Tabriz Journal of Electrical Engineering (TJEE), vol. 54, no. 3, Autumn 2024 Serial no. 109
DOI: 10.22034/tjee.2024.58463.4727

Leukovit: An efficient vision transformer-based model for automatic classification of leukocytes

Z. Asgharzadeh bonab[1], S. Shamekhi[1]*

Faculty of Biomedical Engineering, Sahand University of Technology, Tabriz, Iran[1].
shamekhi@sut.ac.ir
*Corresponding author

Received: 22/09/2023, Revised: 22/12/2023, Accepted: 04/02/2024.

Abstract
The identification and evaluation of leukocytes is important to assess the quality of the human immune system; however, the analysis of blood smears depends on the pathologist's expertise. The manual method for analyzing and classifying WBCs is costly and time-consuming and can result in errors in detection. Most deep learning methods use CNN-based models for white blood cell classification. This paper discusses the use of a ViT-based network, for the classification of leukocytes (WBCs) in a blood sample. The Dataset used in this paper consists of 352 images with a size of 320x240, which was augmented through techniques to create a balanced dataset of 12444 images. The augmented data was then used to train a ViT-based architecture to classify the different types of WBCs. As the first step of the proposed algorithm, a convolutional tokenizer has been applied for patch extraction of images. These patches have been flattened and have been used as input for a ViT-based structure to recognize the subclasses in the second step. The results obtained using Leukovit show that the accuracy of the proposed network is 99.04% which outperforms the state-of-the-art networks.

Keywords: White blood cells, Image classification, Deep learning, Convolutional neural network, Vision Transformer.

1. Introduction

Healthy blood is vital for various actions of the body's organs, and consequently human life [1]. White blood cells (WBCs) as one of the essential cells of the blood are responsible for fighting foreign pathogens and protecting the body from infection. Classification of WBC in a blood sample is a very important need because every different type of WBC has a specific function in the fight against virus, fungus, bacteria, and all other kinds of infections in our body. There are five different types of WBCs with different responsibilities, including lymphocytes, neutrophils, eosinophils, monocytes, and basophils [2].

Correct identification of different WBC classes leads to the possibility of counting the different WBCs and assessing their proportions of presence for some conditions and classes, such as healthy, cancerous cells, immune system disorders, and leukemia [3]. The previously used procedure for analyzing and classifying the WBCs by hematology microscopic images was a manual, costly, and time-consuming process that encountered errors in detection [4]. Therefore, automatic image analyzer systems can help pathologists improve the segmentation and classification of images.

Machine learning algorithms have been applied for the classification of WBCs such as k-means clustering, decision tree [5], support vector machine (SVM), and K-Nearest Neighbor (KNN) [6, 7]. Recently, convolutional neural networks (CNNs) have considerably succussed in classification of WBCs [8]. To recognize the types of WBCs, CNN-based architecture has been used [9] for taking advantage of pre-trained models, such as AlexNet [10], VGG16 [11], GoogleNet [12], and ResNet [13]. In some other frameworks that combine RNN and CNN, important features are extracted efficiently, with more computational time and higher system requirements. The RNN procedure can model a sequence of data, but pixel values are distributed across the grid and underfitting problems may arise [14, 15]. However, as recent studies demonstrate, combining different deep networks for deep feature extraction increases the complexity of algorithm and the loss of detailed information of WBC structure [16-18].

The paper is structured as follows. "Literature survey" reviews the existing methods related to Transformers and WBC classification. Dataset description and the proposed framework are presented in "Materials and methods". "Results" describe the experimental results. "Discussion" discusses about the whole procedure, and finally, we conclude this paper in "Conclusion".

2. Literature survey

For years, convolutional neural networks (CNN) have dominated the field of medical image processing. But recently, Vision Transformers (ViT) has revolutionized this field, which through the mechanism of attention has proven to achieve excellent results in various tasks.

The general architecture of standard ViT is summarized in 4 main steps:

• The image is divided into non-overlapping patches.

• The patches are flattened, and lower-dimensional linear embeddings are created from the flattened patches.

• Adding positional embedding and class token if needed.

• The sequence of patches is fed into transformer block and the output (label) is got through a classification layer to obtain final output prediction.

In ViTs, the most important part is the transformer block that usually contains multi-head-self-attention (MHSA) and MLP. The transformer block receives 3 types of embeddings (patch embeddings, positional embeddings, and class tokens) as input [19]. Several papers provide WBC detection using deep learning techniques, using CNNs, including LeNet, GoogleNet, AlexNet and ResNet, Inception, and so on [20-22].

Before the standard Vision Transformer architecture was introduced by Alexey Dosovitskiy [19], multiple works tried to combine CNN-based architectures with self-attention [23, 24]. Some replace the convolutions completely [25, 26]. Also, many cases have considered the combination of CNNs with self-attention forms, e.g. by augmenting feature maps for image classification [27] or by processing over the output of a CNN using self-attention [24, 28, 29].

Recently, the use of ViTs in the field of medicine has made great progress. Researchers tried to use transformers to improve the weaknesses of CNN. TransUNet [30] is the first transformer-based medical image segmentation structure, which uses the transformer to encode the global context. Mehboob et al. [31], implemented a self-attention transformer-based model for the COVID19 diagnosis using CT scan images. This was the first work performing research based on a Vision Trthensformer for the detection of COVID. So, this approach based on the self-attention transformer is of great importance for using in CT scan images and diagnosis of COVID19.

Another transformer application, used by Qu et al. [32] is for breast cancer detection. It uses a CNN framework to extract local features and applied a ViT to learn the global relationship between various regions and improve the relevant local features. It comprises a VGG backbone, an FC feature extractor layer, and a squeeze-and-excitation block using breast cancer ultrasound image dataset.

Dai et al. [33] introduced a multi-modal medical image classification based on transformers called TransMed. TransMed by CNN and Transformer combination can be effective for extracting low-level features of images and creating long-range dependencies among modalities and as it has been claimed, this is the first work to apply Transformers to multi-modal medical image classification.

Considering the advantages of convolutions and Transformer, Wang et al. [34] proposed an O-Net architecture to combine CNN and Transformer to learn global and local textual features. In this method, the combination of CNN and Swin Transformer was first used as an encoder and then sent to the CNN-based and the Swin Transformer-based decoder, respectively. The combination of 2 decoders gets a better result. The O-Net performs well in a classification task simultaneously.

As far as we know, few studies have been done in classifying WBCs with the help of Transformers, and we can refer to the method presented by Priscilla Cho, et al. [35] for the classification of 2 categories of normal and malignant for the diagnosis of healthy or cancerous cells. Also, Tripathi et al. [36] combined convolution and attention network named (CoAtNet [37]) outperformed CNN models in cell classification compared to 2 baseline models, EfficientNetV2 and ResNext50. The research investigated the morphology of BM in adults using the dataset of bone marrow samples. Also, Haung et al. [38] showed that integrating attention-aware and manifold learning strategies, along with cascaded basic attentional residual units, facilitates the aggregation of multi-scale deep-level features. This enables effective adaptation to category-relevant image-level features. Wang et al.[39] introduced a model which incorporates feature fusion strategies, leveraging a focused attention mechanism. To enhance the localized attention of the Convolutional Neural Network (CNN), the fusion features from both the initial and final convolutional layers are extracted. This process is achieved through a focalized attention mechanism that combines Squeeze-and-Excitation (SE) and Gather-Excite (GE) modules.

