Computers in Biology and Medicine 184 (2025) 109438

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# Multi-scale multimodal deep learning framework for Alzheimer's disease diagnosis

Mohammed Abdelaziz [a,c], Tianfu Wang [a,*], Waqas Anwaar [b], Ahmed Elazab [a,d]

[a] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 518060, China
[b] Medical Ultrasound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 518060, China
[c] Department of Communications and Electronics, Delta Higher Institute for Engineering and Technology (DHIET), Mansoura, 35516, Egypt
[d] Computer Science Department, Misr Higher Institute of Commerce and Computers, Mansoura, 35516, Egypt

ARTICLE INFO

*Keywords:*
Alzheimer's disease
Multimodal data
Multi-scale representation
Convolutional neural network

ABSTRACT

Multimodal neuroimaging data, including magnetic resonance imaging (MRI) and positron emission tomography (PET), provides complementary information about the brain that can aid in Alzheimer's disease (AD) diagnosis. However, most existing deep learning methods still rely on patch-based extraction from neuroimaging data, which typically yields suboptimal performance due to its isolation from the subsequent network and does not effectively capture the varying scales of structural changes in the cerebrum. Moreover, these methods often simply concatenate multimodal data, ignoring the interactions between them that can highlight discriminative regions and thereby improve the diagnosis of AD. To tackle these issues, we develop a multimodal and multi-scale deep learning model that effectively leverages the interaction between the multimodal and multiscale of the neuroimaging data. First, we employ a convolutional neural network to embed each scale of the multimodal images. Second, we propose multimodal scale fusion mechanisms that utilize both multi-head self-attention and multi-head cross-attention, which capture global relations among the embedded features and weigh each modality's contribution to another, and hence enhancing feature extraction and interaction between each scale of MRI and PET images. Third, we introduce a cross-modality fusion module that includes a multi-head cross-attention to fuse MRI and PET data at different scales and promote global features from the previous attention layers. Finally, all the features from every scale are fused to discriminate between the different stages of AD. We evaluated our proposed method on the ADNI dataset, and the results show that our model achieves better performance than the state-of-the-art methods.

## 1. Introduction

Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder that can significantly impact daily life of elderly [1]. The Alzheimer's Association reports that at least 50 million individuals worldwide are currently living with AD. Furthermore, it is projected that by 2050, the number of people living with AD in the United States will have more than doubled [2]. As the disease progresses, patients often experience a significant decline in cognitive function, including memory loss, difficulty with language, and impaired thinking skills [3]. The impact of AD is not limited to the individual affected, as family members and caregivers also bear a substantial burden in providing care and support for the patients [4]. Even though no therapy helps stop the progression of AD, an early diagnosis of AD is still significant for future therapies that aim to avoid developing cognitive symptoms. Additionally, early AD detection and mild cognitive impairment (MCI) progression predictions are both essential for developing effective interventions and improving patient care [5].

The development of neuroimaging techniques has allowed for the acquisition of both anatomical and functional information about the human brain using various imaging modalities, such as magnetic resonance imaging (MRI) and positron emission tomography (PET) [6]. MRI provides detailed features on the structural integrity of brain tissues and offers high contrast between gray matter and white matter, while PET captures functional changes in the brain, providing valuable insights into metabolic processes. Consequently, MRI and PET are often used

* Corresponding author.
*E-mail address:* tfwang@szu.edu.cn (T. Wang).

https://doi.org/10.1016/j.compbiomed.2024.109438
Received 15 September 2024; Received in revised form 27 October 2024; Accepted 12 November 2024
Available online 22 November 2024

together to study the anatomical differences between the brains of AD patients and normal controls (NC) [7].

Recent studies, including those utilizing machine learning and deep learning techniques, have been developed to differentiate among different stages of the disease with the aid of the rich information in multimodal images enabling more precise and early detection of AD. For example, Goel et al. [8] fused multimodal data, including MRI and PET, using wavelet transform. The random vector functional link then utilized features extracted from ResNet-50 for the early detection of AD. Zhang et al. [9] introduced a feature selection method utilizing multimodal data, including MRI, PET, functional MRI, and diffusion tensor imaging, to identify the most discriminative features for AD classification. Tu et al. [10] developed two deep learning techniques. The first network was a generative adversarial network for PET generation, while the second fused multimodal data, including MRI and PET, for AD classification. Liu et al. [11] proposed a new fusion method for AD diagnosis that combined multimodal and multi-level data from MRI and PET. They used two types of convolutional neural network (CNN) networks to extract high-level features from patched data obtained from both imaging modalities. The extracted features were then combined at a compact level for classification.

Cross-modality attention explores and leverages complementary information from multiple modalities to enhance interaction between them. By aligning and integrating features from different sources, cross-modality attention directs attention to discriminative regions, thereby improving the diagnosis of AD. For example, Lu et al. [12] developed a technique using MRI, genetic, and clinical data to predict the progression from MCI to AD. They introduced a hierarchical attention mechanism to assign weights to each modality and understand the interactions between them. Zhang et al. [13] developed a multi-modal cross--attention algorithm utilizing MRI, PET, and CSF biomarkers to distinguish between AD, MCI, and NC. Their cross-modal attention mechanism integrated both imaging and non-imaging information, while also minimizing discrepancies between the modalities, resulting in improved diagnostic performance. Liu et al. [14] used a cascaded modality transformer architecture with cross-attention to integrate different types of data, including MRI, gender, age, and the mini-mental state examination score for disease classification.

Although previous approaches have demonstrated promising results, significant challenges still remain in diagnosing AD using multimodality data. One key challenge is that these approaches often focus on the interaction between multimodal data at a single scale, while ignoring the interactions across multiple scales of multimodal data [13,15,16]. However, multiscale approaches provide discriminative features that single-scale methods often miss because large-size images contain more detailed and small-scale features, while small-size images capture large-scale positional features. This combination enhances the network's adaptability, reduces scale inconsistency problems, and improves the performance of AD diagnosis [17,18]. Additionally, the heterogeneity between multimodal data, which exhibits minimal correlations, make it difficult to effectively fuse low-level features from different biomarkers for DL techniques [10,19]. Moreover, most existing learning methods still rely on manual pre-defined ROIs, which may not take into account individual differences or capture all of the atrophy-related features spread throughout the entire brain. As a result, many learning methods currently in use rely on domain knowledge and expert input to determine informative regions or patches within MRI to create diagnostic models since most regions offer limited useful information for distinguishing between healthy and diseased brains [20–22]. As the early stage of AD only causes minor structural changes in the brain, developing an effective end-to-end model for AD classification without guidance remains challenging.

To address the challenges mentioned above, we propose a multimodal and multi-scale deep learning model (MMSDL) to improve the diagnosis of AD. The proposed method is trained on MRI and PET images via an end-to-end approach. Specifically, MMSDL comprises three main parts, i.e., modality embedding, multimodal scale fusion (MSF), and cross-modality fusion (CMF). Firstly, we embed each scale of the multimodal images using a CNN, which extracts image feature representations that are learned and transformed into a vector format. Secondly, we propose an MSF that includes a series of multi-head self-attention and multi-head cross-attention mechanisms. The multi-head self-attention mechanism is designed to effectively capture the global interdependencies among these features. Furthermore, the multi-head cross-attention mechanism enables the weighing of the contribution of each modality to the other. Finally, we utilize a cross-modality fusion (CMF) module that includes multi-head cross-attention to fuse MRI and PET data at every scale, leveraging both data types at various scales of the network to improve the model's ability to detect differences between the various stages of AD.

In contrast to the current approaches. Our distinctive contributions can be summarized as follows.

1. We propose a multi-scale, multimodal learning approach that captures complex structural and metabolic changes in brain MRIs and PET scans, respectively, improving prediction accuracy by leveraging local and global features to reduce noise and variability across different scales and modalities.
2. Our method integrates an MSF mechanism with multi-head self-attention to understand intra-modality relationships and multi-head cross-attention for inter-modality interactions. These could enhance feature extraction, offering valuable insights into AD progression and significantly improving diagnostic accuracy.
3. We introduce a CMF module that fuses MRI and PET data at various scales using multi-head cross-attention, leveraging complementary features, to improve feature representation and enhance the detection of various stages of AD.

The remainder of this paper is organized as follows. The related work is presented in section 2. The materials and methods are introduced in section 3. The experimental results and discussion are given in section 4. Lastly, the conclusion is summarized in section 5.

## 2. Related work

### 2.1. Multimodal AD diagnosis with deep learning

In recent research, a variety of studies have focused on classifying AD by leveraging multimodal data [23,24]. For instance, Zhou et al. [25] designed a multimodal deep learning framework aimed at the early diagnosis of AD. This framework addresses the heterogeneity between different types of data, such as MRI and PET scans, by projecting them into a shared latent space. Utilizing these unified latent representations, the framework employs multiple diversified classifiers to enhance the accuracy and reliability of AD classification. Tu et al. [26] introduced a new fusion method that integrates multimodal data, including MRI, patient profiles, gene sequences, and mental state examination data, for AD diagnosis. Initially, they developed a feature transformation technique to reduce heterogeneity among the different data modalities. Subsequently, they combined the transformed features with MRI-derived features to enhance the accuracy of AD diagnosis. Ning et al. [27] proposed a bi-directional mapping technique to project multimodal data—specifically, MRI and PET into a shared representation space. This shared representation is then further projected into another space known as the label space, which is utilized for classification tasks. Yang et al. [28] introduced two distinct models for AD diagnosis utilizing multimodal data, including genotype and phenotype data. The first model is a spectral graph attention model that learns varying weight representations among nodes. The second model is a bilinear aggregation model, designed to improve the degree of abnormalities across different population categories. Additionally, they developed an adaptive fusion module that dynamically integrates the

outputs of both models to improve AD prediction accuracy. Dwivedi et al. [29] utilized multimodal data, including MRI and PET, for AD diagnosis. Initially, they employed a discrete wavelet transform to fuse the MRI and PET images. Subsequently, the fused data was used in ResNet-50 for feature learning. Finally, the extracted features were used for classification employing a robust energy-based least square twin support vector machine classifier. Odusami et al. [30] enhanced the classification of AD using MRI and PET scans by modifying a ResNet18 network to learn the multimodal data. They introduced a 3-channel phase feature learning model which enabled the use of the maximum number of samples, subsequently improving the accuracy of AD diagnosis. Tang et al. [31] developed a 3D CNN to learn multimodal data from MRI and PET scans for the purpose of classifying the disease stages of AD. They enhanced the correlation among the multimodal features using an enhanced Transformer model. Additionally, they integrated the discriminative features extracted from both MRI and PET scans to identify the stages of AD accurately. Wu et al. [32] developed a deep learning network that incorporated a CNN and convolutional block attention modules. This approach aimed to extract discriminative features from multimodal neuroimaging data, specifically MRI and PET scans, to improve the accuracy of AD diagnosis.

### 2.2. Cross-modality attention

Recently, various studies have used a straightforward approach by combining multimodal data through the direct concatenation of features from different modalities. However, there is increasing interest in employing cross-modal fusion techniques to integrate multimodal data more effectively, rather than just concatenating them [33,34]. For instance, Cheng et al. [15] developed a deep learning framework to diagnose AD using MRI and PET imaging. Their framework included two distinct feature extractors: the first, called a feature similarity discriminator, was designed to capture discriminative features from multimodal data. The second, known as the mutual attention feature fusion module, enabled interactions between features, allowing for the adjustment of feature weights within each modality. Zhang et al. [16] introduced a new approach for the early detection of AD by employing adversarial learning to align MRI and PET features into a unified representation space, followed by using a transformer-based cross-attention mechanism to effectively fuse these features, enhancing the diagnostic process. Yu et al. [35] proposed a deep learning approach that utilized MRI and PET scans to predict the progression of cognitive decline. Initially, they used a generative adversarial network to generate missing PET samples. Subsequently, a multimodality feature extraction module was employed to extract highly learned features from both the MRI and PET data. These features were then applied to inter-modality and intra-modality feature selection using a multimodality fusion module, ultimately for classification. Dai et al. [36] introduced an AD classification framework that utilized MRI scans and non-imaging data, such as age and mini-mental state examination scores. They developed two distinct encoders to extract high-level features: the first encoder used a CNN to learn MRI data, while the second employed a linear encoder for the non-imaging data. A joint attention module was then applied to integrate these multimodal data, enabling interaction between them, and the extracted features were inputted into a multi-layer perceptron for classification. Sun et al. [37] employed a multimodal dataset comprising MRI and electronic health records to predict brain degeneration accurately by designing an encoder-decoder framework. Initially, they developed two distinct encoders to integrate the complementary information through both inter- and intra-modality interactions. Subsequently, the extracted features were fed into a decoder module to identify the most discriminative features for predicting brain degeneration. Wang et al. [38] designed a multi-task deep learning framework utilizing MRI and PET imaging to predict the progression of MCI. Initially, they employed adversarial learning to address missing samples within the multimodal dataset. Following this, they integrated two cross-attention blocks. These blocks were designed to leverage associations between the multimodal data, thereby enhancing the accuracy of predicting MCI conversion. Leng et al. [39] developed a multimodal cross-enhanced fusion network, a patch-based 3D CNN that used MRI and PET scans to diagnose AD and its early stages. The network comprised two modules: the first extracted features at various scales, while the second emphasized the correlation and complementarity between the features from MRI and PET, enabling automatic detection of discriminative structural and metabolic regions.

