Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 33(10): 5916–5936.
DOI: 10.3934/era.2025263
Received: 11 July 2025
Revised: 27 August 2025
Accepted: 04 September 2025
Published: 10 October 2025

Research article

MMF-ViT: A multi-scale multi-domain frequency-aware vision Transformer for MRI-based Alzheimer's classification

Ying Liu and XiaoLi Yang*

School of Mathematics and Statistics, Shaanxi Normal University, 620 West Chang'an Street, Chang'an District, Xi'an 710119, China

* **Correspondence:** Email: yangxiaoli@snnu.edu.cn.

Abstract: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that imposes a substantial burden on families and healthcare systems. Mild cognitive impairment (MCI), as an intermediate stage between normal aging and AD, can be further divided into progressive MCI (pMCI) and stable MCI (sMCI) based on follow-up outcomes. Unlike the marked differences observed between cognitively normal (CN) individuals and AD patients, sMCI and pMCI share highly similar characteristics, making early identification of pMCI extremely challenging. Although deep learning methods based on structural magnetic resonance imaging (sMRI) have advanced AD classification, research on predicting MCI progression remains limited due to the high similarity between sMCI and pMCI as well as the substantial cost of prospectively collecting longitudinal data. Accurate early identification of pMCI is essential for timely intervention, slowing disease progression, and reducing healthcare costs. Therefore, this study focused on the early identification of progressive MCI. To address this, we proposed a novel vision Transformer framework, the multi-scale multi-domain frequency-aware vision Transformer (MMF-ViT), which employs a multi-scale cross-domain fusion (MSCDF) module to enable deep interaction between spatial and frequency domain features, thereby enhancing the modeling of fine-grained brain structural variations. The multi-scale frequency encoder (MSFE) and multi-scale context encoder (MSCE) were designed to extract and fuse frequency and spatial information, effectively improving classification performance. Experimental results on the ADNI dataset demonstrate that MMF-ViT achieves an accuracy of 72.84% and an AUC of 72.99% for sMCI versus pMCI classification, significantly outperforming mainstream 2D and 3D models. In AD vs. CN classification, MMF-ViT also achieves an accuracy of 85.59%, highlighting its strong feature representation capability and practical potential.

Keywords: Alzheimer's disease classification; structural MRI; vision transformer; frequency-domain modeling; multi-scale fusion

1. Introduction

Alzheimer's disease (AD) is a progressive disorder characterized by a lengthy course, which imposes substantial pressure and burden on both families and healthcare systems [1]. Mild cognitive impairment (MCI) represents an intermediate stage between normal aging and AD. Based on whether MCI patients progress to AD after a period of follow-up (typically 36 months), current MCI cases can be further divided into two finer subtypes: progressive MCI (pMCI) and stable MCI (sMCI) [2]. Unlike the significant feature distribution differences between cognitively normal (CN) and AD, which make classification tasks relatively easier, sMCI and pMCI share highly similar characteristics, making them difficult to distinguish [3]. Therefore, early identification of pMCI is extremely challenging. Additionally, due to the high intrinsic similarity between sMCI and pMCI data and the high cost of prospectively collecting longitudinal data, research on predicting MCI progression remains limited. Most studies focus on binary classification of AD vs. CN or three-way classification (AD vs. MCI vs. CN) [4–6]. However, early identification and diagnosis of pMCI is crucial for timely intervention, slowing disease progression, and reducing healthcare costs. Our work focuses on the identification of pMCI.

Structural magnetic resonance imaging (sMRI) has proven to be an effective imaging modality for early detection of AD, offering excellent tissue contrast and high spatial resolution [7]. With the rapid development of deep learning, convolutional neural networks (CNNs) have achieved remarkable results in AD classification tasks by extracting discriminative spatial features from MRI data [8, 9]. Some studies have further combined CNNs with attention mechanisms to explicitly enhance local-global dependencies and improve AD classification performance [10]. However, due to their limited receptive field, CNNs inherently struggle to capture long-range dependencies and global context, restricting their ability to model subtle and widespread brain structural changes in early AD. To address this, the Transformer architecture [11], with its powerful global modeling capabilities, has emerged as a promising alternative. Recent research integrates Transformers with CNNs to leverage both local and global features, thus enhancing AD classification performance [12, 13]. Although these hybrid models have achieved success, most employ a serial structure, attaching the Transformer after the CNN features, and lack deep cross-domain interaction and joint modeling between the frequency and spatial domains. Current methods generally rely on fixed-scale feature extraction, making it difficult to fully capture the fine-grained structural changes under multi-scale patterns of brain atrophy. These limitations directly affect the performance and stability of models on sMCI vs. pMCI classification tasks.

In AD classification, several studies incorporate multi-scale concepts into CNN architectures [14, 15] or use them to build attention modules [16]. Tian et al. [17] proposed a multi-scale large-kernel fully separable convolutional neural network (MSCLK) for early diagnosis of Alzheimer's disease, and further extended it to multimodal MRI-PET fusion. Similarly, Yan et al. [18] introduced a multi-scale convolutional network with ensemble learning (MCNEL), which effectively integrates the features of EfficientNet, MobileNet, and DenseNet, demonstrating stronger robustness and superior performance on the ADNI dataset. However, these approaches mostly focus on the spatial domain and lack modeling of frequency domain information. On the other hand, several studies have explored the use of frequency domain features in medical imaging tasks. For example, some works incorporate both spatial and low-frequency domain features for brain MRI segmentation [19]. More recently, Feng et al. [20] proposed

a wavelet transform-based CNN (WTCNN) that integrates sMRI and SNP genetic data, demonstrating that combining spatial and frequency-domain cues can effectively improve AD classification accuracy. Notably, Kushol et al. [21] proposed a fusion Transformer that integrates features from both the spatial and Fourier frequency domains and achieved promising results on the classification between CN and AD task. However, their approach did not incorporate multi-scale modeling and has not been validated in the classification between sMCI and pMCI. Nevertheless, the existing frequency-domain approaches in medical image analysis rarely incorporate multi-scale modeling. They usually extract frequency features at a single resolution, making it difficult to capture fine-grained structural variations across different scales. In addition, very few of these approaches have been evaluated on more challenging fine-grained tasks such as distinguishing sMCI from pMCI, where nuanced structural differences are critical. Similarly, although some multi-scale frameworks have been proposed in the spatial domain, they generally lack frequency-domain representations. As a result, the existing methods either miss multi-scale frequency information or fail to integrate spatial and frequency cues in a unified manner, leading to suboptimal characterization of brain atrophy patterns.

To specifically address these shortcomings—namely the absence of multi-scale mechanisms in frequency-domain methods, the lack of frequency-domain modeling in spatial multi-scale approaches, and the insufficient validation on fine-grained tasks such as sMCI versus pMCI—we propose an efficient hybrid vision Transformer framework, MMF-ViT. Given the limitations of existing brain MRI methods in early AD progression prediction—particularly their insufficient ability to jointly model multi-scale and cross-domain features—we propose the MMF-ViT (multi-scale multi-domain frequency-aware vision Transformer), which integrates convolutional layers for local structural feature extraction and Transformer modules for global context modeling. This hybrid design better captures both fine-grained variations and long-range dependencies in brain images, effectively addressing the insufficient characterization of multi-scale brain atrophy patterns and complex structural changes in current models. To address the limitation of insufficient joint modeling between spatial and frequency domain features, we design a multi-scale cross-domain fusion (MSCDF) module between the convolutional and Transformer stages. Based on attention mechanisms, the MSCDF module enables interactive fusion of spatial and frequency domain features: spatial attention enhances frequency domain representation, while channel attention refines spatial features, effectively improving the model's ability to characterize multi-scale and multi-domain brain structural changes.

The main contributions of this work are summarized as follows:

1) **Framework innovation:** We propose the MMF-ViT framework, which organically combines convolution and Transformer modules to achieve both local and global feature modeling. This enables efficient feature representation at a relatively low computational cost and addresses the limitations of traditional serial or single-domain models in brain MRI analysis.
2) **Fusion module innovation:** We design the MSCDF module to achieve interactive fusion of multi-scale spatial and frequency domain features, improving the model's sensitivity to subtle structural changes and providing more stable and effective performance, particularly in challenging tasks such as the classification between sMCI and pMCI.
3) **Submodule innovation:** Within MSCDF, we introduce a multi-scale frequency encoder (MSFE) and a multi-scale context encoder (MSCE), responsible for capturing structural details in the frequency domain and multi-scale contextual information in the spatial domain, respectively. These two modules are complementary in brain MRI analysis and significantly

enhance regional discrimination.

