Received: 4 January 2023 | Revised: 1 May 2023 | Accepted: 24 May 2023

DOI: 10.1049/ipr2.12841

IET Image Processing

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Multimodal predictive classification of Alzheimer's disease based on attention-combined fusion network: Integrated neuroimaging modalities and medical examination data

Hui Chen | **Huiru Guo** | **Longqiang Xing** | **Da Chen** | **Ting Yuan** | **Yunpeng Zhang** | **Xuedian Zhang**

Key Laboratory of Optical Technology and Instrument for Medicine, Ministry of Education, University of Shanghai for Science and Technology, Shanghai, China

Correspondence
Hui Chen, Key Laboratory of Optical Technology and Instrument for Medicine, Ministry of Education, University of Shanghai for Science and Technology, Shanghai, China.
Email: chenhui@usst.edu.cn

Funding information
National Natural Science Foundation of China (NSFC), Grant/Award Number: 62275156; Medical engineering cross project: auxiliary detection system for Alzheimer's disease.

Abstract

Early diagnosis of Alzheimer's disease (AD) plays a key role in preventing and responding to this neurodegenerative disease. It has shown that, compared with a single imaging modality-based classification of AD, synergy exploration among multimodal neuroimages is beneficial for the pathological identification. However, effectively exploiting multimodal information is still a big challenge due to the lack of efficient fusion methods. Herein, a multimodal fusion network based on attention mechanism is proposed, in which magnetic resonance imaging (MRI) and positron emission computed tomography (PET) images are converted into feature vectors with the same dimension, while the demographic information and clinical data are preprocessed and converted into feature vectors through embedding. This attention model can focus on important feature points, fuse the multimodal information more effectively, and thus provide accurate diagnosis and prediction for different pathological stages. The results show that the model achieves an accuracy of 84.1% for triple classification tasks in normal cognition (NC) versus mild cognitive impairment (MCI) versus AD and 93.9% prediction accuracy in stable MCI (sMCI) versus progressive MCI (pMCI). In contrast to the existing multimodal diagnosis methods, our model yields a state-of-the-art accuracy of AD diagnosis, which is powerful and promising in clinical practice.

1 | INTRODUCTION

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases in the elderly, which is characterized by the symptoms of cognitive impairment and memory loss. It is a progressive pathophysiological process with insidious onset and irreversible damage of the brain. Mild cognitive impairment (MCI) is a transitional state between normal cognition (NC) and AD, which can be considered as an early stage of AD. To date, there is not a definite cure for a subject that diagnosed as AD. Although the early detection and intervention of AD is promising to prevent it, due to the lack of a clear understanding of the etiology, early identification of AD remains a great challenge [1]. Currently, early diagnosis and intervention of AD by medical experts is based on their subjective assessment through the use of multiple neuroimaging modalities such as magnetic resonance imaging (MRI) as well as the clinical records such as age, gender, blood pressure. However, the assessments undoubtedly vary among different experts, depending on their professional experiences.

Recently, the increasing development of machine learning has brought new vitality to the research of AD diagnosis and treatment, especially at an early stage. By using machine learning-based algorithm, quantitative features can be extracted from neuroimaging modalities to construct a robust, objective and automatic system for the assistant diagnosis of AD [2]. These machine learning tools are capable of earlier detection and accurate prediction of AD. Furthermore, machine

learning can describe the relationship between the input medical measurements and clinical results, which contributes to guide clinical decision-making. Given a continuous and progressive process of AD, the diagnosis is aimed to identify or predict NC, MCI and AD patients. The main purpose of the existing research is to improve the classification accuracy. Classification frameworks have been extensively studied using a single imaging modality of MRI. Qing Li et al. [3] developed a deep learning model to distinguish AD or MCI from NC based on the multi-feature kernel supervised within-Class-similar discriminative dictionary learning algorithm (MKSCDDL), which achieved an accuracy of 98.18% for the classification of AD and NC, 78.50% for the classification of MCI and NC, and 74.47% for the classification of AD with MCI. In comparison to single modality, multimodal methods have recently exhibited greater advantages for improving the classification accuracy and the understanding of AD patterns. This is reasonable since different modalities can capture AD information from different perspectives. For instance, positron emission tomography (PET) provides complementary information to MRI, which is able to obtain sensitive measurements of cerebral metabolic rates of glucose via the diffusion of radioactive agents. Manhua Liu et al. [4] proposed to construct cascaded convolutional neural networks to learn the multi-level and multimodal features of MRI and PET brain images for AD classification. Experimental results showed that the proposed method achieves an accuracy of 93.26% for classification of AD versus NC, which was 5%−9% higher than the results using unimodal MRI or PET, and 82.95% for classification pMCI versus NC, which was 4%−5% higher than the results using unimodal MRI or PET. These facts demonstrate that the combination of modalities may be an effective approach to increase the overall classification performance, thus deserving to be intensively studied. It is worth noting that, most existing methods have been proposed to fuse the multimodal features from different neuroimaging modalities through the direct splicing [5]. However, these methods lack their interpretability required to clearly explain the specific meaning of extracted features and how the feature fusion of different modalities can further contribute to the classification and prediction with the quantitative point of view. To make the multimodal fusion more effective and persuasive, more efforts are urgently needed to be devoted [6–11].

In addition, medical examination and evaluation data (MED), including the demographic information, clinical symptom scores and genetic risk factor conferred by APOE4, can refine the prediction of the machine learning model and improve the confidence of the model in the cognition of the pathological state, thereby making more robust prediction. Meanwhile, the acquisition of MED is more convenient and cost-effective over the imaging modalities. To the best of our knowledge, however, there are few studies in which different neuroimaging modalities and MED are combined to construct a auxiliary diagnosis model.

Taking all the above mentioned into consideration, in this paper, a novel classification method based on the unique combination of two neuroimaging modalities and MED is presented, aiming to investigate the effective fusion of multimodal data, and simultaneously improve the accuracy of Alzheimer's disease diagnosis as well as predict the progression of cognitive impairment to dementia. The multimodal Alzheimer's disease problem studied in this article is mainly based on the comprehensive judgment of multiple pathological examination results [PET, MRI, and data] of the same case x. That is, the final result is $y = f(\text{PET}, \text{MRI}, \text{data})$. To be specific, 3D convolutional neural networks are first utilized to extract key features from MRI and PET images and convert them into image feature vectors with the same latitude, while the demographic information, clinical symptom scores and genetic data are normalized and converted into data feature vectors through embedding. Both image feature vectors and data feature vectors are further fused using an attention mechanism approach to generate the latent multimodal correlation features of the MRI/PET images and MED. Finally, these learned features are combined by a fully connected layer followed by the softmax layer to achieve the classification of different stages of AD and further predict the progression trend of MCI, serving as upstream and downstream tasks, respectively. This work presents three distinct advantages: (1) Multimodal data fusion with a self-attentive mechanism is more effective to align and fuse different modal data features in high-dimensional space; (2) Task migration between upstream AD/MCI/NC classification model and downstream pMCI/sMCI model improves the generalization of this classification task, and meanwhile, reduce training costs; (3) Explore to discover the internal connection of different modes and build multi-modal feature map to realize multi-modal joint diagnosis

2 | RELATED WORKS

With the advancement of medical technology research more and more people recognize that Alzheimer's disease is caused by the interaction of multiple pathological factors, and the diagnosis of disease development by analyzing different factors becomes the key of current research. Multimodal representation learning based on deep learning has received much attention in recent years, which has powerful multi-level abstraction representation capability to narrow the heterogeneity gap between different modalities. Medical data from different modalities reflect different aspects of pathological changes in Alzheimer's disease, and using deep learning to fuse multimodal data to obtain complementary information can improve the accuracy of diagnosis and the interpretability of results. Applying multimodal analysis methods to multiple neuroimaging techniques targeting specific pathological processes will allow us to gain a comprehensive understanding of their relative roles, sequences, and causal relationships. There have been many breakthroughs in recent years in multimodal diagnosis of Alzheimer's disease, which are important to improve the diagnosis of neurodegenerative diseases and to understand the pathological processes that lead to the disease.

