Data Science and Engineering (DSE) Record, Volume 6, issue 1.

White Blood Cells Classification Using Machine Learning and Deep Learning

Tidarat Katsanook [1] , Chalermrat Nontapa [2] and Kornprom Pikulkaew [3]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
[3] Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
tidarat_katsanook@cmu.ac.th

Abstract. White blood cell (WBC) classification plays a pivotal role in diagnosing and monitoring various medical conditions, particularly hematological and immune-related disorders. This study explores the application of machine learning (ML) and deep learning (DL) techniques to classify WBCs, leveraging their potential to enhance diagnostic precision and efficiency. Using a dataset of 50,000 2D images from the University of North British Columbia, we develop and evaluate models for categorizing WBCs into four key types: eosinophils, lymphocytes, monocytes, and neutrophils. The proposed methodology integrates data augmentation, feature extraction, and advanced classification algorithms, including Convolutional Neural Networks (CNNs) and other statistical approaches. Performance metrics such as accuracy, precision, recall, and F1-score guide the optimization of model architecture and training processes. Experimental results demonstrate the effectiveness of the developed models in achieving high classification accuracy, offering a reliable and automated tool for WBC identification. This research underscores the potential of AI-driven solutions to improve clinical workflows, particularly in resource-limited settings, by providing accessible and cost-effective diagnostic support.

Keywords: Convolution Neural Network (CNN), Deep Learing, Machine Learning, White Blood Cells (WBCs).

1 Introduction

White blood cells are essential components of the human immune system, playing a critical role in defending the body against infections, allergens, and diseases. Accurate classification of WBCs into their subtypes—eosinophils, lymphocytes, monocytes, and neutrophils—is crucial for diagnosing and monitoring a wide range of medical conditions, including autoimmune disorders, infections, and leukemia. Traditional methods of WBC classification often rely on manual microscopy, which can be time-consuming, labor-intensive, and prone to human error. While automatic blood cell analyzers provide a more efficient alternative, their high cost, complex maintenance requirements,

and limited accessibility in resource-constrained environments pose significant challenges.

Advancements in machine learning and deep learning have opened new avenues for automating WBC classification, offering potential solutions that are both accurate and scalable. These techniques, particularly Convolutional Neural Networks, excel in image analysis tasks by learning complex spatial hierarchies of features directly from data. By leveraging ML and DL, it is possible to develop systems that classify WBCs with high precision and efficiency, even in diverse clinical and resource-limited settings.

This study focuses on the development of a robust WBC classification system utilizing 2D blood smear images sourced from the University of North British Columbia dataset. The research incorporates advanced data augmentation techniques, feature extraction methods, and a combination of ML and DL models to achieve superior classification performance. By addressing key challenges such as data variability and computational efficiency, this work aims to contribute to the growing field of AI-driven medical diagnostics, providing a valuable tool for improving healthcare outcomes.

2 Literature Review

Classifying white blood cells is an important task in medicine, helping doctors diagnose infections and track immune system problems. Many studies have looked into using machine learning and deep learning methods to improve how accurately and quickly WBCs can be classified. Below is key contributions in this domain:

2.1 Traditional Machine Learning Approaches

Mu-Chun S. et al. (2014) proposed a neural network-based approach to automate WBC classification. By leveraging advanced feature extraction methods, their system achieved high effectiveness and robustness, providing a reliable tool for automating medical image diagnostics.

Abdullah E. et al. (2019) utilized six ML algorithms, including Random Forest, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Naive Bayes, Decision Trees, and Multinomial Logistic Regression (MLR), for WBC classification from microscopic blood smear images. Among these, MLR achieved the highest success rate, with an average test accuracy of 95%.

2.2 Deep Learning Approaches

Made Satria W. (2018) conducted a comparative study between deep learning and traditional ML techniques for WBC classification. Their findings revealed that deep learning methods significantly outperformed traditional approaches such as MLP, KNN, and SVM, achieving an accuracy of 99.5%. The study demonstrated the robustness of DL techniques and their reduced dependency on manual feature extraction.

Muhammed Y. et al. (2019) explored the application of CNNs for WBC classification, demonstrating the effectiveness of image preprocessing techniques such as

Gaussian and median filters in enhancing classification accuracy. The study underscored the reliability of CNNs in diagnosing diseases, particularly when combined with appropriate preprocessing.

Sarang S. et al. (2022) implemented the DenseNet121 architecture for automated WBC classification. The model was trained and evaluated using various batch sizes (8, 16, 32, and 64) and optimized with the Adam optimizer over 10 epochs. The study highlighted the model's capability to improve both the efficiency and accuracy of WBC classification.

