Pattern Recognition 180 (2026) 114613

Contents lists available at ScienceDirect

Pattern Recognition

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

PyMIF-Net: Pyramid Mutual Information Fusion Network for multimodal skin lesion classification☆

Chengcheng Li [a,b,1], Huiying Xu [a,b,*,1], Huiling Chen [c,*], Zhihua Wu [d], Zhendong Chen [a,b], Yue Hu [a,b], Yue Chen [a,b], Xinkai Xu [e], Xinzhong Zhu [c,a,*]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, 321004, Zhejiang, China
[b] College of Computer Science and Technology, Zhejiang Normal University, Jinhua, 321004, Zhejiang, China
[c] College of Computer and Artificial Intelligence, Wenzhou University, Wenzhou, 325035, Zhejiang, China
[d] Jinhua Key Laboratory of Biotechnology on Specialty Economic Plants, College of Life Sciences, Zhejiang Normal University, Jinhua, 321004, Zhejiang, China
[e] Hangzhou Med Coll, Yongkang Peoples Hosp 1, Dept Resp & Crit Care Med, Yongkang, 321300, Zhejiang, China

ARTICLE INFO

Keywords:
Multimodal medical image classification
Skin lesion classification
Cross-modal fusion
Pyramid mutual information fusion

ABSTRACT

Multimodal skin lesion classification plays a crucial role in dermatological computer-aided diagnosis by jointly leveraging dermoscopic images and patient-level clinical metadata. However, existing methods face two fundamental challenges. (1) Multi-granularity semantic coupling arises because a shared network parameter space struggles to jointly optimize fine-grained local lesion discrimination and coarse-grained global semantic representation, leading to imbalanced performance across skin lesions of different scales and morphologies. (2) Cross-modal information entropy mismatch often causes fusion collapse, where high-entropy dermoscopic image modalities dominate the joint representation, while discriminative signals from low-entropy clinical features are weakened during multimodal fusion. To address these challenges, we propose the Pyramid Mutual Information Fusion Network (PyMIF-Net) for multimodal skin lesion classification. The proposed framework comprises three key components. A CLIP Semantic Encoder (CSE) serves as a CLIP-based transfer encoder that adapts generic vision–language representations to the dermatological domain through hierarchical Adapter modules, and employs bidirectional cross-attention together with contrastive learning to align dermoscopic images with clinical semantics. A Pyramid Feature Extraction Module (PFEM) constructs a four-level feature pyramid and introduces orthogonality constraints to decouple optimization across different semantic granularities. Furthermore, a Mutual Information Constrained Fusion (MICF) module formulates an entropy-aware fusion objective via variational mutual information estimation and applies dynamic gradient reweighting to enhance the contribution of low-entropy clinical modalities. Experiments conducted on public skin lesion datasets show that PyMIF-Net achieves competitive performance compared with state-of-the-art (SOTA) methods, providing empirical evidence for its effectiveness in multimodal skin lesion classification.

1. Introduction

Multimodal skin lesion classification [1] constitutes a core technology in dermatological computer-aided diagnosis (CAD) and plays a pivotal role in skin lesion classification [2]. In real-world dermatological diagnostic practice, clinicians typically integrate dermoscopic imaging with patient-level clinical metadata to make informed decisions: the former provides visual morphological characteristics of skin lesions, while the latter encodes patients' physiological and pathological states. Effectively fusing these heterogeneous multimodal signals to emulate the clinical decision-making process is therefore crucial for improving the performance of automated skin lesion diagnosis systems [3]. Although deep learning has achieved remarkable success in unimodal dermoscopic image analysis in recent years, efficient and reliable dermoscopic image–clinical metadata fusion still faces fundamental challenges.

Existing multimodal skin lesion classification methods [4,5] are constrained by two key technical bottlenecks. (1) Degradation of multi-granularity semantic coupling. Medical lesions exhibit significant scale variability: small-scale abnormalities require fine-grained local discrimination, whereas diffuse lesions depend more on global semantic

☆ This article is part of a Special issue entitled: 'PR_Biomedical Data' published in Pattern Recognition.

* Corresponding authors.
E-mail addresses: xhy@zjnu.edu.cn (H. Xu), chenhuiling.jlu@gmail.com (H. Chen), zxz@zjnu.edu.cn (X. Zhu).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.patcog.2026.114613
Received 2 March 2026; Received in revised form 11 July 2026; Accepted 4 August 2026
Available online 10 August 2026

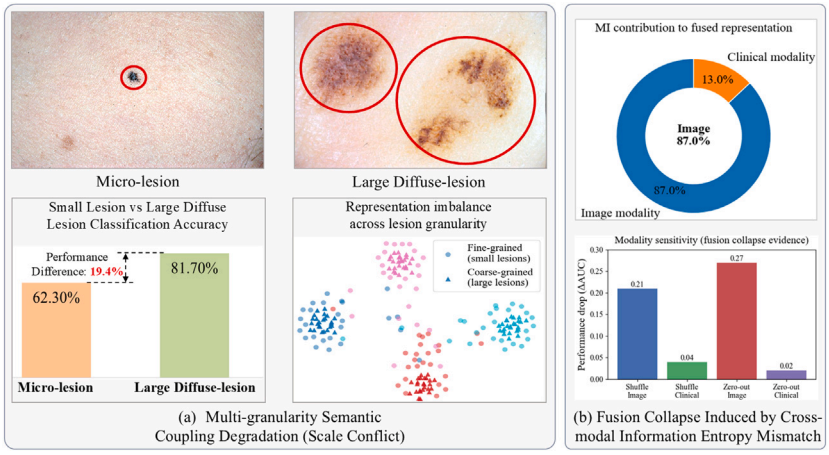

Fig. 1. Illustration of the two fundamental challenges in multimodal medical image classification.

understanding. To verify this, we analyze single-scale features extracted by ResNet-50. As shown in Fig. 1(a), the accuracy for small lesions is only 62.3%, compared to 81.7% for large diffuse lesions, yielding a 19.4 percentage-point gap. t-SNE visualization further reveals that small-lesion samples form dispersed clusters with blurred class boundaries, while large-lesion samples exhibit compact and well-separated distributions. These results suggest that a single shared network struggles to jointly optimize local discriminative sharpness and global semantic compactness, leading to imbalanced performance across lesion scales. (2) Fusion collapse caused by cross-modal information entropy mismatch. A substantial entropy gap exists between image and clinical modalities. Using Mutual Information Neural Estimation (MINE), we measure inter-modal mutual information under naive concatenation-based fusion. The mutual information between the image modality and the fused representation is 3.42 nats, whereas that of the clinical modality is only 0.51 nats. As shown in Fig. 1(b), the image modality contributes 87.0% to the fused representation, while the clinical modality accounts for merely 13.0%. This imbalance indicates that the high-entropy image modality dominates fusion, limiting the contribution of clinical information and causing multimodal learning to degenerate toward image-dominated overfitting. Modality sensitivity analysis further supports this observation: removing image information leads to a larger performance drop, whereas removing clinical information results in relatively smaller changes.

To address the above challenges, we propose PyMIF-Net (Pyramid Mutual Information Fusion Network), a deep learning framework specifically designed for multimodal medical image classification. PyMIF-Net achieves independent optimization of multi-granularity features through hierarchical parameter space decoupling and enforces entropy-balanced multimodal fusion via information-theoretic constraints. The overall architecture consists of three synergistic core modules. (1) The CSE introduces CLIP-ViT as the image encoder and employs a hierarchy-aware Adapter mechanism to facilitate domain transfer from natural images to medical images. Specifically, shallow layers preserve general visual priors, while deeper layers progressively incorporate domain-specific medical knowledge. In addition, a bidirectional cross-attention mechanism is designed to establish fine-grained correspondences between local image regions and clinical semantics. A contrastive alignment loss is further introduced to enhance cross-modal semantic consistency. (2) Hierarchically Decoupled Pyramid Feature Extraction Module. To mitigate the degradation of multi-granularity semantic coupling, the PFEM constructs a four-level heterogeneous feature pyramid. The L1 layer establishes a semantic foundation via linear projection; the L2 layer captures structural relationships using a graph attention network; the L3 layer models global contextual dependencies through multi-head self-attention; and the L4 layer enables deep cross-modal interaction via gated cross-attention. Each layer is associated with an independent parameter subspace, and orthogonality constraints are imposed to decouple optimization across granularities, thereby enabling balanced learning of fine and coarse-grained features. (3) The MICF module estimates and maximizes the mutual information between each modality and the fused representation based on the Donsker–Varadhan variational formulation. This design encourages the fused representation to preserve sufficient discriminative information from all modalities. Moreover, an entropy-aware gradient reweighting mechanism is introduced to dynamically rescale loss weights according to modality entropy ratios, compensating for the optimization disadvantage of low-entropy clinical modalities. Extensive experiments conducted on multiple public skin lesion datasets demonstrate that PyMIF-Net significantly outperforms SOTA methods in terms of classification performance. Comprehensive ablation studies further validate the individual effectiveness and synergistic benefits of the proposed modules. The main contributions of this work can be summarized as follows:

1. We empirically identify two key challenges in multimodal dermoscopy-based skin lesion classification: multi-granularity semantic coupling degradation and fusion imbalance induced by cross-modal information entropy mismatch.
2. We propose PyMIF-Net, a unified framework for dermoscopic image–clinical metadata fusion. Its core components, PFEM and MICF, are designed to alleviate inter-granularity optimization conflict and modality contribution imbalance, while CSE supports early-stage visual–clinical semantic alignment.
3. Experiments on multiple public skin lesion datasets show favorable performance of PyMIF-Net and provide empirical support for the proposed multimodal fusion and multi-granularity representation learning design.