Recently, a number of studies have investigated the performance of ViT-based models in the classification of WBCs. Katar et al. [40] applied the standard ViT model for 5-class WBCs and utilized the Score-CAM algorithm, for visualization of prioritizes during predictions. A pre-trained model, using the ImageNet-21k dataset (consisting of 14 million images and 21,843 classes), was employed to mitigate the considerable expenses associated with training a Vision Transformer (ViT) model from scratch. The ViT model's training duration was 100 epochs. Chen et al. [41] . applied the Shifted Window Vision Transformer (SW-ViT). The SW-ViT architecture undergoes initial pre-training on the ImageNet dataset, followed by fine-tuning specifically on blood cell images for classification. To enhance classification results, two transfer strategies are employed. One involves fine-tuning the complete SW-ViT architecture, while the other focuses on fine-tuning only the linear output layer, with all other parameters frozen. The experiments utilize the BCCD_Dataset. Similarly, Dipto et al. [42] used standard ViT in identifying WBC types. The accuracy for model falls within the range of 83% to 85%.

In convolutional architectures, by adding more convolution layers, performance may be increased, but Transformers are different, they quickly saturate as the architecture deepens. Because the deeper the Transformer becomes, the attention map becomes more similar. CNN’s view is local, and its field of view becomes more global at each layer with a larger field of view because it looks at information already collected by the first layer. But the Vision Transformer presented in [19] succeeds in having a large field of view from the very beginning. Transformers, unlike CNNs, are very

data-hungry, especially because they have the freedom to look all over the picture from the start. In other words, to make accurate predictions, it is initially very unfocused and needs large amounts of data to learn how to focus and what to focus on. Studying recent works, this paper is one of many that surveys whether "Attention" or "convolutions" are necessary? Considering the specific costs that each has, both attentions and convolutions have very favorable qualities for inference and prediction. Considering a small dataset, we sought to introduce a model that can be trained, from scratch, on WBCs of blood smear images. To our knowledge, this is the first novel transformer-based model of WBC subtypes classification on this dataset. To have a compact, small, and efficient model, our goal is to create a simple model, with few parameters, that can be trained fast and efficiently while retaining the state-of-the-art results.

In summary, the following contributions were performed:

1. Our work introduces a novel transformer-based architecture designed for White Blood Cell (WBC) classification in blood smear images. Remarkably, this approach achieves superior performance with reduced computational cost compared to existing deep models.

2. Inducing bias into the model using convolutions in patch extraction, also adding flexibility to the model.

3. Diverging from conventional Vision Transformers (ViTs), our model adopts MLP-sharing instead of Multi-Head Self-Attention (MHSA). This approach observed features in patches on every channel and searched for features on all the channels for each patch.

4. The simplicity of the proposed method translates into a reduced number of parameters, and the core of the architecture was formed by the MLP-layers.

6. Through a comprehensive evaluation employing both qualitative and quantitative analyses, including the use of Grad-CAM, our method demonstrates a notable superiority over the latest techniques in WBC classification.

3. Methods

3.1. Data

In this analytical study, a WBC Dataset including 352 images with a size of 320 × 240 has been used [43]. Since the percentages of basophils in the normal blood (≤ 1%) and the used dataset are very low (3 images), we haven't considered these few images of basophils, and four other types of blood cells have been classified. Some examples of the dataset are depicted in Fig. 1.

A data augmentation, including random rotation, shearing, and flipping, has been applied. A random rotation with an angle between 0 to 45 degrees, and a horizontally/vertically shifting and flipping have been applied. Also, a shearing transformation slants the shape of the image by a range of 0.2. This augmentation led to a balanced dataset with an almost equal number of all classes. The augmented data of this work comprises 12444 WBC images, including 3120 eosinophils, 3103 lymphocytes, 3,098 monocytes, and 3123 neutrophils. It should be noted that the above-mentioned augmentations have already been implemented on the original images of the available dataset. Fig. 2 shows some examples of augmented datasets. The augmented data were split into training and testing sets with a ratio of 80:20, and the images were resized to 64×64×3 for the proposed architecture.

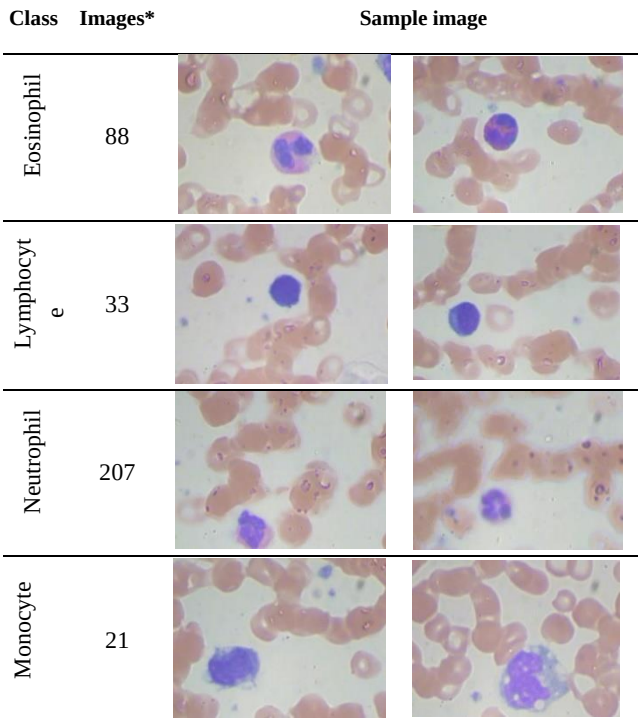


Fig. 1. Sample images of each class from the dataset [46]. *Total number of cell images contained in each class.

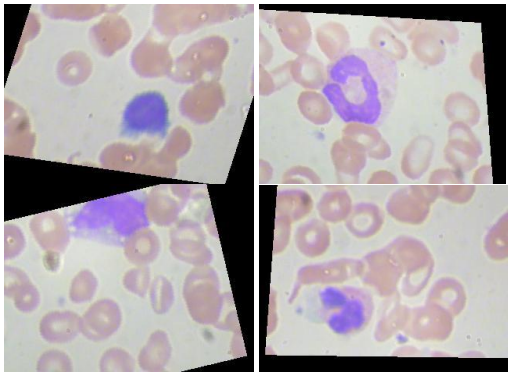


Fig. 2. Augmented data samples using rotation, flipping, shearing, etc.

3.2. Proposed Model

In this section, the proposed architecture for classification of four types of WBC images in blood smear is described. The proposed architecture is depicted in Fig. 3. To design the model, the standard Vision Transformer [19] and the original Transformer [44] structures are followed.

