An XAI Integrated Identification System of White Blood Cell Type Using Variants of Vision Transformer

Chapter · June 2023

DOI: 10.1007/978-3-031-35308-6_26

6 authors, including:

Shakib Mahmud Dipto
University of Liberal Arts Bangladesh (ULAB)
12 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Md Tanzim Reza
BRAC University
46 PUBLICATIONS **159** CITATIONS
SEE PROFILE

Mohammad Zavid Parvez
64 PUBLICATIONS **989** CITATIONS
SEE PROFILE

Prabal Datta Barua
University of Southern Queensland
131 PUBLICATIONS **1,675** CITATIONS
SEE PROFILE

An XAI Integrated Identification System of White Blood Cell Type using Variants of Vision Transformer

Shakib Mahmud Dipto[1], Md Tanzim Reza[1], Md. Nowroz Junaed Rahman[1], Danish Faraz Abbasi[3], and Mohammad Zavid Parvez*[,2,3,4,5,6]

[1]Department of Computer Science and Engineering, BRAC University, Bangladesh
[2] Information Technology, APIC, Australia
[3] Information Technology, Kent Institute, Australia
[4] Information Technology, Torrens University, Australia
[5] Peter Faber Business School, Australian Catholic University, Australia
[6] School of Computing, Mathematics, and Engineering, Charles Sturt University, Australia
{diptomahmud2,rezatanzim,nowrozjunaedrahman}@gmail.com
danish.abbasi@kent.edu.au,mparvez@csu.edu.au

Abstract. White Blood Cells (WBCs) serve as one of the primary defense mechanisms against various diseases. Therefore, in order to detect blood cancer as well as many other disorders, routine WBC monitoring may be necessary. Numerous studies have proposed automated 4 types of WBC detection through Machine Learning and Deep Learning based solutions. However, transformers based applications, which primarily originated from the field of Natural Language Processing, are very scarce. Our proposed study showcases the applications of Vision Transformers (VTs) for WBC type identification. Firstly, a pre-augmented dataset of nearly 12,500 images was taken. Afterward, two variants of VTs were trained and evaluated on the dataset. Our analysis revealed that the accuracy for all the models ranged from 83% to 85%, making the performance of the VTs equivalent to that of the standard Deep Learning models. Meanwhile, VTs have demonstrated significantly faster learning symptoms during the training phase, which can be useful when one wants to maximize learning through fewer epochs, for example, in a federated learning environment. Finally, the application of Explainable AI (XAI) was visualized on the VTs using Gradient-weighted Class Activation Mapping (GradCam).

Keywords: Vision Transformer, GradCam, Blood Cell, White Blood Cell, Transformer, Eosinophil, Neutrophil, Lymphocyte, Monocyte

1 Introduction

Blood cells hold a great deal of significance for the health of the body because they make up the majority of the human physique. Red blood cells (RBCs),

white blood cells (WBCs), and platelets are the three main types of blood cells that circulate throughout the body's many organs. These cell types are distinct in terms of their purposes and functionalities. For instance, RBCs carry oxygen to various parts of the body, platelets help to stop bleeding, and WBCs act as a defense mechanism against diseases. Since an irregular distribution of the WBC elements can cause disruption in bodily functions, a thorough inspection of the WBC components is essential because they ensure the welfare of health. Our proposed study is based on the automated inspection of four types of WBC types; namely Eosinophil, Neutrophil, Lymphocyte, and Monocyte. The majority of earlier studies illustrated the automation process using machine learning (ML) and deep learning (DL) methods. However, transformers-based approaches, which are primarily from the domains of Natural Language Processing and relatively new in the subject of computer vision, have received very little to no research.