2. Related works

In recent years, medical image analysis [3] has gradually shifted from unimodal modeling toward multimodal learning, aiming to enhance diagnostic performance by integrating heterogeneous sources of information, such as medical images and clinical data. Existing studies have demonstrated that unimodal image-based analysis is inherently constrained by the representational capacity of visual data alone [1], whereas multimodal learning can substantially improve model performance by incorporating complementary clinical information, including patient history and examination indicators. Such integration enables a more comprehensive understanding of complex disease patterns and has shown particular advantages in tasks such as early-stage skin lesion diagnosis.

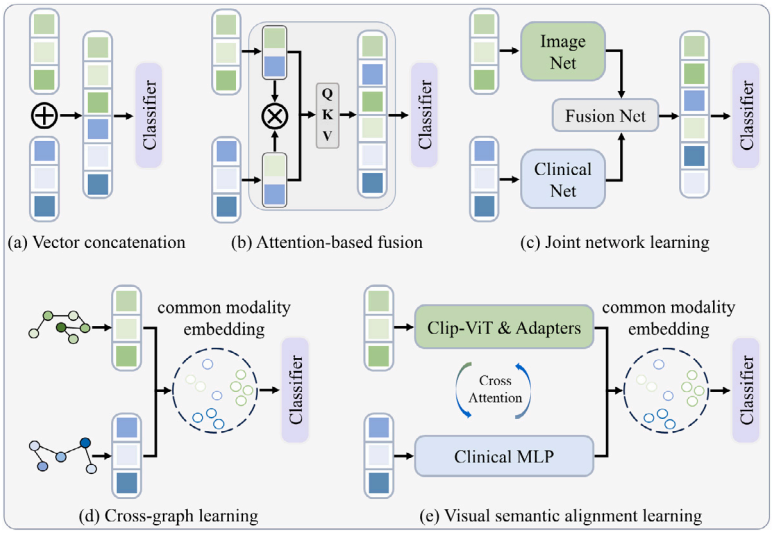

Fig. 2. Comparison of five multimodal fusion methods for medical image classification.

2.1. Cross-modal information fusion

Effective multimodal skin lesion classification requires not only combining dermoscopic image features with clinical metadata, but also aligning their semantic representations. Early approaches [6] primarily adopted feature concatenation or integration strategies, in which feature vectors from multiple modalities are directly combined for classification or regression. Although simple to implement, these methods lack explicit modeling of inter-modal relationships, making it difficult to fully exploit the complementary nature of multimodal information, and thus resulting in limited fusion performance. Subsequently, attention-based multimodal fusion methods [7] and joint network learning [8] approaches were proposed. By explicitly modeling inter-modal correlations or performing joint optimization in high-level feature spaces, these methods enhance cross-modal feature interaction. In addition, graph-based fusion methods model multimodal data as graph structures to capture complex dependencies, showing potential particularly in structured clinical data modeling.

However, most of these methods implicitly assume that the contributions of different modalities are directly comparable, without explicitly modeling differences in their information content. In practical medical scenarios, high-entropy modalities often dominate the optimization process, weakening the effective contribution of low-entropy modalities and leading to degraded fusion representations. More recently, mutual information–based multimodal learning methods have been introduced to promote cross-modal feature alignment from an information-theoretic perspective. For example, CGMCL [9] improves medical image classification by effectively integrating both image and non-image data through cross-modality graphs and contrastive learning, enhancing accuracy, interpretability, and early disease prediction. While these methods alleviate modality heterogeneity to some extent, they typically rely on static constraints and lack mechanisms to dynamically regulate the contributions of different modalities.

Meanwhile, multimodal medical image analysis based on visual pre-trained models, particularly vision–language models such as CLIP [10], has gained increasing attention. Despite their strong transferability in medical scenarios, existing methods generally adopt uniform fine-tuning strategies and perform modality interaction only at the fusion stage, lacking explicit cross-modal semantic alignment during feature encoding. This limitation restricts the consistency and stability of learned multimodal representations. Fig. 2 illustrates different multimodal fusion strategies.

2.2. Multi-granularity feature learning

Skin lesions often exhibit diverse scales, shapes, and morphological patterns. Small lesions or local structures require fine-grained discrimination, whereas diffuse lesions rely more on global contextual understanding. Therefore, multi-granularity feature learning is critical for skin lesion classification. For example, Pacal et al. [11] introduce a MetaFormer-based deep learning model that incorporates focal self-attention mechanisms, effectively integrating coarse-grained global features and fine-grained local features. Similarly, Aruk et al. [12] combine convolutional layers to capture detailed local features and transformer blocks to model global relationships, enabling the joint learning of coarse and fine-grained features for more accurate skin cancer classification. Chen et al. [13] introduce a dual-branch transformer that processes small and large patch tokens with separate branches, using cross-attention to fuse coarse-grained and fine-grained features, enabling stronger image representations for classification. Yun et al. [14] optimizes memory efficiency by using a single-head attention module that reduces computational redundancy, combining global and local information to effectively learn both coarse and fine-grained features. However, since features of different granularities often share the same parameter space, gradient conflicts can arise during training, making it difficult for the model to simultaneously maintain local discriminability and global semantic consistency, with fine-grained features often being overshadowed by coarse-grained ones. To alleviate this problem, various multi-scale modeling methods have been proposed.

To tackle the above challenges, this paper explores solutions from two perspectives: parameter space decoupling and information constraint optimization. On one hand, a hierarchical feature modeling mechanism is introduced to decouple the learning processes of different granularity features, alleviating the coupling problem of multi-granularity semantics within the shared parameter space. On the other hand, from an information theory perspective, the paper combines mutual information constraints and entropy-aware dynamic adjustment strategies to balance the optimization contributions of different modalities during multimodal fusion, thereby improving the stability and discriminative power of feature representations.

3. Methodology

As illustrated in Fig. 3, PyMIF-Net is a deep learning framework designed for multimodal skin lesion classification, aiming to address two core challenges: multi-granularity semantic coupling degradation and cross-modal fusion collapse induced by information entropy mismatch.

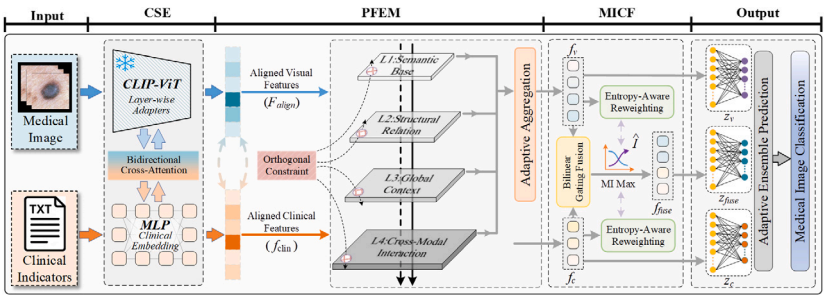

Fig. 3. Overall architecture of the proposed PyMIF-Net framework.

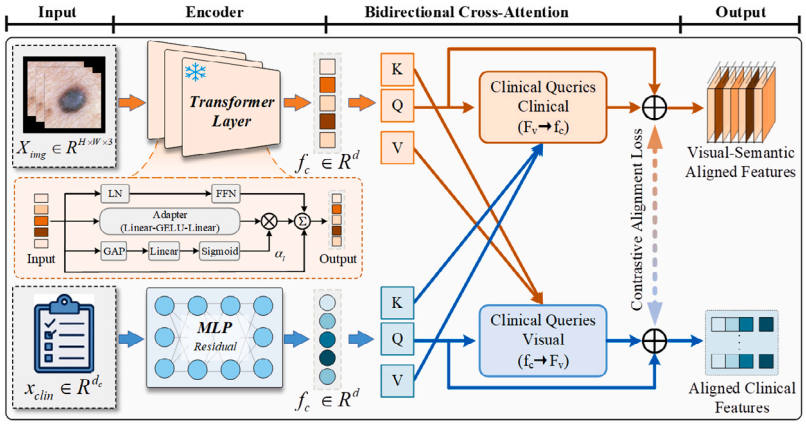

Fig. 4. Overall architecture of the proposed CSE module.

Given an input medical image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a clinical indicator vector $\mathbf{x}_{\text{clin}} \in \mathbb{R}^{d_c}$, PyMIF-Net first extracts semantically aligned visual features $\mathbf{F}_{\text{align}}$ and clinical features $\mathbf{f}_{\text{clin}}$ through the CSE module. Subsequently, the PFEM constructs a multi-granularity feature pyramid and aggregates it into $\mathbf{F}_{\text{pyr}}$. Finally, the MICF module fuses the two modalities to generate the final representation $\mathbf{f}_{\text{fuse}}$ for classification prediction. The detailed designs of each module are described below.

3.1. Visual–semantic aligned CLIP transfer encoder

Fig. 4 shows the schematic diagram of the CSE module. The CSE module leverages the transferable visual representation capability of CLIP-ViT and further learns visual–clinical semantic alignment through clinical feature embedding, bidirectional cross-attention, and contrastive supervision.

3.1.1. Adaptive adapter-based domain transfer

Since CLIP is pre-trained on large-scale natural image-text pairs, directly applying it to medical imaging tasks introduces a significant domain shift. To address this issue, we insert lightweight Adapter modules into each Transformer layer of CLIP-ViT for domain adaptation, while freezing the original backbone parameters to preserve the pre-trained knowledge. The key innovation lies in introducing a layer-wise adaptive scaling mechanism:

$$\mathbf{Z}_l = \text{FFN}(\text{LN}(\mathbf{H}_l)) + \alpha_l \cdot \mathcal{A}_l(\mathbf{H}_l) + \mathbf{H}_l, \tag{1}$$

where $\mathcal{A}_l$ denotes a bottleneck Adapter consisting of a down-projection, a non-linear activation, and an up-projection. The scaling factor

$$\alpha_l = \text{sigmoid}\left(\mathbf{w}_g^{\top} \cdot \text{GAP}(\mathbf{H}_l)\right) \tag{2}$$

is dynamically adjusted according to the feature content. This design allows shallow layers to retain general visual priors, while deeper layers progressively incorporate domain-specific medical information. After L Transformer layers, the visual encoder outputs the feature representation $\mathbf{F}_v \in \mathbb{R}^{(N+1)\times d}$.

