Computers in Biology and Medicine 165 (2023) 107396

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

DE-JANet: A unified network based on dual encoder and joint attention for Alzheimer's disease classification using multi-modal data

Yulan Dai [a,b], Beiji Zou [a,b], Chengzhang Zhu [a,b,*], Yang Li [a,b], Zhi Chen [a,b], Zexin Ji [a,b], Xiaoyan Kui [a], Wensheng Zhang [c]

[a] *School of Computer Science and Engineering, Central South University, Changsha, China*
[b] *Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha, China*
[c] *Institute of Automation, Chinese Academy of Sciences, Beijing, China*

ARTICLE INFO

Keywords:
Alzheimer's disease
Multi-modal data
Convolutional neural network
Joint attention

ABSTRACT

Structural magnetic resonance imaging (sMRI), which can reflect cerebral atrophy, plays an important role in the early detection of Alzheimer's disease (AD). However, the information provided by analyzing only the morphological changes in sMRI is relatively limited, and the assessment of the atrophy degree is subjective. Therefore, it is meaningful to combine sMRI with other clinical information to acquire complementary diagnosis information and achieve a more accurate classification of AD. Nevertheless, how to fuse these multi-modal data effectively is still challenging. In this paper, we propose DE-JANet, a unified AD classification network that integrates image data sMRI with non-image clinical data, such as age and Mini-Mental State Examination (MMSE) score, for more effective multi-modal analysis. DE-JANet consists of three key components: (1) a dual encoder module for extracting low-level features from the image and non-image data according to specific encoding regularity, (2) a joint attention module for fusing multi-modal features, and (3) a token classification module for performing AD-related classification according to the fused multi-modal features. Our DE-JANet is evaluated on the ADNI dataset, with a mean accuracy of 0.9722 and 0.9538 for AD classification and mild cognition impairment (MCI) classification, respectively, which is superior to existing methods and indicates advanced performance on AD-related diagnosis tasks.

1. Introduction

Alzheimer's disease(AD) is an irreversible chronic neurodegenerative disease with symptoms of progressive cognitive impairment [1], and until now there are no any effective treatments. About 60%–80% among AD patients are dementia cases, resulting in serious social problems [2]. Therefore, it is of great importance to diagnose AD early and carry out the clinical intervention in advance to slow down its progression. This has led to a significant amount of research focusing on developing intelligent methods for diagnosing AD [3,4]. And different biomarkers have been developed for the diagnosis of AD and its prodromal stage, i.e. mild cognition impairment (MCI), such as neuroimaging measures and neuropsychological test data [5].

Structural magnetic resonance imaging (sMRI), a typical kind of neuroimaging measure, can reveal the changes in cerebral anatomical structure induced by the AD process [6] as shown in Fig. 1. For example, there is cerebral cortex atrophy and ventricle enlargement in the brains of individuals with MCI and AD when compared to those with normal control (NC). The magnitude of brain changes increases as the disease progression becomes more severe. The features extracted from sMRI, such as cortical thickness, texture, and volume, can capture brain changes and serve as effective biomarkers for diagnosing AD. The existing methods [7] are devoted to capturing these morphological characteristics for the early diagnosis of AD. According to the different feature extraction levels from sMRI, almost all methods based on sMRI are divided into three categories: (1) 3D patch-level methods [8,9] segmenting the whole sMRI into fixed-size patches before feature extraction, (2) 3D regions-of-interest(ROIs)-level methods [10] extracting features from anatomical brain template aligned regions, and (3) 3D subject-level methods [11,12] processing the whole sMRI to obtain voxel-wise pathological features.

Besides neuroimaging measures, the neuropsychological test data can assess the severity of cognitive impairment to estimate the course of AD [5]. For example, the Mini-Mental State Examination(MMSE) score, a neuropsychological assessment variable with a scoring range of 0 to 30, can indicate the degree of cognitive impairment, where a lower MMSE score indicates more severe cognitive impairment [13].

[*] Corresponding author at: School of Computer Science and Engineering, Central South University, Changsha, China.
E-mail address: anandawork@126.com (C. Zhu).

https://doi.org/10.1016/j.compbiomed.2023.107396
Received 27 April 2023; Received in revised form 28 July 2023; Accepted 26 August 2023
Available online 26 August 2023

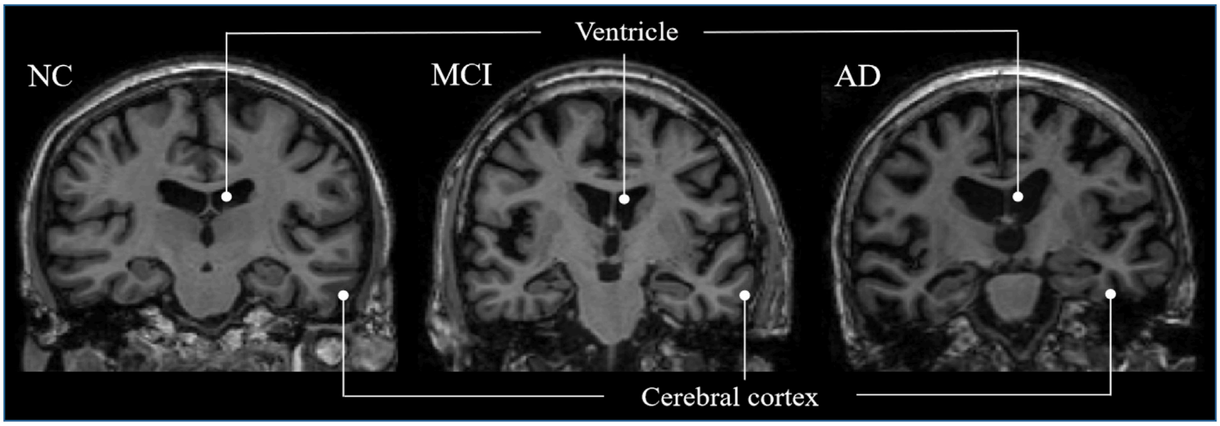

Fig. 1. Examples of sMRI images of NC brain (left), MCI brain (center), and AD brain (right).

Moreover, demographic data, such as age, are also considered a complementary modality. Since aging can cause the brain to shrink and the level of memory impairment to rise, older people are more likely to develop AD [14]. And it has been proven that the prevalence of AD increases with age [15]. Therefore, the magnitude features of age, as well as the MMSE score, can provide the cognitive state information for the diagnosis of AD. The different biomarkers mentioned above expose different pathological information related to AD, resulting in providing comprehending complementary knowledge and facilitating a more accurate diagnosis [16]. This has led to the rapid development of AD classification methods based on multi-modal data [17].

First, among different biomarkers, sMRI is a non-invasive method of obtaining information related to AD by subjectively assessing cerebral morphological changes, which requires rich experience. While age and MMSE score can help evaluate cognitive function in an intuitive quantitative way. By combining these different biomarkers, doctors and researchers can gain a more complete understanding of brain health and cognitive function in individuals. Furthermore, these biomarkers are relatively easy to obtain and cost-effective. Therefore, it is superior to integrate sMRI, age, and MMSE score for AD classification. Second, among the multi-modal-based methods, how to explore the correlation between multi-modal data or construct informative fusion features is the key problem. Whereas, most of the previous methods just simply concatenate multi-modal features and pay less attention to the joint modeling for obtaining multi-modal features [18]. The emergence of Transformer [19] and especially its derivation vision Transformer(ViT) [20] provides new insights for multi-modal data fusion [21]. Owing to its strong self-attention strategy, the data from different modalities can be integrated into a one-dimensional long sequence to capture long-range dependence and be interacted with each other for joint modeling [22]. However, previous studies have found that using tokens obtained by directly encoding inputs with ViT cannot produce satisfactory results for downstream tasks [23]. The reason is that ViT regards the inputs as one-dimensional sequences and solely emphasizes modeling the global features throughout all stages, leading to a lack of local spatial information on the low-level features.

In this paper, we propose a unified AD classification framework, namely DE-JANet, which designs two specific encoders to extract low-level features of the inputted multi-modal data and then adopts ViT to model and fuse these features jointly. Specifically, DE-JANet consists of three modules: (1) dual encoder module, (2) joint attention module, and (3) token classification module. The dual encoder module, including convolutional neural network (CNN) encoder and linear encoder, is designed to separately extract the local spatial features of sMRI and magnitude features of age and MMSE score. The joint attention module defined with ViT is proposed to enable interaction between the extracted low-level features to obtain the fused global features from multi-modal data. Finally, a token classification module is constructed to conduct AD-related diagnoses according to the fused global features. We evaluated DE-JANet on Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets for two AD-related diagnosis tasks which are AD classification, i.e. AD vs NC, and MCI classification, i.e. MCI vs NC. Experimental results prove that our proposed DE-JANet can achieve superior classification performance compared with the existing state-of-the-art methods.

