Expert Systems With Applications 252 (2024) 124241

Contents lists available at ScienceDirect

Expert Systems With Applications

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

MSCLK: Multi-scale fully separable convolution neural network with large kernels for early diagnosis of Alzheimer's disease ☆

Run-Feng Tian [1], Jia-Ni Li [1], Shao-Wu Zhang *

MOE Key Laboratory of Information Fusion Technology, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Keywords:
Mild Cognitive Impairment
Multi-scale convolution
Separable Convolution
Metric Learning
Fusion

ABSTRACT

Alzheimer's disease (AD) is identified as a central nervous system disease that exhibits irreversible degeneration, while mild cognitive impairment (MCI) is viewed as the preliminary stage of AD, and its pathogenesis is notably intricate. MCI contains two stages: early MCI (EMCI), and late MCI (LMCI). EMCI diagnosis can prevent EMCI from progressing to LMCI, and then to AD. Therefore, accurate diagnosis of EMCI/LMCI is crucial for developing the early intervention and treatment strategies of AD. Currently, most existing EMCI/LMCI diagnostic methods use single modality images, while different modality images carry different complementary information that helps for accurate diagnosis of EMCI/LMCI, and the lesion area is usually not limited to a single brain area, which involves multiple regions. In this case, conventional convolution operations cannot be able to accurately extract the pathological features of AD. In this work, we propose a novel Multi-scale fully Separable Convolution neural network with Large Kernels (MSCLK) method to diagnose early Alzheimer's disease with structural Magnetic Resonance Imaging (sMRI) images. MSCLK mainly consists of the multi-scale 3D fully separable convolution modules and the deep metric learning module. The multi-scale convolution that contains both small and large kernels is used to effectively capture the discrimination features of different scale acceptance domains. 3D fully separable convolution is used to reduce parameters and overfitting. The deep metric learning is used to learn hard samples that are similar but belong to different classes. We also propose a variant method of MSCLK (called MSCLK-Fusion MRI and PET, MSCLK-FMP) by adding the pixel-level fusion module and feature-level fusion module into the MSCLK framework to integrate the sMRI image and the Positron Emission Computed Tomography (PET) image for further improving the accuracy of EMCI vs. LMCI classification task. The pixel-level fusion is used to achieve early pixel-level fusion of sMRI and PET images, and the feature-level fusion is used to achieve high-dimensional feature-level fusion of sMRI and PET images. Experimental results on the ADNI database show that the performance of our MSCLK and MSCLK-FMP are superior to other state-of-the-art methods. The accuracy of MSCLK achieves 98.89%, 95.97%, 96.39% and 98.76% for AD vs. EMCI, AD vs. LMCI, EMCI vs. NC and LMCI vs. NC classification tasks, respectively, and MSCLK-FMP achieves 93.93% for EMCI vs. LMCI classification task, indicating that MSCLK/MSCLK-FMP can be effectively used for diagnosing MCI patients. Moreover, our MSCLK-FMP is capable of pinpointing key brain areas involved in the pathological progression of MCI, such as the Temporal_Inf, the Hippocampus, the Precuneus, the Precentral, and the Thalamus. These findings contribute to uncovering the early onset of AD pathogenesis.

1. Introduction

Alzheimer's disease (AD) is the most common irreversible primary degenerative disease of central nervous system in the middle to late stage (Alzheimer's Association, 2023). Early detection of potential AD patients and prediction of AD progression trajectory are of utmost importance. Mild Cognitive Impairment (MCI), an intermediate stage of memory neuron damage, lies between normal aging and early

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

* Corresponding author.
E-mail addresses: tianrunfeng@mail.nwpu.edu.cn (R. Tian), lijiani@mail.nwpu.edu.cn (J. Li), zhangsw@nwpu.edu.cn (S. Zhang).

[1] These authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.eswa.2024.124241
Received 19 January 2024; Received in revised form 17 April 2024; Accepted 13 May 2024
Available online 16 May 2024

Alzheimer's disease in terms of pathophysiology. MCI patients frequently exhibit minor deficits in memory, language, spatial perception, and executive function. Many researchers have shown that early diagnosis and treatment of MCI patients can significantly reduce the conversion rate from MCI to AD (Wang et al., 2019). Mild Cognitive Impairment (MCI) is categorized into two disease stages: early MCI (EMCI) and late MCI (LMCI) (Wang et al., 2019). EMCI is considered the early stage of AD clinical spectrum, while LMCI is the late stage of AD progression. Early diagnosis of MCI can prevent MCI patients from developing into AD (Kang et al., 2020).

With the rapid development of medical imaging technology, neuroimaging technology plays an important role in computer-aided diagnosis. Especially, magnetic resonance imaging (MRI) technology has been widely used in AD diagnosis due to its advantages of high resolution, non-invasive, and multi-directional imaging, which can capture the anatomical and functional features of the brain related to AD (Zhang et al., 2021, 2022; Zheng et al., 2022), distinguish soft tissues (such as white matter and gray matter), and clearly display the anatomical structure of the normal brain (Bi et al., 2020; Liu et al., 2017). For MCI patients, their brain gray matter undergoes changes such as reduced volume and density. However, in the early stages of MCI, the morphological differences in brain lesions detected by MRI are relatively small (Cui et al., 2022). Given that 18F fluorodeoxyglucose positron emission tomography (18F FDG PET) can measure the brain glucose metabolism and detect the molecular-level lesions, it has been reported as an effective biomarker for MCI patients (Cui et al., 2022). Some early MCI diagnostic studies have shown that 18F FDG PET is more effective than other imaging modes (Cui et al., 2022; Hao et al., 2020; Pan et al., 2020; Rubinski et al., 2020). Therefore, in recent years, using PET images to diagnose MCI patients has received widespread attention.

According to the brain partition with different scales, existing CNN-based AD diagnosis methods on brain images can be roughly categorized into 2D slice-level methods (Kang et al., 2021; Ravi et al., 2024; Yagis et al., 2021), 3D patch-level methods (Qiang et al., 2023), region-level methods (Liu et al., 2020; Pan et al., 2022), and 3D subject-level methods (Folego et al., 2020; Helaly et al., 2022; Jo et al., 2020). For example, Ravi et al. (2024) proposed a deep learning-based approach to achieve the multi-stage (i.e. NC, EMCI, LMCI, AD) diagnosis of AD by using 2D slice images from 3D MRI. Qiang et al. (2023) proposed a DANMLP model to implement the classification tasks of AD vs. MCI, AD vs. NC, and MCI vs. NC by using 3D patch images. DANMLP allows for extracting multiple patches from a single 3D sMRI image to increase the number of samples while preserving the spatial information. Pan et al. (2022) divided sMRI images into 246 ROIs and inputted them into 246 3DCNN models to achieve the classification tasks of AD vs. NC, MCIc vs. NC, and MCIc vs. MCInc through voting. Helaly et al. (2022) and Jo et al. (2020) inputted the whole brain images into CNN models to achieve the classification tasks of AD/EMCI/LMCI/NC. While 2D slice-level methods can augment the sample size by selecting numerous 2D slices from a 3D image, they can cause a loss of spatial information inherent in the original 3D image, hence negatively impacting the enhancement of performance. For 3D patch-level methods, dividing brain images into multiple patches may affect the structural information of complete brain regions. For the region-level methods, although they can preserve structural information of brain regions by selecting specific regions (ROIs) of brain MRI images, they only concentrate on particular areas, ignoring the association between different brain regions in early diagnosis of AD. Although 3D subject-level methods can fully preserve all information of whole brain images, they may be a risk of overfitting due to the small sample size, and they also have high computational complexity. Therefore, how to maximize the utilization of all information in brain images and reduce the number of parameters is a major challenge.

Although current AD classification methods have shown good results, the diagnostic accuracy of EMCI/LMCI still faces challenges for clinical applications, and most of them cannot provide interpretable findings in the clinical settings. This is primarily due to the fact that most studies on EMCI/LMCI classification (Jo et al., 2020; Ravi et al., 2024) end to use traditional CNN architectures (e.g., Visual Geometry Group (VGG)) with single-scale convolution kernels to extract features from sMRI images. However, the etiology of EMCI/LMCI is complicated and the brain images of EMCI and LMCI patients are highly similar. This makes it is extremely difficult to extract effective discriminative features using single scale convolution. In addition, most studies only use single modal images, while fusing different modal images helps to further improve the diagnostic accuracy of EMCI/LMCI.

In this work, we designed a novel multi-scale fully separable convolution neural network with large kernels (called MSCLK) for early diagnosis of AD on sMRI images. In the MSCLK framework, we introduced the multi-scale convolution that contains both small and large convolution kernels to effectively extract the discrimination features from different size receptive fields in sMRI images, and designed a 3D fully separable convolution module to reduce parameters and overfitting. In order to further enhance the performance of MSCLK, we introduced the deep metric learning to learn hard samples with high similarity but belonging to different categories (i.e., EMCI, LMCI). We also presented a novel MSCLK-FMP method for high-precision diagnosis of EMCI /LMCI by integrating sMRI and PET images. MSCLK-FMP is a variant of MSCLK, which adds the pixel/feature-level fusion modules into the MSCLK framework. The experimental results on the ADNI database show that our MSCLK is effective in the early diagnosis of AD. MSCLK-FMP can accurately diagnose EMCI and LMCI, providing interpretable results.

In summary, the primary contributions of our work can be summarized as follows:

- We designed a multi-scale fully separable convolution neural network with large kernels (called MSCLK) for early diagnosis of Alzheimer's disease. In MSCLK, the multi-scale convolutional with large kernels is used to extract the discriminatory features in different scales, and 3D fully separable convolutions are used for network reparameterization to prevent overfitting due to excessive parameters.
- We introduced the contrastive Bayesian metric learning to learn hard samples with high similarity but belonging to different categories (e.g., EMCI and LMCI).
- We also proposed a variant method of MSCLK (called MSCLK-FMP) to accurately diagnose EMCI and LMCI by integrating sMRI and PET images. The pixel/feature-level fusion strategies introduced in MSCLK-FMP are used to achieve in-depth interaction between sMRI and PET images from EMCI/LMCI patients.
- Experimental results show that our MSCLK can be effectively used for early diagnosis of AD patients. MSCLK-FMP can diagnose EMCI and LMCI patients with high accuracy, and also identify the brain regions that play key roles in the pathological progression of MCI.

