International Multidisciplinary Research Journal Reviews (IMRJR)

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602

Theoretical Framework Study on Blood Cell Cancer Detection using Vision based Transformer

Nishant Tripathi

Dept. of CSE, JAIN university, Bengaluru

Abstract. Blood cell cancer poses a significant threat to the human body. Cancer a prevalent, multifaceted, and perilous blood disorder, underscores the utmost significance of early detection and treatment. The vital role that blood cells play in the human body allows for their utilization in clinical diagnosis. To predict this condition, various semiautomatic systems have been developed using different medical imaging techniques. This publication conducts a extensive research of the literature regarding the use of vision transformer (ViT) based models in the analysis of blood cell cancer images. It discusses the merits and drawbacks of several ViT-based models, including DBN, CNN, MVT (medical vision transformers), SW-Vit (shifted window vision transformers), MR-ViT (multi-Resolution vision transformers), SNN (spiking neural networks) and MLNN (multilayer neural networks). Furthermore, the review highlights that research on blood cell cancer detection has employed diverse deep learning models across various publicly available datasets. Performance evaluations of these models involved metrics such as accuracy, precision, recall, among others.

Keywords: Vit ViT model Vision Transformer· MVT· MR-ViT·SW-ViT· MLNN· CNN· Accuracy· AUC

I. INTRODUCTION

Cancer, a prevalent, multifaceted, and perilous blood disorder, underscores the utmost significance of early detection and treatment. Timely identification has the potential to lower cancer-related fatalities. The vital role that bloods cells play in the human body's metabolism allows for their utilization in clinical diagnosis. The analysis of blood cell cancer is pivotal for disease identification, with blood cell cancer (BCC) being the most common ailment among both males and females. BCC ranks as the world's deadliest disease, claiming thousands of lives annually. Detecting BCC at an early stage can substantially boost patient survival rates. Machine learning works a important role in automating the identification, outlining, computer-assisted diagnosis of malignant lesions. In this paper, we applied a vision transformer model to train for the detection of cancerous blood cell conditions [1]. Vision transformers are rapidly emerging as the predominant architecture for computer vision tasks, but there remains a significant gap in our understanding of their underlying principles and the knowledge they acquire [2]. The Vision Transformer is a powerful deep learning self-attention mechanism, and its versatility in different fields is due to its strong attention mechanism that enriches our comprehension of data characteristics, leading to remarkable outcomes. When it comes to detecting blood cell patterns, using a faster R-CNN network to find and different each on blood cell sections from the original images. By introducing the Vision Transformer, we significantly improved the precision of blood cell cancer classification. The outcomes achieved with the Vision Transformer underscore that our proposed model outperforms previous methods in both Whole Slide Image Analysis and Blood Cell Cancer Detection tasks [3, 4]. The Transformer architecture has gained widespread recognition as the dominant method for addressing tasks in natural language processing. Nevertheless, its adoption in computer vision tasks remains somewhat limited. In the field of computer vision, attention mechanisms are generally used in conjunction with convolutional networks or selectively integrated to replace certain components within these networks, all while maintaining the core structure of these networks [5].

Chen et al., (2023) introducing a novel method for automated blood cell cancer classification, the authors utilized the Shifted Window Vision Transformer (SW-ViT) model as its foundation. Initially, the Vision Transformer (ViT) architecture underwent subsequent and pre-training on the ImageNet dataset fine-tuning using blood cell images for categorization. Similarly, the SW-ViT architecture was initially pre-trained on the ImageNet dataset and then adjust with blood cell cancer images to enable accurate classification. To enhance classification performance, the study employed two distinct transfer strategies. The first approach involved a comprehensive fine-tuning of the entire SW-ViT model, while the second strategy concentrated on adjust only the linear achievement layer of Shifted Window ViT, keeping all additional boundary unchanged [6]. Recent advancements in artificial intelligence for the recognition of dermoscopic images have significantly advanced the early detection and treatment of blood cell cancer. This is particularly crucial as the global incidence of this condition continues to rise annually, posing a significant and growing threat to human health [7]. Saad et al., (2022) proposed the implementation of a Vision Transformer (ViT) model to classify blood cell images, enabling the automated sorting of these images into normal, non-cancerous, and cancerous categories by incorporating self-attention mechanisms into the model [8]. Wang et al., (2022) proposed is a semi-supervised training framework exploiting the Vit , a model known for outperforming CNN models in various classification tasks.

