


Article

Shifted Window Vision Transformer for Blood Cell Classification

Shuwen Chen ¹, Siyuan Lu ², Shuihua Wang ^{2,3,*}, Yiyang Ni ¹ and Yudong Zhang ^{2,3,*} 

¹ School of Physics and Information Engineering, Jiangsu Second Normal University, Nanjing 210013, China; chenshuwen@jssnu.edu.cn (S.C.); niyy@jssnu.edu.cn (Y.N.)

² School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK; siyuan_lu@foxmail.com (S.L.); shuihuawang@ieee.org (S.W.)

³ Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

* Correspondence: yudongzhang@ieee.org; Tel.: +44-754-870-0453

Abstract: Blood cells play an important role in the metabolism of the human body, and the status of blood cells can be used for clinical diagnoses, such as the ratio of different blood cells. Therefore, blood cell classification is a primary task, which requires much time for manual analysis. The recent advances in computer vision can be beneficial to free doctors from tedious tasks. In this paper, a novel automated blood cell classification model based on the shifted window vision transformer (SW-ViT) is proposed. The SW-ViT architecture is firstly pre-trained on the ImageNet dataset and fine-tuned on the blood cell images for classification. Two transfer strategies are employed to generate better classification results. One is to fine-tune the entire SW-ViT, and the other is to only fine-tune the linear output layer of the SW-ViT while all the other parameters are frozen. A public dataset named BCCD_Dataset (Blood Cell Count and Detection) is utilized in the experiments. The results show that the SW-ViT outperforms several state-of-the-art methods in terms of classification accuracy. The proposed SW-ViT can be applied in daily clinical diagnosis.

Keywords: blood cell; computer-aided diagnosis; computer vision; deep learning; vision transformer



Citation: Chen, S.; Lu, S.; Wang, S.; Ni, Y.; Zhang, Y. Shifted Window Vision Transformer for Blood Cell Classification. *Electronics* **2023**, *12*, 2442. <https://doi.org/10.3390/electronics12112442>

Academic Editor: Hyunjin Park

Received: 9 May 2023

Revised: 24 May 2023

Accepted: 26 May 2023

Published: 28 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Blood flows in the vessels everywhere in human bodies. As an essential component, blood can transport nutrients and oxygen to tissues and carry waste out to keep them clean. Blood also helps to maintain the temperature of the human body. Meanwhile, blood is an important part of the immune system, which fights against infections. Mainly, blood is composed of red blood cells, white blood cells, platelets, and plasma. The complete blood count test is a regular method in clinical diagnoses, which classifies the types of blood cells and indicates the percentages they account for. The manual classification of blood cells suffers from low reproducibility, as naked eyes can become fragile when working overtime. Fortunately, blood cell classification can be implemented automatically and fast with the recent advanced computer vision techniques, and computer-aided diagnosis (CAD) has been a hot topic [1]. Once a blood cell classification model is trained, it can generate the same types of blood cell images without human intervention. Over the last decade, practitioners have proposed many CAD methods for blood cell classification.

Acevedo, et al. [2] proposed a deep learning-based method for peripheral blood cell recognition. The pre-trained VGG-16 and InceptionV3 were employed as the backbones to extract features from the peripheral blood cell images without fine-tuning. Then, those features were used to train a support vector machine (SVM) for recognition. They also directly fine-tuned the pre-trained VGG-16 and InceptionV3 on the peripheral blood cell image dataset for comparison. The fine-tuned VGG-16 achieved the best accuracy of 96% in their experiments. Gupta, et al. [3] developed a white blood cell classification system based on traditional machine learning algorithms. Image pre-processing techniques were utilized including thresholding, resizing, and cropping. Then, shape and texture features were

extracted from the cell images. An optimized binary bat algorithm was proposed to remove the redundant features from the entire feature set. Four classical machine learning models were trained as the classifiers including k nearest neighbors, logistic regression, decision tree, and random forest. Their system was trained and evaluated on a dataset with 237 samples and produced satisfactory results. In [4], the researchers put forward a white blood cell classification method based on blood smear images. The original images were segmented to generate regions of interest, and shape, color, and texture features were extracted from the regions of interest. An artificial neural network with three fully-connected layers and an autoencoder were trained for classification. They also leveraged the AlexNet and trained it for classification by both transfer learning and training from scratch. A dataset with six types of images was used in their experiments. Abdulkarim, et al. [5] transferred a pre-trained AlexNet on the blood smear image dataset, which contained fifteen types of images. Their model could classify red blood cells accurately to assist in the diagnosis of sickle cell anemia. Abou El-Seoud, et al. [6] designed a convolutional neural network (CNN) for white blood cell classification. The CNN model contained four convolution layers and was trained using the dropout strategy. There are over ten thousand samples of four types of white blood cells in their dataset in total, but only just over one hundred samples were used for testing. Banik, et al. [7] firstly segmented the blood smear images to obtain the white blood cell nucleus based on k -means and the filtering. Then, a deep CNN model was developed with residual connections to concatenate the feature maps of shallow layers with those of deep layers, which aimed to reduce information loss. The average accuracy of their model was 96% on the public dataset. Baydilli, et al. [8] attempted to develop a white blood cell classification model with domain adaptation. When the CNN model was trained on the source domain, the AutoAugment method and generative adversarial network (GAN) were utilized to improve the representation learning ability of the CNN. Therefore, the CNN fine-tuned on the target domain can generate domain-invariant features and achieve better results for white blood cell classification. Kutlu, et al. [9] employed the regional CNN (R-CNN) for the classification of blood cells. They trained four different CNN models as the backbones of the R-CNN, including AlexNet, VGG-16, GoogLeNet, and ResNet-50. Their model could produce the types of cells and the bounding boxes of the cells simultaneously. Loey, et al. [10] leveraged the pre-trained AlexNet to detect leukemia based on blood cell classification. They used the pre-trained AlexNet as the feature extractor, and the extracted image features were fed into the traditional machine learning models for classification. In comparison, the pre-trained AlexNet was fine-tuned on the blood cell images. Experiment results showed that the fine-tuned AlexNet achieved the best classification performance. Ridoy and Islam [11] presented a lightweight CNN model for white blood cell classification, which consisted of five convolution layers. Batch normalization layers were embedded to handle the covariance shifting problem and accelerate the convergence when training. Their model was trained for thirty epochs on the BCCD_Dataset (Blood Cell Count and Detection), and the classification performance was promising. Sahlol, et al. [12] used a pre-trained VGG-19 as the backbone of their model to extract image features. An enhanced salp swarm algorithm (ESSA) was proposed to select the most important features for classification. Finally, a decision tree was trained with the refined image features. Settouti, et al. [13] put forward a white blood cell segmentation algorithm. Initially, pixel-level classification was implemented to locate the positions of white blood cells based on the color features. Then, a region-growing algorithm was employed to generate the segmentation results. Chen, et al. [14] employed two pre-trained CNN models and fused their feature maps by addition. Then, a squeeze and excitation module was embedded in their model to enhance the representation learning. A residual connection was added between the start and the end of the squeeze and excitation module. Four different blood cell image datasets were employed to evaluate the proposed model. Çınar and Tuncer [15] developed a hybrid system to detect white blood cells. The pre-trained AlexNet and GoogLeNet were used as the feature extractors, and the two feature sets were concatenated to form the ultimate features. The padding strategy was employed to the dimension difference between the

