Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Vision Transformer based Hybrid Approach for White Blood Cells (WBC) Classification

Kiruthiga K.[1], Umamaheswari K.[2]

[1-2]Department of Information Technology, PSG College of Technology, Coimbatore, India

E-mail: [1]23pb32@psgtech.ac.in, [2]hod.it@psgtech.ac.in

Abstract

Proper classification of white blood cells (WBCs) remains a critical task for medical diagnosis in leukemia, infections, and hematological disorders. WBC classification is based mainly on Convolutional Neural Networks (CNNs) as the main methodology, despite the challenges in such methods in scanning microscopic images for distant dependencies. The current study adopts a hybrid model that combines Vision Transformers (ViTs) with CNNs for better WBC classification. The model has a ViT component that uses self-attention mechanisms to obtain whole-image features through global features extraction, while at the same time, neighborhood information is being extracted by the CNN module. The introduced model is evaluated using theBCCD Dataset (Blood Cell Count and Detection) for binary and multi-class separation of granulocytes/agranulocytes. Experimental tests indicate that the combined approach has robust classification accuracy with guaranteed reliable results in various evaluation metrics. The suggested method achieved 99.20% training accuracy, 87.90% test accuracy, a precision of 0.7083, a recall of 0.7000, and an F1-score of 0.6970, affirming its excellent classification precision. This hybrid deep learning model shows enhanced performance and interpretability, supporting its application in clinical diagnostic processes. These findings attest to excellent classification performance, positioning the model as an excellent choice for automated hematological diagnosis. The study opens up new horizons for combining CNNs and ViTs for improved medical image analysis while setting future development goals in WBC classification. Deep learning hybrid models have shown

DOI: https://doi.org/10.36548/jucct.2025.2.004
Received: 29.04.2025, received in revised form: 02.06.2025, accepted: 19.06.2025, published: 04.07.2025

their significance for clinical diagnostics in the study through the development of enhanced detection systems against hematological diseases.

Keywords: Vision Transformers, White Blood Cells, Hybrid Model, F1-Score, Deep Learning.

1. Introduction

The modern combination of medical imaging technology with artificial intelligence systems has transformed blood cell analysis through the automated, precise identification of white blood cells. Medical professionals need precise WBC classification because abnormal WBC counts and morphological variations serve as diagnostic indicators for leukemia, as well as infections and other hematological disorders [17]. The domain of white blood cell classification succeeds through traditional machine learning with CNNs by using hierarchical features for extraction. The ability of CNNs to detect long-range dependencies and global context remains weak because these elements are necessary for complete cell morphology analysis. International Corporation and the University Hospital of Wales have comparable advantages and limitations in their processing techniques for respiratory diseases.

This paper develops a WBC classification method by combining Vision Transformer (ViT) technology with CNN techniques to facilitate their complementary performance. The local spatial features extracted by CNNs combine effectively with ViTs, which use self-attention mechanisms to understand long dependencies and global relationships, leading to higher interpretable classification outcomes. In order to perform more accurate WBC morphology assessments, the hybrid model combines the global contextual learning capabilities of ViTs with CNN local feature extraction technologies to optimize high-resolution microscopic image processing.

The research utilizes a widely recognized WBC classification dataset called the BCCD Dataset (Blood Cell Count and Detection). This model undergoes testing for binary classification of granulocytes versus agranulocytes together with multi-class subtype classification of WBCs to determine actual clinical application performance. A clinical deployment analysis of this hybrid approach depends on performance evaluations through accuracy, precision, recall, F1-score and computational efficiency metrics. This study investigates model interpretability through the analysis of attention maps in ViTs which offers

visual explainability for classification choices to benefit medical clinicians in understanding AI-assisted diagnostic decisions.