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602

Despite the Vision Transformer's (ViT) success in various domains, its utilization in detecting basal cell carcinoma (BCC) has been limited. Presents a customized semi-supervised learning approach that integrates selfconsistency and supervised training to enhance the model's robustness. The method integrates an adaptive token sampling method, allowing for the selective extraction of vital tokens from the input image, leading to significant enhancements in achievement [9].

The focus of this research is to create a Vision Transformer (ViT) method for identifying and locating cancer cells in blood samples. To achieve this, we have employed a vision transformer for cell identification, and our focus for cell detection has been on segmentation models. Objective of this study is to improve the accuracy of our classification model. Vision Transformer (ViT) architectures have shown impressive performance in classifying images of cancerous blood cells, offering potential applications in upcoming clinical trials. Notably, the vision transformer serves as an innovative self-attention-based framework for classifying sequential medical images.

II. RELATED WORKS

Malaviya et al. (2023) propose a technique wherein CT images classified pre-assess-malignant benign nature of cancer. Segmentation is used to divide the images into small patches which form input to the transformer encoder. Subsequently, through a process of several epochs tortured under training that lasted 100 epochs, a vision transformer is created to produce a model that obtained a striking 91.93% accuracy. Their research work is primarily centered on applying image analysis and machine learning techniques for improvement in lung cancer diagnosis, having attained such an outstanding degree of precision in this regard [1]. Enhanced diagnosis of leukemia through image analytic technology extracts pertinent features from images of blood cells with the aid of encoder layers, along with the addition of a sparse attention module based on the Transformer's self-attention mechanism, such that it can pay selective attention to important areas in the images. In addition, a contrastive loss function has been projected to elevate the discrimination of features. This ultimately aimed at improving the accuracy and objectivity associated with leukemia diagnosis, although concrete results or findings were not specified. However, experimental results prove that the module outperformed others by attaining 91.96% accuracy in working trials [3].

A malignant gene expression data categorization technique will use a Vision Transformer network as its basis according to Gokhale et al. (2023). By making extensive use of the attention mechanism, this technology has a chance of being able to assess with more accuracy the case of cancer diagnosis as compared to anything available [4]. Dosovitskiy et al. (2020) implemented the Transformer framework in computer vision for image classification applications. They argue that with pre-training on large datasets and evaluation on a variety of image recognition benchmarks, a pure Transformer model with no custom modifications can match state-of-the-art performance when compared with other recommended methods and transforms [5]. Clinical diagnostics for assessing blood cell health are of utmost importance, as this directly impacts metabolism within the human body. Thus, the classification of blood cells forms another major task characterized by huge amounts of manual analysis. Nonetheless, with recent advances in computer vision, this burden could be alleviated. The blood cell classification was attempted using two possible transfers to improve accuracy. One transfer improved the complete Shifted Window ViT model, whereas the second one sought to fine-tune the linear output layer of SW-ViT only, keeping the parameters of all other layers frozen. These tests made use of the Blood Cell Count and Detection publicly available for evaluation [6]. A number of CAD (computer-aided diagnostic) systems for detecting and classifying breast cancer (BC) have been developed with several imaging modalities tied to the back with ML model.