4) **Superior empirical performance:** Extensive experiments on the ADNI dataset demonstrate that our MMF-ViT framework achieves state-of-the-art results, particularly in the challenging sMCI vs. pMCI classification task, outperforming existing 2D and 3D models on key metrics such as accuracy, F1-score, and AUC, while maintaining high efficiency in terms of model size and computational cost.

The remainder of this paper is organized as follows: Section 2 details the methodology of the proposed MMF-ViT model, including the MSCDF module, the MSFE and MSCE encoders, and the complementary fusion strategy based on spatial and channel attention. Section 3 describes the dataset, preprocessing pipeline, and experimental setup, and reports extensive results including baseline comparisons, ablation studies, slice-direction analysis, and interpretability evaluation. Section 4 provides the conclusions and discusses potential future directions.

2. Methods

This section introduces the overall methodology of the proposed MMF-ViT framework. As shown in Figure 1, the model is built around the MSCDF module, which serves as its core component and consists of three parts: the MSFE, the MSCE, and a complementary fusion strategy based on spatial and channel attention. We then describe the majority voting strategy used to aggregate predictions across multiple slices. Finally, we outline the overall processing pipeline of MMF-ViT.

2.1. Multi-scale cross-domain fusion (MSCDF) module

The MSCDF module constitutes the central contribution of this work. It is designed to jointly model frequency-domain components and multi-scale spatial structures in medical images. As illustrated in Figure 1, the MSCDF module integrates three key sub-components: 1) a MSFE, 2) a MSCE, and 3) a complementary fusion mechanism based on spatial and channel attention.

2.1.1. Multi-scale frequency encoder (MSFE)

To effectively model both high- and low-frequency information in feature maps, we propose a MSFE. The MSFE architecture is illustrated in Figure 1(c). This module utilizes frequency decomposition, multi-scale fusion, and channel attention to capture fine-grained frequency components and enhance feature representation.

At scale s, the frequency-domain representation is:

$$X_s(u,v) = \sum_{m=0}^{H-1}\sum_{n=0}^{W-1} x_s(m,n)\, e^{-j2\pi\left(\frac{um}{H}+\frac{vn}{W}\right)}, \tag{2.1}$$

where $x_s(m,n)$ denotes the input feature at spatial coordinate (m,n) for scale s, (u,v) are frequency coordinates, and H, W are the height and width.

The low- and high-frequency components are combined with gating:

$$\hat{x}^s = \gamma^s \cdot x^s_{\text{low}} + (1-\gamma^s)\cdot x^s_{\text{high}}, \tag{2.2}$$

where $x^s_{\text{low}}, x^s_{\text{high}}$ are low/high-frequency parts, and $\gamma^s = \sigma(W_g * [x^s_{\text{low}}, x^s_{\text{high}}])$ is the gating weight.

Channel attention is applied:

$$w = \sigma(W_2\,\text{ReLU}(W_1 z)), \quad x_{\text{freq}} = w \odot x_{\text{fuse}}, \tag{2.3}$$

where z is the global pooled channel descriptor, W_1, W_2 are projection matrices, w is the channel attention vector, and $\odot$ denotes channel-wise multiplication.

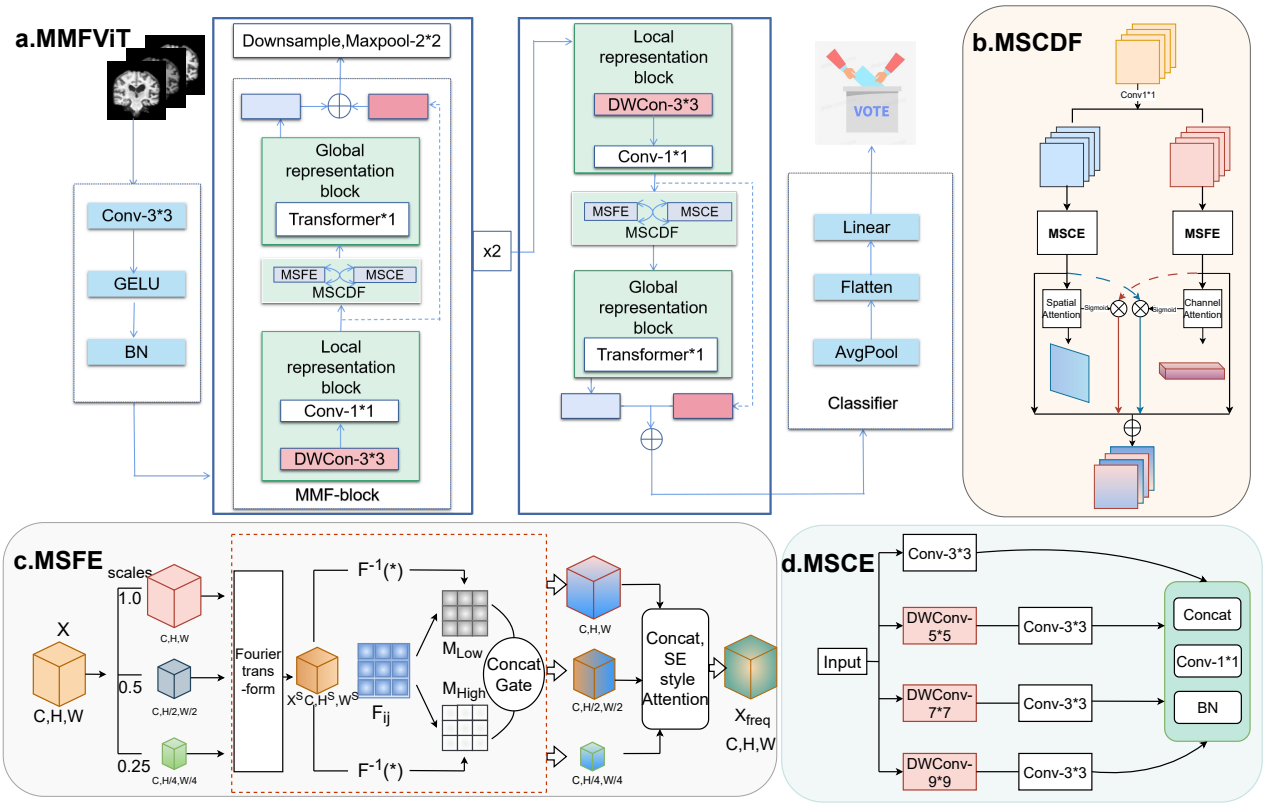

Figure 1. Illustration of the overall architecture of Multi-Scale Multi-Domain Frequency-Aware Vision Transformer and its core components. (a) Overall architecture of multi-scale multi-domain frequency-aware vision Transformer. (b) The multi-scale cross-domain fusion module, which enables effective feature fusion between spatial and frequency domains. (c) The multi-scale frequency encoder module, which captures and encodes detailed structural information from the frequency domain at multiple scales, enhancing sensitivity to subtle brain changes. (d) The multi-scale context encoder module, which aggregates contextual information from the spatial domain at multiple scales to improve regional discrimination.

The proposed MSFE module decomposes input features into frequency components across multiple resolutions. It leverages spatial frequency separation through multi-resolution Fourier transforms and applies adaptive gating mechanisms to model fine-grained details and structural patterns. Subsequently, channel attention is employed to further enhance the frequency-aware representations. This design

enriches the network's capacity to capture diverse frequency characteristics and proves particularly effective for tasks involving texture reconstruction and fine-structure enhancement.

2.1.2. Multi-scale context encoder (MSCE)

To effectively capture both fine-grained structural abnormalities and large-scale spatial variations in Alzheimer's Disease (AD) brain MRI, we introduce the MSCE module. The MSCE architecture is illustrated in Figure 1(d). This component is designed to extract spatial features across multiple receptive fields through parallel convolutional branches with varying kernel sizes.

Four convolutional branches with kernel sizes 3×3, 5×5, 7×7, and 9×9 are used. Their outputs are concatenated and fused as:

$$x_{\text{spatial}} = \text{BN}(W_{1\times1}[b_1, b_2, b_3, b_4]), \tag{2.4}$$

where b_k denotes the output of the k-th branch, $W_{1\times1}$ is a 1×1 convolution kernel, and BN is batch normalization.

