ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

MRW-ViT: Spatial-Frequency Domain Fusion and Optimal Metric for Few-Shot Medical Image Classification

Ying Wu[1], Jin Lu[1,*]

[1]College of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China
**Corresponding author: lujin@sust.edu.cn*

Abstract: *To address the challenges of data scarcity and inadequate lesion representation in medical image classification, this paper proposes a novel few-shot learning approach integrating spatial and frequency domains, termed Multi-Resolution Wavelet Enhanced Vision Transformer (MRW-ViT). The method utilizes two-dimensional discrete wavelet transform (2D-DWT) to decompose medical images, extracting high-frequency features to enhance lesion detail capture. A self-attention mechanism is employed to dynamically integrate global context with local pathological information, improving feature representation completeness. A cross-domain feature fusion module is designed to combine multi-scale features from both spatial and frequency domains, strengthening pathological representation. Furthermore, Earth Mover's Distance (EMD) is introduced to measure subtle inter-class differences, optimizing classification decisions. Experiments were conducted on the MedMNIST dataset, encompassing six classification tasks including PathMNIST, DermaMNIST, and OCTMNIST. Results demonstrate that MRW-ViT achieves an area under the curve (AUC) of 0.990 in colon pathology classification and an AUC of 0.995 in pneumonia detection, outperforming state-of-the-art methods. In breast ultrasound diagnosis with a limited sample size of 780 images, the AUC reaches 0.948. Ablation studies confirm the effectiveness of each module.*

Keywords: *Few-Shot Learning; Medical Image Classification; Earth Mover's Distance; MedMNIST*

1. Introduction

Intelligent analysis of medical imaging is a core technology in advancing precision medicine[1], holding significant value in early disease screening and clinical staging diagnosis. However, existing deep learning-based diagnostic systems face two primary challenges: First, the heterogeneity and scarcity of medical data present a prominent contradiction, with annotation costs varying significantly across imaging modalities (e.g. X-ray, ultrasound, pathology slides). For instance, in the MedMNIST dataset[2], the breast ultrasound dataset (BreastMNIST) contains only 780 samples, while the kidney pathology dataset (TissueMNIST) comprises 236,000 samples. This imbalanced data distribution leads to a sharp decline in the generalization ability of conventional models for rare disease diagnosis. Few-shot learning, by constructing transferable feature spaces through meta-learning frameworks, demonstrates significant advantages in scenarios with scarce annotated data[3][4]. Second, lesion characterization exhibits cross-modal and multi-scale complexity. For example, colon pathology images (PathMNIST) require identification of glandular structure distortions, while chest CT scans (NoduleMNIST3D) demand detection of spatial morphological features of lung nodules. Single feature extraction strategies struggle to balance global anatomical information with local detail differences, whereas attention-based few-shot learning methods enable precise localization of cross-scale pathological markers through dynamic feature calibration[5][6]. Therefore, developing efficient and robust medical image classification techniques is not only a critical pathway to addressing data heterogeneity and pathological complexity but also a cornerstone for achieving early precise intervention and optimizing clinical decision-making processes, with irreplaceable strategic value for improving patient quality of life and healthcare resource allocation efficiency.

In recent years, numerous innovative deep learning-based methods have emerged in the field of medical image classification, yet existing studies exhibit notable limitations. Many algorithms focus on specific classification tasks for single diseases. For example, Pan et al.[7] developed a three-class model for predicting lung adenocarcinoma invasiveness, achieving outstanding performance in classifying pure

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

ground-glass nodules. The lightweight CSDNet proposed by Lahari's team[8] achieved high accuracy in cataract detection, while CNN architectures proposed by Rasheed[9] and Rafiq[10] attained accuracies exceeding 98% and 90% in brain tumor and breast cancer classification, respectively. Although these studies achieved breakthroughs in specific disease classification tasks, their model architectures are often optimized for the morphological features of target lesions, lacking cross-disease generalization and struggling to address the ubiquitous challenges of lesion heterogeneity and multi-scale features in medical imaging data.

With the advancement of deep learning, researchers have begun exploring broadly adaptable medical image classification frameworks. An et al.[11] proposed MCNN, which dynamically captures lesion boundary features through a visual attention mechanism and integrates multi-scale feature fusion strategies, achieving over 99% accuracy in classifying lung nodules, breast masses, and other diseases, validating the effectiveness of attention mechanisms in cross-disease applications. Yang et al.[12] introduced DiffMIC, the first model to incorporate diffusion models into medical classification. By employing a dual-granularity conditional guidance strategy that integrates global anatomical structures with local pathological features and introducing maximum mean discrepancy regularization to enhance cross-modal robustness, DiffMIC achieved significant performance improvements in diverse tasks such as placental maturity grading and skin lesion classification. Manzari et al.[13] developed MedViT, which enhances the capture of subtle pathological features through a channel-spatial dual attention module and leverages contrastive pre-training to improve generalization in few-shot scenarios, achieving strong classification accuracy across cross-modal datasets, including thyroid nodule classification and histopathological analysis.

However, existing general frameworks still suffer from three common deficiencies: First, reliance on simplistic concatenation or stacking of global and local features leads to misalignment between anatomical structural priors and lesion-specific abnormal features. In cross-modal data, such as X-rays and pathology slides, this approach often introduces feature redundancy, weakening the model's ability to discern subtle pathological patterns. Traditional feature fusion strategies lack dynamic calibration mechanisms for spatial semantic correlations, making it difficult to precisely localize local abnormal regions while preserving global anatomical constraints. Second, conventional feature extraction frameworks overly focus on spatial domain information, failing to effectively integrate frequency domain features, which limits their ability to resolve edge ambiguity and micro-texture heterogeneity. Existing methods often employ single-scale convolutional kernels or fixed receptive field designs, unable to adaptively capture multi-scale frequency domain responses of lesions across modalities. Third, the use of fixed distance metrics, such as cosine similarity[14][15], is highly susceptible to data perturbations when data distributions are heavily skewed, leading to blurred decision boundaries and significantly reducing model robustness in rare disease diagnosis. Particularly in few-shot scenarios, traditional metric learning struggles to balance intra-class compactness and inter-class separability, exacerbating the risk of overfitting to noisy features.

To address these challenges, this paper proposes a novel network named Multi-Resolution Wavelet Enhanced Vision Transformer (MRW-ViT), achieving cross-disease generalization through a spatial-frequency domain decoupling architecture. The core innovations include:

(1) A frequency domain feature fusion method based on two-dimensional discrete wavelet transform (2D-DWT), utilizing Haar wavelet multi-directional high-frequency decomposition to enhance cross-disease micro-texture and edge feature resolution, addressing the issue of insufficient frequency domain utilization in traditional models;

(2) A global-local differential fusion strategy based on self-attention dynamic calibration, achieving efficient collaborative modeling of anatomical structures and lesion abnormalities through spatial alignment and redundancy elimination;

(3) The introduction of Earth Mover's Distance (EMD) metric learning, establishing a fine-grained classification space based on optimal transport theory, enhancing classification robustness in few-shot and class-imbalanced scenarios.