In the next sections, we explore the theoretical underpinnings of Vision Transformers, discuss integrating CNNs for hybrid modeling, and examine experimental findings that show how effective the suggested method is. This work opens the door for more automation in clinical decision-making by advancing the development of reliable, accurate, and interpretable AI-driven hematological diagnostic systems. The results demonstrate the promise of hybrid deep learning models in personalized medicine, medical image analysis, and AI-assisted healthcare, with the ultimate goal of enhancing patient outcomes and early disease identification.

2. Research Objective and Contributions

This research aims to develop an interpretable and high-performing hybrid deep learning model that combines CNNs and Vision Transformers for classifying white blood cells. The major contributions are:

- A novel hybrid architecture that fuses CNN-based local feature extraction and ViT-based global attention mechanisms.
- Enhanced interpretability through attention heatmaps.
- Improved accuracy and F1-score across binary and multi-class classification tasks.
- Application of advanced preprocessing techniques, including edge contouring.
- Validation using theBCCD dataset and clinical evaluation metrics.

3. Literature Review

In the study, a review of machine learning methods for classifying white blood cells is conducted as presented by Asghar et al. [1]. Among numerous approaches from multiple datasets and regions this work performs a comprehensive analysis to understand WBC classification method development. Through this study, the authors created a robust base for analyzing algorithms together with performance evaluation challenges in medical imaging.

The study of deep and hybrid learning strategies for disease diagnosis through blood microscopy analysis is presented in detail by Almurayziq et al. [2]. Early identification of WBC disorders achieves improved results through advanced neural architecture applications, according to the authors' emphasis. The research makes an essential contribution to ensemble model development by proving that ensemble methods can strengthen diagnostic precision in medical applications. Rao and Battula [3] developed a new deep learning method that combines MobileNetV3 and ShuffleNetV2 for WBC segmentation and classification purposes. This model performs well yet maintains efficient computations, which enable it to support real-time diagnostic tool requirements. The architectural design stands as an example of efficient yet high-performance neural networks that are suitable for medical usage.

The paper by Yentrapragada [4] proposes the use of CNNs to detect and classify leukocytes automatically. This study proves that deep features enhance both classification accuracy and minimize the frequency of false positive predictions. Deep feature extraction plays a crucial role in making models more resistant to complex blood smear images. Elhassan et al. [5] built a hybrid system that utilizes a deep convolutional autoencoder in the first stage, then applies a CNN framework for atypical WBC classification in acute myeloid leukemia diagnosis. Multi-stage system designs prove essential for challenging datasets in clinical settings, according to their research, which demonstrates the suitability of autoencoders for medical image preparation and noise reduction tasks.

The research by Olayah et al. [6] describes a classification system that unites CNNs with pre-designed features to analyze blood slide images. Research evidence shows that handcrafted features combined with learned features create synergistic effects that lead to better interpretability and performance for predicting outcomes from hematology-based assessments. The research of Cheuque et al. [7] presents a multi-level CNN method to improve WBC classification through network layer features. Architectural depth and multiscale analysis enable an accurate outlook on cellular specifics, which leads to improved generalization when classifying cells.

Jung et al. [8] established a one-of-a-kind system that unites the predictive capabilities of CNN-based classification with generative models to perform dual functions of predictions and synthetic data creation. Medical institutions benefit substantially from this development because it helps to expand scarce datasets while improving training results. The researchers at Davamani et al. [9] created a metaheuristic hybrid learning system to conduct adaptive

blood cell segmentation followed by classification tasks. A combination of optimization techniques and neural networks forms their approach to handilng blood smear image variations for medical diagnostic applications.

A framework for WBC classification based on CNN is enhanced through a PSO algorithm, according to Balasubramanian et al. [10]. Model performance increases by using PSO for hyperparameter optimization, which demonstrates that these computational approaches complement each other. The WBC classification pipeline includes CNN-based deep learning with statistical feature alignment, according to Patil et al. [11]. The implementation of their method produces statistical layers for feature alignment, leading to improved classification accuracy and discrimination capability.