3.1.2. Clinical feature embedding and bidirectional alignment

Clinical indicators are projected into the visual semantic space via a residual MLP:

$$\mathbf{f}_c = \text{MLP}_{\text{res}}(\mathbf{x}_{\text{clin}}) \in \mathbb{R}^d. \tag{3}$$

To establish fine-grained cross-modal associations, we design a bidirectional cross-attention mechanism that allows visual tokens to query clinical semantics, while clinical features simultaneously aggregate relevant visual information:

$$\mathbf{F}'_v = \mathbf{F}_v + \gamma_v \cdot \text{CrossAttn}(\mathbf{F}_v \to \mathbf{f}_c), \tag{4}$$

$$\mathbf{f}'_c = \mathbf{f}_c + \gamma_c \cdot \text{CrossAttn}(\mathbf{f}_c \to \mathbf{F}_v), \tag{5}$$

where γ_v and γ_c are learnable residual scaling factors initialized to zero to ensure training stability. To further enforce semantic consistency across modalities, we introduce a contrastive alignment loss:

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\bar{\mathbf{f}}'_{v,i}, \mathbf{f}'_{c,i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\bar{\mathbf{f}}'_{v,i}, \mathbf{f}'_{c,j})/\tau\right)}, \tag{6}$$

where $\bar{\mathbf{f}}'_v = \text{GAP}(\mathbf{F}'_v)$ denotes the global visual representation and $\tau = 0.07$ is the temperature parameter. The CSE module outputs the semantically aligned features $\mathbf{F}_{\text{align}} = \mathbf{F}'_v$ and $\mathbf{f}_{\text{clin}} = \mathbf{f}'_c$.

The clinical-to-visual attention is designed as a selective aggregation mechanism rather than a direct replacement of visual representations. Since clinical indicators are first projected into the visual semantic space, the clinical feature acts as a compact query to attend to clinically relevant visual tokens. The residual scaling factors γ_v and γ_c are initialized to zero, allowing the model to start from unimodal representations and gradually learn cross-modal interaction, which helps reduce the risk of noisy or unstable alignment.

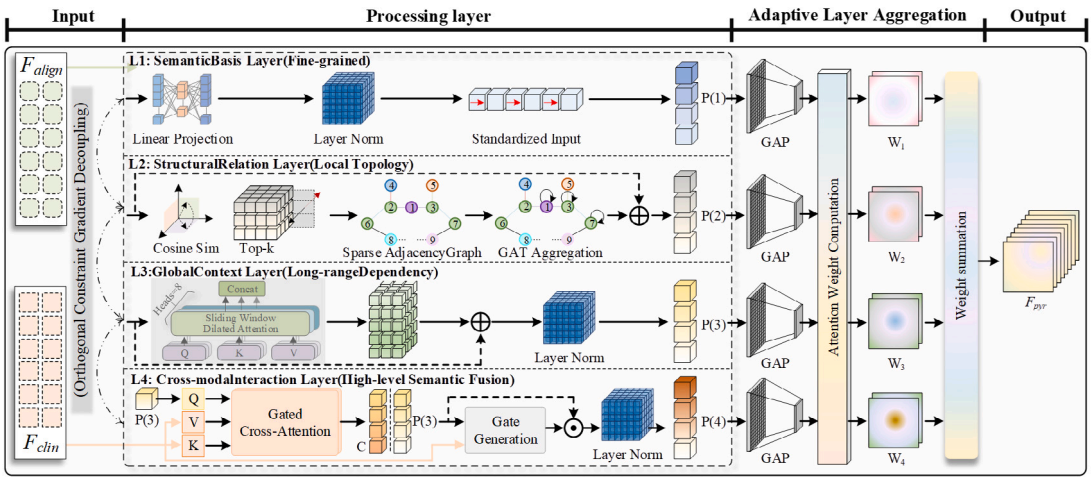

Fig. 5. Overall architecture of the proposed PFEM module.

3.2. Hierarchically decoupled pyramid feature extraction module

Fig. 5 shows the schematic diagram of the PFEM module. To address multi-granularity semantic coupling, PFEM constructs a four-level heterogeneous feature pyramid for modeling lesion patterns at different granularities. Unlike multi-branch Transformer designs that usually aggregate multi-scale feature streams with similar attention mechanisms, PFEM emphasizes inter-granularity decoupling through heterogeneous hierarchical modeling, level-specific parameter subspaces, orthogonality regularization, and gradient decoupling. It should be noted that these parameter subspaces are not strictly independent, since the pyramid levels are sequentially connected and preserve forward dependencies. Instead, "level-specific parameter subspaces" refers to the use of separate learnable transformations and regularization constraints for different granularity levels. This design aims to reduce optimization interference between fine-grained local lesion discrimination and coarse-grained global semantic modeling while maintaining hierarchical information flow, making it suitable for dermoscopy-based skin lesion classification with lesions of diverse scales and morphologies.

3.2.1. Four-level feature pyramid architecture

Based on the semantically aligned features $\mathbf{F}_{\text{align}} \in \mathbb{R}^{N\times d}$ output by the CSE module, where N denotes the number of tokens, PFEM builds a four-level heterogeneous pyramid to capture multi-granularity information ranging from local details to global semantics. Each level is designed with distinct inductive biases and receptive fields, targeting specific granularity requirements in medical image classification.

L1: Semantic base layer. The L1 layer aims to establish a unified semantic representation space that provides normalized inputs for subsequent layers. A learnable linear projection is applied to align the semantic scale of each token:

$$\mathbf{P}^{(1)} = \text{LayerNorm}(\mathbf{W}_1\mathbf{F}_{\text{align}} + \mathbf{b}_1), \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times d}$. This layer introduces no non-linear transformation, thereby preserving fine-grained discriminative information from the original features as a foundation for subsequent multi-granularity extraction. Layer normalization ensures feature distribution stability and prevents gradient vanishing in deeper layers.

L2: Structural relation layer. The L2 layer focuses on capturing local morphological characteristics and spatial topological structures of lesions. In medical images, pathological regions often exhibit specific spatial configurations (e.g., reticular pigmentation patterns), which provide critical diagnostic cues. To model such structural dependencies, we employ a Graph Attention Network (GAT) [15].

First, a sparse adjacency graph is constructed based on cosine similarity between token features. For each token i, only its top-k most similar neighbors are retained ($k = 8$), balancing receptive field coverage and computational efficiency:

$$\mathcal{N}_k(i) = \text{TopK}\left(\left\{\text{CosineSim}(\mathbf{p}_i^{(1)}, \mathbf{p}_j^{(1)})\right\}_{j=1}^{N}\right). \tag{8}$$

Neighborhood information is then aggregated using a standard GAT mechanism. For node i and its neighbor $j \in \mathcal{N}_k(i)$, the attention coefficients are computed as:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^{\top}\left[\mathbf{W}_{\text{gat}}\mathbf{p}_i^{(1)} \parallel \mathbf{W}_{\text{gat}}\mathbf{p}_j^{(1)}\right]\right), \tag{9}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'\in\mathcal{N}_k(i)}\exp(e_{ij'})}. \tag{10}$$

The output of the L2 layer for node i is obtained via weighted neighborhood aggregation with a residual connection to preserve fine-grained information:

$$\mathbf{p}_i^{(2)} = \mathbf{p}_i^{(1)} + \sum_{j\in\mathcal{N}_k(i)} \alpha_{ij}\cdot\mathbf{W}_{\text{gat}}\mathbf{p}_j^{(1)}, \tag{11}$$

where $\mathbf{W}_{\text{gat}} \in \mathbb{R}^{d\times d}$ and $\mathbf{a} \in \mathbb{R}^{2d}$ are learnable parameters. The sparse connectivity and local aggregation strategy enable this layer to specialize in fine-grained structural pattern recognition.

L3: Global context layer. The L3 layer is responsible for extracting coarse-grained global semantics and modeling long-range dependencies. Classification of diffuse lesions requires integration of global contextual information across the entire image. To this end, we adopt multi-head self-attention for global modeling.

The input feature $\mathbf{P}^{(2)}$ is linearly projected into query, key, and value representations:

$$\mathbf{Q} = \mathbf{P}^{(2)}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{P}^{(2)}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{P}^{(2)}\mathbf{W}_V. \tag{12}$$

For each attention head, global token dependencies are then modeled as:

$$\text{head}_h = \text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) = \text{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^{\top}}{\sqrt{d/H}}\right)\mathbf{V}_h. \tag{13}$$

The outputs of all heads are concatenated and projected to obtain the L3 representation:

$$\mathbf{P}^{(3)} = \text{LayerNorm}\left(\mathbf{P}^{(2)} + \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\mathbf{W}_O\right). \tag{14}$$

We use $H = 8$ attention heads with a per-head dimensionality of $d/H = 64$. Unlike the sparse local attention in the L2 layer, the fully connected attention matrix in L3 allows each token to perceive global context, making it suitable for capturing large-scale lesion distributions and holistic semantic patterns.