## 3. Materials and methods

### 3.1. Subjects

In our study, we utilize the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to evaluate our model with a sample of 956 subjects, including both MRI and PET scans. Table 1 provides a detailed overview of the demographic and clinical characteristics of these subjects. Our dataset includes 263 individuals with NC, 272 with stable MCI (sMCI), 203 with progressive MCI (pMCI), and 218 diagnosed with AD.

### 3.2. Data preprocessing

In this study, we followed the pipeline of Zhou et al. [40] to preprocess the MRI and PET data. For MRI preprocessing, we used the MIPAV program to correct the AC-PC, followed by the N3 algorithm to correct the bias field. The brain extraction technique in Ref. [41] was used to strip the skull, and the cerebellum was subsequently removed. We then registered the skull-stripped image with the Montreal Neurological Institute template [42]. In contrast, For PET preprocessing, we applied an affine registration to align each PET data with its corresponding T1 MR image.

### 3.3. Overview of the proposed method

Fig. 1 illustrates the proposed MMSDL framework, which comprises three main components: image embedding, MSF, and CMF. Initially, we employ three different scales at varying resolutions: $64 \times 64 \times 64$, $32 \times 32 \times 32$, and $16 \times 16 \times 16$, from both MRI and PET scans. These scales capture both local and global complementary features of the disease. The larger scale ($64 \times 64 \times 64$) captures the overall structure, while the smaller scales ($32 \times 32 \times 32$ and $16 \times 16 \times 16$) focus on finer details and boundaries. Utilizing multiple scales allows the network to extract rich multi-scale spatial features from multimodal data, thereby enhancing the diagnosis of AD.

Specifically, we employed a CNN to obtain the vector embedding from each scale of each modality. These embedded features were then passed to the MSF module, which includes three multimodal attention sub-modules for each scale. This design enables both inter- and intra-modality interactions within the same scale, as well as intra-scale interactions. Consequently, this approach captures both local and global interactions, thereby highlighting discriminative regions. Next, we passed the features from each scale to the CMF module to enable interaction between MRI and PET at each scale, achieving a deeper

**Table 1**
Characteristics of the ADNI dataset utilized in this study (mean ± standard deviation).

| | NC | sMCI | pMCI | AD | Total |
|---|---|---|---|---|---|
| M/F | 136/127 | 167/105 | 117/86 | 124/94 | 544/412 |
| Age (years) | 74.10 ± 5.76 | 72.41 ± 7.38 | 73.26 ± 7.36 | 74.85 ± 7.82 | 73.61 ± 7.12 |
| MMSE | 29.03 ± 0.69 | 27.39 ± 1.11 | 26.62 ± 1.01 | 23.36 ± 1.36 | 26.76 ± 2.30 |
| CDR-GLOB | 0 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.77 ± 0.16 | 0.42 ± 0.29 |

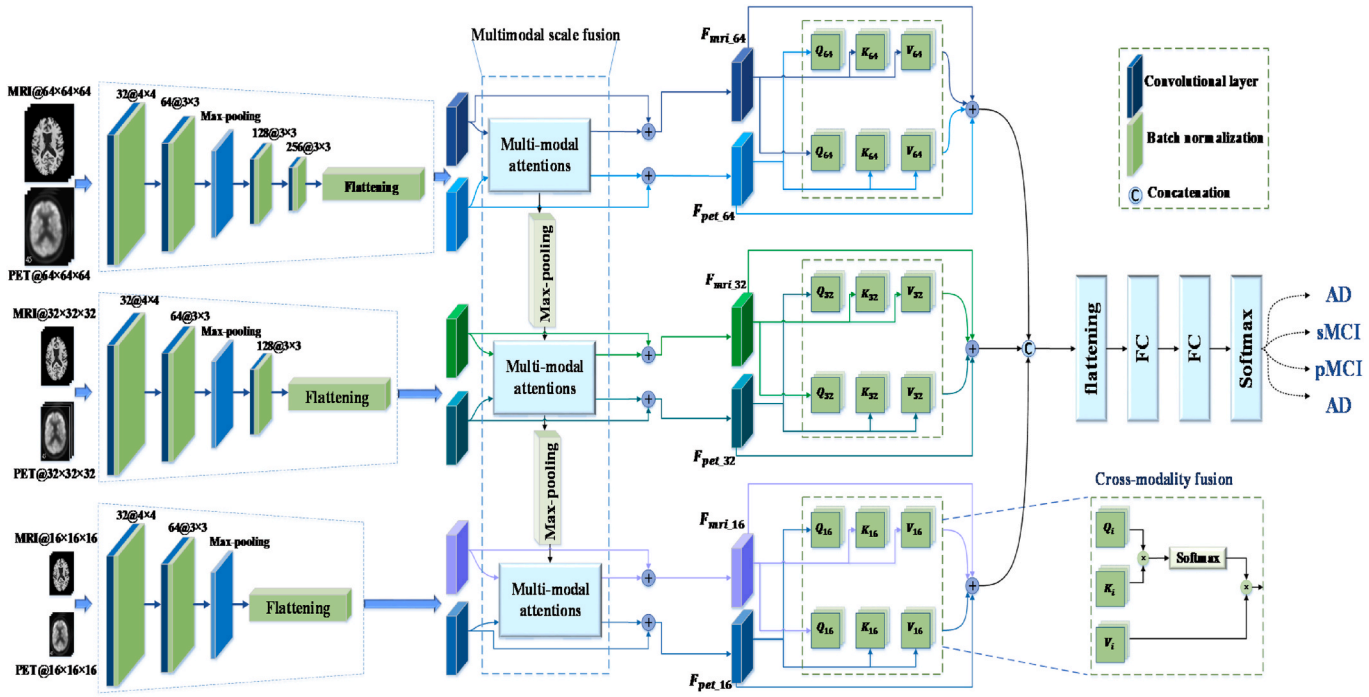


**Fig. 1.** Illustration of the proposed MMSDL architecture which is made up of three main parts; modality embedding, MSF, and CMF.

feature representation and obtaining highly discriminative features. Ultimately, the different scales were fused and passed through fully connected layers, followed by flattening for classification.

### 3.4. Image embedding

We employed a series of convolutional layers to convert each subject from its raw format into appropriate vector embeddings, considering input sizes of 64 × 64 × 64, 32 × 32 × 32, and 16 × 16 × 16. For the 64 × 64 × 64 input size, we utilized four convolutional layers, progressively increasing the number of channels from 32 to 256. For the 32 × 32 × 32 input size, we employed three convolutional layers, increasing the number of channels from 32 to 128. For the 16 × 16 × 16 input size, we used two convolutional layers, incrementing the number of channels from 32 to 64. Ultimately, the feature maps obtained from the final convolutional block were flattened into sequences, yielding 256, 128, and 64 features for the 64 × 64 × 64, 32 × 32 × 32, and 16 × 16 × 16 input sizes, respectively.

The initial convolutional layer utilized a kernel size of 4 × 4 to effectively capture larger spatial features and reduce the spatial dimensions efficiently, while the subsequent convolutional layers utilized a filter size of 3 × 3 to extract more detailed and localized features, enhancing the resolution and detail of the feature maps. To reduce the input size, we utilized a max pooling layer with a filter size of 2 × 2 and a stride length of 2 units. It is worth noting that all convolutional layers were trained with non-zero-padding feature maps using a stride length of 1 unit. Additionally, after each convolutional layer, we applied batch normalization and rectified linear unit activations to ensure the stability and effectiveness of the learning process.

### 3.5. Multimodal scale fusion

The MSF is a cascaded of multimodal attentions modules that include multi-head self-attention and multi-head cross-attention networks, as depicted in Fig. 2. The multi-head self-attention module obtains a query input Q from either the previous layers or the latent feature, depending on whether it is the initial multi-head self-attention layer or not, to create a representation that emphasizes the most relevant features

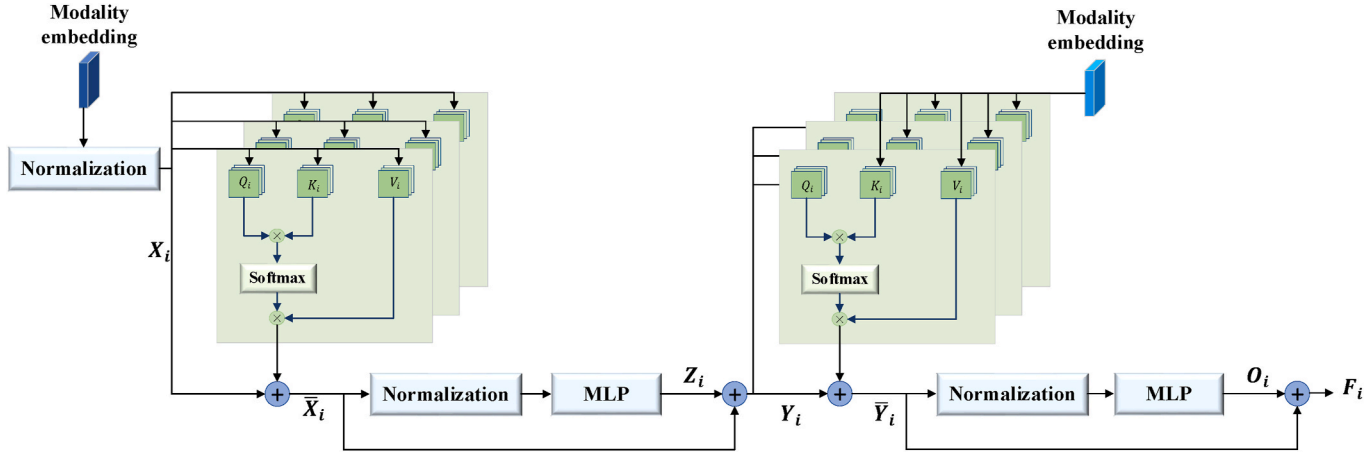


**Fig. 2.** Multimodal attentions module that has a cascaded multi-head self-attention and multi-head cross-attention network.

across inter-scale and inter-modality. This is followed by a multi-head cross-attention layer, which facilitates interactions within intra-scale and intra-modality, enhancing the model's ability to capture complex dependencies and further refine feature representation. Subsequent to each network, a multilayer perceptron (MLP) is employed, comprising two feed-forward networks augmented with the Gaussian error linear unit activation function to inject inductive bias into the network, thus addressing the lack of inductive bias in the self-attention operation.

Given an input sequence of length $L$ (from the output of the previous multimodal attentions module or from $Q_l$ if it is the first modality in the sequence), denoted as $X = \{x_1, x_2, \ldots, x_L\}$, we first transform each input element into a query vector, a key vector, and a value vector using learnable weight matrices:

$$Q = W_Q \times X \tag{1}$$

$$K = W_K \times X \tag{2}$$

$$V = W_V \times X \tag{3}$$

where $W_Q$ , $W_K$, and $W_V$ are learnable weight matrices that transform each input element $X$ into a query vector $Q$, a key vector $K$, and a value vector $V$, respectively.

Next, the attention weights between each pair of Q and K vectors are calculated as follows:

$$Score(Q, K) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{4}$$

where $d_k$ is the dimensionality of the K vector, and the dot product between the query vector $Q$ and the key vector $K$ is normalized by the square root of $d_k$.

The attention weights are then Softmax-normalized over all the key vectors as follows:

$$Attention\ (Q, K, V) = Similarity(Q, K) \times V \tag{5}$$

where $K = \{k_1, k_2, \ldots, k_L\}$ and $V = \{v_1, v_2, \ldots, v_L\}$ are the key and value matrices, respectively. The $Attention\ (Q, K, V)$ represents the weighted sum of the value vectors, where the attention weights are computed based on the similarity between $Q$ and $K$. Next, we pass the $Attention\ (Q, K, V)$ vectors through an MLP with two fully connected layers:

$$Z = MLP\ (Attention\ (Q, K, V)) \tag{6}$$

Then, we compute the output vector $Y$ as follows:

$$Y = Z + \overline{X} \tag{7}$$

where $Z$ is the output of the MLP and $\overline{X}$ is the original input vector $X$ enhanced by adding the self-attended output. This is known as the residual connection, which allows information to flow directly from the input to the output, helping to mitigate the vanishing gradient problem.

Then, we pass the output sequence $Y$ to the cross attention as Q, while the K and V from the embedded sequence. Subsequently, the output from the cross-attention mechanism is passed through an MLP) with two fully connected layers. The final cross-attention output $O$ is computed as the sum of the MLP outputs and $Y$. Then, the final output vector $F$ is computed using:

$$F = O + \overline{Y} \tag{8}$$

where $\overline{Y}$ represents the output sequence $Y$ enhanced by a cross-attended output, forming a residual connection that allows direct information flow from input to output to mitigate the vanishing gradient problem. Also, we perform all the previous steps to each scale in each modality to produce six outputs including, $F_{mri\_64}$ , $F_{pet\_64}$, $F_{mri\_32}$, $F_{pet\_32}$, $F_{mri\_16}$, $F_{pet\_16}$ sequentially from left to right as shown in Fig. 1.