Cinar and Tuncer developed a system that utilizes AlexNet together with GoogleNet, supported by SVM, for WBC subtype identification [12]. The research shows how traditional machine learning approaches combined with deep learning improve multiclass classification capabilities by producing outstanding results, according to its findings. Khan et al. [13] delivered a review that evaluated WBC classification methods through traditional machine learning and deep learning algorithms in blood smear images. The synthesis of traditional and contemporary methods presented in this work offers a comprehensive representation of algorithm development trends that can be referenced for evaluating new framework models.

The research by Malkawi et al. [14] introduces a hybrid WBC classification system based on CNN networks, which demonstrates the current scientific interest in combining multiple architectures to enhance biomedical application reliability and accuracy. The paper by Ozyurt et al. [15] demonstrates a CNN fusion system that employs both MRMR (Minimum Redundancy Maximum Relevance) feature selection alongside an extreme learning machine (ELM). The proposed model illustrates how combining feature selection with classification improves accuracy while highlighting the strong impact of statistical analysis with neural networks in biomedical imaging research.

The study [16] reviews current WBC classification research by focusing on how deep learning technologies merge with engineered features while using optimization methods. A combination of research study findings warrants the creation of the “Vision Transformer-based Hybrid Approach” to advance WBC classification capabilities that meet clinical needs.

The project establishes numerous opportunities for future research along with development activities. Future advancements in Vision Transformer (ViT) architectures that incorporate domain-specific tokenization and adaptive attention methods support higher accuracy rates and increased operational efficiency. The combination of XAI frameworks LIME and SHAP with the model will enhance transparency levels, enabling clinicians to trust the system more. Improving the mobile/embedded hybrid solution holds promising benefits for diagnosing patients at point-of-care locations. Multiple standardized benchmark tests conducted across various WBC datasets will establish general diagnostic performance for different population types and laboratory testing environments. A healthcare implementation needs regulatory compliance, together with integrated data privacy measures easy-to-use interfaces, and interfaces for successful deployment. The model requires real-world clinical testing as well as collaboration with medical experts to validate its performance so that researchers can develop robust, interpretable, and scalable diagnostic systems in hematology.

There are varying numbers of white blood cell (WBC) classification studies that have been established but have had glaring weaknesses. Almurayziq et al. [2] used deep and hybrid learning methods, but these were not tested in real-time clinical settings and were not interpretable. Equally, Rao and Battula [3] proposed a lightweight deep architecture encompassing the use of MobileNetV3 and ShuffleNetV2, but their model exhibited high performance nature at the expense of limited generalizability across environments. Elhassan et al. [5] proposed a two-stage Autoencoder-CNN hybrid for the diagnosis of leukemia, which, although accurate, was not embedded system-friendly, and the training complexity was high. Cinar and Tuncer [12] combined AlexNet, GoogleNet, and SVM in an effort to improve the multiclass classification rate; however, their model tended toward overfitting, especially with a diversified class distribution. Finally, Ozyurt et al. [15] proposed a fusion modeling approach using MRMR feature selection and CNNs with extreme learning machines; the method achieved limited robustness when dealing with skewed or noisy data. All these shortcomings support the necessity of a more balanced solution, like the hybrid ViT-CNN model proposed, which would not only provide enhanced performance in terms of classification but also, due to its increased interpretability, scalability, and clinical usability, would prove to be more clinically applicable.

4. Methodology

In this method, a strong pipeline combines the advantages of Vision Transformers (ViTs) for collecting global contextual dependencies and Convolutional Neural Networks (CNNs) for comprehensive local feature extraction, forming the basis of the WBC classification process. The WBC classification methodology integrates robust pipeline elements that process detailed local features with CNNs while leveraging ViTs to capture overall contextual dependencies. Dataset acquisition from BCCD requires preprocessing as the first step of the workflow. The dataset provides labeled microscopic images that contain the four WBC types: eosinophils, lymphocytes, monocytes, beside neutrophils. Data preprocessing achieves these tasks through three sequential steps: image resolution standardization, pixel value normalization to [0, 1] spectrum and edge detection implementation with Canny and Sobel filters for boundary enhancement.