The proposed study demonstrates the application of two alternative variants of Vision Transformers (VTs), one is the standard version while the other one includes a locality self-attention mechanism, which allegedly helps to provide better performance on small datasets [1]. In addition, by integrating Explainable AI (XAI) leveraging Gradcam [2], the study attempts to display the interpretation of the transformers. The proposed study's main contributions include:

- Demonstration of the usefulness of VTs compared to popular Convolutional Neural Network (CNN) methods in automated WBC identification. As per the authors' knowledge, there haven't been many studies on this topic.
- A comparison of the effectiveness of conventional VTs with self-attention-based VTs
- Utilization of Gradcam to demonstrate the understandability of the transformers based approaches

The paper is broken down into five principal sections. The review of literature is discussed in chapter two after this introductory chapter. The exposition of the proposed model, the analysis of the results, and the conclusion are covered in detail in chapters three, four, and five respectively.

2 Literature Review

Computer vision related research on blood cells covers a wide range of topics, including disease identification and subcellular element detection. Although the utilized tools and methodologies also vary widely, most of them could be generally grouped as Machine Learning (ML) and Deep Learning (DL) based studies. This particular section provides some insight into some of these earlier works and methodologies.

Speaking of general ML based studies, Habibzadeh et al. proposed in their work to apply Dual-Tree Complex Wavelet Transformation to extract wavelet based features. [3] Support Vector Machine (SVM) was then applied to the reduced feature set to identify several types of white blood cells using shape,

intensity, and texture data. Even with poor-quality samples, the method produced considerably accurate results. Meanwhile, to carry out feature selection, Gupta et al. recommended utilizing the Optimized Binary Bat algorithm, an evolutionary algorithm that is an enhanced version of the original Bat algorithm. [4] The authors of this study achieved very high accuracy by using algorithms like Logistic Regression, Decision Tree, KNN, and Random Forest on the chosen features. In another work, Benomar et al. concentrated on offering a system for differentiating WBC counts. [5] In this work, WBCs were identified using a noble color transformation technique. Afterward, utilizing color, texture, and morphological traits, the nucleus and cytoplasm of WBCs were segmented using the controlled watershed algorithm and then classified using the Random Forest algorithm. As we can see, the ML-based approaches generally consist of utilizing two subsequent methods, one for feature extraction and the other for deriving the results based on the extracted features.

The DL based research simplifies the process of feature extraction by automating it, making it popular for research and deployment. Cheque et al. presented an approach of using the Faster R-CNN network along with two parallel Convolutional Neural Networks (CNNs) to classify white blood cells. [6] In this method, mononuclear and polymorphonuclear WBCs were initially separated into two groups using Faster R-CNN. Following the aforementioned procedures, the dataset was divided into two cell groups. For each of the groups, two MobileNet architecture based CNNs leveraging transfer learning were developed and applied for classification. An interesting technique was proposed by Liang et al., where they used a combination of pre-trained CNN with Recurrent Neural Network (RNN) to classify Blood Cell Images. [7] They fed the training data to both the pre-trained CNN and RNN. The resultant extracted feature was then put through a softmax layer to categorize various types of white blood cells. There are more DL-based works that leverage existing popular architectures such as the utilization of VGGNet by Sahlol et al. [8], utilization of VGGNet, Inception V3, LeNet, and XeceptionNet by Sharma et al. [9], segmentation of WBCs using YOLO v3 by Praveen et al. [10], and so forth. There are also works that make use of custom CNN models. For instance, Akram and his co-researchers introduced a novel CNN architecture called multi-scale information fusion network (MIF-Net) for WBC segmentation. [11] They described this architecture as shallow in shape and it can combine internal along with external spatial information to enhance the segmentation process.

A few studies combined ML and DL-based methodologies to maximize their benefits in terms of WBC classification. In these scenarios, the DL component mostly comprises of CNNs that automatically extract features, which the ML classifiers then use for segmentation or classification. The work by Zhang et al. that combines adversarial residual networks with linear Support Vector Machine (SVM) [12] and the combination of AlexNet/GoogleNet with SVM by Cinar et al. [13] are some noteworthy ones.