In order to combine information from multimodal data, the simplest approach is to directly connect high-dimensional features extracted from different modalities [12]. Specifically, first the features from different modalities are normalized and the

TABLE 1 Information of the subjects used in this work.

Class	Num	Age (avg)	Gender (M/F)	Education	APOE4 (0/1/2)	MMSE	ADAS (11/13)	RAVLT			
								immediate	learning	forgetting	perc_forgetting
AD	154 (45)	74 ± 8	29/16	15.3 ± 6	16/23/6	23.1 ± 1	11.4/17.2	26.2	4.0	4.1	54.6
NC	223 (57)	73 ± 10	28/29	16.4 ± 4	39/18/0	29.0 ± 2	9.3/11.3	38.3	6.3	3.8	44.6
pMCI	155 (33)	75 ± 7	20/13	16.9 ± 4	9/9/15	26.1 ± 2	11.3/17.2	33.3	4.4	4.1	54.1
sMCI	386 (92)	72 ± 8	57/35	15.7 ± 5	47/35/10	27.3 ± 3	11.1/16.5	36.6	4.7	4.0	49.5

features from each silent station are directly synthesized into a vector by concatenating them in series or in parallel, using the joint feature vector to train the classifier. However, simple concatenation does not optimally integrate the use of multimodal data, but rather Favors a single modality [13]. To effectively utilize multimodal information complementarily non-linear graph fusion methods can be used [13], and similarly multimodal multitask learning models can jointly predict multiple variables from multimodal data including brain images predicting continuous clinical scores [14] to subsequently determine AD status. In addition to predicting clinical scores, joint learning using multidomain regression and classification tasks [15] can also identify the transition from MCI to AD patients.

In recent studies, most existing methods use imaging data from a single time node to detect pathological changes in AD diagnosis [4, 16]. In fact, longitudinal data collected at follow-up time points often provide useful information about the pathological development of the disease [12, 17]. Consecutive examinations of medical data can make greater use of multimodal neuroimaging and genetic data to improve the accuracy of Alzheimer's disease diagnosis. Also, with the availability of longitudinal image data at multiple time points, it is possible to use them to improve the predictive power of Alzheimer's disease [18]. In addition to traditional medical imaging and bioinformatics that will be used for Alzheimer's multimodal data studies, acoustic, cognitive, and linguistic features can also be used for integrated multimodal learning [19].

3 | MATERIALS AND METHODS

3.1 | Datasets

The data used in the training phase of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI provides research data for those researchers all over the world, who are dedicated to the study of the progression of AD. 3T-MRI and PET image data used in this work were downloaded from the ADNI as pre-processed files. Dataset consisted of multiple sampling results from 227 subjects over three consecutive years, including 45 AD, 92 sMCI, 33 pMCI and 57 NC (see details in Table 1, with multiple test samples per subject). It is worth noting that, for each individual subject, MRI and PET images captured in the same period were collected as multimodal image Dataset. Correspondingly, demographic information and clinical data in the same period were also collected as supplementary information for each subject. Whether MCI subjects will progress to AD (known as pMCI) or not (known as sMCI) within a certain period is particularly vital in practice, therefore, the sMCI and pMCI data are labelled according to the status of case development over a three-year period.

MED includes the demographic information and clinical data. As shown in Table 1, demographic data comprise gender, age, and education level. A previous study [4] indicated that, age, gender, and education may be influenced the brain status, and therefore contribute to AD diagnosis. Clinical data includes APOE4 allele genotyping, Clinical Dementia Rating Scale (CDRS), Alzheimer's Disease Rating Scale, and Auditory Verbal Learning Test (RAVLT).

3.2 | Network architecture

The proposed method here can be divided into three stages, as shown in Figure 1. The first stage is the feature extraction from multimodal data, in which high-dimensional features are extracted from MRI and PET images by the ConvBlock module, respectively. ConvBlock consists of several parts: a 3D volume machine layer (Conv3d), BatchNorm layer, activation function (ELU), 3D pooling layer (MaxPool3d) and Dropout. In addition, genetic information conferred by APOE4, demographic data and clinical symptom scores are normalized and transformed into data feature vectors through embedding.

At the second stage, multimodal feature fusion based on attention mechanism is used to merge complementary information from different imaging modalities and MED, so that a more comprehensive description of the subjects can be obtained in contrast to the individual input data. First, the high-dimensional image features from MRI and PET are merged, while multiple data features from MED are processed in the same way. Subsequently, these two feature vectors are further fused in the middle layer by channel attention mechanism (CA) and spatial attention mechanism (SA). Specifically, the attention mechanism is utilized in the feature fusion stage, which makes the model focus more on the latent information that are decisive for the outcomes. As a consequence, this attention-based fusion method is capable of extracting more critical and important

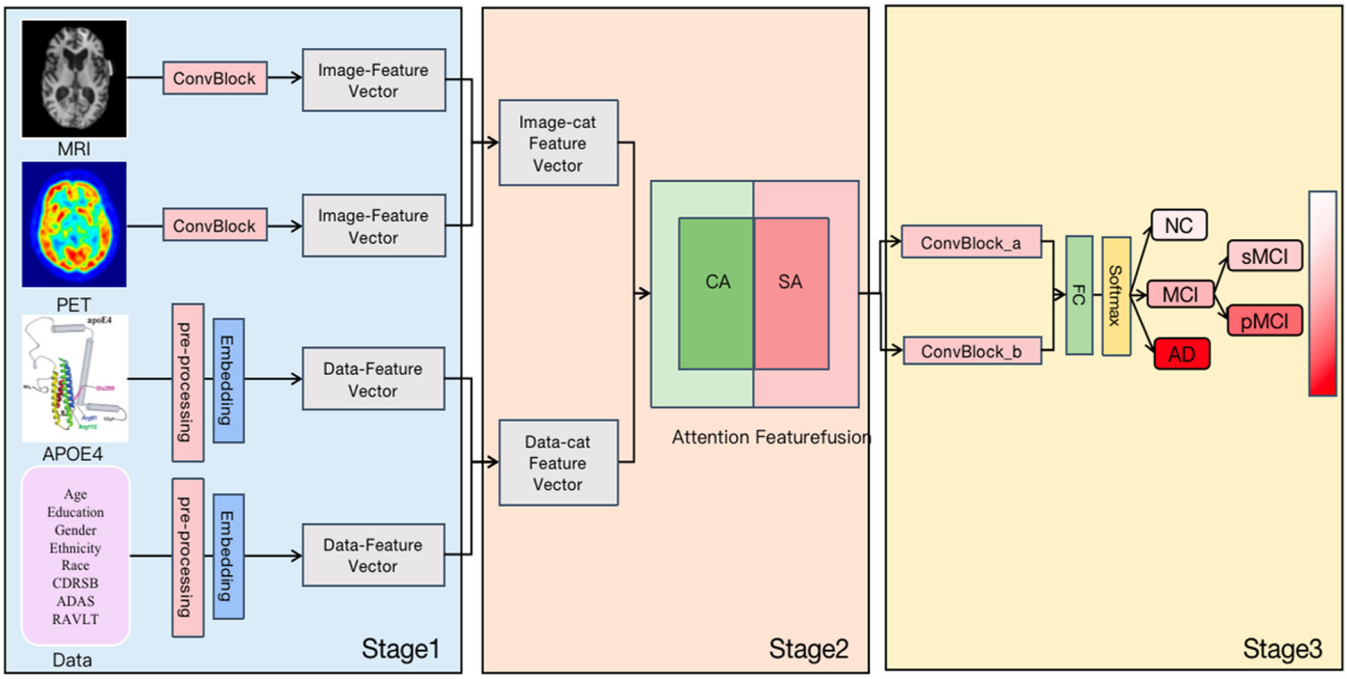

FIGURE 1 Proposed multimodal architecture for AD classification.