2. The proposed method

The proposed MRW-ViT algorithm is built upon the Vision Transformer (ViT) as the backbone network, leveraging wavelet transform to decompose the frequency domain features of medical images, utilizing a spatial-frequency fusion module to enhance lesion representation, designing a global-local

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

feature interaction mechanism, and incorporating EMD to construct an optimal transport metric, thereby improving the robustness of cross-disease few-shot classification. The network architecture of MRW-ViT is illustrated in Figure 1. Initially, two-dimensional discrete wavelet transform is applied to decompose the input support set and query set images into multiple frequency bands, extracting high-frequency edge and micro-texture features. The original images and the frequency domain images obtained from wavelet transform are concatenated at the channel level to form a spatial-frequency joint embedding, which is then fed into the ViT for feature extraction, as shown in the wavelet frequency feature extraction module in Figure 1. Within the ViT, features extracted by the first 11 layers of the Transformer encoder are input into a cross-domain dynamic weighted fusion module to achieve adaptive integration of spatial and frequency features, as depicted in the multi-scale feature interaction and cross-domain fusion section of Figure 1. The final layer of the ViT outputs the global semantic feature [CLS] token (Global features) and local detail features from image patches (Local features). Key local features are dynamically selected using self-attention weights, and differential computation is employed to enhance the discriminability of pathological features. Finally, EMD is introduced to replace the traditional Softmax classification layer, transforming the classification task into a feature distribution matching problem by minimizing the transport cost between the feature distributions of the support set and query samples to achieve class prediction.

Compared to existing medical image classification models such as MCNN and DiffMIC, MRW-ViT enhances sensitivity to edge-blurred lesions and micro-texture heterogeneity through wavelet-based multi-resolution frequency domain analysis, suppresses cross-modal data redundancy via a differential correction mechanism constrained by anatomical priors, and optimizes few-shot classification robustness using the EMD metric. Experimental results demonstrate that this method significantly improves accuracy and generalization performance across multi-disease classification tasks in the MedMNIST dataset, while supporting end-to-end training and being extensible to multi-modal medical image analysis.

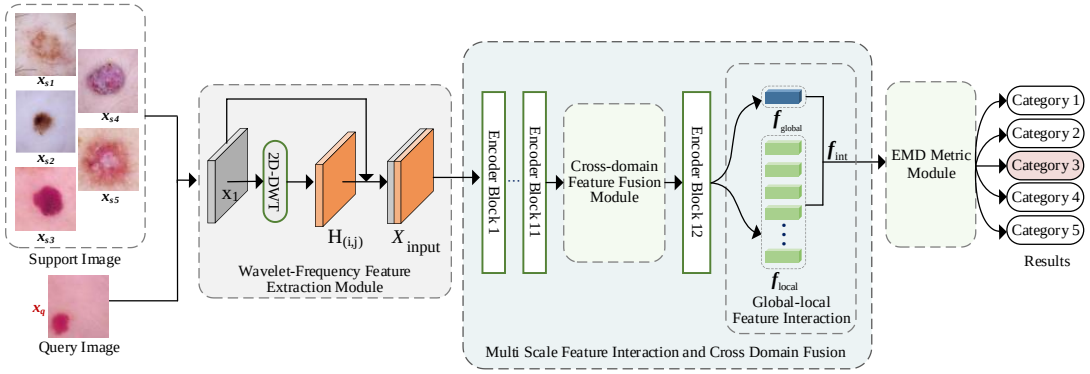

Figure 1: Proposed MRW-ViT Algorithm Overview.

2.1 Few-Shot Learning for Medical Image Classification

In the classification tasks based on the MedMNIST multi-disease image dataset, this paper proposes a few-shot learning method that integrates spatial-frequency domain features to achieve efficient classification of diverse medical images through a meta-learning framework. The core of this method lies in deeply integrating the few-shot learning paradigm with algorithm design to address the challenge of extremely scarce data in real clinical scenarios, constructing a classification model capable of rapid generalization from limited samples. In the specific workflow, the training set is defined as $D_{base} = \{(x_i, y_i) | \ y_i \in C_{base}\}$, where x_i represents the input image (a multi-band fused image processed by 2D-DWT), y_i is its corresponding disease category, and C_{novel} is the set of base class labels, encompassing all disease categories visible during the training phase. The model is first trained on the base class dataset, dynamically integrating spatial texture and frequency-domain microstructural features through the proposed cross-domain feature fusion module to enhance the representation of lesion heterogeneity. During the testing phase, a new class set C_{novel} is introduced (satisfying $C_{\text{novel}} \cap C_{\text{base}} = \varnothing$), and the model predicts the query set Q_{novel} based on the support set $S_{\text{novel}} = \{(x_i, y_i) | \ y_i \in C_{\text{novel}}\}$, where the support set contains *N* unseen categories, each providing *K* samples. The model classifies the query set by computing the similarity between its features and those

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

of the support set for each category, assigning the query set to the nearest category.

To enhance the adaptability of the classification model in cross-category scenarios, this paper introduces an episodic training mechanism. In each training cycle, several different disease categories are dynamically sampled from C_{base} to construct simulated support set and query set pairs. Through iterative training, the model learns to capture critical discriminative features from limited samples, such as textural differences in ground-glass opacities in lung CT or pigment network distribution patterns in dermoscopic images. This training paradigm effectively mitigates the overfitting issues of traditional deep learning methods in the classification of rare diseases or novel lesions.

2.2 Wavelet Frequency Domain Feature Extraction Module

This study employs the Haar wavelet basis function[16] to implement two-dimensional discrete wavelet transform. Due to its orthogonal compact support and computational efficiency, the Haar wavelet can accurately capture edge discontinuities and micro-texture distortion features through horizontal, vertical, and diagonal high-frequency sub-bands while preserving anatomical structures. Compared to other wavelets such as Daubechies[17], the Haar wavelet significantly reduces memory usage, and its binary nature avoids overfitting to specific modalities. Through low-pass filtering(ψ_{low}) and high-pass filtering(ψ_{high}) operations along row and column directions followed by downsampling, the original image is decomposed into four quarter-scale sub-bands. Taking one of the input images as an example, the wavelet transform process is illustrated in Figure 2.