Our contributions are summarized as follows:

- We develop a novel joint modeling strategy that integrates image and non-image features effectively to facilitate downstream tasks.
- The DE-JANet employs a hybrid dual encoder-joint attention structure to compensate for the limitations of ViT in spatial detail extraction, by transferring low-level features from dual encoders to the joint attention module.
- Experimental results demonstrate that DE-JANet outperforms the existing methods on AD classification and MCI classification.

The remainder of this paper proceeds as follows. We give a brief review of related works in Section 2. Then, the studied datasets and our proposed network are introduced in detail in Sections 3 and 4. Next, we evaluate DE-JANet by comparing it to the existing state-of-the-art methods and analyzing its components in Section 5. Meanwhile, the limitations and future work of our study are analyzed. Finally, we draw a conclusion for this paper in Section 6. And for a better understanding of all abbreviations, we summarize them in Table 1.

2. Related work

2.1. Multi-modal-based methods

Many studies have already focused on AD classification using multi-modal data from the ADNI dataset. These multi-modal-based methods incorporate complementary pathological information from multi-modal data to achieve better classification performance by adopting different feature fusion strategies [18]. A straightforward method is to concatenate the multi-modal features from multiple modalities into a unified representation. For example, Jun et al. [24] proposed the multi-modal stacked deep polynomial network to fuse the ROI-based volume features from MRI and average intensity features from Positron Emission Tomography(PET) images. Suk et al. [25] concatenated the ROI-based features from MRI and PET images with the cerebrospinal fluid(CSF) biomarker measures, and then performed weighted feature selection and classifier learning for AD diagnosis. Qiu et al. [26] concatenated the whole sMRI features with normalized age, gender, and MMSE score, and applied multilayer perceptron (MLP) for classification. While the concatenation operation is easy to implement and can handle various modal data, it fails to fully explore the internal relationship between different modal data [27].

Table 1
A summary of abbreviations.

Abbreviation	Full name	Description
DE-JANet	Network based on Dual Encoder and Joint Attention	Our proposed method
AD MCI NC	Alzheimer's Disease Mild Cognition Impairment Normal Control	Class label of clinical status
sMRI MMSE PET CSF	structural Magnetic Resonance Imaging Mini-Mental State Examination Positron Emission Tomograph Cerebrospinal Fluid	Biomarker measures
ADNI	Alzheimer's Disease Neuroimaging Initiative	Dataset
MNI	Montreal Neurological Institute	A standard brain template space
ROIs	Regions-of-Interest	The specific regions in image
ViT CNN MLP FCN 3dConv BN ReLU	Vision Transformer Convolutional Neural Network Multilayer Perceptron Fully Connected Network three-dimensional Convolutional layer Batch Normalization Rectified Linear Unit	Network structure
ACC, PRE, SEN, SPE ROC AUC	Accuracy, Precision, Sensitivity, Specificity Receiver Operating Characteristic the Area Under Curve	Validation metrics

Some studies designed specific feature fusion strategies for exploring the associations among multiple modalities [28]. Tong et al. [29] created pairwise similarity graphs for each modality by utilizing different features from MRI, PET, CSF biomarkers, and genetic data, followed by employing a nonlinear graph fusion technique to merge these similarity graphs into a unified graph for the final classification. Bi et al. [30] proposed "brain region-gene pairs" as multi-modal features by applying canonical correlation analysis methods to capture the associations between MRI images and gene data. Liu et al. [31] adopted a stacked autoencoder framework with a zero-masking strategy for data fusion to extract complementary information from MRI and PET images. Ning et al. [32] and Lei et al. [33] incorporated relational regularization terms, such as Frobenius norm and $l_{2,1}$-norm, into the loss function, aiming to constrain the feature extraction and encourage learning of intrinsic associations inherent in MRI and PET images. However, these specific strategies are sensitive to the allocation of weights to each regularization term or each modality, which often require computationally expensive optimization methods [34]. Furthermore, while these studies do well in the fusion of different modalities of images, there is still significant room for modeling the associations between images and non-image data. In this paper, our joint attention strategy focuses on capturing the interaction information between image and non-image data, without the need to learn sensitive parameters for each modality.

2.2. ViT-based methods

Transformer [19] was first proposed for natural language processing tasks and has made great achievements as its powerful ability in capturing global context information. Subsequently, lots of derivations [20, 35] of Transformer have been developed for computer vision tasks. For example, Alexey et al. [20] designed ViT for image recognition, using only Transformer encoders followed by an MLP head. Moreover, Wang et al. [36] and Chen et al. [37] integrated ViT into 3D U-net-based CNN [38] for medical image segmentation. Xie et al. [39] improved the ViT in the 3D segmentation model by adopting a deformable self-attention strategy, which only performed self-attention on key pixels. These methods have started to integrate Transformer and CNN networks for medical segmentation tasks, while the application of ViT in medical image classification can be further studied. Dai et al. [40] constructed a 2D classification model for preoperative diagnosis of parotid gland tumors by integrating CNN and ViT. However, the 2D slices cannot provide complete structural information, thereby affecting diagnostic accuracy. Moreover, such methods are limited to handling single-modality data.

With regard to processing multi-modal data, Lu et al. [41] proposed a two-stream network called ViLBERT, which first processed different modalities separately and then utilized a cross-Transformer to capture the interactions between modalities. Su et al. [42] proposed a single-stream network called VL-BERT, which simultaneously processed both images and texts in a single Transformer channel. Both ViLBERT and VLBERT can aggregate multi-modal information, but ViLBERT has more parameters than VL-BERT due to its cross-Transformer structure, making it more challenging to train [43]. Moreover, these models require pre-training on large-scale datasets to achieve better generalization [41,42]. When using a single Transformer channel for multi-modal medical data fusion, we design specific encoders for different modalities to capture prior features, making it easier to obtain complementary information during fusion. In this way, we can perform downstream classification tasks without pre-training operations like VL-BERT.

3. Data acquisition and pre-processing

The datasets studied in this paper are acquired from the public Alzheimer's Disease Neuroimaging Initiative [44], namely ADNI-1 and ADNI-2. They are two separate datasets from different phases of ADNI and different subjects so as to eliminate the leakage of test data. The demographic information of all subjects in ADNI-1 and ADNI-2 is presented in Table 2.

In the ADNI-1 dataset, there are 1.5T T1-weighted sMRI scans from 617 subjects. These subjects are divided into three categories: AD, MCI, and NC, according to the standard clinical criteria. To sum up, the ADNI-1 dataset consists of 152 AD, 249 MCI, and 216 NC subjects. The ADNI-2 dataset includes 3T T1-weighted sMRI scans collected from 160 subjects. Similarly, the 160 subjects are divided into 56 AD, 52 MCI, and 52 NC subjects.

The sMRI scans downloaded from ADNI have undergone specific standardized processing steps for eliminating gradient nonlinearity and intensity non-uniformity, including multiplanar reconstruction, 3D Gradwarp correction [45], B1 non-uniformity correction [46], and N3 intensity normalization [47]. On this basis, we perform linear registration of sMRI to the Montreal Neurological Institute (MNI) 152 [48] using the FSL toolbox [49], of which MNI 152 is a standard brain template space to remove global linear differences and unify the coordinate space. After that, the sMRI scans have a uniform size of

Table 2
Demographic information of the subjects involved in the studied datasets (i.e. the ADNI-1 and ADNI-2), including status, gender, age, and MMSE.

Dataset	Status	Gender (Male/Female)	Age (Mean ± Std)	MMSE (Mean ± Std)
ADNI1	AD	77/75	75.53 ± 7.48	23.39 ± 2.06
	MCI	159/90	75.84 ± 7.11	26.26 ± 2.85
	NC	115/101	76.82 ± 5.47	28.91 ± 1.25
ADNI-2	AD	31/25	75.89 ± 8.16	20.30 ± 4.51
	MCI	33/19	80.67 ± 6.67	22.36 ± 4.84
	NC	25/27	77.85 ± 6.04	29.29 ± 0.98

