Blood Cell Classification Based on Feature Importance with the Naïve Bayes Algorithm on the BCCD Dataset

Ardian Nur Rokhani
Informatics Department, Universitas Sahid Surakarta, Surakarta, Indonesia
ardiannur@usahidsolo.ac.id

***Abstract*—Classification of blood cells is of immense importance as a preliminary stage in the diagnosis of several hematological disorders such as leukemia, anemia, and infections. Microscopes and trained personnel often constrain the manual approach due to limited resources, time, and the ability to produce consistent results. In this work, BCCD (Blood Cell Classification Dataset) has been used to solve this problem. This dataset contains 364 images to classify three types of blood cells, which are red blood cells (RBC), white blood cells (WBC), and platelets using the Naïve Bayes algorithm. Feature extraction includes color (RGB), intensity (grayscale), texture (GLCM), shape, and edge density. The data set is split into 80% for training and 20% for testing. Results of the study showed that the Naive Bayes model obtained an accuracy of 98.64% on the test data. The precision of platelets, WBC, and RBC was 85%, 97%, and 100%, respectively. The recalls for the three classes were 98%, 99%, and 99%, respectively. The study demonstrated that the machine learning methods, particularly the Naive Bayes algorithm, constitute a very viable alternative to blood cell classification because of high computation speed, robust performance, and the capability of dealing with high-dimensional data under the simple assumption of feature independence.**

Keywords—blood cells; Naïve Bayes; BCCD Dataset; machine learning

I. INTRODUCTION

Blood cell analysis is a major clinical diagnostic procedure. The leukocyte differential, erythrocyte morphology, and platelet counts are integral to the diagnosis of many pathological diseases, including iron deficiency anemia, infections, autoimmune diseases, and hematological malignancies such as leukemia and lymphoma. Visual inspection is generally done through light microscopy. This practice is still often the case, especially in small to medium-sized laboratories. Disadvantages of such an approach are high interobserver variability, low time efficiency, and dependence on the expert's expertise and physical condition.

In the last decade, machine learning (ML) and deep learning (DL) have shown tremendous potential for medical image classification. Most of the research uses complex DL architectures, such as CNN, R-CNN, and YOLO, that require enormous amounts of data and considerable processing power. Reference [2] states that "while some conventional models achieved up to 99% accuracy, accuracy was shown to decrease in concurrence with decreasing dataset size." This study shows the need to test simple ML algorithms on smaller datasets.

The present work fills several gaps in the current literature: (1) simple ML algorithm evaluations on small medical datasets are scarce; (2) few studies systematically identify the morphological features that contribute most to classification accuracy; (3) rare studies link ML model prediction results with cell morphology-based interpretation; and (4) most publications report only accuracy, failing to analyze metrics per class and feature variance.

This gap has been covered in some previous research projects. [3] in its exhaustive review discusses DL-based segmentation and classification approaches for WBC and RBC, but it does not offer a solution for laboratories with limited resources. [4] did literature review on blood cell classification using ML and did not implement any model on the BCCD dataset. [5] employed an R-CNN framework powered by an autoencoder which requires a lot of data as well. Closest work is [6]. They applied CNN on BCCD and got accuracy 98.75 %. But the calculation needed is substantial.

This study addresses this problem with a low-cost approach, using the Gaussian Naive Bayes (NB) algorithm, a light, fast, and probabilistic classifier [7] easily understood on the BCCD dataset (364 images). This work also contains a model performance evaluation where (1) we analyze the variance of the features to understand the contribution of 22 color, shape, texture, and intensity features to the classification decision; (2) we perform a per-class analysis using the precision, recall, and F1-score metrics; and (3) we compare the performance with state-of-the-art approaches on the same dataset.

