.76	Mutil-modal CNN	362AD + 308NC
Narazani et al. [23]	89.60	–	–	–	Mutil-modal CNN	257AD + 270NC
Zhang et al. [49]	91.07	–	–	94.44	Mutil-modal CNN	157AD + 156NC
Angelica et al. [17]	92.11	92.80	91.33	–	Mutil-modal network	250AD + 250NC
OURS	94.61	92.19	97.22	94.71	Mutil-modal CNN	215AD + 246NC

Table 7
Classification results of sMCI vs. pMCI in studies based on single-modality and multimodal methods.

Studies	ACC	SEN	SPE	AUC	Method	Subjects
Wen et al. [28]	74.00	–	–	–	ROI-based 3D CNN	298sMCI + 295pMCI
Kang et al. [50]	66.70	–	–	–	2D CNN	90sMCI + 126pMCI
Wen et al. [28]	69.00	–	–	–	3D CNN	298sMCI + 295pMCI
Shao et al. [45]	75.48	83.84	63.26	70.34	Hypergraph based	273EMCI + 187LMCI
Gao et al. [46]	73.60	59.10	82.10	73.70	Mutil-modal based	342sMCI + 234pMCI
Huang et al. [15]	72.22	73.44	71.25	–	Mutil-modal CNN	441sMCI + 326pMCI
Zhang et al. [51]	75.50	–	–	–	Mutil-modal CNN	343sMCI + 120pMCI
Lin et al. [16]	74.10	75.00	73.08	76.60	Mutil-modal CNN	183sMCI + 233pMCI
Gao et al. [48]	75.30	77.30	74.10	69.90	Mutil-modal CNN	342sMCI + 234pMCI
OURS	77.19	68.18	81.86	75.56	Mutil-modal CNN	238sMCI + 151pMCI

learned weights by our attention block for two modality images, indicating that the brain regions attended to by our network are different for two brain images, which is consistent with the clinical knowledge that sMRI and PET images of AD appear differently. It also reflects the effectiveness of our method as well as explains why the training approach by a twin-network without shared parameters is more effective.

Moreover, it is not difficult to find that the prediction of sMCI vs. pMCI in current research is not ideal, we analyze the reasons for this as follows: 1) the brain differences between sMCI and pMCI are not obvious, and the early symptoms of mild cognitive impairment are similar in both converters and non-converters, so it is difficult to classify them, which is also a difficult issue in clinical diagnosis; 2) the present work used brain imaging data at the baseline time, and converters need 36 months to determine whether they are converted to AD patients. The brains of the converted subjects may differ significantly from their imaging data at the baseline due to the aggravation of the disease during these 3 years, and it is very difficult to improve the prediction performance for the baseline imaging data without adding more a priori knowledge.

With the development of artificial intelligence, the effective use of multi-modality data for computer-assisted AD diagnosis is the development trend and focus in this field. Meanwhile, multi-modality data is not only limited to the imaging data, but with the accumulation of medical data, some combinations of non-imaging data will also successively become the research objects of multimodal methods. We know that accurate prediction of MCI conversion can assist clinical diagnosis to achieve more accurate exclusion, and play an important role in early intervention of AD. In future work, how to explore the changes in brain regions through machine learning combined with clinical knowledge, and utilize the complementary and rich information of multi-modality medical data to achieve more accurate prediction of MCI conversion is a further challenge to be addressed in the future.

5. Conclusion

In this study, we propose an end-to-end multimodal 3D CNN framework that used sMRI and PET images to predict early Alzheimer's disease. To reduce the loss of brain image information in feature extraction, we integrated multi-level features in the fully connected layer under the attention mechanism. We implemented features learning by a twin-based network via non-shared parameters training for multi-modality images. In multimodal framework, we used the SPP block and further combine the strategy of feature combination and decision fusion for final prediction to achieve a better diagnostic performance. In addition, the effectiveness of our approach is demonstrated by the visualization of the attention map, as the model can focus on important brain regions relevant to AD diagnosis. Compared to several state-of-the-art multimodal-based studies, our proposed method exhibits better or equivalent diagnostic performance in both AD diagnosis and MCI conversion prediction tasks.