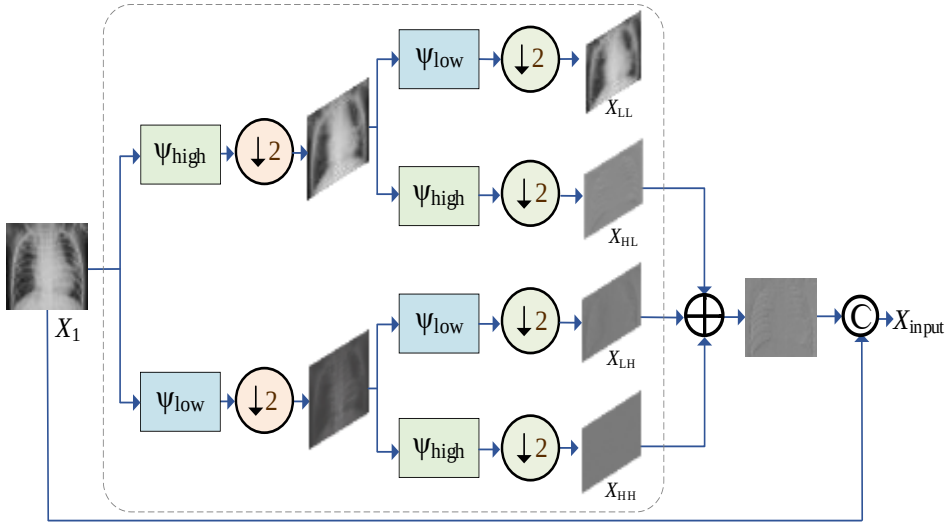

Figure 2: Wavelet-frequency domain feature module.

The LL sub-band (low-frequency information) captures the overall contour and structure of the image. The computational formula is as follows:

$$LL(i,j)=\sum_{m,n=1}^{M,N} f(m,n)\cdot\psi_{\text{low}}(i-m)\cdot\psi_{\text{low}}(j-n) \tag{1}$$

Where M and N represent the row and column dimensions of the input image, respectively, and $f(m,n)$ denotes the original image. The indices i and j represent the spatial coordinates of the sub-band images after wavelet decomposition, with their ranges determined by the input image dimensions and a downsampling factor. $LL(i,j)$ representing the smoothed low-frequency component, ψ_{low} is obtained using a low-pass filter. This sub-band applies low-pass filtering and downsampling to both the rows and columns of the image, filtering out high-frequency noise and details while preserving the global structural information. It effectively highlights the overall contours of different diseases in medical images, facilitating the identification of large-scale features, such as lung contours in chest CT or tissue distribution in pathology slides.

The HL sub-band (horizontal high-frequency information) is obtained through a combination of low-

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

pass filtering in the row direction and high-pass filtering in the column direction, emphasizing edges and details in the horizontal direction. This sub-band is suitable for detecting horizontal boundaries of lesions or horizontal layering of tissues. The computational formula is as follows:

$$HL(i,j)=\sum_{m,n=1}^{M,N} f(m,n)\cdot\psi_{\text{low}}(i-m)\cdot\psi_{\text{high}}(j-n) \tag{2}$$

Where ψ_{high} denotes the high-pass filter, extracting high-frequency changes in the horizontal direction. The HL sub-band focuses on capturing edges and texture details in the horizontal direction, enhancing the representation of subtle features in this orientation. For example, in breast ultrasound images, the HL sub-band aids in detecting the horizontal boundaries of tumors; in colon pathology images,it highlights the horizontal layering features of glandular structures.

The LH sub-band (vertical high-frequency information) is obtained through high-pass filtering in the row direction and low-pass filtering in the column direction, capturing edges and morphological contours in the vertical direction. This sub-band is suitable for analyzing vascular orientations or vertical tissue hierarchies. The computational formula is as follows:

$$LH(i,j)=\sum_{m,n=1}^{M,N} f(m,n)\cdot\psi_{\text{high}}(i-m)\cdot\psi_{\text{low}}(j-n) \tag{3}$$

This sub-band focuses on morphological changes in the vertical direction, enhancing the discriminability of vertical projections of lung nodules in chest CT and selectively amplifying topological features of longitudinal vascular distributions.

The HH sub-band (diagonal high-frequency information) is obtained by applying high-pass filters to both rows and columns, extracting texture and details in the diagonal direction. It is capable of identifying complex edge changes or cross-texture features of lesions. The computational formula is as follows:

$$HH(i,j)=\sum_{m,n=1}^{M,N} f(m,n)\cdot\psi_{\text{high}}(i-m)\cdot\psi_{\text{high}}(j-n) \tag{4}$$

By emphasizing details in the diagonal direction, this sub-band provides a unique perspective, enhancing the model's perception of complex local features and offering specific detection capabilities for cross-texture patterns of microcalcifications.

To evaluate the contribution of each sub-band to classification performance, ablation studies in this research reveal that retaining only the high-frequency components (HL, LH, and HH) while excluding the low-frequency sub-band results in a greater improvement in average classification accuracy on the MedMNIST dataset. This phenomenon is attributed to the fact that low-frequency information primarily encodes the overall contour of the image, whereas high-frequency sub-bands effectively supplement the edge sharpness and texture details absent in the original image. MedMNIST classification tasks rely more heavily on the discrimination of local pathological markers. Consequently, MRW-ViT adopts a high-frequency feature synthesis strategy, spatially stacking the horizontal, vertical, and diagonal high-frequency sub-bands to construct a composite high-frequency feature map, as shown in Figure 2. The process is as follows:

$$\begin{aligned} H_{\text{high-frequency}}(i,j)=H_{\text{horizontal}}(i,j)+H_{\text{vertical}}(i,j)\\ +H_{\text{diagonal}}(i,j)\end{aligned} \tag{5}$$

Where $H_{\text{horizontal}}$, H_{vertical} and H_{diagonal} correspond to the HL, LH, and HH sub-bands, respectively. To meet the input requirements of the ViT, the high-frequency feature map is resized to 224×224 pixels via interpolation and concatenated with the original image X_1 along the channel dimension to form the final input:

$$X_{\text{input}}\in\mathbb{R}^{224\times224\times2} \tag{6}$$

This input combines the global structural information of the original image with the local details of

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

the high-frequency features, providing a rich feature representation for multi-disease classification.

2.3 *ModuleMulti-Scale Feature Interaction and Cross-Domain Fusion*

During the feature extraction phase, MRW-ViT employs dynamic weighted fusion of the frequency domain features extracted in Section 2.2 with the spatial domain features of the original image. Through self-attention and cross-domain fusion, it achieves dynamic calibration of global and local features, as illustrated in the multi-scale feature interaction and cross-domain fusion section of Figure 1. The specific workflow is shown in Figure 3.