Here, we propose Vision Transformer (ViT), which serves as the basis of one semisupervised learning scheme. The ViT model has been shown across many classification benchmarks to outperform convolutional neural network models; however, its applications for BC detection have been rather few. We test our approach on two datasets, which include images from histology and ultrasonography. Results show that ours is the first to maintain state-of-the-art performance alongside CNN baselines in both learning tasks [9]. Aladhadh et al. (2022) designed a two-layer architecture using ML techniques for skin cancer (SC) classification augmentation and countering issues like data insufficiency and low accuracy. Data augmentation was implemented at the initial step. The HAM10000 dataset was tested with application of the Medical Vision Transformer (MVT) model, which proved to be superior in Stage II to existing works [10]. Mali et al. (2022) presented a CRC detection method based on vision transformers. We evaluate performance of our CRC detection method on a well-established benchmark dataset [11].performance of our CRC detection method using a well-established benchmark dataset [11].

PACAL, I., (2022) was involved in early breast cancer detection and reviews deep learning techniques that classify breast ultrasound images. It offers a comparison between traditional CNN architectures such as AlexNet and ResNet with novel transformer models or architectures like Vision Transformers. In the BUSI dataset study, vision transformers have been shown to be very excellent, yielding 88.6% accuracy. These findings establish deep learning as having a role in breast cancer diagnosis with a view to further application in clinical practice [12]. Das et al., (2021) Adding more precision and accuracy in identifying and classifying myeloma cells is the intent with which this research was developed. The Mask-RCNN model gives 93% performance, while the Efficient Net B3 model shows a 94.68% accuracy rate. [13].

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602

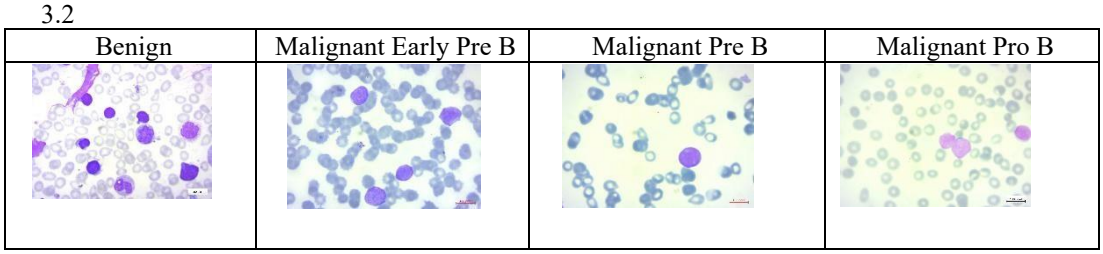
Li et al., (2023) presented interpretations on the technology-based timelines of Distance ViT-National Operations Models, the first using vector embeddings for a continuous countdown in time representation, and the second using a temporal attention model to vary attention weights. Further detailing intended at a more accurate description of longitudinal images of lung nodules with irregular sampling intervals. Experimental results indicate that the way their efforts work better than existing methods of ViTs classifying said images. More specifically, their methods reach AUC scores of, respectively, 0.785 and 0.786, meaning improvements compared to a cross-section on the AUC score with 0.734. Their methods yield matching discriminative performance with the top-performing longitudinal medical imaging algorithm with an AUC score of 0.779 in benign versus malignant classification under cross-validation on screening chest CTs from the NLST dataset. [14]

III. METHODOLOGY

3.1 Dataset

This research aimed to detect Blood Cell Cancer using Vision Transformer. The data collected from Kaggle. Acute Lymphoblastic Leukaemia (ALL) diagnosis is challenging due to its invasiveness and cost. Using peripheral blood smear (PBS) images for initial screening is crucial but prone to diagnostic errors due to vague symptoms. We have a dataset of 3,242 PBS images from 89 suspected ALL patients at Taleqani Hospital in Tehran, Iran. It's divided into benign and malignant classes, with malignant including three subtypes. Images were captured at 100x magnification with a Zeiss camera. The definitive diagnosis was done by a specialist using flow cytometry. An example of Blood Cell Cancer with various diseases is given in fig 1.

3.2



3.3Fig. 1. Image Examples of Blood Cell Cancer used.

