Computers in Biology and Medicine 176 (2024) 108564

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

HAMMF: Hierarchical attention-based multi-task and multi-modal fusion model for computer-aided diagnosis of Alzheimer's disease

Xiao Liu [a], Weimin Li [a,*], Shang Miao [a], Fangyu Liu [b,c,d], Ke Han [e], Tsigabu T. Bezabih [a]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[b] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[c] University of Chinese Academy of Sciences, Beijing, China
[d] BGI-Shenzhen, Shenzhen, China
[e] Medical and Health Center, Liaocheng People's Hospital, LiaoCheng, China

ARTICLE INFO

Keywords:
Alzheimer's disease
Attention mechanism
Transformer
Multi-modal fusion
Deep learning
Multi-task learning

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative condition, and early intervention can help slow its progression. However, integrating multi-dimensional information and deep convolutional networks increases the model parameters, affecting diagnosis accuracy and efficiency and hindering clinical diagnostic model deployment. Multi-modal neuroimaging can offer more precise diagnostic results, while multi-task modeling of classification and regression tasks can enhance the performance and stability of AD diagnosis. This study proposes a Hierarchical Attention-based Multi-task Multi-modal Fusion model (HAMMF) that leverages multi-modal neuroimaging data to concurrently learn AD classification tasks, cognitive score regression, and age regression tasks using attention-based techniques. Firstly, we preprocess MRI and PET image data to obtain two modal data, each containing distinct information. Next, we incorporate a novel Contextual Hierarchical Attention Module (CHAM) to aggregate multi-modal features. This module employs channel and spatial attention to extract fine-grained pathological features from unimodal image data across various dimensions. Using these attention mechanisms, the Transformer can effectively capture correlated features of multi-modal inputs. Lastly, we adopt multi-task learning in our model to investigate the influence of different variables on diagnosis, with a primary classification task and a secondary regression task for optimal multi-task prediction performance. Our experiments utilized MRI and PET images from 720 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The results show that our proposed model achieves an overall accuracy of 93.15% for AD/NC recognition, and the visualization results demonstrate its strong pathological feature recognition performance.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that progressively causes cognitive impairment in patients due to neuronal damage in the brain [1]. AD typically involves changes in brain structure such as cortical atrophy, enlargement of the ventricular area, and a reduction in hippocampal volume [2]. Fig. 1 shows brain images of healthy people and those with AD. According to research by World Health Organization (WHO), more than 55 million people are in the early stages of Alzheimer's disease as of 2021, and Alzheimer's disease related brain degeneration is among the top 10 causes of death worldwide. The global cost of caring for AD patients is $2.8 trillion, placing a huge financial burden on society [3]. Clinical experts say that early diagnosis of patients is extremely important to mitigate the severity of dementia. Early treatment can significantly delay AD progression and enhance patients' quality of life. A range of brain imaging techniques using Magnetic resonance imaging (MRI) [4,5] and positron emission computed tomography (PET) provides a noninvasive and effective examination to help diagnose and understand the anatomical and functional changes associated with AD [6], with promising results in the diagnosis of AD/NC.

Numerous significant studies for building Alzheimer's disease diagnostic models rely heavily on 3D convolution. While 2D convolution can only capture features along the length and width of an image, 3D convolution can extract features across all three dimensions of image depth, width, and length, representing the spatial dependence of the 3D image. This makes 3D convolution more effective at classifying AD development stages. However, the high dimensionality and high resolution of 3D neuroimaging data mean that many disease features

* Corresponding author.
E-mail address: wmli@shu.edu.cn (W. Li).

https://doi.org/10.1016/j.compbiomed.2024.108564
Received 9 January 2024; Received in revised form 15 April 2024; Accepted 5 May 2024
Available online 8 May 2024

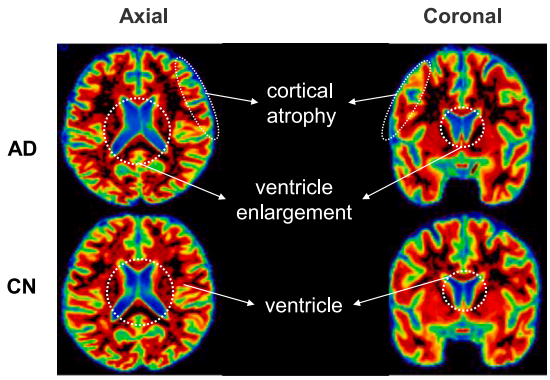


Fig. 1. Comparison of cross-sectional and coronal brain structural changes in two directions between the late Alzheimer's disease brain (top) and normal control brain (bottom).

are hidden within the data, requiring researchers to build deeper models with more parameters to train the data properly and extract the underlying information. Nevertheless, very deep 3D convolutional neural networks can be computationally expensive and consume significant memory. Although residual networks like ResNet [7] and dense blocks like DenseNet [8] can build very deep networks to learn 3D data features, these networks are ineffective at classifying complex brain neuroimaging data, and training them remains time-consuming. On the other hand, multi-modal fusion, with early, late, and hybrid fusion approaches [9], is one frontier area of multi-modal deep learning. Different imaging techniques provide information about different aspects of brain structure and function. Deep learning models are able to fuse these multi-modal data to extract more comprehensive features that can improve the diagnostic accuracy of AD. Fusing multi-modal data has two main benefits: First, the model can make more robust predictions by fusing data from multiple modalities of a given pathological phenomenon. Second, the model can extract complementary information from multiple modalities to improve diagnostic accuracy. However, traditional multi-modal fusion algorithms ignore the correlations between the multi-modal images.

This study takes a three-step approach. The initial stage of this study involves extracting primary (shallow) feature representations from two multi-modal Alzheimer's disease image datasets by pre-training the ResNet network. Second, a novel contextual hierarchical attention module (CHAM) is proposed to fuse the feature representations from multiple modalities. The CHAM module is designed to be robust for disease diagnosis by capturing attention from different modalities. Unlike SENET [10], CHAM addresses the lack of correlation between modalities by focusing on regions of interest using channel and spatial features. Finally, the study employs visualization techniques to assign weights to the model's gradient parameters at each training stage, demonstrating the degree of attention the model pays to different brain regions. The main contributions of this study are outlined below:

- The study proposes a Hierarchical Attention-based Multi-task Multi-modal Fusion (HAMMF) model for the computer-aided diagnosis of Alzheimer's disease. The CHAM module of the model accurately extracts feature data and multi-modal association data from multi-modal 3D neuroimages without significantly increasing the network's redundancy. The module's channel and spatial attention capture the target object's attention weights, which can effectively focus on the focal areas of 3D neuroimages. The module uses a Transformer to capture multi-modal contextual attention weights and automatically learn the correlation between multi-modal hierarchies, which enables the CHAM module to capture the exact location of the lesion.
- The study effectively constructs a multi-task learning framework that uses subjects' demographic characteristics, compensating for the inductive bias that a single task lacks by incorporating losses from different tasks. The classification task serves as the main task to provide guidance, while clinically relevant Alzheimer's disease score indicators and age serve as regression tasks, providing additional evidence for diagnosis. Multi-task joint learning effectively assists disease diagnosis and improves accuracy.
- The model exhibits good generalization and robustness, achieving a detection accuracy of 93.15% on the Alzheimer's Disease Neuroimaging Initiative ADNI [11] dataset. The paper also visualizes MRI images using interpretability methods, with the highlighted activation regions representing the areas of focus of the network, which are also abnormal regions during AD development. Comparing the visualization results of heat maps of different networks verifies the effectiveness of the proposed method.

The remaining part of the paper is organized as follows: An overview of related work on Alzheimer's disease diagnosis research is provided in Section 2. Section 3 presents the dataset used in this study, while Section 4 introduces the proposed method. In Section 5, the model proposed in this paper is evaluated and compared to recent Alzheimer's disease diagnostic models. Section 6 presents the conclusions drawn from the study, while Section 7 outlines some limitations of the current work and proposes directions for future research.

2. Related work

This study section overviews related work in multi-modal diagnostic models for Alzheimer's disease. It begins by reviewing studies on models for diagnosing Alzheimer's disease using multiple modalities (multi-modal), followed by a review of the multi-task learning approach. The section then concludes with a discussion of attention mechanisms and how they have been applied in the field of medical imaging analysis.

2.1. Multi-modal Alzheimer's disease diagnostic model

Convolutional neural networks (CNNs) have been instrumental in multi-modal Alzheimer's disease diagnosis. Previous studies on brain disease diagnosis can be classified into four categories: 2D slice-level, 3D patch-level, ROI-based, and 3D subject-level, depending on the input type of the network. The 2D slice-level approach involves extracting 2D slices from 3D MRI images and inputting them into a 2D CNN to learn relevant features [12,13]. However, this model cannot express the relationships between slices and is impractical due to input clipping. The 3D patch-level method overcomes the limitations of the 2D slice-level method by obtaining partial patches [14–16], but lacks the correlation between different patches. In the ROI-based method [17], ROI templates designed by clinical experts are used to determine brain function or structural blocks. However, this approach also lacks complete brain region correlation information. In deep learning approaches using whole 3D MRI images as input, more research has begun utilizing entire 3D information to discover pathology with stronger correlations and higher accuracy. Zhao et al. [18] obtained three MRI images of different sizes as the input of the model by setting transposed convolutions of different sizes. D2BOF-COVIDNet [19] presented an advanced framework that integrates deep Bayesian optimization and feature fusion techniques to enhance the classification of COVID-19 using chest X-ray and MRI images. Odusami et al. [20] presented an early fusion framework and a modified Resnet18 deep learning architecture for Alzheimer's disease diagnosis, leveraging both MRI and PET scans to improve accuracy, achieving a 73.90% classification rate on the ADNI database and incorporating an Explainable AI model for result interpretation. Odusami et al. [21] proposed a novel multi-modal neuroimaging fusion method for AD diagnosis, combining advanced convolutional

techniques to achieve high classification accuracy with a Mobile Vision Transformer on multiple datasets. In contrast, Ge et al. [22] used Cat12 to extract gray matter, white matter, and cerebrospinal fluid in MRI as input. With the advancement of high-performance computing hardware, more research has focused on end-to-end training of the entire MRI brain to fully utilize overall information for discovering pathological regions with a stronger correlation and higher identification. However, accurately locating the local disease area in the 3D neuroimaging data of Alzheimer's disease remains a significant challenge due to the large number of model parameters and the limited amount of subject data available.