Mohamad A. et al. (2023) investigated advanced deep learning models, including Google Vision Transformer (ViT) and pre-trained CNN models from ImageNet ILSVRC, for WBC classification. Their study focused on peripheral blood smear images from the PBC and BCCD datasets. The results highlighted the superior performance of Google ViT in classifying four types of WBCs. The authors emphasized the importance of balanced data augmentation to avoid negative impacts on classification accuracy.

2.3 Hybrid and Advanced Models

Tulasi Gayatri D. et al. (2024) introduced a hybrid model, ABCNM, which addressed challenges such as dynamic variations in WBC images. This model outperformed conventional CNNs and other state-of-the-art techniques, demonstrating superior performance and reduced computational complexity.

3 Methodology

3.1 Dataset

This research utilizes a dataset sourced from the University of Northern British Columbia (UNBC), containing 349 high-resolution microscopic images of white blood cells. The dataset is separated into four distinct categories based on cell types as follws:

1. Neutrophils: These are polymorphonuclear cells essential for combating bacterial infections as shown in Figure 1.

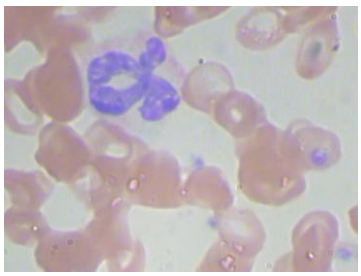

Figure 1. Shows Neutrophils.

2. Eosinophils: Specialized cells involved in allergic reactions and defense against parasites as shown in Figure 2.

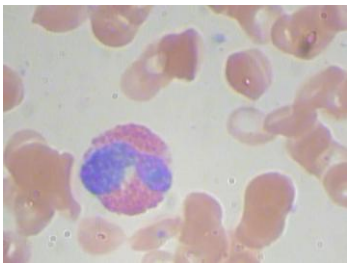

Figure 2. Shows Eosinophils.

3. Lymphocytes: Small white blood cells critical for adaptive immune responses, including antibody production as shown in Figure 3.

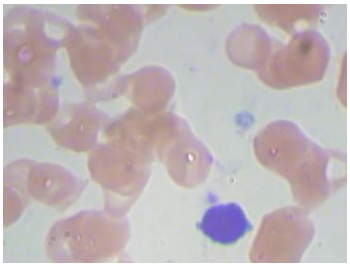

Figure 3. Shows Lymphocytes.

4. Monocytes: These cells differentiate into macrophages and dendritic cells, crucial for inflammation and tissue repair as shown in Figure 4.

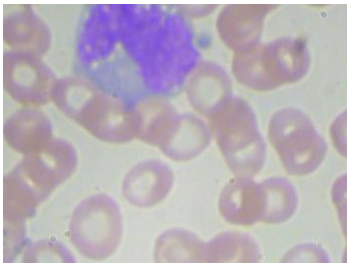

Figure 4. Shows Monocytes.

3.2 Exploratory Data Analysis

The dataset presents a class imbalance as show in Table 1, has many more neutrophils than other types. This imbalance is a common challenge in medical imaging datasets.

Table 3.1. Table shows the number of each type of white blood cell.

Types of white blood cell	Number

Neutrophil	207
Eosinophils	88
Lymphocytes	33
Monocytes	21

3.3 Methodology

In this study has data architechture for creating the machine learning and deep learning model can be separated into three main section that consist of data preprocessing section, feature extraction section, model development section, and proposed model.

3.3.1 Data Pre-processing Section

The section is critical to ensure the dataset is prepared effectively for model training and evaluation. This section have two key components as followings.

1) **Data Augmentation:** According the issue of class imbalance, data augmentation techniques were applied to artificially increase the dataset. Each image in the dataset was transformed using rotation, ensuring that each class contained at least 2,500 images as shown in Table 2. The transformation applied to an image can be mathematically represented as:

$$I' = T(I, \theta) \quad \text{----- (1)}$$

Where I is the original image,

T represents the transformation function,

θ is the rotation angle (e.g., $\theta \in \{-45\circ, -30\circ, 0\circ, 30\circ, 45\circ\}$,

I' is the original image.

Table 3.2. Table shows the number of each classes after data augmentation.

Types of white blood cell	Number
Neutrophil	2499
Eosinophils	2497
Lymphocytes	2483
Monocytes	2478

2) **Dataset Splitting:** The dataset divided into there subsets to facilitate model development as shown in Table 2. First subset is training dataset (60%) used for

training the model. Second subset is validation set (20%) used for hyperparam-eter tuning and performance monitoring and the last subset is test set (20%) resered for evaluating the model on unseen data. To maintain class distribution across subset, Stratified Sampling was used during the splitting process to easure the proportions of each class were consistent. This was achieved usinf the train_test_split function from Scikit-learn.