3.4 Proposed model

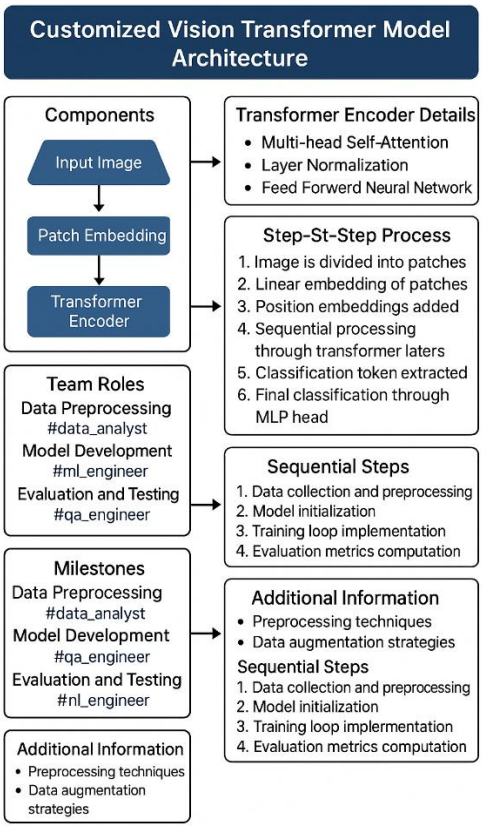


Fig. 2. Customized vision transformer model architecture.

International Multidisciplinary Research Journal Reviews (IMRJR)
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602

3.5 Experimental setup

The following steps were followed to implement the customized Vision Transformer (ViT) model on the dataset:

- **Library Initialization**: All required libraries and frameworks necessary for model construction and training were imported, including those for deep learning and image processing.
- **Hyperparameter Configuration**:
  Key hyperparameters were fine-tuned to optimize model performance. These included:
  - **Number of epochs**: Experimentation was carried out using multiple epoch counts (e.g., 50, 100, 150, 200), with 250 and 500 yielding the most stable performance.
  - **Attention heads**: Set to 2, aligned with the number of classification categories in the dataset to enhance attention-based learning.
  - **Input shape and image size**: Defined based on the dimensions of the preprocessed image data.
  - **Patch size**: Adjusted according to the desired granularity for feature extraction.
- **Data Augmentation**:
  An image augmentation pipeline was implemented to increase dataset diversity and improve generalization. This included standard augmentation techniques such as flipping, rotation, scaling, and brightness adjustments.
- **Patch Extraction**:
  Images were divided into smaller patches to serve as input tokens for the Vision Transformer. For example, for a representative input image with dimensions 32 × 32 pixels and a patch size of 6 × 6, approximately 25 non-overlapping patches were generated per image. Each patch was flattened into a vector to feed into the transformer encoder.

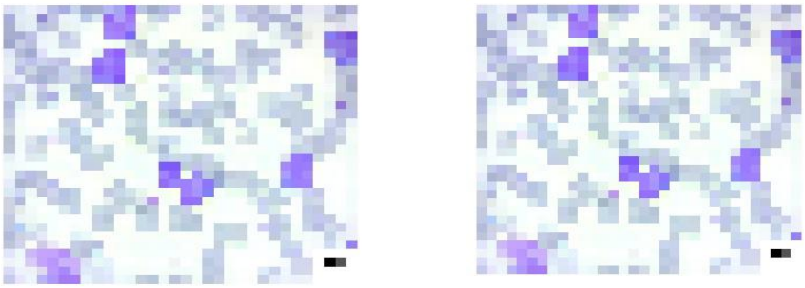


Fig. 3. Output of patch creation.

- **Patch Encoder**
- **Constructing customized Vit classifier.**
- **Run experiment:** The whole experiment ran on the dataset and then retrieved the result of our customized model's performance.

3.6 Customized ViT model

Table 1. A short instance of our customized ViT.

