Hybrid Vision Transformer and MambaVision Network for Leukemia Subtype Classification

Prof. (Dr.) Sandipkumar Ramanlal Panchal[1] *, Prof. (Dr.) Sai Kiran Oruganti*[2]

[1]Postdoctoral Researcher, Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia & Professor, E & C Engineering, School of Engineering and Technology, Dr. Subhash University, Junagadh, Gujarat, India; [2]Professor, Faculty of Engineering and Built Science Lincoln University College, Petaling Jaya, Selangor, Malaysia

pdf.sandipkumar@lincoln.edu.my; saisharma@lincoln.edu.my

Abstract: For the quick diagnosis and effective treatment plan, peripheral blood smear images can be used for leukemia subtype classification. We propose a Hybrid Vision Transformer and MambaVision Network for automated classification of four major leukemia subtypes: Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL) and Chronic Myeloid Leukemia (CML). The proposed framework first processes the input images to make the cellular features more visible. The improved images are further passed to parallel branches: Vision Transformer (ViT) and MambaVision to extract complementary global contextual features and efficient sequential representation. The ViT branch then uses a cross-attention module to utilize the informative features extracted by the MambaVision, concatenates them for feature fusion, and classifies the features using a multilayer perceptron (MLP). The experimental results show that the proposed hybrid architecture has a classification accuracy of 95.20% over all the four types of leukemia, while also outperforming the individual backbone models and provides a powerful and efficient approach towards computer-aided diagnosis of leukemia.

Keywords: Leukemia Subtype Classification; Vision Transformer (ViT); MambaVision; Cross-Attention Feature Fusion; Peripheral Blood Smear Images

Introduction

In medical image analysis field, deep learning has emerged as an effective technique for automated disease diagnosis, including leukemia in the hematological disorders. Proper classification of leukemia subtypes in peripheral blood smear images is very important for early diagnosis, choosing treatment and survival of the patient. Automated classification is a challenging task, however, due to the differences in cell morphology, staining quality, imaging conditions and high similarity among the different leukemia subtypes. In recent years, the deep learning techniques, particularly transformer models, have significantly raised the accuracy of feature learning and classification in leukemia diagnosis [1,2,3,4,5,6,7,10,11].

Convolutional neural networks (CNNs), Vision Transformers (ViTs), hybrid CNN-Transformer architectures, attention mechanisms, explainable learning frameworks, multimodal vision-language networks, and federated learning frameworks have been proposed for leukemia classification by several studies [5,6,12,13,14,15,16,17]. Recently, Mamba based vision architectures have shown great potential in capturing the long-range dependencies with reduced computational complexity, which are applicable to medical image analysis tasks [8,9,18,19,20]. While these methods have shown to be effective, most

current methods use only a single backbone network or simple feature fusion methods, which fails to fully utilize the complementary global contextual and sequential representations.

The proposed framework includes two parallel branches: a Vision Transformer branch and a MambaVision branch, which is to enhance the peripheral blood smear images and process them.

Methodology

The framework uses a Hybrid Vision Transformer (ViT) and MambaVision Network for automatic classification of peripheral blood smear images into leukemia subtypes. The experimental data is taken from sheikhlubna/subtypes-of-leukemia-blood-cell-microscopic-image repository, which is open to the public. There are 600 microscopic images of blood smears for each of the four leukemia subtypes (Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL) and Chronic Myeloid Leukemia (CML)), for a total of 2,400 images. A lot of data augmentation is done to achieve good generalization with reduced overfitting, and this involves random rotation, horizontal/vertical flipping, zooming, translation, brightness adjustment and contrast variation. Each augmented image from each type is 1000 images, thus a balanced set of 4,000 images is created. The data is then split into a training set, a validation set, and a test set, and the data is used to develop the models and to assess their performance.

First, a pre-processing step is performed to resize the images of the blood smears to a fixed resolution and to normalise the images to remove illumination variations and to create uniformity in the features. Assume that the input image is given by,

$$I_n = \frac{I-\mu}{\sigma}, \tag{1}$$

where I indicates the original image, μ is the mean and σ is the standard deviation. $I_n$is the normalized image used as the input to the hybrid network.

