Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

QLViT: A Lightweight Cell Classification Method for Microscope Images Based on MViTv2 and Linear Attention

Panpan Wu[1], Zhangda Liu[1], Ziping Zhao[1*], Rui Guo[2], Hengyong Yu[3]

[1] College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, USA
E-mail: ztianjin@126.com

Received: 4 July 2025; **Revised:** 9 September 2025; **Accepted:** 16 September 2025

Abstract: Accurate cell classification plays a vital role in the diagnosis and treatment of diseases. However, existing methods face challenges such as limited feature learning and excessive computational complexity, resulting in low classification accuracy, prolonged training processes, and slow inference speeds. We propose a novel lightweight method, Quantized Linear Vision Transformer (QLViT), based on the Multiscale Vision Transformers (MViTv2) and linear attention mechanisms, to facilitate cell classification tasks from microscope images. Specifically, QLViT employs a large-kernel convolutional layer and a well-designed feature extraction module called Conv-Linear Attention (CLA) to extract features. It optimizes self-attention with an activation function and utilizes a residual structure to facilitate feature reuse and address gradient issues. The CLA ensures efficient learning of local information via dynamic convolution and employs linear attention to comprehensively capture global features, maintaining a lightweight profile compared to the traditional self-attention. By introducing the Kolmogorov-Arnold Network (KAN) structure, CLA significantly reduces computational complexity and parameter count. Extensive experiments on four public datasets demonstrate the effectiveness of QLViT. We achieve an accuracy of 97.19% on the BioMediTech dataset, 97.35% on the ICPR-HEp-2 dataset, 90.45% on the blood malignancy bone marrow cytology expert-annotated dataset for a six-category classification task, and an impressive accuracy of 99.84% on the white blood cell dataset. Furthermore, our method exhibits a computational efficiency of 1.95 Giga Floating-point Operations (GFLOPs) and utilizes 9.07 million parameters. Our results show that QLViT outperforms current state-of-the-art methods across multiple datasets, demonstrating its superior inference speed, lightweight design, strong feature extraction capabilities and generalizability. This proposed method provides a promising solution in the field of medical image classification.

***Keywords*:** cell classification, linear attention, quantitative methodology, kolmogorov-arnold network

MSC: 68T07, 68T20, 68T45

1. Introduction

Cell classification refers to the process of categorizing cells based on their morphological, functional, gene expression, and surface markers. This task is of great importance in the medical field because it helps to understand the properties,

DOI: https://doi.org/10.37256/cm.7120267713

functions, and roles of different cell types within an organism. Many diseases are associated with abnormalities in specific cell types [1]. Clinical physicians can improve diagnostic accuracies and develop personalized therapies based on the results of cell classification. In the early stages, cell classification primarily relied on the traditional image processing techniques and classical machine learning methods. For example, Computer-Aided Diagnosis (CAD) systems enhanced diagnostic accuracy through methods such as decision trees, Bayesian networks, and rule extraction [2]. These methods are suitable for simpler scenarios and are advantageous due to their low computational complexity. However, due to the limitations of manually designed features, these approaches often fail to analyze images comprehensively, leading to poor generalizability [3].

With the advancement of deep learning (particularly Convolutional Neural Networks (CNNs)), significant breakthroughs have been made in cell classification techniques [4]. Researchers have increasingly turned to CNNs to automatically learn features from cell images [5]. For example, Kutlu et al. [6] and Shafique et al. [7] demonstrated that deep CNNs can accurately identify various morphological cell types in blood images. Promising results have also been reported in the detection of acute lymphoblastic leukemia using pre-trained neural networks such as AlexNet. Furthermore, Song et al. [8] proposed a method that uses deep autoencoder networks within a single architecture to classify irregularly shaped cells in bone marrow histology images accurately. Although CNNs have significantly improved cell classification accuracy by automatically extracting multi-level features, the diversity and complexity of cells may lead to suboptimal performance when using general-purpose networks for specific cell-related tasks.

Since the introduction of the self-attention mechanism [9], self-attention-based models have gradually surpassed CNNs in image processing. Self-attention methods are now widely used to enhance image classification performance. To the best of our knowledge, there have been several studies [10–12] that focus on utilizing the traditional self-attention mechanisms or modified self-attention based on the Vision Transformer (ViT) [13] architectures for cell classification tasks. However, these studies do not fully address the issues associated with excessive computational complexity and limited learning of local information inherent in self-attention mechanisms. Very little research has been conducted to explore self-attention-based models for extracting relevant cell classification information from microscope images.

To address the aforementioned challenges, we propose the Quantized Linear Vision Transformer (QLViT) approach, which primarily leverages a designed feature extraction module Conv-Linear-Attention (CLA) and large kernel convolutions for feature representation. The activation function is used to optimize self-attention, and a residual structure is introduced to mitigate gradient-related issues. In the CLA module, we implement dynamic convolution to flexibly adjust the weights of the convolutional kernels. In addition, we incorporate a linear attention mechanism with lower computational complexity to capture global information, replacing traditional self-attention methods. To further enhance the model's interpretability and feature representation capacity, inspired by the Kolmogorov-Arnold representation theorem [14], we employ the KAN [15] structure, replacing the conventional multilayer perceptrons. This makes the model more lightweight and better equipped to capture both local and global information, further boosting model performance.

The main contributions of this article are as follows. 1) We propose a novel CLA feature extraction module that integrates dynamic convolution, a linear attention mechanism, and the KAN structure. This module can thoroughly learn feature representations while maintaining fewer parameters and lower computational complexity. 2) By introducing the quantization operation to optimize linear layers and activation functions, training cost of the model is significantly reduced and inference speed is notably accelerated. 3) The proposed QLViT approach for the classification of microscope cell images effectively overcomes the limitations of traditional methods, enhancing the performance and generalizabilty, thus offering a new solution and direction for microscope cell image classification.

2. Related work

From the literature, it can be observed that traditional machine learning methods [16, 17] are highly relying on manual feature engineering and exhibit poor scalability when handling large-scale and complex image data. Their performance often lags behind that of deep learning models and they are sensitive to noise and outliers, making it more difficult to tune the parameters [18]. Deep learning models have obtained a range of achievements in the field of medical image processing,

and also in the classification of microscope cell images. Conventional machine learning techniques have gradually been replaced by CNNs due to their superior performance to analyze the cell images captured under a microscope for the identification and classification of different cell types. Asghar et al. [19] noted that while White Blood Cell (WBC) classification was initially carried out based on traditional machine learning methods, it has been shifted towards deep learning methods, particularly CNNs, in recent years. Among 136 relevant studies published between 2006 and 2023, 54.8% employed CNN. For example, Park et al. [9], Yenurkar et al. [10], and Davamani et al. [11] used different CNN-based approaches (integrated models, pre-trained models, and migration learning techniques) to classify cells, respectively. DEV et al. [20] developed several hybrid data-driven models that combined CNNs for feature extraction with two cascaded Recurrent Neural Network (RNN) classifiers to analyze images of red blood cells infected with malaria. Huang et al. [21] proposed the MGCNN model, a hyperspectral imaging-based blood cell classification method that integrates modulated Gabor wavelets with CNN kernels to effectively promote blood cell classification performance in small sample training scenarios. In summary, while there are a series of research works utilizing CNNs for cell classification tasks, the capability of learning global information is insufficient due to the limitations of their receptive fields, particularly when faced with complex data distributions, thus hindering their ability to establish effective global contextual relationships.

The self-attention mechanism enables efficient learning in terms of the importance of different feature information, gradually outperforming CNNs. For example, the ViT model, which was the first to introduce attention mechanisms into the field of computer vision, has achieved outstanding results on datasets such as ImageNet, CIFAR-100, and VTAB. Furthermore, with the ViT model, Halder et al. [12] performed a comprehensive analysis of the MedMNISTv2 dataset that include blood cell microscopy images, showing the potential of self-attention mechanisms in medical image analysis. The MViTv2 model proposed by Li et al. [22] employed a multiscale visual self-attention mechanism and enhanced the models performance by introducing decomposed relative position embeddings and residual pooling connections. However, self-attention mechanisms tend to have high computational complexity in image classification and are more concentrated on global information, often ignoring the learning ability of local features.

Recently, lightweight technologies have also made progress across various domains. For instance, CAS-ViT effectively streamlines the complex Transformer architecture through its innovative CATM module, enabling efficient operation on resource-constrained devices such as mobile platforms while achieving competitive performance across multiple visual tasks [23]. Flat U-Net focuses on domain-specific lightweighting strategy. For solar filament segmentation tasks, it significantly reduces model parameters and enhances segmentation efficiency through optimized convolutional blocks and a flattened network architecture [24]. Furthermore, DyT technology offers an alternative approach to Transformer models by eliminating normalization layers altogether. Through simple element-wise operations, it not only simplifies model structure but also maintains comparable performance to normalized models [25]. Collectively, these efforts advance lightweight technology, providing novel insights and methodologies for efficient, low-resource model deployment in a wide range of applications.

