MOBILE AND FORENSICS

ISSN 2656-6257

Vol. 8, No. 1, March 2026, pp. 65-92

Resource-Efficient Optimization for Multi-Class Hematological Diagnosis: A Hybrid BPSO-Extra Trees Approach with Data Imbalance Handling

Dimas Chaerul Ekty Saputra [a,b,1,*], Zahid Abdullah Nur Mukhlishin [c,2], Affifah Mutiara Pertiwi [a,b,3], Mochammad Zulfikar Alfany [d,4], Irianna Futri [e,5], Raksmey Phann [f,6]

[a] Informatics Study Program, Telkom University, Surabaya Campus, Surabaya 60231, East Java, Indonesia
[b] Center of Excellence for Motion Technology for Safety Health and Wellness, Research Institute of Sustainable Society, Telkom University, Surabaya Campus, Surabaya 60231, East Java, Indonesia
[c] Data Science Study Program, Telkom University, Surabaya Campus, Surabaya 60231, East Java, Indonesia
[d] Industrial Engineering Study Program, Telkom University, Bandung 40257, West Java, Indonesia
[e] Faculty of Nursing, Ratchathani University, Ubonratchathani 34000, Thailand
[f] Department of Data Science, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
[1] dimaschaerulekty@telkomuniversity.ac.id*, [2] zahidabdullahnur@student.telkomuniversity.ac.id, [3] affifahmutiara@telkomuniversity.ac.id,
[4] mochammadzulfikara@student.telkomuniversity.ac.id, [5] irianna.f@kkumail.com, [6] phann.raksmey@ds.seoultech.ac.kr
*corresponding author

ARTICLE INFO

Article history
Received : May 20, 2025
Revised : Nov 10, 2025
Accepted : Mar 24, 2026

Keywords
Complete Blood Count
Hematological Disorders
Feature Selection
Binary Particle Swarm Optimizatio
Computational Efficiency

ABSTRACT

Complete Blood Count (CBC) remains the cornerstone for initial screening of hematological disorders, yet manual interpretation is often challenged by overlapping biological patterns and substantial inter-patient variability. Although machine learning approaches have demonstrated promise for automated diagnosis, many existing studies prioritize classification accuracy while neglecting computational efficiency and the persistent class imbalance inherent in medical datasets. This study develops a lightweight yet effective diagnostic framework for classifying nine hematological conditions using routine CBC parameters. Evaluated on a public dataset of 1,281 records from Kaggle, the proposed model is benchmarked against standard Random Forest, XGBoost, and Support Vector Machine (SVM) classifiers. The approach integrates the Synthetic Minority Oversampling Technique (SMOTE) to mitigate class imbalance, and Binary Particle Swarm Optimization (BPSO) to identify a compact and clinically informative feature subset of exactly 6 parameters, referred to as a clinical fingerprint, optimized for the Extra Trees classifier. Evaluated using ten-fold cross-validation, the BPSO-Extra Trees model achieved an average accuracy of 87.43 percent and an F1 score of 82.75 percent, while demonstrating superior resource efficiency with peak memory consumption of only 0.249 MB, corresponding to a 46.6 percent reduction compared with the standard Random Forest baseline. These findings confirm that swarm intelligence optimized models can effectively balance diagnostic performance with extreme computational frugality, enabling the potential deployment of accurate hematology-based decision support systems on portable devices and in resource-limited laboratory environments.

1. Introduction

Hematological parameters serve as fundamental indicators for diagnosing a wide array of systemic conditions, ranging from nutritional deficiencies to malignancies. The Complete Blood Count (CBC) is the most ubiquitous screening test in clinical practice due to its cost-effectiveness and minimally invasive nature. Its diagnostic utility extends beyond standard blood disorders; recent studies by Yi et al. [1] and Mishra et al. [2] have demonstrated the critical role of CBC parameters

https://doi.org/10.12928/mf.v8i1.15703

in predicting postpartum hemorrhage severity and diagnosing pediatric sepsis in resource-limited settings. However, the interpretation of CBC data is often complicated by the biological heterogeneity of blood disorders. Liu et al. [3] highlight that while modern analyzers generate vast amounts of data, manual assessment remains prone to subjectivity and error. This is compounded by "diagnostic pitfalls" described by Gulati et al. [4] and Kubiak et al. [5], where pre-analytical variables or overlapping symptoms can lead to misclassification in automated systems.

