Article

An Explainable Vision Transformer Model Based White Blood Cells Classification and Localization

Oguzhan Katar [1] and Ozal Yildirim [1,*]

[1] Department of Software Engineering, Firat University, Elazig 23119, Turkey; okatar@firat.edu.tr (O.K.).
* Correspondence: ozalyildirim@firat.edu.tr

Abstract: Blood cell analysis is a crucial diagnostic process in medical practice. In particular, detecting white blood cells (WBCs) is essential for diagnosing of many diseases. The manual screening of blood films is a time-consuming and subjective process, which can lead to inconsistencies and errors. Therefore, automated detection of blood cells can improve the accuracy and efficiency of the screening process. In this study, an explainable Vision Transformer (ViT) model was proposed for the automatic detection of WBCs from blood films. The proposed model utilizes the self-attention mechanism to extract relevant features from the input images and leverages transfer learning by incorporating pre-trained model weights to improve its performance. The proposed model achieved a classification accuracy of 99.40% for five distinct types of WBCs and exhibited potential in reducing the time required for manual screening of blood films by pathologists. Upon examination of the misclassified test samples, it was observed that incorrect predictions were correlated with the presence or absence of granules in the cell samples. To validate this observation, the dataset was divided into two classes, namely Granulocytes and Agranulocytes, and a secondary training process was conducted. The resulting ViT model trained for binary classification achieved an accuracy of 99.70%, recall of 99.54%, precision of 99.32%, and F-1 score of 99.43% during the test phase. To ensure the reliability of the ViT model's multi-class classification of WBCs, the pixel areas that the model focuses on in its predictions are visualized through the Score-CAM algorithm.

Keywords: Vision Transformers; white blood cells; explainable AI models; deep learning; Score-CAM

1. Introduction

White blood cells (WBCs), also known as leukocytes, play a vital role in the body's immune response [1]. They are produced in the bone marrow and are an essential component of the body's defense system against infection and disease. WBCs are classified as either granulocytes, which possess granules in their cytoplasm, or agranulocytes, which lack granules [2]. Granulocytes include neutrophils, eosinophils, and basophils. Agranulocytes include lymphocytes and monocytes. Neutrophils, the most common type of WBCs, are the first to arrive at the site of an infection and are responsible for engulfing and destroying bacteria and other foreign particles [3]. Lymphocytes include T and B cells, which are responsible for cell-mediated and antibody-mediated immunity, respectively [4]. T cells help to identify and attack infected or cancerous cells, while B cells produce antibodies that can neutralize pathogens. Monocytes mature into macrophages, which consume and destroy microorganisms and debris [5]. Eosinophils play a role in the body's response to parasitic infections and allergies [6]. Basophils release histamine and other inflammatory chemicals in response to allergens and other stimuli [7].

WBCs number can increase in response to infection, inflammation, or other stimuli. An abnormal increase in WBCs count is called leukocytosis, while a decrease is called leukopenia [8]. Abnormalities in WBCs counts can indicate a variety of medical conditions, including infections, cancers, and immune system disorders. A complete blood count (CBC) test, which isolates WBCs from a blood sample and studies their number and appearance under a microscope, is commonly used as part of a routine medical check-up [9].

 doi:10.20944/preprints202306.1106.v1

The utilization of artificial intelligence-based systems to automatically classify WBCs in a CBC test can provide several benefits. Firstly, it can enhance the accuracy and consistency of the results by removing the subjective nature of manual classification. Manual classification of WBCs is a complex and time-consuming task that requires a high level of expertise and experience [10]. However, with AI-based systems, the process can be automated, and the results can be more consistent, as the system does not get tired or make mistakes due to human error. Secondly, it can also increase the efficiency of the process by reducing the time required for manual classification. This can be especially beneficial in high-volume settings, such as in hospital laboratories, where a large number of CBC tests are performed daily. Automated classification can also help to reduce the workload of laboratory staff, allowing them to focus on other tasks. Furthermore, AI-based systems can also provide additional information that may not be visible to the human eye, such as detecting rare or abnormal cells, which can assist in the diagnosis of certain blood disorders.

In recent years, there has been a growing interest in using machine learning and artificial intelligence to automate the analysis of WBCs. Deep learning algorithms have been employed to develop automated systems that can identify and segment WBCs in digital images of blood samples, providing a faster and more accurate alternative to manual analysis [11]. To perform WBCs classification using deep learning, a dataset of labeled images is first employed to train a neural network model. The model is subsequently able to make predictions on new images, accurately identifying and classifying various types of WBCs. This approach has demonstrated promising results, with some studies showing the ability to achieve high levels of accuracy and precision in WBCs classification [12–14].

A number of studies have explored the utilization of deep learning for WBCs classification, employing techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to classify various types of WBCs. Cheuque et al. [15] proposed a two-stage hybrid multi-level scheme for efficiently classifying four groups of WBCs (lymphocytes, monocytes, segmented neutrophils, and eosinophils) using a combination of a Faster R-CNN network and parallel CNNs with the MobileNet structure. The proposed model achieved a performance metric of approximately 98.4% in terms of accuracy, recall, precision, and F-1 score. Sharma et al. [16] proposed a deep learning model, specifically the DenseNet121 model, for classifying various types of WBCs in blood cell images. They utilized preprocessing techniques, such as normalization and data augmentation, to optimize the model. The model was evaluated using a dataset from Kaggle containing 12,444 images of various types of WBCs. The results indicated that the model achieved an accuracy of 98.84%, precision of 99.33%, sensitivity of 98.85%, and specificity of 99.61%. Jung et al. [17] proposed a CNN-based method, referred to as W-Net, for WBCs classification, which was evaluated on a large-scale dataset of real images of the five types of WBCs. The proposed method, W-Net, achieved an average accuracy of 97% and demonstrated superior performance compared to other CNN and RNN-based model architectures. The authors also proposed the utilization of Generative Adversarial Networks (GANs) to generate synthetic WBCs images for educational and research purposes. Rustam et al. [18] proposed a hybrid feature set that combines texture and RGB features from microscopic images for classifying various types of WBCs in blood cell images. They utilized a synthetic minority oversampling technique-based resampling to mitigate the influence of imbalanced datasets, which is a common problem in existing studies. The authors also adopted machine and deep learning models for performance comparison using the original dataset, augmented dataset, and oversampled dataset to analyze the performances of the models. The results suggest that a hybrid feature set of both texture and RGB features from microscopic images, yields a high accuracy rate of 97.00% with random forest. Chola et al. [19] proposed a deep learning framework, referred to as BCNet, for the identification of various types of blood cells in an eight-class identification scenario. The proposed BCNet framework is based on transfer learning with a CNN. The dependability and viability of BCNet were established through exhaustive experiments consisting of five-fold cross-validation tests. The performance of BCNet was compared with state-of-the-art deep learning models such as DenseNet, ResNet, Inception, and MobileNet. The BCNet framework achieved the highest performance with the RMSprop optimizer, with 98.51% accuracy and 96.24% F-1 score. CNN-based

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

architectures have been utilized in most studies published in the literature. Nonetheless, CNNs exhibit limitations in managing variable length sequences and capturing long-term dependencies due to the convolution and pooling operations which can result in information loss. Consequently, researchers are in search of alternative methods to CNN architectures.