two feature sets. An SVM served as the classification model. Dincic, et al. [16] proposed a blood cell recognition method based on feature engineering classical machine learning algorithms. The features were extracted from the microscopic images based on color deconvolution, morphological operations, fractal analyses, and gray-level co-occurrence matrix. Afterwards, an SVM was trained for classification. Liao, et al. [17] proposed a red blood cell classification approach based on empirical wavelet transform and SVM. Their method was trained and tested on an ultrasonic dataset. Semerjian, et al. [18] segmented the nucleus from the blood smear images and developed a lightweighted CNN model for classification. Yao, et al. [19] developed a blood cell classification system using two CNN models. The first CNN was trained with high-quality images, and the weights were transferred to the second CNN. Then, deformable convolution layers were added to the second model, and low-quality images were used to train the second CNN. Zhu, et al. [20] used a ResNet-18 as their backbone to generate high-level features from the blood cell images. They ensembled three randomized neural networks as the classifiers of their model. The average accuracy of the proposed BCNet was 96.78% for triple-class classification. Ichim, et al. [21] employed four CNN models for blood cell image classification, including VGG-16, Xception, ResNet-50, and NasNet-Large. A fusion method was proposed based on the validation performance of the four CNN models to obtain the ultimate outcome, which worked with a weighting mechanism. Zhu, et al. [22] proposed a deep CNN model for malaria parasite classification in blood smear images. Their model was verified on a dataset with nearly thirty thousand images. Elhassan, et al. [23] presented an abnormal white blood cell detection method based on deep models. A convolutional autoencoder was trained to generate more training images. Then, the image features were extracted using the latent feature layer of the autoencoder, which were used to train two CNN models for classification. The first CNN classified the samples as normal white blood cells or abnormal ones, while the second CNN classified the abnormal samples into eight different subtypes. Bayat, et al. [24] designed an attention-based CNN with regularization techniques to classify white blood cells. The model could fuse texture features with global features from the blood cell images. Cheuque, et al. [25] firstly used a faster R-CNN to generate a region of interest from the blood cell images. A MobileNet was employed to recognize the sub-classes of the segmented blood cell images. Sharma, et al. [26] transferred a DenseNet-121 for blood cell classification, which was fine-tuned for 10 epochs.

The above works either used traditional handcrafted features with machine learning algorithms or leveraged deep CNN models for classification. However, traditional machine learning algorithms often fail on large datasets, and CNN models have strong inductive biases, such as locality and translational equivalence. Locality assumes that neighboring regions may share similar characteristics, which works with the sliding kernel windows. Translational equivalence means the convolution results for an object can be the same even if the position of the object is changed. The strong inductive biases of CNN models enable them to work well with less training data. Nevertheless, more information is lost with stronger inductive biases during the data flow of CNN models. Recently, transformers have been applied in computer vision tasks, which were proposed for natural language processing. The core idea behind transformers is the attention mechanism [27]. Different from convolution operations, all the patches in an image can interact with each other using an attention mechanism, which preserves more information for classification and segmentation. Therefore, transformers have more potential though more computation cost is required compared with CNN models.

The contributions of this paper are summarized below:

1. A blood cell classification system is presented based on the microscopic blood cell images and a vision transformer.
2. The shifted window vision transformer (SW-ViT) is developed to reduce the computational cost of conventional vision transformers. The SW-ViT was pre-trained on the ImageNet dataset and fine-tuned using two strategies on a public dataset.

3. The performance of the SW-ViT with CNN-based models as well as state-of-the-art methods is compared, and the proposed SW-ViT showed superiority.

The rest of this paper is arranged as follows. The description of the blood cell image dataset is given in Section 2. The details of the proposed SW-ViT are discussed in Section 3. Section 4 is about the simulation results and discussion. Finally, the conclusion and future plan are described in Section 5.

2. Materials

A public blood cell image dataset named BCCD_Dataset (Blood Cell Count and Detection) was used in the verification experiments of the proposed SW-ViT, which is available on the Kaggle website (<https://www.kaggle.com/datasets/paultimothymooney/blood-cells> (accessed on 30 January 2023)). There are four different types of blood cell images in BCCD_Dataset, including eosinophil, lymphocyte, monocyte, and neutrophil, and the numbers of the samples are 3120, 3103, 3098, and 3123, respectively. The size of the blood cell images is 320×240 pixels. Some samples are listed in Figure 1. As the BCCD_Dataset contains sufficient samples, hold-out validation is used instead of k-fold cross-validation. In the experiments, 80% of the images are used for training and the other 20% of the images serve as the testing set.