information, meanwhile improving the effective contributions of demographic information and clinical data as complementary to image features. Noteworthy, due to the fact that the number of feature points of demographic information and clinical data features are much less than that of medical image features, overfitting may occur during the process of feature fusion. We have provided a solution to overfitting in Section 3.4.1.

At the third stage, ConvBlock with a fully-connected (FC) layer and a softmax layer is adopted to obtain the final classification results through the recognition of fused features. During this process, the back propagation mechanism of the rolled-up network continuously optimizes the network parameters to improve the performance of the classification model in both the training and test cohorts. By using this classification model, we execute both binary (AD vs NC, AD vs MCI and MCI vs NC) and triple (AD vs MCI vs NC) classification tasks. In addition, the predictive binary classification is used to classify the MCI group into sMCI and pMCI groups.

To establish such classification model, we employ a 3D convolutional neural network (CNN) to efficiently extract rich spatial information from 3D brain images [4]. In this 3D CNN model, the 3D convolutional layer first convolves each input image with the learned kernel filter, then adds a bias term and applies a nonlinear activation function, and finally generates a feature vector through the filter. The 3D convolutional operation is defined by the formula:

$$u_{kj}^{l}(x,y,z) = \sum_{\delta_x}\sum_{\delta_y}\sum_{\delta_z} F_{k}^{l-1}(x+\delta_x, y+\delta_y, z+\delta_z) \times W_{kj}^{l}(\delta_x,\delta_y,\delta_z) \quad (1)$$

where x, y and z represent the pixel positions of a given 3D image. $W_{kj}^{l}(\delta_x,\delta_y,\delta_z)$ stands for the jth 3D kernel weight connecting the kth feature map of layer $l-1$ and the jth feature map of layer l. $F_{k}^{l-1}$is the kth feature map of the previous $l-1$ layer, and δ_x, δ_y, δ_z are the kernel sizes corresponding to the x, y and z coordinates. The output $u_{kj}^{l}(x,y,z)$ is the convolution response of the 3D kernel filter. After convolution, each convolution layer is followed by using the activation function Sigmoid:

$$F_{j}^{l}(x,y,z) = \text{Sigmoid}\left(b_{j}^{l} + \sum\nolimits_{k} u_{kj}^{l}(x,y,z)\right) \quad (2)$$

where b_{j}^{l} is the deviation term of the jth feature map of the lth layer. The 3D feature vector of the jth lth $F_{j}^{l}(x,y,z)$ layer is obtained by summing the feature maps of different convolution kernels. By capturing spatial correlation using 3D kernels, 3D CNNs can capture a large amount of spatial structure information from the high aspect of medical images, which is critical for feature representation of medical images.

3.3 | Attention mechanism

Attention mechanism has been widely used in different types of machine learning tasks, such as natural language processing, image recognition, speech recognition etc. The essence of attention is to highlight some important features based on the correlations between different parts of the input, which guides the model to reallocate the corresponding weights to each part of the input so that the model can gain reinforced learning ability without incurring computing and storage overheads.

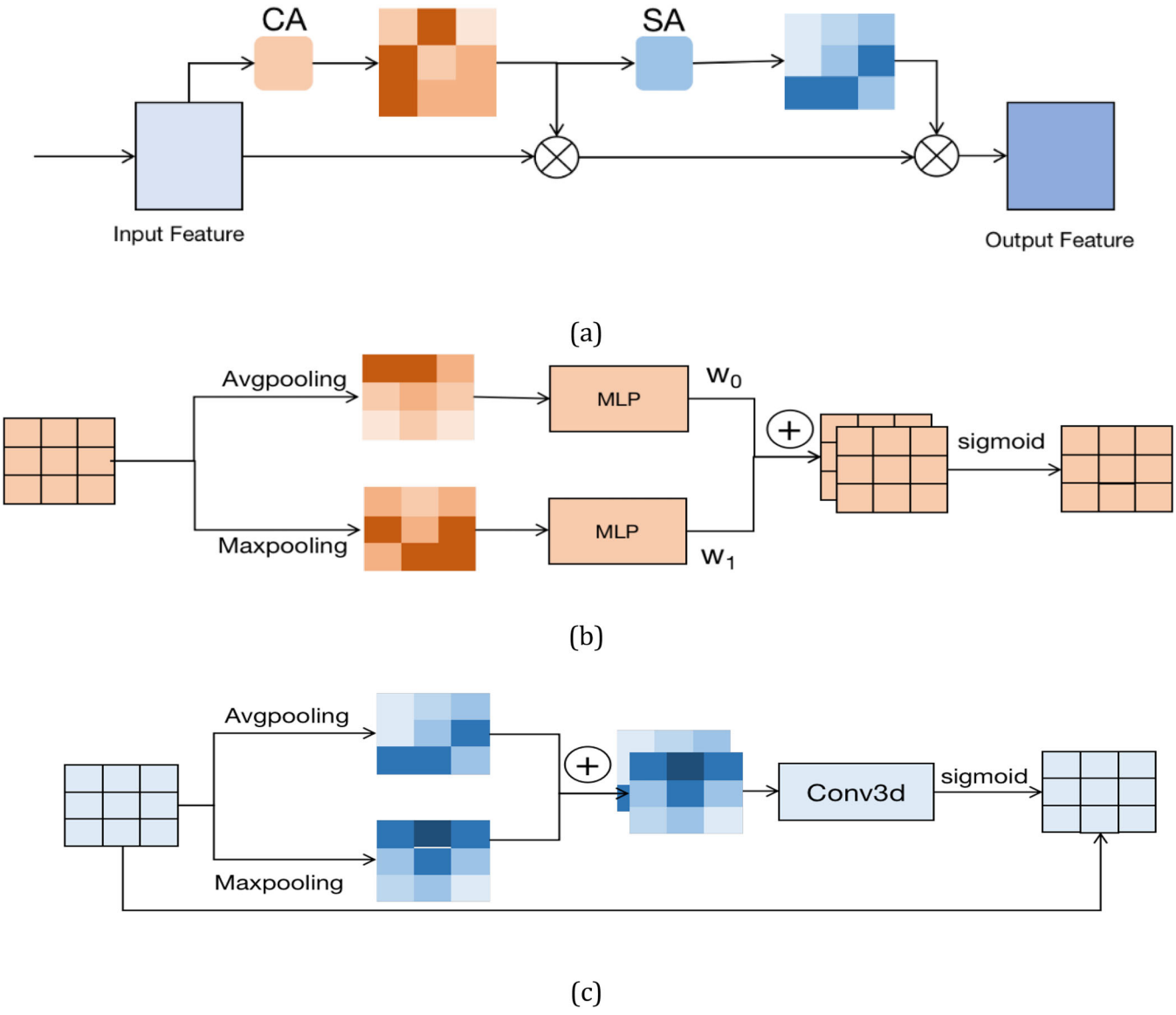

FIGURE 2 (a) Feature fusion based on attention mechanism;(b) channel attention; (c) spatial attention.