Recently, image transformer architectures have been applied to image classification [20]. These architectures, designed to process sequential data, have demonstrated exceptional results in image classification tasks. In contrast to traditional CNNs, which utilize spatial convolutions to extract features from images, image transformers employ self-attention mechanisms to capture relationships between different regions of an image [21]. This allows them to learn global, contextual features that are valuable for classification tasks. Recent studies have shown that image transformers can achieve state-of-the-art performance on various image classification benchmarks [22].

This paper proposes an explainable Vision Transformer (ViT) model for computer-assisted automatic WBCs classification. The ViT model, pre-trained for a distinct task, was fine-tuned to classify WBCs. The model was trained on a public set consisting of 16,633 samples, and its performance was evaluated. The results showed that the model achieved high accuracy rates in both multi-class and binary classification of WBCs. To ensure the proposed method can be confidently applied in clinical settings, the pixel areas upon which the model focuses its predictions have been visualized.

The main contributions of this study can be summarized as follows:

- WBCs classification is achieved without the requirement for any preprocessing or convolutional processes.
- The effectiveness of using ViT for WBC classification, which can potentially outperform traditional CNN architectures.
- The proposed ViT-based method achieves high accuracy rates in both multi-class and binary classification of WBCs, which is crucial for accurate disease diagnosis and treatment.
- An explainable method for WBCs classification, which increases the transparency and trustworthiness of the model's decision-making process.
- Visiualization of the pixel areas upon which the model focuses its predictions, which can facilitate the adoption of the proposed method in clinical settings.

2. Materials and Methods

This study presents a cutting-edge deep learning model for the accurate identification and classification of various subtypes of WBCs using the ViT model. This approach utilizes a dataset of images of WBCs to train the model, enabling it to accurately classify the cells into different subtypes with high accuracy. The model takes an image of a WBC as input and employs its deep learning capabilities to output a prediction of the WBC's subtype. In this prediction process, the Score-CAM algorithm is utilized to demonstrate which regions within the image influence the classification decision. The proposed approach is illustrated in the diagram in Figure 1.

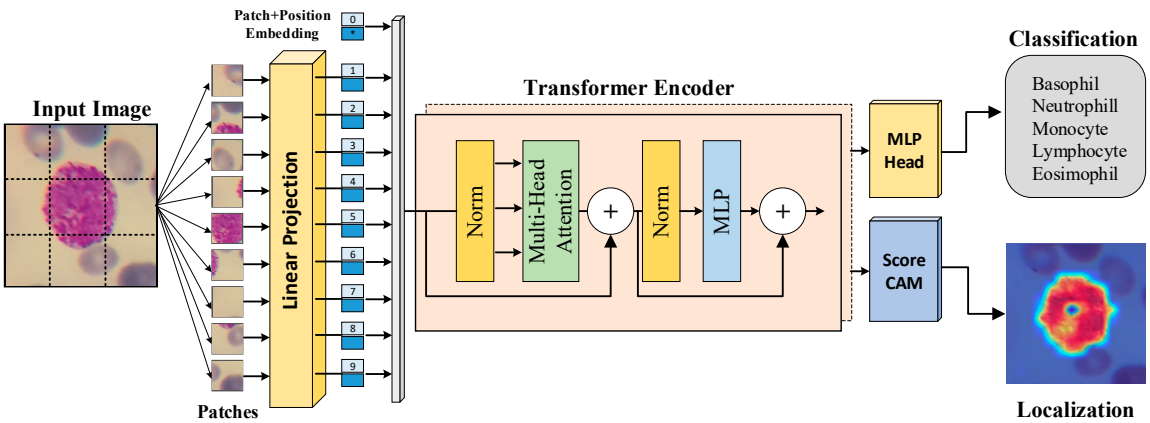

Figure 1. Block representation of the proposed WBC classification and localization method.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

2.1. WBCs Dataset

In this study, we utilized the publicly available dataset Raabin-WBC [23], created from 72 regular peripheral blood films. The samples in the dataset were stained using the Giemsa technique and viewed at 100x magnification using two microscopes. Additionally, smartphones equipped with an adapter designed and fabricated via 3D printing were employed to capture images by mounting the phone to the microscope ocular lens. A total of approximately 23,000 images were acquired and processed utilizing a color filter and a Faster RCNN network to extract WBCs. The data was further cleaned to eliminate duplicate cell images and a comprehensive labeling process was undertaken to accurately determine the cell types. The classes in the dataset obtained after the labeling process and the number of samples they contain are presented in Figure 2.

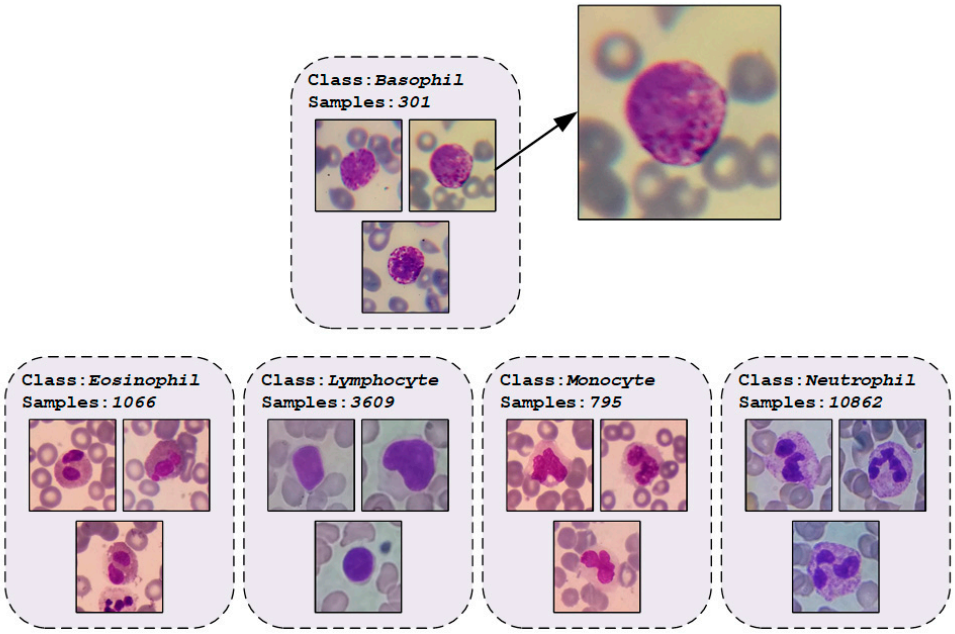

Figure 2. An illustration of the dataset with number of classes and some class images.

Upon examination of the class distributions of the Raabin-WBC dataset samples, it becomes apparent that the dataset is unbalanced. While this may initially be perceived as a problem for use in model training, it is in fact more suitable for real-world data. This is because the distribution of WBCs in the population is not equal, with certain types of WBCs being more prevalent than others. An unbalanced dataset accurately reflects this reality and allows more accurate diagnosis and treatment. Conversely, a balanced dataset may lead to oversampling of less common WBCs, resulting in skewed results and inaccurate conclusions. Additionally, a balanced dataset may lead to oversampling of certain types of WBCs, thereby biasing the training of a machine learning model and resulting in poor performance on real-world data. For this reason, no data augmentation method was applied to the dataset samples. 80% of the dataset samples were reserved for training, 10% for validation, and 10% for testing. The number of data used at each stage as a result of the splitting process is presented in Table 1.

Table 1. Data distribution at each stage of the splitting process for training, validation, and testing.