MRI scans of Alzheimer's patients often exhibit both localized structural degeneration (e.g., hippocampal atrophy, cortical thinning) and global deformation (e.g., ventricular enlargement). To model spatial information at multiple scales, the MSCE module incorporates parallel convolutional branches: the 3×3 branch captures local textures, the 5×5 and 7×7 branches extract regional context, and the 9×9 branch enhances global structural perception. These multi-scale spatial features provide rich structural cues that support the subsequent cross-domain fusion with frequency-domain features, improving the model's capacity to represent complex brain alterations.

2.1.3. Complementary fusion via spatial and channel attention

After frequency and spatial features are extracted by the MSFE and MSCE modules, the model must further emphasize informative cues while suppressing irrelevant or redundant responses. To achieve this, we incorporate both spatial and channel attention mechanisms within the MSCDF module to facilitate complementary feature enhancement prior to final fusion. The architecture is illustrated in Figure 1(b).

These two attention types operate along distinct dimensions—spatial and channel—yet collaboratively enhance the discriminative power of the fused representation. Spatial attention strengthens frequency-derived features by focusing on region-specific activations, while channel attention refines spatial features by adaptively reweighting semantic channels, leading to a more balanced and informative joint feature map.

The spatial attention mechanism takes the output from the spatial convolution branch, denoted as x_{spatial}, and generates a 2D attention map that highlights important spatial locations. This is achieved via a 1×1 convolution for channel compression, followed by batch normalization and a sigmoid activation function:

$$x_{\text{spatial\_att}} = \sigma(\text{BN}(W_s * x_{\text{spatial}})), \tag{2.5}$$

where W_s denotes the learnable weight matrix of the 1×1 convolution used for channel compression. $\text{BN}(\cdot)$ stands for batch normalization, which normalizes activations across a mini-batch to stabilize and accelerate training. $\sigma(\cdot)$ represents the sigmoid activation function, which maps input values into the range $(0, 1)$ and is used to generate attention weights.

The resulting attention map guides the model to focus on regions of anatomical interest, which is especially useful for identifying structural alterations in MRI scans relevant to AD diagnosis, such as cortical thinning or ventricle enlargement.

Simultaneously, the channel attention mechanism is applied to the frequency-based feature map x_{freq}, aiming to capture the relative importance of different frequency channels. A depthwise convolution is first used to encode local channel-wise patterns, followed by global average pooling to aggregate spatial context. The compressed representation is then activated through a sigmoid function to produce the channel-wise attention weights:

$$x_{\text{channel\_att}} = \sigma(\text{GAP}(\text{DWConv}(x_{\text{freq}}))), \tag{2.6}$$

where DWConv$(\cdot)$ denotes a depthwise convolution that applies independent convolutional filters to each channel, enabling local feature extraction per channel. GAP$(\cdot)$ represents global average pooling, which computes the mean activation across the spatial dimensions $H \times W$, yielding a compact channel descriptor. $\sigma(\cdot)$ is the sigmoid activation function, mapping values into the range $(0, 1)$ to produce channel-wise attention weights.

This process enables the model to adaptively amplify channels that are more informative for classification, such as those encoding structural edges or frequency-domain texture variations.

To combine the frequency and spatial features effectively, we introduce an interaction-enhanced fusion strategy. In addition to directly adding the two feature maps, we include two multiplicative attention-guided terms: one that weights frequency features by spatial attention, and another that weights spatial features by channel attention. The final fused representation is computed as:

$$X_{\text{MMF}} = x_{\text{freq}} + x_{\text{spatial}} + x_{\text{freq}} \cdot x_{\text{spatial\_att}} + x_{\text{spatial}} \cdot x_{\text{channel\_att}}. \tag{2.7}$$

Here, the element-wise multiplication terms encode explicit inter-branch interactions, allowing spatial and channel contexts to reinforce one another. Compared to simple addition or concatenation, this formulation better preserves discriminative details and structural coherence in the final representation.

In the context of MRI-based classification tasks, such as distinguishing AD from cognitively normal (CN) subjects or predicting conversion from stable MCI (sMCI) to progressive MCI (pMCI), this attention-guided fusion significantly improves the model's focus on critical anatomical regions. The dual-attention strategy ensures that the resulting features are both context-aware and semantically rich, thereby enhancing the model's performance in structural brain analysis.

2.2. Majority voting strategy

In this study, we adopt a majority voting strategy to obtain subject-level predictions by aggregating slice-level classification results. Each subject is associated with a set of 2D MRI slices (30 slices in our case), and each slice is independently processed by the model to produce a class prediction. Since the clinical diagnosis should consider comprehensive evidence from multiple views, we determine the final subject-level prediction by selecting the class that receives the highest number of votes among all slice-level outputs.

Algorithm 1: End-to-end pipeline of MMF-ViT with MSCDF and majority voting

Input: Subject-level 2D MRI slices $\{x_j^{(i)}\}_{j=1}^{n}$ for subject i.

Output: Subject-level prediction $\hat{Y}^{(i)} \in \{0, 1\}$.

1) Slice-level feature extraction and embedding

for $j = 1, \ldots, n$ **do**

Obtain initial embedding via Conv → BN → GELU.

2) MSFE: multi-scale frequency encoding

(a) Resize the feature map to scales $s \in \{1.0, 0.5, 0.25\}$ (bilinear) to obtain $x^s \in \mathbb{R}^{B\times C\times H_s\times W_s}$, then apply 2D FFT and radial masks with low-frequency threshold $\tau = 0.1$: $x^s_{\text{low}} = \mathcal{F}^{-1}(X^s \odot \mathcal{M}^s_{\text{low}})$, $x^s_{\text{high}} = \mathcal{F}^{-1}(X^s \odot (1-\mathcal{M}^s_{\text{low}}))$.

(b) Gated fusion per scale: $\gamma^s = \sigma(W_g * [x^s_{\text{low}}, x^s_{\text{high}}])$, $\hat{x}^s = \gamma^s \odot x^s_{\text{low}} + (1-\gamma^s) \odot x^s_{\text{high}}$, with $W_g \in \mathbb{R}^{C\times(2C)\times1\times1}$.

(c) Each scale applies two 3×3 convs (ReLU), upsample to (H, W), concatenate across scales to $y \in \mathbb{R}^{B\times(3C)\times H\times W}$, then fuse: $x_{\text{fuse}} = \text{BN}(W_{\text{fuse}} * y)$, with $W_{\text{fuse}} \in \mathbb{R}^{C\times(3C)\times1\times1}$. Apply SE-style channel attention: $z = \text{GAP}(x_{\text{fuse}})$, $w = \sigma(W_2\,\text{ReLU}(W_1 z))$, $x_{\text{freq}} = w \odot x_{\text{fuse}}$, where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, $r = 16$.

3) MSCE: multi-scale context encoding

Extract features using four parallel branches with kernels 3×3, 5×5, 7×7, 9×9; for 5/7/9 branches, use depthwise ($k \times k$, groups $= C$) + pointwise (1×1) conv, followed by an extra 3×3 conv (ReLU).

Concatenate and fuse via 1×1 Conv + BN: $x_{\text{spatial}} = \text{BN}(W_{1\times1}[b_3, b_5, b_7, b_9])$, with $W_{1\times1} \in \mathbb{R}^{C\times(4C)\times1\times1}$.

4) Complementary attention & fusion (MSCDF)

Spatial attention from x_{spatial}: $x_{\text{spatial\_att}} = \sigma(\text{BN}(W_s * x_{\text{spatial}}))$, where $W_s \in \mathbb{R}^{1\times1\times C\times1}$.

Channel attention from x_{freq}: $x_{\text{channel\_att}} = \sigma(\text{GAP}(\text{DWConv}_{k=3}(x_{\text{freq}})))$, DWConv: groups $= C$, kernel $= 3$.

Final fusion: $X_{\text{MMF}} = x_{\text{freq}} + x_{\text{spatial}} + x_{\text{freq}} \odot x_{\text{spatial\_att}} + x_{\text{spatial}} \odot x_{\text{channel\_att}}$.

5) Slice-level prediction

Apply GAP → FC → Softmax to obtain $\hat{y}_j^{(i)}$.

6) Majority voting (subject-level)

Aggregate $\{\hat{y}_j^{(i)}\}$ by majority voting to get $\hat{Y}^{(i)}$.

where: $x_j^{(i)}$: the j-th slice of subject i; MSFE: Multi-Scale Frequency Encoder; MSCE: Multi-Scale Context Encoder; MSCDF: Complementary fusion module; BN: Batch Normalization; GAP: Global Average Pooling; FC: Fully Connected layer.