Figure 2a presents a schematic representation of the attention module. For a given input feature map, both the CA and SA are implemented. CA focuses on the content while SA focuses on the location. To better fuse images with different patterns, the pipelines of CA and SA are concatenated in our work with reference to Convolutional Block Attention Module (CBAM) [20].

Figure 2b shows a schematic diagram of the CA module. The spatial information of the feature mapping is aggregated using the average-pooling and max-pooling operations respectively, and two different spatial context descriptors F^c_{avg} and F^c_{max} correspondingly generated as average-pooled features and max-pooled features. The two descriptors are then passed through the MLP separately to generate the channel attention mapping $M_c \in R^{C\times1\times1}$,and the MLP output features are subjected to element-wise summation-based operation and then sigmoid activation operation to generate the final channel attention feature map. To reduce the parameter overhead, the hidden activation size is set to be $R^{C/r\times1\times1}$,where r is the scaling rate. After applying the shared network to each descriptor, the output feature vectors are merged based on element-by-element addition. The channel attention is calculated as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1(W_0(F^c_{avg})) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \end{aligned} \tag{3}$$

where σ identifies the sigmoid function, $W_0 \in R^{(C/r\times1\times1)}$. W_0 and W_1 are the weights of MLP respectively.

Figure 2c shows a schematic diagram of the SA module. First, the feature map, i.e., the output of the CA module, is used as the input feature map of SA module. Afterwards, global max pooling and global average pooling based on channel were performed to get F^s_{avg} and F^s_{max} respectively, and these two results

were further merged based on channel. Finally, the spatial attention feature is generated by multiplying this feature with the input feature of the module to generate the final feature. The spatial attention is calculated as follows:

$$M_s(F) = \sigma\left(f^{7\times7\times7}\left(\left[AvgPool(F); MaxPool(F)\right]\right)\right) = \sigma\left(f^{7\times7\times7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \quad (4)$$

where σ identifies the sigmoid function, and seven indicates the size of the convolution kernel. Unlike CBAM, herein, a 3D convolution was used.

3.4 | Evaluation

In this experiment, the model was used to execute the various classification tasks in both the training and test cohorts. The performance of the proposed model was assessed by using different evaluation measures, including the accuracy, specificity, sensitivity, F1-score. Note that, the general classification network is assessed by the accuracy, however, the practical project of predicting patients is more concerned with the recall rate. In addition, F1-score, the harmonic mean of precision and recall, was calculated to further characterize the model. These evaluation measures are defined as follow:

$$Accuracy_k = \frac{TP + TN}{total} \quad (5)$$

$$Specificity_k = \frac{TN}{TN + FP} \quad (6)$$

$$Sensitivity_k = Recall_k = \frac{TP}{TP + FN} \quad (7)$$

$$Precision_k = \frac{TP}{TP + FP} \quad (8)$$

$$F1_k = \frac{2 * precision_k * recall_k}{precision_k + recall_k} \quad (9)$$

$$F1 - Score = \frac{1}{n}\sum_{(k=0)}^{n} F1_k \quad (10)$$

where total is the total number of samples, and n is the number of categories, *TP*, *TN*, *FP*, and *FN* represent the number of true positives, true negatives, false positives, and false negatives, respectively. k represents the classification category. *F*1 score was finally obtained by computing an average of the *F*1*k* scores over classes.

All of the symbols used here are defined in the Appendix Table A1 for better clarity.

4 | EXPERIMENT AND RESULTS

4.1 | Experimental settings

All code was implemented using Python 3.7.4 and Pytorch 1.5.1. All experiments were executed under CUDA 10.0 and all experiments were performed using four 11GB NVIDIA Tesla V100 GPUs to train the models. Medical data features are generally sparse matrices with large gradient changes during training, so the network optimizer was set to Adam using adam for better interpretability of hyperparameter settings. The initial learning rate was set to 1e-2 [21] and the batch size was set to 8, the number of epochs was set to 50, and the training time for one task was 10 h. Additionally, the binary cross-entropy was employed as the loss function for the binary classification task, while the categorical cross-entropy was used for the ternary classification task. The classification module uses the Softmax function to compress different type label values within the range of [0,1], and the sum of all vectors after conversion is 1. Softmax has a gradient of 0 when the input is negative, which means it does not backpropagate the negative input, thus avoiding the gradient from disappearing. According to the following formula 11, Softmax preserves each value in a normalized manner.

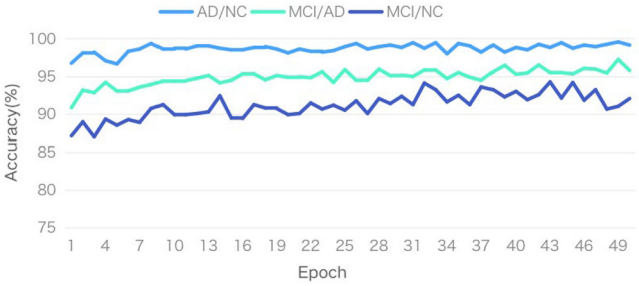

FIGURE 3 Accuracies in binary classification tasks using proposed multimodal framework.

$$\text{Softmax}\ (x) = \frac{\exp(x_i)}{\sum_i \exp(x_i)} \quad (11)$$

5-fold cross-validation strategy used to calculate the model performance, to avoid overfitting problems with limited datasets and simultaneously obtain a fairer comparison. The subjects were randomly divided into five subsets with one subset as the test cohort and other four as the training cohorts. We trained each experiment over 50 calendar hours and updated the learning rate using two strategies. The experimental results were presented as the mean ± SD (standard deviation) of fivefold tests.

4.2 | Performance

4.2.1 | Results for classification

The proposed attention mechanism-based multimodal fusion model was first utilized to execute the binary classification tasks. The classification performance was assessed in terms of the accuracy, sensitivity, specificity and *F*1 scores. As shown in Figure 3, the fivefold cross-validation corresponding to 50 Epoch results demonstrate that, in three binary classification tasks, that are AD versus NC, MCI versus AD, and MCI versus NC, the accuracies reach 97.90%, 92.84%, and 87.85%,

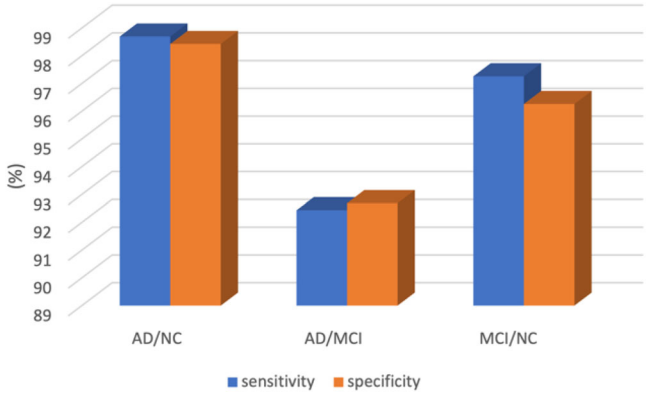

FIGURE 4 Sensitivities and specificizes in binary classification tasks using proposed multimodal framework.