Table 3.3. Table shows the number of subset in data splitting.

Types of white blood cell	Training dataset	Valiation dataset	Test dataset
Neutrophil	1499	500	500
Eosinophils	1498	500	499
Lymphocytes	1489	497	497
Monocytes	1487	495	496

3) **Resizing:** All images were resized to 224×224 pixels to match the input size of the pre-trained models.
4) **Normalization:** The pixel values were scaled to a ranges between 0 and 1 using this formula as Equation 2. This equation ensures the images are in the same format as the pre-trained model expect.

$$Normalizeed\ Pixel\ Value = \frac{Original\ Pixel\ Value255}{255} \quad \text{-----(2)}$$

3.3.2 Feature Extraction Section

This section focuses on extracting importamr patterns from white blood cells (WBCs) images. These patterns, called “features”, are used by the classifier to te’’ the difference between the WBC classes. We used pre-trained models, are deep learning networks trained on large datasets like ImageNet. They have already learned to identify features in millons of images, such as edges, textures, shapes, and objects. These learned features can be reused for classifying with blood cells.

1) VGG16: VGG16 is a deep convolution neural network used for images recognition. It has 16 layers, including 13 convolution layers and 3 fully connected layers as Figure 6. In feature extraction, we focus on the convolution layers, which capture important features of the image.

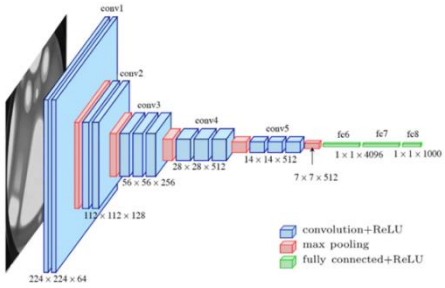

Figure 6. Shows the VGG16 archiyectures.

3.3.3 Model Development Section

This section describes the development of three machine learning models: Decision Tree, Multinomial Logistic Regression (MLR), and Random Forest. Each model was selected based on its suitability for white blood cell (WBC) classification. The development process involved data preprocessing, model training, performance evaluation, and comparative analysis to determine the best-performing approach.

1) Decision Tree:

A decision tree is a simple and powerful model that splits data into branches based on feature values, forming a tree-like structure. It works well for classifying white blood cells (WBCs) into different types, such as Lymphocyte, Monocyte, Neutrophil, and Eosinophil. The main advantages of this model are its interpretability and ability to handle both numerical and categorical data.

Impurity Measures in Decision trees

- Gini Index Formula (used to measure impurity):

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \quad \text{----- (3)}$$

Where p_i is the proportion of samples belonging to class i,
n is the number of classes.

- Entropy Formula (another impurity measure):

$$Entropy = -\sum_{i=1}^{n} p_i log_2(p_i) \quad \text{------(4)}$$

Where p_i is the proportion of samples in class i.
n is the number of classes

- Information gain Formula (used to decide the best split):

$$IG = Entropy_{parent} - \sum_{k=1}^{K} \frac{N_k}{N} \cdot Entropy_k \quad \text{------(5)}$$

Where N_k is the number of WBC samples in child node,
N is the total number of samples in the parent node,
K is the number of child nodes.

2) Random Forest

Random Forest is an advanced ensemble learning model that builds multiple decision trees and aggregates their predictions to improve accuracy and stability. It is particularly effective in classifying WBCs into types, such as distinguishing Eosiniphils from Monocytes. This method reduces overfitting and efficiently captures complex relationships in the data.

Additionally, Random Forest can analyze feature importance, allowing the identification of key factors that influence classification, such as Nucleus size, Cytoplasm color.

- Prediction Formula (Majority Voting Mechanism)

$$y = mode(y_1, y_2, \dots, y_m) \quad \text{------(6)}$$

Where y_i is the prediction from the i-th decision tree,
m is the tltal number of decision trees in the ensemble,
y is the class that appears most frequently.

3) Multinomial Logistic Regression (MLR)

Multinomial Logistic Regression (MLR) is a generalization of logistic regression designed for multi-class classification problems, where the target variable has more than two categories. Unlike binary logistic regression, which differentiates between two classes, MLR assigns probabilities to multiple possible outcomes using the softmax function.