The KAN network is a novel neural network architecture inspired by the Kolmogorov-Arnold Representation Theorem, distinguished by its unique structural design. In KAN, the fixed weight matrices typically used in traditional networks are replaced with learnable univariate functions parameterized by spline curves. By leveraging this formulation, KAN effectively mitigates the difficulty of dimensionality often encountered by conventional neural networks in high-dimensional function approximation tasks. It exhibits stronger expressive power and improved accuracy, while also opening new pathways to improve the interpretability of neural network models.

To this end, our proposed QLViT method incorporates a large-kernel convolutional layer and an innovative CLA module as the feature extractor, enhancing feature representation capabilities while reducing computational complexity in identifying microscope images. By integrating an optimized activation function, dynamic convolution, and linear attention, along with the KAN structure, our approach effectively captures both high-level semantic information and local characteristics, significantly reducing computational overhead and parameter demands.

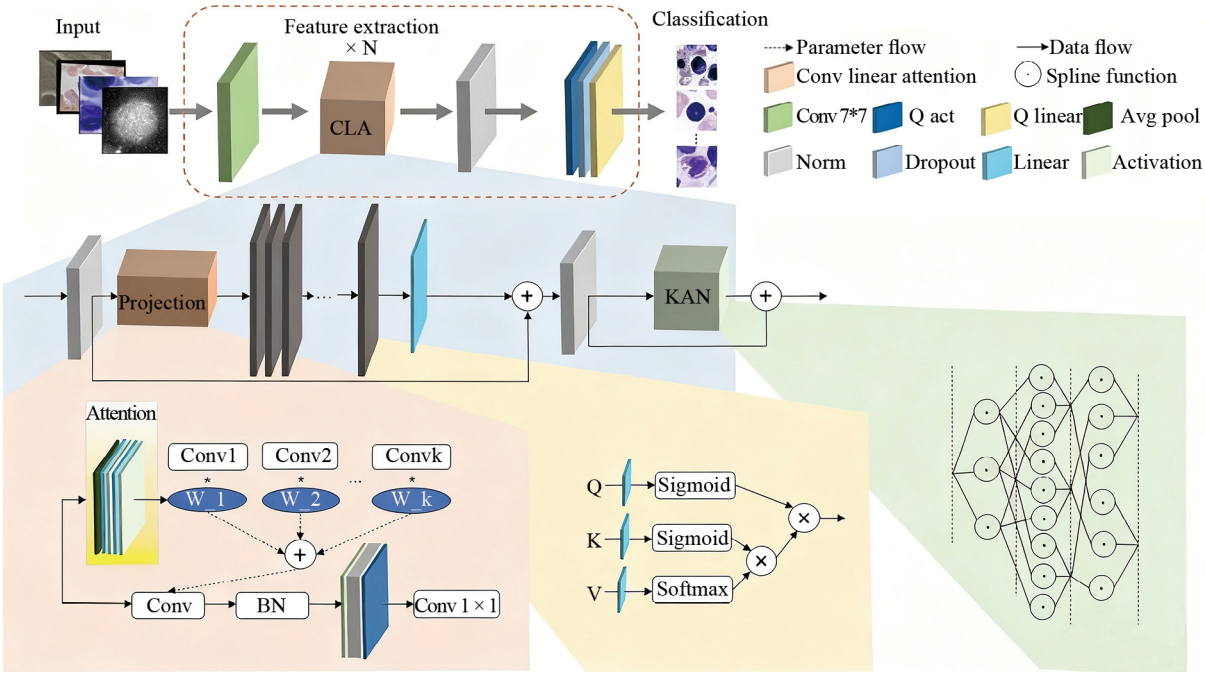

Figure 1. Overall architecture of the proposed QLViT model

3. Research method

3.1 *Overall architecture*

The overall architecture of the proposed QLViT approach is shown in Figure 1. It primarily includes the following three steps: large kernel convolution for coarse feature extraction, the CLA module for refined feature learning, and the output of image classification results. First, the cell images are fed into the large kernel convolution layer, where the extensive receptive field effectively learns the features of each image. Next, the data is passed to the well-assembled CLA module, which first employs dynamic convolution to enhance the detailed features of the image. The quantized activation function is then used to match the performance of traditional activation functions while providing greater computational efficiency. In linear attention, information is integrated into the activation function during computation to alter the order of QKV calculations, achieving linear complexity. A residual structure for feature reuse is introduced to mitigate gradient problems during training, following the self-attention mechanism. The CLA module is repeated N times to learn cell image features thoroughly. Finally, after normalization and the quantized linear layer, the classification results are output.

3.2 *CLA module*

The core component in QLViT is the designed CLA module. After a feature map is input, it first passes through a norm layer to enhance the performance of the neural network by stabilizing the gradients, thereby accelerating model convergence. It also reduces internal covariate shift, which refers to changes in the distribution of input data, thus improving the model's generalizability. The following is the Projection part of the model, which begins with a mapping layer. Dynamic convolution [26] allows the model to better capture complex structures of the input data, thereby increasing its feature representation capabilities. The weights of each convolutional kernel are dynamically adjusted as the attention weights of the input feature map vary. It uses K convolutional kernels $\widetilde{W}$ of the same size, with identical input and output dimensions, and the final convolutional operation is performed by aggregation of the attention weights $\pi_k(x)$. The output of the dynamic convolution can be expressed as:

$$C = A\left(\widetilde{W}^{T}(x)x + \widetilde{b}(x)\right), \tag{1}$$

$$\widetilde{W}(x) = \sum_{k=1}^{K} \pi_k(x)\widetilde{W}_k, \tag{2}$$

$$\widetilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\ \widetilde{b}_k, \tag{3}$$

$$s.t.\ 0 \le \pi_k(x) \le 1,\ \ \sum_{k=1}^{K} \pi_k(x) = 1, \tag{4}$$

where x is the input feature map, and A is an activation function, $\widetilde{W}_k$ and $\widetilde{b}_k$ are k^{th} weight matrix and bias vector. The attention weights $\widetilde{W}_k$ are computed through the compression of global spatial information and a fully connected layer, utilizing the Softmax function to ensure that the sum of all attention weights equals 1. This can be expressed as:

$$\pi_k = \frac{exp(z_k/\tau)}{\sum_{j=1}^{K} exp(z_j/\tau)}, \tag{5}$$

where z_k is the output of the final Linear layer in the attention brunch, and τ is used to control the sparsity of the attention distribution to addressing the efficiency issues associated with deep dynamic convolution. The Qact layer implements activation layer quantization, combining $\log_2$ quantization with iexp which is a polynomial approximation of exponential function [27]. This speeds up the model inference process without affecting performance. The Softmax operation is represented as follows:

$$QA(s \cdot X_k) = 2^b - 1 - \log_2\left\lfloor \frac{\sum_q \text{iexp}(X_q)}{\text{iexp}(X_k)} \right\rfloor, \tag{6}$$

where iexp denotes the iexp function, s is a scaling factor, and X is the input. The term iexp (X_k) provides an integer approximation of the exponential function. This avoids using floating-point numbers in the Softmax operation, allowing efficient hardware execution using integers. Finally, the mapping layer concludes with a 1×1 convolution, which can alter the number of channels in the feature map, facilitating cross-channel information integration and aiding in the extraction of more abstract feature representations. The self-attention mechanism in this module references Flowformer [28], where the outputs Q, K, and V are obtained through the mapping layer. The Q and K matrices are processed through a sigmoid function to ensure the non-negativity of the attention matrix. In contrast to traditional attention calculations, this approach integrates specific activation functions into feature computations, using matrix multiplication optimization techniques. It first performs operations on K and V rather than the conventional approach of first processing Q and K. V represents the importance of each token, and the use of the Softmax function significantly differentiates the levels of importance. Q and K indicate the amount of information each token needs to receive, with the sigmoid function effectively acting as a gating mechanism to control the volume of incoming information. The traditional attention [9] calculation formula is as follows:

$$\text{Attention}\,(Q,\ K,\ V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \tag{7}$$

The attention calculation formulas related to our model are as follows:

$$\text{LinearAttention}\,(Q,\ K,\ V) = \frac{\text{Sigmoid}\,(Q)\,\text{Sigmoid}\left(K^T\right)\text{Softmax}\,(V)}{\sqrt{d_k}}, \tag{8}$$

where d_k is the dimensionality of the data.

$$\text{Sigmoid}(x_{ij}) = \frac{1}{1+e^{-x_{ij}}} \quad i \in (1,\ \ldots,\ n) \quad j \in (1,\ \ldots,\ m), \tag{9}$$

where x_{ij} denotes the element in the i^{th} row and j^{th} column of the matrix $X \in R^{n\times m}$. Each element in the matrix is transformed using the Sigmoid function.

$$\text{Softmax}\,(\,X\,)_j = \left(\left[\frac{e^{x_{1j}}}{\sum_{i-1}^{n} e^{x_{ij}}},\ \frac{e^{x_{2j}}}{\sum_{i-1}^{n} e^{x_{ij}}},\ \ldots,\ \frac{e^{x_{nj}}}{\sum_{i-1}^{n} e^{x_{ij}}}\right]\right)^T. \tag{10}$$

Equation (10) performs the corresponding Softmax operation on each column j of the matrix $X \in R^{n\times m}$, where x_{ij} represents the element in the i^{th} row and j^{th} column of the matrix X.