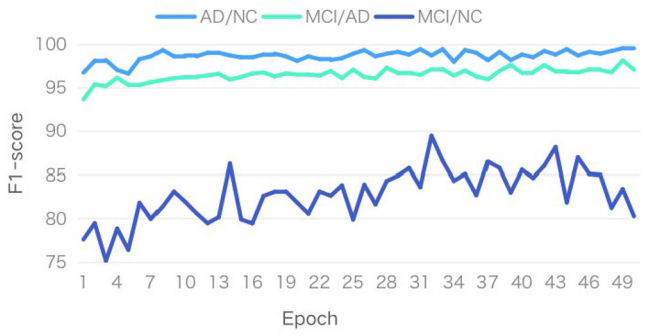

FIGURE 5 *F*1-scores in binary classification tasks using proposed multimodal framework.

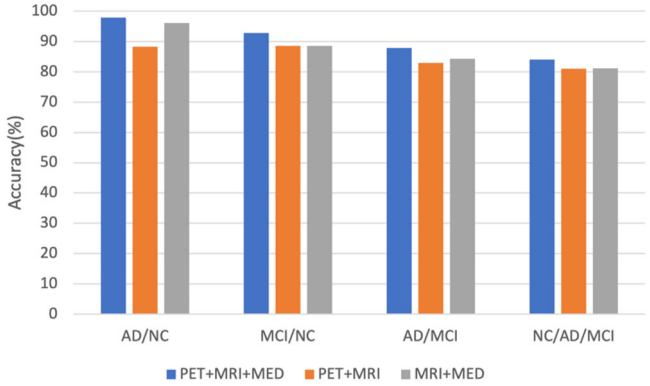

FIGURE 6 Accuracies of various classification tasks through the feature fusion of different data modalities.

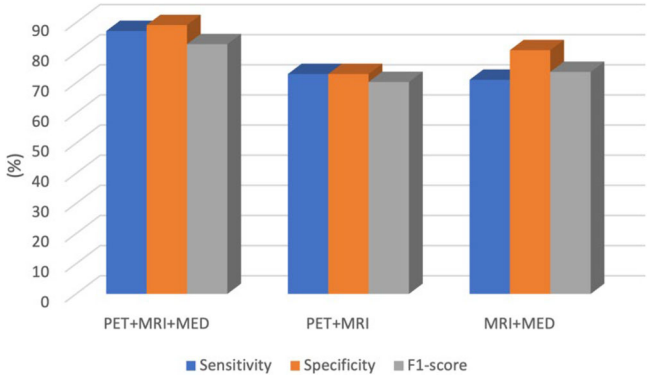

FIGURE 7 Sensitivity, specificity and *F*1 score in a triple classification task using different data modalities.

respectively. According to Figure 4, the model performance has the sensitivity of 98.69%, 97.26% and 92.43% and specificity of 98.42%, 96.26% and 92.68%in the classification of AD versus NC, MCI versus NC and AD versus MCI. Moreover, the model attains the *F*1 scores of 99.5% for AD versus NC, 98.1% for MCI versus AD, 89.5% for MCI versus NC (Figure 5). Given the fact that more data from sMCI were collected in this paper than that from pMCI, the accuracy and *F*1 score for the binary classification of MCI versus AD, are higher than that of MCI versus NC. The results demonstrate an excellent overall performance of the model.

The performance of our model in the triple classification of AD/MCI/NC was further estimated and the results are shown in Figure 6. After 50 Epochs of training the multimodal model triple classification (AD vs MCI vs NC) achieved an average accuracy of 84.10%.

In order to verify the performance of multimodal fusion in multiple tasks, the accuracy of different tasks was further tested by fusing data of different modalities. As shown in Figures 6 and 7, in both the binary and triple classification problems, that are AD versus NC, MCI versus. AD, MCI versus NC, as well as AD versus MCI versus NC, the model using MRI, PET and MED performs superiorly against other counterparts (MRI and PET, MRI and MED) in accuracy, sensitivity, specificity and *F*−1 scores. To be specific, by using MRI and PET modalities, the performance achieves a classification accuracy of 88.33%, 88.57%, 82.98% and 81.05% for AD versus NC, MCI versus AD, MCI versus NC and AD versus MCI versus NC, respectively. When using MRI and MED modalities, the model achieves a classification accuracy of 96.17%, 88.58%, 84.26% and 81.10% for AD vs. NC, MCI vs. AD, MCI vs. NC and AD vs. MCI vs. NC, respectively. However, the combination of MRI, PET and MED modalities effectively improves the classification accuracy, exhibiting 97.90%, 92.84%, 87.85% and 84.10% for AD versus NC, MCI versus AD, MCI versus NC and AD versus MCI versus NC, respectively. Similarly, the sensitivity, specificity and *F*−1 score for various classification tasks were obviously improved using MRI, PET and MED modalities in comparison with that using MRI and PET modalities or MRI and MED modalities (Figure 7). This is reasonable since MED can provide the complementary information to neuroimages and simultaneously the attention-based fusion network can effectively fuse the features extracted from images and MED. As such, the proposed model with imaging and MED modalities can effectively improve the classification accuracy of AD.

Finally, we analyze the effect of different data on the classification performance of the model by ablation experiments, and the results are shown in Table 2. According to the experiments, it is known that in PET features have a greater impact

TABLE 2 Results of the effect of different data on triple classification (AD/MCI/CN).

Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*1-score (%)
MRI+ Data	81.10	79.57	82.66	81.60
PET+Data	81.66	87.50	86.10	86.67
MRI+PET	81.05	79.32	80.03	79.88
MRI+PET+Data	81.10	84.67	84.70	81.16

TABLE 3 Results of the effect of different loss functions on triple classification (AD/MCI/CN).

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*1-score (%)
Adam	81.10	84.67	84.70	81.16
Adadelta	79.16	78.17	77.06	79.67
SGD	66.00	66.51	69.50	66.67
RMSprop	47.30	44.16	44.00	44.18

TABLE 4 Results of the effect of different modules on triple classification (AD/MCI/CN).

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*1-score (%)
NIN-	62.50	61.53	62.04	62.90
CA-	68.70	63.34	68.00	68.10
SA-	59.17	59.10	56.67	59.31
CA-SA-	68.33	58.89	63.68	66.81

on the results and are more accurate in analyzing the results. According to Table 3, by comparing the results of different loss functions on the classification results, Adam loss function is more suitable for this experiment. According to the experiments on the network modules CA, SA, and NIN in Table 4 ("CA-" i.e. remove the CA module from the original network structure), it is known that the SA module has a greater impact on the model performance. Afterwards, we compared the metrics of our method with previously reported multimodal diagnostic methods and the comparative results are shown in Table 5. It is easily shown that, our method achieves better performance compared to the previous studies in the classification of AD. It is especially noted that, the accuracy in the triple classification of AD versus MCI versus NC is significantly improved. This can be attributed to the proposed method has the best feature fusion and discrimination for various classification problems.