For a classification problem with K classes, the probability that a WBC sample x belongs to class j is computed as

$$P(y = j|x) = \frac{e^{\beta_j^T x}}{\sum_{k=1}^{K} e^{\beta_k^T x}} \quad \text{-----(7)}$$

Where $P(y = j|x)$ is the probability that the WBC belongs to class j.
β_j is the coefficient vector for class j (weights assigned to features).
x is the feature vector respresenting the WBC sample.
K is the total number of WBC classes.

- Loss Function (Cross-Entropy Loss): The model is trained by minimizing the negative log-likelihood, also known as the cross-entropy loss:

$$L = -\sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,j} \log \mathrm{P}(y = j|x_i) \quad \text{------(8)}$$

Where N is the total number of WBC samples.
$y_{i,j}$ is 1 if sample belongs to class j, otherwise 0.
$P(y = j|x_i)$is the predicted probability for class j.

4) Convolution Neural Network (CNN)

CNN is a deep learning model designed for analyzing grid-structured data like images. Unlike traditional machine learning methods that depend on handcrafted features, CNN automatically learns spatial feature hierarchies through convolutional layers, activation functions, pooling operations, and fully connected layers as Figure 7. This approach is highly effective for visual pattern recognition tasks, including image classification and medical image analysis.

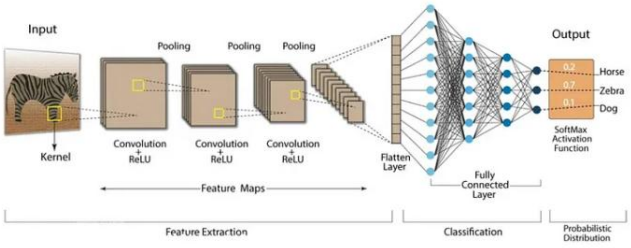

Figure 7. Convolution Neural Network Structure

Main component of CNN consist of:

- Convolutional Layers: Extract local spatial features using filters.
- Activation Functions: Introduce non-linearity, typically using Rectified Linear Unit (ReLU).
- Pooling Layers: Reduce spatial dimensions while retaining key information, commonly using Max Pooling.
- Flatten Layer: Converts multi-dimensional output into a single vector for further processing.
- Fully Connected Layers: Integrate features for prediction by computing weighted sums.
- Output Layer: Produces class probabilities through Softmax.

Training involves minimizing categorical cross-entropy loss, measuring differences between predicted and true class probabilities. The model's final classification is determined by selecting the class with the highest probability.

This structured approach allows CNN to effectively classify white blood cells (WBC), providing insights for practical medical applications and informing the proposed model outlined in the subsequent section.

3.4 Proposed Model

Despite significant progress in automated white blood cell (WBC) classification, there remains a practical gap between high-performing research models and real-world constraints faced by clinical laboratories. Current deep learning models typically require extensive computational resources, long training periods, specialized hardware, and often lack interpretability. Traditional machine learning approaches, although interpretable and efficient, usually achieve limited accuracy. This research gap highlights the need for hybrid frameworks that combine deep learning's robust feature extraction with classical machine learning's efficiency and interpretability.

To address these challenges, this study proposes a hybrid WBC classification pipeline that integrates feature extraction using a pre-trained VGG16 model with lightweight classical classifiers: Decision Tree (DT), Multinomial Logistic Regression (MLR), and Random Forest (RF). This hybrid architecture aims to balance accuracy, efficiency, and interpretability.

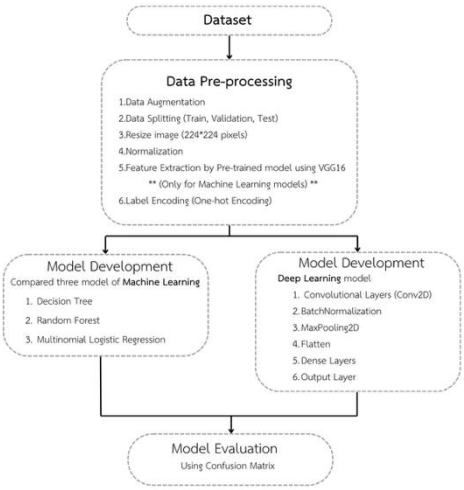

Figure 8. White Blood Cell Classification: Machine Learning and Deep Learning

The proposed system consists of two parallel pipelines:

1. Hybrid Model: A pre-trained VGG16 CNN extracts high-dimensional, discriminative features from WBC images. These features, capturing textures, shapes, and edges, are then classified using interpretable, low-resource machine learning algorithms (DT, MLR, RF), fine-tuned through hyperparameter optimization.
2. CNN Model: An alternative pipeline uses an end-to-end CNN directly trained on raw images, featuring convolutional layers with varying kernel sizes, Batch Normalization, ReLU activations, pooling layers, fully connected dense layers, and dropout for regularization. It uses categorical cross-entropy loss optimized via Stochastic Gradient Descent (SGD) with a learning rate scheduler.