As the reviews listed above demonstrate, Traditional ML and DL techniques dominate the field of WBC type detection from blood cell images. Transformers,

which newly arrived in the realm of computer vision, have not been employed very often to categorize WBC types. The work by Choe et al. on the image transformer is one the most notable of the few works that were discovered. [14] In their work when Vision Transformer was pitted against ResNet in terms of WBC classification, it was able to outperform ResNet. There is also a dearth of research on the interpretability of the working procedures on WBC classification by these transformer based approaches. The lack of data on transformer based analysis combined with the recent domination of the approach in the field of computer vision has led us to research further on it, resulting in the proposed study.

3 Proposed Model

Taking input images, pre-processing, training the models, evaluation and comparison are the five main pillars of the proposed work. Initially, we trained the vision transformer models using the blood cell images. Afterward, necessary metrics were calculated and compared against the results of traditional DNN architecture. Given that our dataset is not particularly large, we opted for the VGGNet designs as traditional DNNs because they are simple to understand, extremely powerful, and better suited for small datasets than some of the deeper architectures. From the VGGNet family, the most popular two architectures, VGG16 and VGG19 were used. A quick introduction to the employed models is given in the following subsection.

3.1 Architecture Details

Vision Transformer [15]: The basic Vision Transformers (VTs) work by dividing images into small patches that are flattened. The flattened patches are next transformed into lower dimension embeddings, which are then sent through a transformer encoder. The encoded information is then passed through a Multilayer Perceptron (MLP) head consisting of numerous fully connected layers. This MLP based head helps to classify the images. In general, this architecture requires a large number of images for training from scratch. Fortunately, pre-trained architectures are available which can be fine-tuned on smaller datasets. Our version of VT is pre-trained on ImageNet-21k dataset, consisting of more than 14 million images. We took the model from Keras libraries, kept the default configurations except for the output layer to match the number of classes, and fined tuned on our small blood cell dataset.

Since the first iteration in 2020, VTs have gone through various modifications. One of these modifications incorporates the addition of Shifted Patch Tokenization (SFT) and Locality Self Attention (LSA). [16] The incorporation of SFT and LSA makes the VT consider the local correlation between image pixels as opposed to regular VTs and hence, reduces the requirements for massive datasets. In our study, we experimented on both the regular VT and the SFT-LSA incorporated VT.

VGGNet [17] : Due to their popularity, power, and simplicity, VGG architectures are among the most widely used CNN architectures currently available. VGG architectures are a combination of convolution and max pool layers sequentially arranged, and the architectures are numbered based on the number of layers. The two most popular VGG variants, with 16 and 19 convolution layers, are VGG16 and VGG19 respectively. We have used these two architectures to provide a comparative performance analysis against the VTs.

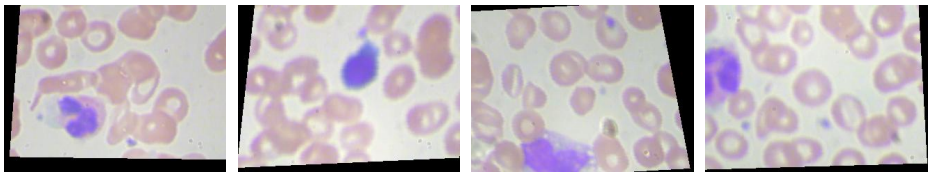
3.2 Dataset Description

The blood cell image dataset consists of microscopic images of the blood tissues. The original dataset is hosted in Github. [18] Meanwhile, a pre-augmented variant is provided in Kaggle, resulting in close to 12,500 images after augmentation. [19] The augmented variant in Kaggle comes in train and test set only, we further created the validation set by splitting the test set and taking nearly 10% of images from it. The overall train-test-validation distribution is provided in table 1.

Table 1: Dataset Train-Test-Validation Distribution

WBC Type	**Train Distribution** (No. of Images)	**Validation Distribution** (No. of Images)	**Test Distribution** (No. of Images)
Eosinophil	2497	66	557
Neutrophil	2499	62	562
Lymphocyte	2483	58	562
Monocyte	2478	58	562

Images from blood tissues spread out on slides were used to create the dataset. Therefore, RBCs, WBCs, and platelets are all represented in a single image. However, the WBCs can be clearly identified thanks to their distinctively large appearance and unique color. In figure 1, a few image samples are provided.