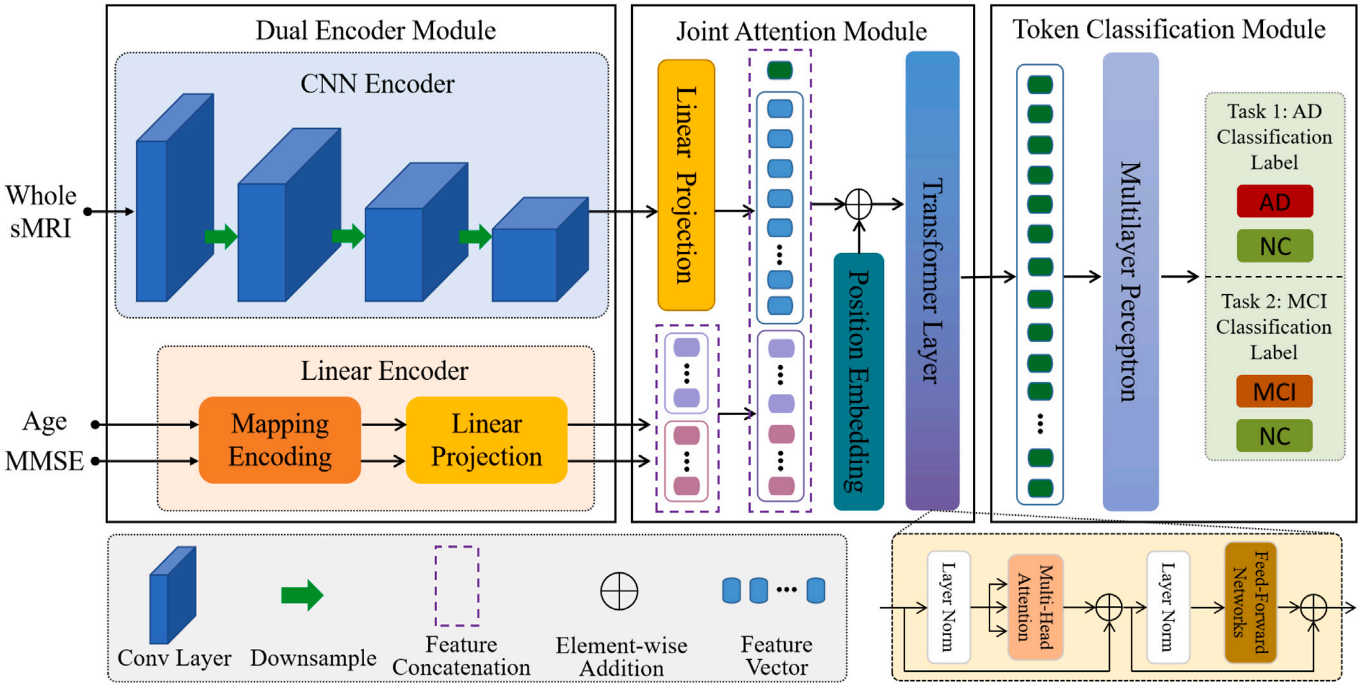

Fig. 2. Illustration of our DE-JANet including three components: (1) dual encoder module, (2) joint attention module, and (3) token classification module.

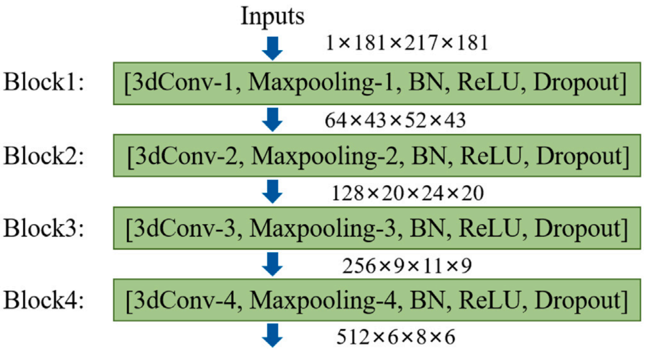
Inputs
1×181×217×181
Block1:
[3dConv-1, Maxpooling-1, BN, ReLU, Dropout]
64×43×52×43
Block2:
[3dConv-2, Maxpooling-2, BN, ReLU, Dropout]
128×20×24×20
Block3:
[3dConv-3, Maxpooling-3, BN, ReLU, Dropout]
256×9×11×9
Block4:
[3dConv-4, Maxpooling-4, BN, ReLU, Dropout]
512×6×8×6

Fig. 3. Illustration of CNN encoder for image data.

181 × 217 × 181 voxels. Then, we further normalize voxels, clip out the intensity outliers and remove background interference to complete the image pre-processing.

4. Methodology

As shown in Fig. 2, our unified network is constructed based on three sequential components, i.e. (1) dual encoder module, (2) joint attention module, and (3) token classification module. Briefly, the dual encoder module consists of a CNN encoder for image data and a linear encoder for non-image data. Then, the image and non-image features interact with each other in the joint attention module to obtain the fused global features that incorporate multi-modal information. Finally, the token classification module generates the class-predicted scores for each subject according to the global features.

4.1. Dual encoder module

4.1.1. CNN encoder for image data

The encoder for image data is built based on an ordinary 3D CNN. As shown in Fig. 3, the CNN encoder comprises four encoding blocks with the same structure. Specifically, each encoding block contains one three-dimensional convolutional (3dConv) layer with different kernel sizes, including 7 × 7 × 7 layer (i.e. 3dConv-1), 4 × 4 × 4 layer (i.e. 3dConv-2), 3 × 3 × 3 layer (i.e. 3dConv-3 and 3dConv-4). The number of channels for 3dConv-1 to 3dConv-4 is 64, 128, 256, and 521, respectively. And the stride of 3dConv-1 is 2, while the rest is 1. Besides, one 3 × 3 × 3 maxpooling operation (i.e. Maxpooling-1) and three 2 × 2 × 2 maxpooling operations (i.e. Maxpooling-2 to Maxpooling-4) are adopted to down-sample the feature maps produced by each 3dConv layer, which are followed by batch normalization (BN) and rectified linear unit (ReLU) activations. And the stride of Maxpooling-4 is 1, while the rest is 2. At the end of each encoding block, a dropout layer with a dropout rate of 0.2 is added to avoid overfitting.

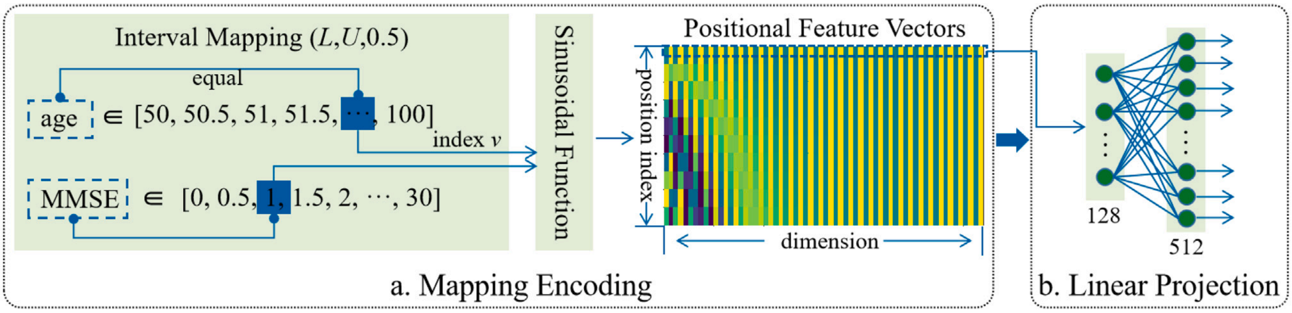
Interval Mapping (L,U,0.5)
equal
age ∈ [50, 50.5, 51, 51.5, ⋯, 100]
index v
MMSE ∈ [0, 0.5, 1, 1.5, 2, ⋯, 30]
Sinusoidal Function
Positional Feature Vectors
position index
dimension
a. Mapping Encoding
128
512
b. Linear Projection

Fig. 4. Illustration of the Linear encoder for non-image data.

As a result, the whole sMRI $I \in R^{(w\times h\times d)}$ is processed by the CNN encoder resulting in a small-size feature representation, denoted as

$$F_{image} = CNN\_Encoder(I) \tag{1}$$

4.1.2. Linear encoder for non-image data

Since concatenating normalized non-image data with image features straightforwardly often results in poorer performance [50], we design a linear encoder inspired by the concept of word embedding to convert each non-image data into a vector. In this way, we can capture the magnitude features of non-image data for the subsequent multi-modal feature interaction.

Specifically, as shown in Fig. 4, the linear encoder is composed of mapping encoding and linear projection. Considering that both age and MMSE score have a range of values, we adopt the relative position of their corresponding values within a fixed interval as their magnitude features. Firstly, we conduct interval mapping for age or MMSE score, which involves projecting these variables into a fixed interval by a lower bound (L), an upper bound (U), and an interval of 0.5, to obtain their position index v with relatively significant differences. Specifically, the MMSE score has a fixed range, with a lower bound (L) of 0 and an upper bound (U) of 30. Based on the age range present in the dataset used, we have set the lower bound (L) for age as 50 and the upper bound (U) as 100. Then, the position index value v is encoded using the sinusoidal function $ME(v,k)$. This function enables capturing relative positioning easily, as $ME(v+o,k)$ can be expressed as a linear function of $ME(v,k)$ for any fixed offset o[24]. $ME(v,k)$ is defined as follows.

$$ME(v,k) = \begin{cases} sin(v/10^{4*k/d}), & k = 2i \\ cos(v/10^{4*(k-1)/d}), & k = 2i+1 \end{cases} \tag{2}$$

where $i = 0,1,2,\ldots,d/2-1$ is the dimension and d is the size of the encoding. After applying the mapping encoding technique, both age and MMSE score are encoded as a positional feature vector of length d. Lastly, the feature vector is further projected linearly through two fully connected layers, with the first layer consisting of 128 neurons and the second layer consisting of 512 neurons, resulting in a final vector size of 1 × 512.