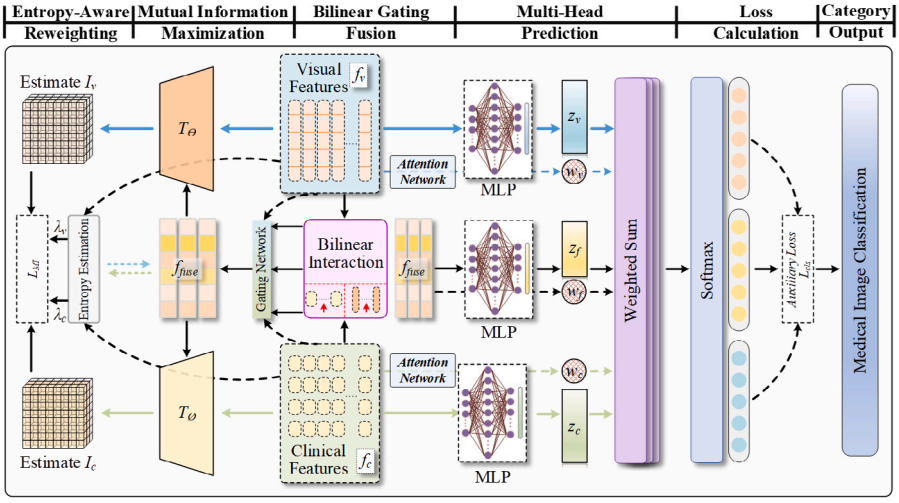

Fig. 6. Overall architecture of the proposed MICF module.

L4: Cross-modal interaction layer. At the highest semantic level, the L4 layer integrates visual and clinical modalities through deep cross-modal interaction to enhance discriminative capability. Clinical indicators (e.g., age, medical history) provide valuable prior constraints for image interpretation. This layer employs gated cross-attention to enable selective information exchange between modalities.

Visual features query clinical semantics via cross-attention, where the clinical feature $\mathbf{f}_{\text{clin}}$ serves as key–value pairs:

$$\mathbf{Q}_v = \mathbf{P}^{(3)}\mathbf{W}_{Q_v}, \quad \mathbf{K}_c = \mathbf{f}_{\text{clin}}\mathbf{W}_{K_c}, \quad \mathbf{V}_c = \mathbf{f}_{\text{clin}}\mathbf{W}_{V_c}, \tag{15}$$

$$\mathbf{C} = \text{softmax}\left(\frac{\mathbf{Q}_v\mathbf{K}_c^\top}{\sqrt{d}}\right)\mathbf{V}_c. \tag{16}$$

Since not all visual regions require modulation by clinical information, we introduce an adaptive gating mechanism to dynamically control cross-modal information flow. The gating vector is generated by jointly considering global visual features and clinical features:

$$\mathbf{G} = \sigma\left(\mathbf{W}_g\left[\text{GAP}(\mathbf{P}^{(3)}); \mathbf{f}_{\text{clin}}\right] + \mathbf{b}_g\right) \in \mathbb{R}^d. \tag{17}$$

The output of the L4 layer is obtained via gated fusion of the original visual features and the cross-modally enhanced features:

$$\mathbf{P}^{(4)} = \text{LayerNorm}\left((1-\mathbf{G}) \odot \mathbf{P}^{(3)} + \mathbf{G} \odot \mathbf{C}\right). \tag{18}$$

This design enables the model to adaptively regulate the strength of cross-modal interaction: when clinical information is beneficial for visual discrimination, $\mathbf{G} \to 1$; otherwise, the model preserves purely visual representations with $\mathbf{G} \to 0$.

3.2.2. Orthogonality constraint and gradient decoupling

To encourage lower correlation among parameter subspaces across different pyramid levels, we introduce an orthogonality regularization loss on level-specific projection matrices. Here, $\mathbf{W}^{(l)}$ does not denote all parameters of the lth PFEM operation, since different levels involve heterogeneous modules such as linear projection, graph attention, multi-head self-attention, and gated cross-attention. Instead, $\mathbf{W}^{(l)} \in \mathbb{R}^{d \times d_p}$ denotes a learnable granularity projection matrix with unified dimensionality, which maps the output representation of the lth pyramid level into a common d_p-dimensional comparison space. Therefore, the orthogonality regularization is computed only among these level-specific projection matrices:

$$\mathcal{L}_{\text{ortho}} = \sum_{l \neq m} \left\| \mathbf{W}^{(l)\top}\mathbf{W}^{(m)} \right\|_F^2. \tag{19}$$

Since all $\mathbf{W}^{(l)}$ are defined in the same projection space, the Frobenius-norm-based comparison is dimensionally consistent. This regularization encourages different pyramid levels to learn less correlated projection directions, rather than imposing strict orthogonality on the heterogeneous operations themselves.

In addition, we adopt a lightweight gradient-stopping strategy to reduce inter-granularity gradient interference. Specifically, lower-level features are detached in cross-level residual paths when constructing higher-level representations:

$$\mathbf{P}^{(l)} = \mathcal{F}^{(l)}\left(\text{sg}\left(\mathbf{P}^{(l-1)}\right)\right) + \mathbf{P}^{(l-1)}, \tag{20}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. All pyramid levels are still supervised by the final classification objective after adaptive hierarchical aggregation, rather than by separate layer-specific objectives. Thus, the stop-gradient operation is used to reduce direct cross-level gradient interference, while $\mathcal{L}_{\text{ortho}}$ encourages parameter-space decoupling.

3.2.3. Adaptive hierarchical aggregation

Features from different pyramid levels are adaptively aggregated using an attention-based mechanism. First, global pooling is applied to each layer to obtain $\mathbf{s}^{(l)} = \text{GAP}(\mathbf{P}^{(l)})$. Hierarchical weights are then computed as:

$$\mathbf{w} = \text{softmax}\left(\mathbf{W}_2 \cdot \text{ReLU}\left(\mathbf{W}_1 \cdot [\mathbf{s}^{(1)}; \ldots; \mathbf{s}^{(4)}]\right)\right). \tag{21}$$

The aggregated pyramid feature is given by

$$\mathbf{F}_{\text{pyr}} = \sum_{l=1}^{4} w_l \cdot \mathbf{s}^{(l)}. \tag{22}$$

This mechanism enables the network to dynamically adjust the contribution of each granularity according to sample characteristics: small lesions rely more on fine-grained layers, whereas diffuse lesions benefit primarily from coarse-grained layers.

3.3. Mutual information constrained adaptive fusion module

Fig. 6 shows the schematic diagram of the MICF module. MICF is designed to alleviate fusion imbalance caused by cross-modal information entropy mismatch, where the high-entropy dermoscopic image modality may dominate naive fusion and weaken the contribution of clinical metadata. To address this issue, MICF integrates four complementary components: bilinear gated fusion for cross-modal interaction modeling, variational mutual information maximization for information preservation, entropy-aware gradient reweighting for modality-balanced optimization, and multi-head adaptive ensemble for decision-level fusion.

3.3.1. Bilinear gated fusion

Given the visual feature $\mathbf{f}_v = \mathbf{F}_{\text{pyr}}$ and the clinical feature $\mathbf{f}_c = \mathbf{f}_{\text{clin}}$, we design a gated fusion mechanism based on bilinear interactions. Both modalities are first projected into a common feature space, after which low-rank bilinear pooling is employed to capture second-order cross-modal relationships:

$$\mathbf{B}(\mathbf{f}_v, \mathbf{f}_c) = \sum_{k=1}^{K} (\mathbf{u}_k^\top \mathbf{f}_v) \cdot (\mathbf{v}_k^\top \mathbf{f}_c) \cdot \mathbf{w}_k. \tag{23}$$

The gating vector integrates both first-order and second-order information:

$$\mathbf{g} = \sigma\big(\mathbf{W}_g[\mathbf{f}_v; \mathbf{f}_c; \mathbf{B}]\big), \tag{24}$$

and the fused feature is computed as

$$\mathbf{f}_{\text{fuse}} = \mathbf{g} \odot \mathbf{f}_v + (1 - \mathbf{g}) \odot \mathbf{f}_c + \mathbf{B}(\mathbf{f}_v, \mathbf{f}_c). \tag{25}$$

3.3.2. Variational mutual information maximization

To ensure that the fused representation preserves discriminative information from both modalities, we introduce a mutual information maximization constraint. Based on the Donsker–Varadhan variational formulation, a neural estimator T_θ is used to estimate a lower bound of mutual information:

$$\hat{I}(\mathbf{f}_v; \mathbf{f}_{\text{fuse}}) = \mathbb{E}_{\text{joint}}[T_\theta(\mathbf{f}_v, \mathbf{f}_{\text{fuse}})] - \log \mathbb{E}_{\text{marginal}}\left[e^{T_\theta(\mathbf{f}_v, \mathbf{f}_{\text{fuse}})}\right]. \tag{26}$$

The statistics network T_θ is implemented as a three-layer MLP, taking as input the concatenation of the two features and their element-wise product to capture interaction information. The clinical–fusion mutual information $\hat{I}(\mathbf{f}_c; \mathbf{f}_{\text{fuse}})$ is symmetrically estimated using an independent network T_ϕ with a smaller learning rate, gradient clipping, numerical stabilization of the exponential term, and an exponential moving average for the marginal expectation, while treating the MI term as a regularization objective to encourage information preservation rather than an exact mutual information estimate.

3.3.3. Entropy-aware gradient reweighting

To compensate for the entropy mismatch across modalities, we design an entropy-aware loss reweighting strategy. The differential entropy of each modality is estimated on each mini-batch via kernel density estimation:

$$\hat{H}(\mathbf{f}) = -\frac{1}{B} \sum_{i=1}^{B} \log\left(\frac{1}{B} \sum_{j=1}^{B} \kappa_h(\mathbf{f}^{(i)} - \mathbf{f}^{(j)})\right), \tag{27}$$

where κ_h denotes a Gaussian kernel. Based on the estimated entropies, modality balancing weights are defined as

$$\lambda_v = \frac{\hat{H}(\mathbf{f}_c)}{\hat{H}(\mathbf{f}_v) + \hat{H}(\mathbf{f}_c)}, \quad \lambda_c = \frac{\hat{H}(\mathbf{f}_v)}{\hat{H}(\mathbf{f}_v) + \hat{H}(\mathbf{f}_c)}. \tag{28}$$

Notably, the low-entropy modality receives a larger weight, compensating for its disadvantage in gradient propagation. The mutual information loss is therefore defined as

$$\mathcal{L}_{\text{MI}} = -\lambda_v \hat{I}(\mathbf{f}_v; \mathbf{f}_{\text{fuse}}) - \lambda_c \hat{I}(\mathbf{f}_c; \mathbf{f}_{\text{fuse}}), \tag{29}$$

which encourages balanced contributions from both modalities during fusion-layer optimization.