As mentioned in Section 1, ViTs are data-hungry (large datasets are needed), and are computationally hungry; therefore, training by inadequate amounts of data, does not generalize well [19]. However, CNNs can learn even

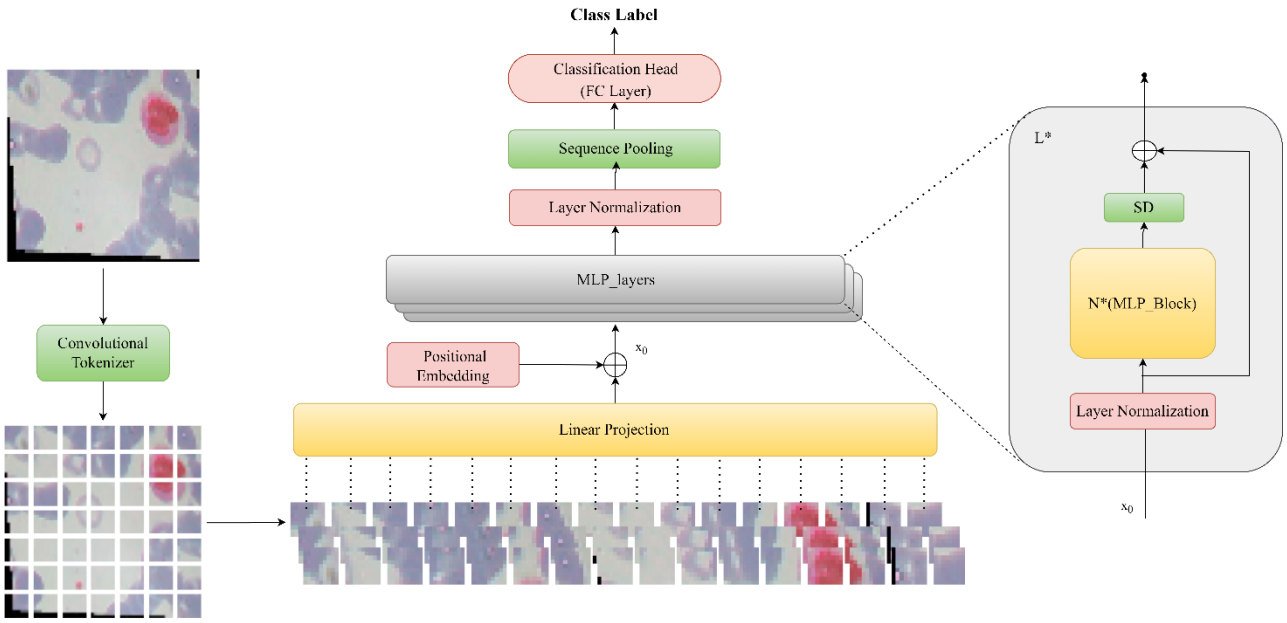


Fig. 3. The architecture of Leukovit, consist of a patch extractor, MLP structure, and classification head

with relatively small amounts of data, which is because of the existence of inductive biases [45]. As a result, CNNs are still used for smaller data sets. They are efficient compared to Transformers because they are better both in terms of computation and memory. Also, local inductive bias is more important for working with smaller images. While they require less time and data for training, they also use fewer parameters to accurately fit the data. In this work, we tried to keep the above-mentioned issues into account and create an architecture that can both focus on important features within the images and be spatially invariant, as we have sparse interactions and weight sharing. This allows a transformer-based framework to be trained from scratch on small datasets.

The first step starts with changing the way patches are extracted. A standard ViT takes a sequence of vectors, called tokens as input, that are obtained from patches. These patches are extracted by dividing the input image into small parts of fixed size. To do this, standard ViT partitions an image into non-overlapping square patches as shown in Fig. 4.

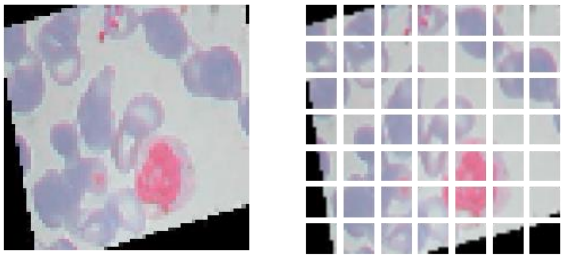


Fig. 4. Example images with patch extraction in standard ViT.

Considering the patch size of P, the sequence of patches is flattened into the one-dimensional vector and transformed into latent vectors of dimension d. It works like a convolution layer which has d filters and the kernel size and stride is P.

In this way of patch extraction, some issues may arise. One is that it is difficult for the network to understand the boundary information between patches. In this case, the network only performs attention to each patch and then determines how those attentions relate. It is referred as related inductive bias. Whereas the proposed network uses convolutions that overlap, which is done by tuning kernel size and stride. This leads to inducing some *bias* into the model and allows the network to learn information that is better embedded. It is an important step because the overlapping convolutions are invariant to spatial translations and have low *related inductive bias*. This is

one of the reasons for the success of CNNs in vision. The patch extraction using convolutions led to generating richer tokens and preserving local information since it is good at encoding relationships between patches comparing the standard ViT. The convolutional token extractor which consists of a single convolution, *ReLU* activation function, and a max-pooling layer. Given an image or feature map $x \in \mathbb{R}^{H\times W\times C}$ (the original input image has resolution (H, W) and C channels:

$$x_0 = MaxPool(RelU(Conv2D(x))) \quad (1)$$

Where the convolution operation has d filters, and the same number as the embedding dimension. However, by using this convolutional block for patch extraction, the model is more flexible than standard ViT, this is because it is no longer dependent on the input resolution, which is strictly divided by the preset patch size. Since the convolutional extracted patches can be adjusted in terms of kernel size, stride, and padding, and are repeatable.

The next step is adding positional embedding to attach spatial information to the sequence. Since the model does not actually know much about the spatial relationship between the tokens, adding extra information can be helpful. Generally, like standard ViTs, a learned

embedding is applied. Then these patches are inputs for the MLP-layer.

MLP-Layer applied in this work has a simple structure that is not like the one used in the transformer encoder (a self-attention followed by multi-layer perceptron structure). Its architecture is completely based on MLPs, which are repeatedly implemented in spatial locations or feature channels in each MLP block. The question arises that how it is possible? The development timeline of classification methods was from MLPs, then CNNs, and Residual CNNs, i.e., ResNets, DenseNets, ViT, and so on. Now, is it true to come back to MLP? The answer is yes, while convolution and attention are both adequate for acceptable performance, but neither is necessary.

The MLP block comprises a sequence of layers that all have the same structure: a linear layer applied across the patches, followed by a layer applied across the channels, and each layer parallelized by a skip-connection [46, 47]. As shown in the Fig. 5, the MLP block accepts a sequence of flattened patches (referred to as tokens earlier) as input and maintains the dimensions in the form of patches × channels. The MLP block uses MLPs in 2 stages: channel-sharing MLPs and token-sharing MLPs. Channel-sharing MLPs allow communication between different channels. They operate on each token independently and take individual rows as input. Token-sharing MLPs enable communication between different spatial locations (tokens). They operate independently on each channel and take separate columns as input. This part is seeking features only in that patch and associates them with the channel, whereas the first part is searching for features in all the channels. This structure aims to separate the per-location (channel-sharing) operations and cross-location (token-sharing) operations. Both operations are done by MLPs.