To overcome these limitations, Machine Learning (ML) has been increasingly adopted to assist pathologists. Recent studies have demonstrated the versatility of ML across various medical domains; for instance, Hong et al. [6] successfully utilized XGBoost and Random Forest to predict disease severity in COVID-19 patients, while Esmaily et al. [7] compared decision trees for diabetes risk stratification. In the specific context of hematology, Airlangga [8] and Amjad et al. [9] demonstrated that hybrid ML models could significantly improve diagnostic precision for anemia and leukemia by capturing complex data patterns that traditional statistical methods miss. However, a major bottleneck in deploying these models to real-world clinical settings is the computational demand. Airlangga [10] emphasizes the critical need for "resource-aware" diagnostics that can operate on low-power hardware without sacrificing accuracy.

Despite the promise of ML in hematology, a critical research gap remains in balancing diagnostic accuracy with operational efficiency. Recent studies increasingly favor computationally expensive image-based methods (e.g., deep learning for blood smear analysis) or rely on heterogeneous, non-routine parameters. Conversely, existing tabular models using routine Complete Blood Count (CBC) data often overlook severe class imbalance, leading to biased predictions for rare conditions. Medical datasets are inherently plagued by this imbalance, where healthy samples vastly outnumber pathological ones. Recent systematic reviews by Yang et al. [11] and Hairani et al. [12] point out that standard classifiers often fail in this context, biasing predictions toward the majority class. Zhu [13] reinforces this in the context of medical imaging, noting that imbalance can lead to model collapse. To address this, Dhibar et al. [14] and Kumar et al. [15] recently proposed advanced optimization models for chronic and heart disease prediction, proving that metaheuristic algorithms can effectively navigate imbalanced high-dimensional data. Therefore, there is an urgent need for a unified, resource-aware framework that exclusively leverages routine CBC parameters to extract a minimal yet highly discriminative 'clinical fingerprint', enabling efficient deployment on low-power hardware.

Although these findings collectively highlight the strength of ML across diverse clinical applications, existing studies remain largely disease-specific and often rely on heterogeneous datasets that lack a unified analytical framework. Studies on chronic kidney disease [16], cancer diagnosis [17], and metabolic syndromes [18] consistently report that without appropriate intervention, diagnostic sensitivity for critical conditions is compromised. Moreover, the demand for rapid screening tools necessitates models that are not only accurate but also computationally efficient, avoiding the "curse of dimensionality" associated with redundant features [19], [20]. As a result, there is still a limited understanding of how hybrid evolutionary algorithms combined with ensemble learning can optimize diagnostic precision while maintaining operational efficiency on imbalanced data.

To address these limitations, the present study aims to systematically develop and evaluate a hybrid machine learning framework for hematology-based disease classification. Specifically, this research investigates the performance of a wrapper-based feature selection approach that utilizes Binary Particle Swarm Optimization (BPSO) and Binary Grey Wolf Optimization (BGWO) when integrated with the Extra Trees Classifier. In addition, the study evaluates the diagnostic reliability of the proposed framework using comprehensive performance metrics under ten-fold cross-validation, while strategically applying the Synthetic Minority Oversampling Technique (SMOTE) to restore balance within the feature space. Through this approach, the study seeks to identify the

most suitable model for cost-effective diagnostic applications and to determine whether the proposed BPSO-Extra Trees model outperforms established industry standards such as Random Forest and Support Vector Machine. Based on these objectives, the key contributions of this study can be summarized as follows:

1) Development of a high-efficiency framework is achieved through the identification of a minimal hematological feature subset, known as a clinical fingerprint, that maintains diagnostic accuracy above 82% while reducing memory usage to under 0.25 MB for cost-effective solutions in resource-constrained laboratories.
2) Advanced metaheuristic integration involves the application of BPSO and BGWO optimization algorithms as wrappers to capture non-linear feature interactions that are often missed by traditional feature selection methods.
3) Comprehensive validation entails a rigorous comparative analysis between ensemble and non-ensemble machine learning models under identical preprocessing, resampling, and cross-validation conditions to demonstrate the superiority of Extra Trees in balancing accuracy and speed.
4) Resilience to data imbalance is empirically demonstrated by showing that the integration of SMOTE with optimized Extra Trees effectively handles majority class bias without compromising precision for rare disease categories [21], [22]