This work aims to show the ability of simple algorithms to reach more than 95% accuracy (up to 98.64% with excellent feature engineering), be quicker to build, be faster to train, and be easier to interpret than complex deep learning. The major contributions of this paper are as follows: (1) a comprehensive framework for classification of three types of blood cells using naive Bayes, (2) an analysis of the most discriminative morphological features, (3) a comprehensive evaluation with a confusion matrix and metrics for each class, and (4) empirical evidence of traditional ML as an effective alternative for small-sized datasets, providing a basis for the development of affordable computer-aided diagnosis (CAD) systems in healthcare facilities with limited resources, which remain underrepresented in current ML-based diagnostic research [8].

The paper is structured as follows. The second section outlines the research methodology, which includes a description of the BCCD dataset, the procedure for cell extraction from

XML annotations, methods for color feature extraction (RGB), intensity (grayscale), texture (GLCM), shape, edge density, the Naïve Bayes algorithm, and the evaluation design. The experimental results are shown in section 3, such as accuracy (98.64%), confusion matrix, evaluation metrics (Precision: RBC 100%, WBC 97%, Platelets 85%; Recall: RBC 99%, WBC 99%, Platelets 98%; F1-Score: RBC 99%, WBC 98%, Platelets 91%), and essential feature analysis based on variance values. Section 4 presents the key findings and future research directions.

II. METHODOLOGY

A. BCCD Dataset

BCCD dataset: The BCCD dataset is a free dataset of medical pictures for researchers to classify blood cells. The dataset contains 364 photos (JPEG) and 364 XML files with the annotations for the photos. The XML files contain the bounding box information of each cell, i.e., the bounding box coordinates (xmin, ymin, xmax, ymax) and the class label of each object present in the cell. The total number of extracted training and test data points from this dataset is 3930 and 956, respectively. The class distribution is RBC with 3329 training data points and 824 test data points. WBC: Training data points: 298, Testing data points: 74, Platelets have 303 training data points and 58 test data points. The dataset is imbalanced in classes where the RBCs are much more than the WBCs and platelets.

B. Cell Extraction from XML Annotations

The cell extraction is done by parsing the XML annotation file using xml.etree.ElementTree parser extracts the object elements, such as class name and bounding box coordinates. The procedure is as follows: (1) Parse the XML file to get the corresponding image file name; (2) open the original JPEG image with OpenCV; (3) crop the bounding box region with a 5-pixel margin on each side to include the whole cell area; (4) ignore cell segments that are too small (size < 20 × 20 pixels); (5) resize each segment to 80 × 80 pixels for uniformity a size large enough to capture morphological details but not too large to be computationally expensive; and (6) save the images in separate directories according to class and train/test data division. The data is split into 80% training data and 20% testing data, using a stratified random split with scikit-learn's train_test_split tool.

C. Feature Extraction

The purpose of feature extraction is to obtain numerical features from cell images that are used as input to machine learning methods. In this work, 22 features were extracted and grouped into five categories. This number of features is comparable to the 22 most significant features selected from a larger feature pool (shape, texture, histogram, correlogram, and morphology) in a similar Naïve Bayes-based medical image classification study [9]. The five categories are: (1) color features (RGB) [10]. Each image is converted to the RGB color space, and 3 descriptive statistics are computed for each channel: mean, standard deviation, and median for a total of 9 features (3 channels × 3 statistics). (2) Gray Scale Intensity features. The image is converted to grayscale. The five statistics from the image are the mean, standard deviation, median, minimum, and maximum. This property contains information about the overall brightness and contrast level. These five features are new. (3) Texture Features (GLCM) Erythrocytes are non-nucleated and smooth. White blood cells. Multilobed, granular. The platelets are small with granules. Blood cells are of different shapes. GLCM produces a total of four features, which are contrast, correlation, energy, and homogeneity. (4) Features of shape. The cells are separated from the background using Otsu thresholding. This method automatically finds the best threshold. From the largest contour extracted, three features are computed: (1) area (area of the cell region in square pixels); (2) circularity; and (3) aspect ratio (ratio between the width and height of the bounding box of the cell). (5) Edge density Features. The Canny algorithm is used to detect edges. It has a lower bound of 50. The upper bound is 150. The edge density is the ratio of the number of edge pixels to the number of picture pixels. This property allows us to distinguish between cells with well-defined boundaries (e.g., red blood cells) and cells with complex surfaces (e.g., white blood cells).