(a) Eosinophil (b) Lymphocyte (c) Monocyte (d) Neutrophil

Fig. 1: Sample images from the dataset

3.3 Model Description

The diagram in figure 2 gives a detailed summary of our study.

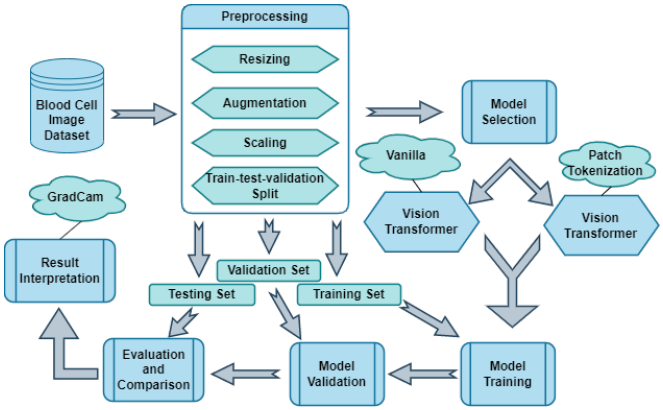


Fig. 2: Proposed Model

In our study, blood cell images were collected and preprocessed. Initially, the input images were divided into three parts: training, testing, and validation sets. Afterward, input images were passed through in 112x112 resolution and the preprocessing layer resized it up to 224x224 resolution. We normalized the pixel values within a range between 0-1, augmented the input images through horizontal flipping, zooming in between 20%, and rotating by 2% randomly.

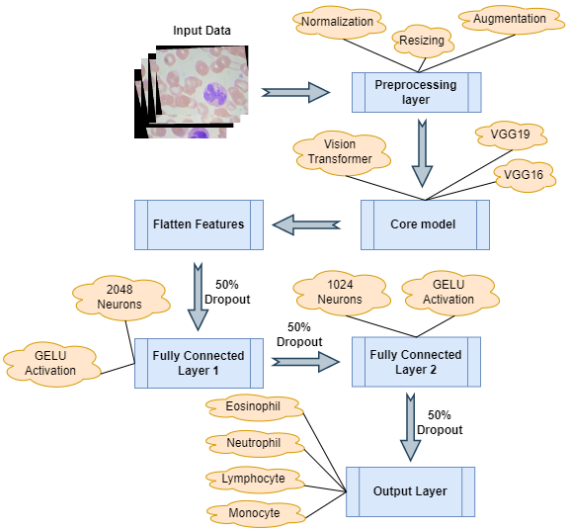


Fig. 3: Experimental Setup

Afterward, both the standard VT and the VT with SFT-LSA integration were applied to the input images. For each VT, the training simulation was run for 100 epochs, and the accuracy results, as well as precision, recall, F1 metrics,

and other data, were extracted. Following that, the same metrics were extracted using 100 epochs of training on the VGG models. The retrieved metrics from all the models were then examined and compared. For the fairness of comparison, we kept the experimental setups and the overall architectures of the models exactly the same. For our experiment, we performed analysis on a workstation consisting of a 3.9 GHz AMD Ryzen 9 5950X 16-core processor, 64 GB ram, and RTX 3090 24 GB. The overall model setups are provided in figure 3.

4 Result Analysis

50 epochs of training through vanilla VT and SFT-LSA incorporated VT show gradual improvement of training performance. As visible in figure 4, The validation loss reaches a plateau after around 15 epochs and no notable improvement was observed after that. The spikes in the validation loss curve are perhaps caused by the small validation set, where small variations in classification-misclassification scores caused big differences in terms of scores.