Funding

This paper was supported in part by the Key Research and Development Program of Sichuan Province (Grant no. 2022YFS0098), the Sichuan Science and Technology Bureau-Key Research & Develop Support Program (Grant no. 22GJHZ0044), the Chengdu Major Technology Application Demonstration Project (Grant no. 2019-YF09-00120-SN) and the Sichuan Science and Technology Program (Grant no. 2023YFS0195).

CRediT authorship contribution statement

Yanteng Zhang: Methodology, Software, Writing – original draft. **Xiaohai He:** Writing – review & editing, Investigation. **Yixin Liu:** Resources. **Charlene Zhi Lin Ong:** Validation, Writing – review & editing. **Yan Liu:** Resources, Project administration. **Qizhi Teng:** Writing – review & editing, Supervision.

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

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. 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/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. More details can be found at adni.loni.usc.edu.

References

[1] M.A. Deture, D.W. Dickson, The neuropathological diagnosis of Alzheimer's disease, Mol. Neurodegener. 14 (1) (2019) 1–18, https://doi.org/10.1186/s13024-019-0333-5.
[2] R.C. Petersen, Mild cognitive impairment, Lancet 367 (9527) (1979) 2006, https://doi.org/10.1016/S0140-6736(06)68881-8.
[3] L. Robinson, E. Tang, J.P. Taylor, Dementia: timely diagnosis and early intervention, BMJ 350 (June) (2015) 1–6, https://doi.org/10.1136/bmj.h3029.
[4] L. Zhang, M. Wang, M. Liu, D. Zhang, A survey on deep learning for neuroimaging-based brain disorder analysis, Front. Neurosci. 14 (October) (2020) 1–19, https://doi.org/10.3389/fnins.2020.00779.
[5] A. Alberdi, A. Aztiria, A. Basarab, On the early diagnosis of Alzheimer's Disease from multimodal signals: a survey, Artif. Intell. Med. 71 (2016) 1–29.
[6] H.Q. Ontario, The appropriate use of neuroimaging in the diagnostic work-up of dementia: an evidence-based analysis, Ont. Health Technol. Assess. Ser. 14 (1) (2014) 1–64.
[7] K. Ishii, H. Sasaki, A.K. Kono, Comparison of gray matter and metabolic reduction in mild Alzheimer's disease using FDG-PET and voxel-based morphometric MR studies, Eur. J. Nucl. Med. Mol. Imaging 32 (8) (2005) 959–963, https://doi.org/10.1007/s00259-004-1740-5.
[8] A. Myoraku, G. Klein, S. Landau, D. Tosun, Regional uptakes from early-frame amyloid PET and 18F-FDG PET scans are comparable independent of disease state, Eur. J. Hybrid Imaging 6 (1) (2022), https://doi.org/10.1186/s41824-021-00123-0.
[9] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3D brain MRI classification, in: International Symposium on Biomedical Imaging(ISBI, 2017, pp. 835–838.
[10] E. Yagis, L. Citi, S. Diciotti, C. Marzi, SW. Atnafu, A.G.S. De Herrera, 3D Convolutional neural networks for diagnosis of Alzheimer's disease via structural MRI, in: Proc. - IEEE Symp. Comput. Med. Syst., vol. 2020-July, 2020, pp. 65–70, https://doi.org/10.1109/CBMS49503.2020.00020.
[11] L. Chen, H. Qiao, F. Zhu, Alzheimer's disease diagnosis with brain structural mri using multiview-slice attention and 3D convolution neural network, Front. Aging Neurosci. 14 (April) (2022), https://doi.org/10.3389/fnagi.2022.871706.
[12] Z. Qin, Z. Liu, Q. Guo, P. Zhu, 3D convolutional neural networks with hybrid attention mechanism for early diagnosis of Alzheimer's disease, Biomed. Signal Process. Control 77 (January) (2022), 103828, https://doi.org/10.1016/j.bspc.2022.103828.
[13] T. Illakiya, K. Ramamurthy, M.V. Siddharth, R. Mishra, A. Udainiya, AHANet: adaptive hybrid attention network for Alzheimer's disease classification using brain magnetic resonance imaging, Bioengineering 10 (6) (2023) 714, https://doi.org/10.3390/bioengineering10060714.
[14] M.A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review, Comput. Methods Prog. Biomed. 187 (2020), 105242, https://doi.org/10.1016/j.cmpb.2019.105242.
[15] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network, Front. Neurosci. 13 (MAY) (2019), https://doi.org/10.3389/fnins.2019.00509.
[16] 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 (April) (2021) 1–13, https://doi.org/10.3389/fnins.2021.646013.
[17] A.I. Aviles-Rivero, C. Runkel, N. Papadakis, Z. Kourtzi, C.B. Schönlieb, Multi-modal hypergraph diffusion network with dual prior for Alzheimer classification, in: Lect. Notes Comput. Sci. (*Including* Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 13433, LNCS, 2022, pp. 717–727, https://doi.org/10.1007/978-3-031-16437-8_69.
[18] T. Illakiya, R. Karthik, Automatic detection of Alzheimer's disease using deep learning models and neuro-imaging: current trends and future perspectives, Neuroinformatics 21 (2) (2023) 339–364, https://doi.org/10.1007/s12021-023-09625-7.