2. Related work

2.1. EMCI/LMCI classification with deep learning

Deep learning (DL) models, particularly convolutional neural networks (CNNs), have demonstrated superior feature learning capabilities, finding extensive use in numerous medical image analysis applications (Bernal et al., 2019; Gu et al., 2020; Khairandish et al., 2022). CNNs not only can eliminate the tedious process of manual feature extraction and reduce the dependence on data preprocessing, but also efficiently learn higher-level features from brain imagery, significantly enhancing the accuracy of brain disease diagnosis (Anwar et al., 2018). For example, Fan et al. (2021) proposed a U-net style model for AD diagnosis by introducing the deep supervision connections, using the modified skip-connections and symmetric encoders/decoders. For four binary classification tasks of AD vs. NC, EMCI vs. LMCI, AD vs. LMCI, and NC vs. EMCI, the accuracy of U-net on 3D MRI images achieves 95.71%, 90.14%, 90.05%, and 87.98%, respectively. For the multi-class classification task of AD vs. NC vs. EMCI vs. LMCI, its accuracy achieves 86.47%. Mehmood et al. (2021) proposed a layer-wise transfer learning model with a VGG-19 architecture for early diagnosis of AD on MRI images. In their model, Mehmood et al. applied the transfer learning to freeze the convolutional layers, and progressively froze the blocks of layers for training with and without augmentation dataset, and the classification accuracy achieves 83.72% for EMCI vs. LMCI task. Cui et al. (2022) introduced a bilinear pooling module into the CNN framework to extract the inter-region representation features of the whole brain, and used the deep metric learning to learn the hard samples in embedding space, then they proposed a BMNet method for classifying EMCI and LMCI with FDG-PET images. The classification accuracy of BMNet achieves 79.64% for EMCI vs. LMCI task. Odusami et al. (2023) proposed a concatenation-based feature fusion framework to fuse the MRI and PET images, and then used the modified Resnet18 network to classify EMCI and LMCI with 73.9% accuracy. Most existing deep learning-based early AD diagnosis methods usually use mono-modality data (e.g., MRI or PET), and there is a significant gap between the classification accuracy of these methods and the clinical application of accurately diagnosing EMCI and LMCI patients. Although some deep learning-based multi-modality fusion methods have been developed to improve the classification accuracy of EMCI vs. LMCI, they just use the simple concatenated feature fusion strategy, and their improvement in classification accuracy is limited. Therefore, in this work we present the pixel-level and feature-level fusion strategies for deep interaction between sMRI and PET images, and introduce the deep metric learning to distinguish hard samples with high similarity but belonging to a different category (i.e., EMCI, LMCI), thereby achieving a high-precision diagnosis of EMCI/LMCI patients.

2.2. Multi-scale CNNs for medical imaging

Due to the different positions of medical image information, it is necessary to choose convolution kernels of different sizes for CNN models. That is, if the medical image information is global, choose a large convolution kernel; otherwise, choose a smaller convolution kernel. Using different scale convolution kernels (e.g., 1×1, 3×3, and 5×5) and pooling in the same layer network can obtain different size receptive fields of the medical images, and the features from different scales are fused and passed to the next layer for improving model performance. In view of that multi-scale CNNs can extract different features of medical images at different scales, multi-scale CNNs have been used to improve the performance of medical image classification models (Bakkouri & Afdel, 2019; Ding et al., 2022; Liu et al., 2018, 2022; Pei et al., 2022). For example, Liu et al. (2018) proposed a method based on multi-view multi-scale CNNs for classifying the lung nodule type from CT images. This method adopts a multi-scale way to ensure the information capture of nodules and their surroundings. Bakkouri and Afdel (2019) proposed an approach based on multi-scale CNN to distinguish between normal and abnormal breast tissue by utilizing multi-scale convolutional deep layers to extract and select the most significant features. Liu et al. (2022) proposed a multi-scale convolutional neural network (MSCNet) with the aim of improving the model's feature representation ability for diagnosing Alzheimer's Disease (AD). MSCNet uses multiple convolution kernels of different sizes and changes the size of the convolution kernel by setting different dilation rates to extract multi-scale features, and fuse the features between channels for obtaining more comprehensive information. Pei et al. (2022) collected the information from multiple convolution kernels for capturing different scales of the target object, thereby proposed a multi-scale attention-based pseudo 3D convolution neural network for AD diagnosis. However, existing multi-scale CNN models for AD diagnosis generally adopt the small size convolutional kernels, such as 1×1, 3×3, and 5×5, and their receptive fields are still small receptive fields, while the design of large kernels requires fewer layers to obtain large effective receptive fields (ERFs), as well as avoids the optimization issue brought by the increasing layer depth. Using a few large kernels instead of many small kernels can result in larger ERF more efficiently, boosting CNN model performances (Ding et al., 2022). Therefore, in this work we design a novel multi-scale convolution with large kernels for AD early diagnosis, designing a novel 3D fully separable convolution to solve the problem of excessive parameters caused by large kernel convolution.

Table 1
Statistics information of AD, LMCI, EMCI, and NC subjects.

Subject	Numbers	Age	Gender (Male/Female)
AD	250	75.46 ± 7.65	119/131
LMCI	72	73.77 ± 5.85	39/33
EMCI	110	73.36 ± 6.25	58/52
NC	250	75.67 ± 5.95	134/116

3. Materials and methods

In this section, we first present the data used in this work, introducing how to preprocess these brain images. Then, we show our MSCLK and MSCLK-FMP models in detail. Finally, we provide some metrics for evaluating the performance of models.

3.1. Dataset and processing

We downloaded the sMRI and PET images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) (Jr et al., 2008), and selected the subjects with sMRI and PET images. Then, we chose the magnetization prepared rapid gradient-echo (MPRAGE) T1-weighted three-dimensional (3D) sMRI images, and the radioactive drug Fluorodeoxyglucose (18F-FDG) PET images. Finally, the sMRI and PET images from 682 subjects (i.e., 250 AD subjects, 72 LMCI subjects, 110 EMCI subjects, and 250 NC subjects) are selected as the training and testing samples. Each subject has one 3D sMRI image and one 3D PET image. The statistics information of the experimental subjects used in this work is shown in Table 1.

Due to the differences in collection equipment and time, the original medical images obtained from the ADNI database may have certain differences among subjects. Therefore, before performing subsequent classification tasks, preprocessing of these images is a necessary step. For this purpose, we used a medical image processing tool package of Statistical Parameters Mapping 12 (SPM12) (Ashburner et al., 2014) to preprocess the MRI and PET images.

The preprocessing process of sMRI images is shown in Fig. 1a. Firstly, we use the Computational Anatomy Toolbox12 in SPM12 to perform anterior committee (AC)–posterior committee (PC) correction on sMRI images. Next, we utilize the Tissue Probability Map (TPM) template in SPM12 to segment the MRI images into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) images. Finally, we use affine linear registration to map the GM, WM, and CSF images into the Montreal Institute of Neurology (MNI) standard space, and also normalize these images. In this work, only GM images are used, as it has been proven that significant brain atrophy in AD patients is mainly concentrated in the gray matter region. As a result, we obtain the GM images with resolution.

The preprocessing process of PET is shown in Fig. 1b. Firstly, we register the PET images of each subject with their corresponding MRI images, as there are differences in spatial resolution and morphology between PET and MRI images. Then, we use the average template generated by sMRI images of all the subjects to spatially normalize all PET images, reducing subject differences. Finally, all images are mapped into the MNI space. As a result, we obtain PET images with resolution. The pipeline of PET images preprocessing is shown in Fig. 1b.

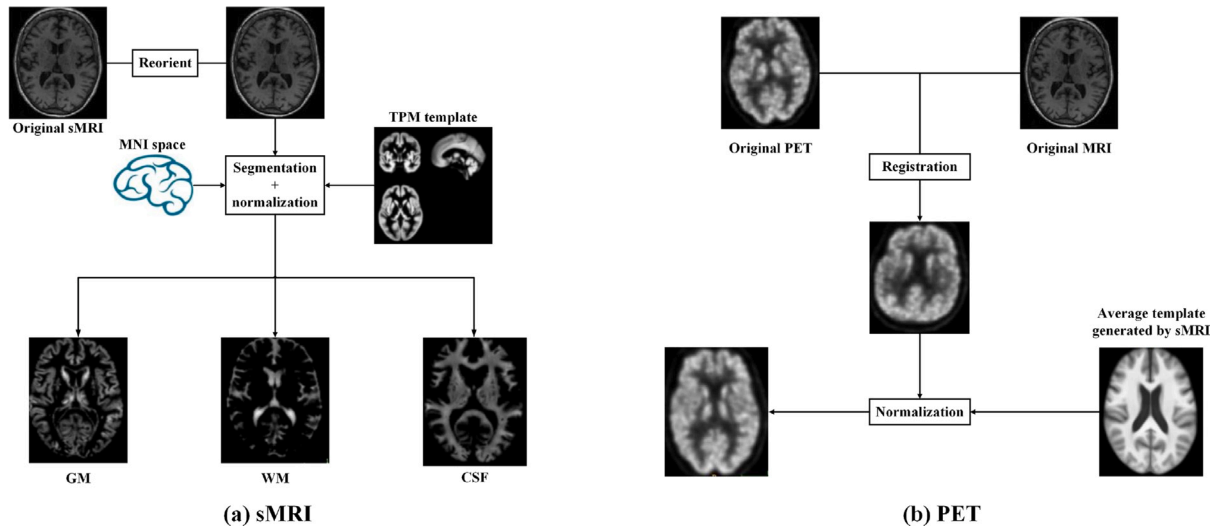

Fig. 1. The pipeline of images preprocessing. (a) sMRI images, (b) PET images.

