Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085136

ARTICLE

BIAC-Net: Bidirectional Global-Local Communication for Feature Refinement in Medical Image Classification

Muhammad Naeem Zafar[1], Yunfei Yin[1,*], Junaid Abbas[2], Bayan Alabdullah[3], Khaled Alnowaiser[4], Yunyoung Nam[5] and Zepa Yang[5,*]

[1]College of Computer Science, Chongqing University, Shapingba, Chongqing, China
[2]School of Big Data and Software Engineering, Chongqing University, Shapingba, Chongqing, China
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[4]Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[5]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Republic of Korea

*Corresponding Authors: Yunfei Yin. Email: yinyunfei@cqu.edu.cn; Zepa Yang. Email: yangzepa@sch.ac.kr

Received: 06 May 2026; Accepted: 29 June 2026; Published: 13 August 2026

ABSTRACT: Accurate medical image classification increasingly relies on the joint modeling of global contextual semantics and fine-grained local structural cues, since many lesions are only reliably recognized when subtle local details are interpreted within their broader anatomical context. However, most recent hybrid CNN–Transformer and global–local frameworks still extract these features in separate streams and merge them only through late-stage static fusion, without explicit bidirectional interaction during representation learning. As a result, global context cannot effectively guide the refinement of subtle local structures, and local discriminative cues cannot recalibrate higher-level semantic reasoning before classification, which limits reciprocal feature refinement and undermines robust integration of contextual and structural information in challenging medical images. To address this limitation, we propose BIAC-Net, a bidirectional global–local feature communication network for medical image classification. BIAC-Net employs a Vision Transformer (ViT) backbone to capture contextual representations and introduces two complementary refinement branches that enhance global saliency and preserve local structural details. Unlike conventional hybrid designs that merge features directly, the proposed framework introduces a bidirectional cross-stream residual refinement mechanism between global and local representations, allowing contextual features to guide local refinement while local structural cues recalibrate global reasoning before fusion. The reciprocally refined representations are then integrated through an adaptive gated fusion mechanism that learns input-dependent feature weighting. Experimental results on the Kvasir and ISIC 2018 datasets demonstrate that BIAC-Net achieves 97.84% and 90.91% accuracy, respectively, outperforming several representative baseline models. Ablation studies further support the contribution of bidirectional communication, while Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations provide preliminary qualitative evidence that the model attends to diagnostically relevant image regions.

KEYWORDS: Medical image classification; vision transformer; global-local feature interaction; bidirectional attention; adaptive gated fusion; explainable AI

1 Introduction

Medical image classification is a core component of computer-aided diagnosis systems and supports clinical decision-making across a wide range of modalities, including magnetic resonance imaging (MRI),

computed tomography (CT), ultrasound, histopathology, and dermoscopy. Deep learning has substantially advanced medical image analysis by enabling models to learn task-relevant representations directly from data, thereby reducing dependence on handcrafted features and improving performance across classification and related tasks [1,2].

Among deep learning methods, convolutional neural networks (CNNs) have been widely used in medical imaging because of their strong ability to extract local spatial patterns and texture-sensitive features. However, convolutional processing is inherently local, which can make modeling long-range dependencies and global structural context more difficult, especially in complex medical images where diagnostically important evidence may be spatially distributed [3,4]. To address this limitation, transformer-based architectures have been increasingly adopted in medical imaging. Following the success of the Vision Transformer (ViT), which demonstrated the effectiveness of self-attention for image recognition, medical imaging research has extensively explored transformer-based models for classification, segmentation, detection, reconstruction, and related tasks [2,5,6]. Recent comprehensive reviews consistently report that transformers are particularly attractive in medical imaging because self-attention provides a principled mechanism for modeling long-range dependencies and global contextual relationships [5,7].

Because CNNs and transformers offer complementary strengths, hybrid architectures that combine convolutional feature extraction with transformer-based contextual modeling have become an active research direction. In such designs, convolutional modules are typically used to preserve fine-grained local structure, whereas transformer components model broader semantic context. Recent medical classification frameworks continue to explore this complementary design through staged attention, coarse-to-fine reasoning, and fusion-based CNN-transformer pipelines [8,9]. At the same time, explainability has become an important requirement in medical AI, since clinically useful models should not only achieve high accuracy but also provide interpretable evidence that can be inspected by domain experts. Recent surveys in medical image analysis emphasize that explainable AI methods, including saliency-based visualization and clinical evaluation protocols, are increasingly important for transparency, trust, and safe deployment in healthcare settings [10–12].

Despite recent advances in hybrid medical image classification, the integration of global contextual semantics and fine-grained local structural cues remains limited in many existing designs. Most conventional frameworks process global and local features independently and combine them only through late-stage static fusion, without allowing explicit interaction during representation learning [13,14]. Consequently, global context cannot guide local detail refinement, and local discriminative cues cannot recalibrate global reasoning before classification. Since clinically meaningful evidence often depends on both subtle local structures and broader contextual organization, this lack of reciprocal interaction restricts effective feature integration [8,15,16]. As a result, the potential for one representation to refine, recalibrate, or correct the other during feature learning may remain underexploited. This issue is particularly relevant in medical image classification, where diagnostically meaningful patterns often depend on both subtle local detail and broader contextual organization [17].

To address this limitation, we propose BIAC-Net, a bidirectional global-local feature communication framework for medical image classification, as illustrated in Fig. 1. BIAC-Net is designed to promote structured information exchange between global and local feature streams before final integration. Specifically, the framework employs a Vision Transformer backbone for contextual feature extraction, a Convolutional Block Attention Module (CBAM)-based saliency refinement branch, and a depthwise separable convolution-based local structural refinement branch. A bidirectional communication mechanism is then introduced to enable controlled reciprocal interaction between the two streams before fusion, allowing contextual representations to guide local refinement while local discriminative cues simultaneously recalibrate global

reasoning. The refined representations are subsequently combined through adaptive gated fusion, which dynamically balances global and local information according to the input. Through this design, BIAC-Net aims to achieve more cooperative feature refinement and more effective integration of contextual semantics and fine-grained structural cues for medical image classification. The main contributions of this paper are as follows:

1. We identify the limitation of conventional hybrid medical image classification architectures in which global contextual modeling and local structural refinement are typically processed independently and integrated only through late-stage fusion, restricting reciprocal refinement between complementary feature representations.
2. We propose BIAC-Net, a bidirectional global-local feature communication framework that enables reciprocal refinement between global contextual and local structural representations through a lightweight cross-stream residual communication mechanism prior to feature fusion.
3. We introduce an adaptive gated fusion strategy that integrates reciprocally refined representations using input-dependent weighting, allowing dynamic balancing between contextual semantics and discriminative structural cues.
4. We conduct extensive experiments and ablation studies on the Kvasir and ISIC 2018 datasets to evaluate the effectiveness of the proposed design, demonstrating consistent improvements over representative baseline models and providing interpretable predictions through Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 reports the experimental results. Sections 5 and 6 provide discussion and conclusion, respectively.

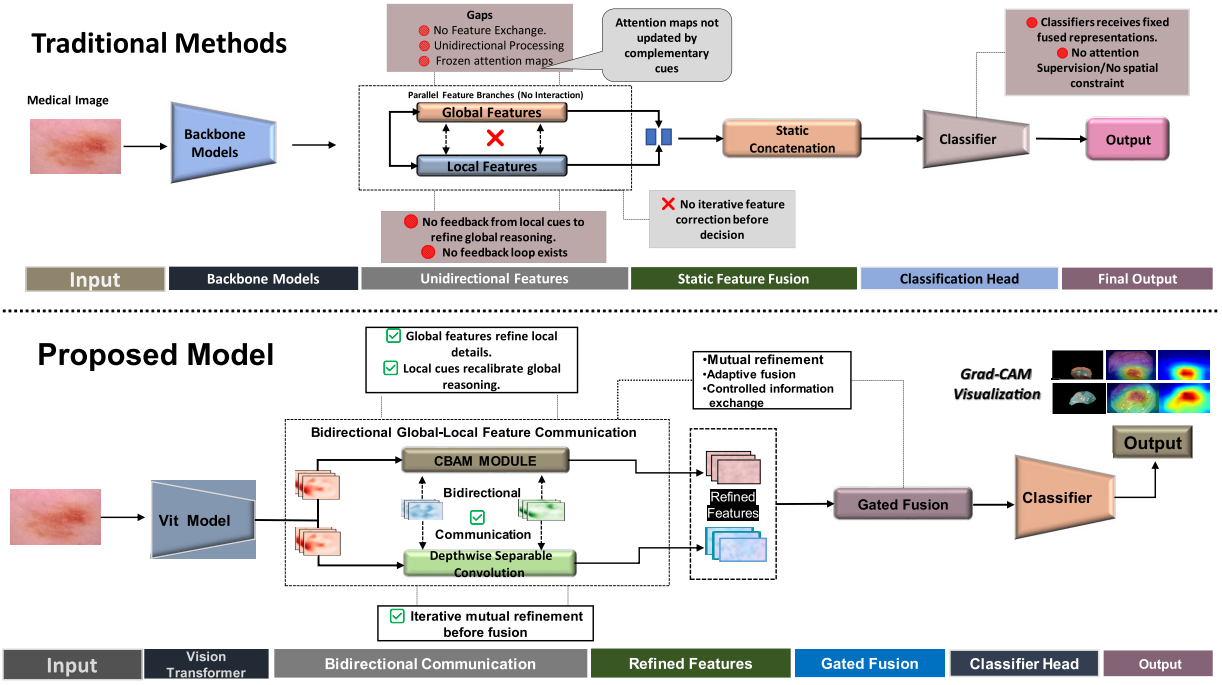

Figure 1: Comparison between traditional hybrid methods and the proposed BIAC-Net. Traditional methods process global and local features separately and rely on static fusion, whereas BIAC-Net enables bidirectional feature refinement and adaptive gated fusion before classification.

2 Related Work

2.1 Traditional CNN Methods

Recent advances in medical image classification have increasingly focused on CNN-based attention mechanisms to enhance feature discriminability by emphasizing diagnostically relevant regions while suppressing background noise [18,19]. Channel-wise and spatial attention strategies have been widely integrated into convolutional backbones to improve sensitivity to subtle visual patterns across diverse medical imaging modalities. More recent attention-augmented CNN frameworks have demonstrated improved robustness by dynamically reweighting feature responses during training [20]. However, these approaches primarily rely on localized receptive fields and hierarchical aggregation, limiting their ability to model long-range dependencies [21]. Furthermore, most CNN-based attention models employ static or loosely coupled fusion strategies without enabling structured interaction or mutual refinement between attention streams [22]. These limitations have motivated the exploration of transformer-based architectures and more advanced attention interaction mechanisms for medical image classification. In complex clinical images, diagnostically relevant cues may appear in spatially distant yet semantically related regions, which are difficult to capture effectively through purely convolutional processing. As a result, CNN-based models may struggle when accurate prediction requires simultaneous understanding of local lesion characteristics and broader anatomical context. Although attention-enhanced CNNs improve local discriminability, they still lack an explicit mechanism for global contextual reasoning across the entire image. This shortcoming becomes more evident in challenging classification tasks involving small lesions, heterogeneous textures, or visually ambiguous disease patterns. Therefore, more expressive frameworks that combine local sensitivity with long-range contextual modeling are needed to further improve medical image classification performance.