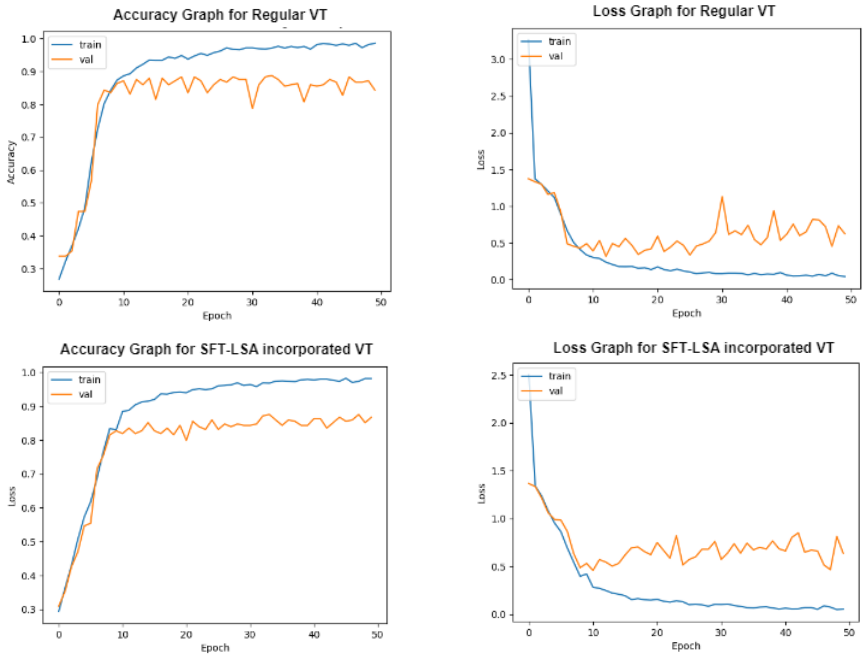


Fig. 4: Accuracy and loss graph for the two variants of VTs

Overall, in terms of training performance, the two variants of VT show very few differences. Rather, we noticed the SFT-LSA incorporated variant taking more time to train, approximately 10 seconds per epoch in contrast to the 8

seconds per epoch for the regular VT. Therefore, SFT-LSA incorporated VT did not serve any major advantage for our case.

Additionally, the confusion matrices generated by the VTs are provided in figure 5.

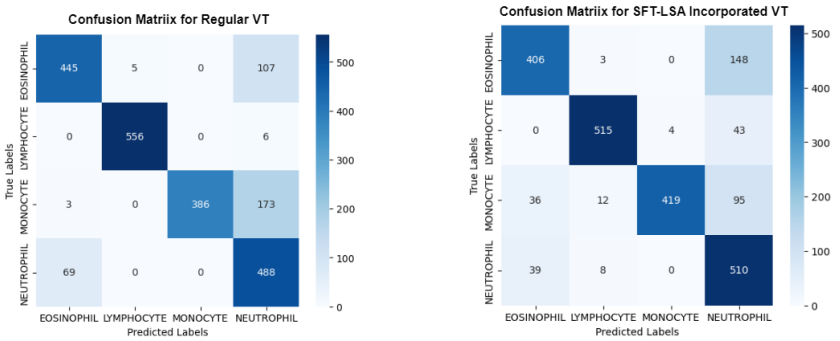


Fig. 5: Confusion matrices for the two variants of VTs

As noticeable in figure 5, the matrices generated by the two variants of VTs show quite similar patterns. Both the models struggle to classify Monocyte and Eosinophil labeled images compared to the Lymphocyte and Neutrophil labeled images. A lot of Monocyte and Eosinophil images are classified as Neutrophils. As a result, it appears that the patterns of monocyte-eosinophil and monocyte-neutrophil pairings overlap. This circumstance is reflected in the results metrics generated by the VTs which are given in figure 6.