Hyper-parameters / dimensions (as implemented)

Scales $\{1.0, 0.5, 0.25\}$ (bilinear); low-frequency threshold $\tau = 0.1$; $W_g \in \mathbb{R}^{C\times(2C)\times1\times1}$; $W_{\text{fuse}} \in \mathbb{R}^{C\times(3C)\times1\times1}$; SE reduction $r = 16$ with $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$; spatial attention $W_s \in \mathbb{R}^{1\times1\times C\times1}$; DWConv kernel $k = 3$, groups $= C$; input split via $W_e \in \mathbb{R}^{(2C)\times C\times1\times1}$ then channel-wise halving to MSFE/MSCE.

Let a subject i have n slices denoted as $\{x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\}$, with corresponding predicted class labels $\{\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \ldots, \hat{y}_n^{(i)}\}$, where $\hat{y}_j^{(i)} \in \{0, 1\}$ represents the predicted class of the j-th slice (e.g., 0 for control and 1 for Alzheimer's disease).

The final subject-level prediction $\hat{Y}^{(i)}$ is defined as:

$$\hat{Y}^{(i)} = \arg\max_{c \in \{0,1\}} \sum_{j=1}^{n} \mathbb{I}(\hat{y}_j^{(i)} = c), \tag{2.8}$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the condition holds and 0 otherwise. This strategy effectively aggregates all slice-level predictions and assigns the subject to the class with the maximum number of votes.

The majority voting strategy aggregates multi-slice information, effectively reducing the impact of misclassified slices, enhancing robustness and stability at the subject level, and aligning well with the clinical diagnostic process, which relies on comprehensive assessment across multiple views.

2.3. Overall pipeline of MMF-ViT

To provide a concise overview of the proposed model, we summarize the entire end-to-end process of MMF-ViT in Algorithm 1. This algorithm highlights the step-by-step pipeline from slice-level feature extraction, through the MSCDF module, to subject-level prediction via majority voting. While the detailed mathematical formulations of MSFE, MSCE, and attention mechanisms are presented in the preceding subsections, this algorithm offers a compact representation of the workflow for better clarity.

3. Experiments

This section introduces the dataset, preprocessing steps, and experimental setup, followed by results on multiple classification tasks. We report comparisons with 2D and 3D baselines, ablation studies, slice-direction analysis, and interpretability evaluation, and conclude with a discussion of the findings.

3.1. Dataset

The imaging data used in this study were obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database [22]. ADNI is a large-scale, multi-phase study (ADNI-1, GO, 2, and 3) launched in 2004, aiming to facilitate early diagnosis, disease monitoring, and treatment evaluation of AD and its prodromal stage, MCI, through the collection of imaging, biomarkers, and clinical assessments.

MCI subjects are categorized based on their conversion status within a 36-month follow-up: those who progress to AD are labeled as progressive MCI (pMCI), and the rest as stable MCI (sMCI). In this study, we select T1-weighted MRI scans from the ADNI dataset, including 260 pMCI, 288 sMCI, 362 AD, and 436 cognitively normal (CN) subjects. Summary statistics are presented in Table 1.

Table 1. Demographic and clinical characteristics of all groups.

Group	Number	Male/Female	Age (mean ± std)	MMSE (mean ± std)	APOE4 (-/+)
AD	362	188/174	74.91 ± 7.78	21.98 ± 3.96	113/249
CN	436	212/224	74.66 ± 5.78	28.65 ± 1.84	303/133
sMCI	288	181/107	72.54 ± 7.70	27.31 ± 2.69	171/117
pMCI	260	157/103	73.81 ± 7.05	24.49 ± 3.88	83/177

3.2. Data preprocessing pipeline

3.2.1. Data preprocessing

Following a standard preprocessing pipeline [4, 23], we processed MRI data using FreeSurfer [24] and registered them to the MNI152 template space via the FSL package. The MRI preprocessing involved the following five steps:

1) Alignment to the anterior and posterior commissure (AC-PC line);

2) Intensity non-uniformity correction;

3) Skull stripping;

4) Registration to the MNI152 template space;

5) Outlier clipping and z-score normalization.

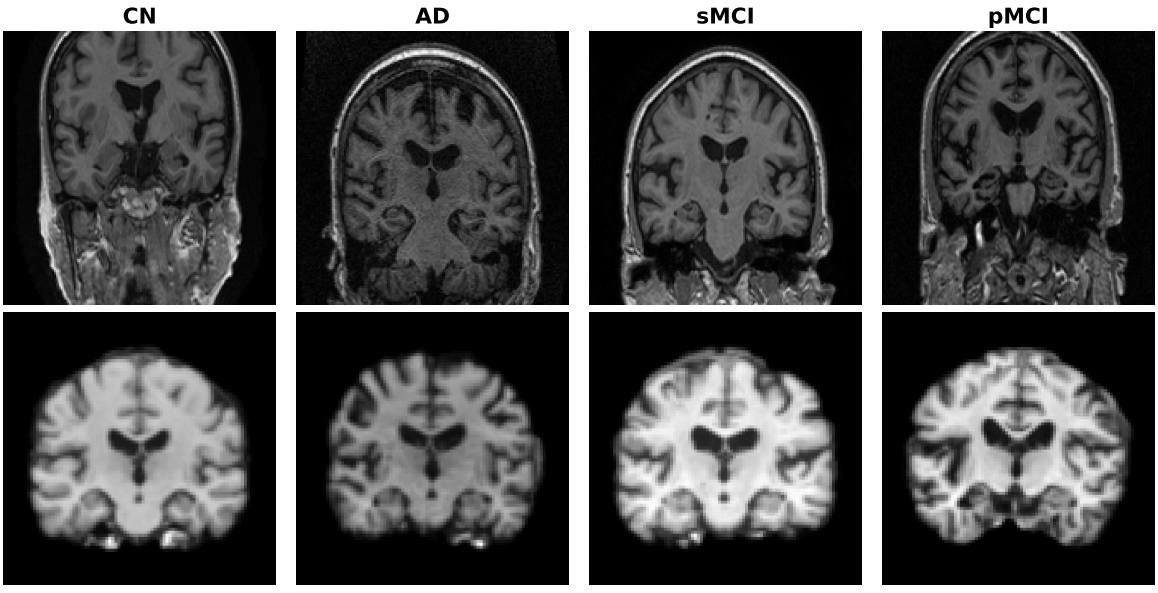

Figure 2. Representative coronal slices of brain MRI. Each column corresponds to one group: cognitively normal (CN), Alzheimer's disease (AD), stable mild cognitive impairment (sMCI), and progressive mild cognitive impairment (pMCI). The top row shows the original slices, and the bottom row shows the preprocessed slices.

The first three preprocessing steps were performed using FreeSurfer, and the registration was implemented with the FLIRT tool from the FSL package. A comparison of the MRI volumes before and after preprocessing is shown in Figure 2. Z-score normalization plays an important role in MRI analysis, as it reduces inter-subject intensity variation, ensures comparability across scans, and stabilizes training, thereby improving the generalization ability of the model.

3.2.2. 3D to 2D conversion

Due to the limited number of medical imaging samples, directly applying high-capacity Transformer models tends to result in overfitting. To mitigate this issue, we extracted 30 central slices from each subject's 3D T1-weighted MRI along the three orthogonal planes: sagittal, axial, and coronal. This strategy not only serves as an effective form of data augmentation, but also significantly reduces the input dimensionality and model complexity, thereby alleviating the risk of overfitting.

In all experiments, we adopted the coronal plane as the primary input view, as it simultaneously captures critical brain structures such as the cortex, ventricles, and hippocampus [25]. In addition, slices from the other two directions were retained for subsequent comparative experiments to comprehensively assess the modeling capability across different views.

3.2.3. Dataset split

To ensure fair evaluation and strong generalization capability, we employed a 5-fold cross-validation strategy for all experiments. The data splitting was strictly performed at the subject level, meaning that all slices belonging to the same subject appear exclusively in either the training or validation set. This protocol effectively prevents data leakage caused by inter-slice redundancy and ensures unbiased performance assessment. At the same time, we also set the random seed to ensure that the data splitting process is reproducible.

3.3. Experimental setup

All experiments were conducted on a server running Ubuntu 22.04, equipped with six NVIDIA L40S GPUs. The models were implemented using PyTorch 2.6.0 with CUDA 12.4 support. Training was performed using the Adam optimizer with a batch size of 128. The initial learning rate was 0.001 and scheduled dynamically using the CosineAnnealingLR strategy. Model performance was evaluated using 5-fold cross-validation. The evaluation metrics included Accuracy, Precision, Recall, F1-score, and AUC, and the final results were reported as the mean ± standard deviation across the five folds.