3.3.4. Multi-head prediction and adaptive ensemble

To fully exploit discriminative information at different abstraction levels from the visual feature $\mathbf{f}_v$, the clinical feature $\mathbf{f}_c$, and the fused feature $\mathbf{f}_{\text{fuse}}$, we design a multi-head prediction mechanism. Each feature branch independently predicts class logits:

$$\mathbf{z}_v = C_v(\mathbf{f}_v), \quad \mathbf{z}_c = C_c(\mathbf{f}_c), \quad \mathbf{z}_{\text{fuse}} = C_{\text{fuse}}(\mathbf{f}_{\text{fuse}}), \tag{30}$$

where C_* denotes a two-layer MLP classifier with a hidden dimension equal to half of the input feature dimension, GELU activation, and a dropout rate of 0.3.

For adaptive integration, an attention-based prediction aggregation mechanism is introduced:

$$\boldsymbol{\omega} = \text{softmax}\left(\mathbf{W}_\omega[\mathbf{f}_v; \mathbf{f}_c; \mathbf{f}_{\text{fuse}}]\right), \tag{31}$$

$$\mathbf{z} = \omega_v \mathbf{z}_v + \omega_c \mathbf{z}_c + \omega_{\text{fuse}} \mathbf{z}_{\text{fuse}}, \tag{32}$$

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}). \tag{33}$$

This design enables the network to dynamically adjust the contribution of each branch according to input characteristics: for samples where unimodal information is sufficient, the corresponding modality branch is emphasized; for samples requiring cross-modal complementarity, the fusion branch dominates the prediction.

During training, all three branches are supervised. The classification loss is defined as

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{cls}}^{\text{main}}(\mathbf{z}, \mathbf{y}) + \mu\left[\mathcal{L}_{\text{cls}}^{\text{aux}}(\mathbf{z}_v, \mathbf{y}) + \mathcal{L}_{\text{cls}}^{\text{aux}}(\mathbf{z}_c, \mathbf{y}) + \mathcal{L}_{\text{cls}}^{\text{aux}}(\mathbf{z}_{\text{fuse}}, \mathbf{y})\right], \tag{34}$$

where the main loss $\mathcal{L}_{\text{cls}}^{\text{main}}$ adopts a class-balanced label-smoothed cross-entropy (described below), and the auxiliary losses $\mathcal{L}_{\text{cls}}^{\text{aux}}$ are standard cross-entropy losses. The weight μ is set to 0.4 and linearly decayed to zero in the later training stage to avoid over-regularization.

3.4. Loss function and optimization

The overall training objective is composed of four terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{MI}} + \beta\mathcal{L}_{\text{align}} + \gamma\mathcal{L}_{\text{ortho}}. \tag{35}$$

Classification loss. The classification loss $\mathcal{L}_{\text{cls}}$ is defined in Section 3.3 and consists of the main prediction loss and three auxiliary branch losses. The main loss adopts a class-balanced label-smoothed cross-entropy, where n_c denotes the number of samples in class c:

$$w_c = \frac{1/\sqrt{n_c}}{\sum_{k=1}^{C} 1/\sqrt{n_k}}, \quad \mathcal{L}_{\text{cls}}^{\text{main}} = -\sum_{c=1}^{C} w_c \tilde{y}_c \log \hat{y}_c, \quad \tilde{y}_c = (1 - \epsilon) y_c + \epsilon/C. \tag{36}$$

The label smoothing parameter is set to $\epsilon = 0.1$. The class-balanced weights encourage rare classes to receive higher loss weights, alleviating class imbalance.

The remaining loss components correspond to the three core constraints of the proposed model. The *mutual information loss* $\mathcal{L}_{\text{MI}}$ (Section 3.3) maximizes the mutual information between each modality and the fused representation, $\hat{I}(\mathbf{f}_v; \mathbf{f}_{\text{fuse}})$ and $\hat{I}(\mathbf{f}_c; \mathbf{f}_{\text{fuse}})$, with entropy-aware weights λ_v and λ_c, ensuring that discriminative information from both modalities is preserved during fusion. The *alignment loss* $\mathcal{L}_{\text{align}}$ (Section 3.1.2) adopts an InfoNCE-style contrastive objective to establish discriminative correspondences between local visual features and clinical semantics during encoding. The *orthogonality loss* $\mathcal{L}_{\text{ortho}}$ (Section 3.2.2) minimizes the inner products between parameter matrices of different PFEM layers, $\|\mathbf{W}^{(l)\top}\mathbf{W}^{(m)}\|_F^2$, enforcing independent optimization of multi-granularity features in orthogonal parameter subspaces.

4. Experiments

4.1. Dataset

To comprehensively evaluate the proposed method under diverse data distributions and realistic clinical scenarios, experiments are conducted on three widely used multimodal skin lesion benchmarks: HAM10000 [16], ISIC 2019 [17], and Derm7pt [18]. The HAM10000 dataset contains 10,015 dermoscopic images spanning seven diagnostic categories with pronounced class imbalance, where NV is the dominant class (6705 images) and DF is the smallest (115 images). In addition

to images, it provides clinical metadata such as patient age, sex, lesion location, and diagnostic method, enabling multimodal analysis. ISIC 2019 is a larger-scale benchmark comprising 25,331 dermoscopic images from eight categories, also exhibiting significant class imbalance (NV: 12,875 images; DF: 239 images), and includes clinical information such as age, sex, and anatomical site, with an approximately balanced gender distribution. To further assess robustness and fine-grained feature modeling capability, the Derm7pt dataset is employed. It contains 1011 cases with dermoscopic images and clinical metadata, annotated according to the seven-point checklist, forming seven dermoscopic feature classification tasks, including both binary and multi-class tasks, along with a five-class diagnostic task (DIAG). Collectively, these datasets support multi-class, multi-label, and multimodal learning settings, enabling comprehensive evaluation of the proposed method in large-scale screening and fine-grained diagnostic scenarios. The dataset is partitioned according to the official method.

4.2. Implementation details

All experiments were implemented in PyTorch 2.8.0 with Python 3.12 on Ubuntu 22.04 and CUDA 12.8, using a single NVIDIA GeForce RTX 5090 GPU with 32 GB VRAM. We adopt a two-stage training strategy. In the first stage, the original CLIP parameters are frozen, and only the Adapter modules, PFEM, and MICF are optimized with a learning rate of $\eta_1 = 10^{-4}$. In the second stage, the CLIP backbone is unfrozen and fine-tuned end-to-end with a reduced learning rate of $\eta_2 = 0.1\eta_1$. The batch size is set to 32 for all main experiments and ablation studies. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$ and a weight decay of 10^{-4}, together with cosine annealing and warm-up. The model is trained for 100 epochs.

PyMIF-Net contains 87.51 M parameters and requires 4.45 GFLOPs, with an inference latency of 1.15 ms on the RTX 5090 GPU. Although its CLIP-based visual backbone increases the parameter count, the graph-attention and cross-modal fusion backend remains lightweight by operating in a low-dimensional feature space. For comparison, CGMCL [9] has fewer parameters (15.07 M) and an estimated latency of 1.8 ms, but requires 68.052 GFLOPs. These results suggest that PyMIF-Net achieves a favorable trade-off between model capacity, computational cost, latency, and classification performance. The training time is approximately 42.5 s per epoch, and the GPU memory footprint is 2150.4 MB.

4.3. Evaluation metrics

Following common practice in skin lesion classification, we evaluate all methods using Accuracy, Precision, Recall, and F1-score. These metrics provide complementary perspectives on overall correctness, positive prediction reliability, sensitivity to positive samples, and the balance between Precision and Recall, respectively. Since they are standard evaluation criteria widely used in this field, we omit detailed formula derivations.

4.4. Quantitative experiments

To comprehensively evaluate PyMIF-Net, we compare it with recent SOTA methods on the ISIC 2019, HAM10000, and Derm7pt datasets. As shown in Tables 1, 2, and 3, PyMIF-Net consistently achieves superior performance across lesion classification, fine-grained dermoscopic feature recognition, and multi-class diagnosis. On ISIC 2019 and HAM10000, the proposed method achieves approximately 0.95 Accuracy and F1-score, outperforming representative CNN-, Transformer-, and multimodal-based approaches. On Derm7pt, PyMIF-Net achieves an Accuracy and F1-score of 0.93 with stable performance across five independent runs, while also obtaining the best results on all seven dermoscopic feature subtasks, with particularly notable improvements on DaG, PIG, and PN. Moreover, it reaches an Accuracy of 0.98 on the DIAG task, surpassing the strongest competing methods at 0.94. These results demonstrate that the proposed multi-granularity feature decoupling and multimodal fusion strategy effectively improves both overall classification and fine-grained diagnostic performance across different datasets.

Table 1
Comparison with existing methods on ISIC 2019 and HAM10000. The best performance results are highlighted in **bold**, and underlines indicate the second-best methods based on the original unrounded values.