Finally, the residual structure [29] is used to alleviate the gradient vanishing problem and aid gradient flow while training a deeper network model. For example, as shown in 1, the features that pass through both the norm layer and the KAN layer involve residual structures. In common neural networks, MLP is used as the base component of the model. However, in our model, this practice is abandoned and a KAN structure based on the Kolmogorov-Armold representation theorem is used, which is more interpretable and has a smaller number of parameters. Unlike MLPs, the inputs in the KAN structure are directly and nonlinearly transformed before the inputs are combined. Multiple curves are combined to simulate arbitrary functions using spline functions. Parametric learning with spline functions is more difficult than with linear functions, but nonlinear representations are much better and can achieve higher accuracy with fewer nodes. The following is a representation of the KAN structure. Let f: $[0,\ 1]^n \to R$ be a smooth multivariable function. According to the Kolmogorov-Arnold

$$f\,(x) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,\ p}\,(\,x_p\,)\right), \tag{11}$$

where $\phi_{q,\ p}$: $[0,\ 1] \to R$ and $\Phi_q : [0,\ 1] \to R$ are univariate continuous functions, and $x = (x_1,\ \ldots,\ x_n)$ represents the input variables.

$$\Phi = \left\{\phi_{q,\ p}\right\}, \qquad p = 1,\ 2,\ \ldots,\ n, \quad q = 1,\ 2,\ \ldots,\ m, \tag{12}$$

where n and m represent the input and output dimensions of the KAN layer, respectively, and each univariate function contains trainable parameters.

For a deep KAN structure, multiple KAN layers can be stacked, with each KAN layer defined as a matrix composed of a set of univariate functions, as shown in Equation (13). The general form of KAN is as follows:

$$\mathrm{KAN}(X) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \cdots \circ \Phi_2 \circ \Phi_1) X. \tag{13}$$

3.3 *Linear layer quantization operation*

In QLViT, the final layer employs a quantized linear layer to further enhance the model's inference speed. The quantized linear layer utilizes a uniform quantization method [30, 31], mapping floating-point numbers X to the nearest quantization bin. The definition of the quantizer is as follow.

$$Q(X|b) = clip\left(\left\lfloor \frac{X}{s} \right\rceil + z_p,\ 0,\ 2^b - 1\right), \tag{14}$$

$$s = \frac{\mu - l}{2^b - 1}, \tag{15}$$

$$l = \min(X), \tag{16}$$

$$\mu = \max(X), \tag{17}$$

$$z_p = clip\left(-\left\lfloor \frac{X}{s} \right\rceil,\ 0,\ 2^b - 1\right), \tag{18}$$

where, b is the quantization bit-width, which determines the precision and range of values. l and μ define the range of quantized values, while the parameters s (scale) and z_p (zero-point) are determined by the upper bound μ and lower bound l of X. The "clip" function ensures that the quantized value does not fall outside this range. If the quantized value is less than 0, it is set to 0; If the quantized value is greater than or equal to $2^b - 1$, it is set to $2^b - 1 - 1$.

3.4 *Loss function*

In this approach, cross-entropy loss is used as the loss function. This is based on the concept of cross-entropy defined in information theory, whereby the distance between two probability distributions is measured. The aim is to minimize the loss by making the predicted distribution as similar as possible to the true distribution, with the difference between the two distributions calculated. The formula for cross-entropy loss is as follow.

$$\text{Cross Entropy Loss} = -\sum_{i=1}^{C} y_i \log(p_i), \tag{19}$$

where C is the number of classification categories and y_i is the distribution of the true labels, which is 1 for the correct category and 0 for the others. p_i is the output probability of the model, representing the probability that the sample belongs to category i. The cross-entropy loss function is well-suited for gradient descent and backpropagation due to its smooth properties. Based on probability distributions, it gives model output a clear probabilistic interpretation, which aids in understanding and interpreting the model.

4. Experiments

4.1 *Dataset*

The details of the four datasets used in the experiment are summarized in Table 1.

Table 1. Summary of the datasets used in the experiment

Name	Number of categories	Cell type	Quantity and category names for each category
BioMediTech dataset [32]	4	Retinal cells	216 Fusiform, 547 Epithelioid, 949 Cobblestone, 150 Mixed
White blood cell dataset [33]	4	Blood cells	3,133 Eosinophils, 3,108 Lymphocytes, 3,095 Monocytes, 3,171 Neutrophils
ICPR-HEp-2 dataset [34]	6	HEp-2 cells	2,495 Homogeneous, 2,831 Speckled, 2,598 Nucleolar, 2,741 Centromere, 2,208 Nuclear membrane, 724 Golgi
Hematological malignancy bone marrow cytology expert annotation dataset [35]	6	Bone marrow cells	5,883 Eosinophils, 3,055 Degenerative myelocytes, 4,040 Monocytes, 6,557 Bone marrow, 3,538 Non-Identifiable elements, 2,740 Proerythroblasts

1) The BioMediTech dataset is composed of 195 initial microscopic images of retinal pigment epithelial cells captured at various stages. These images are segmented into 16 smaller images through a 4×4 grid division, yielding a total of 1,862 images after removing those that are cluttered, blurry, or consist solely of background. Each sub-image is classified by two professional annotators.

2) The ICPR-HEp-2 dataset comprises HEp-2 cells that display a variety of nuclear antigens, which renders it a perfect medium for conducting Indirect Immunofluorescence (IIF) experiments. This dataset, sourced from the University of Salerno, features fluorescence microscopy images of HEp-2 cells exhibiting a range of morphological characteristics.

3) The dataset for hematological malignancies in bone marrow cells encompasses over 170,000 anonymized cells that have been annotated by experts. These cells were obtained from the bone marrow smears of 945 patients. The smears were stained using the May-Grünwald-Giemsa/Pappenheim technique, and the images were taken with a 40 × oil immersion microscope. The full dataset covers a variety of conditions across 21 different cell types. However, given the uneven distribution of the data, we chose six cell types that have a more balanced representation.

4) The leukocyte classification dataset contains 12,500 enhanced images of blood cells, each accompanied by cell type labels. The diagnosis of blood-related diseases usually involves the identification and characterization of patient blood samples, and methods for the automatic detection and classification of blood cell subtypes hold significant value for medical applications.

For each dataset, the data is partitioned using stratified random sampling. Within each category, samples are randomly selected at a 20% ratio to form the testing set, while the remaining data automatically comprises the training set. Regarding data augmentation, we randomly apply rotations within the range of -15 to $+15$ degrees without resizing the images, followed by an automatic enhancement strategy. Additionally, all images are normalized across the three channels using the mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225].

4.2 *Evaluation metrics*

In this experiment, three commonly used evaluation metrics were used to assess the model performance: Accuracy (Acc), Recall, and F1 Score. For comprehensive evaluation across all classes in the multi-class setting, micro-averaging was applied to these metrics. Notably, under micro-averaging, Recall and F1 Score yield identical values to Accuracy. Thus, we primarily focus on Accuracy for performance interpretation. The formulas of accuracy is as follow.

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FP_i + FN_i)}, \tag{20}$$

where C is the total number of categories, and TP_i, FP_i, and FN_i are the number of true instances, false positive instances, and false negative instances of the i^{th} category, respectively.

4.3 *Comparative experiments*

4.3.1 *Experimental setup*

In this study, all baseline models are initialized using weights pre-trained from ImageNet . This process endows the models with robust feature extraction capabilities, enabling them to efficiently adapt to the complex features present in cell images. Through fine-tuning on four cell classification datasets, these models further optimize their performance for specific tasks. We conducted comparative experiments on the aforementioned four datasets with our proposed method and representative methods in the field of image classification, including ResNet [32], RegNet [36], ShuffleNetV2 [37], Vision Transformer (ViT) [38], SwinTransformer [39], ConvNext [36], and MViTv2. During the training stage, all comparative models with pre-trained weights are fine-tuned for 50 epochs on the same dataset. Because our redesigned model does not leverage pre-trained weights, it underwent a longer training process of 500 epochs to achieve convergence and competitive performance.

ResNet addresses the issue of model degradation in deep networks by employing shortcut connections. Unlike conventional networks, ResNet introduces shortcut mechanisms that facilitate residual learning, enabling deeper networks to function effectively. RegNet is a family of simple and fast networks derived from a network design space for various FLOP levels. ShuffleNetV2 is a model with high efficiency and accuracy, and it is developed based on four proposed criteria: (1) minimizing memory cost (MAC) through equal channel width, (2) avoiding excessive group convolutions that increase MAC, (3) reducing network fragmentation to enhance parallelism, and (4) considering the impact of element-wise operations. ViT divides the input images into multiple patches, which are then fed into a Transformer model. It excels in tasks that involve high-resolution images and require capturing long-range dependencies. Swin Transformer combines the strengths of Transformers and CNN. This architecture demonstrates strong generalizability. ConvNeXt integrates some advantages of Transformer architecture while maintaining the efficiency and intuitiveness of CNNs. MViTv2 is a versatile model designed for image and video classification as well as object detection. It achieves superior performance compared to previous work by employing multiscale self-attention, decomposed relative position embeddings, and residual pooling connections.