For model initialization, we followed the default scheme of PyTorch: convolutional layers are initialized with Kaiming normal initialization, and biases (if present) are set to zero; the scaling factors of batch normalization layers are initialized to 1, and the biases are initialized to 0. We did not adopt any additional customized initialization methods.

3.4. Comparison with 2D and 3D baselines

To evaluate the effectiveness of the proposed method, we conducted comparative experiments with representative models in computer vision and medical image analysis, including ResNet-18 [26], ShuffleNetV2 [27], MobileNetV2 [28], EfficientNet-B0 [29], and RepVGG [30], as well as two lightweight Transformer-based models: EfficientViT [31] and MobileViT [32, 33]. Additionally, we

extended the comparison to 3D models using full MRI volumes, including 3D EfficientNet-B0 [29], 3D ResNet18 [26], 3D CoAtNet-0 [34], and 3D MaxViT-T [35].

All models were trained and evaluated under the same dataset and experimental settings to ensure fairness. It should be emphasized that although many studies on AD classification are based on the same public dataset (e.g., ADNI), the exact data selection and partition strategies vary across works, and many state-of-the-art models do not provide open-source implementations, making direct reproduction under identical conditions difficult. To ensure reproducibility and fair comparison, we focused on classical and open-source models as baselines, and retrained all models from scratch under a unified dataset and five-fold cross-validation framework. This strategy guarantees the fairness and consistency of our comparative evaluation. In addition, we report the number of parameters (Params) and floating-point operations (FLOPs) for each model (Tables 2 and 3), which are widely recognized indicators of computational efficiency.

Table 2. Performance comparison of 2D and 3D models for sMCI vs. pMCI classification with majority voting.

Model	Params (M)	FLOPs (G)	ACC	Precision	Recall	F1	AUC
2DResNet-18 [26]	11.169	4.673	69.72 ± 3.12	69.71 ± 5.97	66.18 ± 8.33	67.25 ± 2.46	69.56 ± 2.70
2DShuffleNetV2 [27]	1.255	0.027	70.83 ± 4.24	71.63 ± 4.99	62.97 ± 8.44	66.81 ± 6.26	70.36 ± 4.42
2DMobileNetV2 [28]	2.226	0.057	69.72 ± 3.33	68.61 ± 6.41	68.08 ± 3.26	68.04 ± 1.88	69.63 ± 2.99
2DEfficientNet-B0 [29]	3.968	0.069	70.28 ± 4.98	70.36 ± 5.71	63.79 ± 6.25	66.87 ± 5.86	69.92 ± 5.05
2DRepVGG [30]	26.761	1.034	69.36 ± 4.33	76.60 ± 7.03	50.54 ± 8.75	60.52 ± 7.26	68.32 ± 4.57
2DEfficientViT [31]	2.131	0.020	69.72 ± 2.78	70.06 ± 3.04	62.65 ± 6.31	65.98 ± 4.07	69.35 ± 2.93
2DMobileViTv1 [32]	4.926	0.260	71.01 ± 4.66	70.15 ± 7.66	71.55 ± 9.83	69.86 ± 3.81	71.02 ± 4.16
2DMobileViTv2 [33]	1.102	0.065	69.72 ± 3.80	72.38 ± 8.92	62.22 ± 11.63	65.41 ± 6.41	69.29 ± 3.77
3DEfficientNet-B0 [29]	4.537	2.216	67.33 ± 3.39	65.63 ± 3.68	65.77 ± 7.63	65.46 ± 4.73	67.29 ± 4.31
3DResNet-18 [26]	33.161	31.521	65.33 ± 1.25	66.91 ± 3.60	54.62 ± 7.15	59.67 ± 3.43	64.21 ± 2.51
3DCoAtNet-0 [34]	26.659	29.403	66.06 ± 2.23	69.39 ± 9.64	58.85 ± 20.15	60.21 ± 10.35	67.84 ± 2.24
3DMaxViT-T [35]	30.468	53.890	62.59 ± 2.87	62.24 ± 4.02	55.38 ± 12.72	57.75 ± 6.49	64.26 ± 4.43
Ours (MMF-ViT)	2.469	0.541	72.84 ± 4.00	70.25 ± 5.69	75.48 ± 5.65	72.43 ± 3.04	72.99 ± 3.68

Table 3. Performance comparison of 2D and 3D models for AD vs. CN classification with majority voting.

Model	Params (M)	FLOPs (G)	ACC	Precision	Recall	F1	AUC
2DResNet-18 [26]	11.169	4.673	84.72 ± 2.11	84.57 ± 4.87	88.77 ± 3.78	86.42 ± 1.49	84.32 ± 2.49
2DShuffleNetV2 [27]	1.255	0.027	84.84 ± 2.05	84.80 ± 1.81	88.08 ± 3.33	86.37 ± 1.96	84.52 ± 1.99
2DMobileNetV2 [28]	2.226	0.057	85.22 ± 1.72	85.03 ± 2.97	88.77 ± 3.48	86.77 ± 1.55	84.87 ± 1.82
2DEfficientNet-B0 [29]	3.968	0.069	82.46 ± 3.81	81.08 ± 3.67	88.76 ± 5.31	84.65 ± 3.52	81.83 ± 3.81
2DRepVGG [30]	26.761	1.034	82.71 ± 2.24	81.00 ± 3.99	89.91 ± 4.26	85.05 ± 1.65	81.99 ± 2.58
2DEfficientViT [31]	2.131	0.020	79.70 ± 3.21	79.48 ± 3.39	84.86 ± 2.88	82.05 ± 2.77	79.18 ± 3.29
2DMobileViTv1 [32]	4.926	0.260	81.96 ± 3.16	82.60 ± 2.09	84.85 ± 5.72	83.62 ± 3.34	81.65 ± 2.96
2DMobileViTv2 [33]	1.102	0.065	79.96 ± 3.59	80.56 ± 2.84	83.47 ± 5.73	81.91 ± 3.59	79.58 ± 3.46
3DEfficientNet-B0 [29]	4.537	2.216	73.56 ± 1.71	73.50 ± 2.57	81.21 ± 6.14	76.96 ± 2.04	77.34 ± 2.61
3DResNet-18 [26]	33.161	31.521	77.19 ± 1.19	76.87 ± 1.15	83.50 ± 4.95	79.94 ± 1.70	81.89 ± 2.24
3DCoAtNet-0 [34]	26.659	29.403	79.08 ± 5.80	80.93 ± 6.15	81.19 ± 8.21	80.81 ± 5.75	83.92 ± 6.19
3DMaxViT-T [35]	30.468	53.890	81.20 ± 2.40	79.79 ± 4.40	88.54 ± 4.15	83.75 ± 1.75	86.14 ± 2.56
Ours (MMF-ViT)	2.469	0.541	85.59 ± 1.51	86.52 ± 3.30	87.61 ± 5.37	86.87 ± 1.66	85.39 ± 1.43

As shown in Tables 2 and 3, our MMF-ViT consistently outperformed all baselines on both AD vs. CN and sMCI vs. pMCI classification tasks, achieving the best or highly competitive results in key

metrics such as accuracy, F1-score, and AUC. Meanwhile, MMF-ViT achieved this performance with only 2.47 M parameters and 0.54 G FLOPs, which confirms its computational efficiency. Notably, MMF-ViT demonstrated robust performance even on the more challenging sMCI vs. pMCI task, indicating its strong capability in capturing subtle structural differences. The relatively small performance gaps across models are largely attributed to our use of rigorous five-fold cross-validation, which effectively mitigates data partition bias and enhances the stability and comparability of results. These findings further confirm the generalization ability and feature modeling strength of MMF-ViT under a consistent evaluation framework.

In the sMCI vs. pMCI classification task (Table 2), MMF-ViT achieved an accuracy (ACC) of 72.84%, a recall of 75.48%, and an AUC of 72.99%, all of which are the highest among the evaluated models. Compared to the second-best model, 2DMobileViTv1 (with an accuracy of 71.01% and an AUC of 71.02%), MMF-ViT achieved improvements of 1.83 and 1.97 percentage points, respectively, demonstrating its clear advantage in tackling this challenging and fine-grained classification task.

In the AD vs. CN classification task (Table 3), MMF-ViT also achieved the best performance across multiple key metrics, including an accuracy (ACC) of 85.59%, an F1-score of 86.87%, and a precision of 86.52%. All these values are higher than those of competing models.