The preprocessed image is concurrently sent to the two parallel feature extraction branches. The Vision Transformer (ViT) is a vision model that uses self-attention to learn long-range dependencies and divides the image into non-overlapping patches to capture the global contextual relations. Meanwhile, the MambaVision backbone of the model is able to capture the sequential and spatial dependencies of the data with the help of state-space representations, which are very efficient for extracting features.Meanwhile, the MambaVision back bones of the model are able to capture sequential and spatial dependency of the data using the help of state space representation which are very efficient in extracting features without losing any global information. The extracted feature representations from both branches are represented by,

$$F_{ViT} = f_{ViT}(I_n), F_{Mamba} = f_{Mamba}(I_n), \tag{2}$$

where F_{ViT} and F_{ViT} correspond to the features learned by the Vision Transformer network and MambaVision network, respectively.

To exploit the complementary information from both branches, a cross-attention module enables the ViT features to attend to the informative representations generated by MambaVision. The refined features are then combined together to get a combined representation,

$$F_{fusion} = F_{ViT} \oplus F_{Mamba}, \tag{3}$$

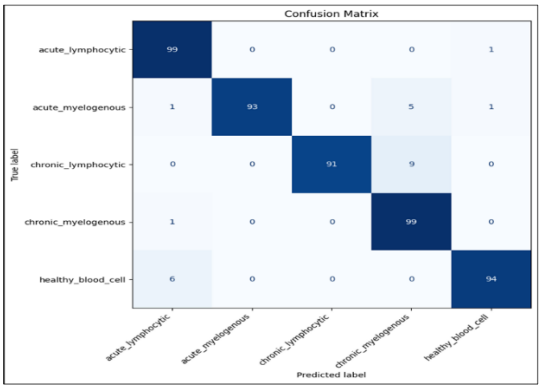
The "⊕" symbol is used to denote feature concatenation. This fused feature vector is then fed into a Multilayer Perceptron (MLP) classifier with fully connected layers and Softmax activation function, which provides estimates of probabilities of the respective leukemia sub-types. The predicted class is given by

$$\hat{y} = \arg\max(\text{Softmax}(\text{MLP}(F_{fusion}))), \quad (4)$$

Where, $\hat{y} \in$ {ALL,AML,CLL,CML}. Overall classification accuracy of 95% is achieved while the proposed hybrid architecture is capable of achieving better discrimination among leukemia subtypes, the key advantage of Vision Transformers over the MambaVision model.

Results

The proposed model was implemented in Python 3.12, with PyTorch framework and using Google Colab with GPU acceleration. The confusion matrix is shown in **Figure 1**, where most predictions are on the diagonal, showing minimal confusion between the four leukemia types (ALL, AML, CLL, CML), with few predictions being misclassified. The representative testing results on the proposed hybrid framework are shown in **Figure 2**, indicating that the hybrid framework can accurately classify the leukemia subtype.



Classification Report

	precision	recall	f1-score	support
acute_lymphocytic	0.9252	0.9900	0.9565	100
acute_myelogenous	1.0000	0.9300	0.9637	100
chronic_lymphocytic	1.0000	0.9100	0.9529	100
chronic_myelogenous	0.8761	0.9900	0.9296	100
healthy_blood_cell	0.9792	0.9400	0.9592	100
accuracy			0.9520	500
macro avg	0.9561	0.9520	0.9524	500
weighted avg	0.9561	0.9520	0.9524	500