Both paths use identical datasets and preprocessing, allowing empirical comparison of accuracy, computational efficiency, and interpretability, making the approach suitable for diverse clinical environments.

3.5 Model Evaluation

The model evaluation process aims to measure how well the machine learning models classify white blood cells (WBCs) into their respective categories. This step is critical to ensure the model's predictions are accurate, reliable, and generalizable to new data.

To evaluate the performance of each model, several standard metrics were used the ecaluation metrics as below:

1) Accuracy: Measure the proportion of correctly classified WBCs among all samples.

$$Accuracy = \frac{Number\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples} \quad \text{------(9)}$$

2) Precision: Evaluates the proportion of correctly identified WBCs for each type.

$$Precision = \frac{True\ Positive(TP)}{Ture\ Positives\ (TP)+False\ Positive\ (FP)} \quad \text{------(10)}$$

3) Recall (Sensitivity): Measures the ability to correctly identifiey all WBCs of a specific type.

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive(TP)+False\ Negative(FN)} \quad \text{------(11)}$$

4) F1-Score: Balances precision and recall into a single metric.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Rescision + Recall} \quad \text{------(12)}$$

5) Confusion Matrix: Provides a summary of predictions, showing the true and predicted classes of WBCs.

4 Results

This chapter summarizes the experimental results for white blood cell (WBC) classification using four models: Decision Tree (DT), Random Forest (RF), Multinomial Logistic Regression (MLR), and Convolutional Neural Network (CNN). All models were trained, validated, and tested on the same dataset to ensure comparability.

4.1 Individual Model Results

4.1.1 Decision Tree (DT)

The DT model, optimized via grid search achieved validation accuracy of 57.93% and test accuracy of 58.23%.

Table 4.1 Classification Report of Decision Tree Model on Test Set

Class	Precision	Recall	F1-Score	Support
EOSINOPHIL	0.50	0.52	0.51	499
LYMPHOCYTE	0.65	0.64	0.65	497
MONOCYTE	0.68	0.67	0.68	496
NEUTROPHIL	0.51	0.50	0.50	500

4.1.2 Random Forest (RF)

The RF model demonstrated strong generalization with validation accuracy of 85.19% and test accuracy of 83.79%.

Table 4.2 Classification Report of Random Forest Model on Test Set

Class	Precision	Recall	F1-Score	Support
EOSINOPHIL	0.79	0.68	0.73	499
LYMPHOCYTE	0.89	0.94	0.92	497
MONOCYTE	0.86	0.95	0.91	496
NEUTROPHIL	0.80	0.78	0.79	500

4.1.3 NMultinomial Logistic Regression (MLR)

The MLR model achieved excellent performance with validation accuracy of 91.42% and test accuracy of 90.76%.

Table 4.3 Classification Report of MLR Model on Test Set

Class	Precision	Recall	F1-Score	Support
EOSINOPHIL	0.86	0.82	0.84	499
LYMPHOCYTE	0.94	0.98	0.96	497
MONOCYTE	0.94	0.98	0.96	496
NEUTROPHIL	0.88	0.85	0.87	500

4.1.4 Convolution Neural Network (CNN)

The CNN model featured a deeper and wider network architecture with multiple convolutional layers incorporating 1×1 and 3×3 kernels, increased filter counts, fully connected layers with enhanced units (1024), and the integration of Batch Normalization and Dropout for effective regularization. The model was trained using the Stochastic Gradient Descent (SGD) optimizer with a learning rate scheduler.

The final CNN achieved a test accuracy of 92.62%, indicating robust generalization and high discriminative power.

Table 4.4 Classification Report of CNN Model on Test Set

Class	Precision	Recall	F1-Score	Support
EOSINOPHIL	0.87	0.87	0.87	499
LYMPHOCYTE	0.99	0.97	0.98	497
MONOCYTE	0.98	1.00	0.99	496
NEUTROPHIL	0.87	0.86	0.87	500

Figure 9 displays the loss curves for the final CNN model. Both training and validation loss decrease steadily and converge, with no significant gap between them, indicating robust generalization and effective regularization.