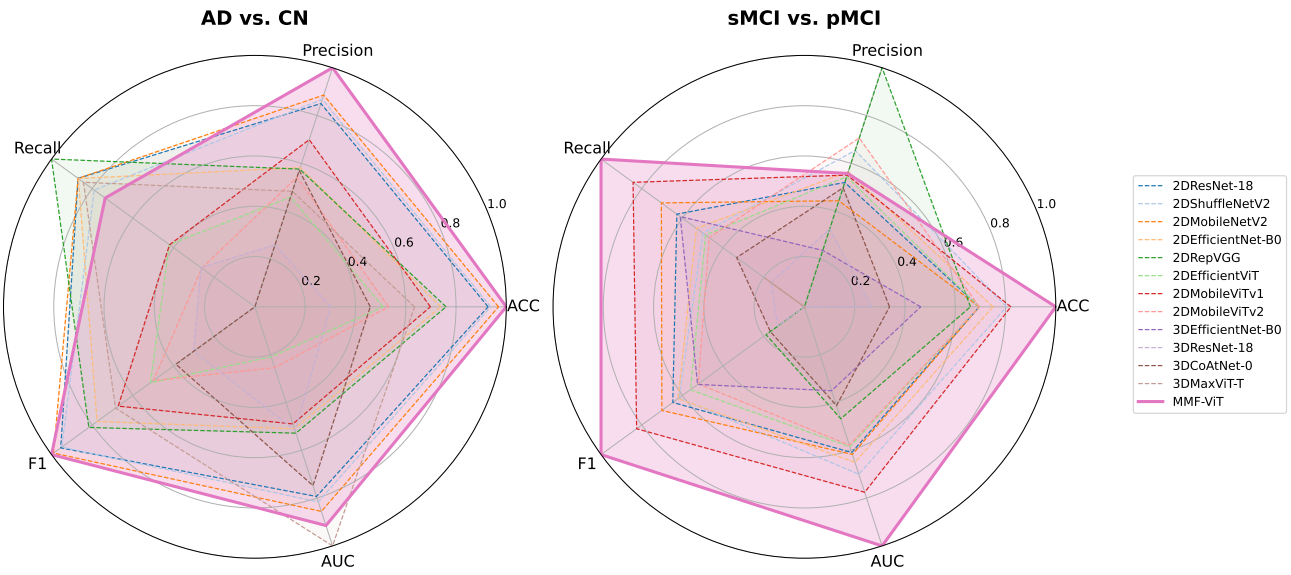

Figure 3. Model performance across AD vs. CN and sMCI vs. pMCI classification tasks.

It is worth noting that although 3D models are designed to leverage the complete volumetric information of MRI data, our experimental results show that they generally perform worse than lightweight 2D models. In particular, 3D MaxViT and 3D CoAtNet, despite incorporating more sophisticated attention mechanisms, fail to demonstrate clear advantages. This is mainly due to the higher risk of overfitting under a relatively limited dataset size, since their parameter counts and computational complexity are substantially larger. In contrast, 2D methods combined with majority

voting tend to provide more stable performance improvements. Looking ahead, hybrid 2.5D approaches may offer a promising compromise by capturing volumetric context while alleviating overfitting risks, and they could be further explored in conjunction with our multi-scale spatial- and frequency-domain fusion strategy to enhance discriminative power.

The comparative performance of all models is comprehensively illustrated by the two radar charts presented in Figure 3, which provide a visual summary of the evaluation metrics across the respective classification tasks.

Figure 4 presents the confusion matrices of MMF-ViT for the AD vs. CN (left) and sMCI vs. pMCI (right) classification tasks, providing a detailed illustration of the distribution between the predicted results and the ground truth labels in both experiments.

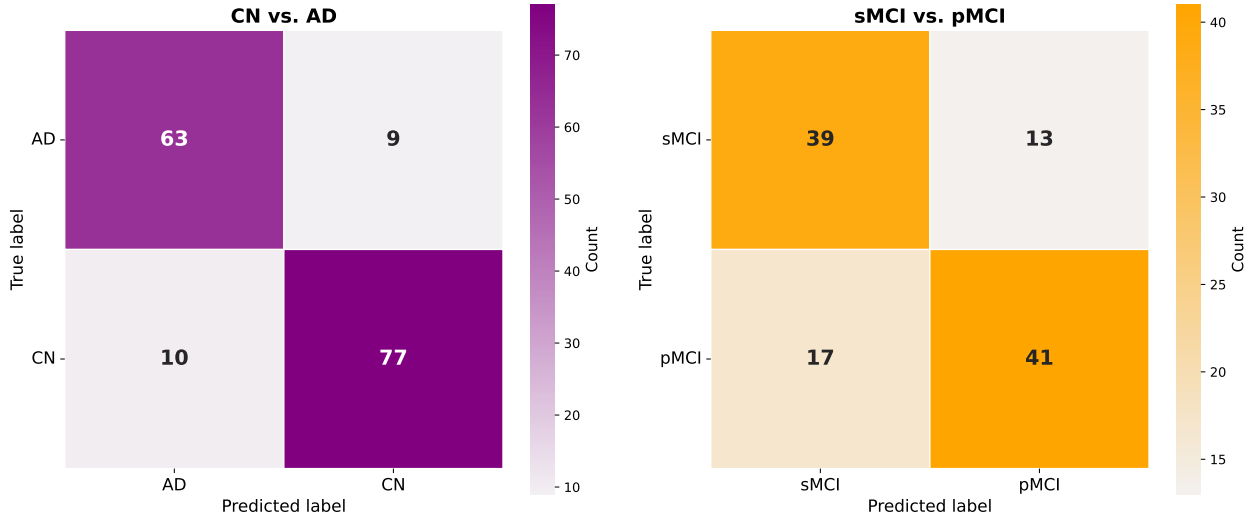

Figure 4. Confusion matrices of MMF-ViT on AD vs. CN (left) and sMCI vs. pMCI (right) classification tasks.

Table 4. Class-specific Precision, Recall, and F1-scores of MMF-ViT on AD vs. CN and sMCI vs. pMCI classification tasks.

Task	Class	Precision	Recall	F1-score
AD vs. CN	AD	86.30	87.50	86.90
AD vs. CN	CN	89.53	88.51	89.02
sMCI vs. pMCI	sMCI	69.64	75.00	72.22
sMCI vs. pMCI	pMCI	75.93	70.69	73.21

Table 4 presents the class-specific precision, recall, and F1-scores for the AD vs. CN and sMCI vs. pMCI classification tasks. The results show that for AD vs. CN, the model achieves high and balanced performance across both classes, reflecting reliable discrimination ability. For sMCI vs.

pMCI, although the performance is relatively lower, the metrics remain balanced between sMCI and pMCI. Importantly, the recall for pMCI does not exhibit abnormal inflation, suggesting that the model maintains sensitivity without bias toward one class.

3.5. Ablation study

To evaluate the effectiveness of each component in our model, we conducted ablation experiments on CN vs. AD and sMCI vs. pMCI classification tasks, as shown in Tables 5 and 6. Starting from a baseline without frequency-domain or spatial-domain modules, we first added the MSFE, which improved performance on both tasks, highlighting the value of frequency-based spatial representations. Adding the MSCE, which integrates multi-scale convolutions with spatial and channel attention, also led to further gains, especially in recall for the AD classification.

Table 5. Results of the ablation study for sMCI vs. pMCI classification.

Model	MSFE	MSCE	Cross	ACC	Precision	Recall	F1	AUC
None				69.36 ± 3.31	71.32 ± 6.44	61.11 ± 8.91	65.09 ± 3.91	68.94 ± 3.09
Only MSFE	✓			70.64 ± 3.07	68.88 ± 4.18	69.62 ± 5.97	69.03 ± 3.43	70.58 ± 3.01
Only MSCE		✓		70.28 ± 4.63	72.59 ± 7.28	61.45 ± 14.07	65.33 ± 8.62	69.79 ± 4.83
Both without cross	✓	✓		71.38 ± 4.55	73.47 ± 6.39	62.27 ± 6.99	67.14 ± 5.34	70.91 ± 4.59
Ours (MMF-ViT)	✓	✓	✓	72.84 ± 4.00	70.25 ± 5.69	75.48 ± 5.65	72.43 ± 3.04	72.99 ± 3.68

Table 6. Results of the ablation study for CN vs. AD classification.