### 3.6. Cross-modality fusion

The proposed module incorporates several sub-modules to enhance the feature representation and accuracy of the model. In the third sub-module, a new fusion module named CMF was developed based on a multi-head cross-attention mechanism to effectively combine the encoded features from every scale, as shown in Fig. 1. Specifically, the MSF module generates two types of features for each scale, which are then passed to the CMF. In the CMF, the cross-attention layer obtains the output sequence from the MRI scale as K and V and the Q from the PET scale, and vice versa, to identify the discriminative features from each modality and effectively combine them. Additionally, to further improve the feature representation, the cross-attentioned output is passed through an MLP with two fully connected layers for each scale, which apply non-linear transformations and refinements to capture detailed and high-level features. After extracting features from different scales, the resulting features are fused together and then fed into a series of fully connected layers for classification.

## 4. Results and discussion

### 4.1. Experimental settings

We performed experiments on a computer system with Ubuntu, 506 Nvidia GTX Titan Xp x2, and i7-6800K installed, using the Keras library with TensorFlow as the backend. We demonstrated the effectiveness of our approach by implementing different configurations for three binary tasks (NC vs. AD, NC vs. MCI, and sMCI vs. pMCI). We then compared our results with those obtained from several state-of-the-art machine learning and deep learning techniques. We evaluated the effectiveness of the proposed method using different classification metrics, including accuracy (ACC), specificity (SPE), sensitivity (SEN), precision (PRE), receiver operating characteristic (ROC), and F1-score (F1).

In our experiments, we initialized the network randomly with a mean of 0 and a standard deviation (SD) of 1 to avoid any bias from predetermined values. We used binary cross-entropy as the loss function and the Adam optimizer for its fast and efficient convergence during training. We set the number of epochs to 50, selected a batch size of 64, assigned a learning rate of $10^{-4}$, and determined the value of k to be 10 for k-fold cross-validation. We conducted ten independent experiments and computed the mean and standard deviation of the results to enhance the reliability and robustness of our findings.

### 4.2. Comparison with the conventional methods

Table 2 compares the performance of the MMSDL approach with several deep learning techniques, including ResNet [43], ResNext [44], MobileNet [45], ShuffleNetv2 [46], and EfficientNet [47], where the best method is in boldface and the second-best is underlined. Fig. 3 presents the ROC curves for the proposed method and these deep learning models, highlighting their overall performance. Across all three classification tasks—NC vs. AD, NC vs. MCI, and sMCI vs. pMCI—our proposed method demonstrates superior performance across various metrics, such as ACC, SEN, SPE, F1-score, and AUC. For the NC vs. AD task, MMSDL achieves the highest ACC (95.25 %), F1-score (95.74), and AUC (94.68), outperforming all conventional methods. In the more challenging NC vs. MCI classification, MMSDL continues to perform the best, achieving the highest ACC (85.22 %) and SPE (94.00 %), while maintaining a competitive AUC (81.36). Even in the most complex task of sMCI vs. pMCI, MMSDL shows notable improvements over baseline methods, with an ACC of 75.16 % and a SPE of 93.81 %, outperforming others in these key metrics.

However, there are inherent challenges in interpreting AUC in tasks like NC vs. MCI and sMCI vs. pMCI due to the subtle distinctions between classes and the potential class imbalances. AUC measures the trade-off between true positives and false positives across different thresholds,

**Table 2**
Classification comparison between the proposed method and different conventional methods for three different tasks (mean ± SD %).

| Symbol | Quantity | ACC | SEN | SPE | F1 | AUC |
|---|---|---|---|---|---|---|
| NC vs. AD | ResNet50 | 91.20 ± 3.26 | 97.15 ± 3.13 | 86.54 ± 7.10 | 91.35 ± 3.81 | 93.34 ± 1.20 |
| | ResNext | 91.89 ± 3.20 | 96.91 ± 2.63 | 87.76 ± 6.72 | 92.10 ± 3.57 | 93.77 ± 1.13 |
| | MobileNet | 86.04 ± 2.43 | 90.09 ± 3.36 | 82.04 ± 3.14 | 86.80 ± 2.32 | 86.75 ± 2.37 |
| | ShuffleNetv2 | 90.14 ± 3.07 | **99.09 ± 1.29** | 83.05 ± 5.73 | 90.08 ± 3.40 | 94.14 ± 1.13 |
| | EfficientNet | 88.61 ± 1.68 | 97.62 ± 1.49 | 81.16 ± 2.45 | 88.62 ± 1.85 | 89.62 ± 2.14 |
| | MMSDL | **95.25 ± 0.72** | 97.35 ± 2.00 | **92.71 ± 3.14** | **95.74 ± 0.61** | **94.68 ± 0.86** |
| NC vs. MCI | ResNet50 | 77.81 ± 3.00 | 80.47 ± 8.79 | 77.63 ± 3.34 | 61.62 ± 7.85 | 81.14 ± 3.19 |
| | ResNext | 78.81 ± 4.85 | 86.98 ± 9.69 | 77.63 ± 5.15 | 60.86 ± 14.15 | **85.63 ± 2.47** |
| | MobileNet | 79.78 ± 5.07 | **92.72 ± 11.59** | 79.46 ± 8.93 | 61.36 ± 16.36 | 84.27 ± 1.93 |
| | ShuffleNetv2 | 78.85 ± 3.19 | 78.13 ± 11.17 | 80.95 ± 4.67 | 66.46 ± 8.00 | 80.25 ± 3.54 |
| | EfficientNet | 74.68 ± 2.92 | 67.38 ± 8.87 | 80.65 ± 4.11 | 63.59 ± 4.66 | 75.08 ± 2.06 |
| | MMSDL | **85.22 ± 3.91** | 69.36 ± 14.93 | **94.00 ± 3.70** | **76.21 ± 8.71** | 81.36 ± 6.26 |
| sMCI vs. pMCI | ResNet50 | 69.35 ± 3.56 | **86.48 ± 11.19** | 66.82 ± 3.91 | 48.82 ± 12.20 | 76.06 ± 3.69 |
| | ResNext | 68.72 ± 5.03 | 84.08 ± 9.03 | 66.64 ± 5.04 | 46.92 ± 17.16 | 75.32 ± 3.24 |
| | MobileNet | 69.12 ± 16.34 | 71.94 ± 21.00 | 84.98 ± 11.37 | **69.09 ± 7.73** | **80.78 ± 3.79** |
| | ShuffleNetv2 | 65.92 ± 5.01 | 65.15 ± 10.00 | 70.75 ± 6.54 | 56.34 ± 11.99 | 68.31 ± 3.56 |
| | EfficientNet | 66.60 ± 5.50 | 62.17 ± 7.35 | 72.73 ± 4.53 | 61.71 ± 6.34 | 68.00 ± 4.18 |
| | MMSDL | **75.16 ± 5.82** | 50.20 ± 17.16 | **93.81 ± 8.85** | 61.90 ± 13.04 | 71.87 ± 6.61 |

making it more sensitive to misclassifications in borderline cases, especially in imbalanced datasets. While MMSDL shows a slightly lower AUC in this task, its higher ACC and SPE indicate that it correctly classifies the majority of cases, outperforming others in these key metrics. This demonstrates that our method can effectively handle the complexities of challenging cases while maintaining strong performance in metrics like ACC and SPE. The use of multi-modality fusion and multi-scale feature extraction enables MMSDL to capture subtle differences between classes better than other models, proving its ability to compete effectively in challenging classification tasks.

Furthermore, Fig. 4 compares the t-SNE visualizations from various deep learning models—ResNet50, ResNext, EfficientNet, MobileNet, ShuffleNetv2, and MMSDL—applied for the three different classification tasks. The results highlight clear differences in the models' ability to separate class clusters. ResNet50 and ResNext show moderate separation, with some overlap, particularly in distinguishing between NC and MCI, as well as sMCI vs. pMCI. EfficientNet demonstrates significantly better separation, especially for NC vs. AD, due to its architecture's ability to scale effectively across multiple dimensions, allowing for more efficient feature extraction and class distinction. This enables EfficientNet to capture subtle differences between normal and diseased brain images more effectively than ResNet-based models, which, while deep, may not scale as well across various parameters. In contrast, MobileNet and ShuffleNetv2 exhibit scattered clusters with substantial overlap, especially in more complex comparisons like sMCI vs. pMCI, reflecting their trade-off between efficiency and the ability to identify complex features in the data. Meanwhile, MMSDL outperforms the other models by producing the most distinct and well-separated clusters across all categories, especially in the challenging task of differentiating stable MCI from progressive MCI, suggesting that it is more proficient at capturing the fine-grained distinctions required in medical imaging diagnostics.

### 4.3. Effectiveness of attention modules

Table 3 presents a comparison of the performance of the MMSDL approach with its corresponding counterparts. Namely, the model with only embedding (E-MMSDL), the model with only PET (PET-MMSDL), the model with only MRI (MRI-MMSDL), the model with only multi-head cross-attention in MSF (C-MMSDL), the model with only multi-head self-attention in MSF (S-MMSDL), the model with only CMF, and the model with only MSF, to assess the efficiency of the MSF module and determine its impact on the overall effectiveness of the MMSDL method. On the other hand, comparing these models enables us to identify the impact of each attention module and determine the optimal configuration for a given task.

Our findings revealed that the proposed attention modules could significantly enhance classification performance. Specifically, our MMSDL approach, which incorporates both cross-modality and cross-scale attention mechanisms, has demonstrated its superiority over other models in our study, emphasizing the importance of fusing complementary information from different modalities and scales to achieve better performance.

Generally, the proposed method showed higher levels of accuracy, specificity, precision, recall, and F1 scores across all three tasks. Specifically, in the NC vs. AD task, our approach demonstrated an accuracy rate of 95.25 %, a sensitivity of 97.35 %, a specificity of 92.71 %, an F1-score of 95.74 %, and an AUC of 94.68 %, which were significantly higher than those achieved by other configurations of our method. For the NC vs. MCI task, our method delivered high accuracy of 85.22 %, a sensitivity of 69.36 %, a specificity of 94.00 %, an F1-score of 76.21 %, and an AUC of 81.36 %, which were significantly better than the results of other versions. Lastly, in the sMCI vs. pMCI task, our method demonstrated significant performance with an accuracy of 75.16 %, a sensitivity of 50.20 %, a specificity of 93.81 %, an F1-score of 61.90 %, and an AUC of 71.87 %, which were significantly higher than those of other configurations of our method.

These classification metrics highlight the high effectiveness of our proposed methodology in accurately distinguishing the disease. The superior performance is attributed to the integration of the MSF and CMF fusion modules. The MSF attention module is designed to capture interactions between different modalities and scales, ensuring the model effectively learns from both. Meanwhile, the CMF attention module focuses on extracting and fusing features from various scales. By incorporating these attention modules into the MMSDL model, we can leverage complementary information from different modalities and scales, significantly enhancing the model's overall performance.

We also performed t-SNE feature visualizations for different configurations of the MMSDL method across three classification tasks: NC vs. AD, NC vs. MCI, and sMCI vs. pMCI, as shown in Fig. 5. The configurations include E-MMSDL, PET-MMSDL, MRI-MMSDL, C-MMSDL, S-MMSDL, CMF, MSF, and the proposed MMSDL. The variation in performance arises from the specific design and characteristics of each configuration. For example, the PET-MMSDL model, which uses only PET data, shows more scattered points and overlaps, indicating that PET alone may not capture all necessary features for clear class separations. In contrast, the MRI-MMSDL and proposed MMSDL, which integrate both MRI and PET data, demonstrate clearer cluster separations, particularly in NC vs. AD. The superior performance of the proposed

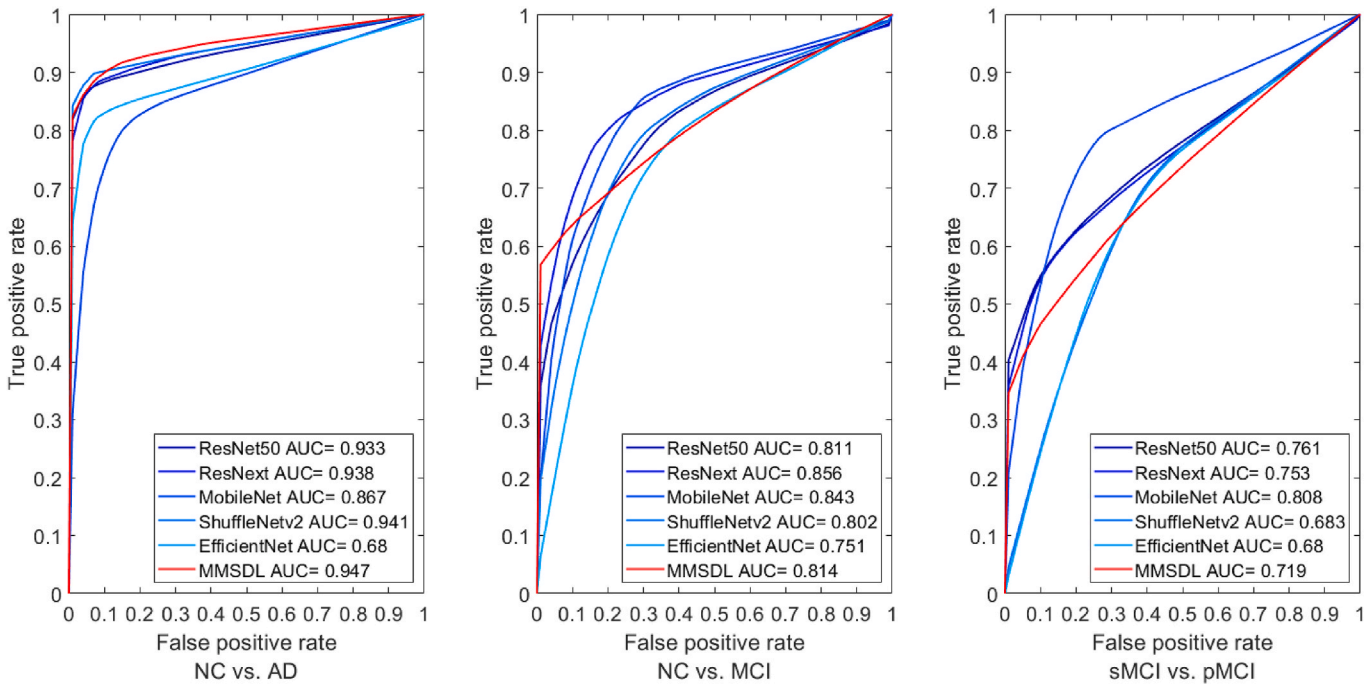


**Fig. 3.** ROC curves comparison between he proposed method and different conventional method for three the different classification tasks.