2.2 Vision Transformer and Global Self-Attention Modeling

ViTs have emerged as a strong alternative to CNNs [23–25] for medical image analysis by enabling global dependency modeling through self-attention mechanisms [26]. Unlike CNNs, which rely on localized receptive fields and hierarchical feature aggregation, ViTs represent images as sequences of embedded patches, allowing each spatial region to interact directly with all others within a single attention operation [27]. This ability to capture holistic contextual relationships has proven particularly valuable in medical imaging tasks, where diagnostic decisions often depend on long-range structural dependencies rather than isolated local features. Recent studies have demonstrated the effectiveness of transformer-based architectures across diverse medical image classification problems, including disease recognition and tumor classification [28,29]. Transformer models have also been integrated with explainability mechanisms to improve clinical interpretability, addressing a key limitation of traditional convolutional approaches [30,31]. In addition, hybrid transformer frameworks that incorporate attention fusion strategies have shown improved robustness and feature representation capacity in challenging medical imaging scenarios [32,33]. Adversarial learning schemes combined with ViTs have further enhanced model generalization and stability in medical image classification tasks [31]. These advances indicate that ViTs are not only powerful for capturing global semantic context but also highly adaptable to different clinical imaging conditions and diagnostic objectives. Their patch-based representation mechanism provides a flexible foundation for integrating complementary refinement modules in downstream architectures. Moreover, global self-attention enables the model to capture spatially distant yet semantically related regions, which is especially useful when pathological patterns are distributed across multiple areas of an image. This characteristic makes transformer-based representations particularly suitable for hybrid frameworks that aim to jointly exploit global context and local structural detail. Therefore, ViTs serve as an effective backbone for building more expressive and interpretable medical image classification models.

2.3 Local Feature Enhancement via Depthwise Separable Convolutions

Although global self-attention mechanisms effectively capture long-range contextual dependencies, fine-grained local details remain crucial for accurate medical image classification. Subtle texture variations, boundary structures, and localized intensity patterns often carry essential diagnostic information that may be diluted when relying solely on global representations. To address this limitation, separable convolution-based mechanisms have gained increasing attention because of their ability to enhance local feature discrimination while maintaining computational efficiency. By decomposing standard convolutions into depthwise and pointwise operations, these mechanisms enable independent spatial filtering and channel-wise feature interaction, resulting in more expressive yet lightweight local representations [34]. Recent studies have demonstrated that integrating depthwise separable convolutions with attention modules significantly improves the modeling of fine-scale structures in medical images [35], leading to enhanced sensitivity to lesion boundaries and local texture patterns [36]. In particular, depthwise separable convolution frameworks have shown strong performance gains in medical image analysis tasks by effectively suppressing background noise and amplifying diagnostically relevant regions [37].

2.4 Bidirectional Global-Local Feature Communication and Gated Fusion

Studies in medical image analysis have highlighted the importance of structured interaction between multiple feature representations to capture complementary contextual information [38–40]. Attention-guided interaction has been used to bridge semantic gaps across feature scales, where multi-scale dependency modeling enables features at different resolutions to guide each other during refinement, improving coherence and boundary awareness [15]. Dual-path attention architectures further demonstrate that parallel attention streams with controlled interaction preserve complementary information more effectively than independent processing [41]. Beyond single-modality settings, bidirectional information exchange has been extensively explored in multimodal learning [42], where deep fusion frameworks show that reciprocal guidance between feature streams yields more robust representations than unidirectional or static fusion strategies [43]. Recent multimodal medical frameworks reinforce this observation by allowing fine-grained cross-stream refinement through feature-level interaction mechanisms [44].

In contrast to existing approaches as described in Table 1, the proposed BIAC-Net framework explicitly models complementary global and local representations using a Vision Transformer backbone, a depthwise separable convolution branch, and a convolutional block attention module. A bidirectional global–local feature communication mechanism enables structured information exchange between attention pathways, while a gated fusion strategy adaptively balances their contributions. Furthermore, Gradient-weighted Class Activation Mapping (Grad-CAM) is used as a post-hoc visualization tool to qualitatively inspect the spatial regions contributing to model predictions. This design addresses the limitations of prior methods by jointly improving classification performance, robustness, and explainability in medical image analysis. By encouraging the network to align its predictions with visually relevant pathological regions, the proposed framework also promotes more reliable and visually coherent attention behavior. As a result, BIAC-Net provides a more interpretable and practically relevant solution for medical image classification tasks.

Table 1: Comparison of representative attention-based and hybrid medical image classification methods. CNN: Convolutional Neural Network; XAI: Explainable Artificial Intelligence; CBAM: Convolutional Block Attention Module; DW: Depthwise convolution.

Method	Attention Type	Fusion	XAI	Communication/ Feature Refinement
Transformer Survey [5]	Multi-head Self-Attention	Concatenation or Summation	No	Independent Local and Global Streams.
DualAttendMed [9]	Dual attention	Parallel branches	Yes	No cross-stream communication.
Hybrid Analysis [23]	CNN–Transformer hybrid	Sequential fusion	Yes	Feature interaction remains sequential.
TransAtten [32]	Global + local attention	Unidirectional fusion	No	Lacks reciprocal feature refinement.
DDSNet [35]	None	None	No	Only local feature modeling.
CE-RS-SBCIT [45]	Channel + spatial	Sequential fusion	Limited	Static interaction between attention modules.
Cross-Scale Attention [46]	Multi-scale spatial	Hierarchical fusion	No	No explicit global–local interaction.
Hybrid Learnable-Fusion [47]	CNN + Transformer	Static fusion	No	Global and local features processed independently.
Hybrid-Fusion [48]	Transformer attention	Multimodal fusion	Limited	No structured feature interaction.
HiFuse [49]	CNN + Transformer attention	Static fusion	No	Global and local features fused without reciprocal interaction.
Decoupled feature extraction [50]	Implicit attention	Correspondence learning	No	Not designed for classification interaction.
Ours	Global (CBAM) + Local (DW)	Adaptive Gated Fusion	Yes	Bidirectional global–local refinement.

Note: XAI is marked as "Yes" when the method explicitly includes an interpretability or visualization component, "Limited" when interpretability is indirect or only partially discussed, and "No" when no explicit interpretability analysis is reported.

2.5 Recent Mamba-Based and Reliability-Oriented Classification Methods

Recent studies have further advanced medical image classification through hybrid CNN–Transformer–Mamba architectures, visual state space models, and reliability-oriented transfer losses. MedCTM introduces a CNN–Transformer–Mamba hybrid network that combines convolutional local feature extraction, transformer-based global dependency modeling, and Mamba-based long-range representation learning for medical image classification [51]. MediTEDNet explores a visual state space model for thyroid eye disease classification, showing the potential of state-space representation learning for capturing discriminative medical image patterns [52]. Similarly, DSA Mamba investigates a Mamba-based architecture for advanced medical image classification and highlights the ability of state-space sequence modeling to support efficient feature extraction [53]. In addition to architectural advances, recent studies have also focused on reliability-oriented transfer learning objectives. TET Loss introduces a temperature–entropy calibrated transfer loss to improve reliability and reduce overconfident predictions in medical image classification [54]. MiT Loss further incorporates medical image-aware transfer calibration to enhance classification performance under transfer learning settings [55]. These methods are closely related to BIAC-Net because they address global contextual modeling, efficient long-range representation learning, and reliable medical image classification. However, most of them focus on backbone-level sequence modeling or calibration-aware optimization, whereas BIAC-Net focuses on explicit bidirectional communication between refined global and local feature streams before adaptive fusion. Therefore, the proposed method is complementary to recent Mamba-based architectures and calibration-aware loss functions.

2.6 Interpretability and Post-Hoc Visualization

Interpretability is a critical requirement for deploying deep learning models in clinical environments [56], where transparent and trustworthy decision-making is essential [57]. Gradient-based visualization techniques, particularly Grad-CAM, have been widely adopted to provide post-hoc explanations by highlighting class-discriminative regions in medical images [58,59]. Recent studies have used Gradient-weighted Class Activation Mapping (Grad-CAM) and related saliency methods as post-hoc tools to inspect class-discriminative regions in medical images [60,61]. In contrast to existing approaches as described in Table 1, the proposed BIAC-Net framework explicitly models complementary global and local representations using a Vision Transformer backbone, a depthwise separable convolution branch, and a convolutional block attention module. A novel bidirectional global–local feature communication mechanism enables structured information exchange between attention pathways, while a gated fusion strategy adaptively balances their contributions. Furthermore, Gradient-weighted Class Activation Mapping (Grad-CAM) is used as a post-hoc visualization tool to qualitatively inspect the spatial regions contributing to model predictions. This design addresses the limitations of prior methods by jointly improving classification performance, robustness, and explainability in medical image analysis.

3 Methodology

3.1 Method Overview

Medical image classification requires the joint modeling of global contextual information and fine-grained structural patterns. However, conventional architectures typically refine global and local representations independently and combine them only at the final stage. This design limits the ability of one representation to correct or guide the other during feature learning. To address this limitation, we propose BIAC-Net, which performs dual-stream refinement with explicit bidirectional communication between global and local feature representations. The proposed framework first employs a Vision Transformer backbone to extract spatially organized feature representations. These features are then processed by

two complementary branches: a global branch that enhances contextual saliency and a local branch that preserves fine-grained structural patterns. Unlike conventional hybrid architectures that fuse branch outputs directly, BIAC-Net introduces an iterative bidirectional cross-stream residual communication mechanism that enables the two streams to exchange projected feature information and refine each other before fusion. Finally, an adaptive gated fusion module integrates the refined global and local representations for classification. In addition, a gradient-based visualization procedure provides class-discriminative localization maps that highlight the image regions responsible for a predicted class. The overall architecture of the proposed method is illustrated in Fig. 2.

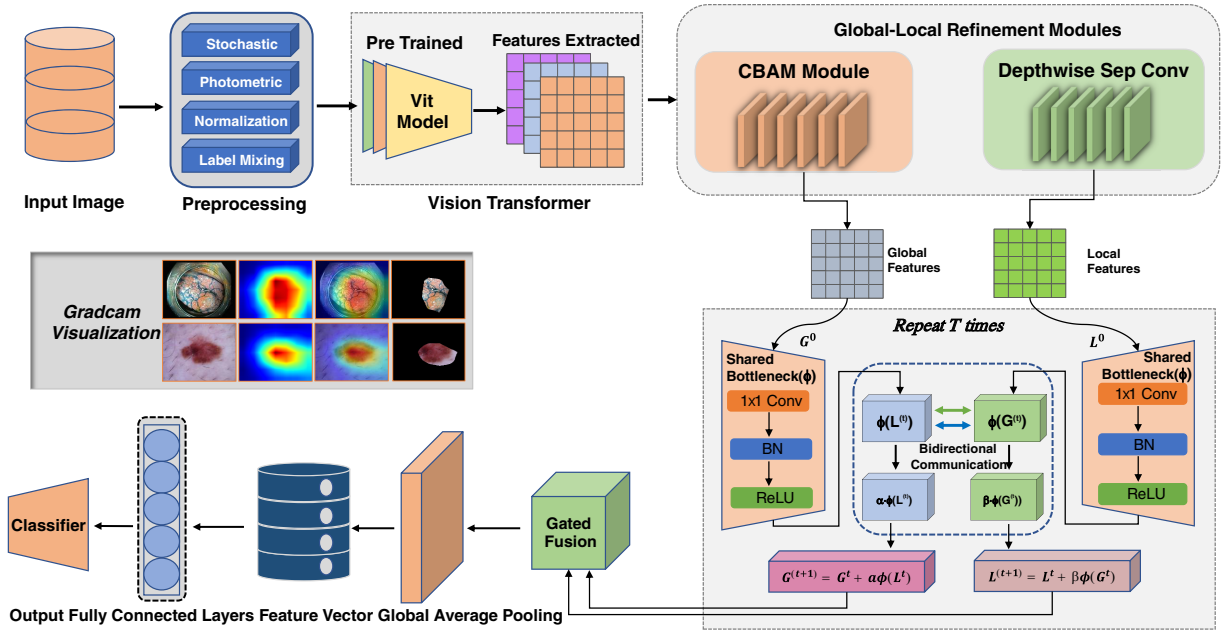

Figure 2: Overview of the proposed BIAC-Net architecture. A Vision Transformer (ViT) backbone extracts spatial feature representations, which are refined through a Convolutional Block Attention Module (CBAM)-based saliency refinement branch and a depthwise separable convolution-based local structural refinement branch. The two streams are coupled through bidirectional global-local feature communication and integrated using adaptive gated fusion.