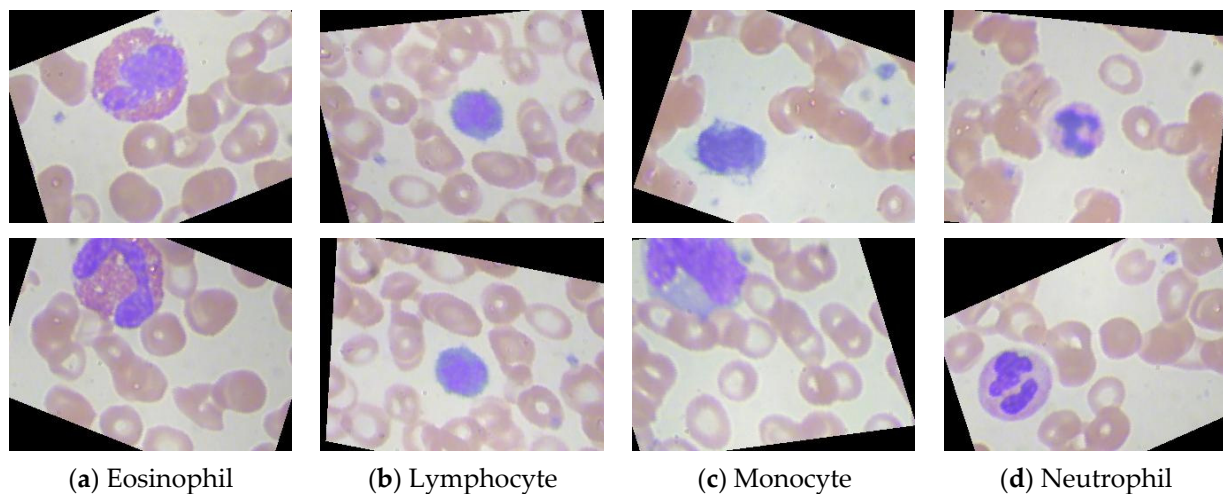


Figure 1. Samples in the BCCD_Dataset (every column represents one type of blood cell image).

3. Methods

Attention-based models, termed transformers, were originally proposed for natural language processing, which achieved great success [27]. Compared with images, the information density of words is much higher because they are human-generated signals. Additionally, the mask tokens and positional encodings in transformers are not naturally compatible with convolution operations. Therefore, CNN models were more preferred over transformers in computer vision until the advent of the vision transformer (ViT) [28]. ViT achieved a better performance than CNN models on a group of benchmark datasets, which showed its potential for image classification and segmentation. Though the computational cost is higher than CNN, it can be foreseen that transformers can provide a unified learning framework for both computer vision and natural language processing. In this study, a novel SW-ViT was employed for the classification of blood cell images.

3.1. SW-ViT

The training and inferring of the ViT are time-consuming because of the self-attention operations. In a ViT, an input image is divided into a set of patches to generate patch embeddings, and each embedding interacts with all the other embeddings using self-attention

operations. Inspired by CNN models, SW-ViT is proposed to improve the efficiency of the original ViT by hierarchical feature learning [29].

In an SW-ViT, an input image is divided into a set of non-overlapping patches, and a patch embedding is generated from each patch, which can be implemented by convolution. Then, a window is used to group the patches, and the self-attention is only computed within the patches in the same group instead of globally. The self-attention operation can be expressed as

$$\text{SELF_ATTENTION}(K, Q, V) = \text{softmax}(KQ^T)V \quad (1)$$

In which K , Q , and V stand for key, query, and value generated from the patches using convolution operations or a fully-connected layer, respectively. Therefore, the patches in different windows will not be used for self-attention computation even if they are adjacent in the image, which reduces the computational complexity substantially. However, there will be information loss as the patches in different windows cannot interact. To handle this issue, in the next self-attention layer, the windows will shift both horizontally and vertically so that the patches will be grouped differently. This window and shifted window mechanism appear in pairs in the SW-ViT for self-attention computation. The basic progress of the shifted window mechanism is depicted in Figure 2.

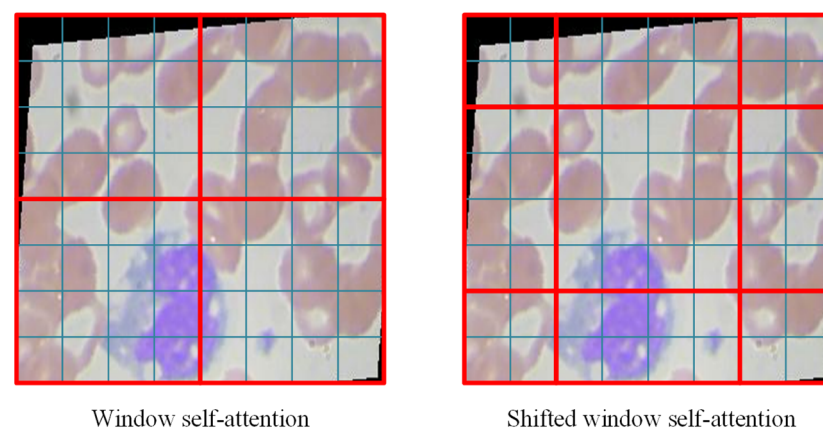


Figure 2. The window and shifted window mechanism. (The small blue boxes represent patches, and the red boxes represent the windows. The window self-attention and shifted window self-attention modules appear in pairs in the SW-ViT).

It can be observed that in the window self-attention computation, the image is divided into four groups of the same size, but there are nine groups of patches in the shifted window. In addition, the shapes of the nine groups are different. The cyclic shift method and masking strategy are employed to calculate the shifted window self-attention efficiently.

The dimension of the patches also changes in the SW-ViT to obtain multi-resolution features. From the input layer to the output layer, the height and width of the patches become halved three times, but the channel of the patches doubles three times. The main modules in the proposed SW-ViT are shown in Figure 3. The patch embedding can be implemented by convolution and layer normalization, and the patch merging can be implemented by layer normalization and linear projection. The dimension of the patches changes in the patch merging layers. The ratio of the four blocks is 1:1:3:1, which can be also seen in [30].

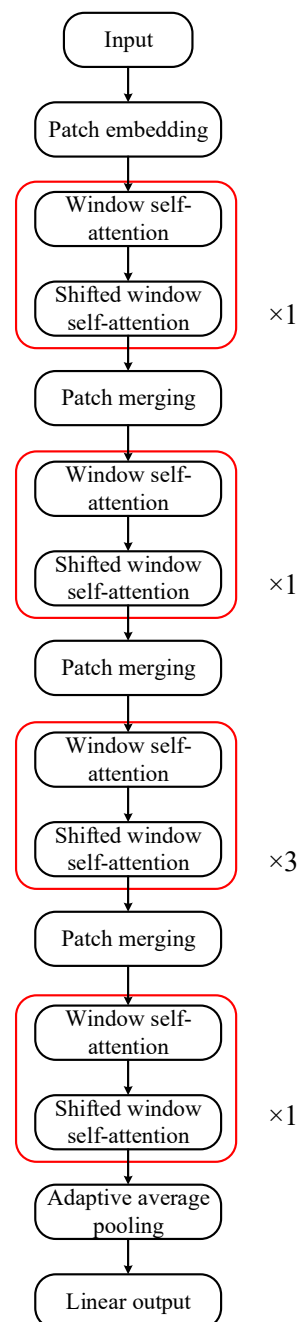


Figure 3. Main modules in the proposed SW-ViT.