D. Naïve Bayes Algorithm

For this study, the algorithm used was the Gaussian Naïve Bayes (GNB), a Naïve Bayes variant specifically designed to deal with continuous (numerical) characteristics and assume that the data in each class follow a normal (Gaussian) distribution. This assumption makes GNB very suitable for image-based feature extraction data with continuous values such as color, texture, and shape.

The Naive Bayes classifier is based on Bayes' theorem, which calculates the probability of a sample belonging to a class, using the prior probability of the class and the conditional probability of the attributes of the sample. In this research, Gaussian Naïve Bayes is used. Gaussian Naïve Bayes estimates the probability of a continuous feature value by using the Gaussian probability density function. The feature value of each class is assumed to follow a normal distribution.

E. Evaluation Plan

The evaluation is conducted on a separate test data set, which consists of 20% of the total data set. The employed metrics are accuracy, precision, recall, F1-score, and the confusion matrix, using the standard formulations widely adopted in the machine learning literature [11]. Accuracy measures the overall proportion of correctly classified samples; precision measures the proportion of positive predictions that are correct; recall measures the proportion of actual positives correctly identified; and F1-score is the harmonic mean of precision and recall. The confusion matrix presents a tabulation of correct and incorrect predictions for each class. All metrics are computed for each class and subsequently averaged in a macro manner.

III. RESULT AND DISCUSSION

A. Cell Extraction Result and Data Splitting

The cell extraction procedure from the BCCD dataset resulted in 3.930 cells for the training dataset and 956 cells for the testing dataset. This dataset has class imbalance characteristics close to the real world, where the number of RBCs is much higher than WBCs and platelets in peripheral blood smears. The distribution of classes is shown in Table 1.

TABLE I. DISTRIBUTION OF TRAINING AND TESTING DATA

Class	Training Data	Testing Data
RBC	3.329	824
WBC	298	74
Platelets	303	58
Total	**3.930**	**956**

B. *Naïve Bayes Model Classification Result*

The Gaussian Naïve Bayes model was trained on 3930 data points with 22 features and tested on 956 unseen data points. The results of the assessment show an overall accuracy of 98.64%. Table 2 shows the classification report-based evaluation metrics for each class.

TABLE II. EVALUATION METRICS FOR THE NAÏVE BAYES MODEL

Class	Precision	Recall	F1-Score	No. of Testing Data
RBC	1.00	0.99	0.99	824
WBC	0.97	0.99	0.98	74
Platelets	0.85	0.98	0.91	58
Average (Macro)	0.94	0.99	0.96	956

The overall accuracy is very satisfactory, but the performance is different from one class to another, as seen in Table 2. RBC precision was 1.00 (perfect), and platelets 0.85 (lowest). It means that the algorithm often misclassifies some of the platelet samples into other categories, probably due to the limited data available.

C. *Confusion Matrix*

The confusion matrix, shown in Fig. 1, tells us how many times we got the prediction right and how many times we got it wrong for each class. The results obtained are as follows: 1. RBC: Out of 824 samples, 813 were predicted correctly, resulting in a recall value of 0.99. 2. WBC: Indicating 73 out of 74 samples were correctly predicted. 3. Platelets We correctly predicted 57 samples, and we predicted one sample incorrectly out of 58.