[19] M. Khojaste-Sarakhsi, S.S. Haghighi, S.M.T.F. Ghomi, E. Marchiori, Deep learning for Alzheimer's disease diagnosis: a survey, Artif. Intell. Med. 130 (May) (2022), 102332, https://doi.org/10.1016/j.artmed.2022.102332.
[20] D. Cheng, M. Liu, CNNs based multi-modality classification for AD diagnosis, in: Proc. - 2017 10th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2017, vol. 2018-Janua, no. 61375112, 2018, pp. 1–5, https://doi.org/10.1109/CISP-BMEI.2017.8302281.
[21] T.D. Vu, H.J. Yang, V.Q. Nguyen, A.R. Oh, M.S. Kim, Multimodal learning using convolution neural network and sparse autoencoder, in: 2017 IEEE Int. Conf. Big Data Smart Comput. BigComp 2017, 2017, pp. 309–312, https://doi.org/10.1109/BIGCOMP.2017.7881683.
[22] A. Shoeibi, et al., Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: a review, Inf. Fusion 93 (December 2022) (2023) 85–117, https://doi.org/10.1016/j.inffus.2022.12.010.
[23] M. Narazani, I. Sarasua, S. Pölsterl, A. Lizarraga, I. Yakushev, C. Wachinger, Is a PET all you need? A multi-modal study for Alzheimer's disease using 3D CNNs, in: MICCAI 2022 Med. Image Comput. Comput. Assist. Interv., 2022, pp. 1–12.
[24] V.M. Anderson, J.M. Schott, J.W. Bartlett, K.K. Leung, D.H. Miller, N.C. Fox, Gray matter atrophy rate as a marker of disease progression in AD, Neurobiol. Aging 33 (7) (2012) 1194–1202, https://doi.org/10.1016/j.neurobiolaging.2010.11.001.
[25] C.M. Cabello, et al., Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI), Alzheimer's Dement 46 (2) (2005) 55–66.
[26] C.R. Jack, et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging 27 (4) (2008) 685–691, https://doi.org/10.1002/jmri.21049.
[27] E. Yagis, et al., Effect of data leakage in brain MRI classification using 2D convolutional neural networks, Sci. Rep. 11 (1) (2021) 1–14, https://doi.org/10.1038/s41598-021-01681-w.
[28] J. Wen, et al., Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation, Med. Image Anal. 63 (2020), 101694, https://doi.org/10.1016/j.media.2020.101694.
[29] N.J. Tustison, et al., N4ITK: improved N3 bias correction, IEEE Trans. Med. Imaging 29 (6) (2010) 1310–1320, https://doi.org/10.1109/TMI.2010.2046908.
[30] A. Klein, et al., Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration, Neuroimage 46 (3) (2009) 786–802, https://doi.org/10.1016/j.neuroimage.2008.12.037.
[31] B.B. Avants, N.J. Tustison, M. Stauffer, G. Song, B. Wu, J.C. Gee, The insight ToolKit image registration framework, Front. Neuroinform. 8 (APR) (2014) 1–13, https://doi.org/10.3389/fninf.2014.00044.
[32] A. Routier, et al., Clinica: an open-source software platform for reproducible clinical neuroscience studies, Front. Neuroinform. 15 (2021), https://doi.org/10.3389/fninf.2021.689675.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, 2016, pp. 770–778.
[34] R.L. Gollub, S.N. Murphy, R.L. Robertson, P.E. Grant, Brain age estimation using LSTM on children's brain MRI, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 420–423.
[35] Z. Ling, S. Yang, F. Gou, Z. Dai, J. Wu, Intelligent assistant diagnosis system of osteosarcoma MRI image based on transformer and convolution in developing countries, IEEE J. Biomed. Heal. Inform. 26 (11) (2022) 5563–5574, https://doi.org/10.1109/JBHI.2022.3196043.
[36] S. Woo, J. Park, J. Lee, I.S. Kweon, Convolutional_Block_Attention', in: Eccv, 2018, p. 17 [Online]. Available: http://files/737/Wooetal.-ConvolutionalBlockAttentionModule.pdf.
[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141, https://doi.org/10.1109/CVPR.2018.00745.
[38] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916, https://doi.org/10.1109/TPAMI.2015.2389824.
[39] P. Msonda, S.A. Uymaz, S.S. Karaağaç, Spatial pyramid pooling in deep convolutional networks for automatic tuberculosis diagnosis, Trait. Signal 37 (6) (2020) 1075–1084, https://doi.org/10.18280/TS.370620.
[40] Z. Zhu, et al., 3D pyramid pooling network for abdominal MRI series classification, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2022) 1688–1698, https://doi.org/10.1109/TPAMI.2020.3033990.
[41] J. Tian, et al., Modular machine learning for Alzheimer's disease classification from retinal vasculature, Sci. Rep. 11 (1) (2021) 1–11, https://doi.org/10.1038/s41598-020-80312-2.
[42] K. Ishii, PET approaches for diagnosis of dementia, Am. J. Neuroradiol. 35 (11) (2014) 2030–2038, https://doi.org/10.3174/ajnr.A3695.
[43] Y. Duan, R. Wang, Y. Li, Aux-ViT : classification of Alzheimer's disease from MRI based on vision transformer with auxiliary branch', in: 2023 5th Int. Conf. Commun. Inf. Syst. Comput. Eng, 2023, pp. 382–386, https://doi.org/10.1109/cisce58541.2023.10142358.