Despite progress in prior research, several key gaps remain unaddressed and are directly confronted in this work. Most previous studies focus on single diseases or rely on diverse feature groups, leaving a need for research that systematically investigates multi-class performance using a unified hematology-only dataset. Additionally, limited research has simultaneously examined predictive accuracy and computational efficiency, two factors that are critical for practical implementation in healthcare systems [23], [24]. Finally, the integration of advanced resampling with wrapper-based feature selection remains underexplored in many traditional ML diagnostic models. By addressing these gaps, this study advances the development of hematology-based machine learning diagnostics and reinforces the role of hybrid evolutionary algorithms as a powerful and practical approach for automated disease classification.

To guide the flow of this study, this paper is organized into five main sections. Section 1 presents the introduction, outlining the background of the problem, motivations, research gaps, objectives, and key contributions. Section 2 discusses the related works derived from an extensive review of current literature. Section 3 describes the materials and methods employed in this study, including data collection, preprocessing procedures, machine learning algorithms, and evaluation strategies. Section 4 presents the experimental results, followed by a comprehensive discussion that interprets the findings and compares them with existing studies. Finally, Section 5 concludes the study by summarizing the major outcomes and offering recommendations for future research directions.

The application of machine learning in hematology has evolved from simple statistical tests to complex predictive modeling. In the domain of anemia, Vohra et al. [25] and Putri and Mukti [26] evaluated various classifiers and found that ensemble methods like XGBoost generally outperform single classifiers like Naive Bayes. Similarly, Haider et al. [27] successfully applied predictive modeling to differentiate leukemia subtypes using extended CBC parameters. Saputra et al. [28] recently advanced this field by implementing Extreme Gradient Boosting for multi-class blood disease classification, achieving high accuracy but highlighting the need for further computational optimization. Alternative approaches have also been explored; Ameen et al. [29] utilized fuzzy logic to handle the ambiguity of disease boundaries.

A significant portion of hematology research focuses on image-based analysis. Sazak and Kotan [30] employed the YOLOv11 architecture for blood cell detection, while Alam and Islam [31] utilized YOLO for automatic cell counting. However, image-based methods are computationally

expensive. Annisa et al. [32] attempted to reduce computational load using statistical feature selection (ANOVA/Chi-Square), but noted that filter-based methods often overlook non-linear feature interactions. To address feature redundancy more effectively, metaheuristic optimization has gained significant attention. Similarly, Muharni et al. [33] successfully optimized Naive Bayes using PSO for heart disease prediction.

Recent advancements have focused on hybrid metaheuristics due to their superior convergence speed. Wijayanto et al. [34] applied hybrid algorithms for feature selection in EEG signal classification, achieving high accuracy with reduced features. In the medical text domain, Subkhi et al. [35] utilized Hybrid Binary Grey Wolf Optimization (GWO), while Saabia and Frikha [36] proposed a GWO-Simulated Annealing hybrid for high-dimensional medical datasets. The versatility of these methods is further evidenced by Alamsyah et al. [37], who enhanced XGBoost with Genetic Algorithms for diabetes prediction, and Gürkan Kuntalp et al. [38], who compared various metaheuristics for respiratory disease classification. Khah et al. [39] provided a survey on these approaches, confirming that swarm-based wrappers are state-of-the-art for selecting the "clinical fingerprint." Additionally, Stephan et al. [40] and Liu and Tian [41] introduced novel GWO variants that balance exploration and exploitation, which are crucial for our proposed framework.

In terms of classification algorithms, tree-based models have shown resilience in medical diagnostics. Hidaka et al. [21] and Yang et al. [42] argued that tree ensembles are particularly robust for imbalanced data. Specifically, the Extra Trees classifier has demonstrated superior efficiency compared to Random Forest. Erni and Sa'adah [43] reported that Extra Trees achieved the highest accuracy (86.93%) in heart disease detection compared to Naive Bayes and Random Forest. This finding is supported by Al Ghifari et al. [44] and Nzenwata et al. [23], who observed faster execution times and higher stability with Extra Trees in liver and cardiac disease prediction, respectively. This speed advantage is theoretically supported by Melanson [24], who explained that the randomization of split points in Extra Trees drastically reduces computational overhead.