Phase	Basophil	Eosinophil	Lymphocyte	Monocyte	Neutrophil	Total
Training	241	852	2,887	637	8,688	**13,305**
Validation	30	107	361	79	1,087	**1,664**
Testing	30	107	361	79	1,087	**1,664**
Total	**301**	**1,066**	**3,609**	**795**	**10,862**	**16,633**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

2.2. Transformer-based Image Classification Method

Transformers are a type of neural network architecture that have been widely utilized in natural language processing (NLP) tasks, such as language translation and language modeling [24]. The transformer architecture is based on the concept of self-attention, which enables the model to weigh the importance of different parts of the input when making predictions. Self-attention is implemented through an attention mechanism, which calculates a weighted sum of the input values based on their relationships to a particular position or query [25]. This mechanism allows the transformer to learn relationships between different parts of the input, which is particularly beneficial in NLP tasks where the meaning of a word or phrase depends on its context. Furthermore, the transformer architecture also employs an encoder-decoder structure [26]. The encoder takes in the input and generates a set of hidden representations, which are then passed to the decoder to generate the output. Both the encoder and decoder are composed of multiple layers, each containing multiple self-attention mechanisms. This enables the model to learn and extract information from the input at multiple levels of abstraction, which is essential for understanding the meaning of the input. The block diagram depicting the multi-head self-attention (MHA) mechanism employed in this study is presented in Figure 3.

In the context of MHA, each patch of the input vector undergoes a self-attention process, wherein it is transformed into three distinct vectors: query (Q), key (K), and value (V), through the use of weight matrices [27]. The saliency of each patch is determined by calculating the dot product of its query and key vectors, producing a score matrix. The softmax activation function is then applied to the score matrix, generating attention weights. Finally, the attention weights are multiplied with the value vector, resulting in the self-attention output. After the self-attention mechanism to each patch and the computation of resulting self-attention matrices, they are aggregated and processed by a linear layer. Subsequently, a regression head is employed to generate the final output of the MHA mechanism.

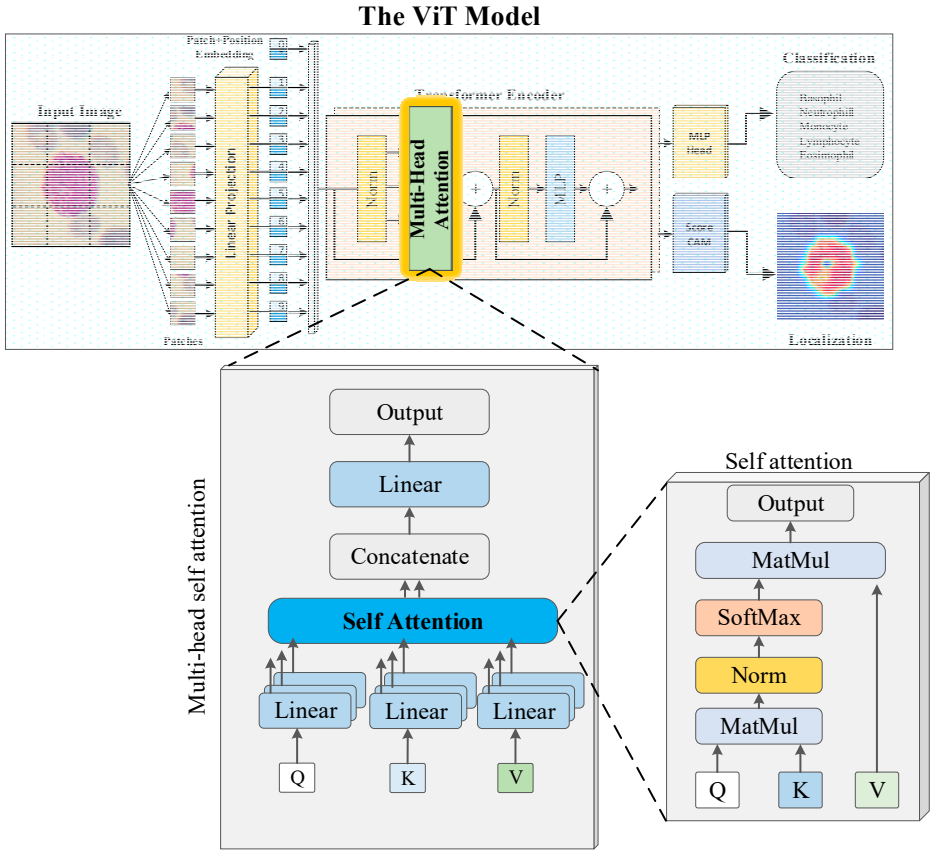

Figure 3. The block diagram depicts the multi-head self-attention (MHA) mechanism.

 doi:10.20944/preprints202306.1106.v1

Dosovitskiy et al. [28] proposed the ViT model, which is based on transformers and comprises self-attention blocks and multilayer perceptron (MLP) networks. ViTs are similar to standard transformers, but are designed to handle images as input. Specifically, ViTs divide images into smaller non-overlapping patches and then use a transformer architecture to process each patch separately [29]. This enables the model to learn and extract information from the images at multiple levels of abstraction, similar to how the transformer architecture works for NLP tasks. A key feature of ViTs is their ability to handle images as input without the need for a pre-processing step, such as applying convolutional layers. Instead, ViT divide images into smaller non-overlapping patches and then use a transformer architecture to process each patch separately. A particularly noteworthy aspect of ViT is their use of a linear projection to reduce the dimensionality of image patches before feeding them into the transformer network. This approach may seem counterintuitive as it reduces the information in the image patches. However, this linear projection serves a crucial role in the ViT architecture and has several benefits. It allows for a more computationally efficient model. By reducing the dimensionality of the image patches, the model can be trained on larger images without requiring a large amount of computational resources, making ViT more accessible to researchers and practitioners and allowing for more widespread usage and experimentation.

3. Experiments

3.1. Experimental Setup

In this study, a pre-trained 'vit-base-patch16-224' model with the ImageNet-21k (14 million images, 21,843 classes) dataset was utilized due to the high cost of training a ViT model from scratch with random weights. The output layer of the model was modified to match the number of classes in the dataset. As the default input size of the pre-trained ViT model is 224×224, all of the dataset samples were resized to this size. The model was then trained on the samples allocated for the training set using the AdamW optimizer and the CrossEntropyLoss function. During training, the pre-trained weights of the model were constantly updated to better fit the set of WBCs. The batch size value was kept constant at 16, and the learning rate at 0.00002. The maximum number of epochs was set to 100, but an early stopping function was defined to monitor the validation loss value. If there was no decrease in the validation loss for five consecutive epochs, the early stopping function would terminate the model training, and the weights of the epoch with the highest classification ability would be recorded. After training, the model was evaluated on the test set to measure its performance on the new dataset.