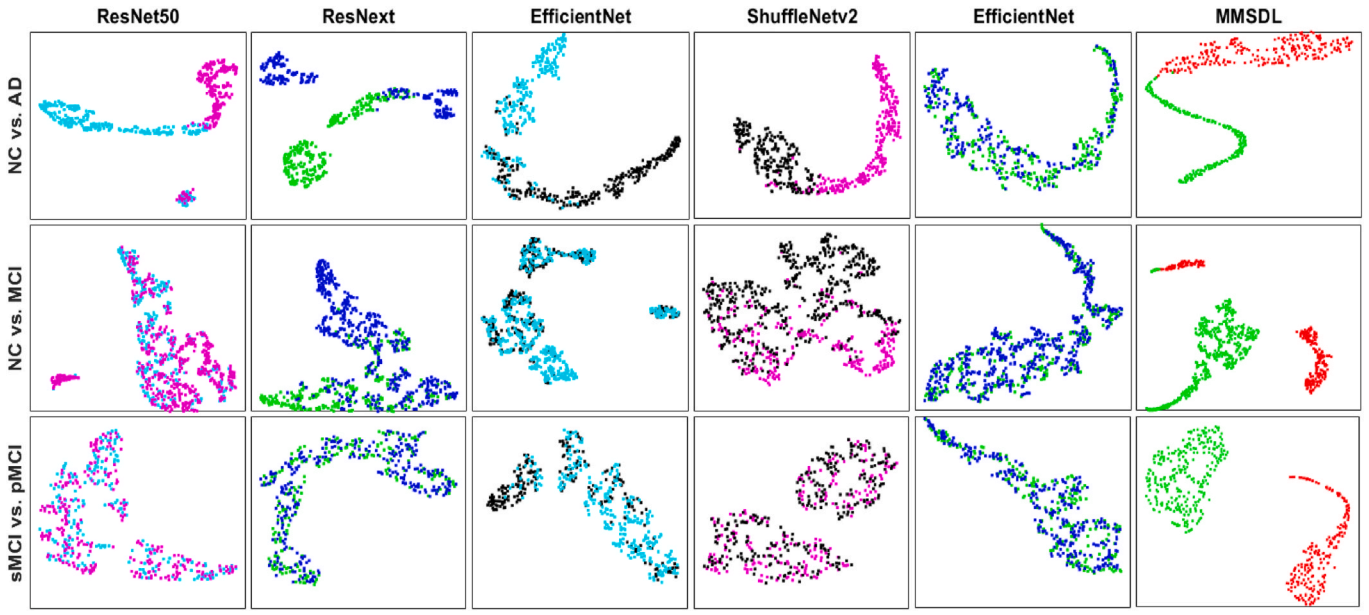


**Fig. 4.** The t-SNE visualization comparison of features between the proposed method and different conventional methods for the three different classification tasks.

MMSDL can be attributed to its combination of multi-head self-attention, which captures global interdependencies across features, and multi-head cross-attention, which weighs the contributions of different modalities, such as MRI and PET. Additionally, the CMF module in the proposed MMSDL effectively fuses MRI and PET data at various scales, allowing the model to leverage features from both modalities to improve classification accuracy. For NC vs. MCI, E-MMSDL and MRI-MMSDL show better separation of clusters compared to PET-MMSDL, as MRI data provides more detailed structural information that aids in class distinction. In the challenging task of sMCI vs. pMCI, the proposed MMSDL stands out with the most distinct cluster separations, while other configurations like MRI-MMSDL and MSF show some overlap, further highlighting the importance of integrating multi-modal data and attention mechanisms for enhanced feature extraction and classification. Overall, the proposed MMSDL consistently achieves superior results due to its advanced multi-head attention mechanisms and effective cross-modality fusion.

Additionally, we evaluated the ROC curves, presented in Fig. 6, to assess the performance of different configurations of our method. The ROC curves provide a visual representation of the true positive vs. the false positive rates for each configuration across the three tasks: NC vs. AD, NC vs. MCI, and sMCI vs. pMCI. These configurations include: E-MMSDL, PET-MMSDL, MRI-MMSDL, C-MMSDL, S-MMSDL, CMF, MSF, and the proposed MMSDL model. As shown in Fig. 6, our proposed method achieves the highest AUC in all classification tasks, indicating superior performance in distinguishing between classes. For the NC vs.

**Table 3**
Classification comparison between different configurations of our method for three different tasks (mean ± SD %).

| Symbol | Quantity | ACC | SEN | SPE | F1 | AUC |
|---|---|---|---|---|---|---|
| NC vs. AD | E-MMSDL | 84.63 ± 2.42 | 97.12 ± 1.62 | 69.49 ± 4.72 | 87.39 ± 1.86 | 83.20 ± 2.53 |
| | PET-MMSDL | 88.42 ± 2.57 | 83.37 ± 6.49 | **94.54 ± 5.53** | 88.66 ± 2.85 | 88.79 ± 2.37 |
| | MRI-MMSDL | 92.55 ± 1.98 | 97.08 ± 1.75 | 87.06 ± 5.08 | 93.48 ± 1.62 | 91.77 ± 2.19 |
| | C-MMSDL | 88.65 ± 2.35 | 96.82 ± 2.89 | 78.76 ± 5.24 | 90.35 ± 1.93 | 87.56 ± 2.47 |
| | S-MMSDL | 89.23 ± 3.00 | 92.08 ± 8.73 | 85.78 ± 6.74 | 90.20 ± 3.43 | 88.73 ± 2.58 |
| | CMF | 90.73 ± 2.73 | 96.06 ± 3.20 | 84.27 ± 8.42 | 91.96 ± 2.04 | 89.91 ± 3.14 |
| | MSF | 94.11 ± 1.21 | 95.76 ± 2.57 | 92.11 ± 4.08 | 94.69 ± 1.02 | 93.61 ± 1.35 |
| | MMSDL | **95.25 ± 0.72** | **97.35 ± 2.00** | 92.71 ± 3.14 | **95.74 ± 0.61** | **94.68 ± 0.86** |
| NC vs. MCI | E-MMSDL | 76.93 ± 4.47 | 44.05 ± 14.85 | 95.14 ± 8.91 | 56.44 ± 12.03 | 69.29 ± 5.50 |
| | PET-MMSDL | 78.69 ± 4.33 | 57.01 ± 12.08 | 90.69 ± 5.69 | 65.14 ± 8.27 | 73.58 ± 5.53 |
| | MRI-MMSDL | 83.36 ± 4.33 | 60.87 ± 12.84 | 95.81 ± 2.20 | 71.61 ± 9.32 | 78.05 ± 6.18 |
| | C-MMSDL | 76.90 ± 4.11 | 44.89 ± 12.27 | 94.61 ± 2.33 | 57.21 ± 10.98 | 69.42 ± 5.84 |
| | S-MMSDL | 78.57 ± 2.91 | 47.08 ± 9.23 | 95.99 ± 1.06 | 60.54 ± 8.10 | 71.18 ± 4.33 |
| | CMF | 77.57 ± 4.74 | 43.71 ± 13.74 | **96.31 ± 1.44** | 56.96 ± 13.96 | 69.74 ± 6.70 |
| | MSF | 83.62 ± 4.61 | 63.82 ± 12.32 | 94.57 ± 2.56 | 72.95 ± 9.39 | 78.87 ± 6.25 |
| | MMSDL | **85.22 ± 3.91** | **69.36 ± 14.93** | 94.00 ± 3.70 | **76.21 ± 8.71** | **81.36 ± 6.26** |
| sMCI vs. pMCI | E-MMSDL | 68.37 ± 3.51 | 45.37 ± 25.49 | 85.55 ± 23.21 | 83.21 ± 17.42 | 52.10 ± 11.59 |
| | PET-MMSDL | 68.03 ± 2.55 | 39.41 ± 10.99 | 89.41 ± 7.19 | 50.53 ± 8.01 | 64.32 ± 3.09 |
| | MRI-MMSDL | 72.31 ± 5.92 | 37.50 ± 13.86 | **98.31 ± 1.02** | 52.27 ± 15.66 | 67.90 ± 6.77 |
| | C-MMSDL | 66.75 ± 7.01 | 32.11 ± 24.90 | 92.64 ± 14.04 | 87.47 ± 15.52 | 40.05 ± 24.75 |
| | S-MMSDL | 70.00 ± 5.03 | 40.34 ± 21.37 | 92.16 ± 21.00 | **91.37 ± 14.83** | 51.44 ± 10.79 |
| | CMF | 71.49 ± 3.39 | 38.28 ± 12.47 | 96.30 ± 6.23 | 52.39 ± 10.16 | 67.29 ± 4.16 |
| | MSF | 72.12 ± 6.86 | 40.64 ± 19.61 | 95.64 ± 7.48 | 52.72 ± 20.49 | 68.10 ± 8.09 |
| | MMSDL | **75.16 ± 5.82** | **50.20 ± 17.16** | 93.81 ± 8.85 | 61.90 ± 13.04 | **71.87 ± 6.61** |

AD task, MMSDL reaches an AUC of 0.947, outperforming other configurations such as MRI-MMSDL and MSF, which also show strong results with AUCs of 0.918 and 0.936, respectively. In the NC vs. MCI task, MMSDL again performs best with an AUC of 0.814, while MSF and MRI-MMSDL follow with competitive AUCs of 0.789 and 0.78. For the most challenging task, sMCI vs. pMCI, MMSDL continues to demonstrate superior performance with an AUC of 0.719, while the other configurations show lower AUC values, indicating that MMSDL can handle the subtle differences between stable and progressive MCI more effectively.

### 4.4. Effectiveness of multi-scale models

Figs. 7–9 present a comparison of the performance of four different scales of MMSDL on three distinct brain image analysis tasks: NC vs. AD, NC vs. MCI, and sMCI vs. pMCI. Specifically, the six scales evaluated are denoted as MMSDL_16, MMSDL_32, MMSDL_64, MMSDL_64_32, MMSDL_64_16, and MMSDL_32_16, where the numbers indicate the input image size and the scales fused during the training process. Notably, MMSDL_16, MMSDL_32, MMSDL_64, MMSDL_64_32, MMSDL_64_16, and MMSDL_32_16 correspond to the input image sizes of 6x16x16, 32x32x32, 64x64x64, fusing 64x64x64 with 32x32x32, fusing 64x64x64 with 16x16x16, and fusing 32x32x32 with 16x16x16, respectively. The model with different scales was found to outperform the single-scale models regarding classification accuracy. Moreover, the results demonstrate that the fusion of different scales can further improve classification accuracy.

For instance, the study found that in the NC vs. AD task, our proposed method achieved an accuracy of 95.25 %, which was higher than the accuracies achieved by single or double scale fusion techniques models. On the other hand, the MMSDL_16, MMSDL_32, MMSDL_64, MMSDL_32_16, MMSDL_64_16, and MMSDL_64_32 achieved accuracies of 83.92 %, 87.86 %, 90.66 %, 91.8 %, 91.93 %, and 93.82, respectively. Similarly, in the NC vs. MCI task, our proposed method achieved an accuracy of 85.22 %, compared to 70.75 %, 72.82 %, 80.7 %, 75.49 %, 82.66 %, and 83.53 % for the MMSDL_16, MMSDL_32, MMSDL_64, MMSDL_32_16, MMSDL_64-16, and MMSDL_64-32, respectively. Likewise, in sMCI vs. pMCI, our proposed method achieved an accuracy of 75.16 %, outperforming the MMSDL_16, MMSDL_32, MMSDL_64, MMSDL_32_16, MMSDL_64-16, and MMSDL_64-32 models, which achieved accuracies of 58.16 %, 58.11 %, 71.11 %, 63.63 %, 73.12 %, and 72.32 %, respectively. These results verify the effectiveness of multi-scale fusion and its ability to extract features from multiple scales and combine them to enhance classification performance.

Our findings highlight the importance of incorporating multi-scale information in AD tasks, as the disease involves subtle changes at various levels. Specifically, MMSDL_64 demonstrates higher accuracy compared to MMSDL_32 and MMSDL_16 due to its higher resolution, which facilitates more precise detection and analysis of variations present in AD. However, integrating any two scales together yields higher accuracy than using a single scale. This suggests that combining multi-scale information can effectively capture a broader range of disease features, leading to more robust and accurate AD diagnosis. Therefore, our results emphasize the essential role of multi-scale approaches in enhancing the performance of AD detection models.