As a result, the magnitude features of the non-image data are well extracted and projected to the same dimension, expressed as:

$$F_{Age} = Linear\_Projection(ME(Age)) \tag{3a}$$

$$F_{MMSE} = Linear\_Projection(ME(MMSE)) \tag{3b}$$

4.2. Joint attention module

The dual encoder module extracts the local spatial features from sMRI and magnitude features from age and MMSE score, respectively. However, it does not yet capture the long-range dependence and the correlation between multi-modal data. Therefore, in this section, we introduce ViT and model these multi-modal features jointly to compensate for the limitations. In this way, we effectively bridge the CNN encoder, linear encoder, and Transformer to enhance feature robustness.

In order to facilitate capturing long-range dependence of image volumes, the image volume features are first projected into a long linear sequence with the same channel as the non-image features. And then, in order to facilitate capturing the correlation between image and non-image features, the flattened image features are concatenated with the encoding features of age and MMSE score to form concatenated multi-modal features which do not contain any interaction information. Before joint modeling, we perform position embedding on the concatenated multi-modal features uniformly to maintain positional information. Besides, we prepend a learnable class token in front of the concatenated multi-modal feature sequences to capture the unified global information between image and non-image features. The class token is defined by random trainable parameters. As for position embedding, we generate random parameters that satisfy the standard normal distribution, and then perform element-wise addition on the generated position parameters and the concatenated multi-modal features, as well as the class token. Last, we input the embedded multi-modal features into the Transformer layer to model jointly. The Transformer layer includes a multi-head attention layer and a feed-forward layer, both following the layer normalization. The multi-head attention layer is a crucial component of the Transformer model, which leverages a self-attention strategy to enable internal interaction and capture correlations among multi-modal features. Specifically, The exact implementation process for the Transformer layer is identical to that for ViT [20].

After being processed by Transformer, the learned class token is obtained and is used to predict the class of the subject. We can explain the joint attention module with the following formulas.

$$F_{non-image} = [F_{Age}; F_{MMSE}] \tag{4a}$$

$$X_{fusion} = Trans([x_{cls}; Linear\_Projection(F_{image}); F_{non-image}] \bigoplus PE)[:,0] \tag{4b}$$

where X_{fusion} represents a fused global multi-modal feature. x_{cls} is a class token. PE is the position embedding parameter. $[;]$ performs concatenate operation. $[:,0]$ denotes the elements with an index of 0 in the Transformer outputs, that is x_{cls} after being processed by Transformer. $Trans$ represents the Transformer layer.

4.3. Token classification module

We get the fused global multi-modal feature X_{fusion} from the joint attention module. Owing to the internal interaction between the image and non-image features, X_{fusion} integrates the complementary information and correlations between them. For example, image features provide the pathological representations of the sMRI images, i.e. the changes in cerebral anatomical structure, and non-image features supply the clinical criteria reference and cognitive function assessment.

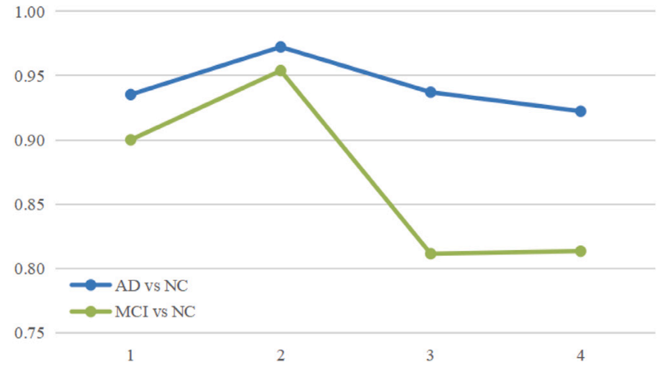

Fig. 5. The variation of classification accuracy (y-axis) with the number of Transformer layers(x-axis).

Both of these features make a great contribution to the AD-related classification tasks and form the fused global feature vector X_{fusion} jointly. In this section, X_{fusion} is fed into the token classification module which adopts MLP as the classifier to yield class predictive scores. Specifically, the fused feature vectors with a size of 1 × 512 are processed by the MLP which consists of two fully connected layers. Finally, the MLP outputs the predicted probabilities of each class.

5. Experiments and analyses

Our experiments focus on two AD-related binary classification tasks, i.e. AD vs NC and MCI vs NC. We first compare our DE-JANet with several state-of-the-art methods to prove its superiority. Then, we perform ablation studies to validate the effectiveness of the components in DE-JANet, including the linear encoder, the multi-modality, and the joint attention.

5.1. Experimental settings

Our DE-JANet is implemented using Python based on the PyTorch package and is trained 100 epochs on a single GPU (i.e. NVIDIA GeForce GTX 1080). The size of the mini-batch is set as 4, and the number of the Transformer layer is 2. We adopt weighted cross-entropy loss to learn our DE-JANet and Adam optimizer for minimizing the loss function with an initial learning rate of 10^{-5}. We verify that the model with these hyperparameter configurations achieves the highest accuracy. For example, Fig. 5 illustrates the variation of classification accuracy with the number of Transformer layers, indicating that the best performance is achieved when the number of layers is 2. The DE-JANet is trained on ADNI-1 and then tested on the other independent dataset, i.e. ADNI-2, which demonstrates the generalization of our method. This process is repeated five times, and performance is presented as mean over the model runs.

For AD classification, we train our model using 152 AD subjects and 185 NC subjects from the ADNI-1 dataset. For MCI classification, we train our model using 249 MCI subjects and 216 NC subjects from the ADNI-1 dataset. We randomly select about 25 percent training samples as the validated set and tune our model according to the validation performance in terms of five typical metrics, including accuracy (ACC), precision (PRE), sensitivity (SEN), specificity (SPE), and F1 score, that are formulated as:

$$ACC = (TP + TN) / (TP + TN + FP + FN) \tag{5a}$$

$$PRE = TP / (TP + FP) \tag{5b}$$

$$SEN = TP / (TP + FN) \tag{5c}$$

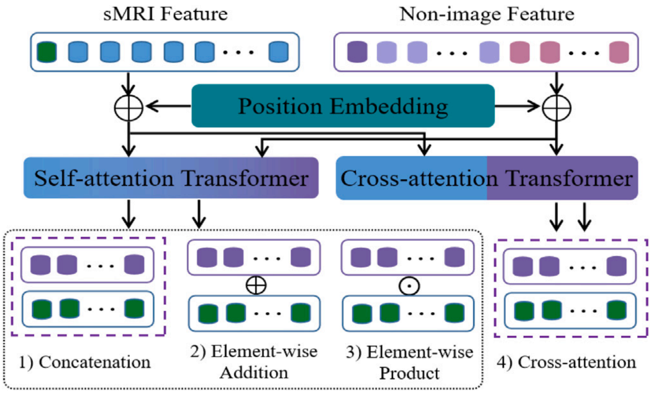

Fig. 6. The flow diagram of the two-stream attention strategy.

$$SPE = TN / (TN + FP) \tag{5d}$$

$$F1 = 2 \times TP / (2 \times TP + FP + FN) \tag{5e}$$

where TP, TN, FP, and FN denote, respectively, the true positive, true negative, false positive, and false negative values. Besides, receiver operating characteristic (ROC) curves and the area under curve (AUC) value are also provided for further evaluation.

5.2. Comparison experiments

In the comparative experiments, the entire experimental process, parameter settings, and data used are consistent with those described in Section 5.1. When conducting the comparisons, we simply replaced our feature fusion strategy or model with other comparative methods.

5.2.1. Comparisons with different feature fusion strategies

We adopt a single-stream joint attention strategy for feature fusion, where the position embedding is uniformly added to the concatenated multi-modal features, with a class token prepend to them. After joint attention operation, we directly obtained the fused global features for classification. To evaluate the performance of our single-stream joint attention strategy, we try another kind of two-stream self-attention strategy with different feature fusion ways, including element-wise addition, element-wise product, and concatenation. In the two-stream self-attention strategy, the position embeddings are added to image features and non-image features respectively, and then the image features and non-image features are separately fed into the self-attention Transformer layer for modeling their respective global features. Finally, the two global feature vectors from image and non-image data are fused in the abovementioned three ways. Besides, we try another two-stream cross-attention strategy [51] for feature fusion. The flow diagram of the two-stream attention strategy and their performance on the ADNI-2 dataset are shown in Figs. 6 and 7, respectively.