3.2. Training Strategies

There are 29 million parameters in the proposed SW-ViT model, which requires large datasets for training, such as the ImageNet-1K dataset. However, the BCCD_Dataset contains only about 12,000 images of four types. Training the SW-ViT on small datasets from scratch may cause the overfitting problem. Meanwhile, it is time-demanding to train deep models even with dedicated graphic cards.

Fortunately, the pre-trained parameters of deep models are often available online for everyone, as researchers from big companies and organizations are willing to share their work. Transfer learning is preferred in downstream tasks when using deep learning models, which often yields satisfactory results. In this study, transfer learning is chosen for training the proposed SW-ViT, as shown in Figure 4.

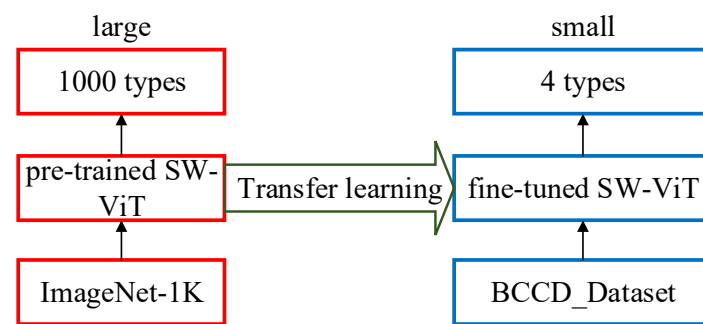


Figure 4. Transfer learning of the SW-ViT.

The parameters in the SW-ViT are pre-trained on the ImageNet-1K dataset, which includes over 12 million images of 1000 different types. The pre-trained SW-ViT can generate complex semantic representations from the images which are beneficial for classification. The types in the ImageNet-1K are common objects, such as animals, fruits, and vehicles, which are obviously different from the blood cell images. However, deep models are data hungry, and the representation learning capability obtained from pre-training is helpful for downstream tasks due to the versatility of big datasets. The images from the ImageNet-1K and BCCD_Dataset look different in the spatial domain, but the distribution patterns could be similar in some latent representation space of the SW-ViT. Another advantage of transfer learning is efficiency. Usually, training a deep model from scratch requires hundreds of epochs. Nevertheless, with pre-trained parameters, the model is likely to converge within dozens of epochs or even several epochs.

Two different strategies are used to fine-tune the pre-trained SW-ViT. The first one is to fine-tune all the parameters in the SW-ViT on the blood cell images. This is helpful for the SW-ViT to learn the distribution patterns of the blood cell images. The second strategy is to only update the parameters in the linear output layer and freeze all the other parameters. This strategy employs the pre-trained SW-ViT as a representation extractor and trains a linear fully-connected layer for classification. Compared with the first strategy, the second method is more time-efficient but usually less accurate. In the experiments, the classification performance of the two training strategies will be compared, and the results from the SW-ViT, which is trained from scratch, will also be compared using the same hyper-parameter setting and platform.

4. Results and Discussion

The implementation of the proposed SW-ViT is based on Python 3.9 with torch 2.0, and all the experimental results are obtained on a personal computer with Intel i9 13900HX CPU and NVIDIA RTX4080 GPU (MECHREVO, Nanjing, China.). The accuracy, precision, recall, and F1-score are employed as the performance metrics for blood cell classification. As there are four different types in the BCCD_Dataset, when computing the precision and recall of a certain type, all the other three types are regarded as negatives.

4.1. Hyper-Parameter Setting

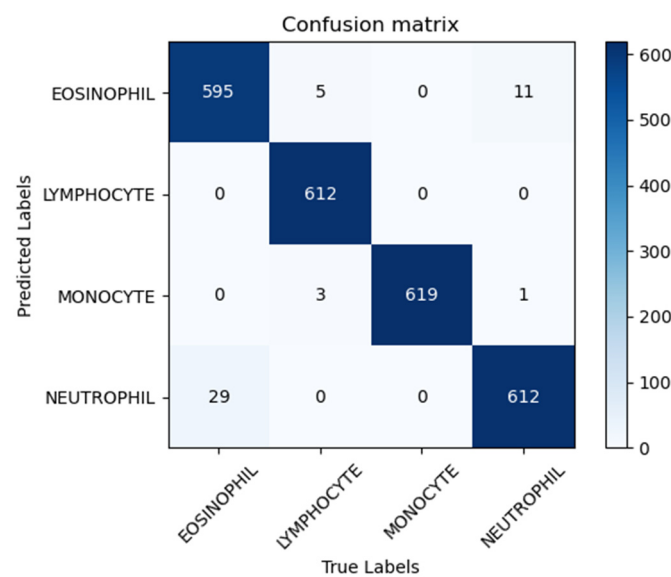
The hyper-parameters to train the SW-ViT are listed in Table 1. The batch size is set as 64 according to the maximum capacity of the 12 GB memory in the GPU. The max epoch value is only 3 in order to avoid overfitting. The optimizer is AdamW, which is the combination of the vanilla Adam and L2 regularization. The values of learning rate and weight decay are 1×10^{-4} and 5×10^{-2} , respectively, which are both common settings.

Table 1. Hyper-parameters.

Hyper-Parameter	Value
Batch size	64
Max epoch	3
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	5×10^{-2}

4.2. Classification Performance of the SW-ViT

The confusion matrix of the SW-ViT is shown in Figure 5, and the performance metrics are listed in Table 2. The classification results for different types are different. The training strategy of the SW-ViT is to fine-tune the entire model. The proposed SW-ViT could classify lymphocyte and monocyte cells nearly perfectly, which yields nearly 100% F1-scores. However, for eosinophil cells, the recall is 95.35%, which is obviously lower, and the precision is at a high level. On the contrary, the precision of neutrophil cell images is merely over 95%, and its recall is above the average performance. In all, the proposed SW-ViT produced an overall accuracy of 98.03% and an F1-score of 98.04% within only three epochs of training, which cost less than 7 min on the platform. The results reveal that the proposed SW-ViT is accurate and efficient for blood cell classification.