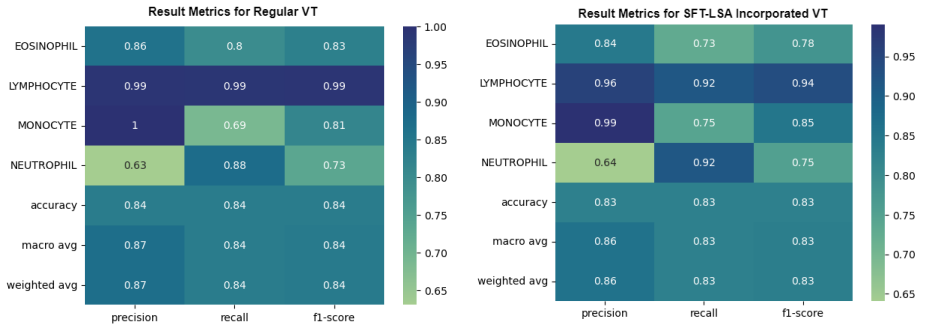


Fig. 6: Result metrics from the two variants of VTs

As we can see in figure 6, the recall score and the precision score for the labels Monocyte and Neutrophil are quite low respectively, showing the difficulty in Monocyte identification and high false positive for Neutrophil detection.

Meanwhile, a comparison against the regular VGGNet models against VT shows similar performance in terms of min loss and max accuracy. However, as visible in figure 7, VGG models reach the max training plateau late compared to the VT models. The loss and accuracy fluctuation is also quite higher compared to what we can see in the training graphs of the VTs.

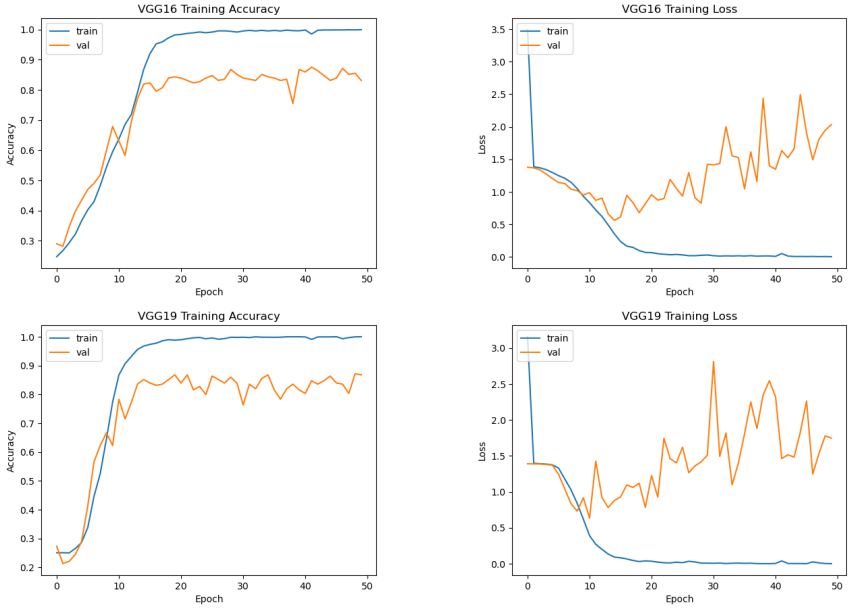


Fig. 7: Accuracy and loss graph for VGG16 and VGG19

Finally, we analyzed the scores of the used architectures on the test set. For our trial case, VGG19 achieved the highest accuracy score of 85%, VGG16 scored the lowest of 82%, and both variants of the VTs scored in the middle.

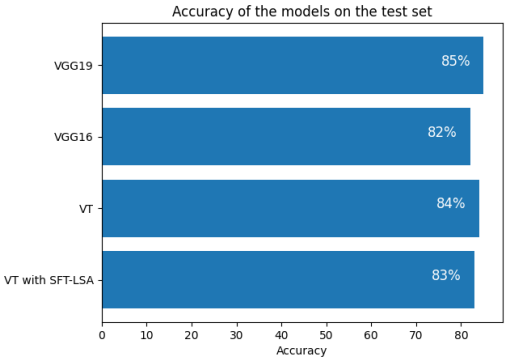


Fig. 8: Comparison of the performance on the test set

The difference between the scores themselves, however, is very little and the order of performance may easily vary due to tumbling into different local maxima

at each trial case. Therefore, we can conclude that in terms of test results, all the models perform fairly close to each other.

Finally, we applied GradCam on the dataset in an attempt to visualize the attention map of the VTs. Some of the resultant maps from the correctly classified images are given in figure 9.