4.3.2 *Experimental results*

Table 2 present the quantitative comparison results of QLViT against classical models in four datasets. It can be seen that our method significantly outperforms others in terms of four metrics on all four datasets. In Particular, on the BioMediTech dataset, the classification accuracy of our method far exceeds that of the state-of-the-art models. For example,compared to ShuffleNet, it increased by 10%. Furthermore, in the white blood cell classification dataset, our model raises the ACC to 99.84%. Our method consistently yields better results compared to the baseline on all four datasets. Notably, on the BioMediTech dataset, our method surpassed the baseline MViTv2 by 7% in classification accuracy. This improvement is likely due to the incorporation of dynamic convolution, the improved linear attention mechanism, and the KAN structure for feature extraction, which provide more effective and computationally efficient

learning. These experimental results indicate that by introducing the novel CLA module, our QLViT method exhibits superior feature representations, leading to notable promotion over other methods for cell classification on the microscope images.

To comprehensively evaluate the model's classification performance, confusion matrices are provided for each dataset in Figure 2. These matrices offer detailed insights into the model's predictive behavior by disaggregating the number of true positives, true negatives, false positives, and false negatives across all categories. The model achieves perfect classification on balanced datasets and categories with distinctive features, such as the White Blood Cell dataset in Figure 2b and the Cobblestone category in Figure 2a-the latter characterized by a distinct dobblestone-like texture pattern. However, on imbalanced dataset, the confusion matrix also reveals the model's tendency to misclassify samples from the minority class. For instance, the model shows notably poor performance in classifying the Golgi category in Figure 2c, with a pronounced error rate. This behavior aligns with common challenges posed by class imbalance, likely resulting from the model developing a bias toward the majority class during training.

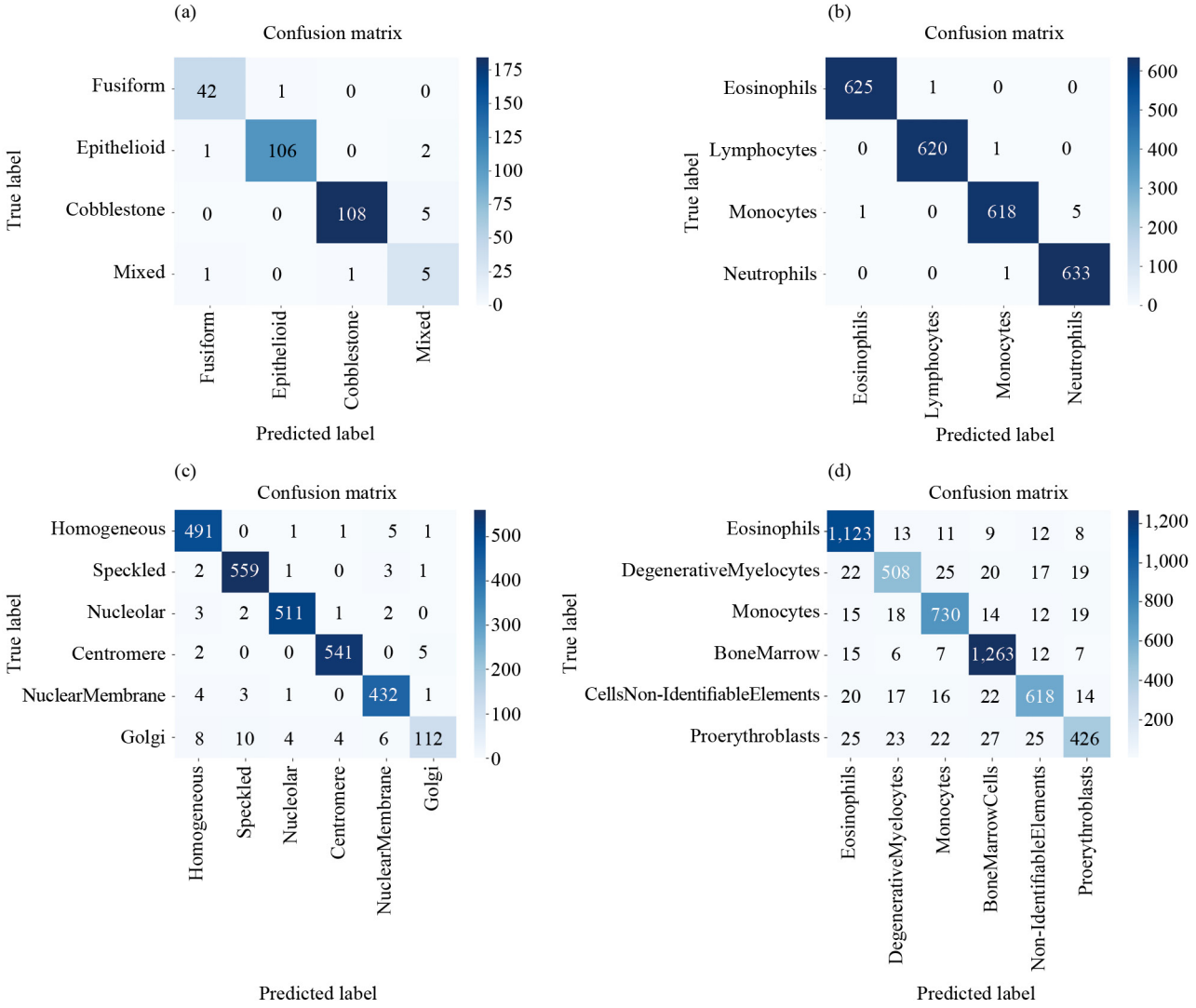

Figure 2. Confusion matrices for different datasets. (a) BioMediTech, (b) White blood cell, (c) ICPR-HEp-2, and (d) Bone marrow

Table 2. Comparison of quantitative results from various models across four datasets

Model	Dataset (Acc (%))			
	BioMediTech	WhiteBloodCell [30]	ICPR-HEp-2 [31]	BoneMarrow [32]
Swin transformer [35]	91.11	93.08	95.49	89.09
Vision transformer [13]	90.03	97.84	96.67	88.14
ResNet [29]	91.11	94.84	96.35	88.45
ConvNext [36]	91.91	94.84	94.70	88.49
ShuffleNet [34]	87.06	93.24	87.46	87.42
RegNet [33]	91.37	95.51	96.06	88.14
MViTv2 [25]	90.49	98.19	95.38	87.21
Ours	97.19	99.84	97.35	90.45