WBC contour extraction generates a controlled cell area, which cropping transforms into the model's processing zone. The model receives better generalization capabilities by undergoing various image augmentation procedures, including flipping, rotations, brightness alteration, and scaling. The images undergo demographic procedures before they are divided into training, validation, test dataset segments.

Mathematical Formulation of Proposed Model:

Let the input image be denoted as,

$$I \in R^{H*W*C}$$

where H, W, C represent the height, width, and number of channels, respectively

The Convolutional Neural Network (CNN) is first applied to extract local spatial features:

$$F_{CNN}=CNN(I)$$

The output $F_{CNN} \in R^{N*D}$ is then flattened and passed as sequential patch tokens to the Vision Transformer (ViT). The embedding of each token is defined as follows:

$$T= \text{Flatten } (F_{CNN})* W_E+PE$$

where W_E is the trainable embedding matrix and PE is the positional encoding added to retain spatial information. The ViT applies the self-attention mechanism, calculated as:

$$\text{Attention } (Q, K, V) = \text{Softmax}(QK^T/\sqrt{(d_k)})V$$

with query, key, and value matrices derived from the input token T:

$$Q = TW^Q, K = TW^K, V = TW^V$$

After several transformer encoder layers, the final feature representation is passed through a fully connected layer for classification:

$$y = \text{Softmax } (W_O * F_{ViT})$$

where WO is the output weight matrix and FViT is the transformed feature embedding. This pipeline allows for both local pattern extraction and global context modeling, which are essential for accurate WBC classification. The hybrid model structure initiates its operations by utilizing a CNN block to obtain high-quality local features. The convolutional layers analyze edge features together with texture elements and morphological properties, whereas pooling layers minimize complexity while decreasing dimensions. The robustness of the system derives from feature extraction methods that process details at different scales simultaneously. The robust solution is possible through the multi- scale feature extraction process which handles microscopic and macroscopic information. The Vision Transformer receives for processing flattened output from the CNN. The ViT obtains global dependencies between spatial features by utilizing self- attention as well as cross-attention mechanisms. Spatial position information through encoding serves an essential role in WBC type discrimination. Successive encoder layers aid the refinement process of cell characteristics representation. The feedback loop system allows numerous iterations that improve the detection of minor morphological changes.

5. Implementation

The implemented hybrid model relied on the Python programming language, along with TensorFlow and Keras deep learning libraries, to build and train the model. The project development occurred through Google Colab, which it provided access to NVIDIA Tesla K80/TPU hardware accelerators. The dataset was connected to Google Drive for easy management. OpenCV operated during the first stage to execute various image processing

operations, which included resizing, edge detection, contour extraction, and binary masking. The WBC segmentation required the edge contouring process illustrated in Fig. 1 which separated WBC cells from other background elements. The processed images entered the CNN module to perform local feature recognition.

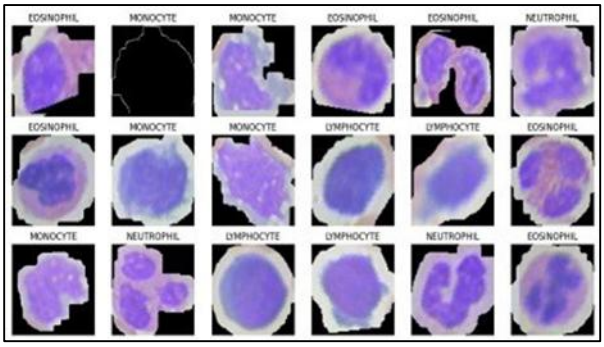

Figure 1. Edge Contouring of WBC Images

A WBC dataset-trained VGG16 architecture served as a basis for CNN backbone operations. Specialized adaptation to medical domain content occurred after selecting which layers to freeze within the VGG16 pre-trained architecture. Sequential patch tokens were derived from the CNN output before entering the Vision Transformer. The Transformer component integrates multiple encoder layers made up of multi-head self-attention and feed-forward sublayers. Strategies to embed positions helped the model retain proper spatial relationships.