3.2. Overall architecture of MSCLK and MSCLK-FMP

Let $X = \{(x_i, y_i) \,|\, i = 1, 2, \ldots, N\}$ denotes the subject dataset, $x_i^{MRI} \in R^{D\times H\times W} (i = 1, 2, \ldots, N)$ denotes the sMRI image of i-th subject, $x_i^{PET} \in R^{D\times H\times W} (i = 1, 2, \ldots, N)$ denotes the PET image of i-th subject, and y_i denotes the label of i-th subject. Our MSCLK (Fig. 2A) consists of five components: multi-scale convolution with small and large kernels (MCK), 3D fully separable convolution (FSC), skip connections, contrastive Bayesian metric learning (CBML), and fully convolution neural network (FCN). The MCK blocks and FSC blocks form one multi-scale 3D fully separable convolution (MKSC) module. The MCK block is used to effectively capture the discrimination features of different scale acceptance domains. FSC block is used to reduce parameters and overfitting. The skip connections are used to associate low-dimensional features with high-dimensional features. The CBML module is used to learn the hard samples that have high similarity but belong to different classes. FCN is used to output the predicted labels of subjects. Our MSCLK-FMP (Fig. 2B) mainly consists of five parts: pixel-level fusion module, feature-level fusion module, MSCLK networks, CBML module, and FCN module. The pixel-level fusion module is used to achieve early pixel fusion of sMRI and PET images, and the feature-level fusion module is used to achieve high-dimensional feature fusion of sMRI and PET images. The inputs of MSCLK are the sMRI images, while the inputs of MSCLK-FMP are the sMRI and PET images.

3.2.1. MKSC module

MKSC (Fig. 2a) is composed of four FSC blocks with different large/small convolutional kernels to extract the features from images. Convolutional operations with large/small kernels can obtain the acceptance domain of different scales and capture the discrimination features of multiple scales more effectively. In MKSC, we adopt a channel parallel approach to integrate the features of different scales.

$$\begin{aligned} G1 &= FSC\ Block_1(E1) \\ G2 &= FSC\ Block_2(E2) \\ G3 &= FSC\ Block_3(E3) \\ G4 &= FSC\ Block_4(E4) \end{aligned} \tag{1}$$

$$E' = concat\,(G1, G2, G3, G4) \tag{2}$$

where E is the input of MSCLK module, $G1$ and $G2$ are the output of FSC Block with small convolutional kernels (e.g., $1\times1\times3$, $1\times5\times1$, $7\times1\times1$, $1\times1\times9$), $G3$ and $G4$ are the output of FSC Block with large convolutional kernels (e.g., $1\times1\times11$, $1\times13\times1$, $15\times1\times1$, $1\times1\times17$), $E' \in R^{C'\times D\times H\times W}$ is convolution feature map, and the 'concat' is a concatenation operator.

In MKSC, the channel self-attention mechanism (Hu et al., 2018) is used to identify the important features at different channels. We first obtain the $E'' \in R^{C'\times1\times1\times1}$ feature map by performing the AdaptiveAvgPool3d operation, and then calculate the attention score $S_A \in [0, 1]$ through convolutional layers and activation functions. Finally, multiplying the channel attention weight score S_A with the convolutional feature map E' yields the output Q of the MKSC module.

$$E'' = AdaptiveAvgPool3d\,(E') \tag{3}$$

$$S_A = Sigmoid(Conv3d(\mathrm{ReLU}(Conv3d(E'')))) \tag{4}$$

$$Q = S_A \cdot E' \tag{5}$$

3.2.2. FSC block

FSC block (Fig. 2b) is designed to re-parameterize 3D convolutional kernels, which uses the asymmetric convolutional kernels to sequentially perform the depth-wise convolution along the direction of width (with $1\times1\times k$ kernel), height (with $1\times k\times1$ kernel), and depth (with $k\times1\times1$ kernel), then execute the pointwise convolution with $1\times1\times1$ kernel. Compared to normal kernels, 3D fully separable convolution can reduce the parameters to $3/\mathrm{k}^2$, where k is the size of normal kernel (in this work, k = 3, 5, 7, 9, 11, 13, 15, 17).

3.2.3. CBML module

Compared with AD and NC subjects, the images of EMCI and LMCI subjects are very similar, making it difficult to accurately distinguish EMCI and LMCI based on the images. Therefore, we introduce contrastive Bayesian metric learning (Kan et al., 2022) to design the CBML module (Fig. 2c) to learn the hard samples with high similarity but belonging to EMCI or LMCI subjects. Let $X^+ = \{(x_i, x_j) | y_i = y_j\}$ denotes the set of positive sample pairs, $X^- = \{(x_i, x_j) | y_i \neq y_j\}$ denotes the set of negative sample pairs; $m(h_i, h_j)$ represents the metric (e.g., cosine similarity) between feature vectors h_i and h_j that are extracted from subjects x_i and x_j by flattening z_i and z_j, respectively, where z_i and z_j are the final feature maps extracted from subjects x_i and x_j. For a given subject x_i, let $X_i^+ = \{x_j | (x_i, x_j) \in X^+\}$ is the set of positive samples for x_i, $X_i^- = \{x_j | (x_i, x_j) \in X^-\}$ is the set of negative samples for x_i. Thus, for the subject x_i, we can define its set of hard positive sample pairs ${}^*X_i^+$, and its set of hard negative sample pairs ${}^*X_i^-$ as follows,

$$\begin{aligned} {}^*X_i^+ &= \{x_j \in X^+ | m(h_\mathrm{i}, h_\mathrm{j}) < \Gamma^{\max}_{X_\mathrm{i}^-} + \varepsilon\} \\ {}^*X_i^- &= \{x_j \in X^- | m(h_\mathrm{i}, h_\mathrm{j}) > \Gamma^{\max}_{X_i^+} + \varepsilon\} \end{aligned} \tag{6}$$

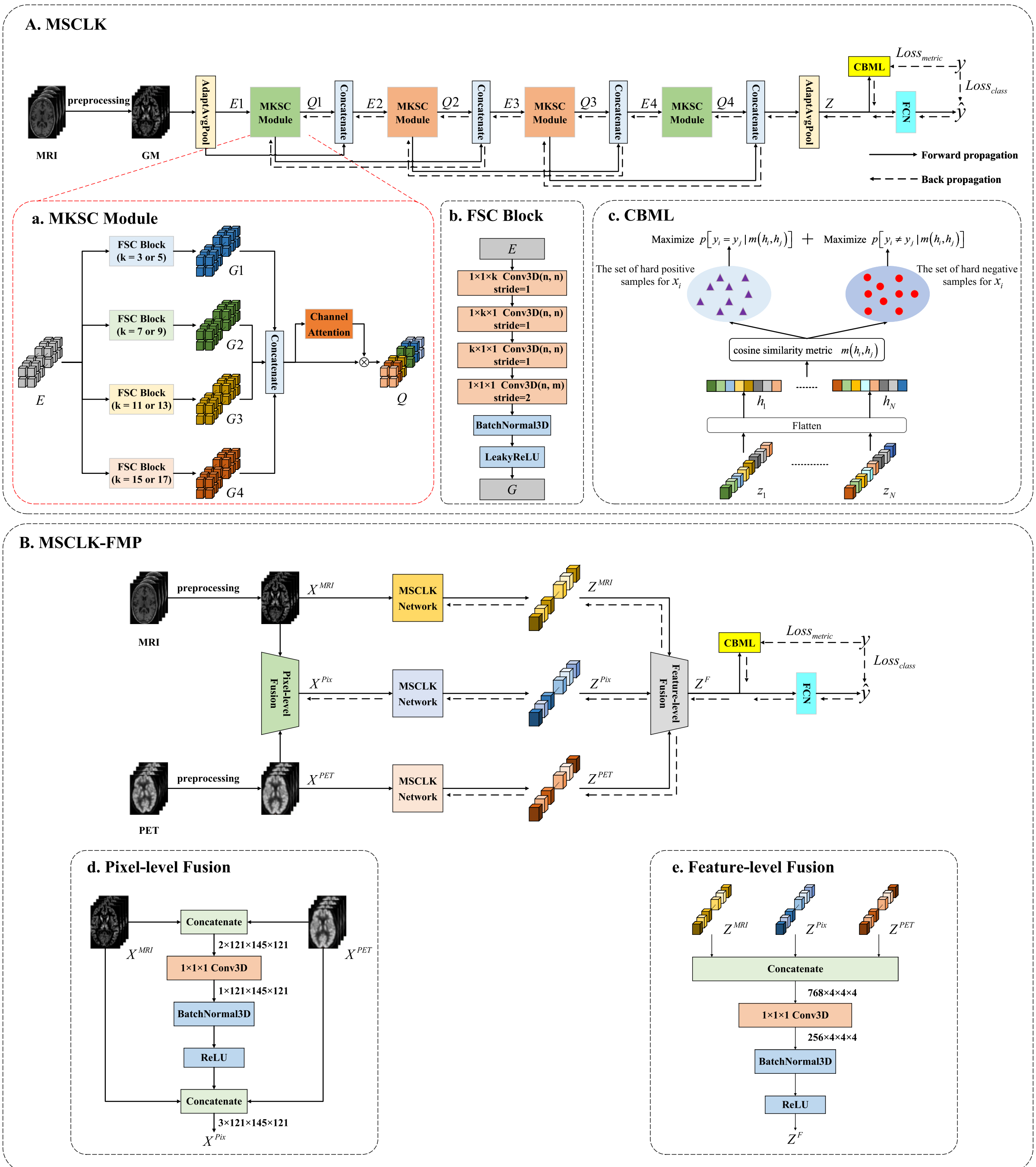

Fig. 2. The Schematic of MSCLK and MSCLK-FMP models. (A) MSCLK. It consists of three key components: multi-scale convolution, 3D fully separable convolution (FSC), and contrastive Bayesian metric learning (CBML). FSC blocks with different small and large kernels form one multi-scale 3D fully separable convolution (MKSC) module. (B) MSCLK-FMP. It consists of four key components: pixel-level fusion, feature-level fusion, MSCLK Network, and CBML. (a) MKSC module is used to extract features from images. (b) FSC block is used to alleviate overfitting caused by excessive parameters. (c) CMBL module is used to learn the hard samples that have high similarity but belong to different classes. (d) Pixel-level fusion module is used for early pixel fusion of sMRI and PET images. (e) Feature-level Fusion module is used for the fusion of high-dimensional features of sMRI and PET images.

where $\Gamma^{\max}_{X_i^-} = \max\left\{m(h_i,h_j)|x_j \in X_i^-\right\}$, $\Gamma^{\max}_{X_i^+} = \max\left\{m(h_i,h_j)|x_j \in X_i^+\right\}$, and ε is a marginal threshold to control the number of selected examples (here we set it to 1e-5 by experience).