2.2. Multi-task learning

Multi-task learning is a technique that can effectively address the issue of low training accuracy of a model under a single task. It reduces the number of model parameters and improves generalization by designing the neural network architecture to share portions of parameters for multiple tasks. In disease classification, the clinical score is an important evaluation index that reflects disease severity [23,24]. Using it as one of the model's predictive tasks can better represent the underlying data relationship. There are two multi-task learning scenarios in deep learning: hard parameter sharing and soft parameter sharing. In the hard parameter sharing approach, the first few convolutional layers of the CNN are shared from the bottom to the top. This allows the model to learn abstract features common to each task, reducing the number of parameters to be learned for each task and improving generalization. In the soft parameter sharing approach, each task owns its parameters independently, and a small number of weighted parameters are shared among tasks.

2.3. Attention mechanism

Attention mechanisms assign importance weights to different parts of features based on their contributions to the overall performance. Researchers have proposed attention mechanisms that weigh the importance of different features in various parts of the feature graph to improve overall classification performance. In natural language tasks, the attention mechanism focuses on weighting the importance of information at different locations in a sentence, allowing the model to concentrate on relevant features [25,26]. For example, attention has been used to extract sentiment features [27]. In computer vision, attention mechanisms capture task-relevant attention features from images to improve performance accuracy. Attention mechanisms have been widely used in medical image analysis. Researchers have proposed attention modules for tasks like classifying or segmenting lesion regions in medical image analysis. For example, Masood et al. [28] introduced a CenterNet-based framework using a ResNet34 model with an attention block for precise brain tumor localization and classification. Ramya et al. [29] introduced a novel AD classification method using MRI data that combines various image processing techniques, including 2D-ABF, ECLAHE, EEM, AH, GLCM, PCA, and Logistic Regression, achieving a high accuracy rate of 96.92%. Odusami et al. [30] investigated the integration of MRI and PET images for AD diagnosis using Pareto optimized deep learning models (VGG11, VGG16, VGG19), with VGG19 showing superior performance on the ADNI dataset based on various image quality metrics. Qin et al. [31] proposed a 3D HA-ResUNet network that uses spatial and channel attention for early dementia diagnosis. Xie et al. [32] designed a cross-attention model to identify high-risk regions to eliminate noise in chest disease diagnosis. Banerjee et al. [33] developed a new patch attention module to learn e most discriminative patches of the fingerprint image for fingerprint spoofing detection. Transformer models use self-attention and multi-headed attention [34]. They have been applied to vision tasks like image classification (ViT [35] and DeiT [36]) and object detection (DETR) [37]. While traditional image attention and Transformer-based approaches have performed well separately, few studies have combined the two for Alzheimer's disease diagnosis.

Compared to models integrated with CHAM, traditional CNNs might lack the flexibility to adaptively adjust the network's focus, potentially falling short in identifying complex patterns associated with AD. Furthermore, pure Transformer models might be less effective at handling details compared to models that combine CNNs due to a lack of focus on local features. Single-task learning might require training multiple independent models when handling multiple related tasks, leading to inefficient resource utilization and an inability to capitalize on potential correlations between tasks. Integrated models with CHAM and Transformers balance local and global feature extraction better than traditional CNNs and standalone Transformer models, providing more powerful and flexible feature representation.

The key to solving the problem of missing modality correlation in Alzheimer's disease image data fusion is to combine the pixel features of different modality images into an overall hierarchical feature representation across modalities. Different modality images contain different amounts of information, so it is not reasonable to combine multiple characteristics equally. The self-attention mechanism can automatically determine the weights of each modality. With this in mind, this study proposes a hierarchical attention-based multi-task multi-modal fusion (HAMMF) model. The model uses an attention mechanism to efficiently fuse Gray Matter (GM), White Matter (WM), Cerebrospinal Fluid (CSF) imaging modalities, and PET imaging modalities, in addition to using multi-task learning to jointly guide computer-aided diagnosis of Alzheimer's disease.

3. Materials and equipment

3.1. Datasets and data preprocessing

The dataset utilized in this paper was obtained from the Alzheimer's Disease Neuroimaging Initiative [11] (ADNI) dataset, which is accessible at http://adni.loni.usc.edu. The ADNI dataset contains data from four different periods: ADNI-1, ADNI-GO, ADNI-2 and ADNI-3, each with a different number of subjects' brain examination data. There are 1193 images of 1.5T/3T T1-weighted structural MRI (sMRI) and PET scans, as well as patient age, gender, Mini-Mental State Evaluation Scale (MMSE) scores, and other basic information. According to clinical standards, subjects were classified into AD, Mild Cognitive Impairment(MCI) and normal control(NC) groups. The image data for each subject in the dataset is a grayscale image with one channel and image size $117 \times 130 \times 110$.

In neuroscience and medical research, selecting an equal number of participants with AD, NC, and MCI primarily aims to ensure a balanced study design, enhance the statistical power of analyses, and reduce variability caused by uneven sample sizes. Such balanced designs facilitate more equitable and comparable group comparisons by simplifying the data analysis process. Moreover, maintaining equal sample sizes also reflects a commitment to research ethics, ensuring that all participants have equal opportunities regarding potential benefits or risks. Additionally, this approach enhances the representativeness of each study group, making the research findings more generalizable and credible. While practical considerations such as participant availability and research budgets might impose constraints, researchers generally strive to achieve sample size balance to ensure the robustness of their study conclusions. Our study selected an equal number of participants across the Alzheimer's Disease AD, MCI, and NC groups, each with both MRI and PET images. We chose the smallest common number among the three groups, 240, as the sample size for each group to ensure a balanced comparative analysis. This approach standardizes the number of subjects for statistical consistency and maximizes the use of available data from the imaging modalities for all groups involved. This study, 720 subjects were selected from the ADNI dataset, including 240 NC subjects, 240 MCI subjects, and 240 AD subjects. Each subject had both

Table 1
Demographic information of subjects in the study dataset. Including group type, number of groups, gender, age, and MMSE score.

Group	NC	MCI	AD
Sample size	240	240	240
Gender (Male/Female)	104/136	132/107	128/112
Age (Mean ± Std)	75.72 ± 6.90	72.31 ± 7.14	77.65 ± 7.82
MMSE (Mean ± Std)	29.12 ± 1.02	26.78 ± 1.23	23.34 ± 1.10

Fig. 2. Cross-sectional, coronal, and sagittal plane images of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) obtained by segmenting MRI with the CAT12 tool.

sMRI and PET image information, and demographic information for the three disease stages is presented in Table 1. The groups described in Table 1 represent three different stages(NC, MCI, AD).

To achieve better feature learning and classification in subsequent steps, the raw structural MRI and PET data obtained from ADNI must undergo pre-processing. The raw 3D NIIfTI format images are initially normalized using three non-uniformity intensity corrections of CAT12 [38]. These corrections include 3D gradient distortion, gradient nonlinear geometric correction, and B1 non-uniformity correction to eliminate noise signals generated by the electromagnetic field in the original images during scanning. Secondly, linear registration is performed using AAL templates [39] to remove global linear differences, such as global translation, scaling, and rotation differences, on all structural MR images. This registration process aligns the brain scan image to the middle position of the entire image. Once the standard template is registered, the data size of MRI and PET images is standardized to $1 \times 113 \times 137 \times 113$ pixels. In the third step, CAT12 removes the cerebellum, peels off the skull, and eliminates irrelevant image noise from the results. In the fourth step, the SPM software package's CAT12 is used to segment the MRI image into three different tissue type: gray matter, white matter, and cerebrospinal fluid. The three different tissues after segmentation are shown in Fig. 2. By combining the pre-processed PET image modality, two modality data types that can reflect brain disorders are obtained. The above pre-processing steps help the model better learn and distinguish the main features of AD.

To utilize MMSE scores as a regression task, the text must incorporate data cleaning methods to restore missing MSSE scores. This process aims to reduce the regression error of the model. MMSE serves as a rubric for early AD stages, and clinicians use scores obtained from the short scale to classify the disease stages.

The data corresponding to different stages has various interval scores, and the mean filling method was used to fill in the missing values. The average MMSE scores of all AD patients were calculated and used to fill in the missing MMSE scores of those AD patients. Subjects' ages with missing values were also imputed in the same way, by the average age of subjects with the same AD stage.

3.2. Equipment

The following hardware and software were used to process all experiment data.

Hardware: The CPU was an Intel(R) Xeon(R) Silver 4210 processor running at 2.2 GHz with 32 GB of memory. The graphics card was an Nvidia GeForce RTX 2080Ti with 11 GB of Video memory and the operating system was Ubuntu 18.04.

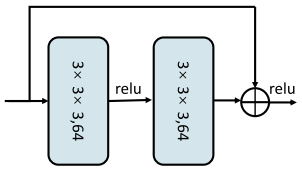


Fig. 3. ResNet architecture.