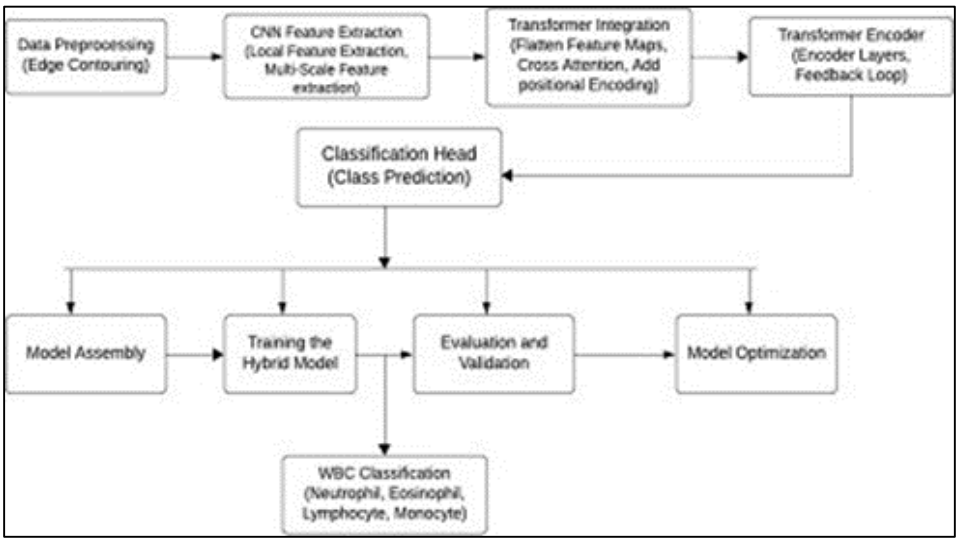

Figure 2. Hybrid Architecture

The model employed cross-attention procedures to achieve contextual relationships between elements located at various measurement levels. Feature representation underwent multiple rounds of refinement through the feedback loop. This depiction in Fig. 2 demonstrates how the integration of CNN and ViT features performs feature extraction. The training consisted of 25 epochs using 32 instances per batch with the Adam optimizer, applying categorical cross-entropy as the loss calculation method. The ReduceLROnPlateau method was used for a dynamic learning rate adjustment scheme during training. The combination of early stopping and checkpoint callbacks served as anti-overfitting measures. Training and validation curves served to monitor the performance of the implemented model.

6. Result and Performance Analysis

The training data accuracy achieved 99.20% according to the model, which exhibited good test generalization at 87.90%. The training and validation accuracy curves in Fig. 3 demonstrate both steady convergence and good avoidance of overfitting. A consistent decrease in loss data reached a final value of 0.0267 as illustrated in Fig. 4. The model displays balanced performance in WBC class detection according to precision (0.7083) and recall (0.7000) together with F1-score (0.6970). The true and false classification spreads for WBC classes can be viewed in the confusion matrix provided in Fig. 5. The model achieved high true positive rates for all types of cells because false positive and false negative results remained low.

Performance Metric Equations:

- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1-score = (2*Precision*Recall)/(Precision +Recall)

Training Configuration:

- Framework: TensorFlow and Keras on Google Colab
- Hardware: NVIDIA Tesla K80
- Epochs: 25, Batch size: 32, Optimizer: Adam
- Loss: Categorical Cross-Entropy

- Callbacks: EarlyStopping, ReduceLROnPlateau
- Augmentation: Random-flips, Brightness, Rotation, Edge-Detection (Sobel and Canny)

The CNN used in WBC classification focused its attention on essential nucleus and cytoplasmic areas as demonstrated by feature maps in Fig. 6. Analysis of the Vision Transformer attention maps in Fig. 7 indicated that the model could evaluate both extensive inter-element dependencies and extensive spatial image structures.