4.2.2 | Results for prediction

As is known, MCI is an intermediate stage between NC and AD, which can be divided into two subgroups sMCI and pMCI. Meanwhile, given the fact that the difference between sMCI and pMCI is very subtle, pMCI versus sMCI is considered to be very challenging. However, differentiating between pMCI from sMCI is of great importance to predict whether MCI subjects will progress to AD. Here, the features extracted from the upstream triple classification task (AD vs MCI vs NC) are also of benefit to differentiate pMCI and sMCI, which were further transferred to the classification of sMCI versus pMCI via incremental learning. The results shown in Figure 8 demonstrate that the model can discriminate a high intra-class difference for the classification of pMCI versus sMCI, which presents a classification accuracy of 93.94%, sensitivity of 96.28%, specificity of 95.51% and F1 score of 95.29% when combining MRI, PET and MED modalities. Figure 9 shows the performance of the model on the sMCI and pMCI classification tasks. The multimodal model combining MRI, PET and MED performs well overall and can be used to predict the progression of MCI. Meanwhile, the change trend of MCI can be used as a valid predictor of AD. Meanwhile the changing trend of MCI can be used as an effective predictor of AD.

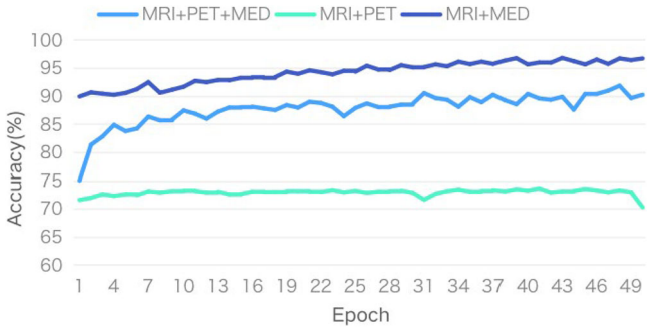

FIGURE 8 classification accuracy of sMCI and pMCI (50 epochs).

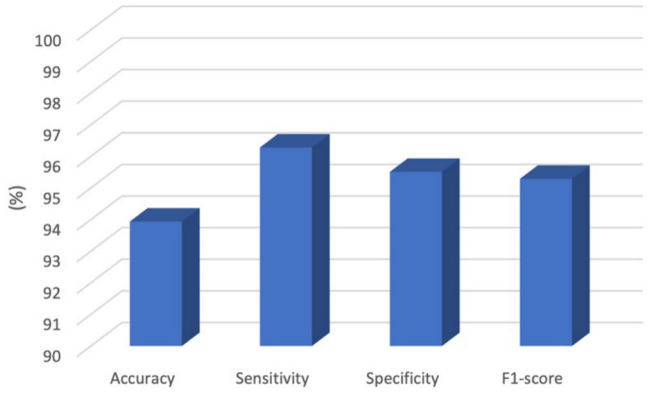

FIGURE 9 classification performance of sMCI versus pMCI (Average results of accuracy, sensitivity, specificity and F1 score).

5 | DISCUSSION

5.1 | Parameter analysis

Here, 3D images are used as network inputs, which requires a large number of convolutional layers to obtain the desired

TABLE 5 Classification accuracies comparison between previously reported methods and proposed method in different tasks.[7]

Method	Data sources	Dataset	Accuracy (%)			
			AD vs NC	MCI vs NC	MCI vs AD	AD vs MCI vs NC
Fang et al. [9]	ADNI	MRI+PET	87.69	80.70		66.29
Liu Q. et al. [22]	ADNI	MRI+PET	91.4	82.1	–	–
Li et al. [23]	ADNI	MRI+PET	91.4	77.4	–	–
Song et al. [24]	ADNI	MRI+PET	94.11	88.48	84.83	74.54
Our method	ADNI	MRI+PET	**88.33**	**82.98**	**88.57**	**81.05**
Tong et al. [8]	ADNI	MRI+PET+Data	91.8	79.5	–	60.2
Zhu et al. [25]	ADNI	MRI+PET+Data	88.02	84.14	–	–
Zhang et al. [7]	ADNI	MRI+PET+Data	96.58	90.11	97.43	–
Our method	ADNI	MRI+PET+Data	**97.90**	**92.84**	**87.85**	**84.10**

expressiveness. However, building such a large number of layers usually suffers from limited computational resources and memory. To solve this problem, depth-separable convolution was used to reduce model parameters and achieve a delicate network framework. Unlike the standard convolution, depth-separable convolution is implemented by Depthwise Convolution and Pointwise Convolution. To be specific, Depthwise Convolution is first used for each channel to convolve the output and the number of channels separately. Furthermore, a $1 \times 1 \times C$ convolution kernel (pointwise kernel) is used to obtain the final value.

The parameters of the separable convolution are calculated to be $192 \times 3 \times 3 \times 3 \times 128$, that are 663,552 parameters in total. In this calculation, the size of input feature map is $28 \times 28 \times 28 \times 192$.

In Depthwise Convolution, channels and convolution kernels correspond to each other. So, a three-channel image is computed to generate 192 Feature maps with 5184 parameters.

Pointwise Convolution does the convolution again for the three channels with convolution kernel $1 \times 1 \times 128$, and the same 128 Feature maps are output with the same output dimension as the regular convolution with 24576 parameters.

The number of parameters of Separable Convolution is much smaller than that of conventional convolution for the same input and Feature maps. Therefore, Separable Convolution can effectively reduce the network parameters under the premise of network determination.

5.2 | Optimized overfitting

Considering that medical images available for training the deep learning model are limited, the overfitting problem is easily encountered during the training. To overcome this problem, several strategies such as batch normalization, Dropout and K-fold are used to combat overfitting in this experiment.

Batch normalization performs a batch normalization operation on a 5D input (N,C,D,H,W) consisting of small batches of 3D data, calculates the mean and standard deviation of each dimension of the input, and normalizes the output of the layer

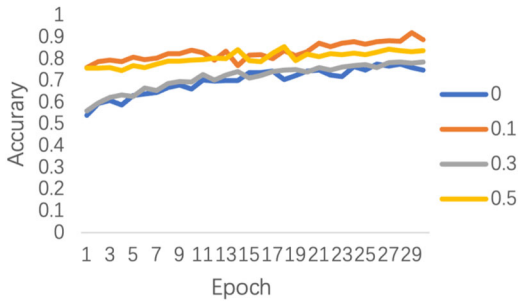

FIGURE 10 Accuracies of the classification of AD versus NC versus MCI with different Dropouts.

by subtracting the mean and dividing it by the standard deviation. This 3D normalization process enforces a fixed activation distribution, thus stabilizing and accelerating the training speed of the deep neural network.

During the training, Dropout was used to actively drop certain feature units and their connections, which keeps the network from relying on certain local features, improves the robustness of hidden neurons to random fluctuations, and learns useful information. Figure 10 demonstrates the effect of different Dropouts on the triple classification of AD versus NC versus MCI. It is clear that, the best performance of the triple classification was obtained when the Dropout was set to 0.1, so that the Dropout is set to 0.1 here.

Traditional dataset partitioning techniques split the data into training and test sets for training, and this static approach leads to the risk of overfitting on the test set, and the evaluation metrics are then not a true reflection of the model generalization performance. Further partitioning the dataset into training, validation and test sets reduces the number of samples used to learn the model, and the results depend on a specific random selection of the (training, validation) sets. To make the model more effectively learn from the limited data, the K-Fold cross validation method [26] is used in this paper, which is capable of improving the data utilization, and preventing the overfitting problem.