Dataset	Method	Year	ACC	F1-Score	Precision	Recall
ISIC 2019	GoogleNet [17]	2020	0.85	0.80	0.80	0.80
	InSiNet [19]	2022	0.92	0.84	0.88	0.90
	DenseNet-201 [20]	2022	0.93	0.94	0.94	0.94
	Inception-ResNet [21]	2022	0.86	0.82	0.85	0.80
	ConvNeXt [22]	2023	0.87	0.88	0.87	0.89
	ConvNeXt [12]	2025	0.93	0.90	0.92	0.89
	ConvNeXtV2 [23]	2025	0.94	0.91	0.92	0.90
	MetaFormer [11]	2025	0.93	0.89	0.90	0.88
	ConvNeXtV2 [24]	2025	0.93	0.92	0.93	0.91
	PyMIF-Net (Ours)	–	**0.95**	**0.95**	**0.95**	**0.94**
HAM10000	DCNN [25]	2021	0.91	**0.95**	**0.97**	0.94
	VGG-16 [26]	2021	0.91	0.77	0.88	0.74
	CrossViT [13]	2021	0.93	0.90	0.91	0.89
	TinyViT [27]	2022	0.93	0.90	0.89	0.91
	DenseNet [28]	2023	0.95	0.94	**0.97**	0.92
	ALBEF [29]	2024	0.94	0.90	0.91	0.90
	ConvNeXt [12]	2025	0.94	0.91	0.93	0.90
	MetaFormer [11]	2025	0.95	0.93	0.95	0.92
	EffiCAT [30]	2025	0.94	0.94	0.94	0.94
	PyMIF-Net (Ours)	–	**0.95**	**0.95**	0.95	**0.95**

4.5. Qualitative experiments

To intuitively analyze the model's prediction behavior, we visualize confusion matrices for the DIAG task and seven dermoscopic feature prediction tasks in Fig. 7. Overall, most samples are concentrated along the diagonal, suggesting good discriminative performance across both diagnostic and attribute-level classifications. For the DIAG task, only limited confusion is observed between visually similar disease categories. For dermoscopic feature prediction, clear diagonal patterns are also observed in both binary tasks (e.g., BWV, RS) and multi-class tasks (e.g., PN, VS, PIG, STR, DaG). Most misclassifications occur between semantically or morphologically related categories, reflecting the intrinsic difficulty of fine-grained dermoscopic feature recognition. These results suggest that the proposed method achieves stable predictions across diagnostic and attribute-level tasks.

To further support the quantitative results, we visualize attention distributions of different models in Fig. 8, covering diffuse-type, conventional, and micro-sized lesions. Compared with ResNet-50 and Graph-ViT, the proposed method produces more lesion-focused and spatially coherent activations. For diffuse-type lesions, our method captures more complete global lesion regions with fewer redundant background responses. For conventional lesions, it provides more concentrated activations around diagnostically relevant structures. For micro-sized lesions, it better highlights compact lesion areas, suggesting improved fine-grained localization. Overall, these visualizations suggest that the proposed framework better balances global context modeling and local detail preservation.

4.6. Ablation study

4.6.1. Module-level ablation study

To validate the contribution of the three core components in PyMIF-Net, we conduct module-level ablation experiments on the Derm7pt

Table 2
Comparison of diagnostic classification performance on the Derm7pt dataset. The best performance results are highlighted in **bold**, and underlines indicate the second-best methods.

Models	Backbone	Image	Clinical	ACC	F1-score	Precision	Recall
Logistic	–	–	✓	0.80	0.80	0.80	0.80
XGBoost	–	–	✓	0.80	0.78	0.78	0.77
AdaBoost	–	–	✓	0.79	0.78	0.78	0.78
Unimodal	–	2-layer CNN	✗	0.86	0.86	0.87	0.86
Unimodal	–	VGG19	✗	0.87	0.87	0.88	0.86
Unimodal	–	ResNet18	✗	0.90	0.90	0.91	0.89
CoBFormer [4]	Graph-Transformer	–	✓	0.90	0.89	0.90	0.89
GTC [31]	Graph-Transformer	–	✓	0.89	0.89	0.91	0.88
Cluster-GT-T [32]	Graph-Transformer	–	✓	0.88	0.88	0.91	0.86
Cluster-GT-L [32]	Graph-Transformer	–	✓	0.87	0.87	0.89	0.86
MHCA [5]	Transformer	3D-CNN	✓	0.87	0.88	0.85	**0.92**
DeAF [33]	Transformer	3D-CNN	✓	0.88	0.88	0.90	0.87
TriFormer [34]	Transformer	3D-CNN	✓	0.83	0.84	0.83	0.86
GEAET [35]	Graph-Transformer	ResNet18	✓	0.90	0.90	0.91	0.89
CGMCL [9]	Graph-Transformer	ResNet18	✓	0.90	0.91	0.91	0.90
PyMIF-Net (Res)	Hierarchically-Pyramid	ResNet18	✓	0.92	0.92	0.92	0.91
PyMIF-Net (Ours)	Hierarchically-Pyramid	CLIP	✓	**0.93**	**0.93**	**0.94**	**0.92**

Table 3
Comparison of the accuracy of multi-class classification performance on Derm7pt dataset. The best performance results are highlighted in **bold**, and underlines indicate the second-best methods.

Model	Backbone	BWV	DaG	PIG	PN	RS	STR	VS	DIAG
Logistic	–	0.83	0.58	0.51	0.68	0.77	0.73	0.83	0.75
XGboost	–	0.82	0.49	0.68	0.67	0.76	0.74	0.84	0.74
AdaBoost	–	0.83	0.54	0.64	0.60	0.77	0.68	0.84	0.66
ResNet18	CNN	0.84	0.48	0.59	0.65	0.74	0.64	0.82	0.75
2-layer	CNN	0.80	0.43	0.60	0.56	0.71	0.69	0.80	0.72
7-point [18]	CNN	0.85	0.60	0.63	0.69	0.77	0.74	0.82	0.73
HcCNN [36]	CNN	0.87	0.66	0.69	0.71	0.81	0.72	0.85	0.74
AMFAM [37]	GAN	0.88	0.64	0.71	0.71	0.81	0.75	0.83	0.75
FusionM4Net [38]	CNN	0.89	0.66	0.72	0.69	0.81	0.76	0.82	0.76
SHViT [14]	ViT	0.82	0.56	0.71	0.62	0.70	0.78	0.83	0.93
BIIGMA [39]	Vector	0.84	0.64	0.71	0.72	0.70	0.77	0.80	0.92
PLRR [40]	ViT	0.86	0.66	0.74	0.76	0.77	0.78	0.87	0.94
CGMCL [9]	ResNet18	0.87	0.67	0.74	0.75	0.78	0.77	0.87	0.94
PyMIF-Net (Res)	ResNet18	0.90	0.79	0.80	0.81	0.85	0.83	0.89	0.96
PyMIF-Net (Ours)	CLIP-ViT	**0.93**	**0.85**	**0.88**	**0.87**	**0.92**	**0.87**	**0.91**	**0.98**

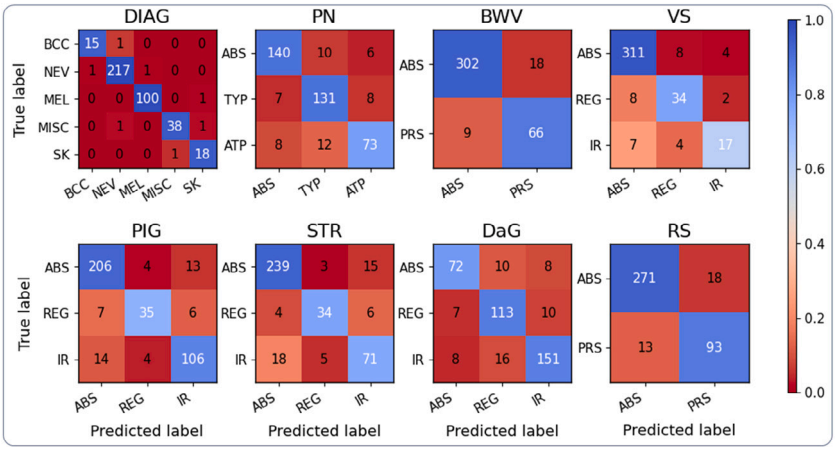

Fig. 7. Confusion matrices for the seven dermoscopic feature prediction tasks and the DIAG task on the Derm7pt dataset.

dataset. Specifically, we remove CSE, PFEM, and MICF respectively while keeping all other training settings unchanged. As shown in Table 4, removing the CSE module results in the most significant performance degradation, with ACC dropping from 0.94 to 0.84 and F1-score decreasing from 0.93 to 0.85. This indicates that early-stage cross-modal semantic alignment plays a critical role in reducing modality discrepancy and facilitating effective feature fusion. When the PFEM module is removed, the performance declines moderately (ACC: 0.87, F1-score: 0.88), suggesting that multi-granularity pyramid modeling enhances the representation of lesions with varying spatial scales. Similarly, excluding the MICF module leads to consistent performance degradation (ACC: 0.90, F1-score: 0.91), validating the importance of mutual-information-guided fusion and entropy-aware balancing in preventing modality dominance during optimization. The results demonstrate that the three core modules are complementary and jointly contribute to the superior performance of PyMIF-Net.