Fig. 10 shows the t-SNE feature visualization across various fusion scales of our method for the classification tasks of NC vs. AD, NC vs. MCI, and sMCI vs. pMCI, respectively. In the NC vs. AD task, MMSDL_64 and the proposed MMSDL demonstrate clearer separations, while configurations like MMSDL_16 and MMSDL_32_16 show scattered clusters and overlaps, indicating that smaller input sizes may not capture all relevant features for accurate class distinction. For NC vs. MCI, MMSDL_64_32 and the proposed MMSDL show better separation compared to MMSDL_16 and MMSDL_64_16, suggesting that the fusion of larger input sizes improves feature representation. In the sMCI vs. pMCI task, MMSDL_64 and MMSDL_64_32 show the most distinct cluster separations, while smaller input sizes like MMSDL_16 and MMSDL_32 display more scattered and overlapping clusters. Overall, the fusion of larger input sizes, as seen in MMSDL_64_32 and the proposed MMSDL, which incorporates all scales, consistently produces superior results across all classification tasks. This highlights the importance of capturing detailed multi-scale information for effective feature extraction and classification, with the extensive fusing of multi-modality inputs achieving better separation and enhancing the model's ability to distinguish between disease stages.

Furthermore, Fig. 11 presents the ROC curves for different fusion scales of our method, providing a comprehensive comparison across the three classification tasks. The fusion scales compared include MMSDL_16, MMSDL_32, MMSDL_64, MMSDL_32_16, MMSDL_64_16, MMSDL_64_32, and the proposed MMSDL. As shown in Fig. 11, the proposed MMSDL configuration consistently achieves the highest AUC across all tasks, indicating its superior ability to fuse multiple scales and extract relevant features for accurate classification. For the NC vs. AD task, the proposed MMSDL achieves an AUC of 0.947, significantly outperforming lower-scale configurations such as MMSDL_16 (AUC = 0.827) and MMSDL_32 (AUC = 0.876). In the NC vs. MCI task, the

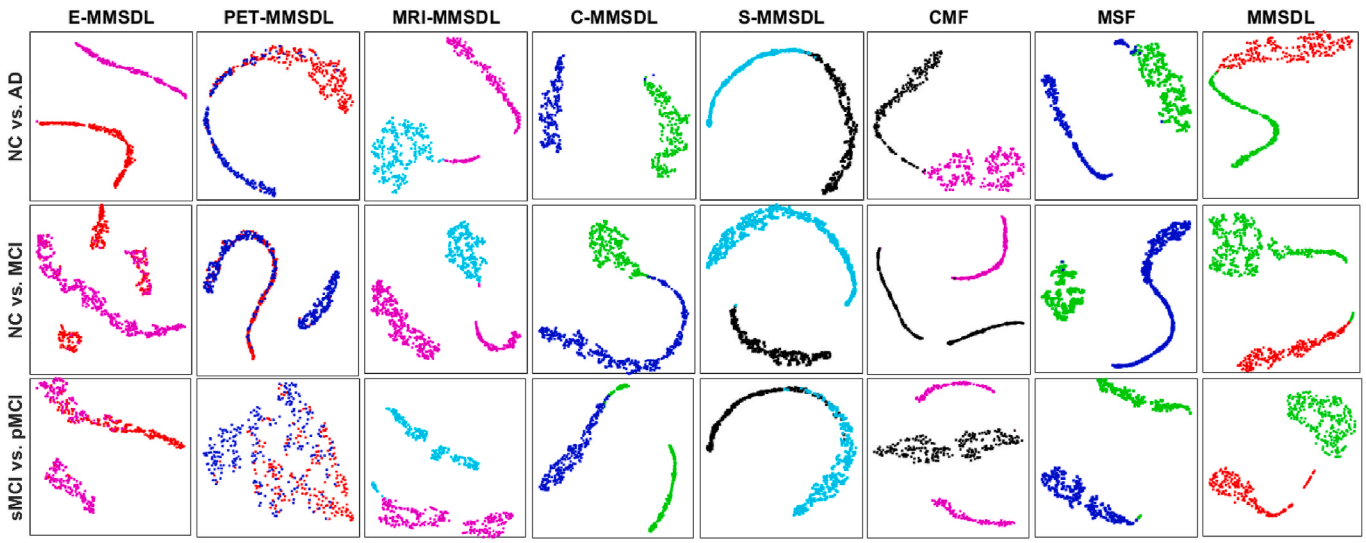


**Fig. 5.** The t-SNE visualization comparison of features between different configurations of our method for the three different classification tasks.

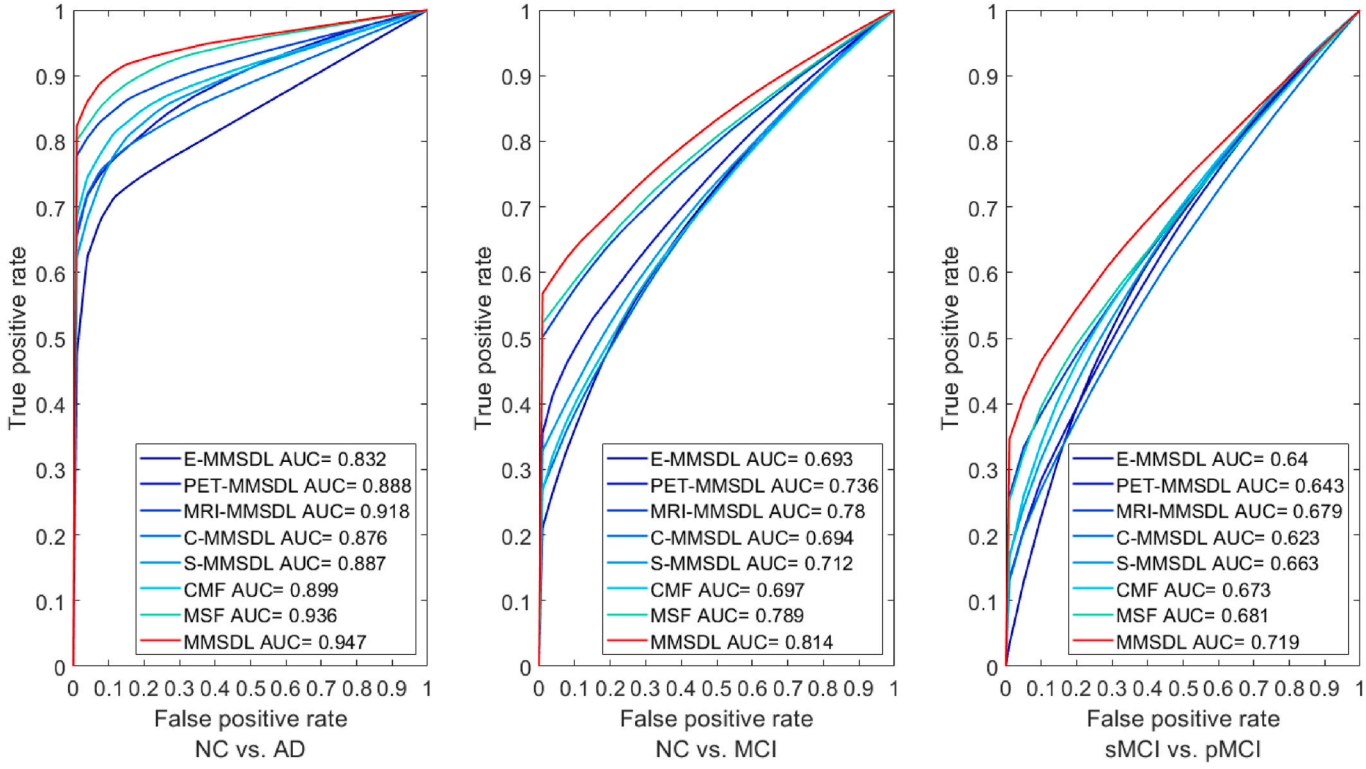


**Fig. 6.** ROC curves comparison between different configurations of our method for three the different classification tasks.

proposed MMSDL again demonstrates its effectiveness with an AUC of 0.814, outperforming other configurations such as MMSDL_32_16 (AUC = 0.686) and MMSDL_64 (AUC = 0.75). The sMCI vs. pMCI task, which presents the most significant challenge, shows the proposed MMSDL achieving the highest AUC of 0.719, while other configurations, such as MMSDL_64 (AUC = 0.671) and MMSDL_64_32 (AUC = 0.698), perform slightly lower. These results demonstrate that the fusion of larger scales, as seen in the proposed MMSDL, enhances the model's ability to capture subtle features and improve classification performance across all tasks. The ROC curves highlight the advantage of combining multiple scales, leading to better feature representation and superior classification results.

### 4.5. Discriminative ROIs

Figs. 12–14 illustrate the top 10 ROIs identified for MRI and PET data across the three distinct classification tasks. The top 10 ROIs were determined by registering the extracted network weights to the automated anatomical labeling (116 ROIs) and ordering them in descending order based on the average values for each ROI [48]. Additionally, Tables 4 and 5 present the top 10 names of significant MRI and PET ROIs, along with their weighted averages, as detected by the proposed method for the three tasks. The ROIs identified in this study, including the hippocampus, cingulum, and occipital regions, have been previously associated with AD pathology. Notably, these identified ROIs align with