As Fig. 7 shows, our single-stream joint attention strategy is superior to the two-stream self-attention strategies using the abovementioned three feature fusion methods, as well as the two-stream cross-attention strategy. Although the two-stream self-attention strategies achieve slightly better results on the metric SEN, our method still performs better in the comprehensive evaluation metric F1 score. These reflect that the fusion features we constructed are more beneficial for downstream classification tasks. This is because the joint attention strategy enables deep interaction between low-level features of the image and non-image data, which helps to explore complementary information among multi-modal features. In contrast, the two-stream self-attention strategies only use simple operations to integrate multi-modal

Table 3
Results of AD classification and MCI classification obtained by our DE-JANet and the competing methods [26,50,51] on ADNI-2 datasets. The best results are in bold.

Method	AD vs NC					MCI vs NC				
	ACC	PRE	SEN	SPE	F1	ACC	PRE	SEN	SPE	F1
Qiu et al. [26] 2020	0.9457	0.9180	0.9756	0.9179	0.9456	0.8506	0.7979	**0.9436**	0.7577	0.8639
Liu et al. [50] 2020	0.9648	**0.9666**	0.9615	**0.9679**	0.9636	0.8558	0.8503	0.8654	0.8462	0.8575
Golovanevsky et al. [51] 2022	0.8407	0.8059	0.8962	0.7893	0.8458	0.7846	0.7607	0.8269	0.7423	0.7904
Our DE-JANet	**0.9722**	0.9623	**0.9808**	0.9643	**0.9714**	**0.9538**	**0.9802**	0.9269	**0.9808**	**0.9523**

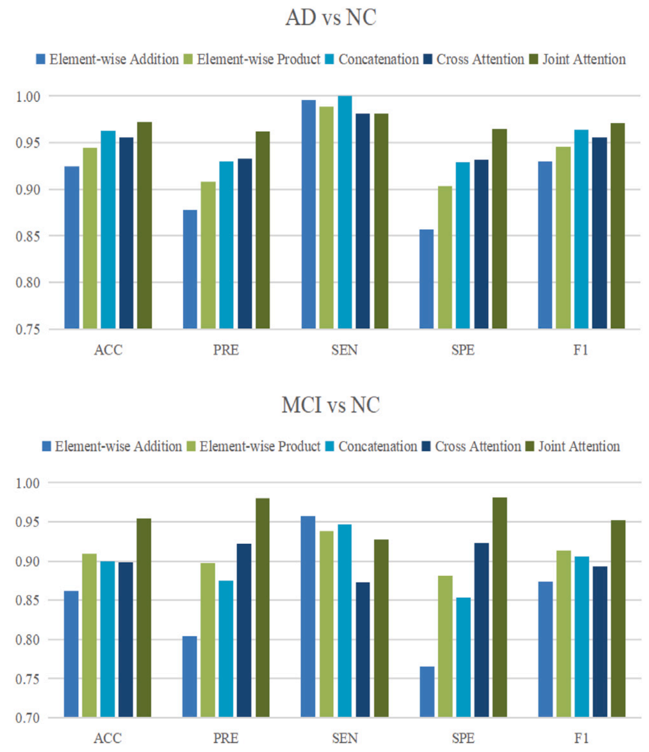

Fig. 7. Performance comparisons of different feature fusion strategies on AD classification (top) and MCI classification (bottom) on ADNI-2 dataset.

features. Besides, although the two-stream cross-attention strategy also captures the interaction between different modal features, it involves separately processing the sequences and then combining the tokens together, which does not result in a prominent joint modeling effect.

5.2.2. *Comparisons with competing methods*

We strictly compare our DE-JANet with three multi-modal-based methods [26,50,51] by repeating these methods in a unified experimental setting and sharing the same multi-modal training and test data including sMRI, age, and MMSE score. The key differences among them are the encoding schemes of non-image data and the fusion methods of multi-modal data. Specifically, Qiu et al. [26] designed a two-step deep learning framework. Firstly, the fully connected network (FCN) model is developed for extracting disease probability maps using randomly selected fixed-size patches from the whole sMRI. Then, the high-risk disease probability maps are selected and integrated alongside non-image data to develop an MLP fusion model for AD classification. Liu et al. [50] concatenated the encoded non-image features to the output of an ordinary 3D CNN trained using sMRI for AD classification. And Golovanevsky et al. [51] used cross-modal attention to capture the interaction between sMRI and non-image data after being processed by CNN or FCN and self-attention, respectively. Table 3 presents the results of the two binary classification tasks obtained by our method and the competing methods [26,50,51].

The following observations can be summarized from Table 3. (1) Our DE-JANet outperforms three competing methods [26,50,51] on both classification tasks on ADNI-2 datasets, especially for MCI classification. For example, the ACC and F1 score of DE-JANet and [26] for MCI classification are 0.9538 versus 0.8506 and 0.9523 versus 0.8639, respectively. One reason for the performance improvement is that we design the linear encoder to extract magnitude features of age and MMSE score considering the sensitivity of the magnitude of age and MMSE score for AD diagnosis. The linear encoder can encode the positional information of age and MMSE score in a fixed interval more effectively to capture the magnitude features by using interval mapping and sinusoidal function, thereby benefiting downstream classification tasks. While Qiu et al. [26] only normalize non-image data and Golovanevsky et al. [51] just project non-image data into a long vector by FCN. Another reason is that we adopt the joint attention strategy to capture the interaction of multi-modal features considering the significance of multi-modal complementary information for AD diagnosis. By jointly modeling one-dimensional long multi-modal feature sequences, the joint attention strategy obtains global features that integrate multi-modal interaction information, which is beneficial for AD classification. While Qiu et al. [26] and Liu et al. [50] just concatenated non-image data to image features directly to achieve multi-modal data fusion. (2) Among the three competing methods [26,50,51], the classification results of Qiu et al. [26] and Liu et al. [50] are relatively better. Although Qiu et al. [26] directly concatenated normalized non-image data to image features, it also achieved relatively good results. This is because Qiu et al. [26] screened the sMRI features obtained by FCN before the concatenation operation, and only the disease probability maps with high risks are used in the next stage of classification. These discriminative probability maps contribute to improving classification performance. Compared with our DE-JANet, Liu et al. [50] fused the multi-modal features by simple concatenation and behaved slightly worse performance. This reflects the superiority of our proposed method in terms of multi-modal data fusion. In addition, Golovanevsky et al. [51] encoded the non-image data by FCN and self-attention and fused the multi-modal features by cross-attention, but the performance is still the worst, as it only used three central slices in each dimension of sMRI for feature encoding and interacting. (3) The MCI classification performance is slightly less significant. It is perhaps due to the fact that the MCI classification is relatively more challenging since MCI subjects show less obvious lesion characteristics compared with AD subjects.

Besides, from Table 3, we find that our method does not achieve the highest results on all metrics. Therefore, in order to provide a comprehensive evaluation of our methods and three competing methods [26,50,51], we also generate ROC curves of AD classification and MCI classification based on the results of one training session for further analysis. From Fig. 8, we can observe that our DE-JANet method consistently achieves the highest AUC score among all the methods evaluated, indicating superior performance in classification. Although the competing methods [26,50] show similar superiority with our DE-JANet in AD classification with AUC scores of 0.9856 and 0.9876 on the ADNI-2 dataset, respectively, our method achieves much higher AUC values in MCI classification with values of 0.9937, reflecting the excellent ability of the DE-JANet model for early AD screening.