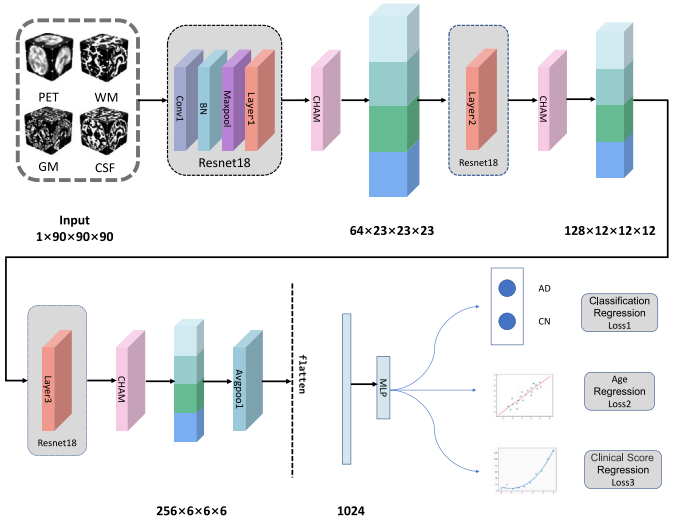


Fig. 4. Architecture of hierarchical attention-based multi-task multi-modal fusion (HAMMF) model.

Software: The software consisted of PyCharm as the IDE, Python 3.8.0 for programming, the PyTorch 1.8 library for deep learning, and the Nibabel and Sklearn libraries. MRI and PET data preprocessing was performed using CAT12 and SPM12 of MATLAB R2017a.

4. Methods

Alzheimer's disease can cause cognitive impairment in the elderly, leading to dementia or even death, often accompanied by varying degrees of brain atrophy and lesions as physical function declines gradually. Deep learning techniques can be utilized to identify the stages of Alzheimer's disease early and accurately, enabling appropriate disease prevention measures. This paper proposes a deep learning model to identify disease types effectively based on hierarchical attention-based multi-task multi-modal fusion. The subsequent sections provide a detailed explanation of the proposed model's specific components.

4.1. Proposed deep learning model

He et al. [7] proposed that the ResNet network architecture using the residual structure allows the model to deepen without encountering gradient degradation problems. The constant mapping of the residual structure (as illustrated in Fig. 3) ensures that the entire network converges in one direction, facilitating quicker learning of image features.

This paper proposes an attention fusion model (shown in Fig. 4) to extract and fuse features from MRI and PET images. The model, called Hierarchical Attention-based Multi-modal Fusion (HAMMF), has four inputs: PET, gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) images of Alzheimer's patients.

Pretrained ResNet18 is used to extract shallow features for each modality. The proposed residual block consists of three stages, each corresponding to a filter number of 64, 128, and 256, respectively. Each

residual block is followed by a CHAM block that extracts the feature attention maps for each branch. The multi-modal feature maps are then reweighted and fused to assign comprehensive attention weights to each modality.

The HAMMF model is primarily used for the classification task of AD/NC. The model also has two regression tasks as auxiliary tasks: predicting subjects' age and Mini Mental State Examination (MMSE) scores. These auxiliary tasks help improve the main classification task.

The model learns a nonlinear representation of the fused features through a multilayer perceptron (MLP). A softmax activation function acts on the binary classification task as follows:

$$y = \mathrm{softmax}\left(\mathrm{MLP}\left(C^0, C^1\right)\right) \tag{1}$$

The predicted disease classification vector is represented by y, while C^0 and C^1 represent the AD and NC classifications, respectively. $\mathrm{MLP}\left(C^0, C^1\right)$ represents the multi-layer perceptron operating on the AD and NC classifications.

MMSE scores and age are continuous data, with the former indicating the severity of AD development, which typically worsens with age. To extract the MMSE score and age information, a supervised regression task was utilized. A multi-layer perceptron (MLP) with ReLU activation is used to predict subjects' age represented by S_{Age} and MMSE scores represented by $S_{ClinicalScore}$ based on their AD and NC classifications:

$$S_{Age} = \mathrm{ReLu}\left(\mathrm{MLP}\left(C^0, C^1\right)\right) \tag{2}$$

$$S_{ClinicalScore} = \mathrm{ReLu}\left(\mathrm{MLP}\left(C^0, C^1\right)\right) \tag{3}$$

4.2. Contextual hierarchical attention module

Due to the slow and insidious progression of Alzheimer's disease, MRI and PET images of subjects show small interclass differences in disease classification. This can cause confusion when using ResNet networks to identify different subject image samples. While transformers can extract fine-grained AD image features by relying on long-range dependencies, using transformers directly at the 3D pixel level would result in excessive model parameters, which is not conducive to training. To address these issues, we propose the Contextual Hierarchical Attention Module (CHAM), which aims to suppress redundant information in input images, reduce the parameters of the Transformer in the model, and improve the recognition accuracy of the model. CHAM accurately locates lesions in images, strengthens the association between multiple modalities, reduces embedded transformer parameters, improves model accuracy, and is a hierarchical multi-modal fusion attention module that can be well integrated with ResNet networks.

4.2.1. Improving channel attention

The original channel attention module (as illustrated in Fig. 5) receives an input feature map with four dimensions information: height (H), width (W), depth (D), and channels (C). To create two distinct feature dimensions, the input feature map $\mathbf{F}$ of shape $H \times W \times D \times C$ undergoes average pooling ($\mathbf{FC}_{avg}$) and maximum pooling ($\mathbf{FC}_{max}$). The output is then fed to the MLP (multi-layer perceptron) layer.

To enhance the learning of MRI pathology feature information attention assignment and improve the interaction between different modalities, the original MLP layer was replaced with a Transformer module, as shown in Fig. 6. Let the input modal image data be $N_{in} = \{N_1, N_2, \ldots, N_i\}$ where i is the number of input modalities, and the dimension is $D_i \in \mathbb{R}^{1\times1\times1\times C}$ after average pooling and maximum pooling. Each modality undergoes a Transformer operation that relies on a linear projection of each feature in D_i using a scaled dot product to compute a set of attention-related vectors: the query matrix $\mathbf{S}_Q$, the key matrix $\mathbf{S}_K$ and the value matrix $\mathbf{S}_V$ as given below.

$$\mathbf{S}_Q = D_i \mathbf{W}_q \tag{4}$$

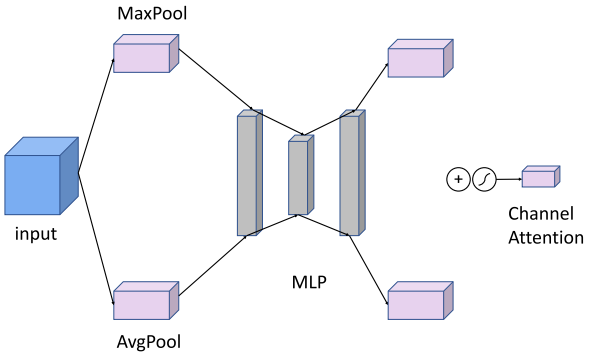


Fig. 5. Original channel attention architecture.

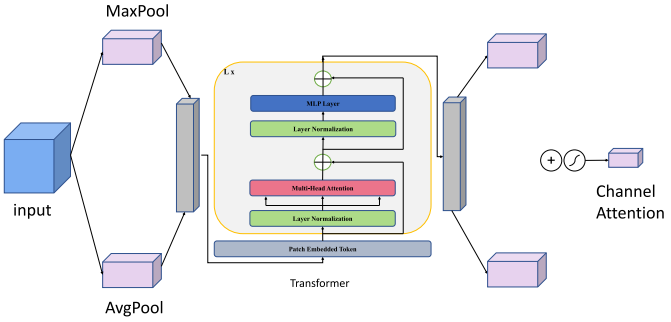


Fig. 6. Improved channel attention architecture.

$$\mathbf{S}_K = D_i \mathbf{W}_k \tag{5}$$

$$\mathbf{S}_V = D_i \mathbf{W}_v \tag{6}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are learnable hyperparameter matrices whose dimension size is denoted as $\mathbf{W}_q \in \mathbb{R}^{D_s * D_q}$, $\mathbf{W}_k \in \mathbb{R}^{D_s * D_k}$, $\mathbf{W}_v \in \mathbb{R}^{D_s * D_v}$ respectively. The query matrix assigned to each value in the sequence and the corresponding key matrix are utilized to obtain the attention weights by dot product operation, and the associated attentions of different sequence values are denoted as follows:

$$\mathrm{Att} = \mathrm{softmax}\left(\frac{\mathbf{S}_Q \left(\mathbf{S}_K\right)^T}{\sqrt{D_k}}\right)\mathbf{S}_V \tag{7}$$

The parameter $\sqrt{D_k}$ is used to normalize the values and prevent the gradient from vanishing.

The Transformer is an attention feature weight assignor. The parameter size of the Transformer was experimentally analyzed qualitatively to better detect feature information of different sizes in $\mathbf{FC}_{avg}$ and $\mathbf{FC}_{max}$ when taking certain values [40]. Next, a join operation merges the properties of $\mathbf{FC}_{max}$ and $\mathbf{FC}_{max}$ to create a new channel attention map $MA_C(\mathbf{F}) \in \mathbb{R}^{1\times1\times1\times C}$. It was observed in the experiment that the attention map $MA_C(\mathbf{F})$ represents the fraction of the original feature map ($\mathbf{F}$) containing pathological and non-pathological regions, which is the ratio of abnormal to normal points in the image. However, if the output of $MA_C(\mathbf{F})$ is directly fed into the next convolution module, the computation will not learn the correlation of various modalities. To address this, a weighting operation is introduced to emphasize the significance of various modal lesion locations. The ratio of the output from $MA_C(\mathbf{F})$ to the original picture is multiplied, and each channel is then weighted individually. This technique further amplifies information about disease characteristics in the original image while suppressing information about non-disease aspects. The resulting feature maps from the weighting operation are input data for the subsequent layer to enhance the model's ability to recognize Alzheimer's disease.