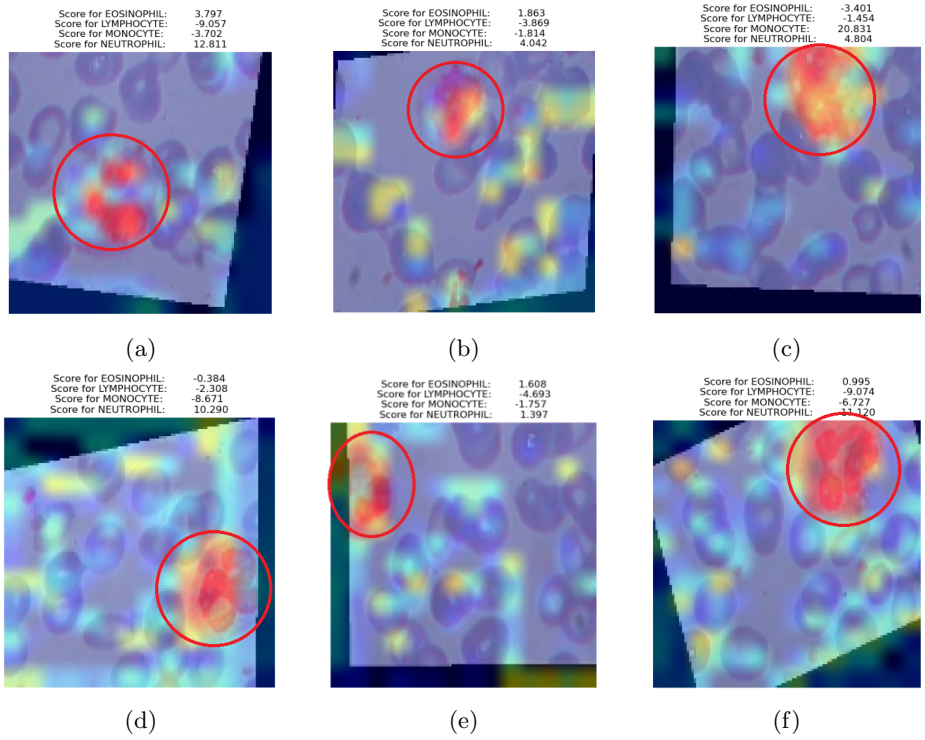


Fig. 9: GradCam activation maps for regular VT (Top) and SFT-LSA incorporated VT (Bottom)

As visualized in figure 9, the gradient weighted class activation mapping for the transformers can be quite sparsely mapped throughout the images. There are still traces of correct activation mapping where the WBC gets highlighted but there are also sparsely distributed activation mapping around the platelets. Generally, this sparse distribution of attention maps is caused by the fact that the mapping only tells about the areas the network paid attention to, but not taking into account the areas that were actually used to make the final classification. Thus, resulting in sparse mapping all around the images. This can be improved by removing the lowest attention values using min fusion [20]. However, that particular implementation is outside of the scope of the current study. After all the analysis and evaluation, our overall findings are as follows:

- Even on a small dataset of 12,500 images, fined-tuned VTs performed quite competitively against regular CNNs. If the dataset was larger, the perfor-

mance of VTs would have likely outperformed that of conventional CNNs due to the massive volume of data that VTs usually require
- Pre-trained VTs typically approach the maximum validation accuracy fairly quicker than regular CNNs, albeit having similar overall performance. The VTs converged on the validation set in only 9–10 epochs as opposed to 17–18 for the VGG models. When models have a limited number of learning iterations from a set of data, this rapid convergence might be extremely helpful. For instance, in a federated setting or during few-shot learning
- The SFT-LSA incorporated VT could not show better performance compared to the vanilla pre-trained VT, despite having slightly larger training time. Perhaps the dataset needed to be even smaller for proper utilization of the SFT-LSA incorporated VT
- GradCam results on VTs can be serviceable but sporadic. The GradCam results can be improved as suggested by other articles. [20]