The sequence of image patches as input (I), each one projected to a desired hidden dimension (C), results in a two-dimensional input, $\boldsymbol{X} \in \mathbb{R}^{I \times C}$, the token-sharing MLP acts on columns of $\boldsymbol{X}$ (it is applied to a transposed input $\boldsymbol{X^T}$) and is shared across all columns. The second one is the channel-sharing MLP which acts on rows of $\boldsymbol{X}$ and is shared across all rows. Each MLP block consists of 2 FC layers and a nonlinear applied independently to each row of its input data stensor. The equations can be written as follows:

$$U_{*,i} = X_{*,i} + W_2\sigma\left(W_1 LayerNorm(X)_{*,i}\right), i = 1, \dots, C \quad (2)$$

$$Y_{j,*} = U_{j,*} + W_4\sigma\left(W_3 LayerNorm(U)_{j,*}\right), \quad i = 1, \dots, I \quad (3)$$

Here, for σ, the nonlinear GELU is used. I and C are tunable widths in the token and channel-sharing MLPs, respectively. Therefore, in terms of the number of input patches, the computational complexity of the network is linear, unlike ViT, whose complexity is quadratic. The input size is the same in each MLP-Block. This part is similar in structure to Transformers or deep RNNs, which also use fixed width, but all CNNs have a pyramid structure. As shown in proposed architecture, Fig. 3, stochastic depth (SD) is just before the residual connection. Stochastic depth is a regularization technique that randomly drops a set of layers, which is very similar to dropout. Unlike dropout, that works on individual nodes in a layer, this method works on a block of layers [46, 48].

Many works have considered the best approach to apply normalization in a Transformer. While Layer Normalization is always the preferred method of normalization, there are two variants of how it is applied: Pre-Norm and Post-Norm. Post-Norm normalizes the output of the sum of skip-connection and the residual, while Pre-Norm normalizes the representation of the residual branch before the applying any projections within. The standard implementation of ViTs uses Post-Norm, which has a benefit of delivering stronger task performance than Pre-Norm under the setting when the default learning rate is used.

$$\textbf{Accuracy} = \frac{\boldsymbol{True\ Positive + True\ Negative}}{\boldsymbol{Total}} \quad (4)$$

$$\textbf{Precision} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive + False\ Positive}} \quad (5)$$

$$\textbf{Recall} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive + False\ Negative}} \quad (6)$$

$$\textbf{F1 Score} = 2 \times \frac{\boldsymbol{Precision \times Recall}}{\boldsymbol{Precision + Recall}} \quad (7)$$

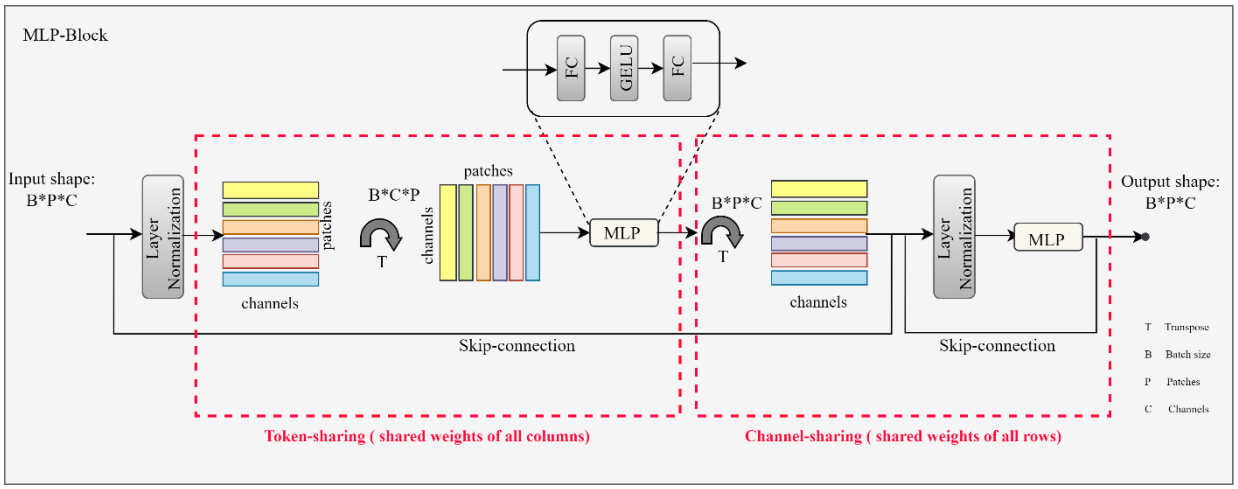


Fig. 5. MLP-Block showing MLP-layer structure in detail

For classification head, ViTs usually add an additional token (CLS class token) to the embedded patch sequence, containing latent information, and collect additional information about the sequence through self-attention, which is later used for classification. ViTs also use average pooling output tokens but find no significant difference in performance. Therefore, it can be replaced by another pooling method. Sequence-Pooling (SeqPool) [49], an attention-based method, pools the whole sequence of embedded patches. Considering that the output sequence contains relevant information in different parts of the input image, the performance can be improved by preserving this information without any additional parameters. Which is simpler and more efficient in contrast to the learnable token. Additionally, this operation reduces the computation a bit, because one less token is being forwarded. SeqPool allows the network to weigh the sequential embeddings of the latent space generated by the MLP-Layer. This operation involves mapping the output sequence using the T transform. $T: \mathbb{R}^{b\times n\times d} \rightarrow \mathbb{R}^{b\times d}$. b is batch size, n is sequence length, and d is the embedding dimension. Then, this output can be sent to a classifier. Some pooling methods, including learnable and static methods, indicate that learnable pooling performance is better, and sequence pooling outperforms others in the case of small-scale training. This allows the model to weight tokens according to their information relevance since each embedded patch does not have the same amount of entropy. The final step is a classification head which comprises FC layer to predict the class labels.

4. Results

For ViTs, studies show that some models get their best results using Adam and AdamW optimizers, while the proposed model performed best with Nadam [50] and learning_rate of 0.001, resulting faster training time than Adam. Nadam uses a decaying step size and first moment hyperparameters that can improve performance. The model trained from scratch, for 50 epochs and the batch size set to 32. Label smoothing [51] commonly used as regularization tool operating on the labels to increase robustness and improve classification problems that prevents the network from becoming over-confident. In this model, the label smooth parameter was assigned to 0.1. Two convolutional layers, each comprising a 3x3 convolution with stride=1 and padding =1, ReLU activation and He_normal as kernel initializer is used; the token extractor is appended with a max pooling layer with the “same” pooling stride equal to 2. Layer normalization epsilon values implemented in every block set to 1 - e6. We used the dropout rate of 0.1 to regularize the model. The proposed algorithm has been implemented on hardware NVIDIA GeForce RTX3060. All steps have been performed in Python Version: 3.9.16 using Keras with TensorFlow backend (Version: 2.10.0). The Equations 4-7 were used to evaluate performance parameters. Performance results on different structures tested and mentioned in Table I, II. This analysis indicates, while the standard ViT is used, the performance is good, since applying attention to classification tasks improves the performance but adding a convolutional patch extractor enhances performance because of additional local information and bias considered. The proposed model outperformed of the other models with no attention layer (i.e., MHSA), the main part of model structure consists of MLPs, while the model size is smaller too. Fig. 6 shows the confusion matrix, loss and accuracy curves in training and validation of the proposed model. These graphs show that the validation loss in certain periods decreases, and simultaneously, the validation accuracy increases. Therefore, the proposed model showed better performance by increasing the classification accuracy with less validation loss. The results of WBC subtypes classification are shown in Table III. Also, Monte Carlo Cross-Validation (MCC) is employed to comprehensively assess the model's performance. This resampling technique is particularly well-suited for handling dataset variability, ensuring robustness in evaluation. Averaging results over multiple random splits (5 was chosen), enhances the generalization assessment and mitigates the impact of randomness in our dataset, contributing to a more thorough and dependable evaluation of model. According to this resampling, average validation accuracy increased to 99.16%.