3.2. Performance Evaluation Metrics

In the evaluation of deep learning models for classification tasks, confusion matrix-based metrics are commonly employed. A confusion matrix illustrates the correspondence between the model's predicted class label for a given input image and the true class label of that image. In artificial intelligence-based classification studies, various situations may arise in the output layer of the model. However, these situations can be encapsulated by four metrics: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Figure 4 illustrates the placement of these metrics in the fields of multi-class and binary classification studies, with respect to a random class.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

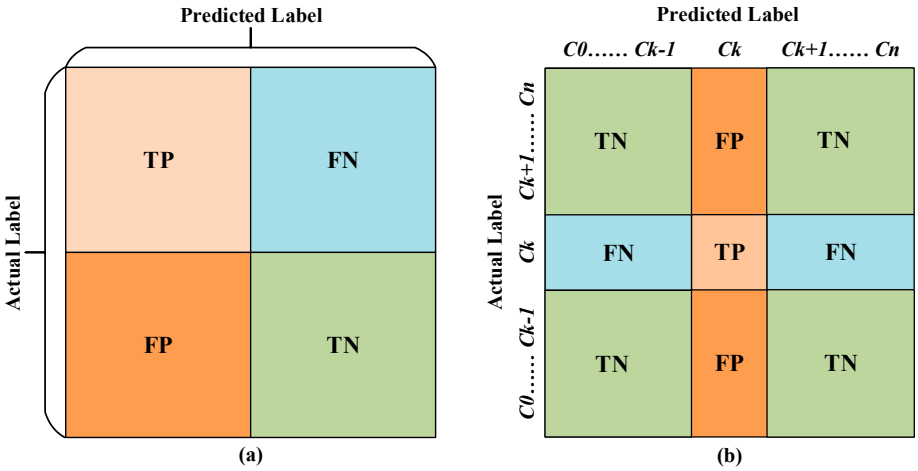

Figure 4. Confusion matrices for a) binary-class and b) multi-class scenarios.

Evaluation of an image classifier deep learning model can be accomplished utilizing a variety of performance metrics, including accuracy, precision, recall, and F-1 score. Accuracy, which is the proportion of correct predictions made by the model, is calculated as the number of correct predictions divided by the total number of predictions. Recall, which is a measure of the model's ability to detect all positive instances, is calculated as the number of true positive predictions divided by the total number of actual positive instances. Precision is a measure of the model's ability to correctly identify positive instances, and is determined by dividing the number of true positive predictions by the total number of positive predictions. The F-1 score, which is the harmonic mean of precision and recall, is a useful metric for balancing precision and recall when they are in conflict. The mathematical definitions for these measures are provided below.

$$\text{Accuracy (Acc)} = (TP + TN)/(TP + FP + FN + TN) \quad (1)$$

$$\text{Recall (Rec)} = TP/(TP + FN) \quad (2)$$

$$\text{Precision (Pre)} = TP/(TP + FP) \quad (3)$$

$$\text{F-1 Score (F1)} = (2\times(Pre\times Rec))/((Pre + Rec)) \quad (4)$$

3.3. Results

Using a pre-trained model can have several implications for the training phase. The pre-trained model provides a starting point that is already optimized, reducing training time and saving computational resources compared to training a model from scratch. Pre-trained models often perform better than models trained from scratch as they have already learned general representations from a large dataset, providing a strong foundation for fine-tuning the target task.

Utilizing the pre-trained models, the training duration for the ViT model was set to a maximum of 100 epochs. However, the activation of the early stopping function resulted in the cessation of training at the 10th epoch, and the weights were saved in the '.pth' format. The accuracy and loss graphs generated during the training phase of the model are illustrated in Figure 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

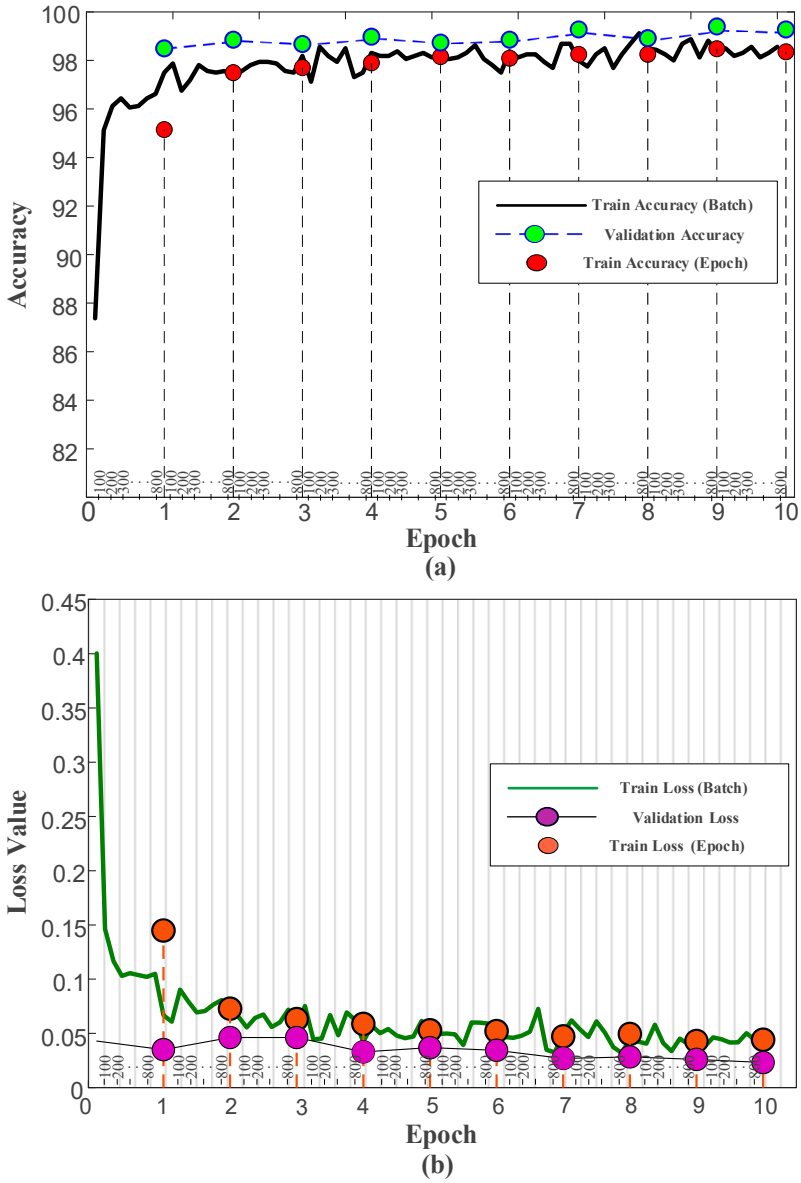

Figure 5. The ViT model performance curves during training, (a) Accuracy values and (b) Loss values.

The performance of the proposed ViT model on both the training and validation sets was consistent with the expectations based on the pre-trained models and reached the desired level. However, test images were utilized to assess the model's capability to generalize to new and unseen scenarios, which is crucial in practical applications. The confusion matrix and recall, precision, and F-1 score graphs obtained from the testing phase are presented in Figure 6. The ViT model achieved an accuracy of 99.40%, with only 10 misclassifications on a set of 1,664 test images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