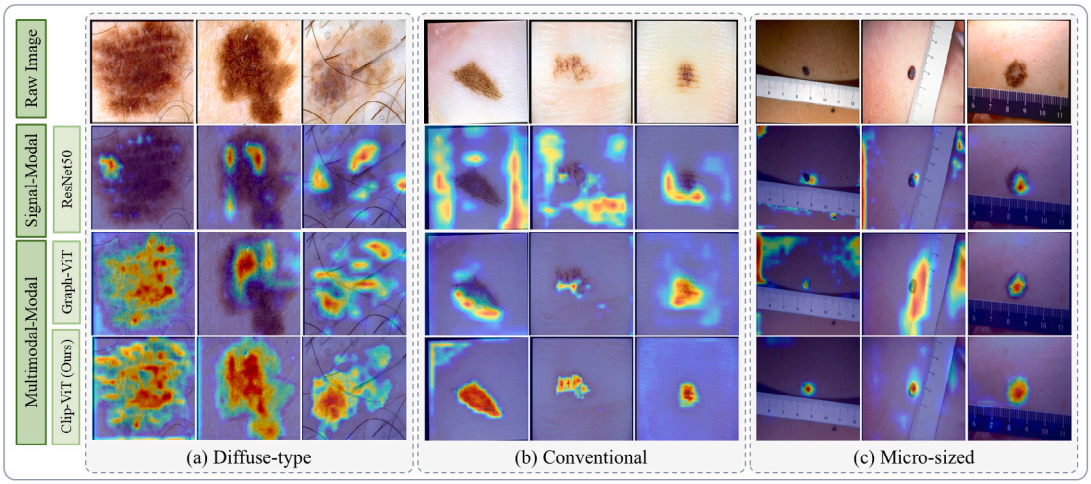

Fig. 8. Grad-CAM visualization of multimodal feature maps across lesion types.

Table 4
Module-level ablation results on the Derm7pt dataset.

Model variant	ACC	F1-score	Precision	Recall
Full PyMIF-Net	**0.94**	**0.93**	**0.94**	**0.92**
w/o CSE	0.84	0.85	0.83	0.86
w/o PFEM	0.87	0.87	0.86	0.87
w/o MICF	0.90	0.90	0.89	0.90

To provide a more fine-grained evaluation of the proposed components, we further conduct a component-level ablation study on the Derm7pt dataset, as shown in Table 5. For CSE, we individually remove the Adapter-based domain transfer module, bidirectional cross-attention, and contrastive alignment loss. For PFEM, we ablate the key functional levels, including the L2 structural relation layer, the L3 global context layer, and the L4 cross-modal interaction layer, as well as the orthogonality constraint. The L1 semantic base layer is retained in all variants because it mainly serves as the basic feature projection and normalization layer, and removing it would change the input dimensional consistency of subsequent layers. For MICF, we remove bilinear gated fusion, mutual information loss, entropy-aware weighting, and multi-head adaptive ensemble separately. The results show that removing any of these components leads to performance degradation, indicating that each component contributes to the final performance. Specifically, the CSE components improve medical-domain adaptation and cross-modal semantic alignment; the PFEM components contribute to local structural modeling, global contextual reasoning, and high-level cross-modal interaction; and the MICF components respectively enhance cross-modal interaction modeling, information preservation, modality-balanced optimization, and decision-level adaptive fusion.

4.7. Ablation study on loss function combinations

To analyze the contribution of each optimization objective, we conduct loss-combination ablation experiments on the Derm7pt dataset. The complete loss function is described in Section 3.4. In this study, we progressively add the proposed loss terms to $\mathcal{L}_{cls}$. As reported in Table 6, using only the classification loss $\mathcal{L}_{cls}$ yields the weakest performance, indicating that the model suffers from insufficient cross-modal consistency and unstable multi-objective optimization without additional regularization. Adding the alignment loss $\mathcal{L}_{align}$ brings a clear improvement, suggesting that explicit cross-modal semantic alignment directly benefits the discriminative learning objective. Introducing the mutual-information regularization $\mathcal{L}_{MI}$ also improves performance, confirming its effectiveness in preserving informative cross-modal dependency during fusion. The orthogonality constraint $\mathcal{L}_{ortho}$ provides consistent but relatively smaller gains, implying that it mainly contributes to optimization stability by mitigating representation entanglement. When combining $\mathcal{L}_{align}$ with $\mathcal{L}_{MI}$ (and further with $\mathcal{L}_{ortho}$), the performance improves steadily, demonstrating that these objectives are complementary. The full objective achieves the best results, validating the necessity of jointly optimizing alignment, mutual-information preservation, and feature decoupling for robust multimodal diagnosis on Derm7pt.

4.8. Ablation study of modality sensitivity

To verify whether the proposed method alleviates image-dominated fusion and enhances the contribution of clinical metadata, we conduct a modality sensitivity analysis on the Derm7pt dataset. As shown in Table 7, the naive fusion baseline shows only a slight performance decrease when clinical metadata is removed, with ACC dropping from 0.89 to 0.88. In contrast, removing the image modality leads to a much larger decrease, with ACC dropping from 0.89 to 0.80. This suggests that naive fusion relies mainly on dermoscopic image features and makes relatively limited use of clinical metadata. For PyMIF-Net, removing clinical metadata causes a more noticeable performance drop, with ACC decreasing from 0.93 to 0.90. This indicates that the proposed framework makes better use of clinical information during multimodal fusion. Meanwhile, removing the image modality still results in a clear degradation, with ACC decreasing from 0.93 to 0.84, which is expected because dermoscopic images remain the primary diagnostic source for skin lesion classification. Overall, these results provide empirical support that PyMIF-Net helps alleviate image-dominated fusion and enhances the contribution of clinical metadata.

4.9. Hyper-parameter ablation of loss weights

To analyze the effect of loss weighting, we conduct a hyper-parameter ablation study on the Derm7pt dataset by varying α, β, and γ in the overall objective, while keeping the other coefficients fixed. As shown in Table 8, the best performance is achieved at (α=1.0, β=1.5, γ=0.5). The results show that removing or overemphasizing any regularization term degrades performance, suggesting that mutual-information preservation, cross-modal alignment, and feature decoupling should be properly balanced. Specifically, $\mathcal{L}_{align}$ benefits from a relatively larger weight because it strengthens early visual–clinical semantic alignment, $\mathcal{L}_{MI}$ uses a moderate weight to balance information preservation and discriminative learning, and $\mathcal{L}_{ortho}$ requires only a smaller coefficient because it mainly acts as a structural regularizer and may over-constrain feature learning if weighted too heavily.

Table 5
Fine-grained ablation study of PyMIF-Net on the Derm7pt dataset.

Module	Variant	ACC	F1-score	Precision	Recall
Full Model	Full PyMIF-Net	0.94	0.93	0.94	0.92
CSE	w/o Adapter	0.88	0.87	0.87	0.88
	w/o Bidirectional Cross-Attention	0.90	0.90	0.89	0.90
	w/o Contrastive Alignment	0.91	0.90	0.90	0.91
PFEM	w/o L2 Structural Relation Layer	0.90	0.89	0.89	0.90
	w/o L3 Global Context Layer	0.91	0.90	0.90	0.91
	w/o L4 Cross-Modal Interaction Layer	0.89	0.89	0.88	0.89
	w/o Orthogonality Constraint	0.92	0.91	0.91	0.92
MICF	w/o Bilinear Gated Fusion	0.91	0.90	0.90	0.91
	w/o MI Loss	0.91	0.91	0.90	0.91
	w/o Entropy-aware Weighting	0.90	0.89	0.89	0.90
	w/o Multi-head Adaptive Ensemble	0.92	0.92	0.91	0.92

Table 6
Loss-combination ablation results on the Derm7pt dataset.

Loss setting	ACC	F1-score	Precision	Recall
$\mathcal{L}_{cls}$	0.89	0.88	0.88	0.89
$\mathcal{L}_{cls} + \mathcal{L}_{align}$	0.91	0.90	0.90	0.91
$\mathcal{L}_{cls} + \mathcal{L}_{MI}$	0.90	0.89	0.89	0.90
$\mathcal{L}_{cls} + \mathcal{L}_{ortho}$	0.90	0.89	0.89	0.90
$\mathcal{L}_{cls} + \mathcal{L}_{align} + \mathcal{L}_{MI}$	0.92	0.92	0.91	0.92
$\mathcal{L}_{cls} + \mathcal{L}_{align} + \mathcal{L}_{ortho}$	0.92	0.91	0.91	0.92
$\mathcal{L}_{cls} + \mathcal{L}_{MI} + \mathcal{L}_{ortho}$	0.91	0.91	0.90	0.91
$\mathcal{L}_{cls} + \mathcal{L}_{MI} + \mathcal{L}_{align} + \mathcal{L}_{ortho}$ (Full)	**0.93**	**0.93**	**0.94**	**0.92**

Table 7
Modality sensitivity analysis on the Derm7pt dataset.

Setting	ACC	F1-score	Precision	Recall
Naive Fusion (Image + Clinical)	0.89	0.89	0.90	0.88
Naive Fusion (w/o Clinical)	0.88	0.88	0.89	0.87
Naive Fusion (w/o Image)	0.80	0.80	0.81	0.79
PyMIF-Net (Image + Clinical)	0.93	0.93	0.94	0.92
PyMIF-Net (w/o Clinical)	0.90	0.90	0.91	0.89
PyMIF-Net (w/o Image)	0.84	0.84	0.85	0.83

Table 8
Hyper-parameter ablation of loss weights on the Derm7pt dataset.