The Fig. 7 shows precision-recall and ROC curves for each of the classes. In the precision-recall curve, the trade-off between accuracy and recall can be seen. A lower false positive and false negative rate results in a higher area under the curve.

We analyzed the proposed model using Grad-CAM [52] to determine which regions of the input images are important for network classification decisions. Grad-CAM technique by generating Class Activation Maps (CAMs), is a good way to show what the model focuses on in the given data. As illustrated in Fig. 8, for all 4 classes, the Grad-CAM and Attention Map show the focus of the model on the desired cell, in the center of the image, and not on the background and red cells.

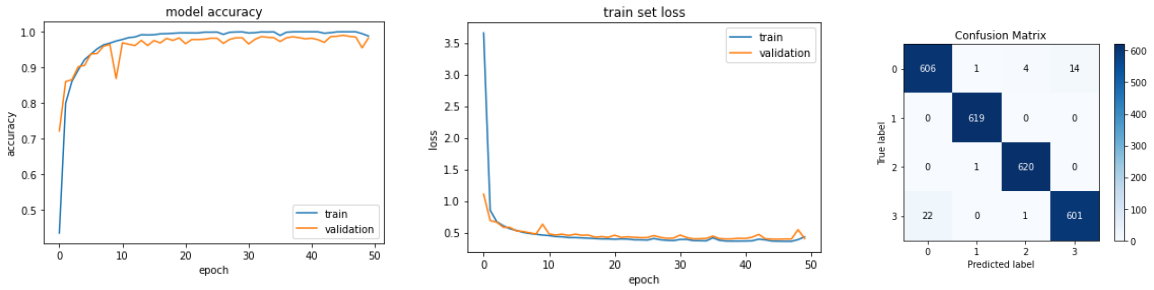


Fig. 6. Performance evaluation of the proposed model on WBC dataset.

Table I. Performance analysis on 50 epochs.

Methods	Pars. (million)	Acc (%)	Val_acc (%)	Val_loss
Standard ViT	0.60	96.05	94.48	0.138
Standard ViT+Convolutional patch extractor	0.86	98.70	97.20	0.406
Standard ViT+Convolutional patch extractor +MLP-Layer	0.88	99.50	97.80	0.399
Lekovit (Proposed)	0.37	100	99.04	0.393

Table II. Performance analysis using different optimizers.

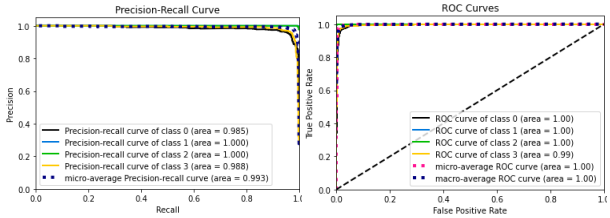
Methods	Time (per epoch)	Acc (%)	Val_acc (%)	Val_loss
Adam	39s	98.70	97.70	0.422
AdamW	26s	97.00	96.10	0.464
Nadam	20s	100	99.04	0.393

Table III. Evaluation parameters for each subtype

Class	Precision (%)	Recall (%)	f1-score (%)
NEUTROPHIL	96	97	97
MONOCYTE	100	100	100
LYMPHOCYTE	99	100	100
EOSINOPHIL	98	96	97
Average accuracy: 99.04%			
Average accuracy using MCC: 99.16%			

Table IV compares the classification results of WBC and state-of-the-art models on the same dataset where there is no difference in the number of classes and the number of images. This indicates, the latest methods have provided lower validation or test accuracy, even using segmentation, pre-processing, or ROI extraction before the classification phase. Also, the proposed model achieved higher performance than all except for one method presented in [17]. The framework of this model comprised a pre-trained AlexNet model, feature fusion, FS strategies, and ELM. This model adopts a pre-trained CNN as a backbone to extract convolutional features and all the deep features from different layers are combined to generate a 1352-dimensional fusion vector. by following, the FS strategies are used to select the discriminative feature vector, and then the selected features are fed to the ELM for WBC-type identification with 19,500 hidden neurons. The performance of the mentioned model is slightly higher, because none of the compared techniques applying intermediate layers with a feature selection strategy. Furthermore, ELMs often lack fine-tuning capabilities compared to neural networks. The hidden layer parameters, such as the number of hidden neurons and activation functions, are typically randomly initialized and not fine-tuned through iterative optimization. However, MLPs that are used in the proposed model can adapt to various tasks and data distributions by adjusting the architecture and hyperparameters. In addition, it should be noted that the size of the images in the proposed method (64×64) is much smaller compared to the mentioned article, which is 227×227.

One of the concerns during training and testing a model is the time taken by the system, which usually was high because of memory limitations. The number of epochs was decreased to around 50 in this work and while the performance is better.



(a) ***(b)***

Fig. 7. Classification curves. (a) Precision-recall and (b) ROC curves of the different classes**.**

5. Discussion

The diagnosis of various pathologies, such as leukemia or other hematological disorders, relies on the classification of subtypes of WBCs, also known as **leukocytes**. Despite several biological techniques to identify leukocytes, microscopic examination of blood smears is often critical to confirm the diagnosis. In this study, a vision transformer-based model using MLPs was proposed to automatically classify WBC subtypes in blood smear images.

We observed that the training of the model with MLP-Layers is more stable than ViTs. This stability comes from replacing self-attention with linear layers. In a similar way to token-sharing MLPs, Wu et al. [53] used parameter sharing in the depth-wise convolutions for natural language processing. It was presented that we could improve the performance of the model by having a small number of parameters, which is possible because of the use of MLP linear layers. In the best case for training data, the accuracy reached 1, which is an impressive result with about 0.3 million parameters.