Total Images	3242	Train	2391
		Validation	526
		Test	325
Number of classes on the dataset	4		
Number of head	2(selected),4,8,10		
Number of Epoch	500,1000,1500,2000,2500(selected),5000		
Optimized	Adam		
Learning rate	0.0001		
Weight decay	0.00001		
Batch size	256		
Projection dim	128		
Augmentation model	Keras.Sequential		
Mlp Head Unit	[2048,1024]		
Dropout	0.2		
Dropout rate	0.2		

International Multidisciplinary Research Journal Reviews (IMRJR)
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602

Experiment result

Table 2. Accuracy of validation and test set, Training time of the customized model according to different image and patch size.

Image size	Patch size	Accuracy [%]	Training time [s]
32	6	98.15%	271
48	6	97.85%	263
56	7	97.23%	198
64	12	97.54%	200
32	8	98.77%	200
28	7	97.54%	263
21	7	97.54%	199
28	4	96.62%	139

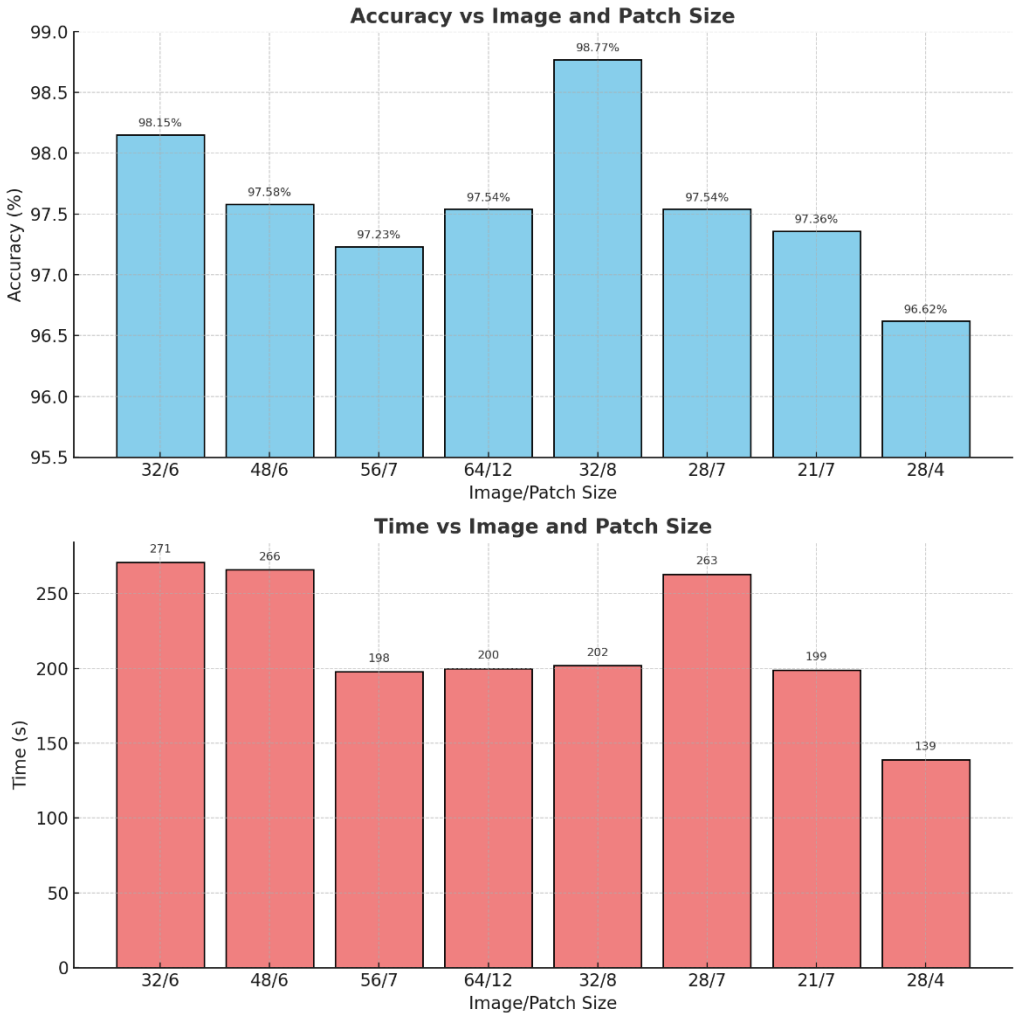
Table 3. Classification report of Validation and Testing according to image size 32 and patch size6.