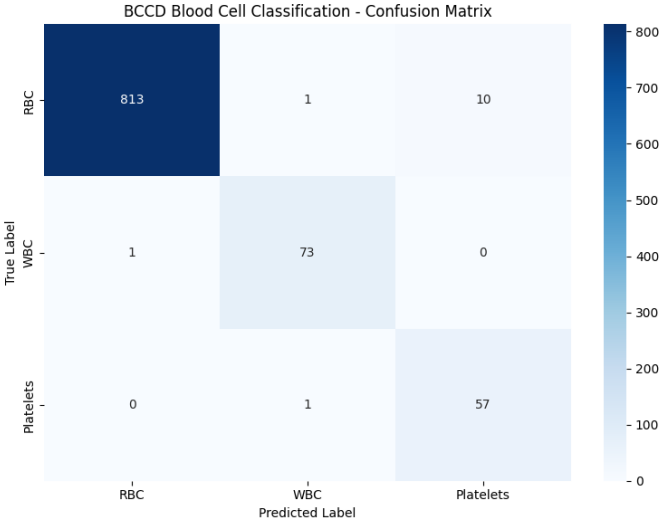

Fig. 1. Confusion Matrix Result

D. *Feature Importance Analysis*

After normalization, we conducted a feature variance analysis to identify which morphological characteristics most significantly influence the classification decision. Table 3 presents the top ten features with the highest variance according to the system.

TABLE III. TOP TEN MOST IMPORTANT FEATURES

Rank	Feature	Category	Variant
1	Aspect Ratio	Shape	1.0000
2	Edge Density	Edge Density	1.0000
3	Red Median	Color (R)	1.0000
4	Gray Maximum	Intensity	1.0000
5	Green Mean	Color (G)	1.0000
6	Gray Standard	Intensity	1.0000
7	Blue Mean	Color (B)	1.0000
8	Contrast (GLCM)	Texture	1.0000
9	Blue Median	Color (B)	1.0000
10	Green Standard	Color (G)	1.0000

The experimental results show that the form feature (aspect ratio) and the edge density feature (edge density) are the major discriminators. This is in agreement with our understanding of hematological morphology. RBCs are biconcave discs (aspect ratio near 1, smooth edges), WBCs are larger and irregular (more complex edges), and platelets are small. The second rank of the edge density feature indicates that the sharpness and complexity of the cell edges contribute to the differentiation between cells with smooth surfaces (RBC) and those with granular or lobed surfaces (WBC).

E. *Comparison with Previous Research*

Table 4 shows the accuracy comparison with previous studies [12] on the BCCD dataset or similar datasets.

TABLE IV. COMPARISON WITH PREVIOUS RESEARCH

Research	Method	Accuracy	Dataset
[6]	DL (CNN)	98.75%	BCCD (364 Images)
[3]	Decision Tree, SVM, etc	97%-98%	Artificial Dataset (Leukocyte)
This research	Naïve Bayes + 22 Features	98.64%	BCCD (364 Images)

The results indicate that the Naïve Bayes model with manual feature extraction achieves an accuracy of 98.64%, which is comparable to the advanced deep learning model published by [6] with an accuracy of 98.75%. The proposed method's advantages are significant: (1) quick training time (seconds), (2) no need for special GPUs, and (3) higher interpretability, since each feature has a clear morphological meaning.

F. Discussion

1) High Classification Accuracy for Red Blood Cells (RBC)

The model showed a very good performance in the identification of red blood cells. The precision was 1.00, and the recall was 0.99. This outcome is an example of the distinctive morphological characteristics of red blood cells, i.e., biconcave disc shape (aspect ratio around 1), lack of nucleus, and smooth cell edges (low edge density), which were reliably detected by the model with high accuracy.

2) Lower Precision in Platelet Classification

Platelets proved the most challenging class to classify, yielding the lowest precision of the three categories (0.85). The model most often confused platelets with RBCs, a pattern that reflects the platelets' minute size (2–3 microns) and their frequent overlap with, or adherence to, surrounding red blood cells, both of which obscure the shape and color cues the model relies on. Encouragingly, recall remained high at 0.98, meaning only one in fifty-eight platelet samples was missed, a level of sensitivity well suited to an initial screening role.