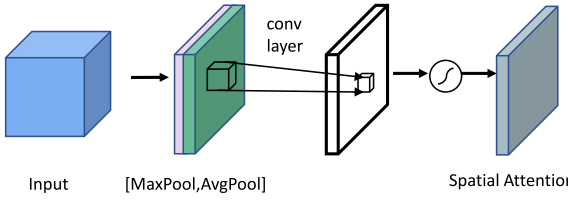


Fig. 7. Original spatial attention architecture.

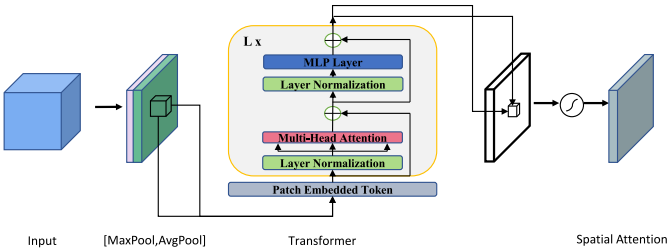


Fig. 8. Improved spatial attention architecture.

4.2.2. Improving spatial attention

In the next network layer, the original spatial attention mechanism (as depicted in Fig. 7) is upgraded by allowing the model to allocate attention weights to multi-modal data features within the spatial dimension of the feature map.

The input feature map dimension is $\mathbf{F}_{in} \in \mathbb{R}^{H\times W\times D\times C}$, and the feature maps are average-pooled and maximum-pooled for each channel to obtain the sizes $\mathbf{FS}_{avg}$ and $\mathbf{FS}_{max}$. Improving the convolutional network's original spatial attention module with a Transformer module is shown in Fig. 8, increasing the interaction of multi-modal feature data in the spatial dimension. The new approach involves first average-pooling and maximum-pooling the feature map to obtain a dimension of $\mathbf{D}_i \in \mathbb{R}^{H\times W\times D\times 1}$. Then the Transformer reassigns spatial attention weights, and the features of $\mathbf{FS}_{avg}$ and $\mathbf{FS}_{max}$ are combined using a join operation to generate a new spatial attention map $\mathbf{MA}_S(\mathbf{F}) \in \mathbb{R}^{H\times W\times D\times 1}$. Finally, the attention map is dot-multiplied with the original image, matching each original image pixel with multi-modal attention weights to obtain a spatial hierarchical fusion feature map.

4.2.3. CHAM model architecture

The CHAM model architecture is designed to improve the diagnostic assessment of Alzheimer's scan brain maps, as the original Convolutional Block Attention Module(CBAM) network is found to underperform in detecting specific pathological points. The updated channel attention network and spatial attention were combined to improve the recognition of different sizes and shapes of pathological points in the modality. This CHAM module, illustrated in Fig. 9, enables the model to capture features of each mode more effectively by reassigning attention weights to different modal channels and spatial dimensions. Furthermore, integrating a Transformer at the attention level allows for efficient feature learning and reduces the model's parameters.

Integrating Transformer modules within CHAM for both channel and spatial attention mechanisms involves reshaping the feature map to accommodate the Transformer module, which then learns the dependencies between channels for channel attention. The encoder's output is transformed through an activation function like sigmoid to generate the channel attention map. For spatial attention, the feature map is compressed along the channel dimension and input into another Transformer encoder, capturing the interactions between different spatial positions. The output from this encoder is similarly activated to form

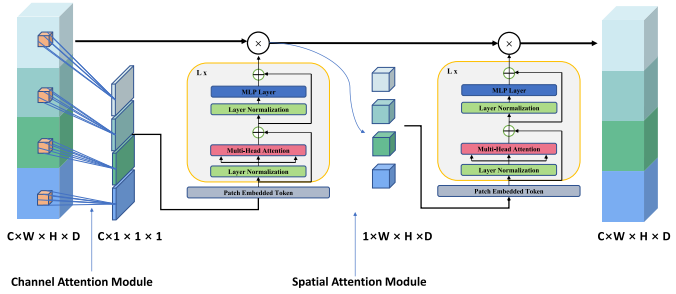


Fig. 9. CHAM module.

the spatial attention map. This integration allows the model to leverage the strengths of Transformers, capturing more complex dependencies at both channel and spatial levels.

To integrate Transformer modules into the spatial and channel attention modules of CHAM with dimensions (C, W, H, D), feature extraction is first performed on the four-dimensional feature maps. In the channel attention module, the feature map is reshaped to $(C, N)(N = W \times H \times D)$, and then passed through a Transformer encoder to capture the dependencies between channels, outputting a channel attention map processed by a sigmoid function. In the spatial attention module, the feature map channels are compressed and stacked into a shape of $(2, W, H, D)$, then reshaped to $(N, 2)$ for input into another Transformer encoder, which learns the dependencies between spatial positions. Reshape the output back, reduce it through a convolution layer to a $(1, W, H, D)$ spatial attention map, and apply a sigmoid function. These modifications enable the model to learn intricate attention patterns across channels and spatial dimensions effectively, leveraging the Transformer's ability to model long-range dependencies.

4.3. Loss function

The model training in this study involves three primary tasks: image categorization, age regression, and clinical score regression. For the categorization task, the cross-entropy loss function is utilized, which is defined as follows:

$$L_s = -\frac{1}{M}\sum_{m=1}^{M}\left(y'_m \log\left(y_m\right) + \left(1-y'_m\right)\log\left(1-y_m\right)\right) \tag{8}$$

where M represents the number of classifications, y_m is the true label, and y'_m is the predicted label assigned by the model.

The objective of the regression task in this study is to predict clinical scores that significantly influence Alzheimer's disease. This study uses root mean square error (RMSE) loss to reduce the effect of outliers on model updates since intra-class clinical score differences cannot be ignored due to the high specificity of Alzheimer's disease. The regression loss for age and clinical scores is defined as:

$$L_r = \frac{1}{2M}\sum_{i=0}^{M}\left(S_i - S'_i\right)^2 \tag{9}$$

where S_i is the predicted clinical score or age value, and S'_i is the corresponding real clinical score or age data. The overall target loss equation of the model is an aggregation of weights for the classification and regression tasks expressed as:

$$L = L_s + \left(1-\lambda_1\right)L_{r1} + \lambda_2 L_{r2} \tag{10}$$

Since the primary goal of the model is classification, the classification loss carries the largest weight. The learnable weights for the two regression tasks of the model are denoted by λ_1, and λ_2, respectively, and their sizes can be adaptively adjusted during model training.

Table 2
Training parameters.

Parameter	Value
Optimizer	Adam
Loss function	Cross-Entropy+L2
Batch size	4
Epochs	100
Learning rate	0.00001

5. Experiments and results

This section first details the experimental setup to demonstrate how the components and multi-modal fusion in the proposed technique contribute to performance. Then, it describes the ablation study findings. Next, the proposed approach is compared to other attention-based methods. Finally, visualization techniques illustrate the different attentional networks and explain why this paper's proposed method is superior.

5.1. Experimental setting

To overcome the limited amount of data available for each subject when using 3D data as input, a data augmentation method of random scaling, cropping, and flipping was used to expand the single subject data. The initial learning rate was set at 0.00001, and the Adam optimizer was used to adjust the learning rate during training dynamically. The model was trained and predicted using a batch size of 4 and 100 iterations on an NVIDIA RTX2080Ti graphics card. We utilized a 5-fold cross-validation method for model evaluation. The learning rate began at 0.00001 and was scheduled to decrease by half after every 10 training epochs. Additionally, the batch size was adjusted to 10% of its original size after every 25 epochs to improve training dynamics. To prevent overfitting, the early stop was implemented. Table 2 summarizes the individual training settings.

To evaluate our model, we divided the dataset into five parts and optionally chose one of the subsets as the test set and the other four as the training set. The training model was evaluated on the test set. This process was repeated five times to ensure each subset was used as a test set.

5.2. Evaluation metrics

The model's training outcomes are evaluated using several metrics, including accuracy, precision, sensitivity, and balanced F-score (F1-SCORE). These are calculated using the equations shown below:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$PRE = \frac{TP}{TP + FP} \tag{12}$$

$$SEN = \frac{TN}{TN + FP} \tag{13}$$

$$REC = \frac{TP}{TP + FN} \tag{14}$$

$$F_1 = 2 \times \frac{PRE \times REC}{PRE + REC} \tag{15}$$

For instance, accuracy measures the proportion of correctly predicted diseases to all diseases, as demonstrated in the example (11), where it is used to determine the accuracy of Alzheimer's disease prediction. Sensitivity, on the other hand, is the percentage of the number of correctly predicted Alzheimer's disease cases, as seen in the example (13). Precision, recall, and F1-score are also used to evaluate model performance, with the latter being a measure of the effectiveness of the model categorization. TP, TN, FP, and FN represent truly positive, truly negative, false positive, and false negative, respectively.