Thus, we select the hard positive/negative sample pairs from the current mini-batch to iteratively learn the feature representation. That is to say, for the hard positive sample pairs $(x_i,x_j) \in {}^*X^+$, we need sure that the learned feature representation can maximize the conditional posterior probability $p\left[y_i = y_j|m\left(h_i,h_j\right)\right]$; for the hard negative sample pairs $(x_i,x_j) \in {}^*X^-$, maximize the conditional posterior probability $p\left[y_i \neq y_j|m\left(h_i,h_j\right)\right]$.

$$p\left[y_i = y_j|m\left(h_i,h_j\right)\right] = \frac{1}{1+\exp\left\{\frac{\alpha^{X^+}-m(h_i,h_j)}{\beta^{X^+}}\right\}\cdot\frac{\left|X_i^-\right|}{\left|X_i^+\right|}} \tag{7}$$

$$p\left[y_i \neq y_j|m\left(h_i,h_j\right)\right] = \frac{1}{1+\exp\left\{\frac{m(h_i,h_j)-\alpha^{X^-}}{\beta^{X^-}}\right\}\cdot\frac{\left|X_i^+\right|}{\left|X_i^-\right|}} \tag{8}$$

$$m(h_i, h_j) = \frac{\sum_{d=1}^{D} h_{id} h_{jd}}{\sqrt{\sum_{d=1}^{D} h_{id}^2}\sqrt{\sum_{d=1}^{D} h_{jd}^2}} \tag{9}$$

$$H = [h_1, h_2, \dots, h_i, \dots, h_N] = Flatten(Z) \tag{10}$$

$$\begin{aligned} Z &= [z_1, z_2, \dots, z_i, \dots, z_\mathrm{N}] \\ &= AdaptiveAvgPool3d\,(concat(Conv3d(Q3), Q4)) \end{aligned} \tag{11}$$

where $\alpha^{X^+}, \alpha^{X^-}, \beta^{X^+}, \beta^{X^-}$ are the parameters for the exponential function, $|\cdot|$ is the size of the sample set, $Q3$ and $Q4$ are the output of 3-th and 4-th MKSC modules, 'AdaptiveAvgPool3d' is the adaptive average pooling layer, the 'concat' is a concatenation operator, the 'Conv3d' is the convolution operator with $1 \times 1 \times 1$ kernel, 'Flatten' is a flattening operation.

3.2.4. Pixel-level fusion module

We designed the pixel-level fusion module (Fig. 2d) to achieve pixel fusion of sMRI and PET. In this module, we first use $1 \times 1 \times 1$ convolution to fuse the sMRI and PET images of i-th subject at the corresponding position to generate the fusion image x_i^*, then concatenate x_i^* with x_i^{MRI} and x_i^{PET} on channels to obtain the pixel-level fusion image x_i^{Pix} of i-th subject.

$$x_i^* = \mathrm{Re}\,LU\left(B\mathrm{N}\left(Conv3d\left(concat\left(x_i^{MRI}, x_i^{PET}\right)\right)\right)\right) \tag{12}$$

$$x_i^{Pix} = concat\left(x_i^*, x_i^{MRI}, x_i^{PET}\right) \tag{13}$$

where, $x_i^{MRI} \in R^{1\times121\times145\times121}$ is the sMRI image of i-th subject, and $x_i^{PET} \in R^{1\times121\times145\times121}$ is the PET image of i-th subject, $x_i^* \in R^{1\times121\times145\times121}$ is the fusion image by fusing sMRI and PET images of i-th subject with $1 \times 1 \times 1$ convolution kernel, $x_i^{Pix} \in R^{3\times121\times145\times121}$ is pixel-level fusion image of i-th subject, 'BN' is the Batch Normalization operator, the 'Conv3d' is the convolution operator with kernel, and the 'concat' is a concatenation operator.

3.2.5. Feature-level fusion module

We also designed a feature-level fusion module (Fig. 2e) to fuse the high-dimensional features extracted from x^{MRI}, x^{PET} and x^{Pix}. In this module, we use $1 \times 1 \times 1$ convolution to fuse and compress the discriminative features (i.e., z_i^{MRI}, z_i^{PET}, z_i^{Pix}) extracted from x_i^{MRI}, x_i^{PET}, and x_i^{Pix} to obtain the feature-level fusion feature z_i^F of i-th subject.

$$z_i^F = \mathrm{Re}\,LU\left(B\mathrm{N}\left(Conv3d\left(concat\left(z_i^{MRI}, z_i^{PET}, z_i^{Pix}\right)\right)\right)\right) \tag{14}$$

where $z_i^{MRI} \in R^{256\times4\times4\times4}$, $z_i^{PET} \in R^{256\times4\times4\times4}$, $z_i^{Pix} \in R^{256\times4\times4\times4}$ and $z_i^F \in R^{256\times4\times4\times4}$.

3.2.6. Loss function of MSCLK and MSCLK-FMP

The loss function $Loss$ of MSCLK and MSCLK-FMP is composed of the cross-entropy loss $Loss_{class}$ for classification tasks, and the deep metric loss $Loss_{\mathrm{metric}}$ for learning the hard samples.

$$Loss = Loss_{class} + \varepsilon Loss_{metric} \tag{15}$$

$$Loss_{class} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log \hat{y}_i + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right) \tag{16}$$

$$\begin{aligned} Loss_{metric} = {} & \mathbb{E}_{x_i\in X} \log\left\{1 + \delta^+ \sum_{x_j \in {}^*X_i^+} \exp\left[\frac{\alpha^{X^+} - m(h_i,h_j)}{\beta^{X^+}}\right]\right\} \\ & + \mathbb{E}_{x_i\in X} \log\left\{1 + \delta^- \sum_{x_j \in {}^*X_i^-} \exp\left[\frac{m(h_i,h_j) - \alpha^{X^-}}{\beta^{X^-}}\right]\right\} \end{aligned} \tag{17}$$

$$\delta^+ = |X^-| / |X^+|^2, \delta^- = |X^+| / |X^-|^2 \tag{18}$$

where X is the training subject set, N is the number of the training subjects, y_i is the true label of i-th subject, $\hat{y}_i$ is the predicted label of i-th subject; $\mathbb{E}_{x_i\in X}\{\cdot\}$ represents the arithmetic average operation for the set X, X^+ represents the set of positive subject pairs, X^- represents the set of negative subject pairs, $|\cdot|$ is the size of the subject set, $m\left(h_i, h_j\right)$ represents the cosine similarity between feature vectors h_i and h_j extracted from subjects x_i and x_j; ${}^*X_i^+$, ${}^*X_i^-$ are the sets of hard positive subject pairs and the hard negative subject pairs for subject x_i, respectively; $\alpha^{X^+}, \alpha^{X^-}, \beta^{X^+}, \beta^{X^-}$ are the parameters for the exponential function, ε is the hyperparameter (here we set it to 0.2 by experience).

3.2.7. Evaluation metrics

To evaluate the performance of our MSCLK/ MSCLK-FMP and other methods in the classification tasks of AD vs. NC, AD vs. EMCI, AD vs. LMCI, EMCI vs. NC, LMCI vs. NC, and EMCI vs. LMCI, we use three metrics of accuracy (ACC), sensitivity (Sn), and specificity (Sp), which are defined as follows,

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{19}$$

$$Sn = \frac{TP}{TP + FN} \tag{20}$$

$$Sp = \frac{TN}{TN + FP} \tag{21}$$

where TP is the number of predicted true positive subjects, FP is the number of predicted false positive subjects, TN is the number of predicted true negative subjects, and FN is the number of predicted false negative subjects.

4. Results and discussion

4.1. Experimental settings

All experiments are conducted on Linux OS server with 24G × 4 NVIDIA RTX 3090 GPU, 40 × 2.4 GHz Intel Xeon CPU, and 128 GB RAM. Python 3.9 and PyTorch 1.12.1 are adopted to implement MSCLK and MSCLK-FMP. The five-fold cross validation (5CV) test approach is used to evaluate the performance of our MSCLK and MSCLK-FMP and other comparative methods. In the 5CV test, all subjects are divided randomly into five equally-sized groups, approximately. Each group is selected as the test set in turn, while the remaining subjects in four groups are further randomly divided into the training subjects and the validation subjects in a 3:1 ratio.

According to some Refs. Liu et al. (2018), Liu et al. (2022) and Pei et al. (2022) and the experimental results of some combinations with different kernel sizes, we choose the kernel size for FSC blocks in four MKSC modules. Due to the numerous combination schemes with different kernel sizes, it is impossible to traverse all combination experiments. Thus, we set the upper and lower bounds of FSC block 1 and block 4 to 3 and 15 respectively, set the upper and lower bounds of block 2 and block 3 to 3 and 15 respectively, and also set the kernel size of block 2 and block 3 by adding 2 to the kernel size in FSC block 1 and block 4. Figure. S1 gives the results of four different scale convolutional kernel combinations, here we set the kernel size of MKSC module 1 and MKSC module 4 to be equal, and the kernel size of MKSC module 2 and MKSC module 3 to be equal. From Figure. S1, we can see that the kernel size in FSC blocks has an impact on ACC of our MSCLK, and ACC of FSC blocks with kernel size [3,7,11,15] (for MKSC module 1, 4) and [5,9,13,17] (for MKSC module 2, 3) achieves the highest (i.e., 85.71%). Therefore, in our work we set the kernel size of FSC blocks to [3,7,11,15] for MKSC module 1 and 4, [5,9,13,17] for MKSC module 2 and 3.

To investigate how the convolutional kernel larger than 17 in FSC blocks affects the performance, we implemented MSCLK with different convolutional kernel sizes larger than 17 in FSC blocks, as shown in

Table 2
Results of MSCLK and other four comparative methods on sMRI images.