5.3. *Ablation studies*

In order to evaluate the effectiveness of the modules and multi-modal data used in DE-JANet, we design multiple comparing models,

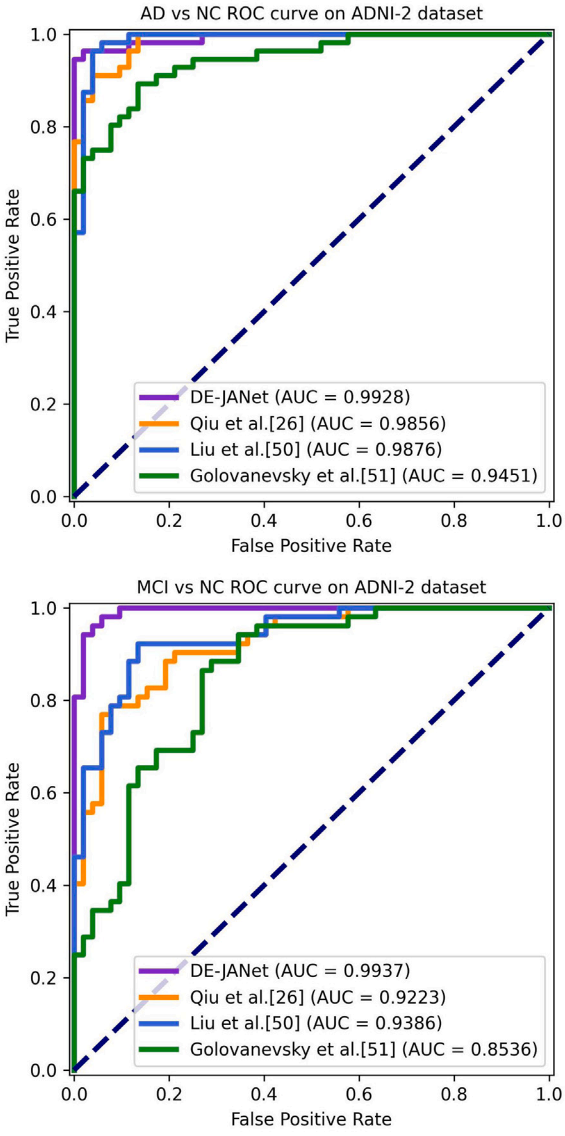

Fig. 8. ROC curves for AD classification (top) and MCI classification (bottom) on ADNI-2 dataset.

i.e. the counterparts of DE-JANet, by removing the related modules or data. As shown in Table 4, Dual-Encoder-Trans represents our DE-JANet, and the rest are its counterparts. The results of ablation studies are reported in Table 5.

5.3.1. Effectiveness of linear encoder

To evaluate the effectiveness of linear encoder, we compare the classification results of Single-Encoder-Trans and Dual-Encoder-Trans. In Single-Encoder-Trans, the age and MMSE score are normalized and then projected into a 512-dimensional vector respectively by FCN before being fed into the joint attention module.

From Table 5, we can observe that Dual-Encoder-Trans outperforms Single-Encoder-Trans in all metrics for both AD and MCI classification, which indicates that the linear encoder module can greatly improve the classification performance. With the common FCN encoder, Single-Encoder-Trans achieves only 0.6481 and 0.7846 ACC for AD and MCI classification, respectively. Besides, compared with CNN-Encoder-Trans, Single-Encoder-Trans does not show a significant improvement in performance even with the addition of non-image data, indicating that only normalizing and projecting non-image data is not sufficient to enhance the classification performance. Therefore, it is necessary to set a reasonable encoding scheme to extract the low-level features of non-image data before multi-modal fusion. In the linear encoder, we adopt a sinusoidal function to extract the magnitude features of age and MMSE score. Since there is a potential correlation between the magnitude of age and MMSE score and the development of AD, for example, the lower the MMSE score, the higher the model's confidence in predicting AD, the linear encoder extracting magnitude features is helpful for improving the performance of AD classification.

5.3.2. Effectiveness of multi-modality

In order to verify the effectiveness of multi-modality, we make the performance comparisons among CNN-Encoder-Trans, Linear-Encoder-Trans, and Dual-Encoder-Trans.

From Table 5, we can have the following observations. (1) Our Dual-Encoder-Trans outperforms the two variant methods on both classification tasks, which is consistent with the assumption that the multi-modal-based methods perform better than the single-modal-based methods. This also implies that the features from different modal data can support each other to improve classification performance. (2) The Linear-Encoder-Trans shows better classification performance than CNN-Encoder-Trans. For example, the ACC for AD and MCI classification is 0.9630 versus 0.6519 and 0.8942 versus 0.7500, respectively. This suggests that the magnitude features of age and MMSE score extracted by Linear-Encoder-Trans have a significant impact on both classification tasks, while the sMRI features extracted by CNN-Encoder-Trans do not contribute much to the improvement of performance. This is because we only use an ordinary CNN in Linear-Encoder-Trans, which may be difficult to explore significant lesion information on sMRI for accurate classification. As a result, The ordinary sMRI features are needed to be combined with other modal features to obtain satisfied classification performance. (3) The performance of Linear-Encoder-Trans is close to that of Dual-Encoder-Trans in AD classification. It may be because the significant difference in the distribution of MMSE score between AD and NC subjects provides discriminative features for AD classification, resulting in a good performance.

In addition, we also set up comparative experiments to analyze the effects of age and MMSE score. As shown in Table 5, after the encoded age and MMSE features are separately supplied into CNN-Encoder-Trans, the performances of both resulting Dual-Encoder-Trans-Na and Dual-Encoder-Trans-Nm are improved compared with CNN-Encoder-Trans. And compared with Dual-Encoder-Trans, the Dual-Encoder-Trans-Na without age and Dual-Encoder-Trans-Nm without MMSE score perform less well. Fig. 9 clearly demonstrates the performance improvement of the models after being separately supplied with MMSE score and age features, revealing a significant upward trend. All these indicate that both age and MMSE score can help improve the performance of AD classification. From Fig. 9, we can also observe that the MMSE score has a more significant performance improvement effect compared to age.

5.3.3. Effectiveness of joint attention

In this group of experiments, we evaluated the effectiveness of joint attention by comparing the performance of Dual-Encoder and Dual-Encoder-Trans. The Dual-Encoder simply concatenates the features extracted by the CNN encoder and linear encoder respectively and inputs them into the token classification module.

From Table 5, we can discover that Dual-Encoder-Trans outperforms Dual-Encoder on both classification tasks, especially for MCI classification. For example, the ACC for MCI classification is 0.9538 versus 0.9077. This intuitively shows that the joint attention module can further improve classification performance. In the joint attention module, we adopt a Transformer-based self-attention structure that highlights the important features relevant to AD diagnosis and captures the interaction between multi-modal data, so as to facilitate the classification tasks. Thus, it is reasonable that the joint attention module makes sense.

Table 4
Models with different modules and data for ablation studies.

Model	CNN encoder	Linear encoder	Image data (sMRI)	Non-image data (MMSE)	Non-image data (Age)	Joint attention
Single-Encoder-Trans	✓		✓	✓	✓	✓
CNN-Encoder-Trans	✓		✓			✓
Linear-Encoder-Trans		✓		✓	✓	✓
Dual-Encoder-Trans-Na	✓	✓	✓	✓		✓
Dual-Encoder-Trans-Nm	✓	✓	✓		✓	✓
Dual-Encoder	✓	✓	✓	✓	✓	
Dual-Encoder-Trans	✓	✓	✓	✓	✓	✓

Table 5
Results of AD classification and MCI classification obtained by DE-JANet and its counterparts on ADNI-2 dataset. The best results are in bold.

Model	AD vs NC					MCI vs NC				
	ACC	PRE	SEN	SPE	F1	ACC	PRE	SEN	SPE	F1
Single-Encoder-Trans	0.6481	0.6239	0.7692	0.5357	0.6779	0.7846	0.7455	0.8692	0.7000	0.8008
CNN-Encoder-Trans	0.6519	0.6241	0.7192	0.5893	0.6650	0.7500	0.7124	0.8500	0.6500	0.7708
Linear-Encoder-Trans	0.9630	0.9615	0.9615	**0.9643**	0.9615	0.8942	0.9767	0.8077	**0.9808**	0.8842
Dual-Encoder-Trans-Na	0.9278	0.8839	**0.9808**	0.8786	0.9294	0.9423	0.9554	0.9308	0.9538	0.9415
Dual-Encoder-Trans-Nm	0.6667	0.6176	0.8077	0.5357	0.7000	0.7596	0.7234	0.8500	0.6692	0.7801
Dual-Encoder	0.9630	0.9444	**0.9808**	0.9464	0.9623	0.9077	0.8727	**0.9577**	0.8577	0.9128
Dual-Encoder-Trans	**0.9722**	**0.9623**	**0.9808**	**0.9643**	**0.9714**	**0.9538**	**0.9802**	0.9269	**0.9808**	**0.9523**

Table 6
Referential comparisons of the state-of-the-art studies using ADNI dataset on AD classification, MCI classification, and 3-class classification. The best results are in bold. '-' indicates that this result was not reported in the literature.