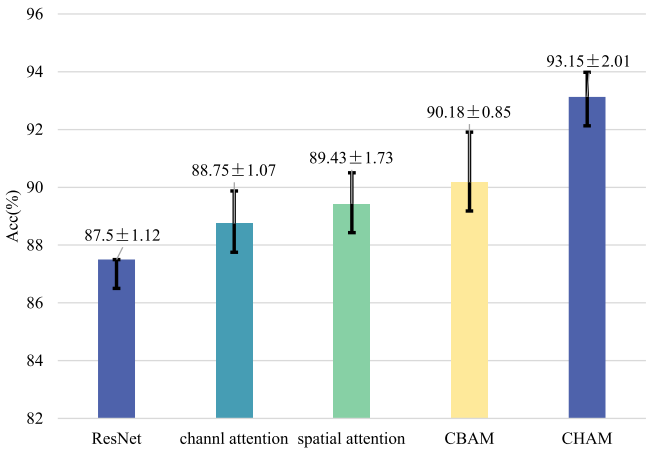


Fig. 10. Data results of different models trained and tested.

5.3. Ablation study

A series of ablation tests were conducted to evaluate the effectiveness of the proposed approach. These ablation experiments were carried out using data from ADNI1, ADNI-GO, ADNI2, and ADNI3.

5.3.1. Effectiveness of CHAM module in the diagnosis of AD

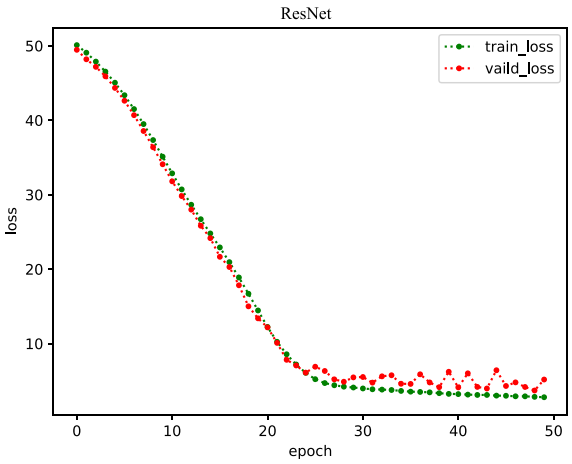
This study aims to analyze the impact of the CHAM module on the classification performance of Alzheimer's disease. Specifically, the study aims to evaluate the effectiveness of using different network configurations, including ResNet only, embedding channel attention, embedding spatial attention, embedding CBAM module, and embedding the CHAM module, on the recognition rate of Alzheimer's disease. The training results are presented in Fig. 10.

The experimental results indicate that using only the ResNet network leads to the model learning the potential representation of each branch independently, rather than for the joint representation of multi-modal data, with 87.50% ± 1.12% accuracy for five-fold cross-validation. When using channel attention and spatial attention separately, both improved compared to ResNet, with spatial attention outperforming channel attention in the Alzheimer's dataset. This is because each subject's image data is a grayscale map and contains more information in the slices. When the CBAM module is embedded, although each branch performs attention learning for its modality data to better represent potential features, it does not learn well for multi-modal data interactions, achieving the accuracy for five-fold cross-validation higher than that of ResNet alone at 90.18% ± 0.85%. Finally, when ResNet is embedded in the CHAM module, the recognition accuracy of five-fold cross-validation achieves the best value of 93.15% ± 2.01%. The training loss curves in Fig. 11 show that the ResNet model and models with channel/spatial attention modules fit slower between 20 and 30 epochs. The network combined with the CHAM module fits faster and better than the network combined with the CBAM module.

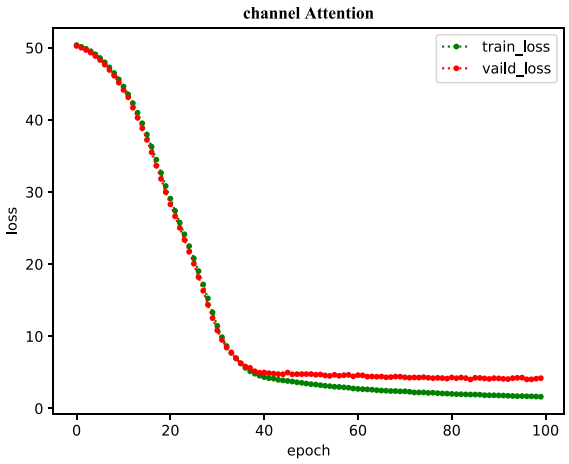
Moreover, the impact of the Transformer with different layers in the CHAM module was trained and compared as depicted in Fig. 12. The corresponding values are presented in Table 3. The study revealed that the model's identification capacity steadily increased as the number of layers increased, and the model achieved maximum accuracy when the number of layers was 4.

5.3.2. Effectiveness of multi-task in the diagnosis of AD

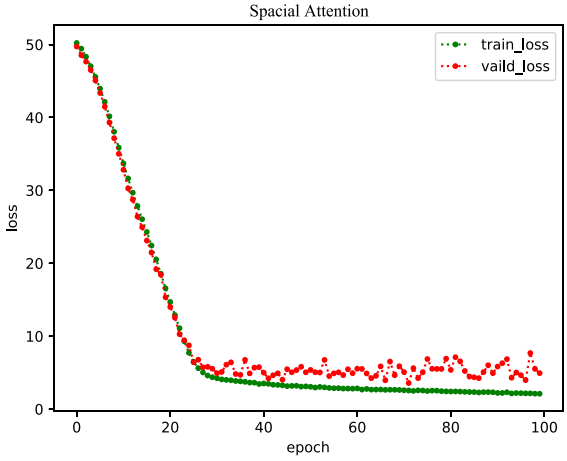
The previous study proposed a multi-task learning approach to improve the generalization ability of the model while reducing its complexity through a hard constraint-sharing method. During the collection of Alzheimer's data, the ADNI database was utilized to determine the



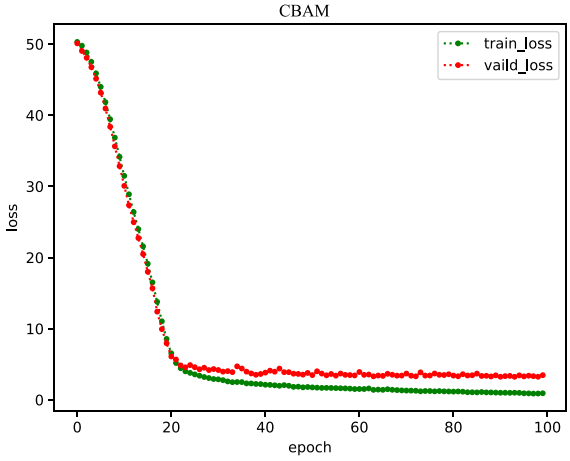
(a) Resnet



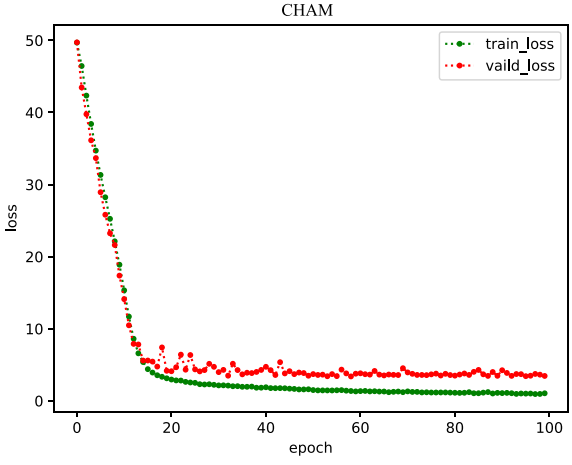
(b) Embedding Channel Attention



(c) Embedding Spatial Attention



(d) Embedding CBAM module



(e) Embedding CHAM module

Fig. 11. Training loss curves for the five training models.

stage of the disease based on various clinical questionnaire scores. Consequently, our experiments used the MMSE clinical scores as a model task. Additionally, brain volume tends to decrease with age in Alzheimer's patients, making age another relevant task. Each task was analyzed in this study, and the results are presented in Table 4. As depicted in Fig. 13, using dichotomous classification alone resulted in weaker recognition outcomes. However, when age or MMSE score data were incorporated as relevant complementary tasks, the accuracy of the model improved, confirming that age and score data impact Alzheimer's disease detection. Performance improved significantly

Table 3
Evaluation values of the Transformer with different layers in CHAM modules. In the results, the former represents the mean and the latter the standard deviation.

Layer number	Acc (%)	F1 (%)
1	90.00 ± 0.69	90.00 ± 0.60
2	90.92 ± 1.05	90.90 ± 1.03
3	92.41 ± 1.31	92.40 ± 1.25
4	**93.15 ± 2.01**	**93.14 ± 1.96**
5	92.70 ± 1.81	92.67 ± 1.91
6	93.00 ± 0.64	93.00 ± 0.72
7	92.55 ± 0.80	92.55 ± 0.78
8	90.92 ± 1.01	90.91 ± 1.10

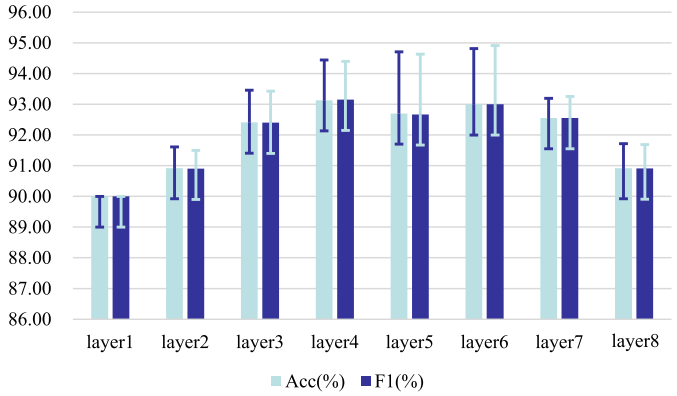


Fig. 12. Classification effect of the Transformer in CHAM module with different layers.

Table 4
Scoring metrics for identifying Alzheimer's disease by different tasks. In the results, the former represents the mean and the latter the standard deviation.