3.2 Data Preprocessing and ViT Feature Extraction

A primary limitation in clinical datasets is that limited data and acquisition variability can cause overfitting and unstable generalization. Furthermore, class imbalance may bias learning toward majority classes and reduce minority-class recall. To improve robustness, we used separate preprocessing pipelines for training and evaluation.

Let an RGB image be denoted by

$$x \in \mathbb{R}^{H\times W\times 3}. \tag{1}$$

During training, each image is first resized and augmented using stochastic spatial and photometric transformations. The training augmentations include random resized cropping, horizontal flipping, vertical flipping, random rotation, color jitter, and random erasing. In our implementation, horizontal and vertical flips are applied with probability 0.5, random rotation is applied within 20°, and random erasing is applied

with probability 0.25. During evaluation, deterministic resizing and center cropping are used to ensure stable inference.

All images are normalized using ImageNet statistics, with mean $\mu = (0.485, 0.456, 0.406)$ and standard deviation $\sigma = (0.229, 0.224, 0.225)$. The normalized image is computed as

$$\tilde{x}_c = \frac{x_c - \mu_c}{\sigma_c}, \quad c \in \{1, 2, 3\}. \tag{2}$$

To reduce the effect of class imbalance, class-weighted cross-entropy is used when class weighting is enabled. Let n_k denote the number of samples in class k. The class weight is computed according to inverse class frequency, where the rebalancing exponent $\gamma > 0$ controls the strength of minority-class up-weighting. For a sample i with label $y_i \in \{1, \ldots, K\}$, logits $z_i \in \mathbb{R}^K$, and softmax probability

$$p_{ik} = \frac{\exp(z_{ik})}{\sum_{j=1}^{K} \exp(z_{ij})}, \tag{3}$$

the weighted cross-entropy loss is

$$\mathcal{L}_{\mathrm{wCE}}(i) = -\sum_{k=1}^{K} w_k \mathbf{1}[y_i = k] \log(p_{ik}), \tag{4}$$

where w_k is the class weight and $\mathbf{1}[\cdot]$ is the indicator function.

Label mixing is applied using mixup to reduce overconfident predictions and improve generalization. Two training samples (x_i, y_i) and (x_j, y_j) are combined using a mixing coefficient $\lambda \in [0, 1]$ sampled from a Beta distribution. In our implementation, mixup uses $\alpha = 0.4$ and is applied with probability 0.4. CutMix is not used in the reported configuration. The resulting soft label is denoted by $\tilde{y}_i$, where each element $\tilde{y}_{ik}$ represents the mixed target probability assigned to class k. The soft-label loss is

$$\mathcal{L}_{\mathrm{soft}}(i) = -\sum_{k=1}^{K} \tilde{y}_{ik} \log(p_{ik}). \tag{5}$$

For feature extraction, we use the Vision Transformer ViT-B/16 as shown in Fig. 3. The image is partitioned into non-overlapping patches of size $P \times P$ and each patch is projected to an embedding.

The patch embeddings are processed by Transformer encoder layers to model long-range dependencies. Let the embedding dimension be C and the number of patches be $N = hw$. ViT produces a token sequence (to support class token); To support downstream spatial refinement, we reshape the patch tokens, excluding the class token, into a 2D grid

$$F \in \mathbb{R}^{C \times h \times w}. \tag{6}$$

This produces a spatial feature map interface suitable for subsequent global-local refinement.

3.3 Global and Local Feature Processing

To capture complementary information, BIAC-Net processes the ViT feature tensor F using two refinement branches: a global branch and a local branch. The global branch is designed to strengthen high-level contextual saliency, while the local branch focuses on preserving fine-grained structural patterns. This dual-branch design is important because medical image classification often requires both broad anatomical context and subtle local evidence, such as lesion boundaries, texture variations, and small abnormal regions.

Therefore, instead of relying on a single feature stream, BIAC-Net separately refines global and local representations before allowing them to interact in the later communication stage.

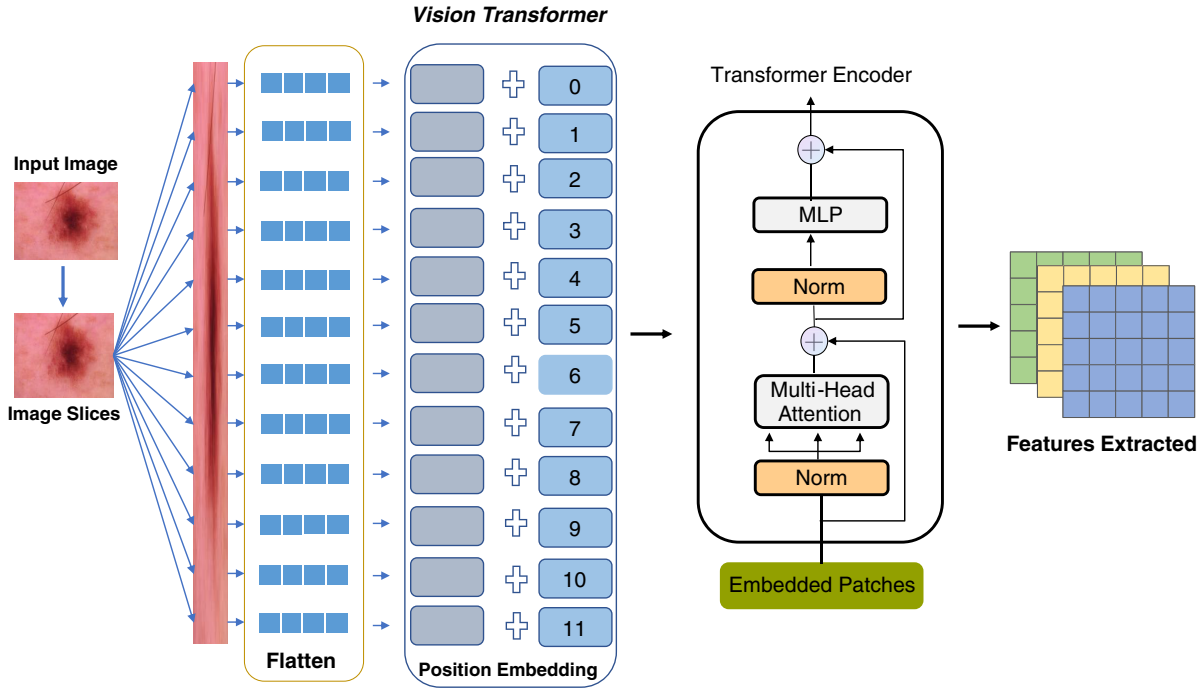

Figure 3: Feature extraction pipeline using a pretrained Vision Transformer (ViT), where image patches are embedded and processed by a Transformer encoder to capture global contextual features.

The global branch enhances contextual saliency through CBAM-based attention, as illustrated in Fig. 4. CBAM is applied because it sequentially models channel-wise and spatial importance, enabling the network to identify which feature channels and spatial regions are more relevant for classification. Channel attention first reweights feature channels according to their importance:

$$F_c = A_c(F) \odot F, \tag{7}$$

where $A_c(F) \in \mathbb{R}^{C\times1\times1}$ and $\odot$ denotes element-wise multiplication with broadcasting. This operation allows the model to emphasize informative semantic channels and suppress less discriminative responses. Spatial attention is then applied to emphasize informative locations:

$$G^{(0)} = A_s(F_c) \odot F_c, \tag{8}$$

where $A_s(F_c) \in \mathbb{R}^{1\times h\times w}$. Through this process, the global branch produces the refined representation $G^{(0)}$, which contains enhanced contextual information and spatially salient diagnostic cues.

In parallel, the local branch preserves fine-grained structural patterns such as lesion boundaries, local textures, and small discriminative regions through depthwise separable convolution, as illustrated in Fig. 5. Unlike standard convolution, depthwise separable convolution decomposes the operation into depthwise and pointwise convolutions. The depthwise convolution performs spatial filtering independently for each channel, allowing the model to capture local structural patterns efficiently. The pointwise convolution then integrates information across channels to generate a compact and discriminative local representation:

$$L^{(0)} = W_p * (W_d \circledast F), \tag{9}$$

where $\circledast$ denotes channel-wise convolution and $*$ denotes standard convolution. This design provides efficient local refinement while preserving discriminative structural cues. As a result, the local branch produces $L^{(0)}$, which complements the global representation by retaining detailed spatial information that may be weakened in transformer-based global modeling.

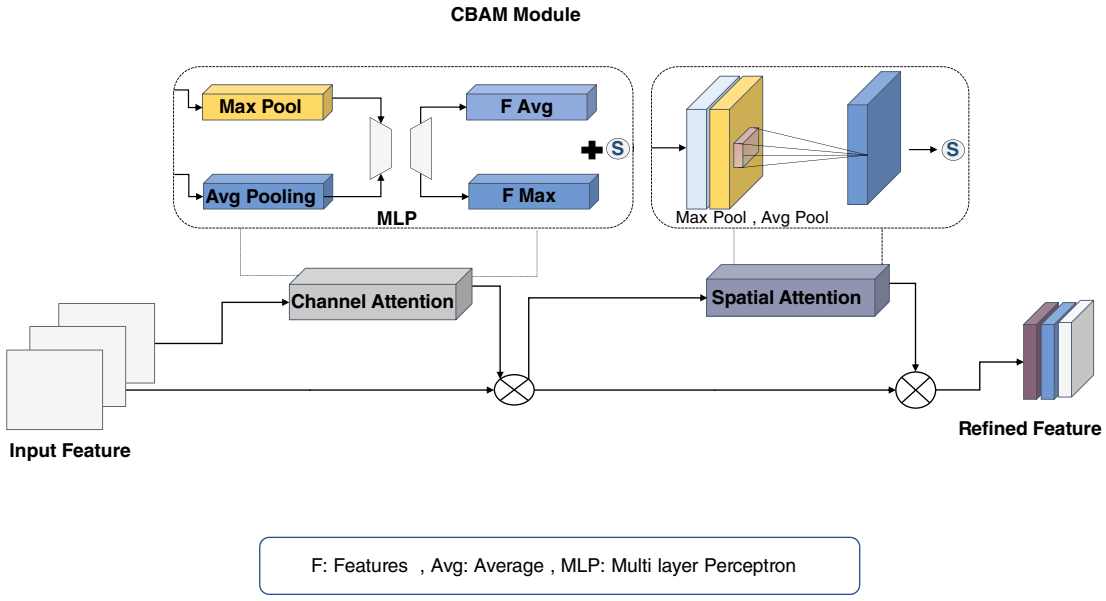

Figure 4: Convolutional Block Attention Module (CBAM)-based global refinement branch. The input feature tensor is sequentially processed by channel attention and spatial attention to enhance contextually salient and diagnostically relevant regions, producing the refined global representation $G^{(0)}$.