**Figure 5.** Confusion matrix of the SW-ViT.**Table 2.** Performance metrics of the SW-ViT.

Cell Type	Precision	Recall	F1-Score	Accuracy
Eosinophil	97.38%	95.35%	96.35%	98.03%
Lymphocyte	100.0%	98.71%	99.35%	
Monocyte	99.36%	100.0%	99.68%	
Neutrophil	95.48%	98.08%	96.76%	
Average	98.06%	98.04%	98.04%	

4.3. Comparison of Different Training Strategies

The performances of three training strategies are presented, including training from scratch, fine-tuning the entire model, and fine-tuning only the linear output layer of the SW-ViT with the same hyper-parameters. The results are demonstrated in Table 3 and Figure 6. It is obvious that the three training strategies produce different results. Fine-tuning only the linear output layer achieves the worst accuracy of 69.44% because the samples from the

ImageNet dataset and BCCD_Dataset are very different. Therefore, the feature learning of the pre-trained SW-ViT cannot produce satisfactory results on the blood cell images. On the other hand, training from scratch achieves better performance with an overall accuracy of 87.29%. However, compared with fine-tuning the entire model, there is still a huge difference of over 10% in average performance, so the parameters in the pre-trained SW-ViT contribute to fast convergence and better accuracy.

Table 3. Results of the proposed SW-ViT with different training strategies.

Training Strategy	Cell Type	Precision	Recall	F1-Score	Accuracy
Training from scratch	Eosinophil	78.46%	94.55%	85.76%	87.29%
	Lymphocyte	87.09%	97.90%	92.18%	
	Monocyte	95.77%	84.17%	89.60%	
	Neutrophil	91.70%	72.60%	81.04%	
	Average	88.26%	87.31%	87.14%	
Fine-tuning the entire model	Eosinophil	97.38%	95.35%	96.35%	98.03%
	Lymphocyte	100.0%	98.71%	99.35%	
	Monocyte	99.36%	100.0%	99.68%	
	Neutrophil	95.48%	98.08%	96.76%	
	Average	98.06%	98.04%	98.04%	
Fine-tuning only the linear output layer	Eosinophil	58.33%	63.94%	61.01%	69.44%
	Lymphocyte	75.77%	83.71%	79.54%	
	Monocyte	83.92%	64.94%	73.22%	
	Neutrophil	63.69%	65.22%	64.45%	
	Average	70.43%	69.45%	69.55%	

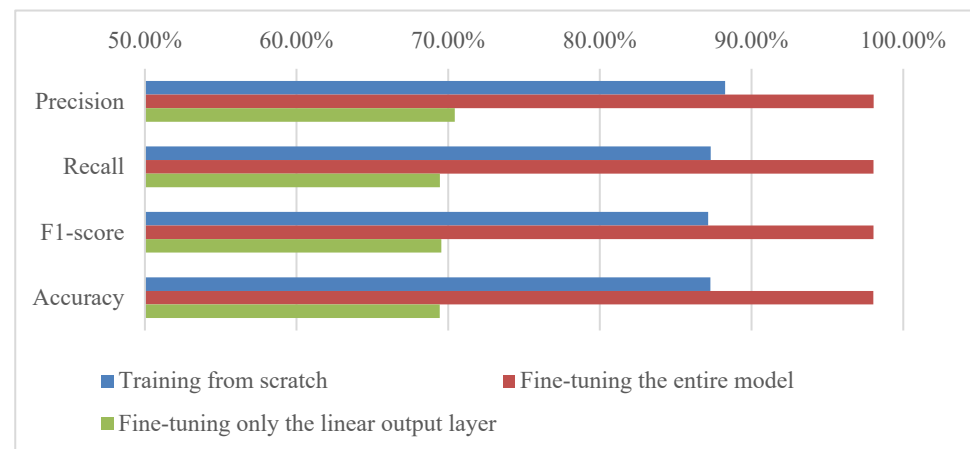


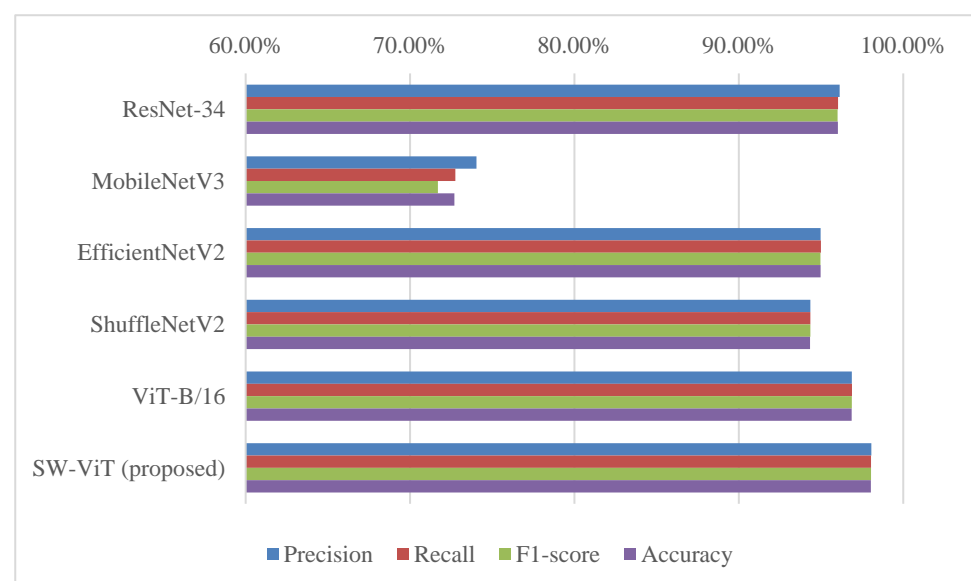
Figure 6. Comparison of different training strategies based on average values.

4.4. Comparison with Other Classification Models

The proposed SW-ViT is compared with other famous image classification models, including ResNet-34, MobileNetV3, EfficientNetV2-B0, ShuffleNetV2, and ViT-B/16. The training strategy involves fine-tuning the entire model, and the models' performances are listed in Table 4 and Figure 7. All the models can achieve over 90% accuracy except for MobileNetV3 which is the most lightweight CNN model among them, and its running time is also the shortest. CNN models can produce satisfactory results, such as ResNet-34, EfficientNetV2-B0, and ShuffleNetV2. The performances of ResNet-34 and ViT-B/16 are close, with accuracies of over 96%, but the running time of ResNet-34 is only one-third of that of ViT-B/16. The proposed SW-ViT achieves the best performance in terms of all four metrics, which reveals that the hierarchical feature learning of the SW-ViT can improve the classification performance of vision transformers. Additionally, the window self-attention in the SW-ViT reduces the running time substantially compared with ViT-B/16.