Task	Metric	Fan et al. (2021)	Mehmood et al. (2021)	Kong et al. (2022)	Sharma et al. (2022)	MSCLK
AD vs. EMCI	ACC (%)	95.56	87.50	96.39	91.39	**98.89**
	Sn (%)	96.84	88.65	97.57	95.47	**98.83**
	Sp (%)	92.84	83.99	93.81	80.67	**99.09**
AD vs. LMCI	ACC (%)	93.79	88.50	93.78	89.75	**95.97**
	Sn (%)	93.08	92.31	97.08	97.64	**98.06**
	Sp (%)	95.49	73.91	80.87	62.60	**88.89**
EMCI vs. NC	ACC (%)	92.50	89.44	93.89	89.17	**96.39**
	Sn (%)	89.02	76.64	90.92	76.20	**91.95**
	Sp (%)	93.97	95.26	95.22	94.76	**98.40**
LMCI vs. NC	ACC (%)	93.48	91.62	94.72	91.30	**98.76**
	Sn (%)	76.01	70.60	80.72	66.52	**97.14**
	Sp (%)	98.39	97.65	98.82	98.39	**99.17**
EMCI vs. LMCI	ACC (%)	84.61	78.63	81.88	80.23	**85.71**
	Sn (%)	**94.91**	84.83	89.19	87.74	93.53
	Sp (%)	66.95	66.96	74.37	69.95	**75.68**

Table S1. From Table S1, we can see that for the EMCI vs. LMCI classification task, the ACC of our MSCLK decreases when the kernel size is larger than 17. Therefore, we set the lower bounds of convolutional kernel size to 17 in FSC blocks.

In addition, we set different FSC block number (i.e., 3, 4, and 5) to investigate the effects of the FSC block number to our MSCLK. As shown in Table S2, we can see that the ACC of MSCLK with 4 block is higher 8.8% and 3.31% than that with 3 and 5 blocks, respectively. Therefore, in this work we set the FSC block number to 4.

The convolution channels of the MKSC module are 64, 64, 128, and 128, and the kernel sizes of the first and fourth MKSC modules are 3, 7, 11, and 15, while the kernel sizes of the second and third MKSC modules are 5, 9, 13, and 17. Other hyperparameters for MSCLK and MSCLK-FMP are set as follows: epoch = 20, batch size = 8, learning rate = 0.0001. We adopt Adam optimizer to train our MSCLK and MSCLK-FMP.

4.2. Comparison of MSCLK and MSCLK-FMP with other methods

To evaluate the performance of our MSCLK for early diagnosis of Alzheimer's disease, we compare our MSCLK with other four state-of-the-art methods, such as Fan's method (Fan et al., 2021), Mehmood's method (Mehmood et al., 2021), Kong's method (Kong et al., 2022), and Sharma's method (Sharma et al., 2022). Fan et al. (2021) introduced skip connections and symmetric encoders/decoders into U-net to design a U-net style model to execute AD-related classification tasks on 3D sMRI images. Mehmood et al. (2021) utilized VGG-19 architecture to design a layer-wise transfer learning model for AD early diagnosis on 2D MRI images. Kong et al. (2022) proposed a three-dimensional convolutional network structure for AD early diagnosis using MRI and PET images. Sharma et al. (2022) proposed a 2D CNN-based ensemble RVFL classifier for AD diagnosis using the slices of MRI and PET images. In this work, we implement Kong's method and Sharma's method just on sMRI images. The results of our MSCLK and other four comparative methods are shown in Table 2.

From Table 2, we can see that for AD vs. EMCI classification task, the ACC of our MSCLK is 3.33%, 11.39%, 2.50%, and 7.50% higher than Fan's method, Mehmood's method, Kong's method and Sharma's method, respectively; for AD vs. LMCI classification task, the ACC of MSCLK is 1.18%, 7.47%, 1.19%, and 6.22% higher than Fan's method, Mehmood's method, Kong's method, and Sharma's method, respectively; for EMCI vs. NC classification task, the ACC of MSCLK is 3.89%, 6.95%, 2.50%, and 7.22% higher than Fan's method, Mehmood's method, Kong's method, and Sharma's method, respectively; for LMCI vs. NC classification, the ACC of MSCLK is 5.28%, 7.14%, 4.04%, and 7.46% higher than Fan's method, Mehmood's method, Kong's method and Sharma's method, respectively; for EMCI vs. LMCI classification, the ACC of MSCLK is 1.10%, 7.08%, 3.83%, and 5.48% higher than Fan's method, Mehmood's method, Kong's method and Sharma's method, respectively. The performance of Mehmood's method (Mehmood et al., 2021) and Sharma's method (Sharma et al., 2022) are obviously lower than other three classification tasks, indicating that 2D slice images lost the spatial information of the original 3D image. These results show that the performance of our MSCLK is superior to other state-of-the-art methods in the five AD classification tasks.

Table 3
Results of MSCLK-FMP and other two image fusion methods on EMCI vs. LMCI task using sMRI and PET images.

Methods	Sn (%)	Sp (%)	ACC (%)
Kong et al. (2022)	90.43	85.63	88.98
Sharma et al. (2022)	91.50	82.59	86.80
MSCLK-FMP_pixel	95.36	87.30	92.85
MSCLK-FMP_feature	95.45	81.32	90.08
MSCLK-FMP	**95.87**	**93.31**	**93.93**

From Table 2, we can also find that the ACC of all methods on EMCI vs. LMCI task is 10.26% 13.28% lower than that on other four classification tasks (i.e., AD vs. EMCI, AD vs. LMCI, EMCI vs. NC, LMCI vs. NC), the Sp lower 13.21% 23.49%, indicating that the sMRI image structures of EMCI and LMCI subjects are very similar, and it is difficult to distinguish EMCI and LMCI patients solely using sMRI images.

Considering that 18F-FDG-PET imaging technology can measure the brain glucose metabolism and detect the molecular level lesions, and it has been documented that PET images can be beneficial in diagnosing patients with Mild Cognitive Impairment (MCI) (Rubinski et al., 2020), we introduced PET images into our MSCLK framework and developed three variant methods of MSCLK (called MSCLK-FMP, MSCLK-FMP_pixel and MSCLK-FMP_feature) for improving the performance of EMCI vs. LMCI classification task. MSCLK-FMP_pixel method fuses the sMRI and PET images at the pixel-level, and the MSCLK-FMP_feature method fuses the sMRI and PET images at the feature-level, while MSCLK-FMP method fuses the sMRI and PET images both at pixel-level and feature-level. The results of MSCLK-FMP, MSCLK-FMP_pixel, MSCLK-FMP_feature and other two image fusion methods (i.e., Kong et al. (2022), and Sharma et al. (2022)) on EMCI vs. LMCI task using sMRI and PET images are shown in Table 3. Kong et al. (2022) used anatomical masks created from gray matter of MRI images to cover the entire registered PET image to obtain GM-PET images, then used the FLIRT component to register and fuse the GM-PET images with the original PET images, and finally used 3D CNN for AD diagnosis. Sharma et al. (2022) used wavelet transform to perform pixel-level fusion on key 2D slices of MRI and PET, and then adopted 2D CNN for AD diagnosis.

From Table 3, we can see that the performance of MSCLK-FMP is superior to that of MSCLK-FMP_pixel, MSCLK-FMP_feature and other

Table 4
Ablation experimental results of MSCLK on sMRI images in 5CV test.

Task	Metric	MSCLK_MS	MSCLK_LK	MSCLK_CA	MSCLK_SC	MSCLK_BM	MSCLK
AD vs. EMCI	ACC (%)	87.78	92.78	95.83	95.56	97.50	**98.89**
	Sn (%)	95.36	95.93	96.73	97.11	98.83	**98.83**
	Sp (%)	71.94	84.77	93.76	91.31	94.78	**99.09**
AD vs. LMCI	ACC (%)	85.69	91.61	94.41	94.11	94.09	**95.97**
	Sn (%)	93.60	95.98	97.67	97.24	97.24	**98.06**
	Sp (%)	57.96	76.11	83.39	75.74	83.33	**88.89**
EMCI vs. NC	ACC (%)	86.94	90.83	94.44	92.78	93.06	**96.39**
	Sn (%)	80.42	81.84	88.08	80.90	86.31	**91.95**
	Sp (%)	90.11	94.39	97.20	97.45	96.05	**98.40**
LMCI vs. NC	ACC (%)	87.58	90.99	93.79	94.72	96.58	**98.76**
	Sn (%)	52.76	73.57	82.82	82.37	91.95	**97.14**
	Sp (%)	97.65	96.04	96.82	98.43	98.04	**99.17**
EMCI vs. LMCI	ACC (%)	78.05	78.57	80.77	81.88	81.31	**85.71**
	Sn (%)	89.21	93.03	92.64	90.66	88.90	**93.53**
	Sp (%)	63.11	55.83	63.28	71.82	70.61	**75.68**

two fusion methods, and the performance of MSCLK-FMP_pixel and MSCLK-FMP_feature are better than that of other two fusion methods, e.g., ACC of MSCLK-FMP achieves 93.93%, which is 4.95%, 7.13%, 1.08%, 3.85% higher than that of Kong's method (Kong et al., 2022), Sharma's method (Sharma et al., 2022), MSCLK-FMP_pixel and MSCLK-FMP_feature, respectively; ACC of MSCLK-FMP_pixel is 3.87%, 6.05% higher than that of Kong' method. Kong et al. (2022), Sharma's method (Sharma et al., 2022), respectively; ACC of MSCLK-FMP_feature is 1.1%, 3.28% higher than that of Kong's method (Kong et al., 2022), Sharma's method (Sharma et al., 2022), respectively. These results show that our methods of MSCLK-FMP, MSCLK-FMP_pixel, and MSCLK-FMP_feature can accurately diagnose EMCI and LMCI subjects. The performance of MSCLK-FMP is superior to that of MSCLK-FMP_pixel and MSCLK-FMP_feature, indicating that fusing sMRI and PET images both at pixel and feature levels can effectively enhance the classification performance of EMCI vs. LMCI task. The performance of MSCLK-FMP_pixel is better than that of MSCLK-FMP_feature, demonstrating that the contribution of pixel-level fusion is more than that of feature-level fusion. From Tables 2 and 3, we can also find that the ACC, Sn, and Sp of MSCLK-FMP are 8.22%, 2.34%, and 17.63% higher than that of MSCLK for EMCI vs. LMCI task, demonstrating that integrating sMRI and PET images can obviously boost the classification performance of EMCI vs. LMCI task.