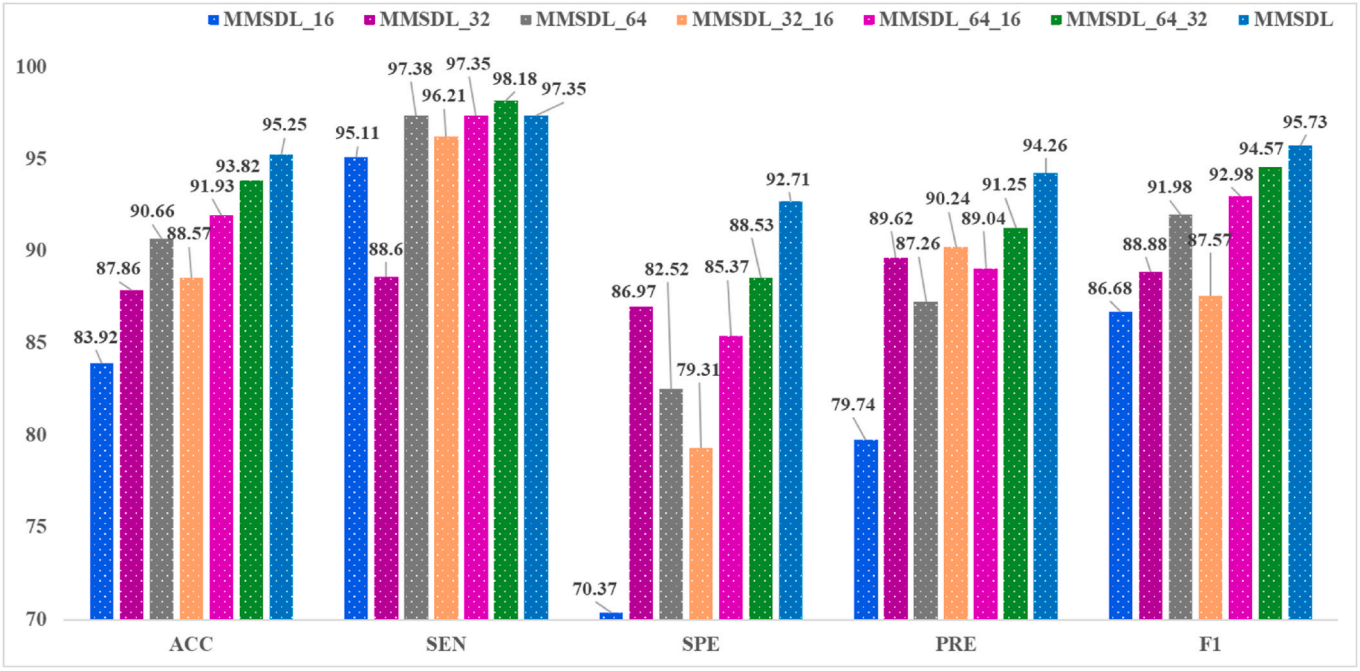


**Fig. 7.** Classification comparison between different fusion scales of the proposed method for NC vs. AD.

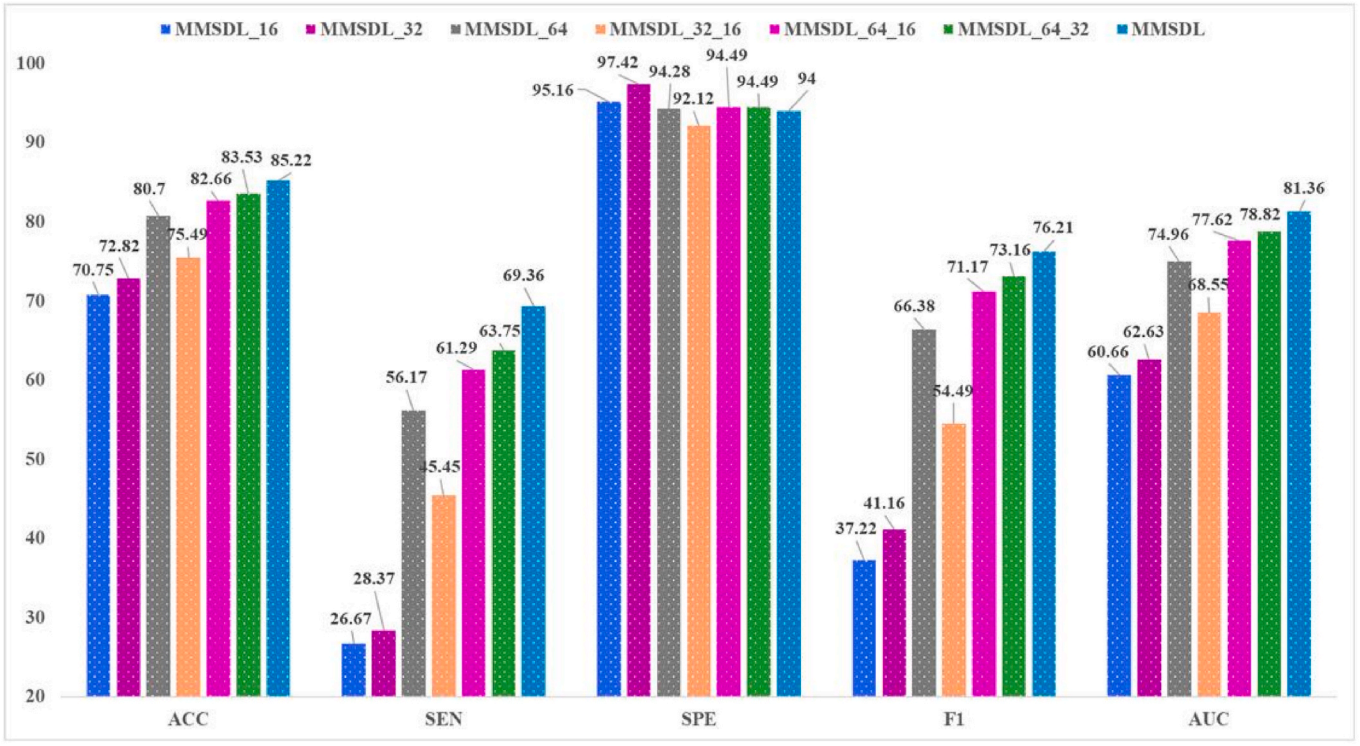


**Fig. 8.** Classification comparison between different fusion scales of the proposed method for NC vs. MCI.

findings from previous studies on AD diagnosis [49–52]. These regions have significant roles in various cognitive and emotional functions of the brain. For example, the hippocampus is one of the earliest and most consistently affected regions in AD, having a vital role in learning and memory. The cingulum, particularly the posterior cingulate cortex, is another region frequently impacted by AD. This region is involved in several cognitive processes, such as attention, emotion regulation, and episodic memory. The occipital lobe, located at the caudal end of the cerebral hemisphere, serves as the primary center for visual processing. Overall, the hippocampus, cingulum, and occipital lobe are essential regions of the brain that serve crucial roles in various cognitive and emotional functions, and their association with AD pathology highlights their importance in AD research.

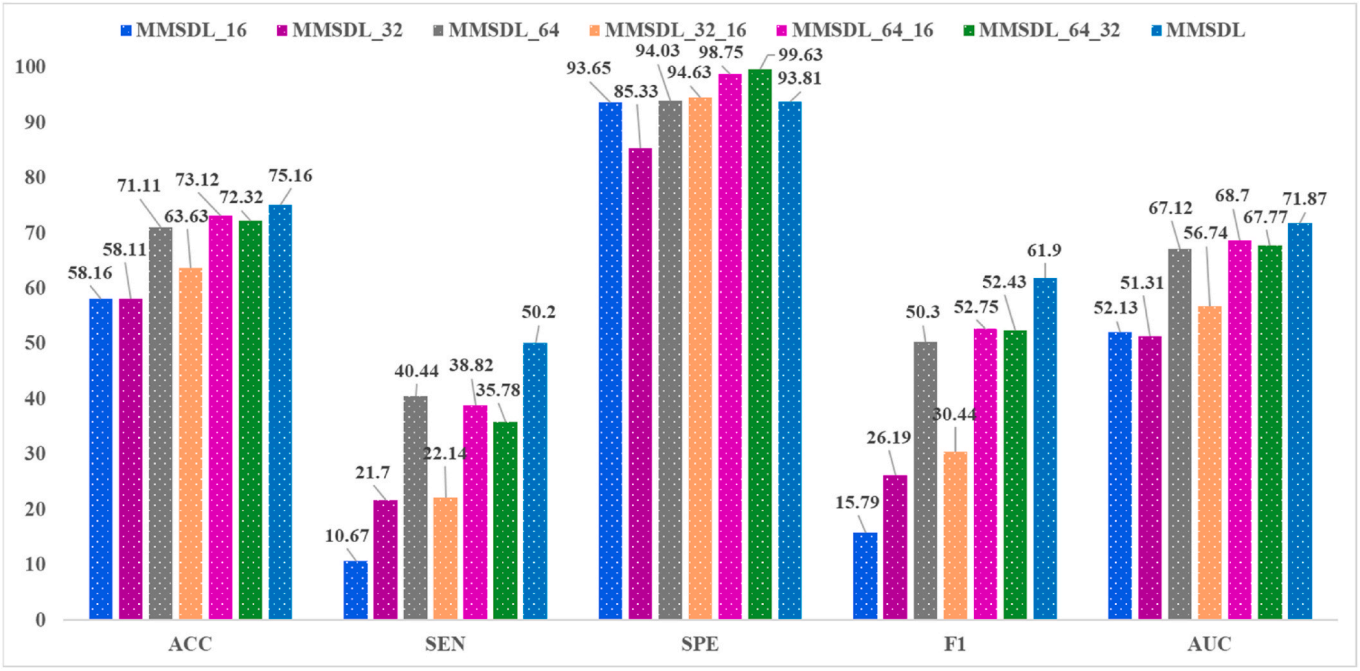


**Fig. 9.** Classification comparison between different fusion scales of the proposed method for sMCI vs. pMCI.

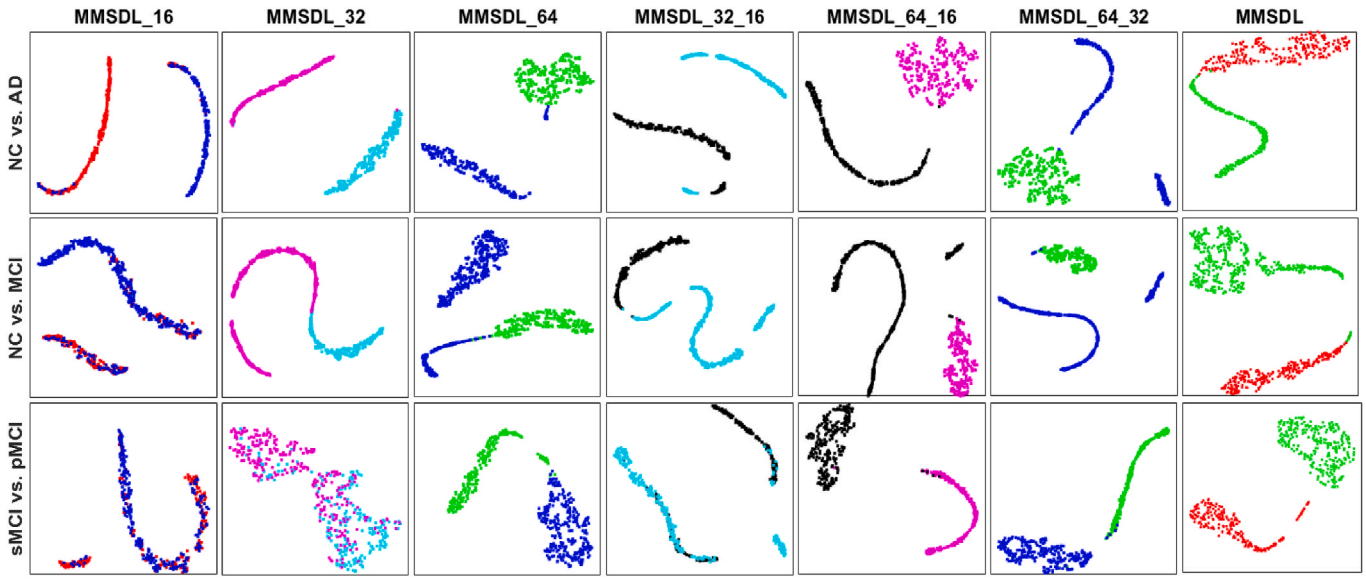


**Fig. 10.** The t-SNE visualization comparison of features between different fusion scales of our method for the three different classification tasks.

### 4.6. *Comparison with the previous studies*

Table 6 compares the classification performance of our approach with previous studies across key metrics, highlighting the effectiveness of our method in various AD classification tasks. In the NC vs. AD classification, the proposed method achieves an ACC of 95.25 %, second only to the top-performing method reported by Zhang et al. [54], which achieved an ACC of 96.68 %. This high ACC indicates the capability of our approach to correctly classify a substantial majority of cases, demonstrating its robustness in distinguishing NC from AD. In terms of SEN, our model attains a rate of 97.35 %, again ranking second among the studies listed. Moreover, our model's SPE of 92.71 % indicates a low false-positive rate, ensuring that individuals without AD are less likely to be misclassified. This balance between SEN and SPE sets our method apart from others, as it not only detects AD accurately but also minimizes false positives, providing more reliable diagnoses.

In the NC vs. MCI task, a more complex classification problem due to the subtle differences between normal aging and mild cognitive impairment, our method achieves an ACC of 85.22 %, the second-highest among the listed studies, and the highest SPE at 94.00 %. This high specificity is especially important for avoiding over diagnosis, as false positives can lead to unnecessary treatments or interventions. The proposed method's ability to maintain high specificity, while still achieving competitive sensitivity, makes it highly effective for clinical settings where precision is essential. Unlike other methods that may trade off SPE to improve SEN, our approach achieves a balance, thereby providing a more reliable diagnostic tool for early intervention in MCI cases.

For the sMCI vs. pMCI task, the most challenging due to the difficulty in distinguishing between stable and progressive MCI, our method

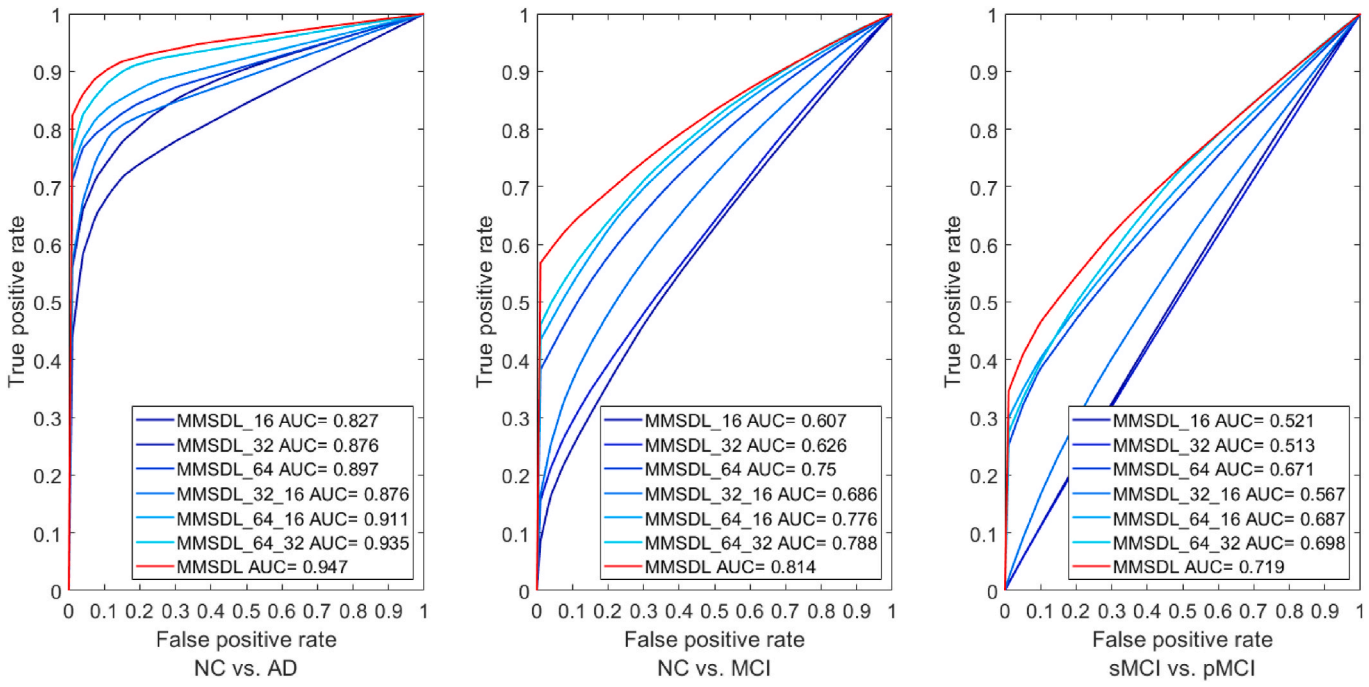


**Fig. 11.** ROC curves comparison between different fusion scales of our method for three the different classification tasks.