Setting	ACC	F1-score	Precision	Recall
Best Setting (α=1.0, β=1.5, γ=0.5)	**0.93**	**0.93**	**0.94**	**0.92**
Varying α (fix β=1.5, γ=0.5)				
α=0	0.89	0.89	0.90	0.88
α=0.5	0.91	0.91	0.92	0.90
α=1.0	**0.93**	**0.93**	**0.94**	**0.92**
α=2.0	0.92	0.92	0.92	0.91
Varying β (fix α=1.0, γ=0.5)				
β=0	0.90	0.90	0.91	0.89
β=1.0	0.92	0.92	0.92	0.91
β=1.5	**0.93**	**0.93**	**0.94**	**0.92**
β=2.0	0.92	0.91	0.92	0.90
Varying γ (fix α=1.0, β=1.5)				
γ=0	0.91	0.91	0.91	0.90
γ=0.5	**0.93**	**0.93**	**0.94**	**0.92**
γ=1.0	0.92	0.92	0.92	0.91
γ=2.0	0.90	0.90	0.90	0.89

5. Conclusion

In this work, we address two key challenges in multimodal skin lesion classification: multi-granularity semantic misalignment and fusion imbalance caused by cross-modal information entropy mismatch. To this end, we propose PyMIF-Net, a pyramid mutual information fusion framework that integrates cross-modal semantic alignment, multi-granularity feature modeling, and entropy-aware fusion. Specifically, CSE promotes visual–clinical semantic alignment, PFEM encourages decoupled multi-granularity lesion representation, and MICF improves modality-balanced fusion between dermoscopic images and clinical metadata. Experiments on HAM10000, ISIC 2019, and Derm7pt show that PyMIF-Net achieves favorable performance compared with SOTA methods, while qualitative results further suggest improved global context modeling and local lesion localization. Overall, this work provides an information-theoretic perspective for dermoscopic image–clinical metadata fusion and offers a promising solution for multimodal skin lesion classification. Nevertheless, several limitations remain, including the reliance on natural-image pre-trained vision–language models, the additional training overhead introduced by mutual information estimation and entropy-aware reweighting, and the need for further validation on multi-center clinical datasets and broader medical scenarios.

CRediT authorship contribution statement

Chengcheng Li conducted the research and wrote the manuscript. Huiying Xu, Huiling Chen, and Xinzhong Zhu supervised the study and revised the manuscript. Zhihua Wu, Zhendong Chen, Yue Hu, Yue Chen and Xinkai Xu contributed to data analysis and experiments. All authors reviewed and approved the final manuscript.

Declaration of generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used in the preparation of this manuscript. All content is the original work of the authors.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376252, 32470395, U25A20450, 62571374); the Mindray Joint Fund of Zhejiang Provincial Natural Science Foundation of China (LMRY26H180015); the Zhejiang Provincial Natural Science Foundation of China (LY24C020002); the Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056); the Zhejiang Province Province-Land Synergy Program (2025SDXT004-3) .

Data availability

All data generated or analyzed during this study are available to readers upon request to the first author.

References

[1] M. Watson, T. Winterbottom, T. Hudson, B. Jones, H.P. Shum, A. Atapour-Abarghouei, T. Breckon, J. Harmsworth King, N. Al Moubayed, Multimodal models for skin cancer classification using clinical freetext and dermatoscopic images, Commun. Med. (2026).

[2] V. Gabani, T. Navamani, K. Shyamala, V.K. Vaswani Rajpal, Multimodal skin lesion classification for early cancer diagnosis using deep learning, Front. Physiol. 17 (2026) 1717517.

[3] M.A. Qazi, A.U.R. Hashmi, S. Sanjeev, I. Almakky, N. Saeed, C. Gonzalez, M. Yaqub, Continual learning in medical imaging: A survey and practical analysis, ACM Comput. Surv. 58 (8) (2026) 1–25.

[4] Y. Xing, X. Wang, Y. Li, H. Huang, C. Shi, Less is more: on the over-globalizing problem in graph transformers, 2024, arXiv preprint arXiv:2405.01102.

[5] J.-E. Ding, C.-C. Hsu, C.-H. Chu, S. Wang, F. Liu, Enhancing multimodal medical image classification using cross-graph modal contrastive learning, 2024, arXiv preprint arXiv:2410.17494.

[6] J. Yap, W. Yolland, P. Tschandl, Multimodal skin lesion classification using deep learning, Exp. Dermatol. 27 (11) (2018) 1261–1267.

[7] X. He, Y. Wang, S. Zhao, X. Chen, Co-attention fusion network for multimodal skin cancer diagnosis, Pattern Recognit. 133 (2023) 108990.

[8] S. Sharma, M. Vardhan, MTJNet: Multi-task joint learning network for advancing medicinal plant and leaf classification, Knowl.-Based Syst. 299 (2024) 112147.

[9] J.-E. Ding, C.-C. Hsu, C.-H. Chu, S. Wang, F. Liu, Enhancing multimodal medical image classification through cross-graph modal contrastive learning, Expert Syst. Appl. (2025) 130566.

[10] Z. Zhao, Y. Liu, H. Wu, M. Wang, Y. Li, S. Wang, L. Teng, D. Liu, Z. Cui, Q. Wang, et al., CLIP in medical imaging: A survey, Med. Image Anal. 102 (2025) 103551.

[11] I. Pacal, B. Ozdemir, J. Zeynalov, H. Gasimov, N. Pacal, A novel CNN-ViT-based deep learning model for early skin cancer diagnosis, Biomed. Signal Process. Control. 104 (2025) 107627.

[12] I. Aruk, I. Pacal, A.N. Toprak, A novel hybrid ConvNeXt-based approach for enhanced skin lesion classification, Expert Syst. Appl. 283 (2025) 127721.

[13] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.

[14] S. Yun, Y. Ro, Shvit: Single-head vision transformer with memory efficient macro design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5756–5767.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.

[16] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (1) (2018) 1–9.

[17] M.A. Kassem, K.M. Hosny, M.M. Fouad, Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning, IEEE Access 8 (2020) 114822–114832.

[18] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE J. Biomed. Health Inform. 23 (2) (2018) 538–546.

[19] H.C. Reis, V. Turk, K. Khoshelham, S. Kaya, InSiNet: a deep convolutional approach to skin cancer detection and segmentation, Med. Biol. Eng. Comput. 60 (3) (2022) 643–662.

[20] J.P. Villa-Pulgarin, A.A. Ruales-Torres, D. Arias-Garzon, M.A. Bravo-Ortiz, H.B. Arteaga-Arteaga, A. Mora-Rubio, J.A. Alzate-Grisales, E. Mercado-Ruiz, M. Hassaballah, S. Orozco-Arias, et al., Optimized convolutional neural network models for skin lesion classification, Comput. Mater. Contin. 70 (2) (2022).

[21] K. Liu, T. Huang, Z. Guo, Classification of pathological images of skin diseases based on deep learning, in: 2022 4th International Conference on Data-Driven Optimization of Complex Systems, DOCS, IEEE, 2022, pp. 1–6.

[22] P. Pranav, S. Sarath, Comparative study of skin lesion classification using dermoscopic images, in: 2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT, IEEE, 2023, pp. 1–5.

[23] B. Ozdemir, I. Pacal, An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms, Results Eng. 25 (2025) 103692.

[24] B. Ozdemir, I. Pacal, A robust deep learning framework for multiclass skin cancer classification, Sci. Rep. 15 (1) (2025) 4938.

[25] M.S. Ali, M.S. Miah, J. Haque, M.M. Rahman, M.K. Islam, An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models, Mach. Learn. Appl. 5 (2021) 100036.

[26] R. Garg, S. Maheshwari, A. Shukla, Decision support system for detection and classification of skin cancer using CNN, in: Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2020, Springer, 2020, pp. 578–586.

[27] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan, Tinyvit: Fast pretraining distillation for small vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 68–85.

[28] S.G. Jasil, V. Ulagamuthalvi, A hybrid CNN architecture for skin lesion classification using deep learning: SPG Jasil, V. Ulagamuthalvi, Soft Comput. (2023) 1–10.

[29] A. Adebiyi, N. Abdalnabi, E.H. Smith, J. Hirner, E.J. Simoes, M. Becevic, P. Rao, Accurate skin lesion classification using multimodal learning on the ham10000 dataset, 2024, MedRxiv, 2024-2005.

[30] A. Sasithradevi, S. Kanimozhi, P. Sasidhar, P.K. Pulipati, E. Sruthi, P. Prakash, EffiCAT: A synergistic approach to skin disease classification through multi-dataset fusion and attention mechanisms, Biomed. Signal Process. Control. 100 (2025) 107141.

[31] Y. Sun, D. Zhu, Y. Wang, Y. Fu, Z. Tian, GTC: GNN-transformer co-contrastive learning for self-supervised heterogeneous graph representation, Neural Netw. 181 (2025) 106645.

[32] S. Huang, Y. Song, J. Zhou, Z. Lin, Cluster-wise graph transformer with dual-granularity kernelized attention, Adv. Neural Inf. Process. Syst. 37 (2024) 33376–33401.

[33] K. Li, C. Chen, W. Cao, H. Wang, S. Han, R. Wang, Z. Ye, Z. Wu, W. Wang, L. Cai, et al., DeAF: a multimodal deep learning framework for disease prediction, Comput. Biol. Med. 156 (2023) 106715.

[34] L. Liu, J. Lyu, S. Liu, X. Tang, S.S. Chandra, F.A. Nasrallah, Triformer: A multi-modal transformer framework for mild cognitive impairment conversion prediction, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–4.

[35] H. Mao, Z. Chen, W. Tang, J. Zhao, Y. Ma, T. Zhao, N. Shah, M. Galkin, J. Tang, Position: Graph foundation models are already here, in: Forty-First International Conference on Machine Learning, 2024, pp. 1–23.

[36] L. Bi, D.D. Feng, M. Fulham, J. Kim, Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network, Pattern Recognit. 107 (2020) 107502.

[37] Y. Wang, Y. Feng, L. Zhang, J.T. Zhou, Y. Liu, R.S.M. Goh, L. Zhen, Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images, Med. Image Anal. 81 (2022) 102535.

[38] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, T. Lasser, FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification, Med. Image Anal. 76 (2022) 102307.

[39] D. Roy, S. Dutta, S. Bose, F. Schwenker, R. Sarkar, Background-invariant independence-guided multi-head attention network for skin lesion classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 34–44.

[40] W. Dong, X. Zhang, B. Chen, D. Yan, Z. Lin, Q. Yan, P. Wang, Y. Yang, Low-rank rescaled vision transformer fine-tuning: A residual design approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16101–16110.