3) Clinical Implications for Computer-Aided Diagnosis

Taken together, an overall accuracy of 98.64% and a recall exceeding 0.98 across all classes point to genuine promise for integrating this model into computer-aided diagnosis (CAD) systems within clinical laboratories. High recall matters most in this setting, since a missed positive during initial screening could delay the detection of serious conditions such as leukocytosis or thrombocytopenia. The comparatively modest precision for platelets (0.85) need not be disqualifying, provided the model is deployed as a decision-support tool whose outputs are confirmed by a specialist rather than as a stand-alone diagnostic.

4) Study Limitations

But there are a number of constraints that temper these results. The BCCD dataset is small (364 original images) and is dominated by RBCs, so the model's ability to generalize to more balanced or heterogeneous datasets remains unknown. The success of the manual feature extraction pipeline relies on the correct segmentation of the cells. Especially overlapping cells can deteriorate the quality of the segmentation and thus the features extracted. Third, the model has not been validated with real-time images directly from digital microscopy equipment, where changes in illumination and resolution might add more difficulty.

IV. Conclusion

From the findings of the experiment and the discussions above, the following are the key conclusions that can be drawn about the classification of red blood cells, white blood cells, and platelets based on the 22 features extracted from the BCCD dataset:

1. The Naïve Bayes classifier attained an accuracy of 98.64%, together with precision, recall, and F1-score measures of 0.94, 0.99, and 0.96, respectively, providing ample proof of its efficacy in blood cell classification, especially for small sample datasets.
2. Aspect ratio and edge density are the most significant features in classifying blood cells, followed by color and intensity.
3. Machine learning is a powerful and effective tool for blood cell classification compared to deep learning. It performs efficiently when handling high-dimensional datasets, yields reliable results, and is less computationally expensive than the latter, all without compromising the level of accuracy of CNNs (98.75%).

The following approaches are capable of being used in future studies to improve the model's dependability and robustness and to manage the limitations found in this study:

1. Large and varied training datasets. The model's performance under various lighting conditions, resolutions, and staining methods should be investigated using a variety of big training datasets.
2. Cross-validation. The use of cross-validation (such as 5-fold and 10-fold) will reduce the likelihood of overfitting and produce objective classification results.
3. Application of ensemble algorithms. Particularly in the case of the platelet class, algorithms like Random Forest and XGBoost, which produce up to 97-98% accuracy on comparable datasets [6], can be used to improve classification accuracy.
4. Segmentation. Application of segmentation methods using a deep learning approach (U-Net, for instance) will allow for more precise segmentation of overlapping cells and feature extraction.
5. Deep feature fusion. Combining features obtained using CNN layers and manually calculated features will help improve discrimination power in the case of difficult classes (platelets, for example).

All the above-mentioned techniques will increase the robustness of the system of blood cell classification.

References

[1] M. Toptaş, B. Toptaş, and D. Hanbay, "Classifying white blood cells using combining different convolutional neural networks," *Multimed. Tools Appl.*, vol. 84, no. 35, pp. 44089–44112, 2025, doi: 10.1007/s11042-025-20879-y.

[2] R. Asghar, S. Kumar, A. Shaukat, and P. Hynds, "Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review," *PLoS One*, vol. 19, no. 6, pp. 1–25, 2024, doi: 10.1371/journal.pone.0292026.

[3] M. Shahzad *et al.*, "Blood cell image segmentation and classification: a systematic review," *PeerJ Comput. Sci.*, vol. 10, pp. 1–35, 2024, doi: 10.7717/peerj-cs.1813.

[4] T. M. Makki and K. Al-Dulaimi, "Blood Cell Microscopic Image Classification in Computer Aided Diagnosis Using Machine Learning: A Review," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 2, 2023, doi: 10.52866/ijcsm.2023.02.02.002.