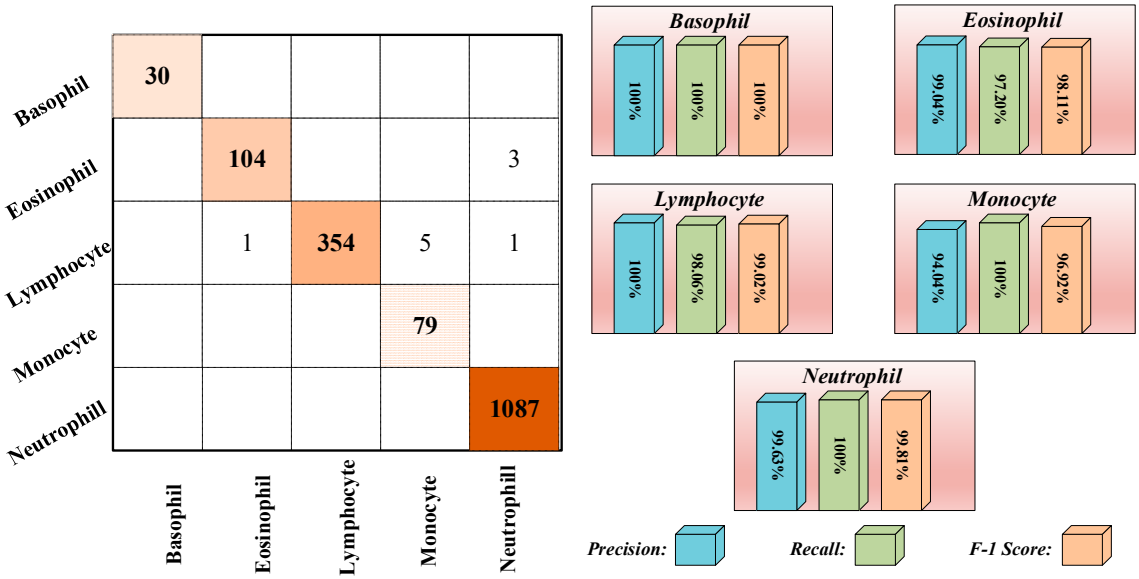

Figure 6. The confusion matrix and some graphs of metrics for multi-class for test dataset.

Upon examination of the test results, it was observed that exceptional performance scores were obtained for each class. However, the precision value for the Monocyte class was lowerthan the other classes. The primary cause for this is the overestimation of samples as Monocytes, despite their actual classification as Lymphocytes. These two classes of samples share similar visual features and do not contain granules. WBCs are commonly classified into five distinct types. However, based on their morphological characteristics, it is feasible to divide the cells into two groups: Granulocytes and Agranulocytes. The cells that constitute Granulocytes and Agranulocytes are depicted in Figure 7.

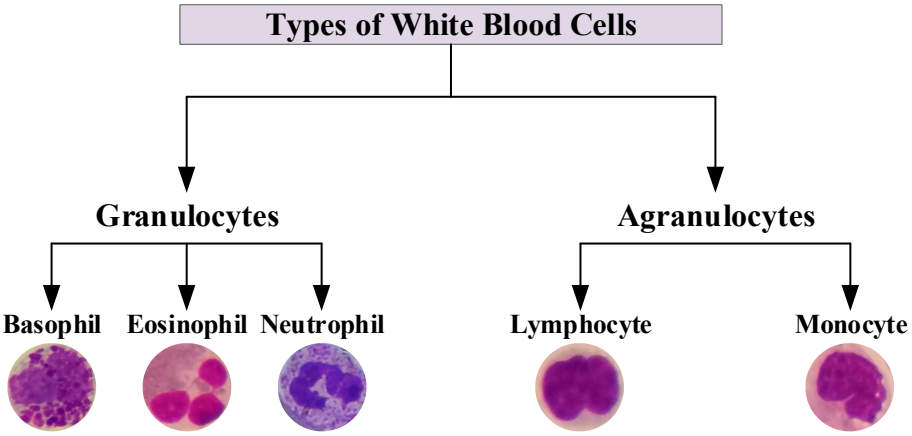

Figure 7. A schematic representation for types of white blood cells.

Samples of datasets for binary classification of WBCs are annotated under the classification scheme depicted in Figure 7. The datasets were trained using the same data split ratios and hyperparameters as in the 5-class classification task. The ViT model demonstrated the capability to differentiate between Granulocytes and Agranulocytes, achieving an accuracy of 99.75% on the test samples. The results of the testing phase, including the confusion matrix and ROC curve, are presented in Figure 8.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

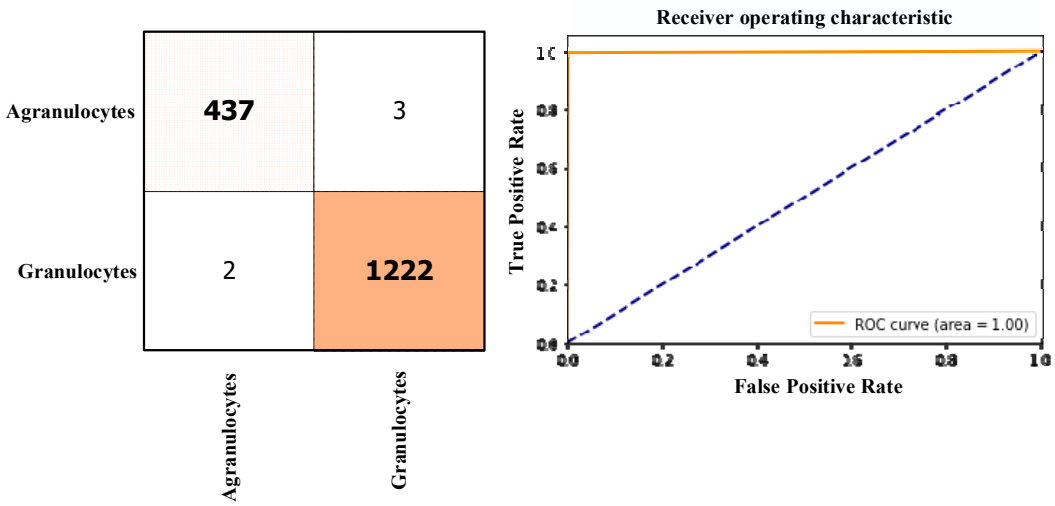

Figure 8. The ViT model performance on Granulocytes and Agranulocytes classification.

The traditional view of deep learning models has been that they are black boxes due to their complex structure, making it challenging to comprehend the internal workings of the models and how they arrive at a specific decision. However, recent advancements in explainability techniques have altered this perception, enabling us to shed light on the inner workings of deep learning models [30, 31]. One such technique is Score-CAM [32], which allows visualization of the most important features of the input data for the model's decision-making process. With the application of these techniques, deep learning models have transformed from being black boxes to becoming more explainable. This is crucial in applications where the model's decisions have significant consequences, as it enables the building of trust and confidence in the model. The aggregate performance measures, such as accuracy, recall and precision, only provide a general view of the model's performance and do not reveal the underlying mechanisms that drive the model's decisions. On the other hand, Score-CAM-like algorithms offer a way to understand the model's behavior and decision-making process, which is vital for establishing accountability and trust in the model. In this study, the predictions made by the ViT model were explained by utilizing the Score-CAM algorithm, which was used to focus on and highlight the areas of interest. ViT models are characterized by the utilization of self-attention mechanisms, which enable the model to focus on relevant parts of the input image while disregarding irrelevant ones. One of the notable characteristics of ViT models is the output of their layers, which is typically of the shape BATCH×197×192. In this dimension, the first element represents the class token, while the remaining 196 elements represent the 14×14 patches in the image. The class token is employed in making the final classification, and the remaining 196 elements can be viewed as a 14×14 spatial image with 192 channels. To integrate the Score-CAM algorithm into vision transformers, it is necessary to reshape them into 2D spatial images. This can be accomplished by passing a reshape transform function to the CAM constructor. By doing so, it is possible to visualize the regions of the input image that are most crucial for the model's final classification. The final classification is based on the class token computed in the last attention block. As a result, the output of the model is not affected by the 14×14 channels in the last layer, and the gradient of the output with respect to them is 0. Therefore, it is recommended to choose any layer prior to the final attention block when generating CAMs to better understand the model's behavior. The operational framework of the explainable ViT model proposed in this study, along with its constituent layers, is depicted in Figure 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