Method	Data	AD vs NC			MCI vs NC			AD vs MCI vs NC		
		ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
Ning et al. [32] 2021	sMRI, PET	0.969	0.957	**0.980**	0.821	0.871	0.723	–	–	–
Zhang et al. [52] 2021	sMRI	0.920	0.903	0.931	0.801	0.782	0.803	0.629	0.645	0.818
Wang et al. [53] 2022	sMRI	0.921	0.962	0.913	0.797	0.772	0.790	0.627	0.661	0.798
Cai et al. [54] 2023	sMRI	0.921	0.917	0.925	0.824	0.836	0.813	–	–	–
Zhu et al. [55] 2023	sMRI, PET, CSF	0.968	**0.986**	0.955	0.866	**0.933**	0.724	–	–	–
our DE-JANet	sMRI, Age, MMSE	**0.972**	0.981	0.964	**0.954**	0.927	**0.981**	**0.725**	**0.724**	**0.862**

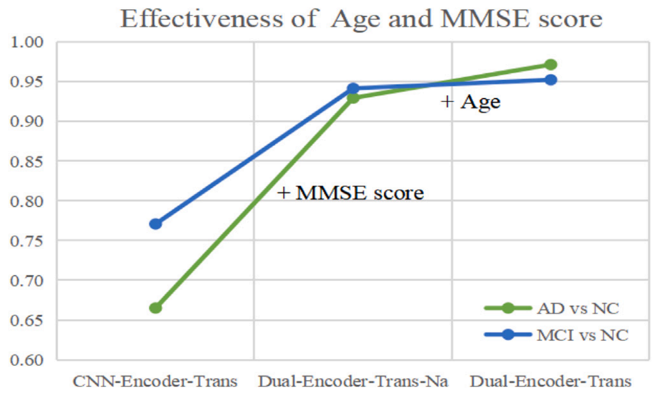

Fig. 9. Performance (F1 score) comparison of DE-JANet and its counterparts on ADNI-2 dataset for analyzing the effectiveness of age and MMSE score. Among these counterparts, supplying the MMSE features into CNN-Encoder-Trans yields Dual-Encoder-Trans-Na, and further supplying age features into Dual-Encoder-Trans-Na yields Dual-Encoder-Trans.

5.4. Discussion

5.4.1. Diagnosis performance

We roughly summarize and compare our results with those of several state-of-the-art methods [32,52–55] using the AD diagnosis results reported in the literature. Besides binary classification, we also show the performance of 3-class classification, i.e. AD vs MCI vs NC. For 3-class classification, our model is trained using all subjects in the ADNI-1 dataset and sharing the same experimental setting as binary classification. As we can observe from Table 6, our proposed method achieves comparable performance on AD classification and advanced performance on MCI classification as well as 3-class classification. Note that due to the differences in data, the direct comparison among these methods mentioned above is impossible and unfair, but we can still draw some conjectures.

First, multi-modal data, which provides more comprehensive information by complementing each other between different modalities, can help improve the performance of models. For example, Zhang et al. [52], Wang et al. [53] and Cai et al. [54] performing classification only based on sMRI show worse performance, while the remaining methods obtain better results by using multi-modal data. Second, for

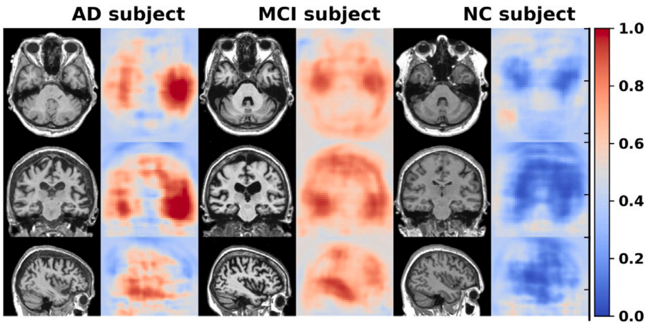

Fig. 10. The visual feature maps from three subjects with clinically confirmed AD, MCI, and NC (from left to right).

multi-modal-based methods, the type of data modalities used for fusion is not crucial provided that the data and methods are effective, due to the fact that the results of [32,55] obtained from fusing images of different imaging ways and that of our DE-JANet obtained from fusing image and non-image data are similar.

5.4.2. Correlation analysis of sMRI features with AD diagnosis

In order to explain model classification decisions, we analyze the correlation between sMRI features and AD diagnosis by visualizing the sMRI features extracted by the CNN encoder as shown in Fig. 10. The visual feature maps highlight high-risk brain regions associated with AD pathology, where the blue and red indicate the low-(<0.5) and high-risk(>0.5) areas of AD. From Fig. 10, we can discover that the sMRI feature map of the first subject with a clinical diagnosis of AD shows much darker red, while that of the third subject with NC is basically blue, and that of the second subject with MCI is in the middle. We can assess the anatomical consistency of AD-suggestive morphology hot spots from these darker area distributions, thus the visual feature maps can act as a means to demonstrate structures most affected by neuropathological changes in AD and explain the model decision to some extent.

5.4.3. Limitations and future work

Though our proposed DE-JANet method achieves satisfactory performance in AD-related diagnosis, there are still several limitations that restrict the exploration of pathological relationships.

First, We analyze the mapping relationship among image data, non-image data, and AD diagnosis. However, we do not explore the correlation between image and non-image data for each subject, which may help reveal underlying pathological responses. Besides, further analysis should be conducted on the features within the site to better support intelligent diagnosis. Second, we analyze the correlation between sMRI features and AD diagnosis to explain the model decision to some extent. However, how the image and non-image features complement each other is unclear in the joint attention module, we can further explore this in the future to support a convincible intelligent diagnosis.

In addition to these improvement efforts, enthusiastic readers may consider expanding this model to the diagnosis of other brain diseases to improve the automation level in the field of neuroimaging analysis. Besides, genetic data and functional imaging data can be introduced to explore more advanced methods for multi-modal data fusion and diagnostic performance improvement.

6. Conclusion

In this study, we propose a unified network DE-JANet consisting of a dual encoder module, a joint attention module, and a token classification module, and adopt multi-modal data consisting of sMRI, age, and MMSE score for AD-related diagnosis. The dual encoder module extracts spatial features of sMRI and magnitude features of age and MMSE score, which serve as a prior basis for the following joint attention module. The joint attention module can effectively fuse image and non-image features and captures the interaction between them, so as to provide significant token features for classification. The comparisons with the existing state-of-the-art methods and ablation studies on the ADNI datasets demonstrate that our method can merge multi-modal data effectively and achieve advanced classification performance. Moreover, we explore the interpretability of the model by analyzing the correlation between sMRI features and AD diagnosis, but we can further analyze the complementarity between the image and non-image features to interpret the model decisions.

Declaration of competing interest

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

Acknowledgments

This work is supported by the National Key R&D Program of China [2018AAA0102100]; the National Natural Science Foundation of China [U22A2034, 62376287]; the Key Research and Development Program of Hunan Province [2022SK2054]; International Science and Technology Innovation Joint Base of Machine Vision and Medical Image Processing in Hunan Province [2021CB1013]; the Natural Science Foundation of Hunan Province [2022JJ30762]. Data used in this work were acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

References

[1] W. Jagust, Vulnerable neural systems and the borderland of brain aging and neurodegeneration, Neuron 77 (2) (2013) 219–234, http://dx.doi.org/10.1016/j.neuron.2013.01.002.

[2] 2021 Alzheimer's disease facts and figures, Alzheimer's Dementia 17 (3) (2021) 327–406, http://dx.doi.org/10.1002/alz.12328.

[3] 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 (2022) 102332, http://dx.doi.org/10.1016/j.artmed.2022.102332.

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

[5] G.M. McKhann, D.S. Knopman, H. Chertkow, B.T. Hyman, C.R. Jack Jr, C.H. Kawas, W.E. Klunk, W.J. Koroshetz, J.J. Manly, R. Mayeux, R.C. Mohs, J.C. Morris, M.N. Rossor, P. Scheltens, M.C. Carrillo, B. Thies, S. Weintraub, C.H. Phelps, The diagnosis of dementia due to Alzheimer's disease: Recommendations from the national institute on Aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's Dementia 7 (3) (2011) 263–269, http://dx.doi.org/10.1016/j.jalz.2011.03.005.

[6] G.B. Frisoni, N.C. Fox, C.R. Jack Jr, P. Scheltens, P.M. Thompson, The clinical use of structural MRI in Alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77, http://dx.doi.org/10.1038/nrneurol.2009.215.

[7] S. Fathi, M. Ahmadi, A. Dehnad, Early diagnosis of Alzheimer's disease based on deep learning: A systematic review, Comput. Biol. Med. 146 (2022) 105634, http://dx.doi.org/10.1016/j.compbiomed.2022.105634.

[8] F. Li, M. Liu, A.D.N. Initiative, Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks, Comput. Med. Imaging Graph. 70 (2018) 101–110, http://dx.doi.org/10.1016/j.compmedimag.2018.09.009.

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

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

[11] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, X. Zhao, Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease, Neurocomputing 333 (14) (2019) 145–156, http://dx.doi.org/10.1016/j.neucom.2018.12.018.

[12] D. AlSaeed, S.F. Omar, Brain MRI analysis for Alzheimer's disease diagnosis using CNN-based feature extraction and machine learning, Sensors 22 (8) (2022) 2911, http://dx.doi.org/10.3390/s22082911.

[13] V.C. Pangman, J. Sloan, L. Guse, An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: Implications for clinical practice, Appl. Nurs. Res. 13 (4) (2000) 209–213, http://dx.doi.org/10.1053/apnr.2000.9231.