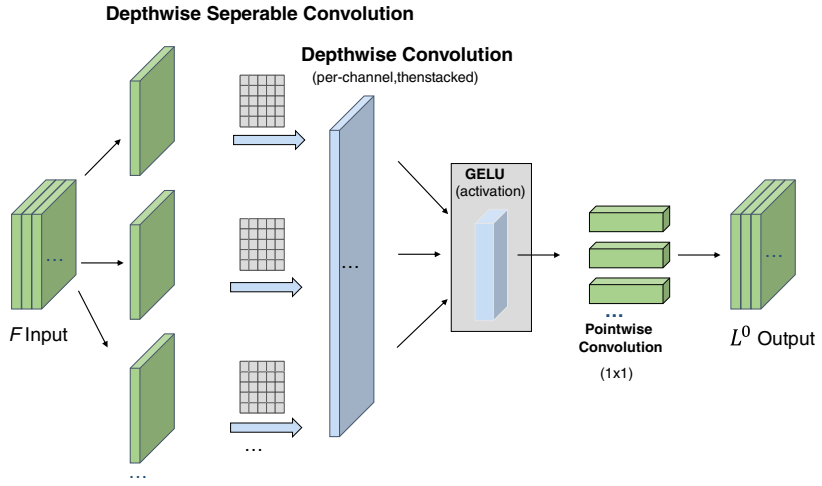

Figure 5: Depthwise separable convolution-based local refinement branch. The input feature tensor is processed by depthwise and pointwise convolutions to preserve fine-grained structural information, including lesion boundaries and texture patterns, producing the refined local representation $L^{(0)}$.

3.4 Bidirectional Cross-Stream Feature Communication

Although dual-stream processing improves representation diversity, independently refined streams may still produce inconsistent or redundant information. To enable cooperative refinement, we introduce an explicit bidirectional cross-stream residual communication mechanism between the global and local streams, as illustrated in Fig. 6. The spatial feature tensor $F \in \mathbb{R}^{C\times h\times w}$ extracted by the Vision Transformer backbone is processed by two refinement branches. The global branch applies CBAM to produce contextual saliency-refined features $G^{(0)}$, while the local branch applies depthwise separable convolution to obtain structural features $L^{(0)}$. These refined representations serve as the inputs to the bidirectional communication module, where reciprocal feature exchange is performed at each communication step. Let $\phi(\cdot)$ denote a shared bottleneck projection implemented using a 1×1 convolution with nonlinear activation, which aligns the feature representations before exchange. At communication step t, where $t = 0, 1, \ldots, T-1$ and $T = 4$ in our implementation, the global and local features are updated as:

$$G^{(t+1)} = G^{(t)} + \alpha\,\phi\left(L^{(t)}\right), \tag{10}$$

$$L^{(t+1)} = L^{(t)} + \beta\,\phi\left(G^{(t)}\right), \tag{11}$$

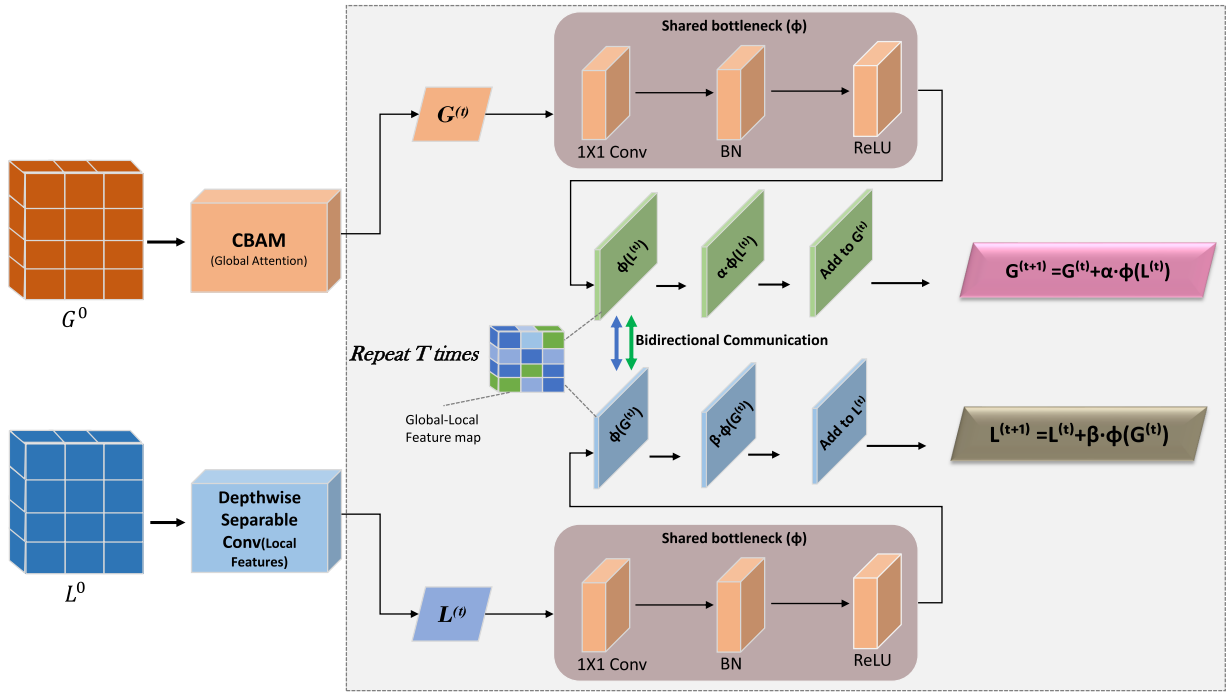

Figure 6: Bidirectional global–local feature communication module. Projected local features are injected into the global stream, while projected global features are injected into the local stream through lightweight residual updates using a shared bottleneck projection.

Unlike conventional hybrid architectures where global and local features are processed independently and fused only at the final stage, the proposed bidirectional update enables iterative reciprocal refinement between contextual and structural representations prior to feature fusion. Here, α and β are learnable scalar parameters that control the strength of cross-stream residual injection. In our implementation, $\phi(\cdot)$ is shared across both communication directions and reused across all communication iterations, while α and β are initialized to zero and optimized during training without additional clipping or positivity constraints. This

design allows the network to begin with independent branch refinement and gradually learn the degree of cross-stream feature exchange. We also compared the learnable coefficients with fixed coupling settings, including $\alpha = \beta = 0$ and $\alpha = \beta = 1$. The learnable coefficients achieved better performance and remained small after training, indicating a modest but adaptive exchange of global and local information. Therefore, the main results are reported using $T = 4$ communication iterations with learnable α and β.

3.5 Classification Head and Visualization

A limitation of simple concatenation or summation is that it assumes equal contribution of global and local evidence across images. In medical imaging, however, the relative importance of contextual and structural cues can vary significantly between cases. To address this issue, we employ an adaptive gated fusion mechanism that dynamically balances the contributions of the two streams. Let $\hat{G} = G^{(T)}$ and $\hat{L} = L^{(T)}$ denote the communicated global and local features. We compute a gate map

$$M = \sigma\left(W_g * [\hat{G}; \hat{L}]\right), \tag{12}$$

where $[\cdot;\cdot]$ denotes channel concatenation, W_g is a 1×1 convolution, and $\sigma(\cdot)$ is the sigmoid function. Given $\hat{G}, \hat{L} \in \mathbb{R}^{B\times C\times h\times w}$, the gate map has the shape $M \in \mathbb{R}^{B\times C\times h\times w}$, enabling channel-spatial-wise adaptive weighting between the two communicated feature streams. The fused representation is then

$$F_{\text{fuse}} = M \odot \hat{G} + (1 - M) \odot \hat{L}. \tag{13}$$

Compared with concatenation, which increases feature dimensionality and leaves feature selection to the classifier, and summation or averaging, which assign fixed contributions to both streams, the proposed gate learns input-dependent fusion weights. Unlike squeeze-and-excitation-style channel recalibration, the gate retains spatial resolution and performs channel-spatial-wise weighting. Therefore, the fusion module provides a lightweight adaptive mechanism for integrating the communicated global and local representations before classification. We apply global average pooling to obtain a vector representation $v \in \mathbb{R}^C$,

$$v_c = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} (F_{\text{fuse}})_{cij}, \tag{14}$$

and then compute logits via a linear classifier

$$z = W_{\text{cls}} v + b_{\text{cls}}, \quad z \in \mathbb{R}^K. \tag{15}$$

The predicted class probabilities are

$$p_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}. \tag{16}$$

Training uses $\mathcal{L}_{\text{wCE}}$ for hard labels or $\mathcal{L}_{\text{soft}}$ for label mixing augmentation, as defined in Section 3.2. For interpretability, we generate gradient-based localization maps from internal activations associated with the fused representation. Let A^k denote the kth channel activation map at a selected layer and let s^c be the score for class c. Following the Grad-CAM principle, we compute channel weights

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial s^c}{\partial A_{ij}^k}, \tag{17}$$

where Z is the number of spatial locations. The class-discriminative heatmap is then

$$H^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{18}$$

The heatmap is normalized and overlaid on the input image to highlight regions most responsible for the decision. In addition, we extract a region of interest visualization by applying a percentile based threshold to the heatmap distribution, which supports qualitative inspection of whether the model focuses on visually plausible disease-related structures.

4 Experiments and Results

This section evaluates the performance of the proposed BIAC-Net architecture for medical image classification using two publicly available datasets: Kvasir and ISIC 2018. The experiments aim to assess the classification performance of the proposed framework, analyze the contribution of its individual components through ablation studies, and examine the effects of architectural design choices and data augmentation strategies. In addition, qualitative visualization results are presented to investigate whether the model focuses on visually relevant pathological regions during prediction. The evaluation is designed to provide both quantitative and interpretability-based evidence for the effectiveness of the proposed bidirectional refinement strategy. By considering balanced and imbalanced datasets with different imaging characteristics, the experiments also assess the robustness and generalization capability of BIAC-Net across diverse medical classification settings.

4.1 Datasets

To evaluate the robustness and generalization capability of the proposed method, experiments are conducted on two publicly available medical imaging datasets with different data characteristics: the Kvasir dataset, which exhibits relatively balanced class distributions shown in Fig. 7, and the ISIC 2018 dataset, which presents significant class imbalance as illustrated in Fig. 8. These datasets represent distinct clinical imaging scenarios and therefore provide a comprehensive evaluation of the proposed model. A statistical summary of the datasets is presented in Table 2.

Kvasir [62]: The Kvasir dataset consists of gastrointestinal endoscopic images collected at Vestre Viken Health Trust in Norway. The dataset includes images of anatomical landmarks such as the Z-line, pylorus, and cecum, as well as pathological findings including esophagitis, polyps, and ulcerative colitis. It also contains images captured during clinical procedures such as polyp resection, including dyed and lifted polyps and resection margins. The image resolution varies from 720 × 576 to 1920 × 1072 pixels. In this study, the dataset was divided into training, validation, and testing subsets using a stratified image-level split with a ratio of 70%/10%/20%. Since Kvasir contains 500 images per class, each class was split into 350 training images, 50 validation images, and 100 testing images.