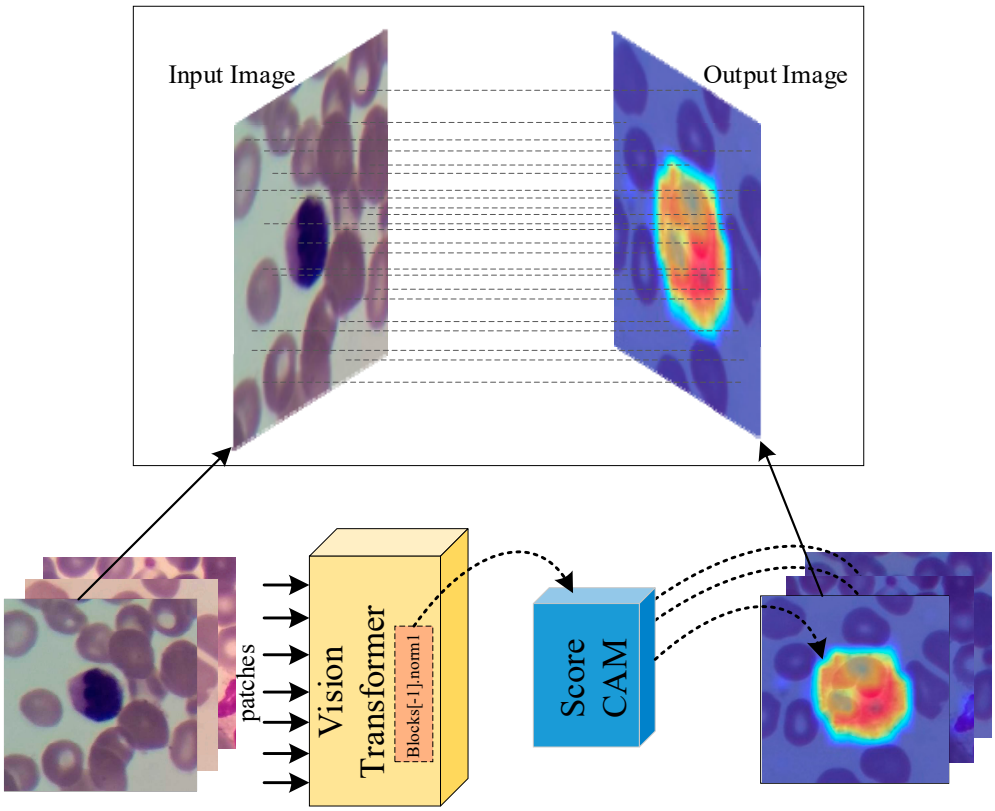

Figure 9. The operational framework of the explainable ViT model.

The areas upon which the model correctly focuses its predictions on the test images are presented in Figure 10. The regions of focus identified by the ViT model exhibit a significant overlap with the areas of WBCs.

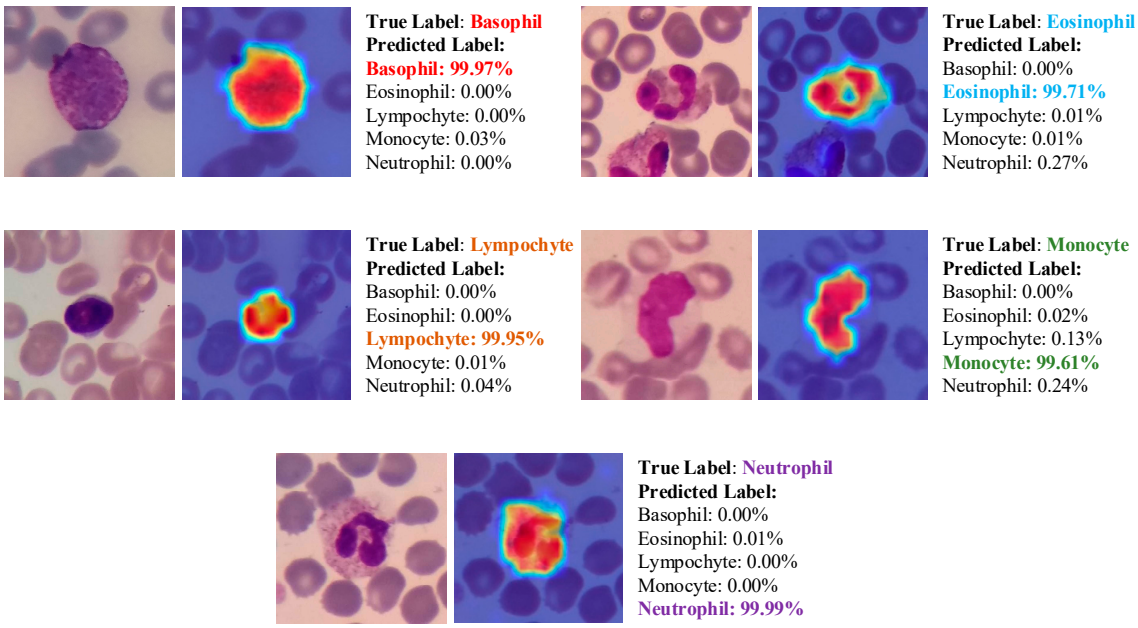

Figure 10. The areas upon which the model correctly focuses its predictions on the test images.

This alignment between the Score-CAM output and the ground truth is a promising indication that the model is effectively learning meaningful features from the input data and utilizing these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1106.v1

features to make accurate predictions. Analysis of the Score-CAM outputs for images misclassified by the ViT model can provide valuable insights into the model's strengths and weaknesses. Figure 11 illustrates a few examples of misclassified images and probabilities.

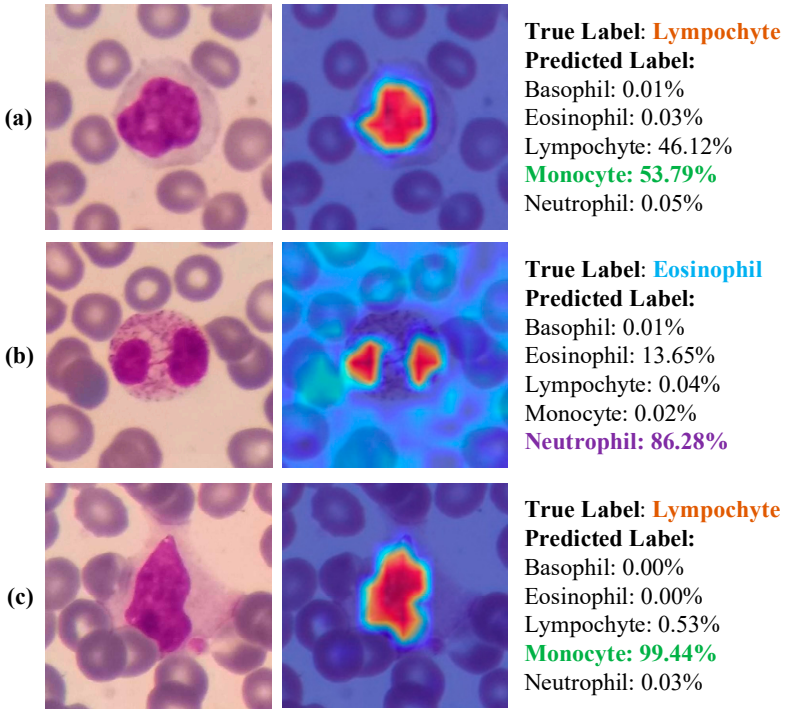

Figure 11. Some examples of misclassified samples and class probabilties during test phase.