It is better to mention that the best performance of ViTs is due to larger images, since larger images in attention mechanism, hold more information and create longer embedding sequences. While large size of images, led to better results, it also has a cost of computation. It increases the parameters to some extent and slows down the training of the network. Therefore, it is important to design the network with good results on smaller images, and this is considered in the proposed model. The results show that the model has been able to perform optimally for smaller-sized images. Because many fields, especially medicine, rarely have available datasets as large as ImageNet, the less the deep learning dependence on large amounts of data, the better. On the other hand, since some events are rare and it is difficult to correctly assign labels, it is even more difficult to create a dataset that has low bias. For example, it is difficult to collect samples for a rare disease considering the associated factors. Furthermore, for a rare disease, there may be only thousands of diseased samples, which is usually not sufficient to train a network with the desired results, unless enough data can be got from pre-trained models,

and also the large data leads to the need for large computing resources. This not only limits the ability to apply the models to different domains but also limits reproducibility. Therefore, we have shown in this paper that with proper configuration, this model can be used on small data sets and outperform convolutional models of the same size.

Leukovit model is also simple and flexible in terms of size. Having shallow convolutional layers at the input tokenization stage makes the model more flexible regarding input image size. The inductive bias introduced can also make the model more efficient in learning and more accurate in the results. Whereas in self-attention, the weights to aggregate information from other patches are data dependent through queries and keys, in the proposed method the weights are not data dependent and are based on absolute positions of patches.

Also, the confusion matrix is a more comprehensive mode of evaluation that provides more insight to the model's performance. As usual, the diagonal elements are the correctly predicted samples. A total of 2446 samples were correctly predicted out of the total 2489 samples. Thus, the overall accuracy is 99.04%.

The rows with 0 in the matrix imply that the model does not confuse samples originally belonging to Monocyte and Lymphocyte with other classes, i.e., the classification features between these classes were learned well by the model. To improve the model's performance, one should focus on the predictive results in Eosinophil and Neutrophil. A total of 23 samples (for Eosinophils) and 19 samples (for Neutrophils) were misclassified by the classifier. Typically, classification predictions made by neural networks do not lend themselves to simple human analysis because they are generated in a data-driven manner based on the training set. However, to gain a deeper understanding of the classifications performed by these algorithms, several analytically approaches have recently been developed to help investigate them [57].

The use of MCC allows us to rigorously validate our model's performance across various subsets of data, offering insights into its consistency and stability. The careful validation process enhances our confidence in the reported results and strengthens the overall credibility of our model evaluation.

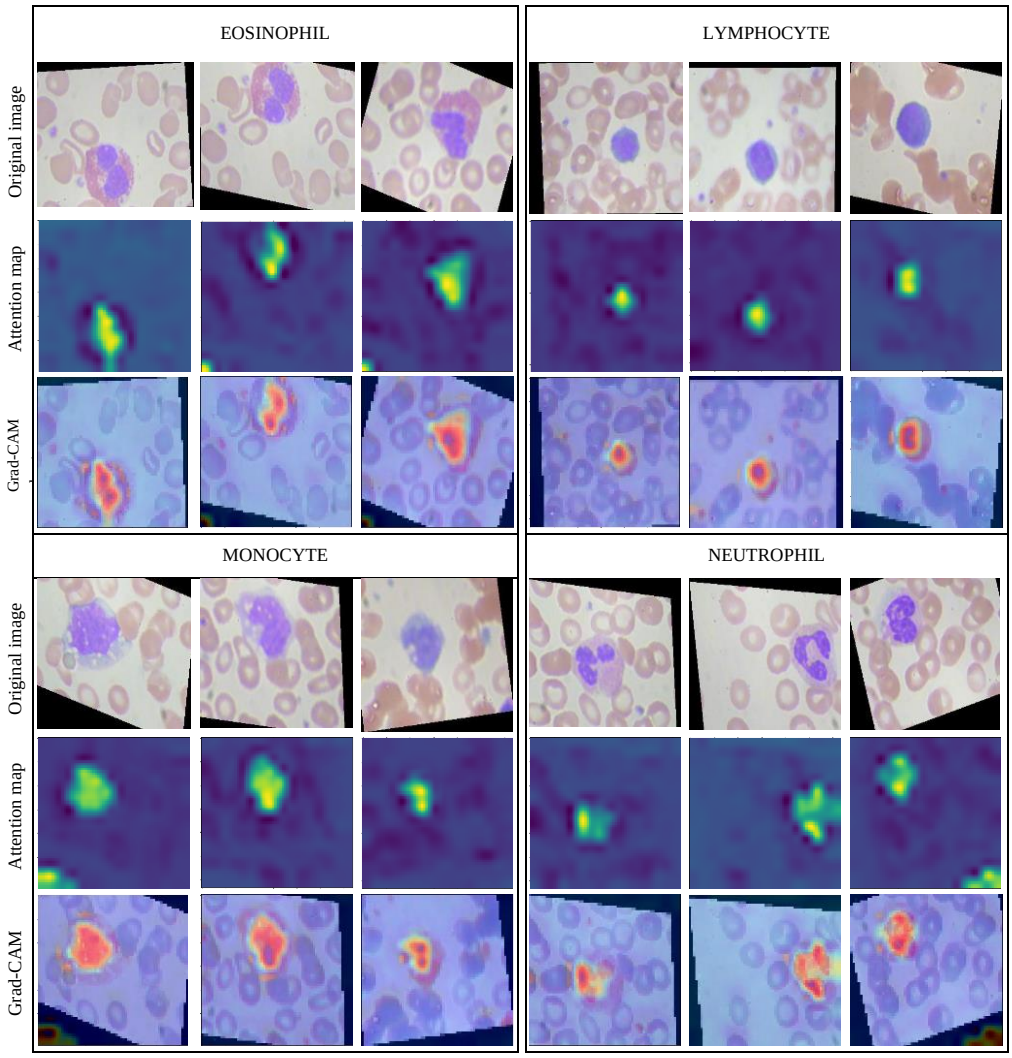


Fig. 8. Each column contains the original WBC image, its attention map and its Grad-CAM heatmap, respectively.

Table IV. Performance comparison of state-of-the-art methods on the same dataset.

Method	Authors	Year	Epoch (#)	Train. Acc. (%)	Valid. Acc. (%)
CNN+RNN(LSTM)	G. Liang et al. [15]	2018	70	-	90.79
LeNet-5	M. Sharma et al. [54]	2019	50	87.06	84.3
Customized CNN	D. Omar, et al [55]	2020	50	-	89.5
Customized CNN	V. Ranga, et al [56]	2020	20	98.1	95.24
deep learning with canonical correlation analysis	A.M.Patila, et al [57]	2020	-	-	95.89
Customized CNN	M. B. Khan et al. [58]	2021	50	98.29	94.82
Extreme learning	A. Khan, et al [17]	2021	-	99.99	99.12
Customized CNN	F Bozkurt [59]	2021	100	-	94
Feature Fusion on transfer learning	S.Parayil,et al [60]	2022	-	89.75	-
SVM by selecting SqueezeNet and LIME properties	E.Başaran [61]	2022	-	-	95.88
Customized CNN	C.Cheuque, et al [62]	2022	-	98.4	98.3
Lekovit (Proposed method)	-	2023	50	100	99.04

Also, the proposed model was also analyzed using Grad-CAM to determine important parts of the input images which are effective for classification decisions of the network. To better interpret the decision of our model, we need to determine which feature in our inputs had the highest contribution to that decision. We created Grad-CAM heatmaps for the last layer of normalization in our model. In theory, the heat map of the last layer will provide the most accurate visual description of the object classified by the model. In terms of gradient mathematics, we realize the importance of all successive feature maps leading up to the last layer.