[44] E. Adeli, K.-H. Thung, L. An, G. Wu, F. Shi, T. Wang, D. Shen, Semi-supervised discriminative classification robust to sample-outliers and feature-noises, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 515–522, https://doi.org/10.1109/tale.2018.8615213.

[45] W. Shao, Y. Peng, C. Zu, M. Wang, D. Zhang, Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease, Comput. Med. Imaging Graph. 80 (2020), 101663, https://doi.org/10.1016/j.compmedimag.2019.101663.

[46] X. Gao, H. Cai, M. Liu, A hybrid multi-scale attention convolution and aging transformer network for Alzheimer's disease diagnosis, IEEE J. Biomed. Heal. Inform. 27 (7) (2023) 3292–3301, https://doi.org/10.1109/JBHI.2023.3270937.

[47] X. Xin, L. Gongbo, Z. Yu, Advit: vision transformer on multi-modality PET images for Alzheimer, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022, pp. 1–4.

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

[49] 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 (April) (2023), 107050, https://doi.org/10.1016/j.compbiomed.2023.107050.

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

[51] Z.C. Zhang, X. Zhao, G. Dong, X.M. Zhao, Improving Alzheimer's disease diagnosis with multi-modal PET embedding features by a 3D multi-task MLP-mixer neural network, IEEE J. Biomed. Heal. Inform. 27 (8) (2023) 4040–4051, https://doi.org/10.1109/JBHI.2023.3280823.