Figure 1. Confusion Matrix and Classification Report

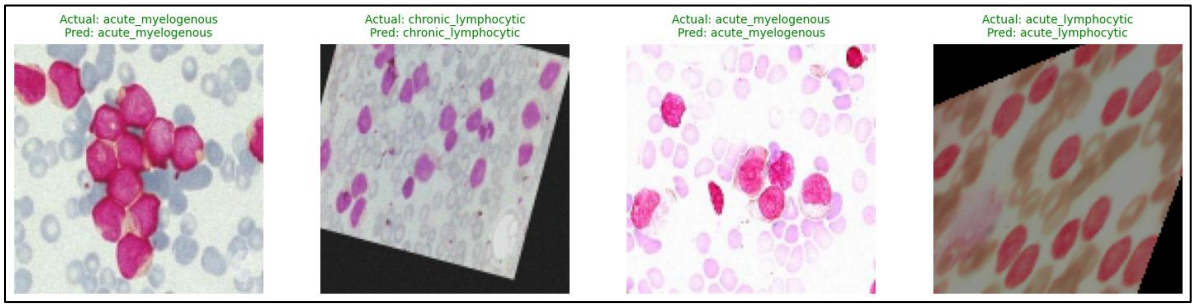


Figure 2. Testing Results

Table 1. Comparative Performance Analysis with Existing Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Multi-Scale Feature Fusion [1]	91.80	91.45	91.20	91.32
LUMINA-Net [2]	92.65	92.81	92.43	92.61
LeukoViT-UNeT [4]	93.74	93.56	93.68	93.62
Hybrid ResNet-50 + ViT [12]	94.21	94.08	94.15	94.11
Hybrid VGG16-ViT [15]	94.67	94.51	94.58	94.54
Proposed Hybrid ViT + MambaVision Network	95.20	95.61	94.20	95.24

Table 1 presents the comparative performance of the proposed Hybrid ViT-MambaVision Network against representative state-of-the-art leukemia classification methods. The proposed approach achieves the highest overall accuracy, precision, recall, and F1-score, demonstrating its effectiveness in accurately

classifying the four leukemia subtypes through complementary feature learning and cross-attention fusion.

Conclusions

The research has proposed a Hybrid Vision Transformer and MambaVision Network model for automated classification of leukemia subtypes based on peripheral blood smear images, which overcame the shortcomings of the previous single backbone deep learning models that fail to capture complementary feature representations. To address these challenges, the proposed framework consists of parallel branches of Vision Transformer and Mamba Vision, in combination with a cross-attention feature fusion mechanism, to learn the global context information and the sequential dependency information effectively for robust subtype discrimination. Experimental evaluation shows that the proposed model can classify four major leukemia types, ALL, AML, CLL and CML, and achieve an overall classification accuracy of 95.20% with balanced precision, recall and F1 score for all classes. The suggested framework has shown good performance but it has been tested with a single public dataset and a set hybrid architecture. Moving forward, further studies are required to validate the model on larger multi-center datasets, integrate the explainable artificial intelligence (XAI) techniques, refine the cross-attention fusion method, and expand it to other hematological diseases and clinical decision support systems (CDSS).

References

[1] W. Yang *et al.*, "A multi-scale feature fusion of deep learning network for classifying acute leukemia genetic subtypes from blood smear images," *Discover Oncology*, vol. 17, no. 1, 2026, doi: 10.1007/s12672-025-03985-z.

[2] O. Attallah, "LUMINA-Net: Acute Lymphocytic Leukemia Subtype Classification via Interpretable Convolution Neural Network Based on Wavelet and Attention Mechanisms," *Algorithms*, vol. 19, no. 4, pp. 1–35, 2026, doi: 10.3390/a19040298.

[3] M. Momenian and S. V. Shojaedini, "Multimodal vision-language framework for text-guided leukemia classification using advanced deep learning architectures," *Discover Artificial Intelligence*, vol. 6, no. 1, 2026, doi: 10.1007/s44163-026-01239-7.

[4] B. Revathi and M. Kaliappan, "LeukoViT-UNeT: A U-Net and Vision Transformer Hybrid approach for Leukemia Cell Segmentation and Classification," *International Journal of Computational Intelligence Systems*, vol. 19, no. 1, 2026, doi: 10.1007/s44196-026-01356-2.

[5] K. Parwez, S. I. Sohail, A. Akram, J. Rashid, G. Atteia, and N. Sarwar, "Explainable federated transformer framework for joint leukemia classification and stage prediction," *Scientific Reports*, vol. 16, no. 1, 2026, doi: 10.1038/s41598-025-34715-8.

[6] F. Jammal, "An explainable meta-learned hybrid CNN-transformer model with dual attention for leukemia diagnosis from peripheral blood smears," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-55606-6.

[7] J. Kasprowicz and A. G. Silva, "Comparative analysis of convolutional and vision transformer models for automated leukocyte classification enhanced by generative color augmentation," *Signal, Image and Video Processing*, vol. 20, no. 5, p. 1, 2026, doi: 10.1007/s11760-026-05309-2.