Index	II-Classify	II-Classify&AGE	II-Classify &MMSE	II-Classify&AGE &MMSE
ACC (%)	91.07 ± 0.85	92.41 ± 1.07	92.56 ± 1.72	**93.15 ± 2.01**
PRE (%)	91.43 ± 0.60	92.48 ± 0.81	92.96 ± 1.49	**93.57 ± 2.00**
REC (%)	91.07 ± 0.71	92.41 ± 1.30	92.56 ± 1.32	**93.15 ± 1.92**
F1 (%)	91.04 ± 0.66	92.41 ± 0.84	92.54 ± 1.56	**93.14 ± 1.96**
AUC (%)	91.07 ± 0.91	92.41 ± 1.10	92.56 ± 1.69	**93.15 ± 2.08**

when combining all three tasks. Each task positively affects Alzheimer's disease recognition, and different classification features are distributed among the tasks. Notably, a single AD classification task cannot share the same feature representation, a limitation. In contrast, considering multiple tasks can guide HAMMF learning more comprehensively.

5.3.3. Effectiveness of different modal characteristics on the diagnosis of AD

To examine how different types of features affect the classification results of a model, the study compared single modality patterns to multiple modality patterns. Table 5 and Fig. 14 presents the results of all the comparisons. The brain's white and gray matter contains many memory related neurons, and researchers have found that the volume of white and gray matter in the brains of Alzheimer's disease (AD) patients is smaller than in healthy subjects. This experiment confirmed that gray and white matter images are crucial in identifying AD, while PET images complement the diagnosis. When multiple modalities were fused for classification, the highest accuracy was achieved by combining CSF, PET, GM and WM image groups (93.15% ± 2.01%) followed by GM, PET, and WM image groups (92.85% ± 1.50%), WM, CSF and PET image groups (92.55% ± 1.83%), GM, WM, and CSF image groups (91.07% ± 1.05%), GM, CSF, and PET image group (90.92% ± 0.98%). The study concluded that the proposed network is more stable in learning white matter features, and the model's accuracy decreases significantly when white matter features are missing. Conversely, the absence of cerebrospinal fluid data less affects the model's

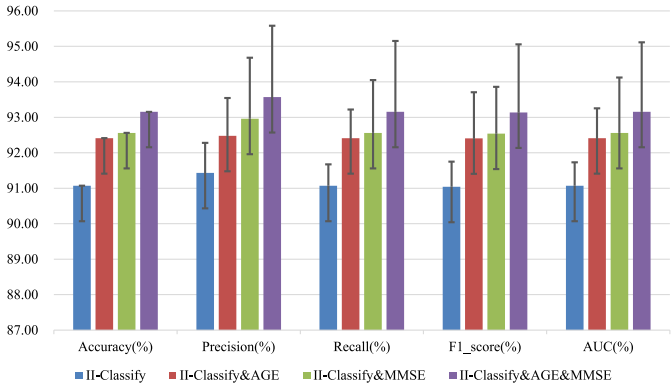


Fig. 13. Classification results of Alzheimer's by different tasks.

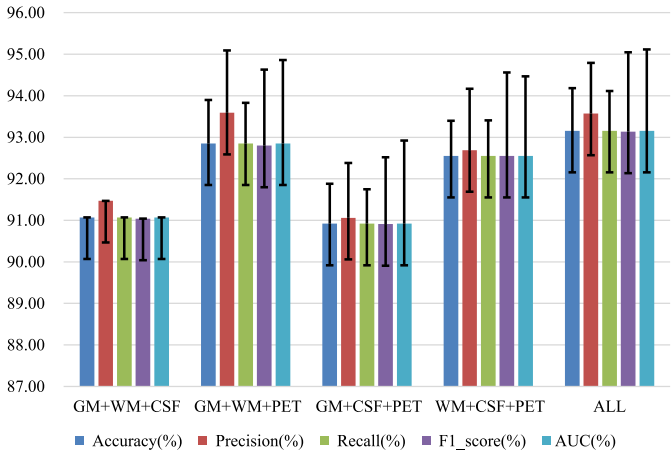


Fig. 14. Classification results of Alzheimer's by different combination characteristics.

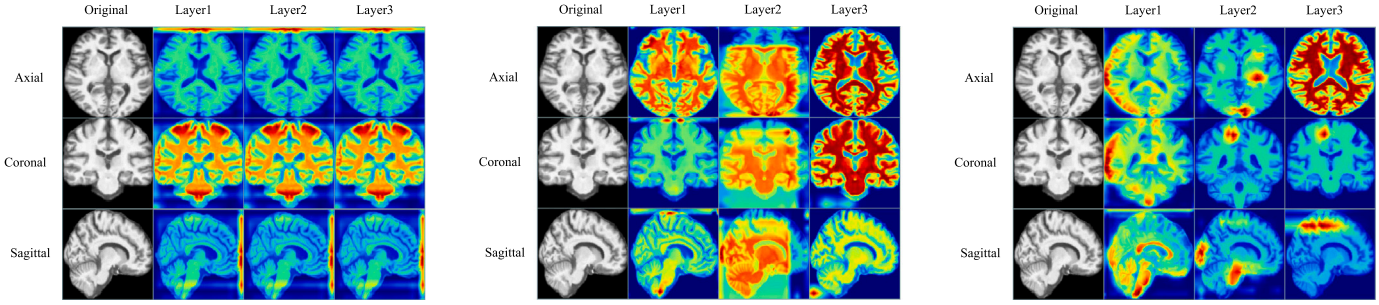
accuracy because fewer features are learned from this modality. Based on this dataset, the degree of modality importance was ranked as follows: white matter images > PET images > gray matter images > cerebrospinal fluid images.

5.4. Comparison of models and existing methods

The paper compares the proposed method with existing methods regarding total training time to evaluate its performance. Table 6 displays the comparison, and the proposed method outperforms existing methods in terms of evaluation metrics. The results show the proposed network exhibits excellent classification performance. The proposed method performs well in classification despite inferior model size and training time compared to the tiny model with 50.654 MB size, 120.30 min training time, and 93.15% ± 2.01% accuracy, model size and time are significantly reduced compared to comparable structured models. Adding the original CBAM module to ResNet increased training duration, model size to 21.6 MB, and accuracy to 90.18% ± 1.06%. Introducing the CHAM module significantly improved accuracy, though training time and parameters increased. The proposed model achieves better results than recent methods in AD classification, and its performance is comprehensive. The proposed model for Alzheimer's recognition was compared to recent methods that use the entire 3D image as input, and it was discovered that the proposed model performs better in AD classification. Additionally, when comparing networks with Transformer structure, it was observed that the network proposed by Dai et al. [41] has a significantly larger number of parameters than the current paper. Still, the difference in training accuracy is negligible. The method of Jang et al. [42] underperforms our method of

Table 5
Results of Alzheimer's classification by different modal characteristics. In the results, the former represents the mean and the latter the standard deviation.

Modals	ACC (%)	PRE (%)	REC (%)	F1 (%)	AUC (%)
GM+WM+CSF	91.07 ± 1.05	91.47 ± 0.96	91.07 ± 0.85	91.04 ± 1.03	91.07 ± 1.01
GM+WM+PET	92.85 ± 1.50	**93.59 ± 1.32**	92.85 ± 1.48	92.80 ± 1.22	92.85 ± 1.45
GM+CSF+PET	90.92 ± 0.98	91.06 ± 0.83	90.92 ± 0.86	90.91 ± 0.96	90.92 ± 1.06
WM+CSF+PET	92.55 ± 1.83	92.69 ± 1.61	92.55 ± 2.01	92.55 ± 1.91	92.55 ± 1.85
ALL	**93.15 ± 2.01**	93.57 ± 2.00	**93.15 ± 1.92**	**93.14 ± 1.96**	**93.15 ± 2.08**



(a) Heat map of ResNet. **(b)** Heat map of CBAM. **(c)** Heat map of CHAM.

Fig. 15. Heat map of attention of three models with different layers and different slice directions.

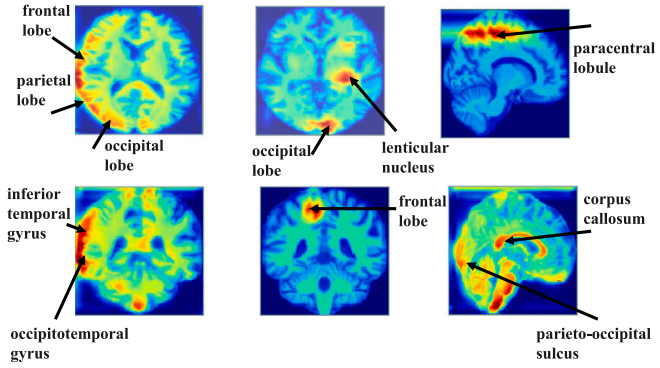


Fig. 16. Labeling of anatomical parts of the brain.

Table 6
Comparison of existing methods with the method proposed in this paper. In the results, the former represents the mean and the latter the standard deviation.

Methods	Size (MB)	Training time (min)	ACC (%)
ResNet [7]	11.670M	84.6 ± 10.4	87.5 ± 2.36
SENet [10]	12.681M	86 ± 15.2	88.75 ± 3.34
CABM [43]	21.683M	87.5 ± 12.3	90.18 ± 1.06
VGG19 [44]	78.1 MB	140.1 ± 31.5	88.2 ± 0.27
H-FCN28 [45]	62.54 MB	12.5 ± 5.5	90.32 ± 0.68
CNN [46]	20.33 MB	60.8 ± 11.4	82.93 ± 0.34
Attention + MIL + CNN [47]	89.67 MB	130.9 ± 30.4	92.4 ± 1.5
TransMed [41]	385 MB	540 ± 60.4	93.61 ± 0.56
M3T [42]	89.22M	188 ± 20.4	92.21 ± 1.03
HAMMF	50.654M	120.3 ± 15.6	93.15 ± 2.01

multimodal Alzheimer's disease fusion. Overall, the study demonstrates that the proposed model delivers exceptional results in terms of overall performance.