In addition, we also investigate whether our MSCLK and other two 3D subject-level methods (i.e., Fan et al. (2021) and Kong et al. (2022)) have overfitting problems. The training and validation loss results of MSCLK and other two methods with sMRI images are shown in Figure S2 (in supplementary), from which we can see that our MSCLK shows consistent trends in training loss and validation loss during the training process for five classification tasks, while Fan's method and Kong's method show inconsistent trends, and the gap between training and validation loss of MSCLK is lower than that of Fan's method and Kong's method. These results indicate that MSCLK did not overfit, while Fan's method and Kong's method are somewhat overfitting. That is to say, 3D fully separable convolutions with large kernels used in our MSCLK framework can effectively avoid overfitting.

4.3. Ablation experiments of diverse architecture components in MSCLK

MSCLK mainly consists of five components: multi-scale convolution in the MKSC module, 3D fully separable convolution (FSC) with large kernels in the MKSC module, channel self-attention in the MKSC module, skip connections, contrastive Bayesian metric learning. To assess the contributions of different components in our MSCLK, we conducted the ablation experiments on sMRI images in the 5CV test by removing each component from MSCLK. The results of MSCLK ablation experiments are shown in Table 4. In Table 4, MSCLK_MS denotes that we remove the multi-scale convolution in MSCLK framework, and adopt the same kernel (i.e., $3 \times 3 \times 3$); MSCLK_LK denotes that we remove the large convolution kernels in MSCLK framework, and adopt $3 \times 3 \times 3$, $5 \times 5 \times 5$, $7 \times 7 \times 7$, $5 \times 5 \times 5$ kernels in FSC block-1, FSC block-2, FSC block-3, FSC block-4, respectively; MSCLK_CA denotes that we remove the channel attention in MSCLK framework; MSCLK_SC denotes that we remove the skip connection in MSCLK framework; MSCLK_BM denotes that we remove the contrastive Bayesian metric learning in MSCLK framework.

As shown in Table 4, we can see that for AD vs. EMCI classification task, the ACC of MSCLK is 11.11%, 7.11%, 3.06%, 3.33%, and 1.39% higher than MSCLK_MS, MSCLK_LK, MSCLK_CA, MSCLK_SC, and MSCLK_BM, respectively; for AD vs. LMCI classification task, the ACC of MSCLK is 10.28%, 4.36%, 1.56%, 1.86%, and 1.88% higher than MSCLK_MS, MSCLK_LK, MSCLK_CA, MSCLK_SC, and MSCLK_BM, respectively; for EMCI vs. NC classification task, the ACC of MSCLK is 9.45%, 5.56%, 1.95%, 3.61%, and 3.33% higher than MSCLK_MS, MSCLK_LK, MSCLK_CA, MSCLK_SC, and MSCLK_BM, respectively; for LMCI vs. NC classification task, the ACC of MSCLK is 11.18%, 7.77%, 4.97%, 4.04%, and 2.18% higher than MSCLK_MS, MSCLK_LK, MSCLK_CA, MSCLK_SC and MSCLK_BM, respectively; for EMCI vs. LMCI classification task, the ACC of MSCLK is 7.66%, 7.14%, 4.94%, 3.83%, and 4.40% higher than MSCLK_MS, MSCLK_LK, MSCLK_CA, MSCLK_SC and MSCLK_BM, respectively.

To demonstrate the effects of 3D convolution operation on parameters and training time, we compared the parameter and training time of two different convolutional kernels with/without 3D FSC, and present the Table S3 to show how much reduction in the parameters and training time is achieved by 3D convolutions for the EMCI vs. LMCI classification task. As shown in Table S3, we can see that the parameters of MSCLK with 3D FSC are 14,710,256 fewer than those without 3D FSC in the convolution kernel size [3,5,3,5] (for modules 1,4) and [3,5,3,5] (for modules 2,3), and the training time with 3D FSC is lower than that without 3D FSC for one epoch; the parameters of MSCLK with 3D FSC are 344,551,920 fewer than those without 3D FSC in the convolution kernel size [3,7,11,15] (for modules 1,4) and [5,9,13,17] (for modules 2,3). For the training time MSCLK of without 3D FSC under the condition of kernel size [3,7,11,15] (for modules 1,4) and [5,9,13,17] (for modules 2,3), due to the limited resources of our server, we are unable to train MSCLK model and provide the training time. From Table S3, we can see that 3D convolution operation can effectively reduce model parameters and training time.

These results show that multi-scale convolution, 3D fully separable convolution (FSC), channel self-attention, skip connections, and contrastive Bayesian metric learning used in our MSCLK are valid and contribute to improve the performance of MSCLK, especially, the contribution of multi-scale convolution and large convolution kernels is more than that of other three components in AD-related five classification tasks. In addition, we also find that the contribution of contrastive

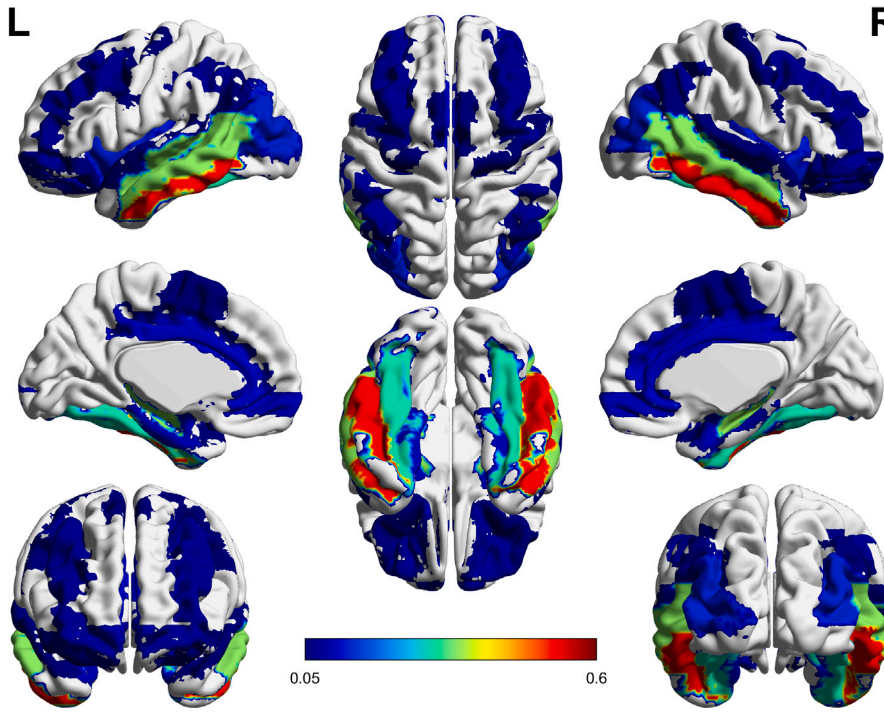

Fig. 3. Visualization of important brain regions identified by MSCLK using sMRI images.

Table 5
Top 10 significant brain regions identified by MSCLK and MSCLK_FMP.

	sMRI		PET		sMRI and PET	
	Brain region	$\Delta ACC\%$	Brian Region	$\Delta ACC\%$	Brain region	$\Delta ACC\%$
1	Temporal_Inf	52.4	Hippocampus	50.8	Hippocampus	74.3
2	Hippocampus	34.1	Insula	49.9	Precentral	40.8
3	Middle temporal gyrus	33.8	Precentral	34.7	Precuneus	30.8
4	Fusiform gyrus	29.0	Caudate	31.7	Thalamus	28.2
5	Occipital_Inf	13.4	Precuneus	29.1	Caudate	27.9
6	Insula	13.2	Heschl	26.5	Supp_Motor_Area	25.5
7	Cingulum_Ant	11.4	Calcarine	26.3	Pallidum	25.5
8	Cingulum_Mid	11.0	Cuneus	24.0	Insula	25.3
9	Frontal_Mid_Orb	10.3	Frontal_Inf_Tri	23.9	Frontal_Mid	23.0
10	ParaHippocampal	10.3	Occipital_Inf	18.6	Temporal_Mid	22.8

Bayesian metric learning in EMCI vs LMCI and EMCI vs. NC tasks is more than in other AD-related three tasks, indicating that contrastive Bayesian metric learning plays an important role in the tasks with high sample similarity but belonging to different classes.

4.4. Discriminative brain regions for EMCI and LMCI

In this part, we used the Automated Anatomical Labeling (AAL) template to divide the entire brain area of subjects into 45 regions, and adopted the occlusion analysis tool (Kwak et al., 2021, 2022) to identify the relative contribution of each brain region to the diagnosis of EMCI/LMCI by masking this brain region. To perform the occlusion analysis for each subject brain region, we masked one brain region on the input images by setting all voxel values to zero in this brain region, and then quantified the contribution of the masked brain region to the performance of the model using $\Delta ACC_q = (ACC - ACC_{\mathrm{q}})/ACC$ ($q = 1, 2, \ldots, 45$), where ACC represents the classification accuracy on the unmasked images, ACC_i represents the classification accuracy on the images masked their ith region. By comparing the model's classification accuracy before and after occluding one region, we are able to quantify the contribution of each brain region to the model's classification accuracy. If occlusion of a brain region leads to a significant decrease in classification accuracy, we consider that this brain region has an important contribution to improving the accuracy of the model for EMCI vs. LMCI classification. In the case of using sMRI images, PET images, and jointly using sMRI and PET images, we visualize the contributions (Figs. 3–5) of 45 brain regions to the classification accuracy and presented the top 10 brain regions (Table 5) with the highest contributions.