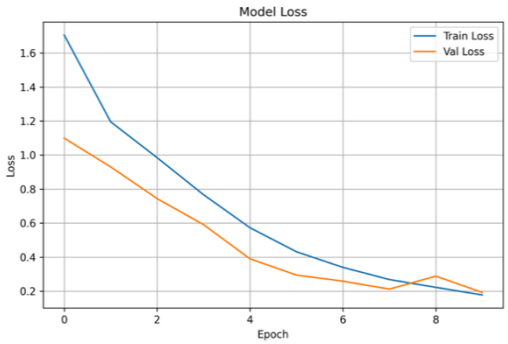

Figure 9. Training and Validation Loss Comparions

4.2 Comparative Evaluation

The comparative analysis indicated CNN as the best-performing model, closely followed by MLR and RF. DT showed the weakest performance. CNN offered optimal accuracy, while MLR and RF provided valuable interpretability and efficiency. These results provide practical insights for model selection in clinical settings.

4.3 Statistical Significance Testing

A paired t-test was used to compare the classification accuracy of the CNN model and the Hybrid model (VGG16 + MLR). Predictions were binarized (1 = correct, 0 = incorrect) for direct sample-wise comparison. The test yielded:

$$t = 2.082, p = 0.037$$

Since the p-value is below 0.05, the result indicates a statistically significant difference in accuracy between the two models at the 95% confidence level. While the Hybrid model (90.85%) offers faster training and lower resource usage, the CNN model (92.62%) achieved significantly higher accuracy.

5 Conclusion

This section summarizes and interprets the performance results for white blood cell (WBC) classification using Decision Tree (DT), Random Forest (RF), Multinomial Logistic Regression (MLR), and Convolutional Neural Network (CNN) models. Emphasis is placed on their strengths, limitations, and particular challenges in differentiating between morphologically similar cell types, especially EOSINOPHIL and NEUTROPHIL.

5.1 Interpretation of the classification results

5.1.1 Decision Tree (DT)

The DT model achieved the lowest accuracy (58.23%) and macro F1-score (0.58). Although highly interpretable, it was unable to handle complex and subtle visual differences effectively. Notably, misclassifications frequently occurred between EOSINOPHIL and NEUTROPHIL due to similar granularity and nuclear lobation features, highlighting the DT's rigid, threshold-based limitations.

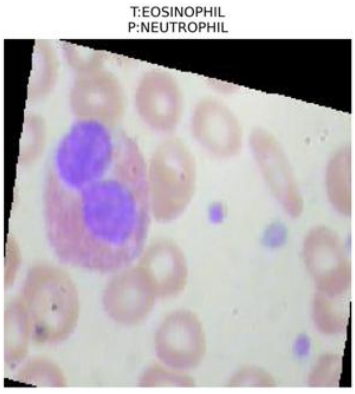

Figure 10. Misclassified of Decision Tree Model

5.1.2 Random Forest (RF)

The RF significantly outperformed DT, achieving 83.78% accuracy and an F1-score of 0.83. Ensemble learning reduced overfitting, but confusion remained pronounced between EOSINOPHIL and NEUTROPHIL, reflecting the persistent challenge of distinguishing visually overlapping morphological features despite improved model complexity.

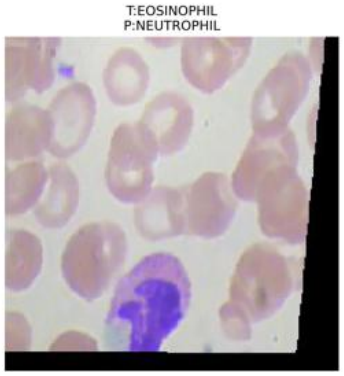

Figure 11. Misclassified of Random Forest Model

5.1.3 Multinomial Logistic Regiression (MLR)

The MLR paired with deep features from VGG16, reached high performance (accuracy 90.76%, macro F1-score 0.91). It effectively leveraged high-quality feature extraction, yet still faced challenges differentiating EOSINOPHIL from NEUTROPHIL due to overlapping morphological characteristics. This result underscores the inherent limitations of linear classifiers in scenarios involving visually subtle distinctions.

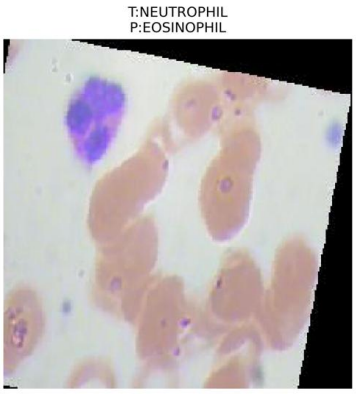

Figure 12. Misclassified of MLR Model

5.1.4 Convolution Neural Network (CNN)

The CNN model achieved the highest performance overall (accuracy 92.62%, macro F1-score 0.93), benefiting from its ability to learn hierarchical visual features directly. The robust architecture, combined with effective regularization (Batch Normalization, Dropout), greatly enhanced generalization. Nevertheless, EOSINOPHIL-NEUTROPHIL misclassification persisted, emphasizing the continuing challenge posed by their close morphological similarities.