Table 4. Results of different classification models.

Model	Cell Type	Precision	Recall	F1-Score	Accuracy	Running Time
ResNet-34	Eosinophil	98.42%	89.74%	93.88%	96.02%	204 s
	Lymphocyte	99.84%	100.00%	99.92%		
	Monocyte	95.22%	99.84%	97.48%		
	Neutrophil	91.05%	94.55%	92.77%		
	Average	96.13%	96.03%	96.01%		
MobileNetV3	Eosinophil	83.61%	49.04%	61.82%	72.70%	124 s
	Lymphocyte	69.11%	94.19%	79.72%		
	Monocyte	74.11%	83.68%	78.60%		
	Neutrophil	69.32%	64.10%	66.61%		
	Average	74.04%	72.75%	71.69%		
EfficientNetV2-B0	Eosinophil	90.24%	91.83%	91.03%	94.97%	368 s
	Lymphocyte	99.84%	99.68%	99.76%		
	Monocyte	97.63%	100.00%	98.80%		
	Neutrophil	92.15%	88.46%	90.27%		
	Average	94.97%	94.99%	94.96%		
ShuffleNetV2	Eosinophil	90.87%	90.87%	90.87%	94.33%	310 s
	Lymphocyte	99.51%	98.23%	98.87%		
	Monocyte	97.29%	98.71%	97.99%		
	Neutrophil	89.73%	89.58%	89.65%		
	Average	94.35%	94.35%	94.35%		
ViT-B/16	Eosinophil	94.15%	95.35%	94.75%	96.86%	632 s
	Lymphocyte	98.10%	99.84%	98.96%		
	Monocyte	100.00%	99.52%	99.76%		
	Neutrophil	95.23%	92.79%	93.99%		
	Average	96.87%	96.88%	96.87%		
SW-ViT (proposed)	Eosinophil	97.38%	95.35%	96.35%	98.03%	388 s
	Lymphocyte	100.0%	98.71%	99.35%		
	Monocyte	99.36%	100.0%	99.68%		
	Neutrophil	95.48%	98.08%	96.76%		
	Average	98.06%	98.04%	98.04%		

**Figure 7.** Comparison of different classification models based on average values.

4.5. Grad-CAM Visualization

Grad-CAM (gradient class activation map) belongs to a visualization method to explain the outputs of deep models. In this study, the Grad-CAM is employed for the interpretation of the proposed SW-ViT. The results are given in Figure 8. The SW-ViT puts more emphasis on the red regions than on the blue areas, and in most of the listed Grad-CAMs, the majority of the heatmaps are in red and yellow, which uncovers that the proposed SW-ViT can generate better global representations. On the other side, the Grad-CAMs of some eosinophil and neutrophil images are not consistent with medical knowledge, which results in relatively lower F1-scores for both types.

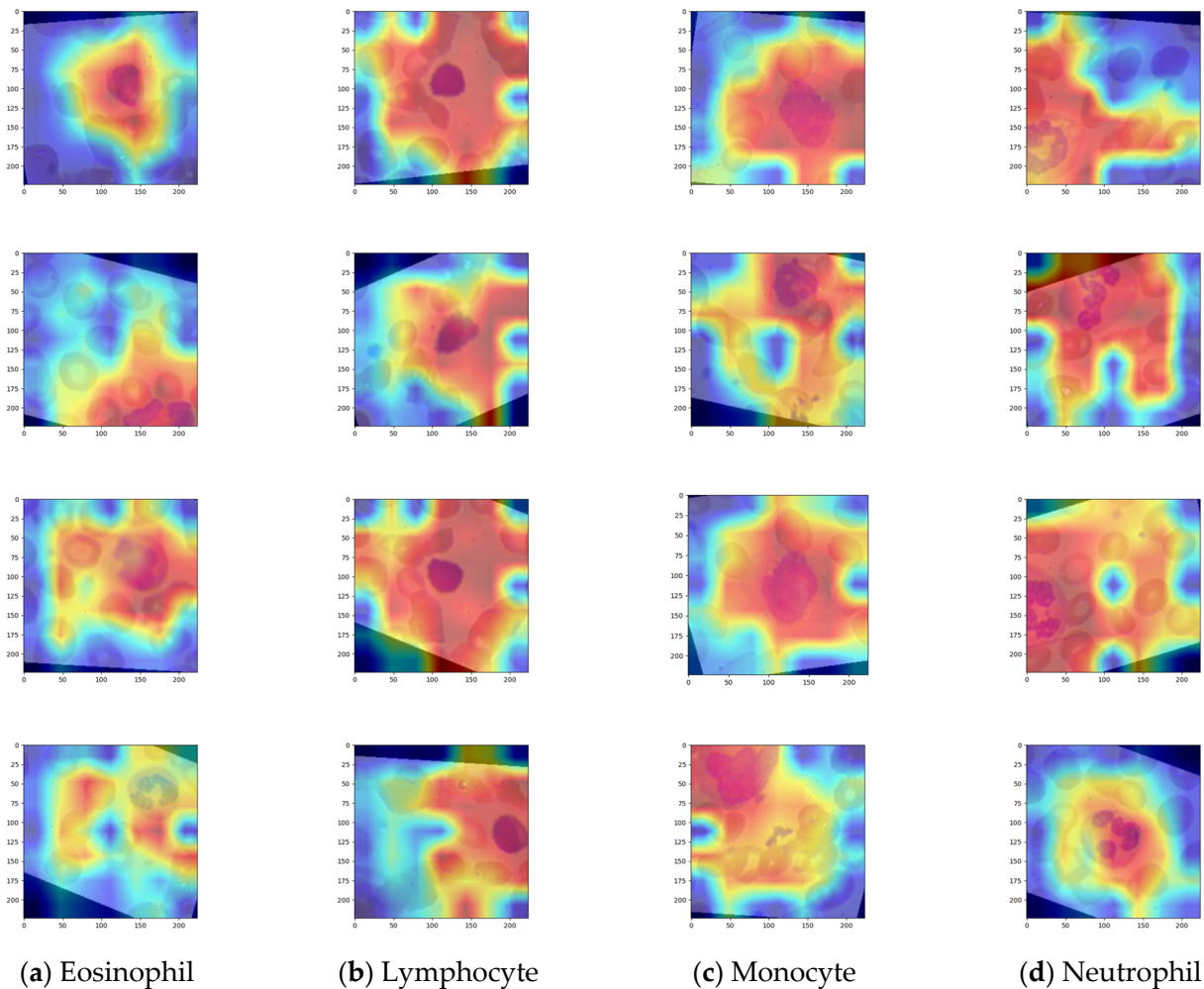


Figure 8. Grad-CAMs of the SW-ViT (every column represents one type of cells).