From Fig. 3 and Table 5, we can find that brain regions such as the inferior temporal gyrus, hippocampus, and middle temporal gyrus play a crucial role in diagnosing EMCI and LMCI in the case of using sMRI alone. Specifically, the hippocampus has been reported to be closely related to long-term memory (Wolf et al., 2003). The inferior temporal gyrus and middle temporal gyrus are important biomarkers, which are brain regions clearly associated with memory and may be the first affected temporal lobe region in the development of Alzheimer's disease (Convit et al., 2000; Rasero et al., 2017; Xu et al., 2016).

From Fig. 4 and Table 5, we can find that the brain regions of the hippocampus, insula, and precentral gyrus play a crucial role in diagnosing EMCI and LMCI in the case of using PET images. He et al. (2015) conducted a meta-analysis of PET images and observed a significant decrease in glucose metabolism levels in the precentral gyrus among subjects with Alzheimer's disease (AD) and mild cognitive impairment (MCI), suggesting that the precentral gyrus is an important biomarker. The insular is responsible for processing a range of higher-level brain functions, including emotions, self-awareness, and cognitive functions. By utilizing PET images, we can observe the activation level

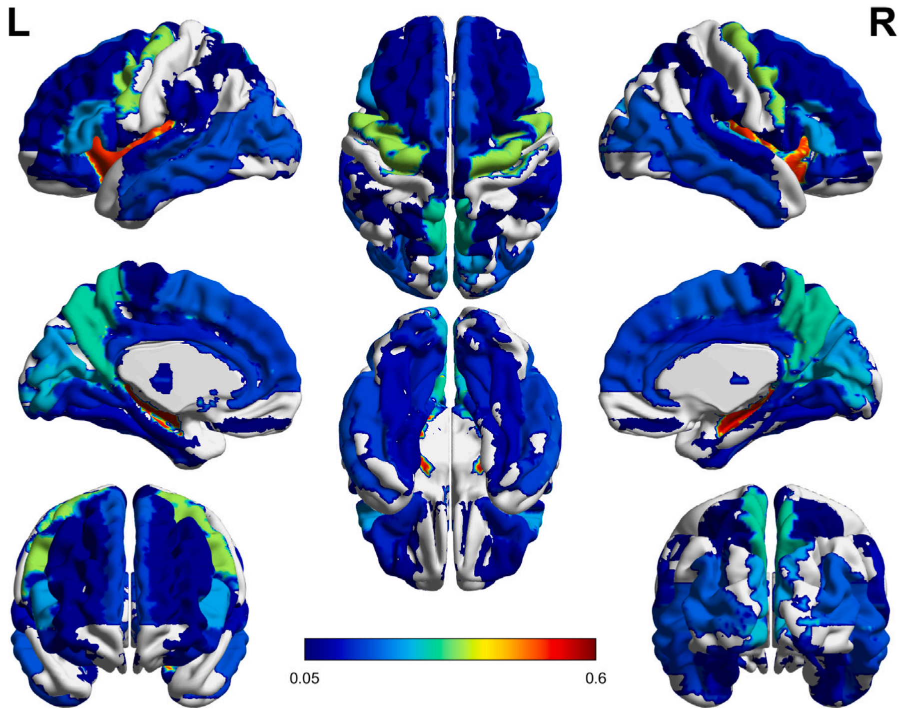

Fig. 4. Visualization of important brain regions identified by MSCLK using PET images.

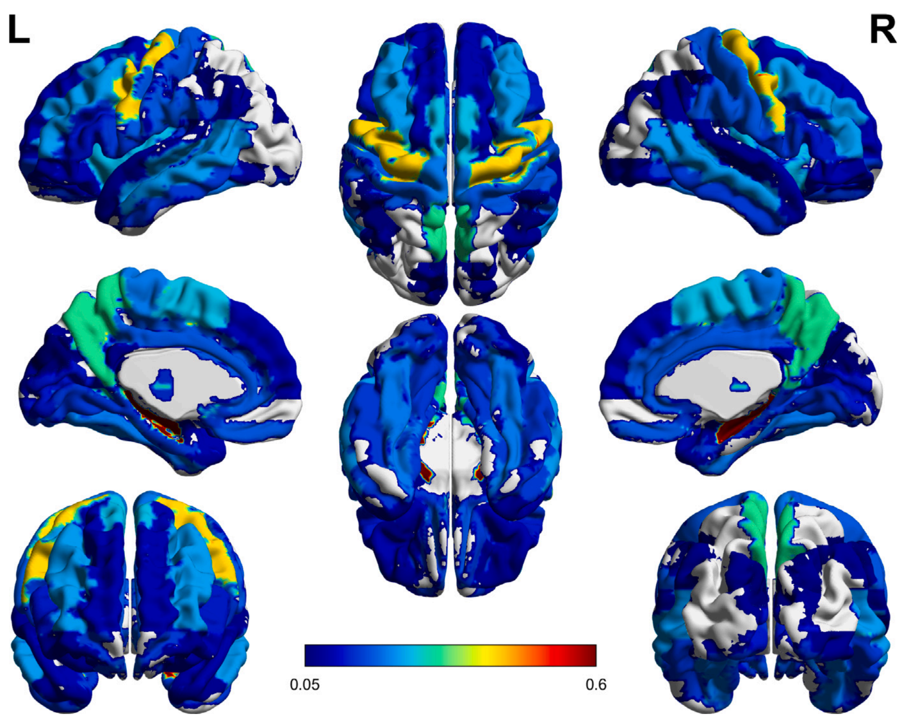

Fig. 5. Visualization of important brain regions identified by MSCLK-FMP using sMRI and PET images.

and functional connectivity of the insular during different brain tasks, thus playing a key role in the diagnosis of EMCI/LMCI (Xie et al., 2012).

From Fig. 5 and Table 5, we can observe that the brain regions of the hippocampus, precentral gyrus, precuneus (Nobili & Morbelli, 2010), and thalamus play a crucial role for diagnosing EMCI and LMCI in the case of using sMRI and PET images simultaneously. These findings are consistent with most studies on EMCI/LMCI. Especially, we find that the thalamus plays an important role in diagnosing EMCI/LMCI. As a part of the sensorimotor network, the thalamus is believed to be closely associated with cognition, information processing speed, emotion, and gait speed (Rosano et al., 2012). Some studies have revealed a strong correlation between the thalamus and subtypes of Mild Cognitive Impairment (Cai et al., 2015; Jitsuishi & Yamaguchi, 2022). These findings indicate that the use of multimodal brain data can help us identify important brain regions that cannot be detected with single modality data. As observed in Table 5, the brain regions ranked higher using multimodal data are predominantly those ranked higher using PET images, indicating that PET images are more effective than other imaging modes in classifying EMCI/LMCI (Cui et al., 2022; Hao et al., 2020; Pan et al., 2020; Rubinski et al., 2020).

Overall, we have identified the important brain regions related to EMCI and LMCI, such as the hippocampus, inferior temporal gyrus, precentral gyrus, precuneus lobe, and thalamus. sMRI provides the information on tissue structure, delineating the shape and positional relationship of different brain regions, organs, or tissues. PET images provide the information on biological activity and function, reflecting glucose metabolism activity in various areas. In addition, the same brain region (e.g., the precentral) has different contributions to the diagnosis of AD in the case of using different imaging modalities. Thus, it is essential to integrate different modality images for diagnosing EMCI and LMCI.

5. Conclusions

Early AD diagnosis is helpful for timely intervention, and the delay of cognitive impairment progression, as well as for early prevention and treatment of AD. In this work, we proposed a novel Alzheimer's disease early diagnosis method (called MSCLK) by designing the multi-scale 3D fully separable convolution neural network with small/large kernels, and introducing the contrastive Bayesian metric learning. The multi-scale 3D fully separable convolution neural network with small/large kernels is used to effectively capture the discrimination features of different scale acceptance domains, and reduce parameters and overfitting of MSCLK. The contrastive Bayesian metric learning is used to learn the hard samples that have high similarity but belong to different classes. Considering that PET images contain the information of brain glucose metabolism and the molecular level lesions, which are helpful for the diagnosis of MCI patients, we also presented the variant methods of MSCLK (called MSCLK-Fusion MRI and PET, MSCLK-FMP) to improve the classification accuracy of EMCI vs. LMCI task by fusing sMRI and PET images at pixel-level and feature-level. Experimental results on the ADNI database show that the performance of our MSCLK is superior to other state-of-the-art methods and all the components introduced in MSCLK can effectively enhance the classification accuracy of early diagnosis of Alzheimer's disease; our MSCLK-FMP achieves the best performance in EMCI vs. LMCI classification task. By analyzing the brain regions related to MCI identified by MSCLK and MSCLK-FMP, we find that the brain regions of Temporal_Inf, Hippocampus, Precuneus, Precentral, and Thalamus play key roles in EMCI vs. LMCI classification task. In the future, we will further investigate the role of these key brain regions and explore their specific mechanisms in disease development.

CRediT authorship contribution statement

Run-Feng Tian: Conceptualization, Methodology, Data collection and processing, Software, Writing – original draft. **Jia-Ni Li:** Methodology, Validation, Visualization. **Shao-Wu Zhang:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

The experimental data used in this work was downloaded from Alzheimer's disease neuroimaging initiative (ADNI). This work was supported by National Natural Science Foundation of China under grants (62173271, 61873202).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2024.124241.

References

Alzheimer's Association (2023). 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *19*(4), 1598–1695.

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical image analysis using convolutional neural networks: a review. *Journal of Medical Systems*, *42*, 1–13.

Ashburner, J., Barnes, G., Chen, C. C., Daunizeau, J., Flandin, G., Friston, K., & Penny, W. (2014). *SPM12 manual. Wellcome trust centre for neuroimaging*. London, UK: 2464(4).

Bakkouri, I., & Afdel, K. (2019). Multi-scale CNN based on region proposals for efficient breast abnormality recognition. *Multimedia Tools and Applications*, *78*, 12939–12960.

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., & Lladó, X. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, *95*, 64–81.

Bi, X. A., Hu, X., Wu, H., & Wang, Y. (2020). Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2973–2983.