Finally, addressing class imbalance remains a priority. Naseriparsa et al. [45] and Alamsyah et al. [46] compared various sampling strategies and concluded that oversampling techniques like SMOTE generally yield better sensitivity than undersampling. Studies on Chronic Kidney Disease by Salau et al. [16] and Cancer diagnosis by Gurcan and Soylu [17] have confirmed that resampling is essential for valid predictions. Specific applications of SMOTE have proven successful in other fields, as seen in the work of Nurhayati and Rahardi [18] on Metabolic Syndrome and Riadi et al. [22] on Diabetes. Building on these diverse findings, including the insights from Akbar et al. [20] on hybrid optimization, this study integrates SMOTE, BPSO, and Extra Trees into a unified framework to classify multi-class hematological disorders efficiently.

Collectively, these studies demonstrate the strength of machine learning across diverse hematological applications. However, current literature remains fragmented. Most studies focus on single diseases (only anemia or leukemia), rely on filter-based feature selection that misses non-linear interactions, or do not fully integrate swarm intelligence with robust resampling for multi-class datasets. There is limited research exploring a unified framework combining Binary Particle Swarm Optimization (BPSO) with Extra Trees to handle class imbalance via SMOTE. These gaps, summarized in Table 1, highlight the need for developing a clinical fingerprint model that relies on a minimal but highly effective feature subset.

Table 1. Detailed Comparison of Recent Literature in Hematological Classification and Optimization

Reference	Disease Domain	Classifier Model	Feature Selection Approach	Imbalance Handling	Research Gap / Limitation
[26]	Anemia (3 Classes)	RF, SVM, XGBoost, Naive Bayes	None (All features)	None	Limited to anemia subtypes; did not address class imbalance or feature redundancy.
[27]	Leukemia (Subtypes)	Deep Learning (CNN)	Expert-defined (Research params)	None	Relies on specialized research parameters not available in standard routine CBC tests.
[8]	Anemia (Binary & Multi)	Stacking Ensemble, LightGBM	VIF, PPS (Filter-based)	None	Focuses solely on anemia; filter methods often miss complex non-linear feature interactions.
[32]	Anemia	KNN, SVM, Naive Bayes	ANOVA, Chi-Square (Filter-based)	None	Filter-based selection ignores classifier interaction; no handling of data imbalance.
[16]	Chronic Kidney Disease	RF, SVM, Decision Tree	None	SMOTE	Demonstrates SMOTE utility but applied to kidney disease, not multi-class hematology.
[3]	General Hematology	XGBoost, RF, LR	None	None	Large-scale study but lacks optimization for minimal feature subset ("clinical fingerprint").
Proposed Method	9 Hematological Classes	Extra Trees Classifier	Binary PSO & GWO (Wrapper)	SMOTE	Integrates swarm optimization for minimal features and SMOTE for imbalance in a unified, lightweight multi-class framework.

2. Method

This section presents the methodological framework used in this study, covering all procedures from data acquisition to model evaluation. Subsection 3.1 describes the data collection process, including dataset characteristics and ethical considerations. Subsection 3.2 explains the research flow, outlining each step of the analytical pipeline from raw data processing to model training. Subsection 3.3 introduces the data-level preprocessing, specifically the Synthetic Minority Oversampling Technique (SMOTE), which is applied to address class imbalance. Subsection 3.4 details the proposed hybrid model, integrating Binary Grey Wolf Optimization (BGWO) for feature selection with the Extra Trees Classifier, along with the justification for its use in analyzing hematological parameters. Finally, Subsection 3.5 presents the evaluation metrics employed to assess model performance, encompassing accuracy-based measures, computational efficiency, and confusion matrix interpretation.

2.1. Data Collection

The dataset used in this study was obtained from a publicly available secondary source hosted on the Kaggle platform (https://kaggle.com/datasets/ehababoelnaga/anemia-types-classification). The dataset contains Complete Blood Count (CBC) records labeled with corresponding anemia diagnoses. All samples were originally collected from real clinical examinations, and the diagnostic labels were assigned manually by medical professionals, ensuring that the dataset reflects authentic hematological patterns associated with various anemia types and related disorders. Since the dataset is openly accessible and does not contain personal identifiers or protected health information, the use of this data complies with ethical standards for secondary data analysis. Formal ethical clearance was not required because the dataset is fully anonymized and provided for research purposes under Kaggle's open licensing policy. However, the study adheres to general ethical principles concerning the responsible use of clinical datasets in machine learning research.