[8] X. Liu, C. Zhang, F. Huang, S. Xia, G. Wang, and L. Zhang, "Vision Mamba: A Comprehensive Survey and Taxonomy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 37, no. 2, pp. 505–525, 2026, doi: 10.1109/TNNLS.2025.3610435.

[9] L. Clifton, X. Tian, D. Palasuwan, P. Watanaboonyongcharoen, P. Rojnuckarin, and N. Anantrasirichai, "Mamba-Based Ensemble Learning for White Blood Cell Classification," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2026-April, pp. 6–10, 2026, doi: 10.1109/ISBI61048.2026.11515403.

[10] D. Thamer and Adnan Mohsin Abdulazeez, "A Review on White Blood Cell Classification for Leukemia Diagnosis Using Deep and Transfer Learning Techniques," *The Indonesian Journal of Computer Science*, vol. 15, no. 1, pp. 1227–1240, Feb. 2026, doi: 10.33022/ijcs.v15i1.5083.

[11] S. B. Mohan, S. Sathya, S. Rajalakshmi, G. Gurumoorthy, and R. Sivanraju, "Efficient convolutional neural networks for acute lymphoblastic leukaemia prediction in computer vision," *Scientific Reports*, vol. 16, no. 1, p. 1135, Dec. 2026, doi: 10.1038/s41598-025-30777-w.

[12] J. Hendrawan, T. F. Halim, J. B. Gusman, and W. Budiharto, "Hybrid deep learning model for acute lymphoblastic leukemia (ALL) detection using pre-trained ResNet-50 and Vision Transformer architecture," *Procedia Computer Science*, vol. 269, pp. 933–942, 2025, doi: 10.1016/j.procs.2025.09.036.

[13] A. Alabduljabbar, M. Awais, T. Akram, and Y. N. Altherwy, "A Hybrid Framework for Acute Lymphoblastic Leukemia Identification Utilizing VIT-CNN Fusion and Immunologically Inspired Deep Feature Selection," *IEEE Access*, vol. 13, no. September, pp. 180355–180374, 2025, doi: 10.1109/ACCESS.2025.3622660.

[14] M. Hameed, M. A. Z. Raja, A. Zameer, H. S. Dar, A. S. Alluhaidan, and R. Aziz, "Acute myeloid leukemia classification using ReLViT and detection with YOLO enhanced by adversarial networks on bone marrow images," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-17891-5.

[15] M. S. Ali *et al.*, "A Hybrid VGG16-ViT Approach With Image Processing Techniques for Improved White Blood Cell Classification and Disease Diagnosis: A Retrospective Study," *Health Science Reports*, vol. 8, no. 6, pp. 1–13, 2025, doi: 10.1002/hsr2.70859.

[16] O. M. Alshehri *et al.*, "A Hybrid CNN-Transformer Framework for Normal Blood Cell Classification: Towards Automated Hematological Analysis," *CMES - Computer Modeling in Engineering and Sciences*, vol. 144, no. 1, pp. 1165–1196, 2025, doi: 10.32604/cmes.2025.067150.

[17] S. K. Yohesha and R. Dheepthi, "Hybrid Deep Learning Approach for Non-Hodgkin's Lymhoma using ViT and ResNet," in *Proceedings of the International Conference on Intelligent Systems and Digital Transformation (ICISD 2025)*, no. Icisd 2025, Atlantis Press International BV, 2025, pp. 36–45. doi: 10.2991/978-94-6463-866-0_5.

[18] A. Hatamizadeh and J. Kautz, "MambaVision: A Hybrid Mamba-Transformer Vision Backbone," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, no. 1, pp. 25261–25270, 2025, doi: 10.1109/CVPR52734.2025.02352.

[19] S. Bansal *et al.*, "A Comprehensive Survey of Mamba Architectures for Medical Image Analysis: Classification, Segmentation, Restoration and Beyond," *Computer Methods and Programs in Biomedicine*, Oct. 2025, [Online]. Available: http://arxiv.org/abs/2410.02362

[20] H. Zhang *et al.*, "A Survey on Visual Mamba," *Applied Sciences*, vol. 14, no. 13, p. 5683, Jun. 2024, doi: 10.3390/app14135683.