5.3 | Multitasking

Multitask learning can be considered as a form of transfer learning by sharing knowledge and model parameters among different tasks to improve the generalization ability and efficiency of the model. In multitask learning, the parameters of the model are designed to be shared by multiple tasks, and this sharing of parameters is considered as a form of transfer learning, where the knowledge learned by the model from one task can be transferred to other tasks. The premise of multi-task learning in model training is that there are multiple related tasks that need to be completed simultaneously and that there is some correlation between these tasks, which from a clinical perspective is more challenging for the diagnosis of MCI (pMCI vs sMCI) compared to AD detection (AD vs NC). Considering the intrinsic connection and common features between these two tasks, using multitask leaning to solve both tasks together is a promising approach [27]. In multitask learning, the model typically contains a shared underlying representation and multiple task-specific output layers. The underlying representation is a generic feature needed to learn all tasks, while the output layer is optimized for the specific task. The benefit of a shared underlying representation is that the model can share the learned knowledge across all tasks, thus increasing the generalization capability of the model. In addition, the model can better handle complex multi-task problems because the shared underlying representation can better capture the relationships between tasks. There are many advantages to using multi task learning. Multi task learning assumes that the features learned in multiple tasks are useful for all tasks, and there is a certain correlation between multiple tasks. By sharing these features, the generalization ability and efficiency of the model can be improved without increasing the number of training samples, thereby reducing the waste of training time and computing resources. Multi task learning can be regarded as a regularization technique, which limits the capacity of the model by sharing model parameters, so as to avoid overfitting problems. Multi task learning is regarded as a non-convex optimization problem, where a specific optimization algorithm can be used to optimize the objective functions of multiple tasks, resulting in an optimal shared model.

Unlike traditional migration learning methods, multitask learning assumes that there are multiple related tasks to be completed simultaneously, and that the generalization ability and efficiency of the model can be improved by sharing correlation and feature information between tasks, that is, the features learned in the upstream tasks for AD versus NC versus MCI classification are beneficial for refining MCI classification. Specifically, here, the classifier is divided into upstream and downstream tasks, with the upstream task focusing on AD versus NC versus MCI classification, while the downstream task performs pMCI versus sMCI classification. As can be seen from Table 6, the use of migration learning only requires modifying the network full-link layer so that the network is assigned shared parameters that are adapted to multiple tasks to train the model, which greatly saves model training time while ensuring model accuracy. The visual representations extracted from the AD versus NC versus MCI classification tasks are further transferred to pMCI versus sMCI classification through multitask learning, thus improving the proposed model's ability to generalize better in pMCI versus sMCI classification. As can be seen from Table 6, using migration learning only requires modifying the network full linkage layer to adapt the network assignment to train the model with shared parameters across multiple tasks, ensuring model accuracy while greatly saving model training time. Visual representations extracted from the AD versus NC versus MCI classification task are further transferred for pMCI versus sMCI classification through multitask learning, which improves a better generalization capability of the proposed model in the pMCI versus sMCI classification.

TABLE 6 Comparison of results (accuracy) and training time for different tasks.

Task	PET+MRI+ DATA	PET+ MRI	MRI+ DATA	Train time
NC/AD/MCI	84.10	81.05	81.10	10 h
sMCI/pMCI	93.94	74.05	81.14	4 h

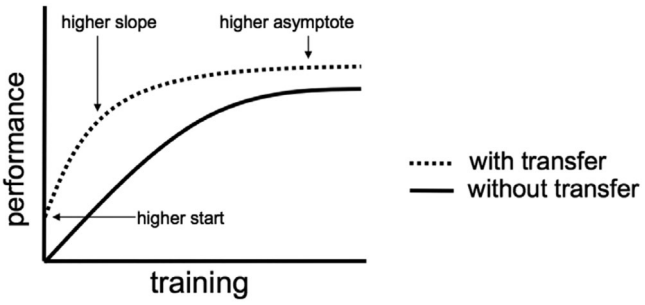

FIGURE 11 Ideal performance of transfer learning model [28].

Multitask learning is a type of migration learning that saves time and achieves better performance. Multitask learning takes more into account the intrinsic correlation between multimodal data and ideally performs as shown in Figure 11, where potential representations of each modality are learned independently in the upstream classification task, and shared model weights allow for a higher initial point of training for the downstream model. The downstream task relies on the existing capability of the model to be able to train the model at a faster rate, reduce the spatial distance of multimodal objects through label-aligned multitask feature selection, and the fusion capability of the trained model is better than direct training, further enhancing the generalization capability of the model.

6 | CONCLUSION

Here, a multimodal image feature fusion method based on a self-attentive mechanism combining neuroimages and MED is proposed for AD diagnosis. By using this model, two neuroimaging modalities, demographic information, clinical symptom scores and genetic data were combined to

generate features with different weights for classification tasks. The results demonstrate that the proposed method can consistently and significantly improve the classification and prediction performance in contrast to single modality-based methods. Specifically, the proposed method achieves 84.1% accuracy in the classification of AD, MCI and NC and 93.9% prediction accuracy for the progression of MCI (pMCI vs sMCI), which outperforms existing multimodal diagnostic methods, especially in the early diagnosis of AD. This outstanding performance is benefit from multimodal learned features and effective feature fusion. Moreover, the proposed model combines the neuroimaging diagnosis with the clinical diagnosis, which makes the whole diagnosis process is much closer to a clinician's diagnosis process. For the application scenario here, it is important to screen out all cases that may be AD, so that a high recall rate is also a key metric in addition to a high accuracy rate. As such, obtaining a network model with a robust accuracy-recall balance is also a valuable direction when we aim to further improve and optimize the diagnosis model in the future.

AUTHOR CONTRIBUTIONS

Hui Chen: Conceptualization; Huiru Guo: Methodology; Longqiang Xing: Data curation; Da Chen: Resources; Ting Yuan: Visualization; Yunpeng Zhang: Software; Xuedian Zhang: Supervision.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (NSFC) (no. 62275156); Medical engineering cross project: auxiliary detection system for Alzheimer's disease.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Alzheimer's disease neuroimaging initiative at http://adni.loni.usc.edu. These data were derived from the following resources available in the public domain: MRI/PET/Clinical Data at ADNI url: http://adni.loni.usc.edu.