ISIC 2018 [63]: The ISIC 2018 dataset released by the International Skin Imaging Collaboration contains dermoscopic images for skin lesion classification. The dataset includes seven diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis/Bowen's disease, benign keratosis, dermatofibroma, and vascular lesions. All images have a resolution of 600 × 450 pixels.

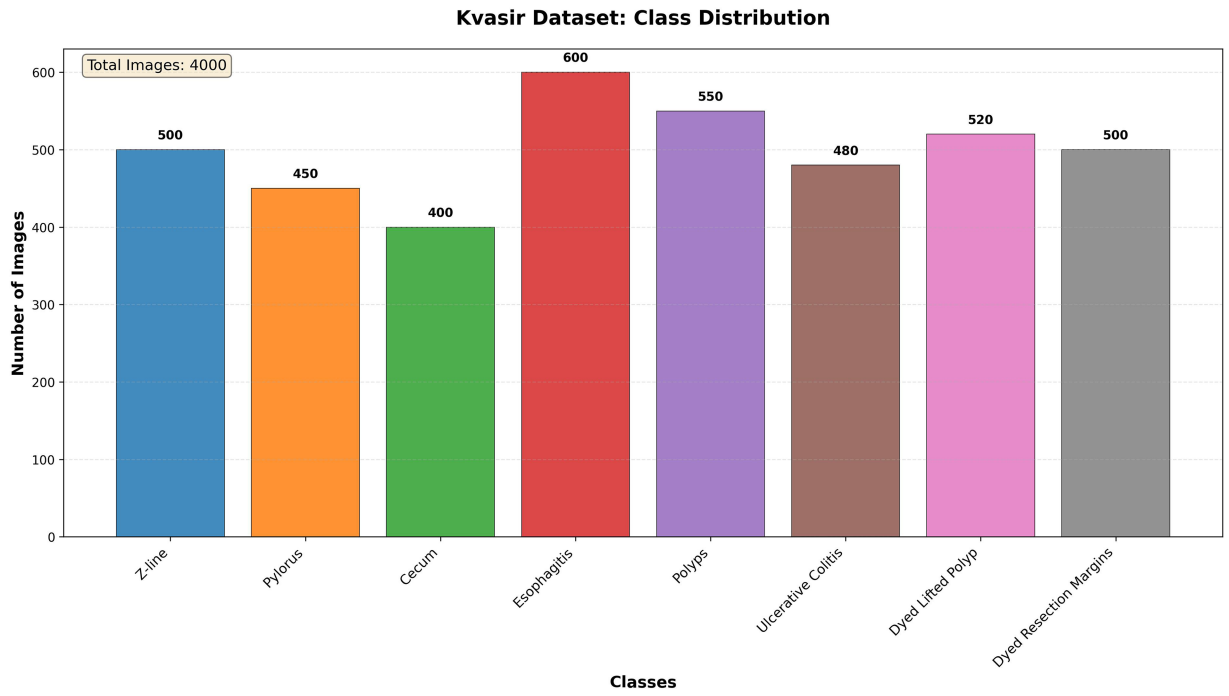

Figure 7: Details of each class contained in the Kvasir dataset.

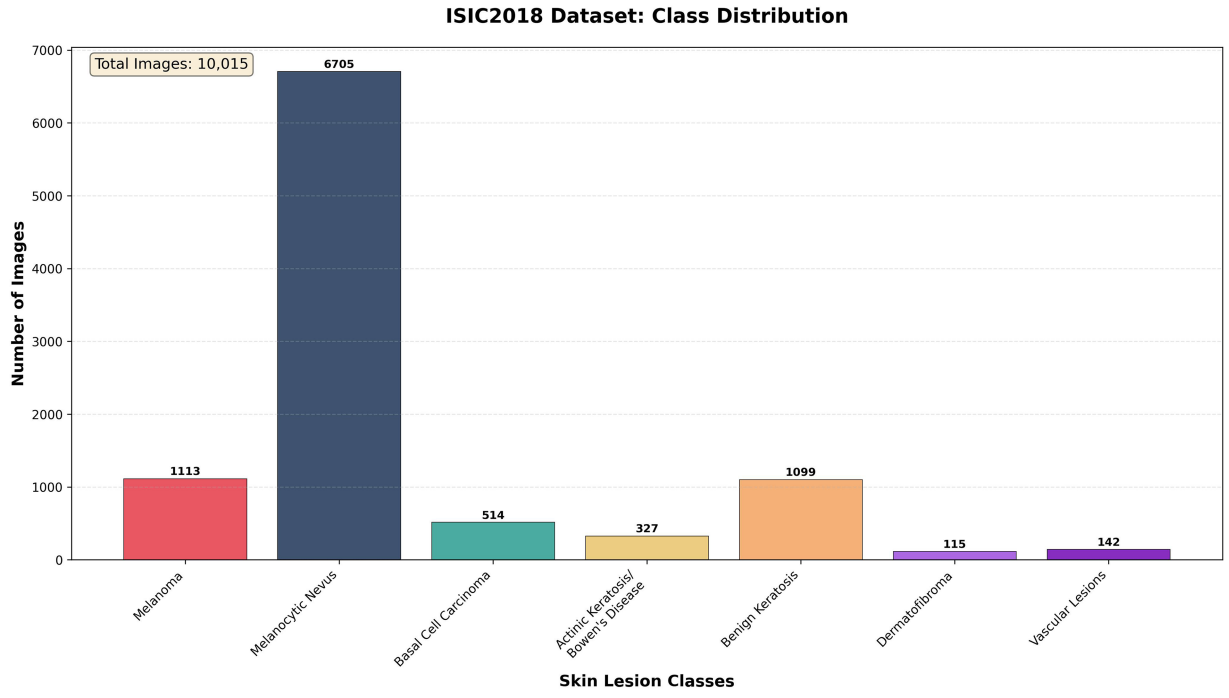

Figure 8: Details of each class contained in the ISIC 2018 dataset.

Table 2: The statistics of datasets.

Dataset	Availability	Data Volume	Categories	Resolution	Image Type
Kvasir	Public	4000	8	Mixed resolutions	Endoscopic GI tract images
ISIC 2018	Public	10,015	7	600 × 450	Dermoscopic images

For consistency with the Kvasir experiments, ISIC 2018 was also divided into training, validation, and testing subsets using a stratified image-level split with a ratio of 70%/10%/20%, preserving the class distribution as much as possible under the severe class imbalance.

For both datasets, the training set was used for model optimization, the validation set for checkpoint selection, and the test set for final evaluation. The same split protocol was applied to all compared methods.

4.2 Implementation Details and Evaluation Protocol

The proposed BIAC-Net framework was implemented in PyTorch. All input images were resized to 224 × 224 pixels to match the input configuration of the Vision Transformer backbone. A ViT-B/16 model initialized with ImageNet-1K pretrained weights was used for feature extraction and was fine-tuned end-to-end together with the newly introduced global-local refinement, bidirectional communication, gated fusion, and classification modules. The backbone parameters were not frozen during training. The feature representation was extracted from the output of the final Transformer encoder block. Since the proposed refinement modules operate on spatial feature maps, the class token was excluded from the downstream refinement process. Only the patch tokens were retained and reshaped into a two-dimensional feature tensor $F \in \mathbb{R}^{C\times h\times w}$ before being passed to the global and local refinement branches. Because the input images were resized to the standard ViT-compatible resolution, no position embedding interpolation was required. In the bidirectional global-local communication module, the number of communication iterations was set to $T = 4$. This setting was selected because it provided stable performance in our experiments while avoiding unnecessary computational growth. The projection function $\phi(\cdot)$ was implemented as a lightweight shared bottleneck consisting of 1×1 convolution and nonlinear activation. The same projection function was shared across both communication directions and reused across all communication iterations. The communication coefficients α and β were implemented as learnable scalar parameters and initialized to zero, allowing the network to begin with independent global and local refinement and gradually learn the strength of cross-stream residual communication during optimization. No additional clipping or positivity constraint was imposed on these parameters. The adaptive gated fusion module receives the communicated global and local feature maps $\hat{G}$ and $\hat{L}$ and generates a gate map from their channel-wise concatenation. Given $\hat{G}, \hat{L} \in \mathbb{R}^{B\times C\times h\times w}$, the resulting gate map has the shape $M \in \mathbb{R}^{B\times C\times h\times w}$, enabling channel-spatial-wise adaptive weighting between the two streams. The fused representation is then passed through global average pooling, dropout, and a fully connected classification layer. The model was trained with a batch size of 32. Training was performed for 25 epochs on the ISIC 2018 dataset and 20 epochs on the Kvasir dataset. Data augmentation strategies described in Section 3.2 were applied during training to improve generalization. The AdamW optimizer was used with an initial learning rate of 0.001 and weight decay for regularization. The learning rate was decayed exponentially by a factor of 0.9 every five epochs to support stable convergence. All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU. The datasets used in this study are publicly available and were used in accordance with their standard ethical guidelines. Model performance was evaluated using Accuracy, Precision, Recall, and F1-score. Let TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. The evaluation metrics are defined as

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

$$\text{Precision} = \frac{TP}{TP + FP} \tag{20}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

These metrics provide a comprehensive evaluation of classification performance, particularly under class imbalance conditions commonly observed in medical imaging datasets.

4.3 Performance Comparison with State-of-the-Art

We ran the official implementations of all baseline methods using the same data partitions and evaluation protocol, including DenseNet-162 [64], ResNet-152 [65], Swin Transformer [66], LS+ [67], HiFuse [49], PSA-MIL [68], FPT [69], and Proto-Non-Param [70]. For fairness, the original architectural settings and pretrained initialization of each baseline were retained, while only dataset-dependent settings, such as the number of output classes, were adjusted. All methods followed the same input resolution, epoch budget, learning-rate schedule, and augmentation protocol unless a method-specific configuration was required by the official implementation. As shown in Tables 3 and 4, BIAC-Net consistently outperforms all baseline methods across Precision, Recall, F1-score, and Accuracy on both datasets. On the Kvasir dataset, BIAC-Net achieves Precision of 97.71%, Recall of 97.58%, F1-score of 97.64%, and Accuracy of 97.84%. Compared to the strongest competing method, Proto-Non-Param with Accuracy of 93.01%, our method improves Accuracy by +4.83%, Precision by +4.71%, Recall by +4.62%, and F1-score by +4.64%. Similarly, on the ISIC 2018 dataset, BIAC-Net achieves Precision of 90.89%, Recall of 90.87%, F1-score of 90.90%, and Accuracy of 90.91%. Compared to Proto-Non-Param with Accuracy of 87.76%, our method improves Accuracy by +3.14%, Precision by +3.19%, Recall by +3.76%, and F1-score by +3.14%. The results show that BIAC-Net maintains consistent performance across Accuracy, Precision, Recall, and F1-score on both datasets. The consistent performance across all evaluation metrics supports the effectiveness of the proposed design. The most significant performance gains arise after introducing the bidirectional inter-attention communication mechanism, where reciprocal global–local refinement substantially enhances feature alignment compared to independent attention modeling. Adaptive gated fusion further stabilizes representation weighting by dynamically balancing complementary cues prior to classification. Together, these components drive the model to achieve superior final accuracies 97.84% on Kvasir and 90.91% on ISIC 2018, supporting the central role of structured feature exchange in the proposed framework.