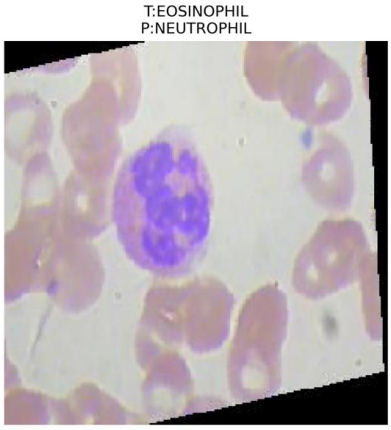

Figure 13. Misclassified of CNN Model

All models consistently faced difficulty distinguishing EOSINOPHIL from NEUTROPHIL, which share overlapping nuclear shapes, granularity, and size. Future research should focus on incorporating additional differentiating features or more sophisticated architectures to address this ongoing challenge.

5.2 Model Benchmarking Study

Our VGG16 + Multinomial Logistic Regression (MLR) and custom CNN models achieved high classification performance on peripheral blood smear images (MLR: 90.76% accuracy, F1-score: 91%; CNN: 92.62% accuracy, F1-score: 93%). When benchmarked against ICSH clinical standards (accuracy ≥ 90%, F1-score 90–95%) and commercial analyzers (e.g., DM96™, DI-60), both models meet baseline clinical requirements.

5.3 Model Benchmarking Study

A paired t-test comparing CNN and Hybrid (VGG16 + MLR) yielded $t = 2.082$, $p = 0.037$, indicating a statistically significant performance difference at the 95% confidence level. While CNN outperformed the Hybrid model in accuracy, the Hybrid

model remains a practical alternative for scenarios requiring faster training, lower resource usage, or greater interpretability.

6 Discussion and Feture Work

This chapter synthesizes experimental findings, highlighting the strengths and limitations of each classification model while suggesting directions for future research in automated white blood cell (WBC) classification.

6.1 Key Strengths

6.1.1 Comprehensive Model Comparison Four classification paradigms

Decision Tree (DT), Random Forest (RF), Multinomial Logistic Regression (MLR), and Convolutional Neural Network (CNN)—were evaluated under identical preprocessing and dataset conditions, enabling fair and meaningful performance comparisons.

6.1.2 Integration of Deep and Classical Features

The hybrid approach combining VGG16-extracted deep features with interpretable MLR achieved 90.76% accuracy, demonstrating a balanced trade-off between performance and explainability for clinical applications.

6.1.3 Rigorous Error Analysis

Detailed confusion matrix and per-class error analyses were conducted to understand model behavior, particularly the recurrent misclassification between morphologically similar classes such as EOSINOPHIL and NEUTROPHIL.

6.1.4 Reproducible Evaluation Protocol

All experiments utilized fixed random seeds, stratified dataset splits, and consistent preprocessing pipelines, ensuring reliable and reproducible benchmarking.

6.1.5 Statistical Significance of Model Comparison

A paired t-test confirmed that the CNN model's superior accuracy (92.62%) compared to the Hybrid model (90.85%) is statistically significant (**$t = 2.082$, $p = 0.037$**), reinforcing the robustness and reliability of the performance difference observed between the two models.

6.2 Limitations

6.2.1 Morphological Overlap

All models faced challenges distinguishing EOSINOPHIL from NEUTROPHIL due to inherent visual similarities in granule density and nuclear structure, suggesting a limitation of appearance-based classification alone.

6.2.2 Dataset Size and Diversity

The dataset included 349 images per class, with limited staining and imaging variability, potentially reducing model generalizability to diverse clinical environments.

6.2.3 Single Imaging Modality

Dependence on standard bright-field microscopy limited the visibility of nuanced cellular features that could be more apparent under alternative imaging techniques.

6.2.4 Simplified Architectures and Preprocessing

The study employed basic CNN architectures and general preprocessing methods. Incorporating cell-specific preprocessing or performing architecture search may yield improved performance.

6.3 Future Research Directions

6.3.1 Multimodal Imaging

Incorporating advanced imaging techniques like fluorescence or phase contrast microscopy could provide additional cues for difficult classifications.

6.3.2 Attention Mechanisms

Adding spatial attention modules or transformer-based architectures could help models better localize key morphological features such as granules or nuclear contours.

6.3.3 Transfer Learning and Fine-Tuning

Utilizing large-scale hematology datasets for pretraining followed by domain-specific fine-tuning may enhance model robustness and cross-protocol adaptability.