The dataset consists of CBC parameters that represent routine hematological measurements commonly used in medical diagnostics. These features include Hemoglobin concentration (HGB), Platelet count (PLT), White Blood Cell count (WBC), Red Blood Cell count (RBC), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Platelet Distribution Width (PDW), and Procalcitonin (PCT). Each parameter plays a distinct clinical role: HGB reflects oxygen transport capacity; PLT contributes to blood clotting; WBC indicates immune response; RBC represents oxygen-carrying potential; MCV, MCH, and MCHC describe red blood cell morphology and hemoglobin content; PDW measures variability in platelet size; and PCT is used to evaluate the risk of sepsis. The target variable, Diagnosis, specifies the anemia type or related condition based on the CBC values.

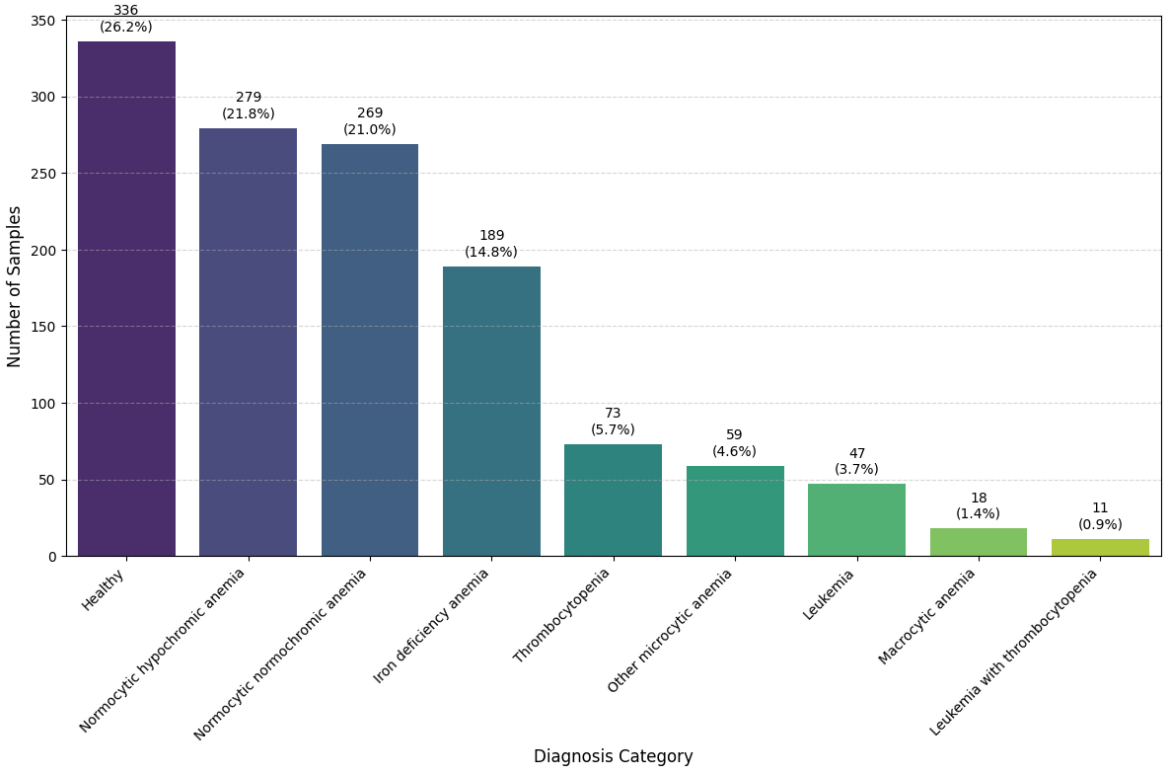

Fig. 1. Class Distribution

As illustrated in Fig. 1, the dataset exhibits a distinct multiclass distribution across nine diagnostic categories. The vertical bar chart clearly highlights the significant class imbalance present in the data. The majority classes include Healthy (336 samples, 26.23%), Normocytic hypochromic anemia (279 samples, 21.78%), and Normocytic normochromic anemia (269 samples, 21.00%), which together constitute the bulk of the dataset. In contrast, the remaining categories show a steep decline in frequency. Intermediate classes include Iron deficiency anemia (189 samples, 14.75%), Thrombocytopenia (73 samples, 5.70%), and Other microcytic anemia (59 samples, 4.61%). Less frequent classes such as Leukemia (47 samples, 3.67%), Macrocytic anemia (18 samples, 1.41%), and Leukemia with thrombocytopenia (11 samples, 0.86%) represent rare conditions but are clinically important categories that contribute to the dataset's complexity and imbalanced structure. This skewed distribution, where the majority class outnumbers the minority class by a factor of over 30, underscores the complexity of the classification task and justifies the necessity of applying data-level preprocessing strategies, such as the Synthetic Minority Oversampling Technique (SMOTE), to prevent model bias and ensure robust sensitivity for rare diseases. In summary, the dataset provides a comprehensive and clinically relevant source of hematological information suitable for developing and evaluating machine learning models for disease classification based on CBC parameters as shown in Table 2.

Table 2. Dataset