2.3.1 Dynamic Weighted Spatial-Frequency Feature Fusion

To further enhance the feature representation capability, MRW-ViT introduces a cross-domain feature fusion module into the ViT architecture. In the feature extraction stage, the input dual-channel images from the spatial and frequency domains are divided into patches and linearly projected, then passed through the first 11 layers of the Transformer encoder to extract feature information. Before passing to the 12th layer encoder, the spatial features($\boldsymbol{f}_{\text{sd}}$) and frequency features($\boldsymbol{f}_{\text{fd}}$) are fused through a weighted sum using learnable weights, dynamically balancing the contributions of both types of features to obtain the fused features, which are then passed to the final encoder layer for processing. The formula for the weighted fusion using learnable weights is as follows:

$$\boldsymbol{f}_{\text{fused}} = w_1 \cdot \boldsymbol{f}_{\text{sd}} + w_2 \cdot \boldsymbol{f}_{\text{fd}} \tag{7}$$

Where the weights $w_1, w_2 \in [0,1]$ are processed by an activation function and adaptively optimized via the backpropagation algorithm, with the numerical constraint that their sum is 1. The fused features achieve collaborative modeling of multi-resolution pathological features through cross-domain fusion, effectively capturing the shape, size, and boundary details of diseases in the image. The fused features from the spatial and frequency domains are deeply integrated into the final global and local features through the 12th encoder layer, achieving dynamic association and calibration of semantics across domains via the self-attention mechanism.

2.3.2 Global-Local Feature Generation and Interaction Mechanism

This section introduces the process by which the ViT generates global and local features through hierarchical encoding and their interaction, as illustrated in the global-local feature interaction part of Figure 1. The input image X_{input} is divided into non-overlapping patches of $P \times P$, each of which is mapped to a *d*-dimensional feature space through linear embedding and augmented with position encoding to generate the initial sequence:

$$\boldsymbol{Z}_0 = [\boldsymbol{z}_{[\text{CLS}]}, \boldsymbol{E}(\boldsymbol{x}_1), \boldsymbol{E}(\boldsymbol{x}_2), \ldots, \boldsymbol{E}(\boldsymbol{x}_N)] + \boldsymbol{E}_{\text{pos}} \tag{8}$$

Where $\boldsymbol{Z}_0$ is the initial input sequence, $z_{[\text{CLS}]}$ is the classification token, $E(x_i)$ is the patch embedding representation of the *i*-th image patch, and E_{pos} is the position encoding. The input sequence $\boldsymbol{Z}_0$ is processed through multiple layers of the ViT encoder, each consisting of layer normalization (LN), multi-head attention (MHA), and a multilayer perceptron (MLP). The [CLS] token is then used to output the global feature, while the image patch-level outputs serve as local features:

$$\boldsymbol{f}_{\text{global}} = \boldsymbol{z}_{[\text{CLS}]}^{(\text{L})} \tag{9}$$

$$\boldsymbol{f}_{\text{local}} = [\boldsymbol{E}_1^{(\text{L})}, \boldsymbol{E}_2^{(\text{L})}, \ldots, \boldsymbol{E}_N^{(\text{L})}] \tag{10}$$

Where L denotes the number of encoder layers, and $\boldsymbol{E}_i^{(\text{L})}$ is the embedding representation of the *i*-th image patch output from the L-th (final) layer of the ViT encoder. Through the self-attention mechanism, the [CLS] token interacts with all image patches to aggregate global context, obtaining the global feature $\boldsymbol{f}_{\text{global}}$, which represents the semantics of the entire image. The local features $\boldsymbol{f}_{\text{local}}$ consist of the final embeddings of all image patches $\boldsymbol{E}_i^{(\text{L})}$, preserving the local details and spatial information of the image.

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

In medical image analysis, there is semantic coupling between global features, such as the overall shape of an organ, and local features, such as the microstructure of lesions. Directly concatenating global and local features may lead to feature redundancy, reducing the model's sensitivity to subtle pathological changes. This paper proposes an attention-guided redundancy suppression strategy. By obtaining the self-attention matrix $\boldsymbol{A}$ from the last encoder layer, the attention weights of the [CLS] token on each image patch are extracted. These weights are then used to generate interactive local features that focus on discriminative regions strongly related to pathology, as follows:

$$\boldsymbol{f}_{\text{int}} = \boldsymbol{f}_{\text{local}} - \text{Softmax}(\boldsymbol{A}) \cdot \boldsymbol{f}_{\text{global}} \tag{11}$$

This mechanism uses the attention weights $\text{Softmax}(\boldsymbol{A})$ obtained from the encoder's multi-head attention to quantify the redundant influence of global features on local features. By employing directional subtraction, it suppresses anatomical commonalities such as organ textures, thereby enhancing pathological markers like lesion edges and calcifications.

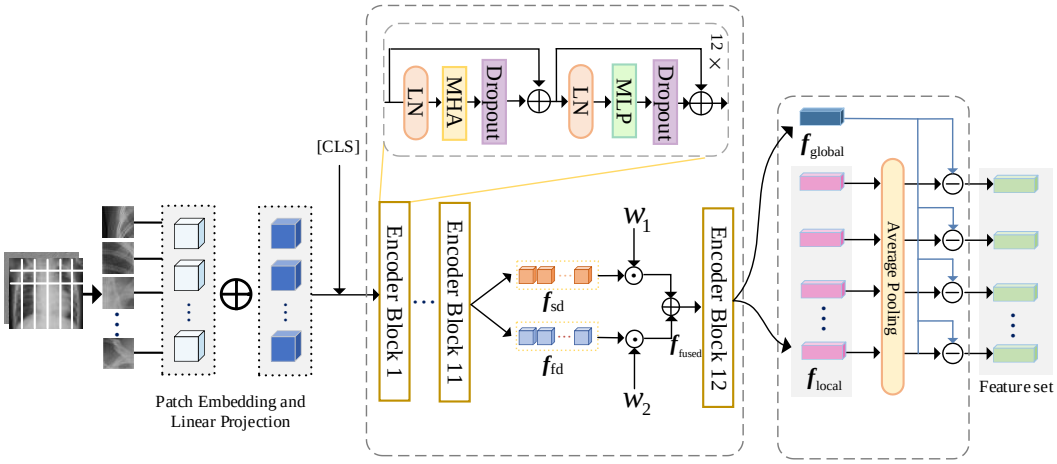

Figure 3: Multi Scale Feature Interaction and Cross Domain Fusion Module.

2.4 Fine-Grained Classification Module Based on EMD

Due to the significant intra-class variations and high inter-class similarities in medical image classification, coupled with the typically limited availability of labeled data, this paper proposes a fine-grained classification framework based on EMD. This framework leverages optimal transport theory to match the feature distributions of the support set and query set, focusing on class-relevant feature information. By comparing the regional features of images in the support set and query set, the most relevant parts are automatically matched and weighted to compute the final similarity score $\text{simi}(x_s, x_q)$.