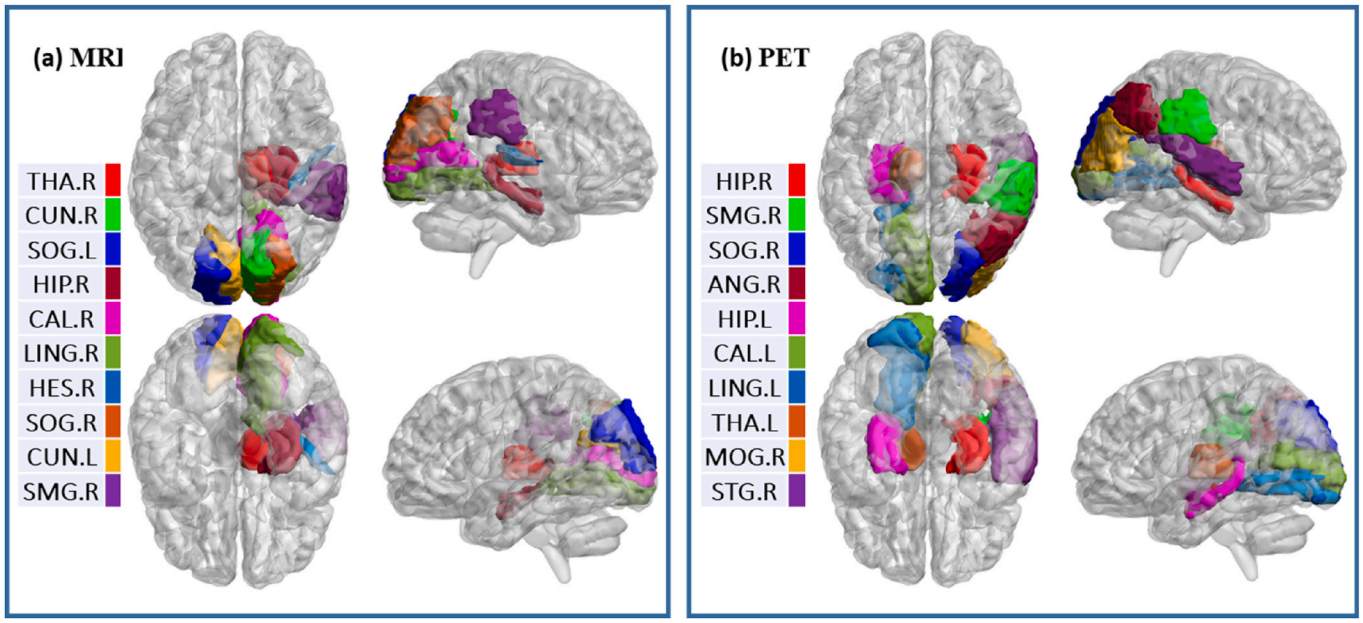


**Fig. 12.** Top 10 brain regions identified in MRI and PET data for NC vs. AD.

achieves an ACC of 75.16 %, the second-highest among the listed studies, and the highest SPE at 93.81 %. Although this task is particularly difficult, the method's performance shows a strong balance between detecting progressive MCI and reducing false positives. This balance is important because patients classified as progressive MCI are at higher risk of developing AD, and early, accurate detection allows for timely intervention. Many previous studies have struggled with this task, often prioritizing sensitivity over specificity or vice versa. The proposed method, however, achieves a balance that maximizes both, demonstrating its effectiveness and robustness even in the most difficult classification tasks.

Overall, the results shown in Table 6 emphasize the superiority of our method. It consistently performs at a high level across all tasks, balancing key metrics like SEN, SPE, and ACC. By leveraging cross-modality attention and multiscale feature extraction, our method captures both local and global features, allowing it to handle complex classification tasks with greater precision. This combination of sensitivity and specificity makes our approach particularly effective for clinical applications, where reliable and balanced performance is essential for early and accurate diagnosis of AD and its early stages. The consistent performance across all tasks, especially in comparison to existing methods, highlights the practical value of our approach and its potential for clinical medical use.

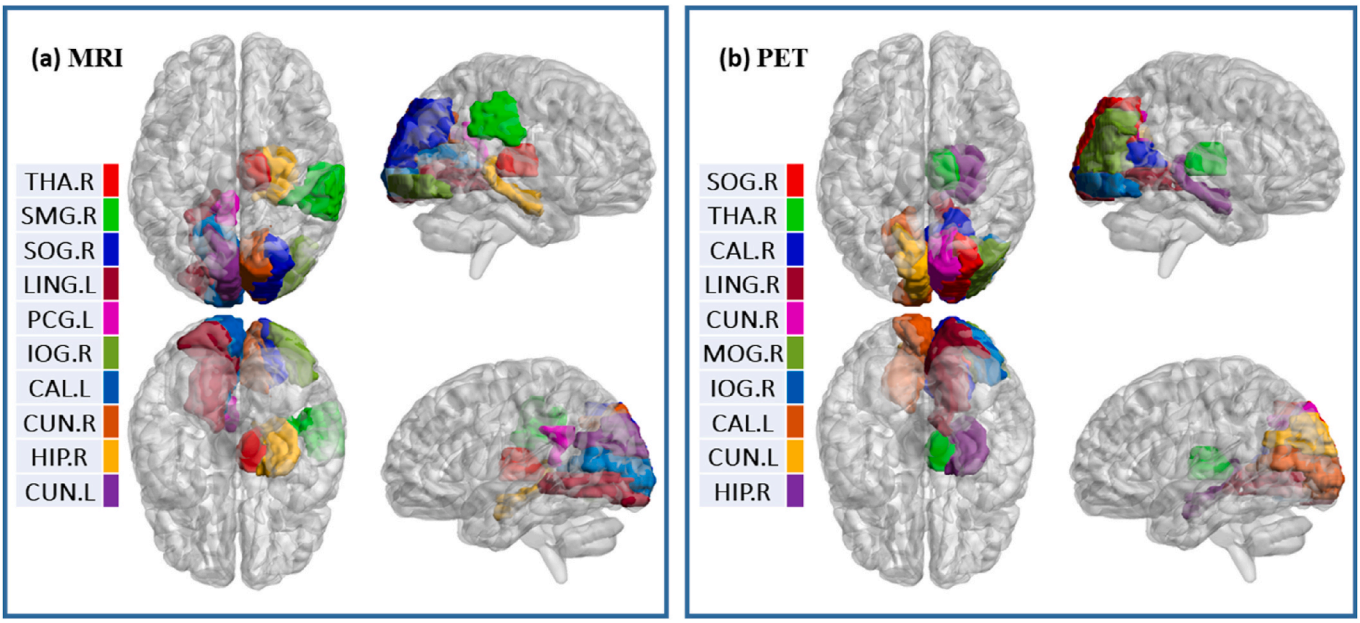


**Fig. 13.** Top 10 brain regions identified in MRI and PET data for NC vs. MCI.

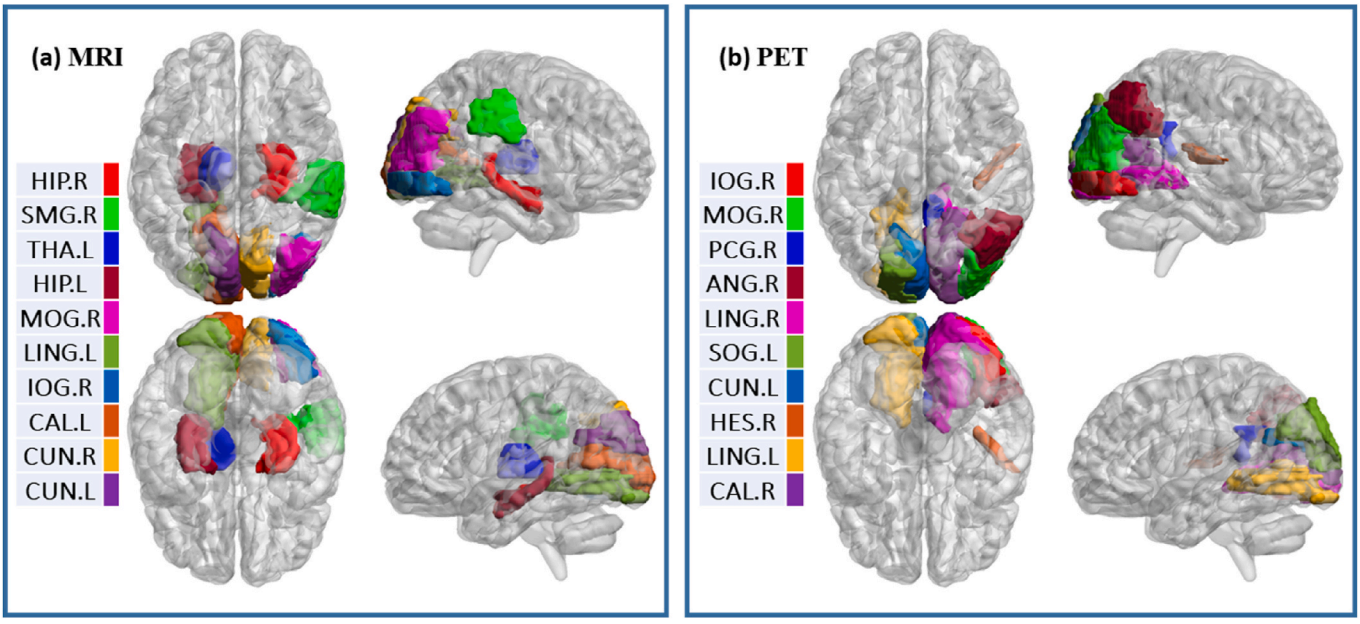


**Fig. 14.** Top 10 brain regions identified in MRI and PET data for sMCI vs. pMCI.

**Table 4**
Top 10 significant MRI ROIs and their weighted averages detected by the proposed method for the three tasks.

| NC vs. AD | | NC vs. MCI | | sMCI vs. pMCI | |
|---|---|---|---|---|---|
| ROIs | Weighted averages | ROIs | Weighted averages | ROIs | Weighted averages |
| Thalamus_R | 0.9057 | Thalamus_R | 0.9069 | Hippocampus_R | 0.9159 |
| Cuneus_R | 0.9028 | SupraMarginal_R | 0.9058 | SupraMarginal_R | 0.9156 |
| Occipital_Sup_L | 0.9024 | Occipital_Sup_R | 0.8989 | Thalamus_L | 0.9145 |
| Hippocampus_R | 0.9014 | Lingual_L | 0.8988 | Hippocampus_L | 0.9090 |
| Calcarine_R | 0.8994 | Cingulum_Post_L | 0.8980 | Occipital_Mid_R | 0.9055 |
| Lingual_R | 0.8988 | Occipital_Inf_R | 0.8973 | Lingual_L | 0.9040 |
| Heschl_R | 0.8975 | Calcarine_L | 0.8969 | Occipital_Inf_R | 0.9040 |
| Occipital_Sup_R | 0.8964 | Cuneus_R | 0.8964 | Calcarine_L | 0.9023 |
| Cuneus_L | 0.8943 | Hippocampus_R | 0.8961 | Cuneus_R | 0.9006 |
| SupraMarginal_R | 0.8935 | Cuneus_L | 0.8936 | Cuneus_L | 0.8994 |

**Table 5**
Top 10 significant PET ROIs and their weighted averages detected by the proposed method for the three tasks.

| NC vs. AD | | NC vs. MCI | | sMCI vs. pMCI | |
|---|---|---|---|---|---|
| ROIs | Weighted averages | ROIs | Weighted averages | ROIs | Weighted averages |
| Hippocampus_R | 0.8563 | Occipital_Sup_R | 0.8273 | Occipital_Inf_R | 0.8050 |
| SupraMarginal_R | 0.8530 | Thalamus_R | 0.8214 | Occipital_Mid_R | 0.7868 |
| Occipital_Sup_R | 0.8377 | Calcarine_R | 0.8171 | Cingulum_Post_R | 0.7845 |
| Angular_R | 0.8357 | Lingual_R | 0.8165 | Angular_R | 0.7742 |
| Hippocampus_L | 0.8348 | Cuneus_R | 0.8109 | Lingual_R | 0.7713 |
| Calcarine_L | 0.8316 | Occipital_Mid_R | 0.8103 | Occipital_Sup_L | 0.7683 |
| Lingual_L | 0.8300 | Occipital_Inf_R | 0.8098 | Cuneus_L | 0.7674 |
| Thalamus_L | 0.8282 | Calcarine_L | 0.8030 | Heschl_R | 0.7673 |
| Occipital_Mid_R | 0.8265 | Cuneus_L | 0.8029 | Lingual_L | 0.7672 |
| Temporal_Sup_R | 0.8243 | Hippocampus_R | 0.8018 | Calcarine_R | 0.7641 |