Cells are distinguished by different characteristics, such as the number of nucleus lobes, the shape of the nucleus, and the color of the cytoplasm. More useful information for classification of cells often comes from the nucleus of leukocytes. The Lymphocyte nucleus is shaped like a “U” and has a multilobed nucleus (typically 3 or 4 lobes) and these lobes may overlap.

The number of lobes can increase according to the age of the cell. Eosinophils can also be identified by their large granules, and the eosinophil nucleus often has 2 lobes. Monocytes are the largest type of WBC. They contain only one nucleus, which is rarely lobed. The shape of the nucleus in monocytes is often curved or kidney-shaped. The visualized attentions almost completely cover the target objects [63]. The Grad-CAM plots, shown in Fig. 8, use feature maps produced by the last layer of the network. Based on these figures, the Grad-CAM and Attention Map of each class of 4 different leukocytes show the focus of the model on the desired cell, in the center of the image, and not on the background and red cells. Grad-CAM shows that this model focuses more on the nucleus, but also considers granules in only the eosinophil subset. We show that the proposed model can outperform other transformer-based models while significantly reducing computational costs and memory constraints.

The observed differences in the results between the model with MLP-Layers and Vision Transformers (ViTs) reveal notable insights. The stability in training observed in the model with MLP-Layers is attributed to the replacement of self-attention with linear layers.

A crucial factor contributing to the superior performance of ViTs is the utilization of larger images. Larger images, in the context of attention mechanisms, accommodate more information and result in longer embedding sequences. However, the advantage of larger images comes with computational costs, as it increases the number of parameters and slows down network training. The proposed model takes into consideration the need for optimal results on smaller images.

The introduced inductive bias contributes to the model's efficiency in learning and accuracy in results. Unlike self-attention, where weights for aggregating information from other patches are data-dependent through queries and keys, the proposed method employs weights that are not data-dependent but based on the absolute positions of patches. This design choice not only enhances efficiency but also addresses challenges related to rare events and limited datasets, showcasing the model's potential applicability in diverse domains with varying data constraints.

According to the aforementioned, focusing on efficiency, simplicity in structure, low computational cost, fewer required parameters, end-to-end training without any preprocessing, and successful performance on small datasets unlike ViT-based models are the advantages of this study. Also, the proposed model outperformed the other models with no attention layer (i.e., MHSA) by applying MLPs. However, it's essential to acknowledge a certain uncertainty regarding its generalizability. The model's assessment is specific to the dataset used, and questions arise about its reliability when faced with larger datasets. Large datasets can potentially help mitigate overfitting, a common concern in machine learning. However, the architecture and regularization techniques used in the MLP play a crucial role in controlling overfitting.

6. Conclusion

The accurate diagnosis of various pathologies, including leukemia, depends on the classification WBCs. While various biological techniques are available to identify leukocytes, automated microscopic examination of blood smears remains crucial in confirming diagnoses. This paper introduces a high-performing network for

classifying four subtypes of leukocytes. The proposed model outperformed state-of-the-art CNN models in cell type classification and achieved exceptional accuracy in precise four-class classification, especially with small image sizes. The exploration of image transformer models is a burgeoning research area, and we have discussed various ViT models in this context. Our analysis, based on the ViT structure for WBC classification, constitutes a significant contribution to our study. As indicated by the results, the proposed architecture exhibited satisfactory performance across all classes, prompting the exploration of effective regions in the image for detecting WBC classes using the Leukovit model. This led to the consideration of the Grad-CAM approach. Our future endeavors include gaining insights into inductive biases within various features and their role in generalization. Additionally, exploring the applicability of such a design on other datasets would be beneficial. We anticipate that future research outcomes will surpass the existing models based on both convolution and attention structures.

7. References

[1] W. King, K. Toler, J. Woodell-May, Role of white blood cells in blood-and bone marrow-based autologous therapies, *BioMed research international*, 2018, 2018.

[2] K. Almezhghwi, S. Serte, Improved classification of white blood cells with the generative adversarial network and deep convolutional neural network, *Computational Intelligence and Neuroscience*, 2020, 2020.

[3] S. Shafique, S. Tehsin, Computer-aided diagnosis of acute lymphoblastic leukaemia, *Computational and mathematical methods in medicine*, 2018, 2018.

[4] V.N. Murthy, S. Maji, R. Manmatha, Automatic image annotation using deep learning representations, Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 603-606, 2015.

[5] K.A.A. Daqqa, A.Y. Maghari, W.F. Al Sarraj, Prediction and diagnosis of leukemia using classification algorithms, 2017 8th international conference on information technology (ICIT), IEEE, pp. 638-643, 2017.

[6] D.M. Ushizima, A.C. Lorena, A. De Carvalho, Support vector machines applied to white blood cell recognition, Fifth International Conference on Hybrid Intelligent Systems (HIS'05), IEEE, pp. 6 pp., 2005.

[7] X. Zheng, Y. Wang, G. Wang, J. Liu, Fast and robust segmentation of white blood cell images by self-supervised learning, *Micron*, 107 55-71, 2018.

[8] سزاوار, فرسی, حسن, محمدزاده, بازیابی تصویر مبتنی بر محتوا با استفاده از شبکه‌های عصبی کانولوشن عمیق, مجله مهندسی برق دانشگاه تبریز, ۴۸ ۱۵۹۵-۱۶۰۳, ۲۰۱۹.

[9] P.P. Banik, R. Saha, K.-D. Kim, An automatic nucleus segmentation and CNN model based classification method of white blood cell, *Expert Systems with Applications*, 149 113211, 2020.

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, 25, 2012.

[11] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[14] M.Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, V.K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, *arXiv preprint arXiv:1802.06955*, 2018.

[15] G. Liang, H. Hong, W. Xie, L. Zheng, Combining convolutional neural network with recursive neural network for blood cell image classification, *IEEE access*, 6 36188-36197, 2018.

[16] P. Tang, H. Wang, S. Kwong, G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition, *Neurocomputing*, 225 188-197, 2017.

[17] A. Khan, A. Eker, A. Chefranov, H. Demirel, White blood cell type identification using multi-layer convolutional features with an extreme-learning machine, *Biomedical Signal Processing and Control*, 69 102932, 2021.

[18] A. Darvish, S. Shamekhi, A hybrid multi-scale CNN-LSTM deep learning model for the identification of protein-coding regions in DNA sequences, *TABRIZ JOURNAL OF ELECTRICAL ENGINEERING*, 52 137-146, 2022.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*, 2020.

[20] Y. Dong, Z. Jiang, H. Shen, W.D. Pan, L.A. Williams, V.V. Reddy, W.H. Benjamin, A.W. Bryan, Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells, 2017 IEEE EMBS international conference on biomedical & health informatics (BHI), IEEE, pp. 101-104, 2017.

[21] M. Imran Razzak, S. Naz, Microscopic blood smear segmentation and classification using deep contour aware CNN and extreme machine learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 49-55, 2017.