Specifically, the EMD computation involves two key inputs: the cost matrix and weights. During the classification prediction phase, the network performs classification based on the similarity between the features of the support set and query set, using the EMD metric to calculate the distance between feature distributions[18][19], as illustrated in Figure 4. Given the feature set $\boldsymbol{P} = \{\boldsymbol{p}_k\}_{k=1}^{M}$ of the support set image x_s and the feature set $\boldsymbol{Q} = \{\boldsymbol{q}_l\}_{l=1}^{M}$ of the query set image x_q, where M is the number of local features per image, $\boldsymbol{p}_k$ denotes the k-th local feature vector of the support set, and $\boldsymbol{q}_l$ denotes the l-th local feature vector of the query set. The EMD computation is simplified as:

$$\text{EMD}(\boldsymbol{P}, \boldsymbol{Q}) = \min \sum_{k=1}^{M} \sum_{l=1}^{M} c_{kl} \cdot f_{kl} \tag{12}$$

$$c_{kl} = 1 - S_{kl} \tag{13}$$

Where the cost matrix $\boldsymbol{C} = [c_{kl}]$ represents the dissimilarity between features $\boldsymbol{p}_k$ and $\boldsymbol{q}_l$; f_{kl} is the matching flow from p_k to q_l; and S_{kl} is the cosine similarity between $\boldsymbol{p}_k$ and $\boldsymbol{q}_l$. The

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

cost matrix with dimensions $M \times M$, is generated through pairwise comparisons, providing the basis for the transport cost in EMD.

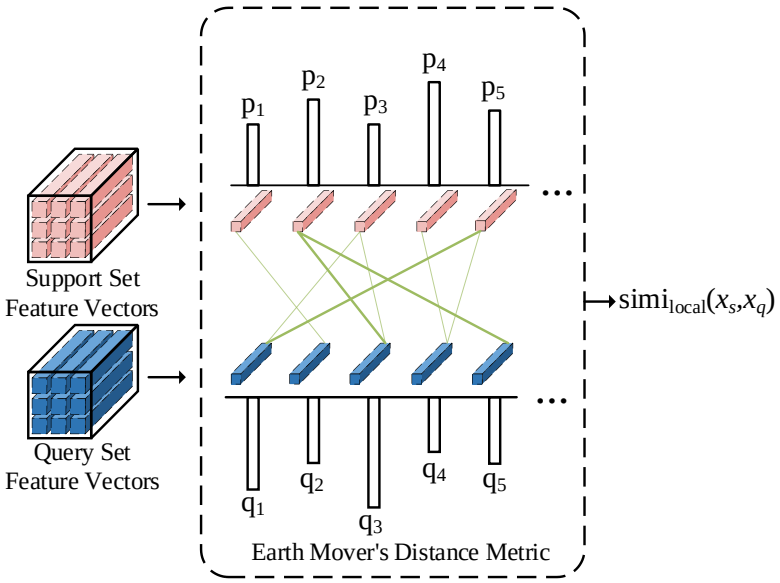

Figure 4: EMD Metric Module.

In this method, weights $\boldsymbol{w}_s = \{\omega_{sk}\}$ and $\boldsymbol{w}_q = \{\omega_{ql}\}$ are introduced to represent the importance of local features in the support set and query set, respectively, generated by an attention mechanism, where each element corresponds to the weight of features $\boldsymbol{p}_k$ and $\boldsymbol{q}_l$. The weights undergo nonlinear transformation and normalization. Based on the cost matrix $\boldsymbol{C}$, weights $\boldsymbol{w}_s$ and $\boldsymbol{w}_q$, EMD solves for the flow matrix $\boldsymbol{F} = [f_{kl}]$ using an optimal transport algorithm. This matrix reflects the optimal matching relationship between the features of the support set and query set. The local similarity score is computed as:

$$\mathrm{simi}_{\mathrm{local}}(\boldsymbol{x}_s, \boldsymbol{x}_q) = \sum_{k=1}^{M}\sum_{l=1}^{M} S_{kl} \cdot f_{kl} \tag{14}$$

$\mathrm{simi}_{\mathrm{local}}$ quantifies the semantic consistency of local feature distributions through weighted summation, leveraging optimal transport to focus on key pathological regions.

The proposed method combines the local similarity from EMD with global feature similarity for classification prediction. The global feature similarity score is calculated as:

$$\mathrm{simi}_{\mathrm{global}} = \cos\left(\boldsymbol{f}_{\mathrm{s}}^{\mathrm{global}}, \boldsymbol{f}_{\mathrm{q}}^{\mathrm{global}}\right) \tag{15}$$

Where $\boldsymbol{f}_{\mathrm{s}}^{\mathrm{global}}$ is the global feature of the support set image, $\boldsymbol{f}_{\mathrm{q}}^{\mathrm{global}}$ is the global feature of the query set image, with their cosine similarity computed. The final similarity score, integrating local and global information, is computed for each support set sample $\boldsymbol{x}_s^i$ and query set sample $\boldsymbol{x}_{\mathrm{q}}^{\mathrm{j}}$:

$$\mathrm{simi}(\boldsymbol{x}_s^i, \boldsymbol{x}_{\mathrm{q}}^{\mathrm{j}}) = \sum_{k=1}^{M}\sum_{l=1}^{M} S_{kl} \cdot f_{kl} \tag{16}$$

Subsequently, the differences across all support set samples for each class are normalized using the softmax function to obtain the probability that the query image belongs to each class:

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

$$P(y=c_i \mid x_q)=\frac{\sum_{j=1}^{K}\exp(\mathrm{simi}(x_{s,i,j},x_q))}{\sum_{m=1}^{N}\sum_{l=1}^{K}\exp(\mathrm{simi}(x_{s,m,l},x_q))} \tag{17}$$

Where $P(y=c_i \mid x_q)$ denotes the probability that the query image belongs to class c_i; K is the number of support set samples per class, and N is the total number of classes; $x_{s,i,j}$ represents the *j*-th support set sample of class c_i. The predicted class is determined by maximizing the probability:

$$\hat{y}=\arg\max_{c_i} P(y=c_i \mid x_q) \tag{18}$$

Where $\arg\max_{c_i}$ represents the category index with the highest probability, and $\hat{y}$ is the final predicted category.

The network is optimized by minimizing the negative log-likelihood loss function:

$$\mathrm{loss}=-\sum_{q=1}^{N}\sum_{i=1}^{N} I(y_q=c_i)\log(p(x_q)_i) \tag{19}$$

Where $I(y_q=c_i)$ is an indicator function that equals 1 if the true label of the query image y_q is class c_i, and 0 otherwise. This loss function drives the model to enhance the semantic alignment of pathological features within the same class while suppressing cross-class interference. By combining EMD-based local structure matching with global feature fusion, this method improves classification accuracy in few-shot fine-grained classification while preserving multi-scale pathological information.

3. Experimental results and analysis

3.1 Dataset and experimental setup

This study utilizes the publicly available MedMNIST dataset, which comprises 12 two-dimensional subsets encompassing various imaging modalities, including X-ray, optical coherence tomography (OCT), ultrasound, CT scans, and electron microscopy. Systematic experiments were conducted on six subsets of MedMNIST-2D, as shown in Figure 5, covering diverse medical imaging modalities such as pathological slides (PathMNIST), dermoscopic images (DermaMNIST), optical coherence tomography (OCTMNIST), pneumonia X-ray (PneumoniamnIST), breast ultrasound (BreastMNIST), and blood cell microscopy (BloodMNIST). This dataset supports various classification tasks, including binary and multi-class classification, with sample sizes spanning multiple orders of magnitude, ranging from 10^2 to 10^5. The heterogeneous characteristics of the dataset establish a multidimensional benchmark for the systematic validation of classification models. Detailed data are presented in Table 1.