mage size: 32				Patch size:6	
	Class	**precisio n**	**recall**	**f1-score**	**support**
Validation	Benign	0.99	0.98	0.98	0.93
	Malignant early Pre B	0.99	0.99	0.99	156
	Malignant Pre B	1.00	1.00	1.00	158
	Malignant Pro B	1.00	1.00	1.00	119
	Accuracy			0.99	526
Testing	Benign	0.94	0.94	0.94	51
	Malignant early Pre B	0.98	0.97	0.97	98
	Malignant Pre B	0.99	0.99	0.99	96
	Malignant Pro B	0.99	1.00	0.99	80
	Accuracy			0.978	325

IV. DISCUSSION

In this research paper, we propose an innovative vision transformer model designed for analyzing a dataset related to mango leaf diseases. We conducted experiments using our custom model with different epoch settings and determined that 250 epochs yielded the best results. Additionally, we explored the effect of varying the figure of attention heads on validity and found that aligning the number of heads with the dataset's class count produced optimal outcomes. Furthermore, we investigated diverse image and patch sizes to fine-tune the dataset and evaluate the performance of our models. We have displayed the accuracy matrix for the entire execution period in Table 2. The conclusion is that, among all patch sizes and picture sizes, our model validates the dataset best when image size and patch size are 32/8, 48/6, and 64/12. Again, the 32/6 patch and image size performed best for testing accuracy. However, 56/7, 21/7, 28/4 required less time to train the dataset. Although it required more time than the 28/7 image and patch sizes, the 48/6 image and patch size had the best validation and testing accuracy of all. Table 4 further demonstrates that while testing results are fewer when using smaller images and patches, larger images and patches acquire more information with time.

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602



V. INFERENCE OF THE CURRENT STUDY

Recent deep learning algorithms are less effective during infections and only capable of detecting Blood Cell Cancer. ViT can identify the Blood Cell Cancer where the disease is present by highlighting key components of the blood cell images, providing with vital information.

VI. FUTURE SCOPES

Additionally, we will analyze this dataset to create confusion matrices for testing and validation, taking into account different picture and patch sizes. In addition, we'll look into how varying picture and patch sizes affect the loss function graphs used for training and validation. However, blood cell cancer disorders pose a serious threat to people and have a profound effect on the body. According to the World Health Organization (WHO), cancer will be the primary cause of over 10 million deaths worldwide in 2020, or one in every six deaths. We therefore plan to extend our suggested methodology to other cancer conditions in order to increase the medical industry's resilience.

VII. CONCLUSION

With a particular emphasis on the development of computer vision techniques, this study explores the changing landscape of cancer solutions. Vision transformers (ViT) stand out as a relatively recent and fascinating development among these cutting-edge innovations. ViT demonstrates the ability to quickly and precisely categorize and identify a variety of Blood Cell Cancer. ViT has emerged as a strong option for image processing in the medical field as a result of its rising popularity across numerous sectors. In the end, these technological advancements aim to address the urgent problem of early diagnosis and efficient control of Blood Cell Cancer.

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 6, June 2025
DOI 10.17148/IMRJR.2025.020602

REFERENCES

[1]. Malaviya, N., Rahevar, M., Virani, A., Ganatra, A., & Bhuva, K. (2023, January). LViT: Vision Transformer for Lung cancer Detection. In 2023 International Conference on Artificial Intelligence and Smart Communication (AISC) (pp. 93-98). IEEE. doi:org/10.21597/jist.1183679Khaskheli, M. I. (2020). Mango Diseases: Impact of Fungicides. Horticultural Crops, 143. doi: 10.5772/intechopen.87081