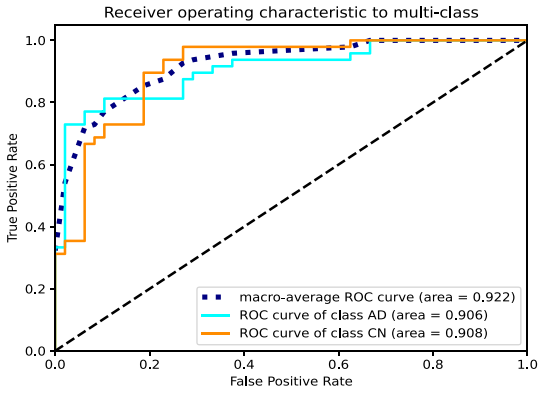
The statistical testing method used in this study is the paired t-test. The method of this paper is compared with the multi-task and multi-modal fusion methods. We conduct statistical tests between the top three methods in terms of accuracy and the proposed method. This study conducts statistical tests between the three methods in terms of accuracy and the proposed method. The HAMMF model exhibits a marginally superior diagnostic accuracy compared to the TransMed (paired t-test: $t = 2.31$, $p < 0.015$), Attention + MIL + CNN (paired t-test: $t = 3.64$, $p < 0.017$), M3T(paired t-test: $t = 2.27$, $p < 0.008$) models in multi-task and multi-modal assisted diagnosis.

5.5. Model visualization comparison

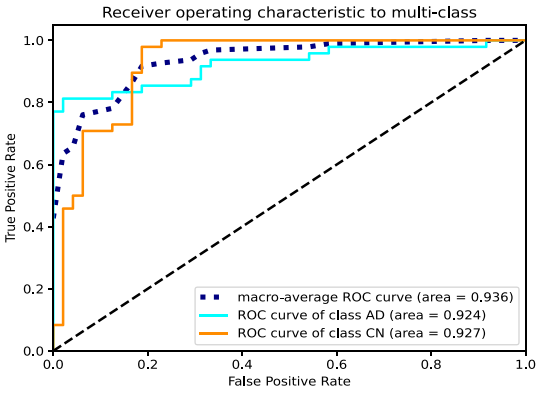
Due to variable sized lesion patches in the brains of Alzheimer's patients, there is more variation between clusters and less variation within clusters. During training, the model focused not just on disease-related information but also on most non-disease information as the number of training cycles increased. The non-disease information in the model's final classification became an extraneous feature that interfered with disease diagnosis, lowering test set classification accuracy. We use Grad-CAM [48] visualization in this study to better illustrate how different models affect network performance and make the model more interpretable. Grad-CAM can show the areas the model focuses on after various modules. In this experiment, Grad-CAM, the CBAM modular network, and HAMMF were used to output heatmaps after each ResNet network level. The heatmap is shown in Fig. 15. The proposed model in this paper was able to represent the data characteristics well in learning Alzheimer's MRI image data, and the black areas represent the regions the model focuses on, likely with severe brain lesions. In contrast, the ResNet network and the combined CBAM module network performed poorly in learning, and their attention regions were inadequate. To demonstrate the model's effectiveness in identifying AD-specific patterns in neuroimaging data, this paper also labeled the attentional map of HAMMF with anatomical parts, as shown in Fig. 16. This paper concludes that the HAMMF network has excellent results in diagnosing Alzheimer's disease, and comparing brain anatomy shows the model pays more attention to frontal, parietal, occipital lobes, lenticular nucleus, and other regions, which to some extent explains the pattern of model learning.

5.6. Model performance evaluation

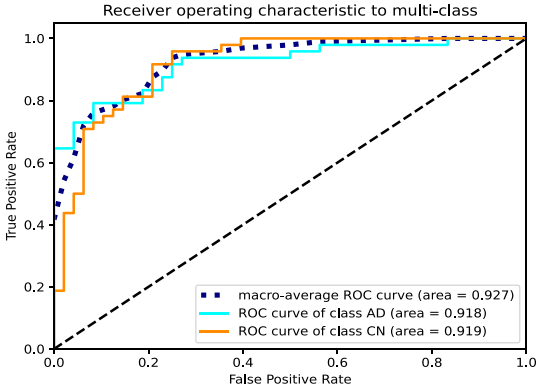
In Fig. 17(a), the area under the macro-average ROC curve is 0.922, indicating that the overall model has high classification performance. The ROC curve area for the AD class is 0.906, while the area for



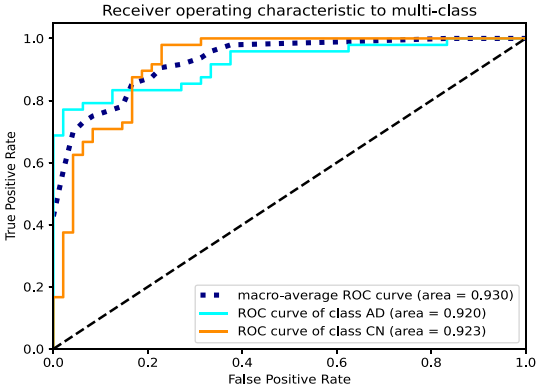
(a) Embedding Channel Attention



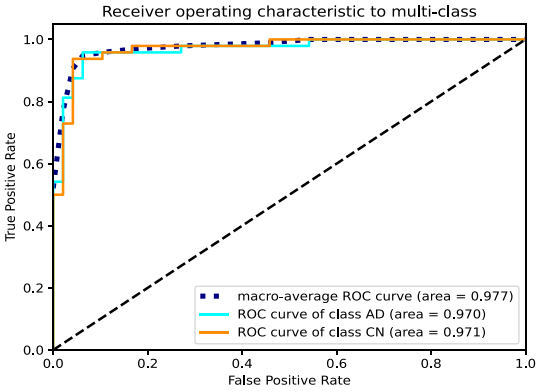
(b) Embedding CBAM module



(c) Resnet



(d) Embedding Spatial Attention



(e) Embedding CHAM module

Fig. 17. ROC curves for the five models. ROC curves typically feature true positive rate (TPR) on the Y axis, and false positive rate (FPR) on the X axis.

the CN class is 0.908; these two values are close and relatively high, suggesting that the model also has good classification ability for these two categories. The closer the ROC curve is to the upper left corner, the better the performance of the model is generally, as it implies achieving a higher true positive rate while maintaining a low false positive rate. Following the above analysis of Fig. 17(b)(c)(d)(e), it can be seen that the CHAM module has the best expressive power among all the models.

6. Conclusion

This study presents a novel approach called HAMMF for early diagnosis of Alzheimer's disease. The proposed model employs a contextual hierarchical attention module to perform secondary extraction of multi-modal disease features obtained through ResNet. The attention module separately extracts channel and spatial attention feature

weights within each modality and then redistributes attention feature weights among different modalities. This multiple attention allocation approach effectively strengthens the correlation between various modalities, enhancing the extraction of potential information and improving diagnostic performance. By implementing the Transformer at the attention level, computational volume and computation time are reduced compared to applying the Transformer directly to multi-modal voxel features. Furthermore, this study incorporates disease-related regression tasks as auxiliary judgments in the classification problem, exploring the connections between subject age, clinical scores, subjects, and the importance of different tasks in multiple ways to inform the final diagnosis.

The proposed method's effectiveness is demonstrated on the ADNI dataset and outperformed traditional CNN and the latest methods with an accuracy of 93.15%. Visualization experiments are also conducted to validate the effectiveness and reliability of the proposed method. In this study, the developed model is interpreted from a visual perspective for AD classification. The proposed model in this paper was able to represent the data characteristics well in learning Alzheimer's MRI image data, and the black areas represent the regions the model focuses on, likely with severe brain lesions. In contrast, the ResNet network and the combined CBAM module network performed poorly in learning, and their attention regions were inadequate. Furthermore, utilizing various evaluation metrics to explain the model's classification performance quantitatively, it can be observed that our classification results perform relatively well overall among all outcomes.

While the proposed model exhibits excellent performance, certain aspects require further consideration. The model may show significant performance variations when applied to new, unseen data compared to its performance on the training data. Moreover, the study is limited to MRI and PET imaging modalities; the efficacy of the model with other modalities is yet to be determined. Addressing these points could provide a more comprehensive evaluation of the model's generalization ability.

7. Limitations and future work

The study proposes a hierarchical attention-based multi-task and multi-modal fusion model that performs well on the ADNI dataset. However, the model has some limitations that need to be addressed. The multi-task learning experiment only considers age and MMSE score data, which is not comprehensive enough since AD clinical score data, such as GDSCALE, FAQ, and other score data, should also be included. Also, biomarkers are essential for AD diagnosis and should be incorporated into the model.