[14] R. Peters, Ageing and the brain, Postgrad. Med. J. 82 (964) (2006) 84–88, http://dx.doi.org/10.1136/pgmj.2005.036665.

[15] X. Gao, H. Cai, M. Liu, A hybrid multi-scale attention convolution and aging transformer network for Alzheimer's disease diagnosis, IEEE J. Biomed. Health Inf. (2023) 1–8, http://dx.doi.org/10.1109/JBHI.2023.3270937.

[16] A. Chartsias, G. Papanastasiou, C. Wang, S. Semple, D.E. Newby, R. Dharmakumar, S.A. Tsaftaris, Disentangle, align and fuse for multimodal and semi-supervised image segmentation, IEEE Trans. Med. Imaging 40 (3) (2021) 781–792, http://dx.doi.org/10.1109/TMI.2020.3036584.

[17] S.P. Yadav, S. Yadav, Image fusion using hybrid methods in multimodality medical images, Med. Biol. Eng. Comput. 58 (2020) 669–687, http://dx.doi.org/10.1007/5117-020-02136-6.

[18] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017, http://dx.doi.org/10.1109/CVPR.2016.90.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, arXiv:2010.11929.

[21] P. Xu, X. Zhu, D.A. Clifton, Multimodal learning with transformers: A survey, 2022, arXiv:2206.06488.

[22] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, ACM Comput. Surv. 54 (10s) (2022) 1–41, http://dx.doi.org/10.1145/3505244.

[23] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, Z.-J. Zha, A battle of network structures: An empirical study of CNN, transformer, and MLP, 2021, arXiv:2108.13002.

[24] J. Shi, X. Zheng, Y. Li, Q. Zhang, S. Ying, Multimodal neuroimaging feature learning with multimodal multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease, IEEE J. Biomed. Health Inf. 22 (1) (2018) 173–183, http://dx.doi.org/10.1109/JBHI.2017.2655720.

[25] H.-I. Suk, S.-W. Lee, Dinggang Shen & The Alzheimer's Disease Neuroimaging Initiative, Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis, Brain Struct. Funct. 221 (5) (2016) 2569–2587, http://dx.doi.org/10.1007/s00429-015-1059-y.

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

[27] T. Wang, X. Chen, X. Zhang, S. Zhou, Q. Feng, M. Huang, Multi-view imputation and cross-attention network based on incomplete longitudinal and multimodal data for conversion prediction of mild cognitive impairment, 2023, arXiv:2206.08019.

[28] T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, D. Shen, Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data, Med. Image Anal. 60 (2020) 101630, http://dx.doi.org/10.1016/j.media.2019.101630.

[29] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, Multi-modal classification of Alzheimer's disease using nonlinear graph fusion, Pattern Recognit. 63 (1) (2017) 171–181, http://dx.doi.org/10.1016/j.patcog.2016.10.009.

[30] X. Bi, X. Hu, H. Wu, Y. Wang, Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest, IEEE J. Biomed. Health Inf. 24 (10) (2020) 2973–2983, http://dx.doi.org/10.1109/JBHI.2020.2973324.

[31] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M.J. Fulham, ADNI, Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, IEEE Trans. Biomed. Eng. 62 (4) (2015) 1132–1140, http://dx.doi.org/10.1109/TBME.2014.2372011.

[32] Z. Ning, Q. Xiao, Q. Feng, W. Chen, Y. Zhang, Relation-induced multi-modal shared representation learning for Alzheimer's disease diagnosis, IEEE Trans. Med. Imaging 40 (6) (2021) 1632–1645, http://dx.doi.org/10.1109/TMI.2021.3063150.

[33] B. Lei, P. Yang, T. Wang, S. Chen, D. Ni, Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis, IEEE Trans. Cybern. 47 (4) (2017) 1102–1113, http://dx.doi.org/10.1109/TCYB.2016.2644718.

[34] F. Yang, H. Wang, S. Wei, G. Sun, Y. Chen, L. Tao, Multi-model adaptive fusion-based graph network for Alzheimer's disease prediction, Comput. Biol. Med. 153 (2023) 106518, http://dx.doi.org/10.1016/j.compbiomed.2022.106518.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, 2020, http://dx.doi.org/10.48550/arXiv.2005.12872.

[36] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol. 12901, 2021, pp. 109—119, http://dx.doi.org/10.1007/978-3-030-87193-2_11.

[37] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, 2021, http://dx.doi.org/10.48550/arXiv.2102.04306.

[38] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol. 9351, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[39] Y. Xie, J. Zhang, C. Shen, Y. Xia, CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol. 12903, 2021, pp. 171–180, http://dx.doi.org/10.1007/978-3-030-87199-4_16.

[40] Y. Dai, Y. Gao, F. Liu, TransMed: Transformers advance multi-modal medical image classification, Diagnostics 11 (8) (2021) 1384, http://dx.doi.org/10.3390/diagnostics11081384.

[41] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: International Conference on Neural Information Processing Systems, Vol. 32, 2019, pp. 13–23, http://dx.doi.org/10.48550/arXiv.1908.02265.

[42] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: Pre-training of generic visual-linguistic representations, in: International Conference on Learning Representations, 2020, http://dx.doi.org/10.48550/arXiv.1908.08530.

[43] C. Li, M. Yan, H. Xu, F. Luo, W. Wang, B. Bi, S. Huang, SemVLP: Vision-language pre-training by aligning semantics at multiple levels, 2021, arXiv:2103.07829.

[44] M.W. Weiner, P.S. Aisen, C.R. Jack Jr., W.J. Jagust, J.Q. Trojanowski, L. Shaw, A.J. Saykin, J.C. Morris, N. Cairns, L.A. Beckett, A. Toga, R. Green, S. Walter, H. Soares, P. Snyder, E. Siemers, W. Potter, P.E. Cole, A.D.N.I. Mark Schmidt, The Alzheimer's disease neuroimaging initiative: Progress report and future plans, Alzheimer's Dementia 6 (3) (2010) 202–211, http://dx.doi.org/10.1016/j.jalz.2010.03.007.

[45] J. Jovicich, S. Czanner, D. Greve, E. Haley, A. van der Kouwe, R. Gollub, D. Kennedy, F. Schmitt, G. Brown, J. MacFall, B. Fischl, A. Dale, Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data, Neuroimage 30 (2) (2006) 436–443, http://dx.doi.org/10.1016/j.neuroimage.2005.09.046.

[46] P.A. Narayana, W.W. Brey, M.V. Kulkarni, C.L. Sievenpiper, Compensation for surface coil sensitivity variation in magnetic resonance imaging, Magn. Reson. Imaging 6 (3) (1988) 271–274, http://dx.doi.org/10.1016/0730-725X(88)90401-8.

[47] J. Sled, A. Zijdenbos, A. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Trans. Med. Imaging 17 (1) (1998) 87–97, http://dx.doi.org/10.1109/42.668698.

[48] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, D. Hawkes, Nonrigid registration using free-form deformations: Application to breast MR images, IEEE Trans. Med. Imaging 18 (8) (1999) 712–721, http://dx.doi.org/10.1109/42.796284.

[49] M. Jenkinson, C.F. Beckmann, T.E.J. Behrens, M.W. Woolrich, S.M. Smith, FSL, NeuroImage 62 (2) (2012) 782–790, http://dx.doi.org/10.1016/j.neuroimage.2011.09.015.

[50] S. Liu, C. Yadav, C. Fernandez-Granda, N. Razavian, On the design of convolutional neural networks for automatic detection of alzheimer's disease, in: Machine Learning for Health NeurIPS Workshop, vol. 116, 2020, pp. 184–201, http://dx.doi.org/10.48550/arXiv.1911.03740.

[51] 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, http://dx.doi.org/10.1093/jamia/ocac168.

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

[53] C. Wang, Y. Wei, J. Li, X. Li, Y. Liu, Q. Hu, Y. Wang, Asymmetry-enhanced attention network for Alzheimer's diagnosis with structural magnetic resonance imaging, Comput. Biol. Med. 151 (2022) 106282, http://dx.doi.org/10.1016/j.compbiomed.2022.106282.

[54] H. Cai, Q. Zhang, Y. Long, Prototype-guided multi-scale domain adaptation for Alzheimer's disease detection, Comput. Biol. Med. 154 (2023) 106570, http://dx.doi.org/10.1016/j.compbiomed.2023.106570.

[55] Q. Zhu, B. Xu, J. Huang, H. Wang, R. Xu, W. Shao, D. Zhang, Deep multi-modal discriminative and interpretability network for Alzheimer's disease diagnosis, IEEE Trans. Med. Imaging 42 (5) (2023) 1472–1483, http://dx.doi.org/10.1109/TMI.2022.3230750.