ORCID

Huiru Guo https://orcid.org/0000-0001-7572-0764

REFERENCES

1. Jia, L.F., Du, Y.F., Chu, L., et al.: Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: A cross-sectional study. Lancet Public Health 5(12), e661–e671 (2020). https://doi.org/10.1016/S2468-2667(20)30185-7
2. Basaia, S., Agosta, F., Wagner, L., et al.: Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. Neuroimage-Clin. 21, 101645 (2019). https://doi.org/10.1016/j.nicl.2018.101645
3. Li, Q., Wu, X., Xu, L., Chen, K., Yao, L.: Classification of Alzheimer's disease, mild cognitive impairment, and cognitively unimpaired individuals using multi-feature kernel discriminant dictionary learning. Front. Comput. Neurosci. 11 (2018). https://doi.org/10.3389/fncom.2017.00117
4. Liu, M., Cheng, D., Wang, K., Wang, Y.: Ulti-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics 16(3–4), 295–308 (2018). https://doi.org/10.1007/s12021-018-9370-4
5. Zhou, T., Thung, K., Zhu, X., Shen, D.: Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum. Brain Mapp. 40(3), 1001–1016 (2018). Portico. https://doi.org/10.1002/hbm.24428
6. Lin, W., Gao, Q., Du, M., Chen, W., Tong, T.: Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. Comput. Biol. Med. 134, 104478 (2021). https://doi.org/10.1016/j.compbiomed.2021.104478
7. Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., Zhang, X.: Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. Neurocomputing 361, 185–195 (2019). https://doi.org/10.1016/j.neucom.2019.04.093
8. Tong, T., Gray, K., Gao, Q., Chen, L., Rueckert, D.: Nonlinear graph fusion for multi-modal classification of Alzheimer's disease. Lect. Notes Comput. Sci. 77–84 (2015). https://doi.org/10.1007/978-3-319-24888-2_10
9. Fang, C., Li, C., Forouzannezhad, P., Cabrerizo, M., Curiel, R.E., Loewenstein, D., Duara, R., Adjouadi, M.: Gaussian discriminative component analysis for early detection of Alzheimer's disease: A supervised dimensionality reduction algorithm. J. Neurosci. Methods 344, 108856 (2020). https://doi.org/10.1016/j.jneumeth.2020.108856
10. Thung, K.-H., Yap, P.-T., Shen, D.: Multi-stage diagnosis of Alzheimer's disease with incomplete multimodal data via multi-task deep learning. Lect. Notes Comput. Sci. Quebec Canada 160–168 (2017). https://doi.org/10.1007/978-3-319-67558-9_19
11. Liu, M., Zhang, J., Adeli, E., Shen, D.: Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Transactions on Biomed. Eng. 66(5), 1195–1206 (2019). https://doi.org/10.1109/tbme.2018.2869989
12. Zhou, T., Thung, K., Zhu, X., Shen, D.: Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum. Brain Mapp. 40(3), 1001–1016 (2018). Portico. https://doi.org/10.1002/hbm.24428
13. Tong, T., Gray, K., Gao, Q., Chen, L., Rueckert, D.: Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. Pattern Recognit. 63, 171–181 (2017). https://doi.org/10.1016/j.patcog.2016.10.009
14. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59(2), 895–907 (2012). https://doi.org/10.1016/j.neuroimage.2011.09.069
15. Cheng, B., Zhu, B., Xiong, J.: Multimodal multi-label transfer learning for early diagnosis of Alzheimer's disease. J. Computer Appl. 36(8), 2282-2286 (2016). http://doi.org/10.11772/j.issn.1001-9081.2016.08.2282
16. Lee, G., Nho, K., Kang, B., Sohn, K.-A., Kim, D., Alzheimer's Disease Neuroimaging Initiative: Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci. Rep. 9(1). (2019). https://doi.org/10.1038/s41598-018-37769-z
17. Candemir, S., Nguyen, X.V., Prevedello, L.M., Bigelow, M.T., White, R.D., Erdal, B.S., Neuroimaging Initiative, A. D. Munich, Germany: Predicting rate of cognitive decline at baseline using a deep neural network with multidata analysis. J. Med. Imaging 7(04), 44501 (2020). https://doi.org/10.1117/1.jmi.7.4.044501
18. Aderghal, K., Khvostikov, A., Krylov, A., Benois-Pineau, J., Afdel, K., Catheline, G.: Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning. 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). Karlstad, Sweden. (2018). https://doi.org/10.1109/cbms.2018.00067
19. Sarawgi, U., Zulfikar, W., Soliman, N., Maes, P.: Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity. Interspeech Shanghai 2020 (2020). https://doi.org/10.21437/interspeech.2020-3137
20. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. Lect. Notes Comput. Sci. 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1

21. Smith, L.N.: Cyclical learning rates for training neural networks. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, CA. (2017). https://doi.org/10.1109/wacv.2017.58
22. Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., ADNI: Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. 62(4), 1132–1140 (2015). https://doi.org/10.1109/tbme.2014.2372011
23. Li, F., Tran, L., Thung, K.-H., Ji, S., Shen, D., Li, J.: A robust deep model for improved classification of AD/MCI patients. IEEE J. Biomed. Health. Inf. 19(5), 1610–1616 (2015). https://doi.org/10.1109/jbhi.2015.2429556
24. Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., Shen, P.: An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. Front. Digit. Health 3 (2021). https://doi.org/10.3389/fdgth.2021.637386
25. Zhu, Q., Yuan, N., Huang, J., Hao, X., Zhang, D.: Multi-modal AD classification via self-paced latent correlation analysis. Neurocomputing 355, 143–154 (2019). https://doi.org/10.1016/j.neucom.2019.04.066
26. Sampath, R., Indumathi, J.: Earlier detection of Alzheimer disease using N-fold cross validation approach. J. Med. Syst. 42(11). (2018). https://doi.org/10.1007/s10916-018-1068-5
27. Oh, K., Chung, Y.-C., Kim, K.W., Kim, W.-S., Oh, I.-S.: Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. Sci. Rep. 9(1) (2019). https://doi.org/10.1038/s41598-019-54548-6
28. Torrey, L., Shavlik, J.: Transfer learning. In Olivas, E.S., Guerrero, J.D.M., Martinez-Sober, M., Magdalena-Benedito, J.R., Serrano López, A.J. (eds.) Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 242–264. (2010). https://doi.org/10.4018/978-1-60566-766-9

How to cite this article: Chen, H., Guo, H., Xing, L., Chen, D., Yuan, T., Zhang, Y., Zhang, X.: Multimodal predictive classification of Alzheimer's disease based on attention-combined fusion network: Integrated neuroimaging modalities and medical examination data. IET Image Process. 17, 3153–3164 (2023). https://doi.org/10.1049/ipr2.12841

APPENDIX

TABLE A1 Definition of the symbols.

Symbols	Definition
u^l_{kj}	Convolutional Response of L-layer 3D Kernel Filter
(x, y, z)	Represents the pixel position of a given 3D image
F^{l-1}_k	kth feature map of layer $l-1$
$\delta_x, \delta_y, \delta_z$	Convolution kernels corresponding to x, y, and z coordinates
W^l_{kj}	Connect the kth feature map of layer $l-1$ to the jth 3D kernel weight of the jth feature map of layer l
$F^l_j(x, y, z)$	The jth 3D characteristic diagram of layer l
Sigmoid	A Common activation function
b^j_j	Offset term of the jth characteristic graph of layer l
$F \in R^{(C/r\times1\times1)}$	Input characteristics
r	Dimension reduction ratio, the proportion of spatial dimension reduction performed when using the global pooling layer
$M_c(F), M_s(F)$	Channel attention module convolution, spatial attention module convolution
F^c_{avg}, F^c_{max}	Channel information descriptors: average pooling feature, maximum pooling feature
F^s_{avg}, F^s_{max}	Spatial information descriptors: average pooling feature, maximum pooling feature
MLP	Multi layer perceptron
AvgPool	Average pooling
MaxPool	Maximum pooling
$f^{7\times7\times7}$	Indicates that the filter size is 7 × seven × Convolution operation of 7
$Accuracy_k$	Accuracy rate of category k: the proportion of the predicted correct quantity in positive and negative cases to the total quantity
$Specificity_k$	Category k specificity: The proportion of negative cases identified as negative cases to all negative cases, which measures the classifier's ability to recognize negative cases
Sensitivity_k	Sensitivity of category k: The proportion of pairs in all positive cases, which measures the classifier's ability to recognize positive cases and is numerically equal to the recall rate $Recall_k$
$Recall_k$	Recall rate of category k: the proportion of correctly predicted positive cases in the total actual positive case samples
$Precision_k$	Accuracy rate of category k: the proportion of True positives in the identified images
$F1_k$	$F1$ value of category k: an indicator that neutralizes accuracy and recall
TP	Positive samples predicted by the model as positive
TN	Negative samples predicted by the model as negative classes
FP	Negative samples predicted by the model as positive
FN	Positive samples predicted as negative by the model