In discussing our model for Alzheimer's disease diagnosis, we must acknowledge the presence of certain limitations, particularly concerning potential data set biases, the generalizability of the model, and its performance in clinical settings. Firstly, data set bias is a significant concern. The dataset our model relies on may not represent the broader population, especially if the data is predominantly sourced from specific demographics (such as particular races, age groups, or geographical regions). This limitation constrains the accuracy and reliability of our model when generalized to a global population. Secondly, the issue of generalizability involves the model's consistent performance across different datasets. While the current model performs well on the training dataset employed, its performance on external datasets remains unknown. To enhance the model's generalizability, we need to employ cross-dataset validation, utilize broader cross-validation techniques, and explore transfer learning strategies to adapt to different data distributions. Lastly, the model's performance in actual clinical settings can be influenced by a variety of factors, including differing diagnostic criteria, data acquisition methods, and variability in clinical practices. Therefore, even if a model exhibits excellent performance in a research setting, it may encounter challenges in clinical practice. To overcome these limitations, empirical studies need to be conducted in clinical settings to test the model's efficacy and iteratively optimize it based on clinical feedback. Moreover, establishing collaborative relationships with clinicians to gain insights from practical applications is crucial for guiding model improvements. In summary, although the model shows potential for AD diagnosis, further research and development are necessary to ensure its accuracy, robustness, and adaptability for application across diverse populations and clinical practice.

Future work should focus on constructing a multi-task representation of the model using more comprehensive tasks and analyzing the impact of different tasks on classification. Secondly, the current model design, though better than similar models, requires expansion due to its size and some missing multi-modal data. Future work needs to generate more supplementary data with the help of adversarial networks to increase the data capacity and thus improve the model's generalization performance.

CRediT authorship contribution statement

Xiao Liu: Writing – review & editing, Methodology, Conceptualization. **Weimin Li:** Writing – review & editing, Supervision, Formal analysis. **Shang Miao:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Fangyu Liu:** Visualization, Validation, Conceptualization. **Ke Han:** Visualization, Supervision. **Tsigabu T. Bezabih:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgment

This work was supported by National Key Research and Development Program of China (No. 2022YFC3302600).

References

[1] A. Association, et al., 2010 Alzheimer's disease facts and figures, Alzheimer's Dementia 6 (2) (2010) 158–194.
[2] W. Jagust, Vulnerable neural systems and the borderland of brain aging and neurodegeneration, Neuron 77 (2) (2013) 219–234.
[3] J. Cummings, G. Lee, K. Zhong, J. Fonseca, K. Taghva, Alzheimer's disease drug development pipeline: 2021, Alzheimer's Dementia: Transl. Res. Clin. Interv. 7 (1) (2021) e12179.
[4] M. Nawaz, T. Nazir, M. Masood, A. Mehmood, R. Mahum, M.A. Khan, S. Kadry, O. Thinnukool, Analysis of brain MRI images using improved cornernet approach, Diagnostics 11 (10) (2021) 1856.
[5] M.S. Ullah, M.A. Khan, A. Masood, O. Mzoughi, O. Saidani, N. Alturki, Brain tumor classification from MRI scans: a framework of hybrid deep learning model with Bayesian optimization and quantum theory-based marine predator algorithm, Front. Oncol. 14 (2024).
[6] G. Chetelat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, J.-C. Baron, Mild cognitive impairment: can FDG-PET predict who is to rapidly convert to Alzheimer's disease? Neurology 60 (8) (2003) 1374–1377.
[7] 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.
[8] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[9] V.D. Calhoun, J. Sui, Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness, Biol. Psychiatry: Cogn. Neurosci. Neuroimaging 1 (3) (2016) 230–244.
[10] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[11] C.R. Jack Jr., M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J. L. Whitwell, C. Ward, et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging: Off. J. Int. Soc. Magn. Reson. Med. 27 (4) (2008) 685–691.

[12] Z. Liu, H. Lu, X. Pan, M. Xu, R. Lan, X. Luo, Diagnosis of Alzheimer's disease via an attention-based multi-scale convolutional neural network, Knowl.-Based Syst. 238 (2022) 107942.
[13] T. Zhang, M. Shi, Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease, J. Neurosci. Methods 341 (2020) 108795.
[14] F. Liu, S. Yuan, W. Li, Q. Xu, B. Sheng, Patch-based deep multi-modal learning framework for Alzheimer's disease diagnosis using multi-view neuroimaging, Biomed. Signal Process. Control 80 (2023) 104400.
[15] S. Ahmed, K.Y. Choi, J.J. Lee, B.C. Kim, G.-R. Kwon, K.H. Lee, H.Y. Jung, Ensembles of patch-based classifiers for diagnosis of Alzheimer diseases, IEEE Access 7 (2019) 73373–73383.
[16] F. Liu, S. Yuan, W. Li, Q. Xu, B. Sheng, Patch-based deep multi-modal learning framework for Alzheimer's disease diagnosis using multi-view neuroimaging, Biomed. Signal Process. Control 80 (2023) 104400.
[17] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, H. Benali, Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI, Neuroradiology 51 (2009) 73–83.
[18] Y. Zhao, C. Sun, X. Xu, J. Chen, RIC-Net: A plant disease classification model based on the fusion of inception and residual structure and embedded attention mechanism, Comput. Electron. Agric. 193 (2022) 106644.
[19] A. Hamza, M.A. Khan, M. Alhaisoni, A. Al Hejaili, K.A. Shaban, S. Alsubai, A. Alasiry, M. Marzougui, D2BOF-covidnet: a framework of deep bayesian optimization and fusion-assisted optimal deep features for COVID-19 classification using chest X-ray and mri scans, Diagnostics 13 (1) (2022) 101.
[20] 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.
[21] M. Odusami, R. Maskeliūnas, R. Damaševičius, Optimized convolutional fusion for multimodal neuroimaging in Alzheimer's disease diagnosis: Enhancing data integration and feature extraction, J. Pers. Med. 13 (10) (2023) 1496.
[22] C. Ge, Q. Qu, I.Y.-H. Gu, A.S. Jakola, Multi-stream multi-scale deep convolutional networks for Alzheimer's disease detection using MR images, Neurocomputing 350 (2019) 60–69.
[23] C. Yu, Z. Gao, W. Zhang, G. Yang, S. Zhao, H. Zhang, Y. Zhang, S. Li, Multitask learning for estimating multitype cardiac indices in MRI and CT based on adversarial reverse mapping, IEEE Trans. Neural Netw. Learn. Syst. 32 (2) (2020) 493–506.
[24] X. Hou, Y. Bai, Y. Xie, Y. Li, Mass segmentation for whole mammograms via attentive multi-task learning framework, Phys. Med. Biol. 66 (10) (2021) 105015.
[25] H. Cheng, S. Yuan, W. Li, X. Yu, F. Liu, X. Liu, T.T. Bezabih, De-accumulated error collaborative learning framework for predicting Alzheimer's disease progression, Biomed. Signal Process. Control 89 (2024) 105767.
[26] J. Wang, W. Li, W. Liu, C. Wang, Q. Jin, Enabling inductive knowledge graph completion via structure-aware attention network, Appl. Intell. 53 (21) (2023) 25003–25027.
[27] A.M. Luvembe, W. Li, S. Li, F. Liu, G. Xu, Dual emotion based fake news detection: A deep attention-weight update approach, Inf. Process. Manage. 60 (4) (2023) 103354.
[28] M. Masood, R. Maham, A. Javed, U. Tariq, M.A. Khan, S. Kadry, Brain MRI analysis using deep neural network for medical of internet things applications, Comput. Electr. Eng. 103 (2022) 108386.
[29] J. Ramya, B.U. Maheswari, M. Rajakumar, R. Sonia, Alzheimer's disease segmentation and classification on MRI brain images using enhanced expectation maximization adaptive histogram (EEM-AH) and machine learning., Inf. Technol. Control 51 (4) (2022) 786–800.
[30] M. Odusami, R. Maskeliūnas, R. Damaševičius, Pareto optimized adaptive learning with transposed convolution for image fusion Alzheimer's disease classification, Brain Sci. 13 (7) (2023) 1045.
[31] Z. Qin, Z. Liu, Q. Guo, P. Zhu, 3D convolutional neural networks with hybrid attention mechanism for early diagnosis of Alzheimer's disease, Biomed. Signal Process. Control 77 (2022) 103828.
[32] H. Xie, X. Zeng, H. Lei, J. Du, J. Wang, G. Zhang, J. Cao, T. Wang, B. Lei, Cross-attention multi-branch network for fundus diseases classification using SLO images, Med. Image Anal. 71 (2021) 102031.
[33] S. Banerjee, S. Chaudhuri, et al., DeFraudNet: End2End fingerprint spoof detection using patch level attention, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2695–2704.
[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.
[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 213–229.
[38] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, Neuroimage 15 (1) (2002) 273–289.
[39] H. Wen, Y. Liu, I. Rekik, S. Wang, Z. Chen, J. Zhang, Y. Zhang, Y. Peng, H. He, Multi-modal multiple kernel learning for accurate identification of tourette syndrome children, Pattern Recognit. 63 (2017) 601–611.
[40] S. Miao, Q. Xu, W. Li, C. Yang, B. Sheng, F. Liu, T.T. Bezabih, X. Yu, MMTFN: Multi-modal multi-scale transformer fusion network for Alzheimer's disease diagnosis, Int. J. Imaging Syst. Technol. (2023).
[41] Y. Dai, Y. Gao, F. Liu, Transmed: Transformers advance multi-modal medical image classification, Diagnostics 11 (8) (2021) 1384.
[42] J. Jang, D. Hwang, M3T: three-dimensional medical image classifier using multi-plane and multi-slice transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20718–20729.
[43] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[44] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[45] C. Lian, M. Liu, J. Zhang, D. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2018) 880–893.
[46] D. Lu, K. Popuri, G. Ding, R. Balachandar, M. Beg, Alzheimer's disease neuroimaging I. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images, Sci. Rep. 8 (1) (2018) 5697.
[47] W. Zhu, L. Sun, J. Huang, L. Han, D. Zhang, Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI, IEEE Trans. Med. Imaging 40 (9) (2021) 2354–2366.
[48] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.