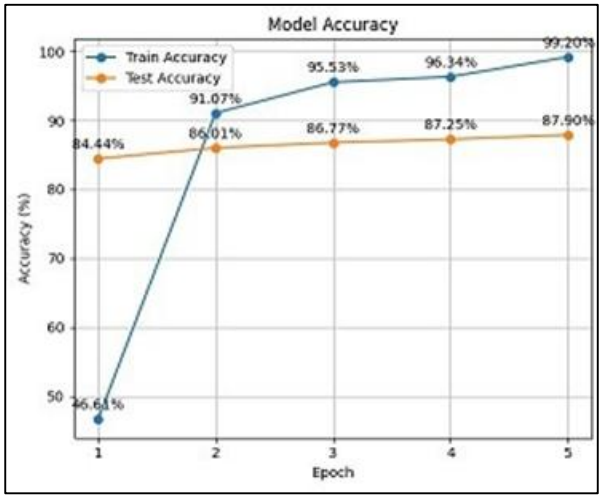

Figure 3. Training and Validation Accuracy

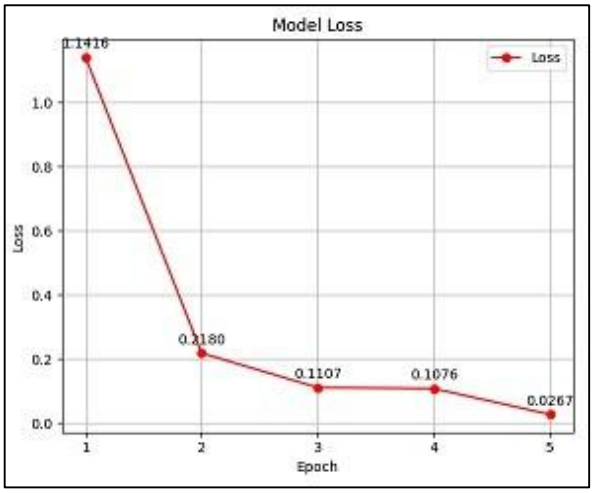

Figure 4. Training and Validation Loss

The model was further tested using real-time image streams because experts needed to confirm its operational speed in clinical settings. The system provided the proper annotations that combined the classification results with confidence data and visualization of the boundary detection. The hybrid model shows potential to serve as an integrated component in diagnostic systems due to its effective functionality.