Model	MSFE	MSCE	Cross	ACC	Precision	Recall	F1	AUC
None				82.71 ± 1.52	82.03 ± 1.95	87.62 ± 2.21	84.70 ± 1.37	82.21 ± 1.58
Only MSFE	✓			84.34 ± 1.50	83.32 ± 1.15	89.22 ± 3.44	86.13 ± 1.57	83.84 ± 1.34
Only MSCE		✓		84.72 ± 2.35	83.31 ± 2.45	90.13 ± 2.49	86.57 ± 2.07	84.16 ± 2.40
Both without cross	✓	✓		85.09 ± 2.15	85.47 ± 2.71	87.84 ± 5.33	86.50 ± 2.21	84.80 ± 2.03
Ours (MMF-ViT)	✓	✓	✓	85.59 ± 1.51	86.52 ± 3.30	87.61 ± 5.37	86.87 ± 1.66	85.39 ± 1.43

We then combined MSFE and MSCE without cross-path interaction, using simple concatenation and a 1×1 convolution, which continued to improve results, indicating the complementarity of the two feature types. Finally, enabling full cross-branch attention-based interaction formed the proposed MSCDF module. The resulting MMF-ViT model achieved the best performance on both tasks, with recall reaching 75.48% in the sMCI vs. pMCI task. These results confirm the importance of each component and demonstrate that effective integration and interaction between branches are crucial for improving model robustness and discriminative power.

3.6. Comparison across slice directions

We evaluated the performance of MMF-ViT across three common slice directions: sagittal, axial, and coronal. The results are summarized in Tables 7 and 8.

Table 7. Comparison of classification performance across three directions on sMCI vs. pMCI using MMF-ViT with majority voting.

Direction	ACC	Precision	Recall	F1	AUC
Axial	67.89 ± 2.39	68.92 ± 2.39	67.15 ± 2.72	66.73 ± 3.13	67.76 ± 3.42
Coronal	72.84 ± 4.00	70.25 ± 5.69	75.48 ± 5.65	72.43 ± 3.04	72.99 ± 3.68
Sagittal	69.17 ± 2.22	70.09 ± 2.76	68.83 ± 2.68	68.49 ± 2.50	70.80 ± 4.02

Table 8. Comparison of classification performance across three directions on CN vs. AD using MMF-ViT with majority voting.

Direction	ACC	Precision	Recall	F1	AUC
Axial	80.96 ± 2.28	81.06 ± 2.24	80.68 ± 2.67	80.69 ± 2.44	86.00 ± 2.95
Coronal	85.59 ± 1.51	86.52 ± 3.30	87.61 ± 5.37	86.87 ± 1.66	85.39 ± 1.43
Sagittal	81.08 ± 1.91	81.60 ± 2.03	80.43 ± 1.90	80.65 ± 1.94	86.92 ± 2.17

Coronal slices consistently achieved the best performance on both tasks, suggesting a superior ability to capture key anatomical structures such as the hippocampus, cortex, and ventricles. These findings highlight the impact of slice orientation on classification outcomes.

3.7. Interpretability analysis

To enhance the interpretability of the model predictions, we employed Grad-CAM (gradient-weighted class activation mapping) to visualize the decision-making process of the model. We selected the 5th, 15th, and 25th coronal slices from each subject as representatives, and presented the results of four typical subjects (AD, CN, pMCI, and sMCI) to intuitively compare the model's attention regions.

As shown in Figure 5, in AD subjects, the model mainly focused on the hippocampus, entorhinal cortex, and adjacent temporal cortex, which are known to undergo atrophy in the early stages of Alzheimer's disease, consistent with neuroimaging and pathological findings. In contrast, the attention of CN subjects was more dispersed without obvious focus. For pMCI subjects, the high-response regions were also concentrated in the hippocampus and temporal cortex, suggesting their association with the risk of AD progression.

This interpretability analysis indicates that the model indeed relies on anatomical regions closely related to AD progression in the classification task, rather than random or irrelevant features. This not only enhances the credibility of the model predictions but also provides support for its potential clinical application.

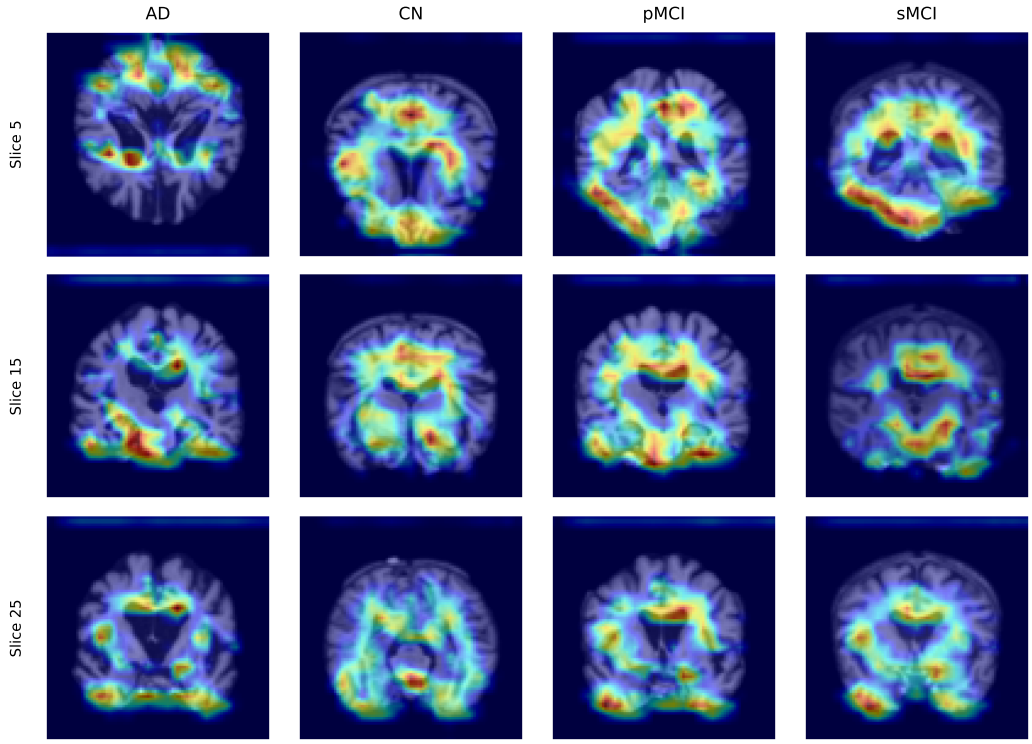

Figure 5. Grad-CAM visualization results of representative AD, CN, pMCI, and sMCI subjects on the 5th, 15th, and 25th coronal slices.

3.8. Discussion

Overall, MMF-ViT achieves strong performance on both AD vs. CN and sMCI vs. pMCI classification tasks, with particularly notable gains in the more challenging sMCI vs. pMCI setting. This improvement can be attributed to the cross-domain fusion mechanism between multi-scale frequency and spatial branches, which enhances both feature representation and discriminative capability.

Moreover, experiments across different slice directions reveal that the coronal view consistently yields the best results, confirming its advantage in preserving critical anatomical structures relevant to Alzheimer's disease.

Looking forward, prior studies have shown that integrating multiple cognitive domains in neuropsychological assessment substantially improves the prediction of MCI conversion [36], while combining cognitive and social functioning measures has demonstrated effectiveness in detecting MCI and dementia in Parkinson's disease [37]. Taken together, these findings highlight that multimodal and multi-domain strategies represent an essential direction for accurate early detection and subtype differentiation. Building on this evidence, our future work will aim to extend the proposed neuroimaging-based framework toward such broader multimodal contexts, integrating imaging with complementary behavioral and functional assessments to further enhance diagnostic precision.

4. Conclusions

This paper proposes a novel vision Transformer framework, MMF-ViT, for fine-grained classification of Alzheimer's disease and its prodromal stages based on structural MRI. The proposed MSCDF module integrates a multi-scale frequency encoder and a multi-scale spatial context encoder, enabling interactive feature fusion through attention mechanisms and enhancing the model's ability to identify subtle structural variations in the brain. Experimental results on the ADNI dataset demonstrate that MMF-ViT achieves competitive performance in both AD vs. CN and sMCI vs. pMCI classification tasks. Specifically, for the more challenging sMCI vs. pMCI classification, MMF-ViT attains an accuracy (ACC) of 72.84% and an AUC of 72.99%, outperforming mainstream 2D and 3D models. For AD vs. CN classification, MMF-ViT also achieves strong performance with an accuracy of 85.59% and an AUC of 85.39%.