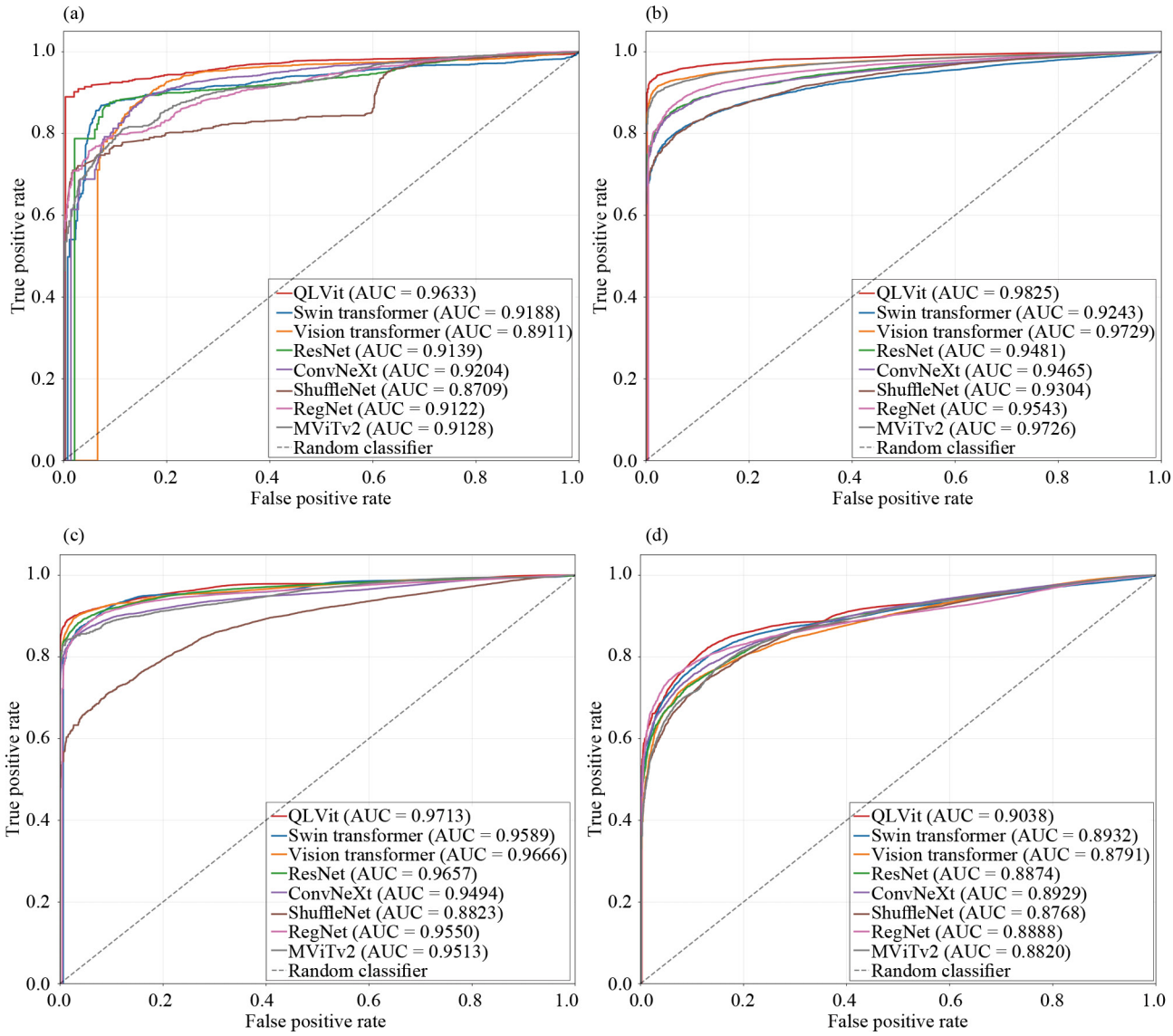

Figure 3. ROC curves of different models across four datasets. (a) BioMediTech, (b) White blood cell, (c) ICPR-HEp-2, and (d) Bone marrow

Figure 3 presents the ROC curves of different models across four datasets. Our model demonstrates outstanding classification performance on all datasets, achieving high AUC values of 0.9633 on BioMediTech dataset, 0.9825 on White Blood Cell dataset, 0.9713 on ICPR-HEp-2 dataset, and 0.9038 on Bone Marrow dataset, with strong and consistent stability compared to other models. Notably, in Figure 3a, the curve rises rapidly in the low false positive rate region, reflecting the model's ability to capture a substantial number of true positives at an early stage.

In several typical misclassification cases, some datasets contain cell images with high visual similarity. For instance, a sample from the Fusiform class (Figure 4a) in the BioMediTech dataset is misclassed as Mixed, and a Monocytes sample (Figure 4c) from the bone marrow dataset is incorrectly predicted as Proerythroblasts. These errors indicate that our model still struggles in distinguishing subtle textures variations. This mainly stems from the fact that texture features consist of multi-scale and multi-dimensional information, including both local and global texture characteristics. Without effectively integration of features across different levels, it remains incomplete for the model to comprehensively understand complex textures. Further improvements could focus on improving attention mechanisms to more efficiently guide convolutional feature extraction and improve multi-level feature fusion.

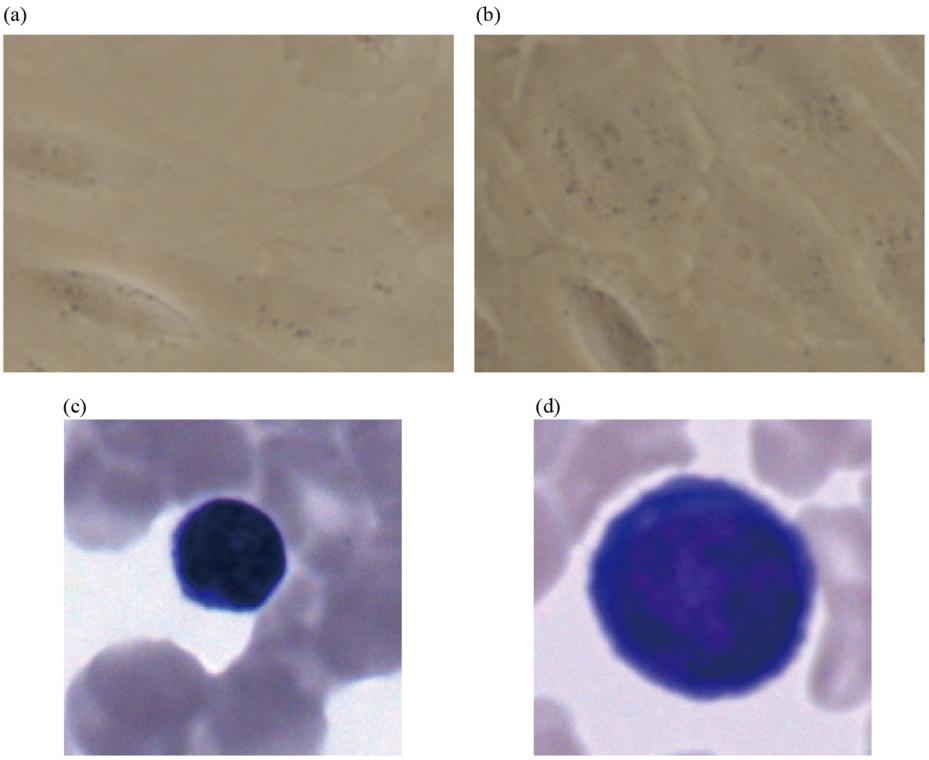

Figure 4. Examples of misclassified images in the dataset. (a) and (b) are misclassed samples from the BioMediTech dataset, belonging to the Fusiform and Mixed categories, respectively. (c) and (d) are misclassified instances from the bone marrow dataset, corresponding to Monocytes and Proerythroblasts, respectively

4.3.3 *Interpretability analysis*

We employ the Grad-CAM tool [40] to generate and visualize class activation maps for two datasets to further demonstrate the effects of our method. The visualization results can clearly exhibit each model's focus on features relevant to the classification task within the cellular images, highlighting the differences in cellular feature recognition among different models and enhancing the interpretability of the models.

Figure 5 shows a set of class activation maps of fusiform Retinal Pigment Epithelium (RPE) cells classification generated by different methods from the BioMediTech Dataset. It can be seen that for the Fusiform and the Epithelioid

images, our model's attention is more focused on each cell nucleus within the images compared to other models. In the Cobblestone category, the cells are tightly arranged under normal physiological conditions, forming a texture reminiscent of cobblestones, which are crucial characteristics for identifying the cell categories. These visualization results clearly demonstrate that our method successfully concentrates attention on the essential characteristics of different types of cells, showing a stronger feature learning abilities over other methods.

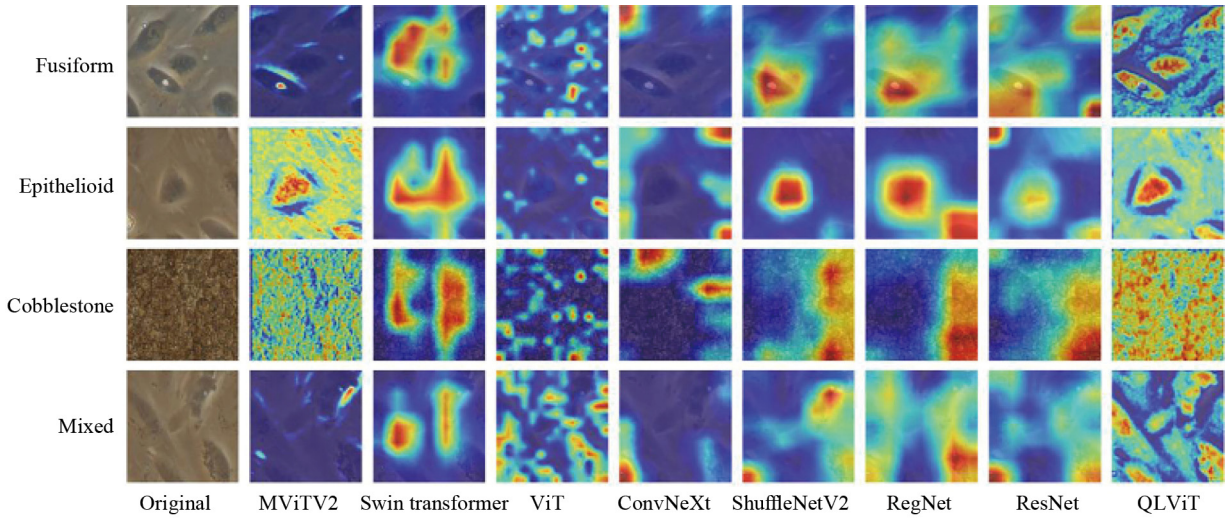

Figure 5. Visualization of feature maps generated by various models on the BioMediTech dataset