PathMNIST is derived from colorectal cancer histopathological slides, comprising 107,180 image patches of 224×224 pixels, with a non-overlapping sampling strategy to ensure histological independence. DermaMNIST contains 10,015 RGB images of 450×600 pixels, used for the differential diagnosis of seven types of skin lesions. OCTMNIST integrates 109,309 retinal OCT images with an original resolution of 384×512 pixels, centrally cropped to preserve the macular region structure, with four pathological labels corresponding to the diabetic macular edema grading system. PneumoniaMNIST includes 5,856 pediatric chest X-rays for a binary classification task to distinguish pneumonia from normal cases. BreastMNIST comprises 780 breast ultrasound images for a binary classification task, distinguishing between benign and malignant cases. BloodMNIST is a dataset for peripheral blood cell classification, encompassing eight blood cell types.

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

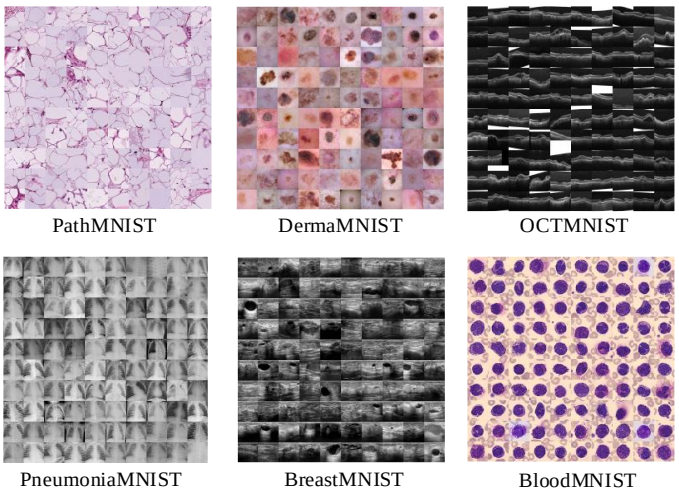

Figure 5: Experimental dataset thumbnail images.

Table 1: MedMNIST dataset.

Name	Data Modality	Task Type	Sample Size	Train/Validation/Test
PathMNIST	Pathological images of colon	Multi classification (9 categories)	107,180	89996 / 10004 / 7180
DermaMNIST	Dermatoscope	Multi classification (7 categories)	10,015	7007 / 1003 / 2005
OCTMNIST	Retinal OCT	Multi classification (4 categories)	109,309	97477 / 10832 / 1000
PneumoniaMNIST	Chest X-ray	Binary classification	5,856	4708 / 524 / 624
BreastMNIST	Breast ultrasound	Binary classification	780	546 / 78 / 156
BloodMNIST	Microscopic image of blood cells	Multi classification (8 categories)	17,092	11959 / 1712 / 3421

The model employs ViT as the backbone network for classification tasks. To leverage the advantages of transfer learning, the model is initialized with ViT weights pretrained on the ImageNet dataset [20], as officially released. Experiments were implemented using the PyTorch framework and trained on an NVIDIA 3090 Ti GPU. In terms of network architecture, ViT divides input images into fixed-size 16×16 pixel patches and learns feature representations through a self-attention mechanism. During training, the Adam optimizer is used with a learning rate set to 1e-5. All models are trained for 1000 epochs, with validation performed every 20 epochs to save the best weights. For multi-class tasks, a 5-way 10-shot paradigm is adopted, where each training task randomly selects five classes from the base class set, with ten support samples per class for learning. The testing phase follows the same 5-way 10-shot configuration. For binary classification tasks, a 2-way 10-shot paradigm is used to evaluate model performance in few-shot learning scenarios.

Evaluation metrics include Accuracy and Area Under the ROC Curve (AUC), assessing model performance from the perspectives of overall classification accuracy and threshold robustness, respectively. Accuracy measures the overall correct prediction rate, while AUC integrates true positive and false positive rates across different classification thresholds, reflecting the model's stable discriminative ability. To ensure the reliability and statistical significance of the results, all test tasks are evaluated using 10,000 randomly generated tasks, with corresponding 95% confidence intervals calculated.

3.2 Analysis of experimental results

3.2.1 Comparison of experimental results

To validate the effectiveness of the proposed method, experiments were conducted on six MedMNIST-2D subsets, systematically comparing it against 11 models, including ResNet-18[22], ResNet-50[21], Auto-sklearn[23], AutoKeras[24], Google AutoML[25], DARTS[26], SNAS[27], HOPNAS[28], MedVIT-S[13], NSGA-Net[29], and MSTF-NAS[30] (table 2). The results demonstrate that MRW-ViT exhibits superior performance in most tasks. Figure 6 visualizes the differences in feature attention regions between ResNet-50 and the proposed MRW-ViT using heatmaps. ResNet-50's feature responses show a diffuse distribution, whereas MRW-ViT, through spatial-frequency domain fusion and attention calibration mechanisms, significantly enhances focus on diagnostically critical regions across modalities.

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

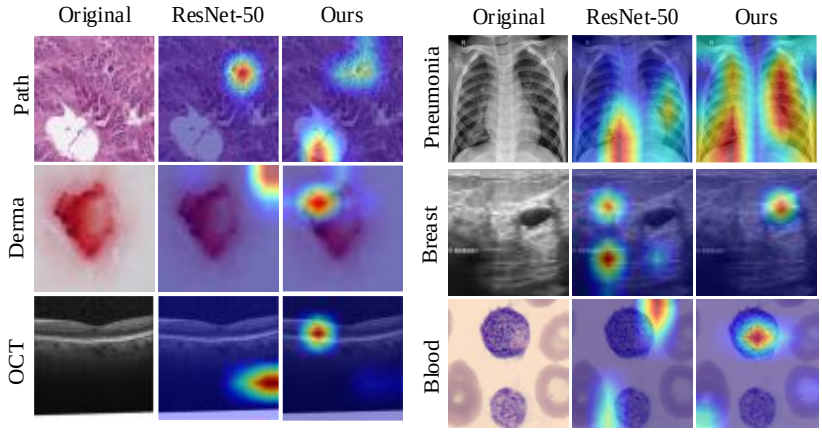

Figure 6: Comparison of ResNet-50 and MRW-ViT Heatmaps.