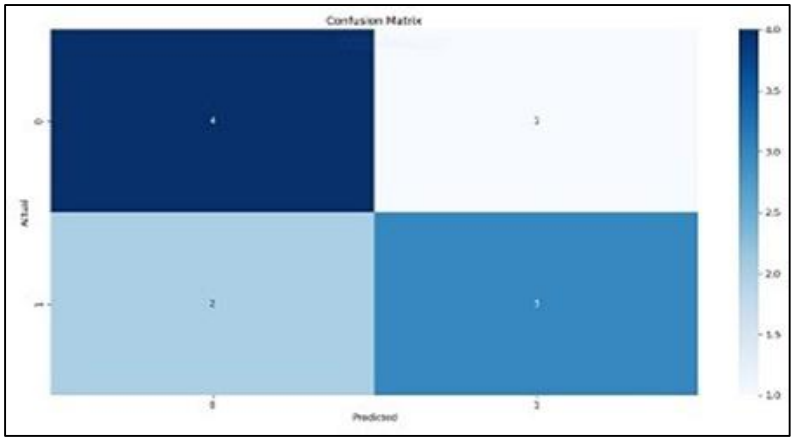

Figure 5. Confusion Matrix

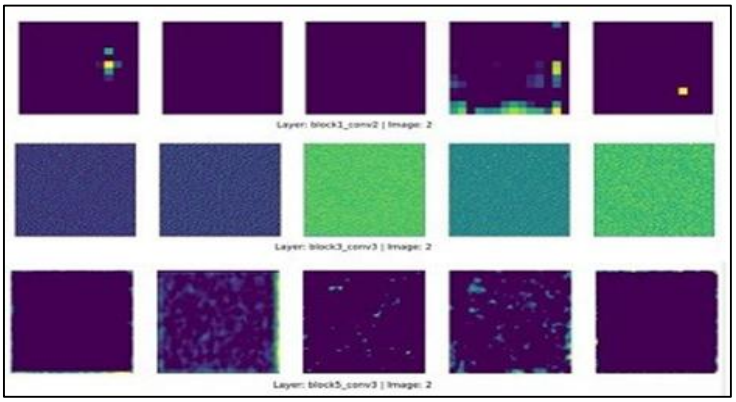

Figure 6. Feature Maps from CNN Highlighting Morphological Features

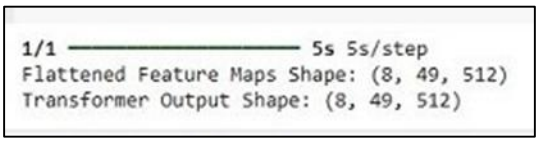

Figure 7. Attention Heatmaps from Vision Transformer Showing Spatial Focus

7. Comparison and Evaluation

The hybrid approach achieved better performance based on evaluations against independent models according to testing results. All performance metrics derived from Table 1 indicate that hybrid model surpassed both CNN and ViT baseline models in their results. Standalone applications of the Vision Transformer produced 82.97% accuracy with a loss value of 0.0701, outperforming the 78.01% accuracy and 0.12 loss resulting from the solitary CNN application. The hybrid model delivered the most optimal combination of precision, recall, and F1-score metrics.

Table 1. Result Comparison

Model	Accuracy (%)	Loss	Precision	Recall	F1-Score
CNN (Baseline)	78.01	0.120	0.6512	0.6345	0.6427
ViT (Baseline)	82.97	0.0701	0.6834	0.6709	0.6771
Proposed Hybrid Model	**87.90**	**0.0267**	**0.7083**	**0.7000**	**0.6970**

Through their collaborative system the hybrid framework employed CNN's detailed feature assessment capabilities together with the Vision Transformer's methods for context-

based knowledge processing. The test-case generalization capability of the system became stronger when random data splits occurred multiple times and experiments were repeated. The model offers transparent decision-making through visual interpretation capabilities that utilize attention maps and activation heatmaps which fulfill medical practitioners' requirements.

8. Conclusion and Future Work

The proposed hybrid model is efficient enough to integrate Convolutional Neural Networks and Vision Transformers in the classification of white blood cells in medical images. It combines global attention and local feature extraction in order to enhance diagnostic accuracy. The model has been trained and tested on the BCCD Dataset, achieving good generalization with an accuracy of 87.90% and F1-score of 0.6970. The tools of explainability based on attention models increase the visibility of the decision process, which is essential in clinical use. The feedback loops with edge-aware preprocessing guarantee improved morphological sensitivity. Future directions can cover multi-disease WBC segmentation, self-supervised learning, and explainable AI using Grad-CAM and SHAP. Federated learning should be considered to enable data non-centralized multi-institutional training and support secure training. Also, the use of this system on embedded systems will facilitate low-resource diagnostics in real-time. These developments are poised to generate scalable, interpretable, and precise AI-based diagnostic frameworks applicable to hematological analysis to fill the gaps between algorithm development and clinical use.

References

[1] Asghar, Rabia, Sanjay Kumar, Arslan Shaukat, and Paul Hynds. "Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review." Plos one 19, no. 6 (2024): e0292026.