[2]. Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., ... & Goldstein, T. (2022). What do vision transformers learn? a visual exploration. arXiv preprint arXiv:2212.06727. doi:org/10.48550/arXiv.2212.06727

[3]. Sun, T., Zhu, Q., Yang, J., & Zeng, L. (2022). An improved Vision Transformer model for the recognition of blood cells. Sheng wu yi xue Gong Cheng xue za zhi= Journal of Biomedical Engineering= Shengwu Yixue Gongchengxue Zazhi, 39(6), 1097-1107. doi:org/10.7507/1001-5515.202203008

[4]. Gokhale, M., Mohanty, S. K., & Ojha, A. (2023). GeneViT: Gene vision transformer with improved DeepInsight for cancer classification. Computers in Biology and Medicine, 155, 106643. doi:org/10.1016/j.compbiomed.2023.106643.

[5]. 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. doi:org/10.48550/arXiv.2010.11929

[6]. Chen, S., Lu, S., Wang, S., Ni, Y., & Zhang, Y. (2023). Shifted Window Vision Transformer for Blood Cell Classification. Electronics, 12(11), 2442. doi:org/10.3390/electronics12112442

[7]. Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., ... & Chen, H. (2022). An improved transformer network for skin cancer classification. Computers in Biology and Medicine, 149, 105939. doi:org/10.1016/j.compbiomed.2022.105939.

[8]. Saad, M., Ullah, M., Afridi, H., Cheikh, F. A., & Sajjad, M. (2022, October). BreastUS: Vision Transformer for Breast Cancer Classification Using Breast Ultrasound Images. In 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 246-253). IEEE. doi: 10.1109/SITIS57111.2022.00027

[9]. Wang, W., Jiang, R., Cui, N., Li, Q., Yuan, F., & Xiao, Z. (2022). Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. Frontiers in Pharmacology, 13, 929755. doi:org/10.3389/fphar.2022.929755.

[10]. Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S., & Islam, M. (2022). An effective skin cancer classification mechanism via medical vision transformer. Sensors, 22(11), 4008. doi:org/10.3390/s22114008

[11]. Mali, M. T., Hancer, E., Samet, R., Yıldırım, Z., & Nemati, N. (2022, September). Detection of Colorectal Cancer with Vision Transformers. In 2022 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE. doi: 10.1109/ASYU56188.2022.9925335

[12]. PACAL, İ. (2022). Deep learning approaches for classification of breast cancer in ultrasound (US) images. Journal of the Institute of Science and Technology, 12(4), 1917-1927. doi:org/10.21597/jist.1183679

[13]. Das, S. K., Islam, K. S., Neha, T. A., Khan, M. M., & Bourouis, S. (2021). Towards the Segmentation and Classification of White Blood Cell Cancer Using Hybrid Mask-Recurrent Neural Network and Transfer Learning. Contrast Media & Molecular Imaging, 2021. doi:org/10.1155/2021/4954854

[14]. Li, T. Z., Xu, K., Gao, R., Tang, Y., Lasko, T. A., Maldonado, F., ... & Lanman, B. A. (2023, April). Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. In Medical Imaging 2023: Image Processing (Vol. 12464, pp. 221230). SPIE. doi:org/10.1117/12.2653911

[15]. Paul, S., & Chen, P. Y. (2022, June). Vision transformers are robust learners. In Proceedings of the AAAI conference on Artificial Intelligence (Vol. 36, No. 2, pp. 2071-2081). doi:org/10.1609/aaai.v36i2.20103

[16]. Kutlu, H., Avci, E., & Özyurt, F. (2020). White blood cells detection and classification based on regional convolutional neural networks. Medical hypotheses, 135, 109472. doi:org/10.1016/j.mehy.2019.109472

[17]. Tiwari, P., Qian, J., Li, Q., Wang, B., Gupta, D., Khanna, A., ... & de Albuquerque, V. H. C. (2018). Detection of subtype blood cells using deep learning. Cognitive Systems Research, 52, 1036-1044. doi:org/10.1016/j.cogsys.2018.08.022