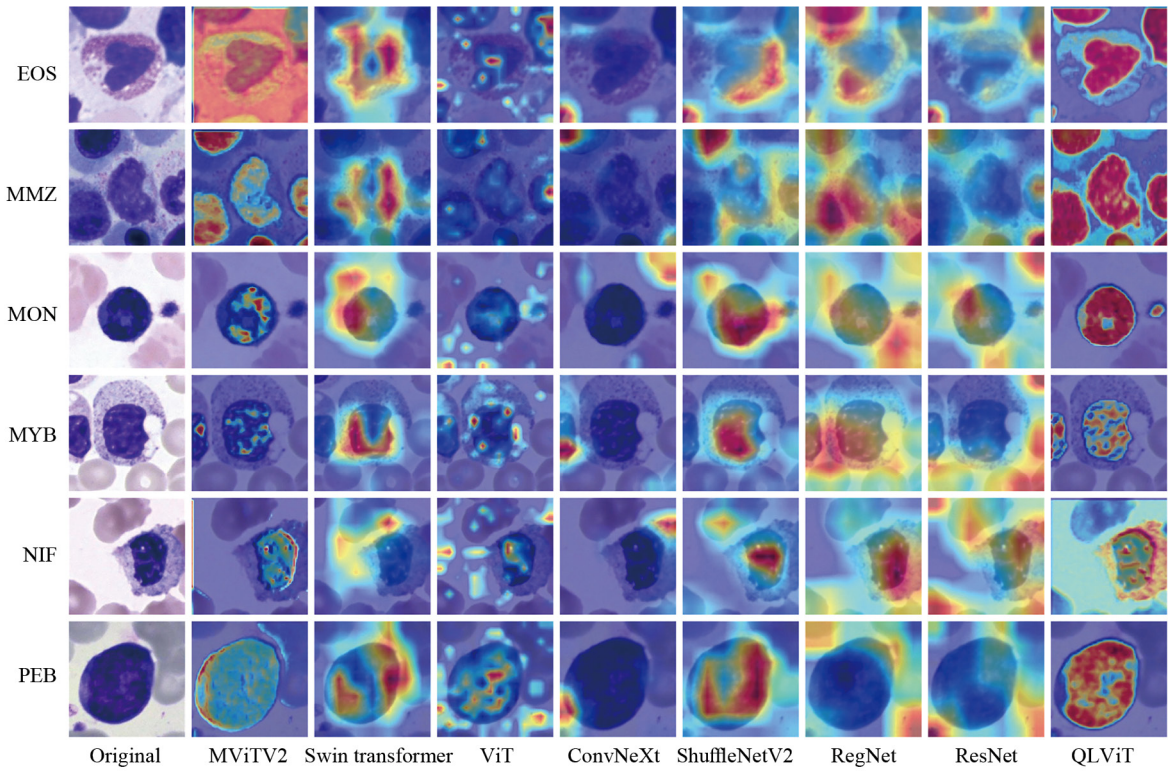

Figure 6. Feature maps of different models on the bone marrow dataset

Activation maps of different models on the bone marrow dataset are shown in Figure 6. It is evident that our method shows the best visualization results regarding the internal feature representations of different types of cells in almost all cases, closely aligning with the key feature regions and excelling in the morphology and boundaries of the nucleus. While except the MViTV2 perform relatively well, other models exhibit blurred boundaries or incorrect feature areas. In summary, these results validate that our method outperforms others in characterizing the classification-related features on the microscope images. It significantly aids in precise classification and identification of cell types, offering higher classification and accuracy and more reliable results.

4.4 *Ablation study*

4.4.1 *Ablation experiment setup*

We performed ablation studies on the BioMediTech dataset to verify the effectiveness of each component in the proposed method. To this end, five versions from QLViT_V0 to QLViT_V4 were constructed, and each model is an improvement of the previous version. All experiments are trained for 500 epochs.

QLViT_V0: Compared to the baseline MViTv2, the V0 version incorporates a convolutional layer and feature processing operations before the standard multi-head self-attention feature mapping.

QLViT_V1: The quantization operation is utilized in the activation functions within the multi-head self-attention mechanism and a dynamic convolution is employed to replace the conventional convolution of the mapping layer.

QLViT_V2: The quantization operation is used on the linear layers and the activation functions except for the core attention module.

QLViT_V3: The QKVs in the self-attention mechanism are processed with the activation function.

QLViT_V4: The KAN structure is adopted to supersede the MLP at the end of the CLA module.

Table 3. Comparison of FLOPs and number of parameters for five versions of the model using the same vectors

Model	Acc (%)	FLOPs	Params
MViTv2 [28]	90.49	4.00 G	23.33 M
QLViT_V0	92.05	4.3 G	24.78 M
QLViT_V1	92.33	4.00 G	23.33 M
QLViT_V2	92.61	4.29 G	25.11 M
QLViT_V3	94.89	4.27 G	25.09 M
QLViT_V4	97.19	1.95 G	9.07 M

To measure the number of parameters and computational complexity of each model, the input vector of the same size $3 \times 224 \times 224$ was fed into each model to record the FLOPs and number of parameters. When the model depth (the repetitions of the backbone CLA module) is 10 layers, the corresponding FLOPs and the parameter quantity are listed in Table 3.

Table 4. Training and inference time of different variant

Models	Training time	Inference time
V0	1 h 23 min 15 sec	15.12 ms
V1	2 h 26 min 25 sec	14.29 ms
V2	1 h 45 min 40 sec	12.25 ms
V3	1 h 17 min 7 sec	11.39 ms
V4	2 h 20 min 56 sec	9.43 ms

The training and inference times for different variants are presented in Table 4. Each variant was evaluated by passing a $3 \times 224 \times 224$ tensor and measuring the full inference duration. The evolution in training time from QLViT_V0 to QLViT_V4 reflects the architectural modifications in each version. QLViT_V1 introduces quantization and dynamic convolution operations. Quantization involves an additional discretization step, while dynamic convolution requires computing input-dependent kernel weights, collectively increasing training time to 2 hours 26 minutes 25 seconds. In QLViT_V2, quantization was extended to linear layers and activation functions. Although quantizing linear layers raised computational complexity, optimizations such as quantization-aware training helped curb this increase, reducing training time to 1 hour 45 minutes 40 seconds. QLViT_V3 applies activation function processing only to the QKV components in the self-attention mechanism. Since activation functions are computationally inexpensive and efficiently implemented, their impact on training time was marginal, leading to a further reduction to 1 hour 17 minutes 7 seconds. Finally, QLViT_V4 replaces the final MLP in the CLA module with a KAN structure, which introduces richer feature interactions and nonlinear transformations. This change increased training time to 2 hours 20 minutes 56 seconds. Overall, the incorporation of complex operations or modules tends to increase training time, whereas structural simplification and optimization can reduce it. Notably, inference time is often more critical in practical applications. Our final model, QLViT_V4, achieves the shortest inference time of only 9.43 milliseconds.

4.4.2 *Results and analysis*

The ablation study results are shown in Table 3. We can see that our ultimate QLViT model achieves the best results in all evaluation metrics. With the addition of convolutional layers, the accuracy of QLViT_V0 increases by 2% over the baseline model. However, as indicated in Table 3, both FLOPs and the number of parameters have a slight increase. There is a further improvement of accuracy with QLViT_V1 by 0.3% compared to QLViT_V0 because of the involvement of the quantization operation and dynamic convolution, while identical FLOPs and parameters with MViTv2 are obtained. QLViT_V2 and QLViT_V3 achieve additional 0.3% and 2.5% increase in accuracy over QLViT_V1 because of the utilization of the quantization operation on the entire model and an improved self-attention mechanism, respectively, at an expense of computational complexity. Ultimately, our method QLViT_V4 further aggregates the KAN structure and achieves the highest accuracy of 97.19%, recall of 97.19% and F1-score of 97.19%, significantly outperforming other configurations. Particularly, the smallest FLOPs of 1.95G and number of parameters of 9.07M are obtained. This further demonstrates the superior lightweight characteristics and outstanding performance of our proposed method.

As shown in Figures 7 and 8, the Grad-CAM visualizations generated by KAN-based QLViT_V4 exhibit clearer and more prominent features compared to those from MLP-based QLViT_V3. This advantage can be largely attributed to a key characteristic of KAN by deploying learnable activation functions on the edges of the network rather than fixed node activation functions. KAN gains greater flexibility in capturing complex relationships within input data, thereby achieving more accurate and robust feature extraction. Furthermore, KAN's design enables the model to achieve comparable or superior accuracy with fewer parameters. These results demonstrate KAN's potential to boost both model performance and interpretability in deep learning models.