In the case of incorrectly predicted images, the model continued to focus on WBCs areas. However, the inaccuracies in the model's predictions could be attributed to visual similarities between cells.

4. Discussion

WBCs classification plays a crucial role in diagnosing of many diseases, including infections and blood-related disorders. However, manual classification of WBCs can be time-consuming and prone to errors due to human subjectivity. Therefore, the use of machine learning algorithms, has the potential to improve the accuracy and efficiency of WBCs classification. Table 2 provides details for a hand-curated selection of research studies on that topic. Tavakoli et al. [33] introduced a novel approach for the classification of white blood cells utilizing image processing and machine learning techniques. The proposed method encompasses three main stages, namely nucleus and cytoplasm detection, feature extraction, and classification through an SVM model. The achieved accuracy rate of the proposed method in categorizing WBCs in the Raabin-WBC dataset was 94.65%. Katar and Kilincer [34] proposed an approach for the automatic classification of WBCs using pre-trained deep learning models, including ResNet-50, VGG-19, and MobileNet-V3-Small. The proposed approach achieved high accuracy rates, with the MobileNet-V3-Small model reaching the highest accuracy of 98.86%. Akalin and Yumusak [35] present a study on the real-time detection and classification of WBCs in peripheral blood smears using object recognition frameworks. YOLOv5s, YOLOv5x, and Detectron2 R50-FPN pre-trained models were used, and two original contributions were made to improve the model's performance. The maximum accuracy rate achieved on the test dataset for detection and classification of WBCs was 98%. Leng et al. [36] present a study on developing a deep learning-based object detection network for leukocyte detection using the detection transformer (DETR) model. The study findings indicate that the improved DETR model outperforms the original DETR and CNN with a mean average precision (mAP) detection performance of up to 96.10%.

 doi:10.20944/preprints202306.1106.v1

Table 2. Comparison of our work with some state-of-the-art study techniques for WBC classification.

Study	Year	Number of Class	Method	Explainability	Performance
Tavakoli et al. [33]	2021	5 (Basophil, Eosinophil, Lymphocyte, Monocyte, Neutrophil)	SVM	Black-box	Acc=94.65%
Katar and Kilincer [34]	2022	5 (Basophil, Eosinophil, Lymphocyte, Monocyte, Neutrophil)	CNN	Grad-CAM	Acc=98.86%
Akalin and Yumusak [35]	2022	5 (Basophil, Eosinophil, Lymphocyte, Monocyte, Neutrophil)	Hybrid	Black-box	Acc=98.00%
Leng et al. [36]	2023	3 (Eosinophil, Monocyte, Neutrophil)	DETR	Black-box	mAP=96.10%
The proposed study	2023	2 (Granulocytes and Agranulocytes)	ViT	Score-CAM	Acc=99.75%
		5 (Basophil, Eosinophil, Lymphocyte, Monocyte, Neutrophil)			Acc=99.40%

In this study, the pre-trained ViT model was utilized for automatic classification of white blood cells. The model trained for five distinct types of white blood cells attained an accuracy rate of 99.75%. In contrast, the fine-tuned model, which classified cells based on their granule content, achieved an accuracy rate of 99.40%. In comparison to the studies listed in Table 3, we achieved higher accuracy and evolved the ViT model into an explainable structure using the Score-CAM algorithm. The ViT model's superior performance can be attributed to its unique architecture, which enables it to capture long-range dependencies between different parts of the image, resulting in better image recognition performance.

The advantages of our explainable model can be summarized as follows:

- The proposed model is based on vision transformers that has become popular research field. Therefore this study is an example to examine vision transformers performance in biomedical image classification.
- This model can classify WBCs images with end-to-end transformer structure. There is no need to use any feature engineering.
- Due to the explainable structure, the proposed model presents focused regions during the classification process. According to these results, experts can validate model performance.
- Due to its high level of classification accuracy, it has the potential to be utilized in clinical applications.

The limitations of our study are outlined as follows. Although the proposed method achieves a high success rate in classifying white blood cells, its response time was not assessed in a real-time study. Additionally, the model's resilience to image variations due to factors such as illumination and the noise was not verified. To address these limitations, future research will involve generating synthetic images using data augmentation techniques and training new models with these images. Furthermore, we will investigate the effectiveness of transformer-based models in detecting various diseases and symptoms. The proposed method can be seamlessly integrated into clinical software and provide invaluable assistance to specialists in WBCs classification.

5. Conclusions

In this study, we propose an explainable method based on the vision transformer for the automatic detection of white blood cells in blood film images. The model is trained and validated using a public five-class dataset of 16,633 samples. The pre-trained ViT model achieved a testing phase accuracy rate of 99.40% for the detection of five different subtypes of white blood cells. Examination of the model's predictions revealed that the most misclassified samples belonged to the Lympochyte subtype, which was predominantly predicted as Monocyte. Since Lympochyte and Monocyte cells lack granules and share similar visual features, this misclassification is understandable. The dataset used to confirm this situation was labeled according to granule presence, and the ViT model was trained using binary classification. The resulting model correctly classified test images with a success rate of 99.70%. The pixel areas focused on in the ViT model's predictions

were visualized using a heat mapping technique with the Score-CAM algorithm, further enhancing the model's reliability. The study's main limitation is the lack of information on real-time performance compared to object detection algorithms. The proposed ViT model can automate the detection of cells in blood films and can be effectively used in medical education due to its explainable structure. Moreover, the model can be fine-tuned for similar tasks, benefiting from the knowledge accumulated during training and achieving high accuracy rates.

Author Contributions: Conceptualization, O.K. and O.Y.; methodology, O.K.,.; software, O.K.,; validation, O.Y.; formal analysis, O.K.; investigation, O.K., and O.Y., ; writing—original draft preparation, O.K., and O.Y.; writing—review and editing, O.K., and O.Y.,; visualization, O.K.,; supervision, O.K.,. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