4.4 Ablation Studies

To analyze the contribution of each component in BIAC-Net, we conduct ablation experiments on the Kvasir and ISIC 2018 datasets. The study evaluates the effects of preprocessing, dual-branch attention refinement (CBAM + depthwise convolution), bidirectional communication, and adaptive gated fusion. The results are summarized in Table 5.

Table 3: Performance comparison of the proposed BIAC-Net against existing state-of-the-art approaches on the ISIC 2018 dataset. All values are reported in percentage (%). "BIAC-Net (Ours)" denotes the result obtained on the standard train/validation/test split, the 5-fold result is reported as mean ± standard deviation.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
DenseNet-162	64.41 ± 1.3	65.16 ± 1.2	64.49 ± 1.2	64.40 ± 1.3
ResNet-152	66.68 ± 1.2	66.61 ± 1.1	66.59 ± 1.1	66.67 ± 1.2
Swin Transformer	71.24 ± 1.1	71.18 ± 1.1	71.12 ± 1.2	71.22 ± 1.1
LS+	77.01 ± 0.8	76.92 ± 0.9	76.90 ± 0.9	77.00 ± 0.8
HiFuse	81.12 ± 0.6	81.11 ± 0.6	81.00 ± 0.7	81.11 ± 0.6
PSA-MIL	84.26 ± 0.6	84.12 ± 0.5	84.17 ± 0.6	84.23 ± 0.6
FPT	86.95 ± 0.4	86.84 ± 0.5	86.88 ± 0.4	86.91 ± 0.5
Proto-Non-Param	87.76 ± 0.3	87.70 ± 0.3	87.11 ± 0.4	87.76 ± 0.4
BIAC-Net (Ours)	90.91 ± 0.2	90.89 ± 0.2	90.87 ± 0.2	90.90 ± 0.3

Table 4: Performance comparison of the proposed BIAC-Net against existing state-of-the-art approaches on the Kvasir gastrointestinal image dataset. All values are reported in percentage (%). "BIAC-Net (Ours)" denotes the result obtained on the standard train/validation/test split, the 5-fold result is reported as mean ± standard deviation.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
DenseNet-162	78.85 ± 0.8	77.86 ± 0.8	77.83 ± 0.7	78.85 ± 0.7
ResNet-152	82.61 ± 0.7	82.59 ± 0.8	82.67 ± 0.7	82.75 ± 0.7
Swin Transformer	86.21 ± 0.7	86.27 ± 0.6	86.45 ± 0.7	86.37 ± 0.8
LS+	89.11 ± 0.4	89.02 ± 0.5	89.05 ± 0.5	89.09 ± 0.4
HiFuse	89.16 ± 0.3	89.24 ± 0.5	89.18 ± 0.3	89.17 ± 0.3
PSA-MIL	91.88 ± 0.3	91.67 ± 0.4	91.73 ± 0.4	91.87 ± 0.4
FPT	92.85 ± 0.3	92.78 ± 0.2	92.81 ± 0.3	92.85 ± 0.2
Proto-Non-Param	93.01 ± 0.2	93.00 ± 0.2	92.96 ± 0.3	93.00 ± 0.3
BIAC-Net (Ours)	97.84 ± 0.1	97.71 ± 0.1	97.58 ± 0.1	97.64 ± 0.2

Table 5: Ablation study of BIAC-Net on the Kvasir and ISIC 2018 datasets. Prep. denotes preprocessing techniques; CBAM denotes Convolutional Block Attention Module; DW denotes depthwise separable convolution; Bi-Com. denotes bidirectional communication; Acc., Prec., Rec., and F1 denote accuracy, precision, recall, and F1-score, respectively.

Backbone	Prep.	CBAM+DW	Bi-Com.	Gated Fusion	Kvasir				ISIC 2018			
					Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
✓	–	–	–	–	76.30%	76.25%	76.24%	76.25%	67.12%	67.10%	67.09%	67.11%
✓	✓	–	–	–	80.17%	80.06%	80.07%	80.08%	71.23%	71.20%	71.16%	71.19%
✓	✓	✓	–	–	86.92%	86.81%	86.87%	86.88%	78.86%	78.80%	78.83%	78.84%
✓	✓	✓	✓	–	92.46%	92.43%	92.42%	92.41%	85.69%	85.66%	85.67%	85.68%
✓	✓	✓	–	✓	89.23%	89.16%	89.18%	89.18%	83.20%	83.19%	83.18%	83.19%
✓	–	✓	✓	✓	93.13%	93.06%	93.05%	93.08%	88.16%	88.12%	88.14%	88.15%
✓	✓	✓	✓	✓	97.84%	97.71%	97.58%	97.64%	90.91%	90.89%	90.87%	90.90%

Backbone Only. Using only the ViT backbone achieves accuracies of 76.30% on Kvasir and 67.12% on ISIC 2018. Without explicit global–local refinement, the backbone relies solely on general feature extraction and lacks mechanisms to simultaneously capture contextual semantics and fine-grained structural details.

Effect of Preprocessing. Introducing stochastic preprocessing improves accuracy to 80.17% on Kvasir and 71.23% on ISIC 2018. Compared to the backbone baseline, this corresponds to improvements of +3.87% on Kvasir and +4.11% on ISIC 2018. This gain indicates that augmentation and dataset-specific normalization reduce overfitting and improve robustness to illumination, scale, and appearance variations commonly observed in medical images.

Effect of Attention Mechanisms (CBAM + DW). Adding the CBAM-based global branch and the depthwise separable convolution (DW) local branch further increases accuracy to 86.92% on Kvasir and 78.86% on ISIC 2018. This represents additional improvements of +6.75% and +7.63%, respectively. The global branch enhances contextual saliency through channel and spatial attention, while the local branch preserves fine-grained structural patterns such as lesion boundaries and texture variations. Their complementary representations significantly improve discriminative capability.

Effect of Bidirectional Communication. Introducing the proposed bidirectional global–local feature communication improves accuracy to 92.46% on Kvasir and 85.69% on ISIC 2018, corresponding to gains of +5.54% and +6.83% over the previous configuration. This improvement verifies that reciprocal interaction between global and local features is critical. Through projected feature exchange, contextual reasoning guides local feature refinement while local discriminative cues recalibrate global attention.

Effect of Gated Fusion. Replacing static fusion with adaptive gated fusion improves performance to 89.23% on Kvasir and 83.20% on ISIC 2018 when used without communication. When combined with all components, the full BIAC-Net achieves the highest accuracy of 97.84% on Kvasir and 90.91% on ISIC 2018, corresponding to additional gains of +5.38% and +5.22% over the communication stage.

Overall, the ablation results indicate that each component contributes positively to the final performance. The most significant improvements arise from the introduction of the dual-branch attention refinement and the bidirectional communication mechanism, which enable effective interaction between contextual and structural representations. The integration of adaptive gated fusion further strengthens feature aggregation, resulting in the best overall performance. The contribution of each component is further illustrated in Fig. 9.

4.5 Qualitative Visualization Analysis

To provide a qualitative visual inspection of the regions contributing to BIAC-Net predictions, we generated Gradient-weighted Class Activation Mapping (Grad-CAM) localization maps from the fused feature representation described in Section 3.5. The heatmaps were normalized and thresholded to produce region-of-interest (ROI) visualizations. In our implementation, the ROI mask was generated by retaining activation values above the 85th percentile was selected as an empirical threshold to retain the most salient activation regions while suppressing diffuse low-response background areas in the visualization. As shown in Fig. 10, the Kvasir examples indicate that the activation responses are generally concentrated around visually relevant pathological regions, such as abnormal tissue and lesion-like areas. Similarly, Fig. 11 shows that the ISIC 2018 heatmaps tend to emphasize lesion regions and surrounding discriminative visual patterns. These results provide qualitative evidence that the proposed model often relies on image regions that are visually consistent with disease-related structures. Overall, the Grad-CAM visualizations provide a qualitative view of the decision regions learned by BIAC-Net. The observed activation patterns suggest

that the fused representation tends to emphasize visually relevant pathological structures while reducing responses from surrounding background areas. These results complement the quantitative classification performance by offering an interpretable visual perspective on the model predictions.

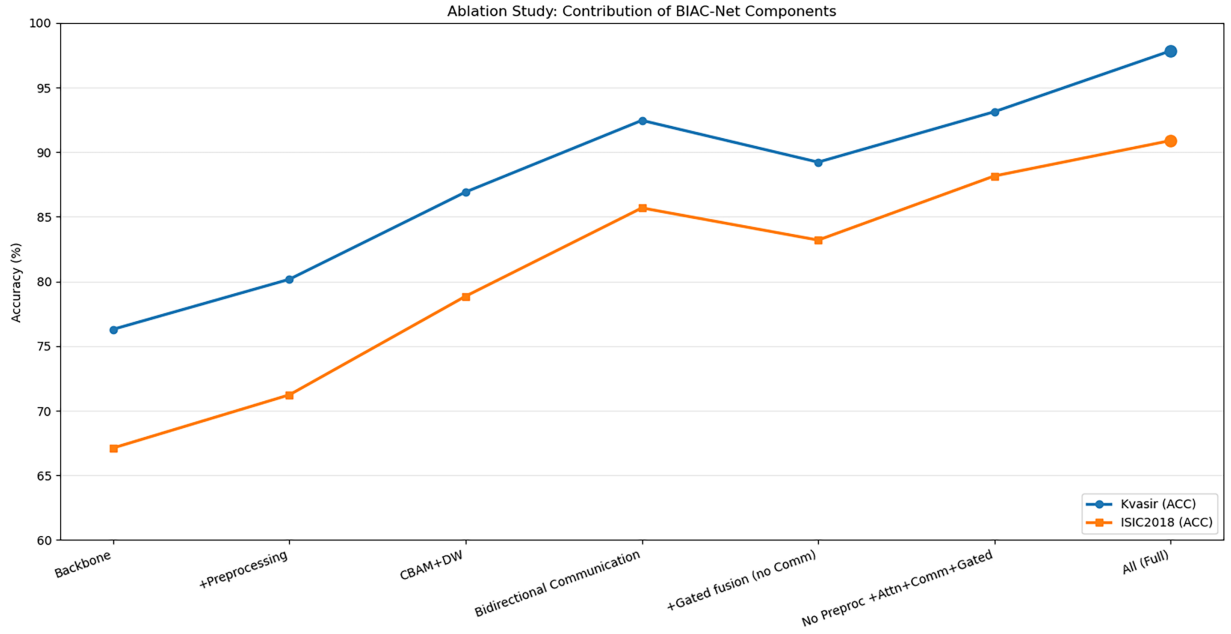

Figure 9: Ablation study results showing the contribution of each component in BIAC-Net on the Kvasir and ISIC 2018 datasets.