4.6. Mis-Classified Samples

Although the proposed SW-ViT can achieve good classification performance, there are still mis-classified samples. Some of the mis-classified images are listed in Figure 9. The eosinophil cells are likely to be classified as neutrophil cells by the model, which can be also found in the confusion matrix. The classification of two types of cell images shall be further investigated in future research.

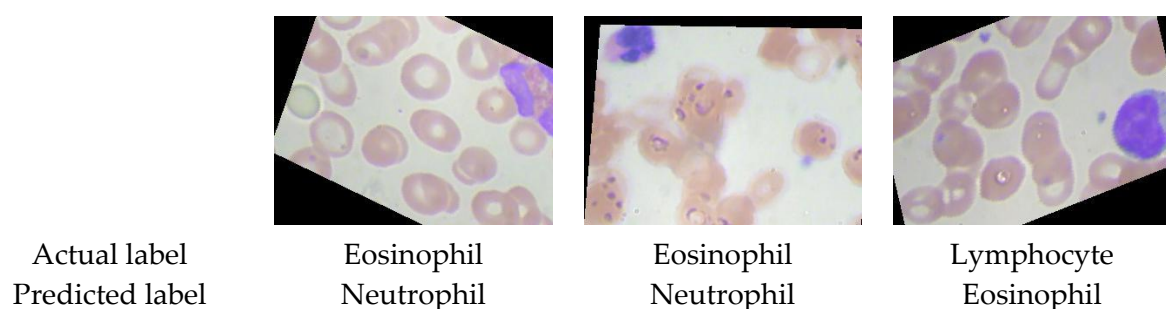


Figure 9. Mis-classified images.

4.7. Comparison with State-of-the-Art Methods

The comparison of the proposed SW-ViT with other state-of-the-art (SOTA) blood cell classification approaches is shown in Table 5. The BCCD_Dataset is also used in Lightweight CNN, Fusion CNN, and BCNet, but only three types of blood cell images are employed to evaluate the BCNet, which explains the good classification performance within only two epochs. Fusion CNN performs much better than the Lightweight CNN, but its max epoch is 100, which is time-consuming. In VGG+ESSA, they use the pre-trained VGG for feature extraction without fine-tuning, and the max epoch is set for the ESSA optimization. The recall of VGG+ESSA is the best out of the listed methods. However, the classifier optimization of 100 epochs is also time-demanding. Meanwhile, the dataset in their experiments contains hundreds of images of only two types. The proposed SW-ViT produces worse results than DenseNet-121 and Multi-level CNN, but the difference between the three is marginal. Meanwhile, the max epoch for DenseNet-121 is 10, and multiple CNNs are trained in Multi-level CNN, which requires more time than the proposed SW-ViT. The max epoch of the SW-ViT is also obviously lower than Lightweight CNN, Fusion CNN, and DenseNet-121, which are evaluated on the BCCD_Dataset. It can be inferred that the self-attention mechanism helps in global feature representation learning so that the proposed SW-ViT can achieve SOTA performance.

Table 5. Comparison with SOTA methods.

Method	Precision	Recall	F1-Score	Accuracy	# Types	Max Epoch
Lightweight CNN [11]	91.75%	91.50%	91.35%	91.64%	4	30
Fusion CNN [7]	96%	96%	96%	96%	4	100
BCNet [20]	97.07%	96.77%	96.78%	96.78%	3	2
VGG+ESSA [12]	93.43%	99.55%	96.22%	96.11%	2	100
DenseNet-121 [26]	99.33%	98.85%	99.09%	98.84%	4	10
Multi-level CNN [25]	98.37%	98.37%	98.36%	98.36%	4	-
SW-ViT (proposed)	98.06%	98.04%	98.04%	98.03%	4	3

5. Conclusions

In this study, a novel blood cell image classification model is presented based on vision transformers. The proposed SW-ViT leverages window self-attention modules for global and hierarchical representation learning, which reduces the computational complexity simultaneously. The transfer learning strategies are also employed to accelerate the convergence. Experimental results from a public dataset suggest that the SW-ViT is accurate and efficient for blood cell classification, which can also be inferred from the Grad-CAMs. The proposed SW-ViT can be a useful tool in clinical diagnoses to assist in verification.

However, there are some drawbacks to the proposed SW-ViT. The classification performance of eosinophil and neutrophil cell images is not satisfactory compared with the other two types. In the future, this problem shall be investigated. Currently, the SW-ViT can only recognize four types of blood cells. Larger datasets with more types of cells will be used to

train the SW-ViT for future research. In addition, blood cell segmentation and counting can be explored in the future.

Author Contributions: S.C.: Conceptualization; software; data curation; writing—original draft; visualization; S.L.: methodology; software, validation; investigation; resources; writing—review and editing; S.W.: methodology; software, validation, writing—original draft; Y.N.: investigation; resources; writing—review and editing; Y.Z.: methodology; formal analysis; investigation; writing—review and editing; supervision; project administration; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is partially supported by MRC, UK (MC_PC_17171); Royal Society, UK (RP202G0230); BHF, UK (AA/18/3/34220); Hope Foundation for Cancer Research, UK (RM60G0680); GCRF, UK (P202PF11); Sino-UK Industrial Fund, UK (RP202G0289); LIAS, UK (P202ED10, P202RE969); Data Science Enhancement Fund, UK (P202RE237); Fight for Sight, UK (24NN201); Sino-UK Education Fund, UK (OP202006); BBSRC, UK (RM32G0178B8).

Data Availability Statement: The dataset can be downloaded at <https://www.kaggle.com/datasets/paultimothymooney/blood-cells> (accessed on 30 January 2023).