5 Conclusion

In conclusion, it is fair to state that VTs hold a lot of promise in the field of computer vision-based medical picture analysis. The application of VTs in image classification is still fairly new, with new variants being introduced on a regular basis. Our proposed study demonstrates that, despite VTs' inherent need for large amounts of data, fine-tuning a pre-trained VT can still yield decent results even on smaller datasets. There are also other variants of transformers that have proven to be performing better, swin transformers and Multi-axis vision transformers for instance. In further work, we plan to apply the other variants of transformers to analyze their performance. In addition, we also wish to improve the performance of Explainable AI by modifying the process of activation mapping. The suggested improvements should provide a broader analysis of the application of transformers on WBC type classification.

References

1. Lee, S. H., Lee, S., & Song, B. C. (2021). Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492.
2. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
3. Habibzadeh, M., Krzyżak, A., & Fevens, T. (2013). Comparative study of shape, intensity and texture features and support vector machine for white blood cell classification. Journal of Theoretical and Applied Computer Science, 7(1), 20-35.
4. Gupta, D., Arora, J., Agrawal, U., Khanna, A., & de Albuquerque, V. H. C. (2019). Optimized Binary Bat algorithm for classification of white blood cells. Measurement, 143, 180-190.

5. Benomar, M. L., Chikh, A., Descombes, X., & Benazzouz, M. (2021). Multi-feature-based approach for white blood cells segmentation and classification in peripheral blood and bone marrow images. International Journal of Biomedical Engineering and Technology, 35(3), 223-241.
6. Cheuque, C., Querales, M., León, R., Salas, R., & Torres, R. (2022). An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification. Diagnostics, 12(2), 248.
7. Liang, G., Hong, H., Xie, W., & Zheng, L. (2018). Combining convolutional neural network with recursive neural network for blood cell image classification. IEEE access, 6, 36188-36197.
8. Sahlol, A. T., Kollmannsberger, P., & Ewees, A. A. (2020). Efficient classification of white blood cell leukemia with improved swarm optimization of deep features. Scientific Reports, 10(1), 1-11.
9. Sharma, M., Bhave, A., & Janghel, R. R. (2019). White blood cell classification using convolutional neural network. In Soft Computing and Signal Processing (pp. 135-143). Springer, Singapore.
10. Praveen, N., Punn, N. S., Sonbhadra, S. K., Agarwal, S., Syafrullah, M., & Adiyarta, K. (2021, October). White blood cell subtype detection and classification. In 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 203-207). IEEE.
11. Akram, N., Adnan, S., Asif, M., Imran, S. M. A., Yasir, M. N., Naqvi, R. A., & Hussain, D. (2022). Exploiting the Multiscale Information Fusion Capabilities for Aiding the Leukemia Diagnosis Through White Blood Cells Segmentation. IEEE Access, 10, 48747-48760.
12. Zhang, C., Wu, S., Lu, Z., Shen, Y., Wang, J., Huang, P., ... & Li, D. (2020). Hybrid adversarial-discriminative network for leukocyte classification in leukemia. Medical physics, 47(8), 3732-3744.
13. Çınar, A., & Tuncer, S. A. (2021). Classification of lymphocytes, monocytes, eosinophils, and neutrophils on white blood cells using hybrid Alexnet-GoogleNet-SVM. SN Applied Sciences, 3(4), 1-11.
14. Cho, P., Dash, S., Tsaris, A., & Yoon, H. J. (2022, April). Image transformers for classifying acute lymphoblastic leukemia. In Medical Imaging 2022: Computer-Aided Diagnosis (Vol. 12033, pp. 633-639). SPIE.
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
16. Lee, S. H., Lee, S., Song, B. C. (2021). Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492.
17. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
18. BCCD Dataset. https://github.com/Shenggan/BCCD_Dataset. Accessed on 14 November, 2022
19. Paul Moony. Blood Cell Images. https://www.kaggle.com/datasets/paultimothymooney/blood-cells. Accessed on 14 November, 2022
20. jacobgil. Explainability for Vision Transformers. https://github.com/jacobgil/vit-explain. Accessed on 15 November, 2022