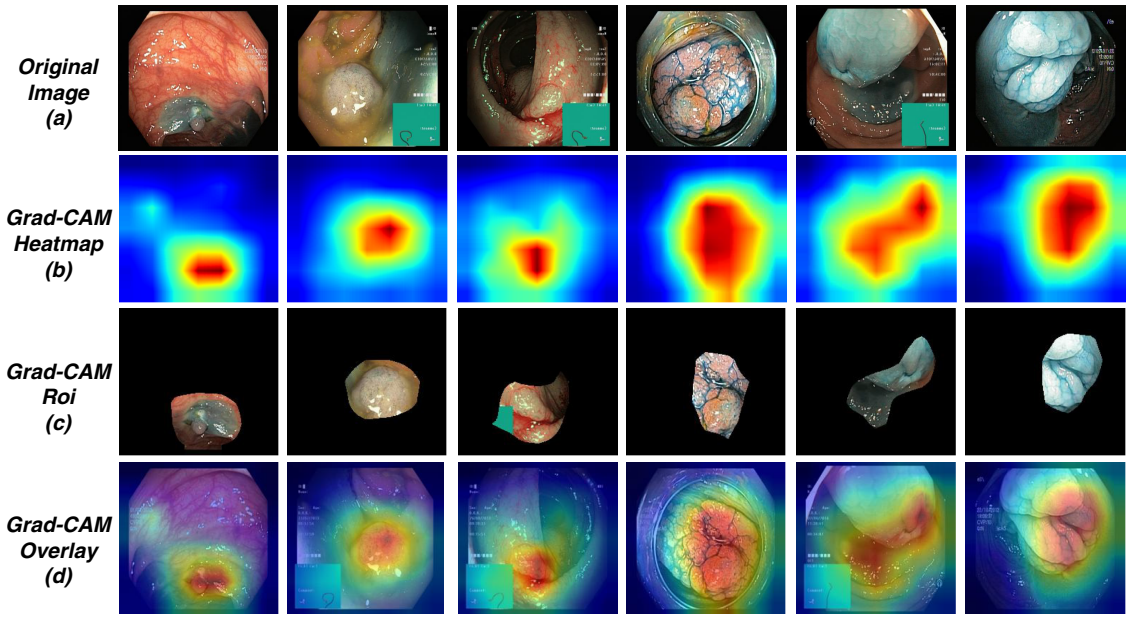

Figure 10: Gradient-weighted Class Activation Mapping (Grad-CAM) on the Kvasir dataset. Rows represent (**a**) input image, (**b**) Grad-CAM heatmap, (**c**) ROI mask, and (**d**) heatmap overlay highlighting pathological regions influencing the prediction.

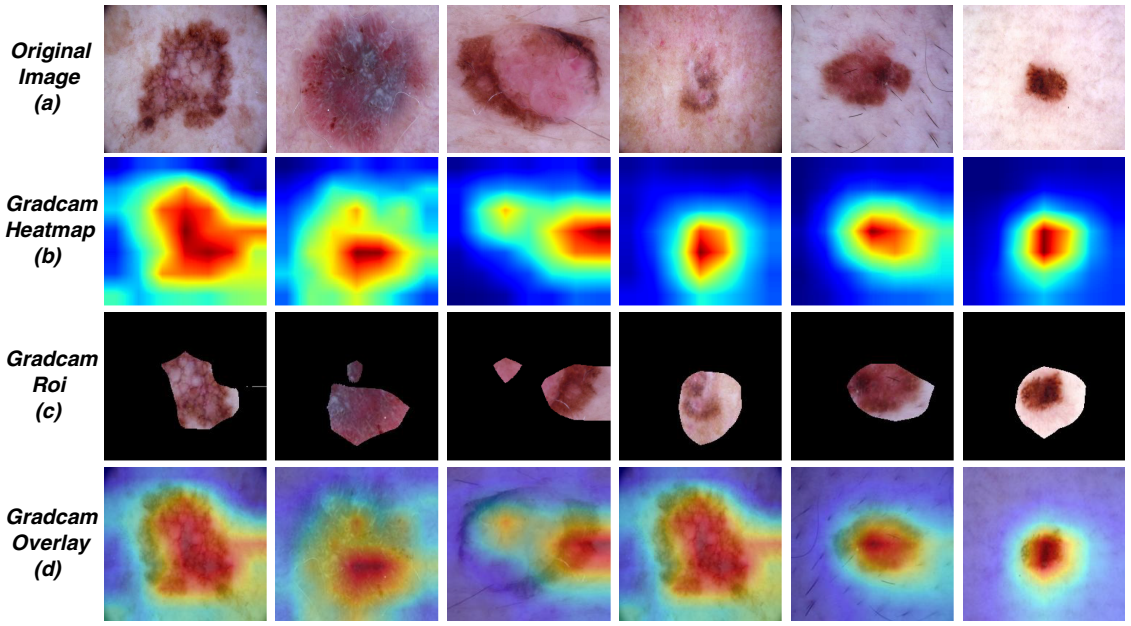

Figure 11: Gradient-weighted Class Activation Mapping (Grad-CAM) on the ISIC 2018 dataset. Rows show (**a**) input image, (**b**) Grad-CAM heatmap, (**c**) extracted ROI mask, and (**d**) heatmap overlay highlighting lesion regions influencing the prediction.

5 Discussion

The experimental results show that BIAC-Net improves classification performance on both Kvasir and ISIC 2018, but the improvement is larger on Kvasir. This difference is mainly related to dataset characteristics. Kvasir is relatively balanced and contains endoscopic classes with more distinguishable anatomical and pathological patterns. Therefore, the proposed bidirectional global–local communication can more effectively combine broad contextual appearance with local structural cues. In contrast, ISIC 2018 is highly imbalanced and contains visually similar lesion categories with substantial intra-class variation, making the improvement more moderate. The proposed design is particularly useful for cases where prediction depends on both global morphology and fine local details. In Kvasir, abnormal mucosal patterns, polyp-like structures, and inflammatory regions benefit from combining contextual scene information with local texture and boundary cues. In ISIC 2018, lesion categories with irregular borders, pigmentation variation, and subtle texture patterns can also benefit from reciprocal global–local refinement. The bidirectional communication mechanism helps reduce errors caused by independent feature streams, where global features may overlook small discriminative structures and local features may respond to visually similar but contextually irrelevant regions. The ablation study indicates that the improvements come from both training and architecture, but the largest architectural gain is obtained after introducing bidirectional communication. Data augmentation improves robustness to image appearance variation, while the CBAM-based saliency branch and depthwise separable convolution branch provide complementary feature refinement. Adaptive gated fusion further improves feature integration by assigning input-dependent weights to the communicated global and local representations. BIAC-Net is built on a ViT-B/16 backbone with about 86.4 million parameters, and the added modules increase the total count only slightly to about 87.2 million. Thus, the performance gains are achieved with limited additional complexity. Several limitations remain. For ISIC 2018, severe class imbalance, visually similar lesion types, low-contrast regions, and ambiguous boundaries may still lead to misclassification. For Kvasir, errors may occur when pathological areas are small, partially occluded, or affected by illumination

artifacts. In addition, the Grad-CAM results provide only qualitative visual evidence and require further quantitative localization evaluation or expert validation. A more detailed analysis of runtime, memory usage, and FLOPs will be explored in future work. Future work will also include additional imbalance-sensitive metrics for highly imbalanced datasets such as ISIC 2018, along with larger multi-center validation, patient-level or lesion-level evaluation when metadata are available, expert-reader comparison, and optimization for computational efficiency.

6 Conclusion

This work presented BIAC-Net, a bidirectional global–local feature communication framework for medical image classification. The proposed method combines a Vision Transformer backbone, CBAM-based saliency refinement, depthwise separable convolution-based local refinement, bidirectional cross-stream communication, and adaptive gated fusion to improve the interaction between contextual and structural features before classification. By enabling reciprocal refinement before final fusion, BIAC-Net reduces the limitation of independently processed global and local streams and provides a more cooperative feature learning strategy. Experiments on Kvasir and ISIC 2018 show that BIAC-Net achieves competitive performance, reaching 97.84% accuracy on Kvasir and 90.91% accuracy on ISIC 2018. The 5-fold evaluation results further indicate stable performance across different data partitions, while ablation studies support the contribution of bidirectional communication and adaptive fusion. Grad-CAM visualizations provide a preliminary qualitative view of the regions contributing to model predictions. These results suggest that explicit global–local communication can improve representation learning across both balanced and imbalanced medical image datasets. Overall, structured global–local feature communication appears to be a promising design direction for medical image classification. Future work will include external validation, patient-level or lesion-level evaluation, additional imbalance-sensitive metrics, quantitative localization analysis, expert-reader comparison, and computational efficiency optimization.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant No. RS-2023-00218176, the Soonchunhyang University Research Fund. Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: Muhammad Naeem Zafar led the overall research work, including conceptualization, methodology, model design, implementation, experiments, result analysis, figure preparation, and manuscript writing. Yunfei Yin supervised the research, provided methodological guidance, and contributed to manuscript revision. Junaid Abbas supported the experiments and implementation. Bayan Alabdullah contributed to validation and manuscript review. Khaled Alnowaiser contributed to experiments, result analysis, and writing, review, and editing. Yunyoung Nam contributed to validation, visualization, and manuscript improvement. Zepa Yang contributed to supervision, review, funding support, and correspondence. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available. The Kvasir dataset is available from its official public repository, and the ISIC 2018 dataset is publicly available through the International Skin Imaging Collaboration (ISIC) archive. The implementation code of BIAC-Net is publicly available at the authors' GitHub repository: https://github.com/Naeem-cqu/BIAC-Net.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. doi:10.1016/j.media.2017.07.005.
2. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digit Med. 2023;6(1):74. doi:10.1038/s41746-023-00811-0.
3. Han Q, Qian X, Xu H, Wu K, Meng L, Qiu Z, et al. DM-CNN: dynamic multi-scale convolutional neural Network with uncertainty quantification for medical image classification. Comput Biol Med. 2024;168:107758. doi:10.1016/j.compbiomed.2023.107758.
4. Papanastasiou G, Dikaios N, Huang J, Wang C, Yang G. Is attention all you need in medical image analysis? A review. IEEE J Biomed Health Inform. 2023;28(3):1398–411. doi:10.1109/JBHI.2023.3348436.
5. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. Med Image Anal. 2023;88(1):102802. doi:10.1016/j.media.2023.102802.
6. Manzari ON, Ahmadabadi H, Kashiani H, Shokouhi SB, Ayatollahi A. MedViT: a robust vision transformer for generalized medical image classification. Comput Biol Med. 2023;157:106791. doi:10.1016/j.compbiomed.2023.106791.
7. Sharma AK, Verma NK. A novel vision transformer with residual in self-attention for biomedical image classification. arXiv:2306.01594. 2023.
8. Kim JW, Khan AU, Banerjee I. Systematic review of hybrid vision transformer architectures for radiological image analysis. J Imaging Inform Med. 2025;38(6):3248–62. doi:10.1101/2024.06.21.24309265.
9. Abbas J, Soomro DB, Huang S, Liu L. DualAttendMed: a coarse-to-fine dual-stage attention framework for interpretable disease localization and classification. Expert Syst Appl. 2025;305:130886. doi:10.1016/j.eswa.2025.130886.
10. Prentzas N, Kakas A, Pattichis CS. Explainable AI applications in the medical domain: A systematic review. arXiv:2308.05411. 2023.
11. Bhati D, Neha F, Amiruzzaman M. A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. J Imaging. 2024;10(10):239. doi:10.20944/preprints202408.0765.v1.
12. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. Med Image Anal. 2023;84:102684. doi:10.1016/j.media.2022.102684.
13. Muhammad A, Jin Q, Elwasila O, Gulzar Y. Hybrid deep learning architecture with adaptive feature fusion for multi-stage Alzheimer's disease classification. Brain Sci. 2025;15(6):612. doi:10.3390/brainsci15060612.
14. Chowdary GJ, Yin Z. Med-former: a transformer based architecture for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin/Heidelberg, Germany: Springer; 2024. p. 448–57.
15. Liu L, Li Y, Wu Y, Ren L, Wang G. LGI Net: enhancing local-global information interaction for medical image segmentation. Comput Biol Med. 2023;167:107627. doi:10.1016/j.compbiomed.2023.107627.
16. Liang C, Huang K, Mao J. Global-local deep fusion: semantic integration with enhanced transformer in dual-branch networks for ultra-high resolution image segmentation. Appl Sci. 2024;14(13):5443. doi:10.3390/app14135443.
17. Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, et al. Advances in medical image analysis with vision transformers: a comprehensive review. Med Image Anal. 2024;91:103000. doi:10.1016/j.media.2023.103000.
18. Chen C, Isa NAM, Liu X. A review of convolutional neural network based methods for medical image classification. Comput Biol Med. 2025;185:109507. doi:10.1016/j.compbiomed.2024.109507.
19. Li X, Li M, Yan P, Li G, Jiang Y, Luo H, et al. Deep learning attention mechanism in medical image analysis: basics and beyonds. Int J Netw Dyn Intell. 2023;2(1):93–116. doi:10.53941/ijndi0201006.
20. Zhou Q, Huang Z, Ding M, Zhang X. Medical image classification using light-weight CNN with spiking cortical model based attention module. IEEE J Biomed Health Inform. 2023;27(4):1991–2002. doi:10.1109/JBHI.2023.3241439.