MRW-ViT achieves superior average AUC and ACC across the six subsets, outperforming other algorithms, as shown in Table 3. In the colon pathology classification task (PathMNIST), MRW-ViT achieves an AUC of 0.990 but an ACC of 0.927, lower than MedViT-S. In the dermoscopic image classification task (DermaMNIST), MRW-ViT's AUC and ACC significantly surpass other algorithms. In the retinal classification task (OCTMNIST), its ACC outperforms ResNet-50, MedViT-S, and MSTF-NAS, highlighting its ability to capture complex pathological features. In binary classification tasks, the proposed method demonstrates stronger adaptability. For the pneumonia X-ray classification task (PneumoniaMNIST), MRW-ViT achieves an AUC of 0.995 and an ACC of 0.948, surpassing all baselines and achieving state-of-the-art performance. For the breast ultrasound image classification task (BreastMNIST), AUC and ACC reach 0.948 and 0.891, respectively, improving by 3.6% and 1.5% over ResNet-18 and significantly outperforming AutoKeras, underscoring its discriminative capability in low-resolution ultrasound images.

Table 2: Comparative Classification Performance.

Network	PathMNIST		DermaMNIST		OCTMNIST	
	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18(224)	0.978	0.860	0.896	0.727	0.960	0.752
ResNet-50(224)	0.976	0.848	0.895	0.719	0.951	0.750
Auto-sklearn	0.500	0.186	0.906	0.734	0.883	0.595
AutoKeras	0.979	0.864	0.921	0.756	0.956	0.736
Google AutoML	0.982	0.811	0.925	0.766	0.965	0.732
DARTS	0.975	0.872	0.913	0.749	0.953	0.712
SNAS	0.969	0.850	0.906	0.737	0.949	0.708
HOPNAS	0.987	0.912	0.899	0.759	0.948	0.761
MedVIT-S	0.984	0.942	0.914	0.773	0.945	0.782
NSGA-NET	0.979	0.866	0.915	0.744	0.954	0.765
MSTF-NAS	0.990	0.910	0.923	0.773	0.952	0.780
Ours(MRW-ViT)	0.990	0.927	0.941	0.797	0.958	0.804
Network	PneumoniaMNIST		BreastMNIST		BloodMNIST	
	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18(224)	0.970	0.861	0.915	0.878	0.991	0.958
ResNet-50(224)	0.968	0.896	0.863	0.833	0.997	0.956
Auto-sklearn	0.947	0.865	0.848	0.808	0.984	0.878
AutoKeras	0.970	0.918	0.833	0.801	0.996	0.961
Google AutoML	0.993	0.941	0.932	0.865	0.996	0.966
DARTS	0.965	0.874	0.912	0.832	0.994	0.953
SNAS	0.974	0.871	0.894	0.811	0.996	0.946
HOPNAS	0.971	0.852	0.907	0.853	0.996	0.958
MedVIT-S	0.991	0.921	0.938	0.883	0.997	0.950
NSGA-NET	0.965	0.907	0.857	0.846	0.997	0.970
MSTF-NAS	0.963	0.912	0.930	0.872	0.998	0.976
Ours(MRW-ViT)	0.995	0.948	0.948	0.891	0.998	0.980

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

Table 3: Average Performance Comparison.

Network	Average AUC	Average ACC
ResNet-18(224)	0.952	0.839
ResNet-50(224)	0.942	0.834
Auto-sklearn	0.845	0.678
AutoKeras	0.943	0.839
Google AutoML	0.966	0.847
DARTS	0.952	0.833
SNAS	0.948	0.821
HOPNAS	0.951	0.849
MedVIT-S	0.962	0.875
NSGA-NET	0.945	0.850
MSTF-NAS	0.959	0.871
Ours(MRW-ViT)	0.972	0.891

Note: Numbers in parentheses indicate the input image spatial resolution (pixels), e.g. ResNet-18(224) denotes an input size of 224×224 pixels. Boldface indicates the best performance for each task.

MRW-ViT demonstrates overall superiority across the six medical imaging tasks, but its accuracy in PathMNIST is slightly lower than MedViT-S. Its spatial-frequency domain decoupling architecture, leveraging Haar wavelet high-frequency enhancement, improves sensitivity to early cancerous markers, with AUC surpassing MedViT-S, aligning with the "sensitivity-first" principle in the NCCN Colorectal Cancer Screening Guidelines[31]. However, low-frequency suppression weakens global feature modeling, limiting precision in highly differentiated lesions. MedViT-S's dual-attention mechanism is better suited for pathological slide modeling. In PneumoniaMNIST and BreastMNIST tasks, MRW-ViT exhibits cross-modal advantages, outperforming baselines. In DermaMNIST and OCTMNIST tasks, MRW-ViT's accuracy surpasses Google AutoML, though its AUC is slightly lower. Its high-frequency enhancement strategy effectively captures local features but is less effective in modeling low-frequency morphological features, whereas Google AutoML more evenly captures low-frequency features. MRW-ViT's decision boundaries align well with clinical standards in class-imbalanced scenarios, achieving an excellent balance between sensitivity and specificity, with comprehensive performance validating its cross-modal advantages.

3.2.2 Ablation experiment

To validate the contribution of the core modules in the proposed method, systematic ablation experiments were conducted on the BreastMNISt dataset. As shown in Table 4, the impact of each component on classification performance was quantitatively analyzed by incrementally introducing the wavelet frequency domain features, spatial-frequency domain fusion, global-local feature interaction, and EMD metric modules.

Wavelet frequency domain feature extraction and spatial-frequency domain fusion establish a foundation for cross-modal feature representation, enhancing discriminative power by resolving high-frequency features. The EMD metric optimizes feature distribution alignment (improving AUC by 0.6%), suppressing modal noise in high-frequency subbands and forming a "representation-metric collaborative optimization" paradigm. The global-local interaction module guides high-frequency feature focus on critical regions (e.g., tumor edges in breast ultrasound) through global attention, while local differentiation enhances lesion specificity, significantly improving AUC by 1.7%. The staged optimization strategy—wavelet transform for multi-scale feature extraction, global-local interaction for anatomical key region selection, and EMD for inter-class difference enhancement—forms a "feature extraction-semantic selection-metric enhancement" pipeline. The effectiveness of multi-dimensional joint modeling was validated on MedMNIST.

Table 4: Ablation results.

Wavelet Frequency Domain Feature Extraction	Global-Local Interaction	Spatial-Frequency Domain Fusion	EMD Metric	AUC	ACC
×	×	×	×	0.899	0.761
√	×	√	×	0.925	0.823
×	√	×	×	0.908	0.789
×	×	×	√	0.916	0.805
√	×	√	√	0.931	0.844
√	√	√	√	0.948	0.891

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

4. Conclusions

To address the challenges of data heterogeneity and cross-modal generalization in medical image classification, this study proposes a novel Multi-Resolution Wavelet-Enhanced Vision Transformer (MRW-ViT) framework. By leveraging two-dimensional discrete wavelet transform to extract multi-band high-frequency features, combined with a self-attention dynamic calibration mechanism, MRW-vit achieves collaborative modeling of global anatomical constraints and local lesion features. Additionally, an EMD metric is employed to optimize decision boundaries for few-shot classification. Experimental validation on the MedMNIST multi-modal dataset demonstrates that MRW-ViT significantly outperforms mainstream models in tasks such as pathology classification and pneumonia detection, with an average classification accuracy improvement of 6.13%, confirming the effectiveness of the spatial-frequency domain decoupling architecture.