[22] C.L. Chen, A. Mahjoubfar, L.-C. Tai, I.K. Blaby, A. Huang, K.R. Niazi, B. Jalali, Deep learning in label-free cell classification, *Scientific reports*, 6 1-16, 2016.

[23] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803, 2018.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, European conference on computer vision, Springer, pp. 213-229, 2020.

[25] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, *Advances in Neural Information Processing Systems*, 32, 2019.

[26] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.-C. Chen, Axial-deeplab: Stand-alone axial-attention for

panoptic segmentation, European Conference on Computer Vision, Springer, pp. 108-126, 2020.
[27] I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q.V. Le, Attention augmented convolutional networks, Proceedings of the IEEE/CVF international conference on computer vision, pp. 3286-3295, 2019.
[28] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3588-3597, 2018.
[29] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, *arXiv preprint arXiv:2006.03677*, 2020.
[30] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306*, 2021.
[31] F. Mehboob, A. Rauf, R. Jiang, A.K.J. Saudagar, K.M. Malik, M.B. Khan, M.H.A. Hasnat, A. AlTameem, M. AlKhathami, Towards robust diagnosis of COVID-19 using vision self-attention transformer, *Scientific Reports*, 12 1-12, 2022.
[32] X. Qu, H. Lu, W. Tang, S. Wang, D. Zheng, Y. Hou, J. Jiang, A VGG attention vision transformer network for benign and malignant classification of breast ultrasound images, *Medical Physics*, 49 5787-5798, 2022.
[33] Y. Dai, Y. Gao, F. Liu, Transmed: Transformers advance multi-modal medical image classification, *Diagnostics*, 11 1384, 2021.
[34] T. Wang, J. Lan, Z. Han, Z. Hu, Y. Huang, Y. Deng, H. Zhang, J. Wang, M. Chen, H. Jiang, O-Net: a novel framework with deep fusion of CNN and transformer for simultaneous segmentation and classification, *Frontiers in Neuroscience*, 16, 2022.
[35] P. Cho, S. Dash, A. Tsaris, H.-J. Yoon, Image transformers for classifying acute lymphoblastic leukemia, Medical Imaging 2022: Computer-Aided Diagnosis, SPIE, pp. 633-639, 2022.
[36] S. Tripathi, A.I. Augustin, R. Sukumaran, S. Dheer, E. Kim, HematoNet: Expert level classification of bone marrow cytology morphology in hematological malignancy with deep learning, *Artificial Intelligence in the Life Sciences*, 2, 2022.
[37] Z. Dai, H. Liu, Q.V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, *Advances in Neural Information Processing Systems*, 34 3965-3977, 2021.
[38] P. Huang, J. Wang, J. Zhang, Y. Shen, C. Liu, W. Song, S. Wu, Y. Zuo, Z. Lu, D. Li, Attention-aware residual network based manifold learning for white blood cells classification, *IEEE Journal of Biomedical and Health Informatics*, 25 1206-1214, 2020.
[39] Z. Wang, J. Xiao, J. Li, H. Li, L. Wang, WBC-AMNet: Automatic classification of WBC images using deep feature fusion network based on focalized attention mechanism, *Plos one*, 17 e0261848, 2022.
[40] O. Katar, O. Yildirim, An Explainable Vision Transformer Model Based White Blood Cells Classification and Localization, 2023.
[41] S. Chen, S. Lu, S. Wang, Y. Ni, Y. Zhang, Shifted Window Vision Transformer for Blood Cell Classification, *Electronics*, 12 2442, 2023.
[42] S.M. Dipto, M.T. Reza, M.N.J. Rahman, M.Z. Parvez, P.D. Barua, S. Chakraborty, An XAI Integrated Identification System of White Blood Cell Type Using Variants of Vision Transformer, International Conference on Interactive Collaborative Robotics, Springer, pp. 303-315, 2023.
[43] D. Parthasarathy, WBC-classification, Google License, Online; accessed 2022, 2017.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems*, 30, 2017.
[45] S. d'Ascoli, H. Touvron, M.L. Leavitt, A.S. Morcos, G. Biroli, L. Sagun, Convit: Improving vision transformers with soft convolutional inductive biases, International Conference on Machine Learning, PMLR, pp. 2286-2296, 2021.
[46] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, Mlp-mixer: An all-mlp architecture for vision, *Advances in Neural Information Processing Systems*, 34 24261-24272, 2021.
[47] M. VASOUJOUYBARI, E. Ataie, M. Bastam, An MLP-based Deep Learning Approach for Detecting DDoS Attacks, *TABRIZ JOURNAL OF ELECTRICAL ENGINEERING*, 52 195-204, 2022.
[48] G. Huang, Y. Sun, Z. Liu, D. Sedra, K.Q. Weinberger, Deep networks with stochastic depth, European conference on computer vision, Springer, pp. 646-661, 2016.
[49] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, H. Shi, Escaping the big data paradigm with compact transformers, *arXiv preprint arXiv:2104.05704*, 2021.
[50] T. Dozat, Incorporating nesterov momentum into adam, 2016.
[51] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, M.-M. Cheng, Delving deep into label smoothing, *IEEE Transactions on Image Processing*, 30 5984-5996, 2021.
[52] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.
[53] F. Wu, A. Fan, A. Baevski, Y.N. Dauphin, M. Auli, Pay less attention with lightweight and dynamic convolutions, *arXiv preprint arXiv:1901.10430*, 2019.
[54] M. Sharma, A. Bhave, R.R. Janghel, White blood cell classification using convolutional neural network, Soft Computing and Signal Processing, Springer, pp. 135-143, 2019.
[55] O. Dekhil, Computational techniques in medical image analysis application for white blood cells classification, 2020.
[56] V. Ranga, S. Gupta, P. Agrawal, J. Meena, Pathological analysis of blood cells using deep learning techniques, *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 15 397-403, 2022.

[57] A. Patil, M. Patil, G. Birajdar, White blood cells image classification using deep learning with canonical correlation analysis, *Irbm*, 42 378-389, 2021.
[58] M.B. Khan, T. Islam, M. Ahmad, R. Shahrior, Z.N. Riya, A CNN Based Deep Learning Approach for Leukocytes Classification in Peripheral Blood from Microscopic Smear Blood Images, Proceedings of International Joint Conference on Advances in Computational Intelligence, Springer, pp. 67-76, 2021.
[59] F. BOZKURT, Classification of Blood Cells from Blood Cell Images Using Dense Convolutional Network, *Journal of Science, Technology and Engineering Research*, 2 81-88, 2021.
[60] S. Parayil, J. Aravinth, Transfer Learning-based Feature Fusion of White Blood Cell Image Classification, 2022 7th International Conference on Communication and Electronics Systems (ICCES), IEEE, pp. 1468-1474, 2022.
[61] E. Başaran, Classification of white blood cells with SVM by selecting SqueezeNet and LIME properties by mRMR method, *Signal, Image and Video Processing*, 1-9, 2022.
[62] C. Cheuque, M. Querales, R. León, R. Salas, R. Torres, An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification, *Diagnostics*, 12 248, 2022.
[63] S. Standring, Gray's anatomy e-book: the anatomical basis of clinical practice, Elsevier Health Sciences2021.