21. Knigge DM, Romero DW, Gu A, Gavves E, Bekkers EJ, Tomczak JM, et al. Modelling long range dependencies in n d: from task-specific to a general purpose CNN. arXiv:2301.10540. 2023.
22. Bhati A, Gour N, Khanna P, Ojha A, Werghi N. An interpretable dual attention network for diabetic retinopathy grading: IDANet. Artif Intell Med. 2024;149(7):102782. doi:10.1016/j.artmed.2024.102782.
23. Sharma RR, Sungheetha A, Tiwari M, Pindoo IA, Ellappan V, Pradeep G. Comparative analysis of vision transformer and CNN architectures in medical image classification. In: Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024); 2025 Dec 30–31; Chennai, India. p. 1343–55.
24. Yuan F, Zhang Z, Fang Z. An effective CNN and Transformer complementary network for medical image segmentation. Pattern Recognit. 2023;136:109228. doi:10.1016/j.patcog.2022.109228.
25. Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, et al. Lvit: language meets vision transformer in medical image segmentation. IEEE Trans Med Imaging. 2023;43(1):96–107. doi:10.1109/TMI.2023.3291719.
26. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, et al. Transformers in medical image analysis. Intell Med. 2023;3(1):59–78.
27. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal. 2024;97:103280. doi:10.1016/j.media.2024.103280.
28. Mzoughi H, Njeh I, BenSlima M, Farhat N, Mhiri C. Vision transformers (ViT) and deep convolutional neural network (D-CNN)-based models for MRI brain primary tumors images multi-classification supported by explainable artificial intelligence (XAI). Vis Comput. 2025;41(4):2123–42. doi:10.1007/s00371-024-03524-x.
29. Abbas J, Soomro DB, Buriro M, Khan MY, Abbas S, Afzal M. LGAF-Net: a local-global attention fusion network for efficient medical image classification. In: Proceedings of the 2025 5th International Conference on Digital Futures and Transformative Technologies (ICoDT2); 2025 Dec 17–18; Islamabad, Pakistan. p. 1–6.
30. Khaniki MAL, Mirzaeibonehkhater M, Manthouri M, Hasani E. Brain tumor classification using vision transformer with selective cross-attention mechanism and feature calibration. arXiv:2406.17670. 2024.
31. Gulsoy EK, Ayas S, Kablan EB, Ekinci M. Enhancing the adversarial robustness in medical image classification: exploring adversarial machine learning with vision transformers-based models. Neural Comput Appl. 2025;37(12):7971–89. doi:10.1007/s00521-024-10516-4.
32. Badar D, Abbas J, Alsini R, Abbas T, Chengliang W, Daud A. Transformer attention fusion for fine grained medical image classification. Sci Rep. 2025;15(1):20655. doi:10.1038/s41598-025-07561-x.
33. Abbas J, Abbas S, Liu L. Gradient-guided causal attention mechanism for interpretable skin lesion classification. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Berlin/Heidelberg, Germany: Springer; 2025. p. 336–49.
34. Soares RR, Monteiro FP, Pinheiro GM, Serrão MKM, Costa Filho CF, Costa MG. Evaluation of depth-wise separable convolution and channel attention mechanism to bacilli segmentation. In: Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2024 Jul 15–19; Orlando, FL, USA. p. 1–5.
35. Huang A, Son J, Xiong Z. DDSNet: a lightweight dense depthwise separable network for tumor classification. In: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing; 2025 Mar 31–Apr 4; Sicily, Italy. p. 1114–21.
36. Zhou Y, Kang X, Ren F, Lu H, Nakagawa S, Shan X. A multi-attention and depthwise separable convolution network for medical image segmentation. Neuro Comput. 2024;564(3):126970. doi:10.2139/ssrn.4495223.
37. Liu S, Wang P, Lin Y, Zhou B. SMRU-Net: skin disease image segmentation using channel-space separate attention with depthwise separable convolutions. Pattern Anal Appl. 2024;27(3):93. doi:10.1007/s10044-024-01307-7.
38. Zheng J, Liu H, Feng Y, Xu J, Zhao L. CASF-Net: cross-attention and cross-scale fusion network for medical image segmentation. Comput Methods Programs Biomed. 2023;229:107307. doi:10.1016/j.cmpb.2022.107307.
39. Yan Q, Liu S, Xu S, Dong C, Li Z, Shi JQ, et al. 3D medical image segmentation using parallel transformers. Pattern Recognit. 2023;138(10):109432. doi:10.1016/j.patcog.2023.109432.
40. Yue Y, Li Z. Medmamba: Vision mamba for medical image classification. arXiv:2403.03849. 2024.

41. Chang Y, Li Z. Attention-enhanced dual-stream registration network via mixed attention transformer and gated adaptive fusion. Med Image Anal. 2025;105(1):103713. doi:10.1016/j.media.2025.103713.
42. Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X. Vision mamba: efficient visual representation learning with bidirectional state space model. arXiv:2401.09417. 2024.
43. Qiu L, Zhao L, Hou R, Zhao W, Zhang S, Lin Z, et al. Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features. Comput Med Imaging Graph. 2023;104(4):102176. doi:10.1016/j.compmedimag.2022.102176.
44. Fang Z, Zhu S, Chen Y, Zou B, Jia F, Liu C, et al. GFE-Mamba: mamba-based AD multi-modal progression assessment via generative feature extraction from MCI. arXiv:2407.15719. 2024.
45. Zahoor MM, Khan SH. CE-RS-SBCIT a novel channel enhanced hybrid CNN transformer with residual, spatial, and boundary-aware learning for brain tumor MRI analysis. arXiv:2508.17128. 2025.
46. Xu T, Xiang Y, Du J, Zhang H. Cross-scale attention and multi-layer feature fusion YOLOv8 for Skin disease target detection in medical images. J Comput Technol Softw. 2025;4(2):14984982. doi:10.5281/zenodo.14984982.
47. Qezelbash-Chamak J, Hicklin K. A hybrid learnable fusion of ConvNeXt and swin transformer for optimized image classification. IoT. 2025;6(2):30. doi:10.3390/iot6020030.
48. Vidhya S, Nithya R. A transformer driven hybrid feature fusion framework for multi-modal medical image analysis. Preprint. 2025. doi:10.21203/rs.3.rs-7712468/v1.
49. Huo X, Sun G, Tian S, Wang Y, Yu L, Long J, et al. HiFuse: hierarchical multi-scale feature fusion network for medical image classification. Biomed Signal Process Control. 2024;87:105534. doi:10.1016/j.bspc.2023.105534.
50. Zhou Z, Fu C, Wang C, Zhu P, Xia K, Qian P. Decoupled feature extraction and correspondence modeling for deformable medical image registration using large kernel attention. Alex Eng J. 2026;134:556–69. doi:10.1016/j.aej.2025.12.033.
51. Pan W, Kang J, Wang X, Lv C, Zhang X, Liu X. MedCTM: a CNN-transformer-mamba hybrid network for medical image classification. Inf Process Manag. 2026;63(7):104867. doi:10.1016/j.ipm.2026.104867.
52. Han M, Chen J, Yao M, Xie B. MediTEDNet: visual state space model for thyroid eye disease classification. IEEE J Biomed Health Inform. 2025;30(3):2550–62. doi:10.1109/JBHI.2025.3608153.
53. Wang Z, Yin H, Yu J, Gong M, Chen Q, Chen Q, et al. DSA mamba: a model for advanced medical image classification. Expert Syst Appl. 2025;299:130064. doi:10.1016/j.eswa.2025.130064.
54. Pan W. TET Loss: a temperature-entropy calibrated transfer loss for reliable medical image classification. J Imaging Inform Med. 2026:1–13. doi:10.1007/s10278-025-01816-9.
55. Pan W, Wang X. MiT Loss: medical image-aware transfer-calibrated loss for enhanced classification. Meas Sci Technol. 2025;36(10):105404. doi:10.1088/1361-6501/ae08d8.
56. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10(5):e1379. doi:10.1002/widm.1379.
57. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923. 2017.
58. Guluwadi S. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Med Imaging. 2024;24(1):1–19. doi:10.1186/s12880-024-01292-7.
59. Raveenthini M, Lavanya R, Benitez R. Grad-CAM based explanations for multiocular disease detection using Xception net. Image Vis Comput. 2025;154(1):105419. doi:10.1016/j.imavis.2025.105419.
60. Kumaran SY, Jeya JJ, Mahesh TR, Khan SB, Alzahrani S, Alojail M. Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam. BMC Med Imaging. 2024;24(1):176. doi:10.1186/s12880-024-01345-x.
61. Wang J, Bhalerao A, Yin T, See S, He Y. Camanet: class activation map guided attention network for radiology report generation. IEEE J Biomed Health Inform. 2024;28(4):2199–210. doi:10.1109/JBHI.2024.3354712.
62. Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, et al. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference; 2017 Jun 20–23; Taipei, Taiwan. p. 164–9.

63. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (isic). arXiv:1902.03368. 2019.
64. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 4700–8.
65. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
66. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22.
67. Sambyal AS, Niyaz U, Shrivastava S, Krishnan NC, Bathula DR. Ls+: informed label smoothing for improving calibration in medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin/Heidelberg, Germany: Springer; 2024. p. 513–23.
68. Peled S, Maruvka YE, Freiman M. PSA-MIL: a probabilistic spatial attention-based multiple instance learning for whole slide image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2026 Jun 4–8; Buena Vista, FL, USA. p. 1211–20.
69. Huang Y, Cheng P, Tam R, Tang X. Fine-grained prompt tuning: a parameter and memory efficient transfer learning method for high-resolution medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin/Heidelberg, Germany: Springer; 2024. p. 120–30.
70. Zhu Z, Fan L, Pagnucco M, Song Y. Interpretable image classification via non-parametric part prototype learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 11–15; Nashville, TN, USA. p. 9762–71.