Despite these promising results, this study is limited to the ADNI dataset and does not include robustness testing under synthetic MRI artifacts. In addition, it is restricted to unimodal structural MRI, while multimodal data such as PET, cognitive assessments, and CSF biomarkers could provide complementary value. In future work, we plan to extend validation to additional datasets (e.g., OASIS, AIBL, and local cohorts), conduct robustness experiments with synthetic artifacts (e.g., motion blur, contrast loss, noise), and integrate multimodal information to further enhance the clinical applicability of MMF-ViT.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 12372062).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. M. A. Better, 2023 Alzheimer's disease facts and figures, *Alzheimer's Dementia*, **19** (2023), 1598–1695. https://doi.org/10.1002/alz.13016
2. C. Wang, Y. Lei, T. Chen, J. Zhang, Y. Li, H. Shan, HOPE: Hybrid-granularity ordinal prototype learning for progression prediction of mild cognitive impairment, *IEEE J. Biomed. Health Inf.*, **28** (2024), 6429–6440. https://doi.org/10.1109/JBHI.2024.3357453
3. Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, Alzheimer's Disease Neuroimaging Initiative (ADNI), Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network, *Front. Neurosci.*, **13** (2019), 509. https://doi.org/10.3389/fnins.2019.00509

4. S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, et al., Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, *Brain*, **143** (2020), 1920–1933. https://doi.org/10.1093/brain/awaa137
5. X. Zhang, L. Han, W. Zhu, L. Sun, D. Zhang, An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, *IEEE J. Biomed. Health Inf.*, **26** (2022), 5289–5297. https://doi.org/10.1109/JBHI.2021.3066832
6. U. Khatri, G. R. Kwon, Diagnosis of Alzheimer's disease via optimized lightweight convolution-attention and structural MRI, *Comput. Biol. Med.*, **171** (2024), 108116. https://doi.org/10.1016/j.compbiomed.2024.108116
7. G. B. Frisoni, N. C. Fox, C. R. J. Jr, P. Scheltens, P. M. Thompson, The clinical use of structural MRI in Alzheimer disease, *Nat. Rev. Neurol.*, **6** (2010), 67–77. https://doi.org/10.1038/nrneurol.2009.215
8. W. Feng, N. V. Halm-Lutterodt, H. Tang, A. Mecum, M. K. Mesregah, Y. Ma, et al., Automated MRI-based deep learning model for detection of Alzheimer's disease process, *Int. J. Neural Syst.*, **30** (2020), 2050032. https://doi.org/10.1142/S012906572050032X
9. B. Y. Lim, K. W. Lai, K. Haiskin, K. A. S. H. Kulathilake, Z. C. Ong, Y. C. Hum, et al., Deep learning model for prediction of progressive mild cognitive impairment to Alzheimer's disease using structural MRI, *Front. Aging Neurosci.*, **14** (2022), 876202. https://doi.org/10.3389/fnagi.2022.876202
10. J. Wu, X. Zhang, Y. Li, Y. Zhang, J. Liu, C. Zheng, et al., A multi-scale feature and dual self-attention mechanism for enhanced Alzheimer's disease classification, in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, (2024), 4300–4306. https://doi.org/10.1109/BIBM62325.2024.10822167
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems (NeurIPS)*, (2017), 5998–6008.
12. J. Jang, D. Hwang, M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2022), 20686–20697. https://doi.org/10.1109/CVPR52688.2022.02006
13. Y. Huang, W. Li, Resizer swin transformer-based classification using sMRI for Alzheimer's disease, *Appl. Sci.*, **13** (2023), 9310. https://doi.org/10.3390/app13169310
14. F. Liu, H. Wang, S. N. Liang, Z. Jin, S. Wei, X. Li, et al., MPS-FFA: A multiplane and multiscale feature fusion attention network for Alzheimer's disease prediction with structural MRI, *Comput. Biol. Med.*, **157** (2023), 106790. https://doi.org/10.1016/j.compbiomed.2023.106790
15. X. Zhang, J. Wu, Y. Zhang, Y. Li, J. X. Liu, Q. Liang, et al., MSFAN: A multi-scale feature attention network for Alzheimer's disease diagnosis, in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, (2024), 5335–5342. https://doi.org/10.1109/BIBM62325.2024.10821998
16. X. Fan, H. Li, L. Liu, K. Zhang, Z. Zhang, Y. Chen, et al., Early diagnosing and transformation prediction of Alzheimer's disease using multi-scaled self-attention network on structural MRI images with occlusion sensitivity analysis, *J. Alzheimer's Dis.*, **97** (2023), 909–926. https://doi.org/10.3233/JAD-230705

17. R. F. Tian, J. N. Li, S. W. Zhang, MSCLK: Multi-scale fully separable convolution neural network with large kernels for early diagnosis of Alzheimer's disease, *Expert Syst. Appl.*, **252** (2024), 124241. https://doi.org/10.1016/j.eswa.2024.124241

18. F. Yan, L. Peng, F. Dong, K. Hirota, MCNEL: A multi-scale convolutional network and ensemble learning for Alzheimer's disease diagnosis, *Comput. Methods Programs Biomed.*, **264** (2025), 108703. https://doi.org/10.1016/j.cmpb.2025.108703

19. H. Ding, J. Lu, J. Cai, Y. Zhang, Y. Shang, SLF-UNet: Improved UNet for brain MRI segmentation by combining spatial and low-frequency domain features, in *Advances in Computer Graphics*, Springer, (2023), 415–426. https://doi.org/10.1007/978-3-031-50075-6_32

20. J. Feng, M. Jiang, H. Zhang, L. Yin, ADNI, Integrating imaging and genetic data via wavelet transform-based CNN for Alzheimer's disease classification, *Biomed. Signal Process. Control*, **104** (2025), 107583. https://doi.org/10.1016/j.bspc.2025.107583

21. R. Kushol, A. Masoumzadeh, D. Huo, S. Kalra, Y. H. Yang, Addformer: Alzheimer's disease detection from structural MRI using fusion transformer, in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, (2022), 1–5. https://doi.org/10.1109/ISBI52829.2022.9761421

22. R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, et al., Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization, *Neurology*, **74** (2010), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

23. C. Wang, S. Piao, Z. Huang, Q. Gao, J. Zhang, Y. Li, et al., Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information, *Med. Image Anal.*, **91** (2024), 103032. https://doi.org/10.1016/j.media.2023.103032

24. B. Fischl, FreeSurfer, *NeuroImage*, **62** (2012), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

25. Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, J. Yang, et al., Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning, *Front. Comput. Neurosci.*, **9** (2015), 66. https://doi.org/10.3389/fncom.2015.00066

26. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

27. N. Ma, X. Zhang, H. T. Zheng, J. Sun, ShuffleNet V2: Practical guidelines for efficient CNN architecture design, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2018), 122–138. https://doi.org/10.1007/978-3-030-01264-9_8

28. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2018), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

29. M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR, (2019), 6105–6114.

30. X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, RepVGG: Making VGG-style ConvNets great again, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 13728–13737. https://doi.org/10.1109/CVPR46437.2021.01352

31. X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, Y. Yuan, EfficientViT: Memory efficient vision transformer with cascaded group attention, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2023), 14420–14430. https://doi.org/10.1109/CVPR52729.2023.01386

32. S. Mehta, M. Rastegari, MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer, in *International Conference on Learning Representations (ICLR)*, (2022), 1–26.

33. S. Mehta, M. Rastegari, Separable self-attention for mobile vision transformers, preprint, arXiv:2206.02680.

34. Z. Dai, H. Liu, Q. V. Le, M. Tan, CoAtNet: Marrying convolution and attention for all data sizes, in *Advances in Neural Information Processing Systems (NeurIPS)*, (2021), 3965–3977.

35. Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, et al., MaxViT: Multi-axis vision transformer, in *European conference on computer vision (ECCV)*, Springer, (2022), 459–479. https://doi.org/10.48550/arXiv.2204.01697

36. S. Belleville, C. Fouquet, C. Hudon, H. T. V. Zomahoun, J. Croteau, Consortium for the Early Identification of Alzheimer's disease-Quebec, Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: A systematic review and meta-analysis, *Neuropsychol. Rev.*, **27** (2017), 328–353. https://doi.org/10.1007/s11065-017-9361-5

37. Y. W. Yu, C. H. Tan, H. C. Su, C. Y. Chien, P. S. Sung, T. Y. Lin, et al., A new instrument combines cognitive and social functioning items for detecting mild cognitive impairment and dementia in Parkinson's disease, *Front. Aging Neurosci.*, **14** (2022), 913958. https://doi.org/10.3389/fnagi.2022.913958

AIMS Press