Cai, S., Huang, L., Zou, J., Jing, L., Zhai, B., Ji, G., & Alzheimer's Disease Neuroimaging Initiative (2015). Changes in thalamic connectivity in the early and late stages of amnestic mild cognitive impairment: a resting-state functional magnetic resonance study from ADNI. *PLoS One*, *10*(2), Article e0115573.

Convit, A., De Asis, J., De Leon, M. J., Tarshish, C. Y., De Santi, S., & Rusinek, H. (2000). Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiology of Aging*, *21*(1), 19–26.

Cui, W., Yan, C., Yan, Z., Peng, Y., Leng, Y., Liu, C., Chen, S., Jiang, X., Zheng, J., & Yang, X. (2022). BMNet: A new region-based metric learning method for early Alzheimer's Disease identification with FDG-PET images. *Frontiers in Neuroscience*, *16*, Article 831533.

Ding, X., Zhang, X., Han, J., & Ding, G. (2022). Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (* (pp. 11963–11975).

Fan, Z., Li, J., Zhang, L., Zhu, G., Li, P., Lu, X., & Wei, W. (2021). U-net based analysis of MRI for Alzheimer's disease diagnosis. *Neural Computing and Applications*, *33*, 13587–13599.

Folego, G., Weiler, M., Casseb, R. F., Pires, R., & Rocha, A. (2020). Alzheimer's disease detection through whole-brain 3D-CNN MRI. *Frontiers in Bioengineering and Biotechnology*, *8*, Article 534592.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., & Zhang, S. (2020). CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, *40*(2), 699–711.

Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S. L., Andrew, J. S., Yao, X., Shen, L., & Alzheimer's Disease Neuroimaging Initiative (2020). Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease. *Medical Image Analysis*, *60*, Article 101625.

He, W., Liu, D., Radua, J., Li, G., Han, B., & Sun, Z. (2015). Meta-analytic comparison between PIB-PET and FDG-PET results in alzheimer's disease and MCI. *Cell Biochemistry and Biophysics*, *71*, 17–26.

Helaly, H. A., Badawy, M., & Haikal, A. Y. (2022). Deep learning approach for early detection of Alzheimer's disease. *Cognitive Computation*, *14*(5), 1711–1727.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (* (pp. 7132–7141).

Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., & Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27*(4), 685–691.

Jitsuishi, T., & Yamaguchi, A. (2022). Searching for optimal machine learning model to classify mild cognitive impairment (MCI) subtypes using multimodal MRI data. *Scientific Reports*, *12*(1), 4284.

Jo, T., Nho, K., Risacher, S. L., Saykin, A. J., & Alzheimer's Neuroimaging Initiative (2020). Deep learning detection of informative features in tau PET for Alzheimer's disease classification. *BMC Bioinformatics*, *21*, 1–13.

Kan, S., He, Z., Cen, Y., Li, Y., Mladenovic, V., & He, Z. (2022). Contrastive bayesian analysis for deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7220–7238.

Kang, L., Jiang, J., Huang, J., & Zhang, T. (2020). Identifying early mild cognitive impairment by multi-modality MRI-based deep learning. *Frontiers in Aging Neuroscience*, *12*(206).

Kang, W., Lin, L., Zhang, B., Shen, X., Wu, S., & Alzheimer's Disease Neuroimaging Initiative (2021). Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis. *Computers in Biology and Medicine*, *136*, Article 104678.

Khairandish, M. O., Sharma, M., Jain, V., Chatterjee, J. M., & Jhanjhi, N. Z. (2022). A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. *Irbm*, *43*(4), 290–299.

Kong, Z., Zhang, M., Zhu, W., Yi, Y., Wang, T., & Zhang, B. (2022). Multi-modal data Alzheimer's disease detection based on 3D convolution. *Biomedical Signal Processing and Control*, *75*, Article 103565.

Kwak, K., Giovanello, K. S., Bozoki, A., Styner, M., & Dayan, E. (2021). Subtyping of mild cognitive impairment using a deep learning model based on brain atrophy patterns. *Cell Reports Medicine*, *2*(12).

Kwak, K., Niethammer, M., Giovanello, K. S., Styner, M., Dayan, E., & Alzheimer's Disease Neuroimaging Initiative (2022). Differential role for hippocampal subfields in Alzheimer's disease progression revealed with deep learning. *Cerebral Cortex*, *32*(3), 467–478.

Liu, X., Chen, W., Hou, H., Chen, X., Zhang, J., Liu, J., Guo, Z., & Bai, G. (2017). Decreased functional connectivity between the dorsal anterior cingulate cortex and lingual gyrus in Alzheimer's disease patients with depression. *Behavioural Brain Research*, *326*, 132–138.

Liu, X., Hou, F., Qin, H., & Hao, A. (2018). Multi-view multi-scale CNNs for lung nodule type classification from CT images. *Pattern Recognition*, *77*, 262–275.

Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., & Alzheimer's Disease Neuroimaging Initiative (2020). A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage*, *208*, Article 116459.

Liu, Z., Lu, H., Pan, X., Xu, M., Lan, R., & Luo, X. (2022). Diagnosis of Alzheimer's disease via an attention-based multi-scale convolutional neural network. *Knowledge-Based Systems*, *238*, Article 107942.

Mehmood, A., Yang, S., Feng, Z., Wang, M., Ahmad, A. S., Khan, R., & Yaqub, M. (2021). A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience*, *460*, 43–52.

Nobili, F., & Morbelli, S. (2010). [18F] FDG-PET as a biomarker for early Alzheimers Disease. *The Open Nuclear Medicine Journal*, *2*(1).

Odusami, M., Maskeliūnas, R., Damaševičius, R., & Misra, S. (2023). Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images. *Journal of Medical and Biological Engineering*, *43*(3), 291–302.

Pan, D., Luo, G., Zeng, A., Zou, C., Liang, H., Wang, J., Zhang, T., Yang, B., & Alzheimer's Disease Neuroimaging Initiative (2022). Adaptive 3DCNN-based interpretable ensemble model for early diagnosis of Alzheimer's disease. *IEEE Transactions on Computational Social Systems*, *11*(1), 247–266.

Pan, X., Phan, T. L., Adel, M., Fossati, C., Gaidon, T., Wojak, J., & Guedj, E. (2020). Multi-view separable pyramid network for AD prediction at MCI stage by 18 F-FDG brain PET imaging. *IEEE Transactions on Medical Imaging*, *40*(1), 81–92.

Pei, Z., Wan, Z., Zhang, Y., Wang, M., Leng, C., & Yang, Y. H. (2022). Multi-scale attention-based pseudo-3D convolution neural network for Alzheimer's disease diagnosis using structural MRI. *Pattern Recognition*, *131*, Article 108825.

Qiang, Y. R., Zhang, S. W., Li, J. N., Li, Y., Zhou, Q. Y., & Alzheimer's Disease Neuroimaging Initiative (2023). Diagnosis of Alzheimer's disease by joining dual attention CNN and MLP based on structural MRIs, clinical and genetic data. *Artificial Intelligence in Medicine*, *145*, Article 102678.

Rasero, J., Alonso-Montes, C., Diez, I., Olabarrieta-Landa, L., Remaki, L., Escudero, I., & Alzheimer's Disease Neuroimaging Initiative (2017). Group-level progressive alterations in brain connectivity patterns revealed by diffusion-tensor brain networks across severity stages in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *9*(215).

Ravi, V., EA, G, & KP, S. (2024). Deep learning-based approach for multi-stage diagnosis of Alzheimer's disease. *Multimedia Tools and Applications*, *83*(6), 16799–16822.

Rosano, C., Bennett, D. A., Newman, A. B., Venkatraman, V., Yaffe, K., Harris, T., & Aizenstein, H. J. (2012). Patterns of focal gray matter atrophy are associated with bradykinesia and gait disturbances in older adults. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, *67*(9), 957–962.

Rubinski, A., Franzmeier, N., Neitzel, J., Ewers, M., & Alzheimer's Disease Neuroimaging Initiative (ADNI) (2020). FDG-PET hypermetabolism is associated with higher tau-PET in mild cognitive impairment at low amyloid-PET levels. *Alzheimer's Research & Therapy*, *12*, 1–12.

Sharma, R., Goel, T., Tanveer, M., Suganthan, P. N., Razzak, I., & Murugan, R. (2022). Conv-ervfl: Convolutional neural network based ensemble RVFL classifier for Alzheimer's disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, *27*(10), 4995–5003.

Wang, H., Shen, Y., Wang, S., Xiao, T., Deng, L., Wang, X., & Zhao, X. (2019). Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. *Neurocomputing*, *333*, 145–156.

Wolf, H., Jelic, V., Gertz, H. J., Nordberg, A., Julin, P., & Wahlund, L. O. (2003). A critical discussion of the role of neuroimaging in mild cognitive impairment. *Acta Neurol Scandinavica*, *107*, 52–76.

Xie, C., Bai, F., Yu, H., Shi, Y., Yuan, Y., Chen, G., & Li, S. J. (2012). Abnormal insula functional network is associated with episodic memory decline in amnestic mild cognitive impairment. *Neuroimage*, *63*(1), 320–327.

Xu, L., Wu, X., Li, R., Chen, K., Long, Z., Zhang, J., & Alzheimer's Disease Neuroimaging Initiative (2016). Prediction of progressive mild cognitive impairment by multi-modal neuroimaging biomarkers. *Journal of Alzheimer's Disease*, *51*(4), 1045–1056.

Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., & Diciotti, S. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, *11*(1), 22544.

Zhang, X., Han, L., Zhu, W., Sun, L., & Zhang, D. (2021). An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5289–5297.

Zhang, F., Pan, B., Shao, P., Liu, P., Shen, S., Yao, P., Xu, R. X., & Alzheimer's Disease Neuroimaging Initiative (2022). A single model deep learning approach for Alzheimer's disease diagnosis. *Neuroscience*, *491*, 200–214.

Zheng, G., Zhang, Y., Zhao, Z., Wang, Y., Liu, X., Shang, Y., Cong, Z., Dimitriadis, S., Yao, Z., & Hu, B. (2022). A transformer-based multi-features fusion model for prediction of conversion in mild cognitive impairment. *Methods*, *204*, 241–248.