**Table 6**
Classification comparison between ours and previous studies (%).

| Algorithm | Subject | Modality | NC vs. AD | | | NC vs. MCI | | | sMCI vs. pMCI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ACC | SEN | SPE | ACC | SEN | SPE | ACC | SEN | SPE |
| Zhang et al. [13] | 129AD+110NC+125MCI | MRI + PET + CSF | 91.07 | 94.44 | 91.11 | 71.26 | 71.47 | 71.67 | – | – | – |
| Kong et al. [20] | 111AD+130NC+129MCI | MRI + PET | 93.21 | 91.43 | **95.42** | **86.52** | **94.34** | 81.64 | – | – | – |
| Huang et al. [22] | 465AD+480NC+567 MC | MRI + PET | 90.10 | 90.85 | 89.21 | – | – | – | 72.22 | 73.44 | 71.25 |
| Song et al. [53] | 95AD+126NC+160MCI | MRI + PET | 94.11 | 93.33 | 94.27 | 85.00 | 84.69 | 85.60 | – | – | – |
| Zhang et al. [54] | 215AD+246NC+211sMC + 120pMCI | MRI + PET | **96.68** | **99.19** | 94.49 | – | – | – | **78.00** | 54.96 | 89.37 |
| Lin et al. [55] | 362AD+308NC+233sMC + 183pMCI | MRI + PET | 92.28 | 90.38 | 94.37 | – | – | – | 74.10 | 75.00 | 73.08 |
| Zhang et al. [56] | 300 NC,120 pMCI, 343 sMCI, 149 AD | MRI + PET | 90.60 | – | – | 75.5 | – | – | – | – | – |
| Gao et al. [57] | 352AD+427NC+342sMC + 234pMCI | MRI + PET | 92.0 | 89.1 | 94.0 | – | – | – | 75.3 | **77.3** | 74.1 |
| Zhu et al. [58] | 93AD+202MCI+101NC | MRI + PET | 91.7 | 91.8 | 91.6 | 74.5 | 42.5 | 90.3 | 72.6 | 84.8 | 58.5 |
| MMSDL | 218AD+264NC+204 pMCI +273 sMCI | MRI + PET | 95.25 | 97.35 | 92.71 | 85.22 | 69.36 | **94.00** | 75.16 | 50.20 | **93.81** |

## 5. Conclusion and future work

In this paper, we developed an end-to-end learning network, named MMSDL, which offered several advantages over state-of-the-art methods. Our method included three different modules: modality embedding, MSF, and CMF. Initially, the modality embedding module transformed neuroimaging data into high-dimensional vector features. The MSF module utilized cascading multi-head self-attention and cross-attention to capture global relations among embedded features, weigh each modality's contribution to another modality, and extract complex relationships across different scales of MRI and PET data. Additionally, we applied an effective CMF to fuse MRI and PET data at each scale, enhancing global features from the previous attention layers. By integrating features from every scale, our model achieved superior performance compared to state-of-the-art methods when evaluated on the ADNI dataset.

Although our approach demonstrated superior performance compared to many recent studies, few limitations need to be addressed. One limitation is that our method utilizes only neuroimaging data, excluding genetic and clinical data. Integrating our proposed method with clinical and/or genetic data could enhance the accuracy of disease diagnosis and prediction. Another limitation is that we do not consider the interdependence or correlation among the ROIs during the training phase. Incorporating the interconnectivity and relationships among the ROIs, particularly those highly associated with AD. In our future analyses, we will focus on providing a deeper understanding of the neural mechanisms underlying AD, ultimately enhancing the accuracy and reliability of AD diagnosis.

## CRediT authorship contribution statement

**Mohammed Abdelaziz:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Tianfu Wang:** Resources, Supervision, Project administration, Funding acquisition. **Waqas Anwaar:** Software. **Ahmed Elazab:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

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

## Acknowledgement

This work was supported in part by the Shenzhen Key Basic Research Project under Grant KCXFZ20201221173213036.

## References

[1] T. Goel, et al., Multimodal Neuroimaging Based Alzheimer's Disease Diagnosis Using Evolutionary RVFL Classifier, 2023.
[2] J. Gaugler, et al., 2022 Alzheimer's Disease Facts and Figures, vol. 18, 2022, pp. 700–789, 4.
[3] Rammohan V. Rao, Kaavya G. Subramaniam, Julie Gregory, Aida L. Bredesen, Christine Coward, Sho Okada, Lance Kelly, Dale E. Bredesen, Rationale for a multifactorial approach for the reversal of cognitive decline in Alzheimer's disease and MCI: a review, Int. J. Mol. Sci. 24 (2) (2023) 1659.
[4] Azadeh EghbalManesh, Asghar Dalvandi, Mohammad Zoladl, The experience of stigma in family caregivers of people with schizophrenia spectrum disorders: a meta-synthesis study, Heliyon 9 (3) (2023) e14333.
[5] M. Bucholc, S. Titarenko, X. Ding, C. Canavan, T. Chen, A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia, Expert Syst. Appl. 217 (2023) 119541.
[6] Zhuopin Sun, Sharon L. Naismith, Steven Meikle, Fernando Calamante, Alzheimer's Disease Neuroimaging Initiative, A novel method for PET connectomics guided by fibre-tracking MRI: application to Alzheimer's disease, Hum. Brain Mapp. 45 (4) (2024) e26659.
[7] S. Matsushita, et al., The association of metabolic brain MRI, amyloid PET, and clinical factors: a study of Alzheimer's disease and normal controls from the open access series of imaging studies dataset, J. Magn. Reson. Imag. 59 (4) (2024) 1341–1348.
[8] T. Goel, R. Sharma, M. Tanveer, P. Suganthan, K. Maji, R. Pilli, Multimodal neuroimaging based Alzheimer's disease diagnosis using evolutionary RVFL classifier, IEEE Journal of Biomedical and Health Informatics (2023).

[9] Yuanpeng Zhang, Shuihua Wang, Kaijian Xia, Yizhang Jiang, Pengjiang Qian, Alzheimer's Disease Neuroimaging Initiative, Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion, Inf. Fusion 66 (2021) 170–183.
[10] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, Z. Wang, D. Wang, Multimodal fusion diagnosis of Alzheimer's disease based on FDG-PET generation, Biomed. Signal Process Control 89 (2024) 105709.
[11] Manhua Liu, Danni Cheng, Kundong Wang, Yaping Wang, Alzheimer's Disease Neuroimaging Initiative, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinformatics 16 (2018) 295–308.
[12] P. Lu, L. Hu, A. Mitelpunkt, S. Bhatnagar, L. Lu, H. Liang, A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease, Biomed. Signal Process Control 88 (2024) 105669.
[13] J. Zhang, X. He, Y. Liu, Q. Cai, H. Chen, L. Qing, Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data, Comput. Biol. Med. 162 (2023) 107050.
[14] Linfeng Liu, Siyu Liu, Lu Zhang, Xuan Vinh To, Fatima Nasrallah, Shekhar S. Chandra, Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data, Neuroimage 277 (2023) 120267.
[15] J. Cheng, H. Wang, S. Wei, J. Mei, F. Liu, G. Zhang, Alzheimer's disease prediction algorithm based on de-correlation constraint and multi-modal feature interaction, Comput. Biol. Med. 170 (2024) 108000.
[16] Y. Zhang, K. Sun, Y. Liu, D. Shen, Transformer-based multimodal fusion for early diagnosis of Alzheimer's disease using structural MRI and PET, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 2023, pp. 1–5.
[17] S.B. Hassen, M. Neji, Z. Hussain, A. Hussain, A.M. Alimi, M. Frikha, Deep learning methods for early detection of Alzheimer's disease using structural MR images: a survey, Neurocomputing 576 (2024) 127325.
[18] Z. Pei, Z. Wan, Y. Zhang, M. Wang, C. Leng, Y.-H. Yang, Multi-scale attention-based pseudo-3D convolution neural network for Alzheimer's disease diagnosis using structural MRI, Pattern Recogn. 131 (2022) 108825.
[19] C. Choudhury, T. Goel, M. Tanveer, A coupled-GAN architecture to fuse MRI and PET image features for multi-stage classification of Alzheimer's disease, Inf. Fusion (2024) 102415.
[20] Z. Kong, M. Zhang, W. Zhu, Y. Yi, T. Wang, B. Zhang, Multi-modal data Alzheimer's disease detection based on 3D convolution, Biomed. Signal Process Control 75 (2022) 103565.
[21] S. Wang, H. Wang, A.C. Cheung, Y. Shen, M. Gan, Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. Deep Learning Applications, 2020, pp. 53–73.
[22] Yechong Huang, Jiahang Xu, Yuncheng Zhou, Tong Tong, Xiahai Zhuang, Alzheimer's Disease Neuroimaging Initiative (ADNI), Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network, Front. Neurosci. 13 (2019) 509.
[23] R. Sharma, T. Goel, M. Tanveer, C. Lin, R. Murugan, Deep learning based diagnosis and prognosis of Alzheimer's disease: a comprehensive review, IEEE Transactions on Cognitive and Developmental Systems (2023).
[24] F. Behrad, M.S. Abadeh, An overview of deep learning methods for multimodal medical data mining, Expert Syst. Appl. 200 (2022) 117006.
[25] 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.
[26] Y. Tu, S. Lin, J. Qiao, Y. Zhuang, P. Zhang, Alzheimer's disease diagnosis via multimodal feature fusion, Comput. Biol. Med. 148 (2022) 105901.
[27] 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. Imag. 40 (6) (2021) 1632–1645.
[28] 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.
[29] S. Dwivedi, T. Goel, M. Tanveer, R. Murugan, R. Sharma, Multimodal fusion-based deep learning network for effective diagnosis of Alzheimer's disease, IEEE MultiMedia 29 (2) (2022) 45–55.
[30] M. Odusami, R. Maskeliūnas, R. Damaševičius, S. Misra, Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI Images, J. Med. Biol. Eng. 43 (3) (2023) 291–302.
[31] Y. Tang, X. Xiong, G. Tong, Y. Yang, H. Zhang, Multimodal diagnosis model of Alzheimer's disease based on improved Transformer, Biomed. Eng. Online 23 (1) (2024) 8.
[32] S. Wu, S. Qin, K. Wang, S. Yang, S. Zhang, Alzheimer's disease detection model based on multimodal data early fusion of medical neuroimaging, in: 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application, HPCC/DSS/SmartCity/DependSys), IEEE, 2023, pp. 801–808.
[33] C. Cui, et al., Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review, Prog. Biomed. Eng. 5 (2) (2023) 022001.
[34] W. Huang, K. Tan, Z. Zhang, J. Hu, S. Dong, A review of fusion methods for omics and imaging data, IEEE ACM Trans. Comput. Biol. Bioinf 20 (1) (2022) 74–93.
[35] M. Yu, et al., Hybrid multimodality fusion with cross-domain knowledge transfer to forecast progression trajectories in cognitive decline, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 265–275.
[36] Y. Dai, et al., DE-JANet: a unified network based on dual encoder and joint attention for Alzheimer's disease classification using multi-modal data, Comput. Biol. Med. 165 (2023) 107396.
[37] X. Sun, W. Guo, J. Shen, Toward attention-based learning to predict the risk of brain degeneration with multimodal medical data, Front. Neurosci. 16 (2023) 1043626.
[38] 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, Expert Syst. Appl. 231 (2023) 120761.
[39] Y. Leng, et al., Multimodal Cross Enhanced Fusion Network for Diagnosis of Alzheimer's Disease and Subjective Memory Complaints, 2023.
[40] T. Zhou, K.H. Thung, X. Zhu, D. Shen, Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis, Hum. Brain Mapp. 40 (3) (2019) 1001–1016.
[41] Rui Min, Guorong Wu, Jian Cheng, Qian Wang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, Multi-atlas based representations for Alzheimer's disease diagnosis, Hum. Brain Mapp. 35 (10) (2014) 5052–5070.
[42] J.L. Lancaster, et al., Automated Talairach atlas labels for functional brain mapping, Hum. Brain Mapp. 10 (3) (2000) 120–131.
[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[44] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[45] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[46] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV), 2018, pp. 116–131.
[47] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[48] L. Sun, et al., Mining brain region connectivity for Alzheimer's disease study via sparse inverse covariance estimation, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 1335–1344.
[49] X. Zhu, H.-I. Suk, S.-W. Lee, D. Shen, Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification 63 (3) (2015) 607–618.
[50] T.M. Nir, et al., Effectiveness of Regional DTI Measures in Distinguishing Alzheimer's Disease, MCI, and Normal Aging, vol. 3, 2013, pp. 180–195.
[51] J. Peng, L. An, X. Zhu, Y. Jin, D. Shen, Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 70–78.
[52] T. Zhou, M. Liu, K.-H. Thung, D. Shen, Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data 38 (10) (2019) 2411–2422.
[53] J. Song, J. Zheng, P. Li, X. Lu, G. Zhu, P. Shen, An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis, Frontiers in digital health 3 (2021) 637386.
[54] Y. Zhang, X. He, Y.H. Chan, Q. Teng, J.C. Rajapakse, Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans, Comput. Biol. Med. 164 (2023) 107328.
[55] W. Lin, et al., Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease, Front. Neurosci. 15 (2021) 646013.
[56] Z.-C. Zhang, X. Zhao, G. Dong, X.-M. Zhao, Improving Alzheimer's disease diagnosis with multi-modal PET embedding features by a 3D multi-task MLP-mixer neural network, IEEE Journal of Biomedical and Health Informatics 27 (8) (2023) 4040–4051.
[57] X. Gao, F. Shi, D. Shen, M. Liu, Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease, IEEE journal of biomedical and health informatics 26 (1) (2021) 36–43.
[58] X. Zhu, H.-I. Suk, D. Shen, Low-rank dimensionality reduction for multi-modality neurodegenerative disease identification, World Wide Web 22 (2019) 907–925.