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

References

1. Al-Dulaimi, K.A.K.; Banks, J.; Chandran, V.; Tomeo-Reyes, I.; Nguyen Thanh, K. Classification of White Blood Cell Types from Microscope Images: Techniques and Challenges. *Microscopy science: Last approaches on educational programs and applied research (Microscopy Book Series, 8)* **2018**, 17–25.
2. Ahmad, Z.; Shah, T.A.; Reddy, K.P.; Ghosh, S.; Panpatil, V.; Kottoru, S.K.; Rayees, S.; Rao, D.R. Immunology in Medical Biotechnology. In *Fundamentals and Advances in Medical Biotechnology*; Springer, 2022; pp. 179–207.
3. Kiboneka, A.N. Basic Concepts in Clinical Immunology: A Review. **2021**.
4. Tripathi, C. Tolerance and Autoimmunity. In *An Interplay of Cellular and Molecular Components of Immunology*; CRC Press, 2022; pp. 207–216 ISBN 1003286429.
5. Otieno, F.; Kyalo, C. Perspective Chapter: Macrophages Plasticity and Immune Metabolism. In *Basic and Clinical Aspects of Interferon Gamma*; IntechOpen, 2022 ISBN 1803558865.
6. Wechsler, M.E.; Munitz, A.; Ackerman, S.J.; Drake, M.G.; Jackson, D.J.; Wardlaw, A.J.; Dougan, S.K.; Berdnikovs, S.; Schleich, F.; Matucci, A. Eosinophils in Health and Disease: A State-of-the-Art Review. In Proceedings of the Mayo Clinic Proceedings; Elsevier, 2021; Vol. 96, pp. 2694–2707.
7. Santos, A.F.; Alpan, O.; Hoffmann, H. Basophil Activation Test: Mechanisms and Considerations for Use in Clinical Trials and Clinical Practice. *Allergy* **2021**, *76*, 2420–2432.
8. Parente, J. Diagnostics for White Blood Cell Abnormalities: Leukocytosis and Leukopenia. *Physician Assist Clin* **2019**, *4*, 625–635.
9. Agnello, L.; Giglio, R.V.; Bivona, G.; Scazzone, C.; Gambino, C.M.; Iacona, A.; Ciaccio, A.M.; lo Sasso, B.; Ciaccio, M. The Value of a Complete Blood Count (CBC) for Sepsis Diagnosis and Prognosis. *Diagnostics* **2021**, *11*, 1881.
10. Wang, Q.; Bi, S.; Sun, M.; Wang, Y.; Wang, D.; Yang, S. Deep Learning Approach to Peripheral Leukocyte Recognition. *PLoS One* **2019**, *14*, e0218808.
11. Khamael, A.-D.; Banks, J.; Nugyen, K.; Al-Sabaawi, A.; Tomeo-Reyes, I.; Chandran, V. Segmentation of White Blood Cell, Nucleus and Cytoplasm in Digital Haematology Microscope Images: A Review–Challenges, Current and Future Potential Techniques. *IEEE Rev Biomed Eng* **2020**, *14*, 290–306.
12. H Mohamed, E.; H El-Behaidy, W.; Khoriba, G.; Li, J. Improved White Blood Cells Classification Based on Pre-Trained Deep Learning Models. *Journal of Communications Software and Systems* **2020**, *16*, 37–45.
13. Patil, A.M.; Patil, M.D.; Birajdar, G.K. White Blood Cells Image Classification Using Deep Learning with Canonical Correlation Analysis. *Irbm* **2021**, *42*, 378–389.
14. Basnet, J.; Alsadoon, A.; Prasad, P.W.C.; Aloussi, S. al; Alsadoon, O.H. A Novel Solution of Using Deep Learning for White Blood Cells Classification: Enhanced Loss Function with Regularization and Weighted Loss (ELFRWL). *Neural Process Lett* **2020**, *52*, 1517–1553.
15. 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.
16. 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*.

17. Jung, C.; Abuhamad, M.; Alikhanov, J.; Mohaisen, A.; Han, K.; Nyang, D. W-Net: A CNN-Based Architecture for White Blood Cells Image Classification. *arXiv preprint arXiv:1910.01091* **2019**.
18. Rustam, F.; Aslam, N.; de La Torre Díez, I.; Khan, Y.D.; Mazón, J.L.V.; Rodríguez, C.L.; Ashraf, I. White Blood Cell Classification Using Texture and RGB Features of Oversampled Microscopic Images. In Proceedings of the Healthcare; MDPI, 2022; Vol. 10, p. 2230.
19. Chola, C.; Muaad, A.Y.; bin Heyat, M.B.; Benifa, J.V.B.; Naji, W.R.; Hemachandran, K.; Mahmoud, N.F.; Samee, N.A.; Al-Antari, M.A.; Kadah, Y.M. BCNet: A Deep Learning Computer-Aided Diagnosis Framework for Human Peripheral Blood Cell Identification. *Diagnostics* **2022**, *12*, 2815.
20. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding Robustness of Transformers for Image Classification. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 10231–10241.
21. Bazi, Y.; Bashmal, L.; Rahhal, M.M. al; Dayil, R. al; Ajlan, N. al Vision Transformers for Remote Sensing Image Classification. *Remote Sens (Basel)* **2021**, *13*, 516.
22. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-Trained Image Processing Transformer. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021; pp. 12299–12310.
23. Kouzehkanan, Z.M.; Saghari, S.; Tavakoli, S.; Rostami, P.; Abaszadeh, M.; Mirzadeh, F.; Satlsar, E.S.; Gheidishahran, M.; Gorgi, F.; Mohammadi, S. A Large Dataset of White Blood Cells Containing Cell Locations and Types, along with Segmented Nuclei and Cytoplasm. *Sci Rep* **2022**, *12*, 1123.
24. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; 2020; pp. 38–45.
25. Guo, M.-H.; Liu, Z.-N.; Mu, T.-J.; Hu, S.-M. Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks. *IEEE Trans Pattern Anal Mach Intell* **2022**.
26. Wang, L.; He, Y.; Li, L.; Liu, X.; Zhao, Y. A Novel Approach to Ultra-Short-Term Multi-Step Wind Power Predictions Based on Encoder–Decoder Architecture in Natural Language Processing. *J Clean Prod* **2022**, *354*, 131723.
27. Thakur, P.S.; Khanna, P.; Sheorey, T.; Ojha, A. Explainable Vision Transformer Enabled Convolutional Neural Network for Plant Disease Identification: PlantXViT. *arXiv preprint arXiv:2207.07919* **2022**.
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 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* **2020**.
29. Lee, S.H.; Lee, S.; Song, B.C. Vision Transformer for Small-Size Datasets. *arXiv preprint arXiv:2112.13492* **2021**.
30. Ekanayake, I.U.; Meddage, D.P.P.; Rathnayake, U. A Novel Approach to Explain the Black-Box Nature of Machine Learning in Compressive Strength Predictions of Concrete Using Shapley Additive Explanations (SHAP). *Case Studies in Construction Materials* **2022**, *16*, e01059.
31. Liang, Y.; Li, S.; Yan, C.; Li, M.; Jiang, C. Explaining the Black-Box Model: A Survey of Local Interpretation Methods for Deep Neural Networks. *Neurocomputing* **2021**, *419*, 168–182.
32. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2020; pp. 24–25.
33. Tavakoli, S.; Ghaffari, A.; Kouzehkanan, Z.M.; Hosseini, R. New Segmentation and Feature Extraction Algorithm for Classification of White Blood Cells in Peripheral Smear Images. *Sci Rep* **2021**, *11*, 19428.
34. KATAR, O.; KILINÇER, İ.F. Automatic Classification of White Blood Cells Using Pre-Trained Deep Models. *Sakarya University Journal of Computer and Information Sciences* **2022**, *5*, 462–476.
35. AKALIN, F.; YUMUŞAK, N. Detection and Classification of White Blood Cells with an Improved Deep Learning-Based Approach. *Turkish Journal of Electrical Engineering and Computer Sciences* **2022**, *30*, 2725–2739.
36. Leng, B.; Wang, C.; Leng, M.; Ge, M.; Dong, W. Deep Learning Detection Network for Peripheral Blood Leukocytes Based on Improved Detection Transformer. *Biomed Signal Process Control* **2023**, *82*, 104518.