[5] N. Markose and M. S. Elayidom, "BLOOD CELL DETECTION AND CLASSIFICATION WITH AUTOENCODER-DRIVEN R-CNN," vol. 5, no. 2, pp. 344–350, 2025.

[6] T. Fang, X. Huang, X. Chen, D. Chen, J. Wang, and J. Chen, "Segmentation, feature extraction and classification of leukocytes leveraging neural networks, a comparative study," *Cytom. Part A* , vol. 105, no. 7, pp. 536–546, 2024, doi: 10.1002/cyto.a.24832.

[7] A. Sulaiman *et al.*, "ResRandSVM: Hybrid Approach for Acute Lymphocytic Leukemia Classification in Blood Smear Images," *Diagnostics*, vol. 13, no. 12, 2023, doi: 10.3390/diagnostics13122121.

[8] R. M. Carrillo-Larco, L. Tudor Car, J. Pearson-Stuttard, T. Panch, J. J. Miranda, and R. Atun, "Machine learning health-related applications in

low-income and middle-income countries: a scoping review protocol," *BMJ Open*, vol. 10, no. 5, pp. 1–5, 2020, doi: 10.1136/bmjopen-2019-035983.

[9] S. Latha, P. Muthu, S. Dhanalakshmi, R. Kumar, K. W. Lai, and X. Wu, "Emerging Feature Extraction Techniques for Machine Learning-Based Classification of Carotid Artery Ultrasound Images," *Comput. Intell. Neurosci.*, vol. 2022, no. Ml, 2022, doi: 10.1155/2022/1847981.

[10] F. Rustam *et al.*, "White Blood Cell Classification Using Texture and RGB Features of Oversampled Microscopic Images," *Healthc.*, vol. 10, no. 11, pp. 1–21, 2022, doi: 10.3390/healthcare10112230.

[11] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[12] P. Jeneessha and V. K. Balasubramanian, "Domain knowledge-infused pre-trained deep learning models for efficient white blood cell classification," *Sci. Rep.*, vol. 15, no. 1, pp. 1–22, 2025, doi: 10.1038/s41598-025-00563-9.

[13] T. Banerjee and İ. Paçal, "A systematic review of machine learning in heart disease prediction," *Turkish J. Biol.*, vol. 49, no. SI-5, pp. 600–634, 2025, doi: 10.55730/1300-0152.2766.

[14] T. Islam *et al.*, "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-57740-5.

[15] O. J. Egwom, M. Hassan, J. J. Tanimu, M. Hamada, and O. M. Ogar, "An LDA–SVM Machine Learning Model for Breast Cancer Classification †," *BioMedInformatics*, vol. 2, no. 3, pp. 345–358, 2022, doi: 10.3390/biomedinformatics2030022.

[16] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification Prediction of Breast Cancer Based on Machine Learning," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, 2023, doi: 10.1155/2023/6530719.

[17] A. Jafari, "Machine-learning methods in detecting breast cancer and related therapeutic issues: a review," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 12, no. 1, 2024, doi: 10.1080/21681163.2023.2299093.

[18] P. Shinde, M. Sanghavi, and T. A. Tran, "A Survey on Machine Learning Techniques for Heart Disease Prediction," *SN Comput. Sci.*, vol. 6, no. 4, 2025, doi: 10.1007/s42979-025-03860-2.

[19] Y. Mao, B. L. Jimma, and T. B. Mihretie, "Machine learning algorithms for heart disease diagnosis: A systematic review," *Curr. Probl. Cardiol.*, vol. 50, no. 8, p. 103082, 2025, doi: 10.1016/j.cpcardiol.2025.103082.

[20] A. La Moglia and K. Mohamad Almustafa, "Breast cancer prediction using machine learning classification algorithms," *Intell. Med.*, vol. 11, no. December 2024, p. 100193, 2025, doi: 10.1016/j.ibmed.2024.100193.