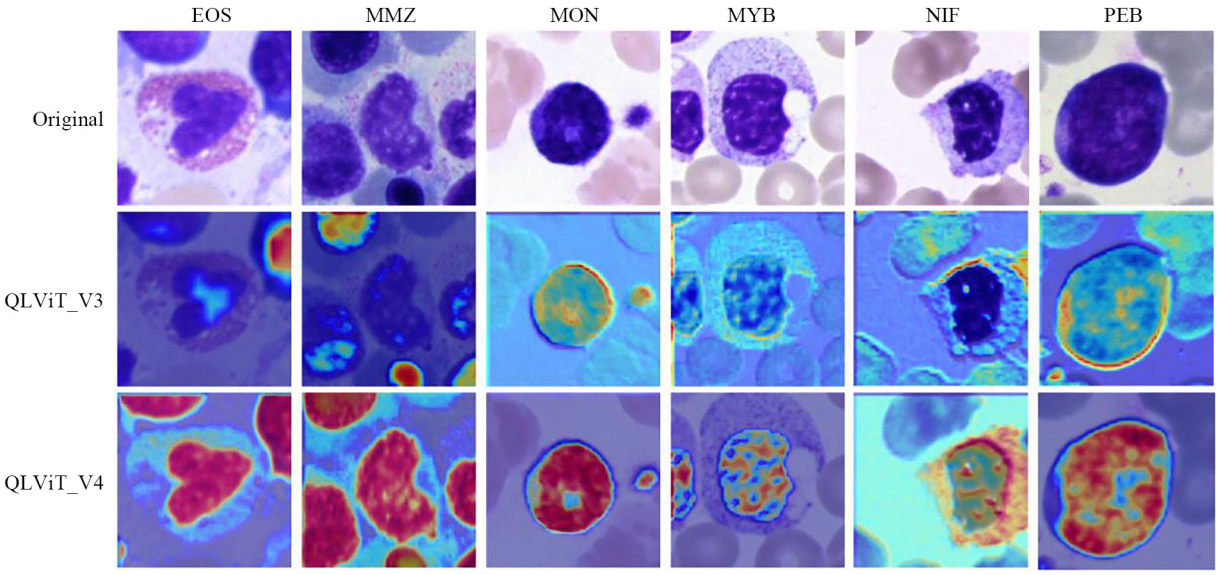

Figure 7. Grad-CAM visualizations for QLViT_V3 and QLViT_V4 on the BioMediTech dataset

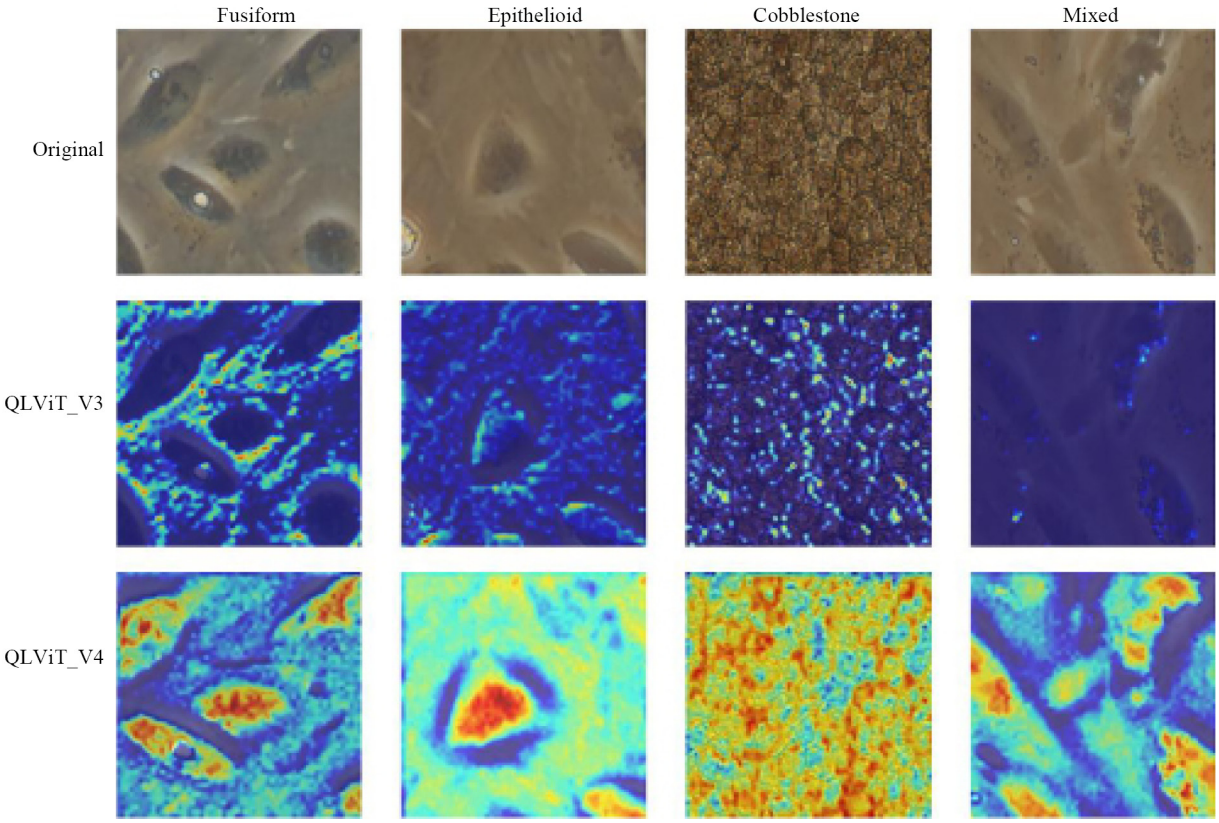

Figure 8. Grad-CAM visualizations for QLViT_V3 and QLViT_V4 on the BioMediTech dataset

5. Discussion

This work presents the QLViT method for cell image classification under the microscope. The method combines large kernel convolutions, a well-designed CLA module, quantized activation functions, and linear layers. Large kernel convolutions are utilized to capture local information extensively in images, thereby enhancing the model's feature extraction capabilities. By integrating dynamic convolution, a linear attention mechanism, and the KAN structure, the CLA module enables the model to achieve better feature representation while maintaining lower computational complexity.

The dynamic convolution allows the convolution kernel weight distribution to be adjusted based on feature information, enabling more comprehensive feature learning. The advantages of the improved linear attention mechanism lie in the inclusion of activation functions during the QKV calculations, which reorders computations compared to traditional self-attention methods. This adjustment brings two key benefits, similar to gating mechanisms: it better controls the information flow and captures global information dependencies, mitigating gradient issues during training, and it selectively retains information relevant to the task at hand, improving the model's memory capacity and predictive performance. Replacing the traditional MLP with the KAN structure further enhances the model's nonlinear learning capability and interpretability. Experimental results (see Table 3) also show a significant reduction in computational load and a decrease in the number of parameters.

Class activation maps from the BioMediTech and bone marrow cell datasets indicate that our method is more sensitive to the microscopic cell images than other the state-of-the-art classification models. It focuses on key areas of each cell type. Based on the comprehensive experimental results on four datasets, it proves that our is capable of handling various cell classification tasks with strong performance and extreme lightweight features.

Due to the variabilities of the employed dataset and different validation procedures, it is difficult to make quantitative comparison with related work. Nevertheless, we believe it is still important to attempt a relative comparison. To this end, we identified several representative literatures. Manju et al. [41] proposes a hybrid method combining CNN, Gray Level Co-occurrence Matrix (GLCM), and Discrete Cosine Transform (DCT) for HEp-2 cell image classification, highlighting the importance of multi-dimensional feature fusion, achieving 96.56% accuracy on the ICPR-HEp-2 dataset. Our QLViT method integrates the KAN architecture and self-attention optimization, yielding a 0.6% increase in accuracy on the same dataset. Pandiraj et al. [42] develops a domain-specific graph-based model for blood cell classification, achieving 99.13% accuracy on the White Blood Cell dataset. Although their innovative approach excels in scenarios with clear morphological patterns, it faces scalability issues in multi-domain images. The utilization of dynamic convolution and self-attention mechanism in our model addresses these limitations, enabling the accuracy rise by 0.71%. Chen et al. [43] develops a ViT-based bone marrow cell classification model that assembles the SCConv [44] and KAN structure, mitigating the problem of local and global feature loss. However, different dataset is adopted in contrast with our model and the large parameter size limits its application. Our proposed method balances model compression and performance by using lightweight strategies on a more compact architecture. Overall, our approach outperforms existing methods for cell classification tasks with strong generalization and lightweight characteristics.

Although our method achieves satisfactory results on four public microscopy cell image classification datasets, future work could expand to include more diverse datasets or other type of classification tasks to further validate its generalizability.

6. Conclusion

In conclusion, this study presents QLViT, a lightweight cell classification method based on the MViTv2 architecture and linear attention mechanism for microscope images. By incorporating the CLA feature extraction module, it achieves superior feature representation with reduced computational complexity. Comprehensive experimental evaluations on four publicly available cell classification datasets validate QLViT's robustness and efficacy, demonstrating its potential for large-scale biological image analysis and broader classification tasks in other domains.

Acknowledgement

This research was funded in part by the National Natural Science Foundation of China under Grant Nos. 61902282 and 62071330.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Rahman S, Wang L, Sun C, Zhou L. Deep learning based HEp-2 image classification: A comprehensive review. *Medical Image Analysis*. 2020; 65: 101764. Available from: https://doi.org/10.1016/j.media.2020.101764.

[2] Wang Z, Luo Y, Xin J, Zhang H, Qu L, Wang Z, et al. Computer-aided diagnosis based on extreme learning machine: A review. *IEEE Access*. 2020; 8: 141657-141673. Available from: https://doi.org/10.1109/ACCESS.2020.3012093.

[3] Nahid AA, Kong Y. Involvement of machine learning for breast cancer image classification: A survey. *Computational and Mathematical Methods in Medicine*. 2017; 2017(1): 3781951. Available from: https://doi.org/10.1155/2017/3781951.