Despite its advantages in cross-modal medical image classification, MRW-ViT has certain limitations:the selection of wavelet basis functions relies on prior knowledge, and the suppression of low-frequency information by Haar wavelets may impact performance in specific modalities, such as MRI soft tissue imaging; the computational complexity of the EMD metric increases significantly with the number of classes, limiting its real-time applicability in large-scale classification scenarios; and the model's ability to model spatiotemporal features in three-dimensional medical imaging remains unverified. Future work will explore adaptive wavelet basis optimization strategies, integrate knowledge distillation to reduce EMD computational complexity, and extend the framework to multi-modal image joint analysis to enhance clinical deployment versatility.

References

[1] LIU J, ZHU M Y, CHEN F, et al. Research on precise diagnosis and treatment in neurosurgery based on intelligent medical image analysis technology[J]. Chinese Medical Equipment Journal, 2018, 39(2): 1-6,28.

[2] YANG J, SHI R, WEI D, et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification[J]. Scientific Data, 2023, 10(1): 41.

[3] WANG Y L, ZHANG S L, LI C J, et al. Text classification method based on TF-IDF and cosine similarity[J]. Journal of Chinese Information Processing, 2017, 31(5): 138-145.

[4] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 4077-4087.

[5] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning, Sydney, Aug 6-11, 2017. PMLR, 2017: 1126-1135.

[6] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. ACM Computing Surveys, 2020, 53(3): 1-34.

[7] CHEN W Y, LIU Y C, KIRA Z, et al. A closer look at few-shot classification[C]//7th International Conference on Learning Representations, New Orleans, May 6-9, 2019. ICLR, 2019.

[8] PAN Z, HU G, ZHU Z, et al. Predicting invasiveness of lung adenocarcinoma at chest CT with deep learning ternary classification models[J]. Radiology, 2024, 311(1): e232057.

[9] PLL, VADDI R, ELISH M O, et al. CSDNet: A novel deep learning framework for improved cataract state detection[J]. Diagnostics, 2024, 14(10): 983.

[10] RASHEED Z, MA Y K, ULLAH I, et al. Automated classification of brain tumors from magnetic resonance imaging using deep learning[J]. Brain Sciences, 2023, 13(4): 602.

[11] RAFIQ A, CHURSIN A, AWAD ALREFAE W, et al. Detection and classification of histopathological breast images using a fusion of CNN frameworks[J]. Diagnostics, 2023, 13(10): 1700.

[12] AN F, LI X, MA X. Medical image classification algorithm based on visual attention mechanism-MCNN[J]. Oxidative Medicine and Cellular Longevity, 2021, 2021: 6280690.

[13] YANG Y, et al. DiffMIC: Dual-guidance diffusion network for medical image classification[C]//26th International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, Oct 8-12, 2023. Cham: Springer, 2023: 221-230.

[14] MANZARI O N, AHMADABADI H, KASHIANI H, et al. MedViT: A robust vision transformer for generalized medical image classification[J]. Computers in Biology and Medicine, 2023, 157: 106791.

[15] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: Relation network for few-shot learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. IEEE, 2018: 1199-1208.

ISSN 2616-5775 Vol. 8, Issue 7: 33-46, DOI: 10.25236/AJCIS.2025.080705

[16] LEPIK Ü, HEIN H. Haar wavelets[M]//Haar wavelets: with applications. Cham: Springer, 2014: 7-20.
[17] VONESCH C, BLU T, UNSER M. Generalized Daubechies wavelet families[J]. IEEE Transactions on Signal Processing, 2007, 55(9): 4415-4429.
[18] RIAZ F, HASSAN A, REHMAN S, et al. EMD-based temporal and spectral features for the classification of EEG signals using supervised learning[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2015, 24(1): 28-35.
[19] DU R S, ZHANG Y N, MENG L D, et al. Few-shot mineral image classification based on EMD distance metric[J]. Journal of Zhengzhou University (Natural Science Edition), 2023, 55(6): 63-70.
[20] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Jun 20-25, 2009. IEEE, 2009: 248-255.
[21] WEN L, LI X, GAO L. A transfer convolutional neural network for fault diagnosis based on ResNet-50[J]. Neural Computing and Applications, 2020, 32(10): 6111-6124.
[22] AYYACHAMY S, ALEX V, KHENED M, et al. Medical image retrieval using ResNet-18[C]//Medical Imaging 2019: Imaging Informatics for Healthcare, San Diego, Feb 17-18, 2019. SPIE, 2019, 10954: 233-241.
[23] FEURER M, EGGENSPERGER K, FALKNER S, et al. Auto-sklearn 2.0: Hands-free automl via meta-learning[J]. Journal of Machine Learning Research, 2022, 23(261): 1-61.
[24] JIN H, SONG Q, HU X. Auto-keras: An efficient neural architecture search system[C]//25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, Aug 4-8, 2019. ACM, 2019: 1946-1956.
[25] ERICKSON N, MUELLER J, SHIRKOV A, et al. AutoGluon-tabular: Robust and accurate automl for structured data[J]. arXiv preprint, 2020. arXiv:2003.06505.
[26] LIU H, SIMONYAN K, YANG Y. DARTS: Differentiable architecture search[C]// Proceedings of the International Conference on Learning Representations, 2019.
[27] XIE S, ZHENG H, LIU C, LIN L. SNAS: Stochastic neural architecture search[C]// Proceedings of the International Conference on Learning Representations, 2019.
[28] ZHANG J, LI D, WANG L, ZHANG L. One-shot neural architecture search by dynamically pruning supernet in hierarchical order[J]. International Journal of Neural Systems, 2021, 31(7): 2150029.
[29] LU Z, WHALEN I, BODDETI V, DHEBAR Y, DEB K, GOODMAN E, BANZHAF W. NSGA-Net: Neural architecture search using multi-objective genetic algorithm[C]// Proceedings of the Genetic and Evolutionary Computation Conference. 2019: 419-427.
[30] WANG Y, et al. MedNAS: Multiscale Training-Free Neural Architecture Search for Medical Image Analysis[J]. IEEE Transactions on Evolutionary Computation, 2024, 28(3): 668-681.
[31] Burt R W, Barthel J S, Dunn K B, et al. NCCN clinical practice guidelines in oncology. Colorectal cancer screening[J]. Journal of the National Comprehensive Cancer Network: JNCCN, 2010, 8(1): 8-61.