[2] Almurayziq, Tariq S., Ebrahim Mohammed Senan, Badiea Abdulkarem Mohammed, Zeyad Ghaleb Al-Mekhlafi, Gharbi Alshammari, Abdullah Alshammari, Mansoor Alturki, and Abdullah Albaker. "Deep and hybrid learning techniques for diagnosing microscopic blood samples for early detection of white blood cell diseases." Electronics 12, no. 8 (2023): 1853.

[3] Rao, Bairaboina Sai Sambasiva, and Battula Srinivasa Rao. "An effective WBC segmentation and classification using MobilenetV3–ShufflenetV2 based deep learning framework." IEEE Access 11 (2023): 27739-27748.

[4] Yentrapragada, Divyateja. "Deep features based convolutional neural network to detect and automatic classification of white blood cells." Journal of Ambient Intelligence and Humanized Computing 14, no. 7 (2023): 9191-9205.

[5] Elhassan, Tusneem A., Mohd Shafry Mohd Rahim, Mohd Hashim Siti Zaiton, Tan Tian Swee, Taqwa Ahmed Alhaj, Abdulalem Ali, and Mahmoud Aljurf. "Classification of atypical white blood cells in acute myeloid leukemia using a two-stage hybrid model based on deep convolutional autoencoder and deep convolutional neural network." Diagnostics 13, no. 2 (2023): 196.

[6] Olayah, Fekry, Ebrahim Mohammed Senan, Ibrahim Abdulrab Ahmed, and Bakri Awaji. "Blood slide image analysis to classify WBC types for prediction haematology based on a hybrid model of CNN and handcrafted features." Diagnostics 13, no. 11 (2023): 1899.

[7] Cheuque, César, Marvin Querales, Roberto León, Rodrigo Salas, and Romina Torres. "An efficient multi-level convolutional neural network approach for white blood cells classification." Diagnostics 12, no. 2 (2022): 248.

[8] Jung, Changhun, Mohammed Abuhamad, David Mohaisen, Kyungja Han, and DaeHun Nyang. "WBC image classification and generative models based on convolutional neural network." BMC Medical Imaging 22, no. 1 (2022): 94.

[9] Davamani, K. Anita, CR Rene Robin, D. Doreen Robin, and L. Jani Anbarasi. "Adaptive blood cell segmentation and hybrid Learning-based blood cell classification: A Meta-heuristic-based model." Biomedical Signal Processing and Control 75 (2022): 103570.

[10] Balasubramanian, Kishore, N. P. Ananthamoorthy, and K. Ramya. "An approach to classify white blood cells using convolutional neural network optimized by particle swarm optimization algorithm." Neural Computing and Applications 34, no. 18 (2022): 16089-16101.

[11] Patil, A. M., M. D. Patil, and G. K. Birajdar. "White blood cells image classification using deep learning with canonical correlation analysis." Irbm 42, no. 5 (2021): 378-389.

[12] Çınar, Ahmet, and Seda Arslan Tuncer. "Classification of lymphocytes, monocytes, eosinophils, and neutrophils on white blood cells using hybrid Alexnet-GoogleNet-SVM." SN Applied Sciences 3 (2021): 1-11.

[13] Khan, Siraj, Muhammad Sajjad, Tanveer Hussain, Amin Ullah, and Ali Shariq Imran. "A review on traditional machine learning and deep learning models for WBCs classification in blood smear images." Ieee Access 9 (2020): 10657-10673.

[14] Malkawi, Areej, Rawan Al-Assi, Taimaa Salameh, Bara'ah Sheyab, Hiam Alquran, and Ali Mohammad Alqudah. "White blood cells classification using convolutional neural network hybrid system." In 2020 IEEE 5th middle east and Africa conference on biomedical engineering (MECBME), pp. 1-5. IEEE, 2020.

[15] Dosovitskiy A, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv (2020): 2010.11929.

[16] Özyurt, Fatih. "A fused CNN model for WBC detection with MRMR feature selection and extreme learning machine." Soft Computing 24, no. 11 (2020): 8163-8172.

[17] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014)