6.3.4 Interpretability and Human Trust

Employing model interpretability tools like Grad-CAM or SHAP can help identify which visual features contribute most to predictions, thereby building clinician trust.

6.3.5 Clinical Validation

A real-world prospective study comparing automated predictions with expert annotations on live clinical data would assess system reliability, diagnostic utility, and real-time integration feasibility.

6.3.6 Advanced Image Processing

Integrating advanced techniques such as cell segmentation, contrast enhancement, and granule-specific filtering could yield cleaner inputs, minimize inter-class confusion, and enhance overall model performance.

References

1. Abou Ali, M., Dornaika, F., & Arganda-Carreras, I. (2023). White blood cell classification: Convolutional neural network (CNN) and vision transformer (ViT) under medical microscope. *Algorithms, 16*(11), Article 525. https://doi.org/10.3390/a16110525
2. Bain, B. J., Clark, D. M., & Wilkins, B. S. (2012). *Dacie and Lewis practical hematology* (11th ed.). Elsevier.
3. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
4. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Cornet, E., Perol, J.-P., & Troussard, X. (2008). *Performance evaluation and relevance of the CellaVision™ DM96 system in routine analysis and in patients with malignant hematological diseases*. *International Journal of Laboratory Hematology, 30*(6), 536–542. https://doi.org/10.1111/j.1751-553X.2007.00996.x
6. Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.
7. Devi, T. G., & Patil, N. (2024). Optimization-based convolutional neural model for the classification of white blood cells. *Journal of Big Data, 11*, Article 90. https://doi.org/10.1186/s40537-024-00949-y
8. Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Wiley.
9. Elen, A., & Turan, M. K. (2019). Classifying white blood cells using machine learning algorithms. *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi, 11*(1), 141–152. https://doi.org/10.29137/umagd.498372
10. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
12. Günay, D. (n.d.). *Random forest* [Diagram]. Medium. https://medium.com/@denizgunay/random-forest-af5bde5d7e1e
13. Kratz, A., Lee, S., Zini, G., Riedl, J. A., Hur, M., & Machin, S. J. (2019). *Digital morphology analyzers in hematology: ICSH review and recommendations*. *International Journal of Laboratory Hematology, 41*(Suppl. 1), 1–19. https://doi.org/10.1111/ijlh.13042
14. Nam, M., Yoon, S., Hur, M., Lee, G. H., & Kim, H. N. (2022). *Digital morphology analyzer Sysmex DI-60 vs. manual counting for white blood cell differentials in leukopenic samples: A comparative assessment of risk and turnaround time*. *Annals of Laboratory Medicine, 42*(1), 19–28. https://doi.org/10.3343/alm.2022.42.1.19
15. Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
16. Shahriar, N. (n.d.). *Convolutional neural network — CNN architecture* [Diagram]. Medium. https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5
17. Shapiro, H. M. (2003). *Practical flow cytometry* (4th ed.). Wiley-Liss.
18. Sharma, S., Gupta, S., Gupta, D., Juneja, S., Gupta, P., Dhiman, G., & Kautish, S. (2022). Deep learning model for the automatic classification of white blood cells. *Computational IntelligenceandNeuroscience,2022*, Article 7384131. https://doi.org/10.1155/2022/7384131
19. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint* arXiv:1409.1556.
20. SmartDraw,Inc.(n.d.). *Structure of a decision tree* [Diagram]. SmartDraw. https://www.smartdraw.com/decision-tree/
21. Su, M.-C., Cheng, C.-Y., & Wang, P.-C. (2014). A neural-network-based approach to white blood cell classification. *The Scientific World Journal, 2014*, Article 796371. https://doi.org/10.1155/2014/796371

22. Tomar, N. R. (n.d.). *Multinomial logistic regression | Softmax regression* [Diagram]. In *Introduction-to-TensorFlow*(GitHubretory). https://github.com/nikhilroxtomar/Introduction-to-TensorFlow/blob/master/06%20-%20%20Multinomial%20Logistic%20Regression%20%7C%20Softmax%20Regression.ipynb
23. University of Northern British Columbia. (2010). *White blood cell image dataset* [Dataset]. https://www.unbc.ca/datasets/white-blood-cell
24. Wibawa, M. S. (2018, October). A comparison study between deep learning and conventional machine learning on white blood cells classification. In *Proceedings of the 2018 International Conference on Orange Technologies (ICOT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICOT.2018.8705892
25. Yildirim, M., & Çinar, A. (2019). Classification of white blood cells by deep learning methods for diagnosing disease. *Revue d'Intelligence Artificielle, 33*(5), 335–340. https://doi.org/10.18280/ria.330502