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

References

1. Mercan, E.; Mehta, S.; Bartlett, J.; Shapiro, L.G.; Weaver, D.L.; Elmore, J.G. Assessment of Machine Learning of Breast Pathology Structures for Automated Differentiation of Breast Cancer and High-Risk Proliferative Lesions. *JAMA Netw. Open* **2019**, *2*, e198777. [CrossRef] [PubMed]
2. Acevedo, A.; Alferez, S.; Merino, A.; Puigvi, L.; Rodellar, J. Recognition of peripheral blood cell images using convolutional neural networks. *Comput. Methods Programs Biomed.* **2019**, *180*, 105020. [CrossRef] [PubMed]
3. Gupta, D.; Arora, J.; Agrawal, U.; Khanna, A.; de Albuquerque, V.H.C. Optimized Binary Bat algorithm for classification of white blood cells. *Measurement* **2019**, *143*, 180–190. [CrossRef]
4. Hegde, R.B.; Prasad, K.; Hebbar, H.; Singh, B.M.K. Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. *Biocybern. Biomed. Eng.* **2019**, *39*, 382–392. [CrossRef]
5. Abdulkarim, H.A.; Razak, M.A.A.; Sudirman, R.; Ramli, N. A deep learning AlexNet model for classification of red blood cells in sickle cell anemia. *IAES Int. J. Artif. Intell. (IJ-AI)* **2020**, *9*, 221–228. [CrossRef]
6. El-Seoud, S.A.; Siala, M.H.; McKee, G. Detection and Classification of White Blood Cells through Deep Learning Techniques. *Int. J. Online Biomed. Eng. (Ijoe)* **2020**, *16*, 94–105. [CrossRef]
7. Banik, P.P.; Saha, R.; Kim, K.-D. An Automatic Nucleus Segmentation and CNN Model based Classification Method of White Blood Cell. *Expert Syst. Appl.* **2020**, *149*, 113211. [CrossRef]
8. Baydilli, Y.Y.; Atila, U.; Elen, A. Learn from one data set to classify all—A multi-target domain adaptation approach for white blood cell classification. *Comput. Methods Programs Biomed.* **2020**, *196*, 105645. [CrossRef]
9. Kutlu, H.; Avci, E.; Ozyurt, F. White blood cells detection and classification based on regional convolutional neural networks. *Med. Hypotheses* **2020**, *135*, 109472. [CrossRef]
10. Loey, M.; Naman, M.; Zayed, H. Deep Transfer Learning in Diagnosing Leukemia in Blood Cells. *Computers* **2020**, *9*, 29. [CrossRef]
11. Ridoy, M.A.R.; Islam, M.R. An Automated Approach to White Blood Cell Classification Using a Lightweight Convolutional Neural Network. In Proceedings of the 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 28–29 November 2020.
12. Sahlol, A.T.; Kollmannsberger, P.; Ewees, A.A. Efficient Classification of White Blood Cell Leukemia with Improved Swarm Optimization of Deep Features. *Sci. Rep.* **2020**, *10*, 2536. [CrossRef] [PubMed]
13. Settouti, N.; Bechar, M.E.A.; Daho, M.E.H.; Chikh, M.A. An optimised pixel-based classification approach for automatic white blood cells segmentation. *Int. J. Biomed. Eng. Technol.* **2020**, *32*, 144–160. [CrossRef]
14. Chen, H.; Liu, J.; Hua, C.; Zuo, Z.; Feng, J.; Pang, B.; Xiao, D. TransMixNet: An Attention Based Double-Branch Model for White Blood Cell Classification and Its Training with the Fuzzified Training Data. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021.
15. Çınar, A.; Tuncer, S.A. Classification of lymphocytes, monocytes, eosinophils, and neutrophils on white blood cells using hybrid Alexnet-GoogleNet-SVM. *SN Appl. Sci.* **2021**, *3*, 503. [CrossRef]
16. Dinčić, M.; Popović, T.B.; Kojadinović, M.; Trbović, A.M.; Ilić, A. Morphological, fractal, and textural features for the blood cell classification: The case of acute myeloid leukemia. *Eur. Biophys. J.* **2021**, *50*, 1111–1127. [CrossRef] [PubMed]
17. Liao, Z.; Zhang, Y.; Li, Z.; He, B.; Lang, X.; Liang, H.; Chen, J. Classification of red blood cell aggregation using empirical wavelet transform analysis of ultrasonic radiofrequency echo signals. *Ultrasonics* **2021**, *114*, 106419. [CrossRef]

18. Semerjian, S.; Khong, Y.F.; Mirzaei, S. White Blood Cells Classification Using Built-in Customizable Trained Convolutional Neural Network. In Proceedings of the 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 5–7 March 2021.
19. Yao, X.; Sun, K.; Bu, X.; Zhao, C.; Jin, Y. Classification of white blood cells using weighted optimized deformable convolutional neural networks. *Artif. Cells Nanomed. Biotechnol.* **2021**, *49*, 147–155. [[CrossRef](#)]
20. Zhu, Z.; Lu, S.; Wang, S.H.; Gorriz, J.M.; Zhang, Y.D. BCNet: A Novel Network for Blood Cell Classification. *Front. Cell Dev. Biol.* **2021**, *9*, 813996. [[CrossRef](#)]
21. Ichim, L.; Iordan, C.-A.; Popescu, D. Multi-Network Blood Cell Classification System Based on Decision Fusion. In Proceedings of the 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 16–18 November 2022.
22. Zhu, Z.; Wang, S.; Zhang, Y. ROENet: A ResNet-Based Output Ensemble for Malaria Parasite Classification. *Electronics* **2020**, *11*, 2040. [[CrossRef](#)]
23. Elhassan, T.A.; Rahim, M.S.M.; Zaiton, M.H.S.; Swee, T.T.; Alhaj, T.A.; Ali, A.; Aljurf, M. Classification of Atypical White Blood Cells in Acute Myeloid Leukemia Using a Two-Stage Hybrid Model Based on Deep Convolutional Autoencoder and Deep Convolutional Neural Network. *Diagnostics* **2023**, *13*, 196. [[CrossRef](#)]
24. Bayat, N.; Davey, D.D.; Coathup, M.; Park, J.-H. White Blood Cell Classification Using Multi-Attention Data Augmentation and Regularization. *Big Data Cogn. Comput.* **2022**, *6*, 122. [[CrossRef](#)]
25. Cheuque, C.; Querales, M.; León, R.; Salas, R.; Torres, R. An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification. *Diagnostics* **2022**, *12*, 248. [[CrossRef](#)] [[PubMed](#)]
26. Sharma, S.; Gupta, S.; Gupta, D.; Juneja, S.; Gupta, P.; Dhiman, G.; Kautish, S. Deep Learning Model for the Automatic Classification of White Blood Cells. *Comput. Intell. Neurosci.* **2022**, *2022*, 7384131. [[CrossRef](#)] [[PubMed](#)]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021.
30. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.