[4] Meng N, Lam EY, Tsia KK, So HKH. Large-scale multi-class image-based cell classification with deep learning. *IEEE Journal of Biomedical and Health Informatics*. 2018; 23(5): 2091-2098. Available from: https://doi.org/10.1109/JBHI.2018.2878878.

[5] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*. 2021; 8: 1-74. Available from: https://doi.org/10.1186/s40537-021-00444-8.

[6] Kutlu H, Avci E, Özyurt F. White blood cells detection and classification based on regional convolutional neural networks. *Medical Hypotheses*. 2020; 135: 109472. Available from: https://doi.org/10.1016/j.mehy.2019.109472.

[7] Shafique S, Tehsin S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technology in Cancer Research & Treatment*. 2018; 17: 1533033818802789. Available from: https://doi.org/10.1177/1533033818802789.

[8] Song TH, Sanchez V, Daly HE, Rajpoot NM. Simultaneous cell detection and classification in bone marrow histology images. *IEEE Journal of Biomedical and Health Informatics*. 2018; 23(4): 1469-1476. Available from: https://doi.org/10.1109/JBHI.2018.2878945.

[9] Ashish V. Attention is all you need. Advances in neural information processing systems. *arXiv:1706.03762*. 2017. Available from: https://doi.org/10.48550/arXiv.1706.03762.

[10] Zedda L, Loddo A, Di Ruberto C. SAMMI: Segment anything model for malaria identification. In: *19th International Conference on Computer Vision Theory and Applications*. Rome, Italy: Institute for Systems and Technologies of Information, Control and Communication; 2024. p.367-374. Available from: https://doi.org/10.5220/0012325500003660.

[11] Kim A, Kim R. Analysis of modern computer vision models for blood cell classification. *arXiv:240700759*. 2024. Available from: https://doi.org/10.48550/arXiv.2407.00759.

[12] Halder A, Gharami S, Sadhu P, Singh PK, Woźniak M, Ijaz MF. Implementing vision transformer for classifying 2D biomedical images. *Scientific Reports*. 2024; 14(1): 12567. Available from: https://doi.org/10.1038/s41598-024-63094-9.

[13] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:201011929*. 2020. Available from: https://doi.org/10.48550/arXiv.2010.11929.

[14] Schmidt-Hieber J. The Kolmogorov-Arnold representation theorem revisited. *Neural Networks*. 2021; 137: 119-126. Available from: https://doi.org/10.1016/j.neunet.2021.01.020.

[15] Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, et al. Kan: Kolmogorov-arnold networks. *arXiv:240419756*. 2024. Available from: https://doi.org/10.48550/arXiv.2404.19756.

[16] De Faria LC, Rodrigues LF, Mari JF. Cell classification using handcrafted features and bag of visual words. In: *Proceedings of the 14th Workshop on Computer Vision*. Brazil: Brazilian Computer Society, Workshop on Computer Vision; 2018. p.68-73.

[17] Wiliem A, Wong Y, Sanderson C, Hobson P, Chen S, Lovell BC. Classification of human epithelial type 2 cell indirect immunofluoresence images via codebook based descriptors. In: *2013 IEEE Workshop on Applications of Computer Vision*. Clearwater Beach, FL, USA: IEEE; 2013. p.95-102. Available from: https://doi.org/10.1109/WACV.2013.6475005.

[18] Dhall D, Kaur R, Juneja M. Machine learning: A review of the algorithms and its applications. Proceedings of ICRIC 2019: Recent innovations in computing. 2020; 597: 47-63. Available from: https://doi.org/10.1007/978-3-030-29407-6_5.

[19] Asghar R, Kumar S, Shaukat A, Hynds P. Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review. *Plos One*. 2024; 19(6): e0292026. Available from: https://doi.org/10.1371/journal.pone.0292026.

[20] Dev A, Fouda MM, Kerby L, Fadlullah ZM. Advancing malaria identification from microscopic blood smears using hybrid deep learning frameworks. *IEEE Access*. 2024; 12: 71705-71715. Available from: https://doi.org/10.1109/ACCESS.2024.3402442.

[21] Huang Q, Li W, Zhang B, Li Q, Tao R, Lovell NH. Blood cell classification based on hyperspectral imaging with modulated Gabor and CNN. *IEEE Journal of Biomedical and Health Informatics*. 2019; 24(1): 160-170. Available from: https://doi.org/10.1109/JBHI.2019.2905623.

[22] Li Y, Wu CY, Fan H, Mangalam K, Xiong B, Malik J, et al. Mvitv2: Improved multiscale vision transformers for classification and detection. *arXiv:2112.01526*. 2022. p.4804-4814. Available from: https://doi.org/10.48550/arXiv.2112.01526.

[23] Zhang T, Li L, Zhou Y, Liu W, Qian C, Hwang JN, et al. Cas-vit: Convolutional additive self-attention vision transformers for efficient mobile applications. *arXiv:240803703*. 2024. Available from: https://doi.org/10.48550/arXiv.2408.03703.

[24] Zhu G, Lin G, Yang X, Zeng C. Flat U-Net: An Efficient ultralightweight model for solar filament segmentation in full-disk Hα images. *arXiv:250207259*. 2025. Available from: https://doi.org/10.48550/arXiv.2502.07259.

[25] Zhu J, Chen X, He K, LeCun Y, Liu Z. Transformers without normalization. *arXiv:2503.10622*. 2025. Available from: https://doi.org/10.48550/arXiv.2503.10622.

[26] Chen Y, Dai X, Liu M, Chen D, Yuan L, Liu Z. Dynamic convolution: Attention over convolution kernels. *arXiv:1912.03458*. 2020. Available from: https://doi.org/10.48550/arXiv.1912.03458.

[27] Kim S, Gholami A, Yao Z, Mahoney MW, Keutzer K. I-bert: Integer-only bert quantization. *arXiv:2101.01321*. 2021. Available from: https://doi.org/10.48550/arXiv.2101.01321.

[28] Wu H, Wu J, Xu J, Wang J, Long M. Flowformer: Linearizing transformers with conservation flows. *arXiv:220206258*. 2022. Available from: https://doi.org/10.48550/arXiv.2202.06258.

[29] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE; 2016. p.770-778. Available from: https://doi.org/10.1109/CVPR.2016.90.

[30] Cai J, Takemoto M, Nakajo H. A deep look into logarithmic quantization of model parameters in neural networks. In: *IAIT '18: Proceedings of the 10th International Conference on Advances in Information Technology*. Thailand: Association for Computing Machinery; 2018. p.1-8. Available from: https://doi.org/10.1145/3291280.3291800.

[31] Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE; 2018. p.2704-2713. Available from: https://doi.org/10.1109/CVPR.2018.00286.

[32] Nanni L, Paci M, Caetano dos Santos FL, Skottman H, Juuti-Uusitalo K, Hyttinen J. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLos One*. 2016; 11(2): e0149399. Available from: https://doi.org/10.1371/journal.pone.0149399.

[33] Mooney P. *BCCD-Blood Cell Classification Dataset*. Available from: https://www.kaggle.com/datasets/paultimothymooney/blood-cells [Accessed 1st May 2024].

[34] Percannella G, Foggia P, Soda P. *ICPR 2013 Dataset*. Available from: https://www.heywhale.com/mw/dataset/5ec3c6883241a100378d5d4a [Accessed 1st May 2024].

[35] Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood*. 2021; 138(20): 1917-1927. Available from: https://doi.org/10.1182/blood.2020010568.

[36] Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P. Designing network design spaces. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE; 2020. p.10428-10436. Available from: https://doi.org/10.1109/CVPR42600.2020.01044.

[37] Ma N, Zhang X, Zheng HT, Sun J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. (eds.) *Computer Vision-ECCV 2018*. 2018. p.122-138. Available from: https://doi.org/10.1007/978-3-030-01264-9_8.

[38] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*. 2021. Available from: https://doi.org/10.48550/arXiv.2103.14030.

[39] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA: IEEE; 2022. Available from: https://doi.org/10.1109/CVPR52688.2022.01167.

[40] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE; 2017. p.618-626. Available from: https://doi.org/10.1109/ICCV.2017.74.

[41] Manju CC, Victor Jose M. HEP-2 specimen image segmentation and classification using GLCM and DCT based feature extraction with CNN classifier. *Advances in Communication Systems and Networks*. 2020; 656: 147-159. Available from: https://doi.org/10.1007/978-981-15-3992-3_12.

[42] Pandiraj J, Balasubramanian VK. Domain knowledge-based deterministic graph traversal method for white blood cell classification. *Machine Learning: Science and Technology*. 2025; 6(1): 015037. Available from: https://doi.org/10.1088/2632-2153/adb126.

[43] Chen Y, Zhu Z, Zhu SH, Qiu LW, Zou BF, Jia F. Sckansformer: Fine-grained classification of bone marrow cells via kansformer backbone and hierarchical attention mechanisms. *IEEE Journal of Biomedical and Health Informatics*. 2024; 29(1): 558-571. Available from: https://doi.org/10.1109/JBHI.2024.3471928.

[44] Li J, Wen Y, He L. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, BC, Canada: IEEE; 2023. Available from: https://doi.org/10.1109/CVPR52729.2023.00596.