No	WBC	LYMp	NEUTp	LYMn	NEUTn	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	PDW	PCT	Diagnosis
1	10.0	43.2	50.1	4.3	5.0	2.77	7.3	24.2	87.7	26.3	30.1	189	12.5	0.17	Normocytic hypochromic anemia
2	10.0	42.4	52.3	4.2	5.3	2.84	7.3	25.0	88.2	25.7	20.2	180	12.5	0.16	Normocytic hypochromic anemia
3	7.2	30.7	60.7	2.2	4.4	3.97	9.0	30.5	77.0	22.6	29.5	148	14.3	0.14	Iron deficiency anemia
4	6.0	30.2	63.5	1.8	3.8	4.22	3.8	32.8	77.9	23.2	29.8	143	11.3	0.12	Iron deficiency anemia
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

1278	5.6	25.8	77.5	1.88	5.14	4.85	15.0	46.1	91.7	31.0	33.8	215	14.3	0.26	Healthy
1279	9.2	25.8	77.5	1.88	5.14	4.47	13.1	46.1	88.7	29.3	33.0	329	14.3	0.26	Healthy
1280	6.48	25.8	77.5	1.88	5.14	4.75	13.2	46.1	86.7	27.9	32.1	174	14.3	0.26	Healthy
1281	8.8	25.8	77.5	1.88	5.14	4.95	15.2	46.1	89.7	30.6	34.2	279	14.3	0.26	Healthy

2.2. Research Flow

The flow of this study is organized into a sequence of methodological stages designed to convert raw hematological data into a reliable machine learning classification model as shown in Fig. 2. The stages include data collection, data preprocessing, feature optimization, and model development and evaluation. Each stage is implemented systematically to ensure that the resulting model is built on high-quality data, processed under consistent analytical procedures, and evaluated rigorously using validated performance measures.

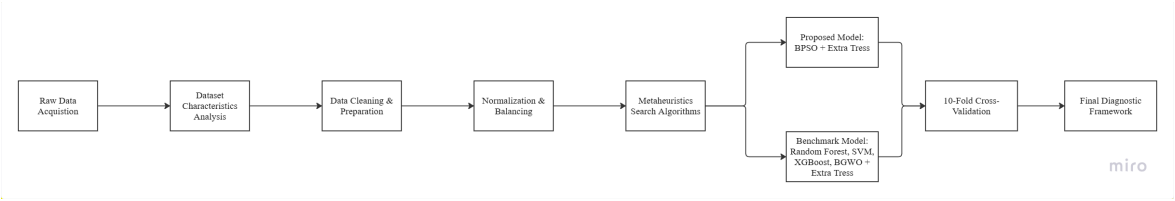

Fig. 2. Research Flow

The second stage, Data Preprocessing, is essential for improving dataset quality and ensuring suitability for machine learning. The process begins with data preparation where relevant features are identified and extracted while the dataset structure is organized for analysis. Data cleaning is then performed to address issues such as missing values, measurement inconsistencies, and noise that may negatively affect model performance. Numerical variables are transformed using Min-Max Normalization to scale all features into the range of 0 to 1 which is critical for the convergence of the subsequent optimization algorithms. Diagnostic labels are converted into numerical form through label encoding to ensure compatibility with machine learning algorithms. Furthermore, because the dataset exhibits significant class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is applied to generate synthetic samples of the minority class such as Leukemia, enabling the learning algorithms to better recognize underrepresented patterns without bias.

The third stage, Feature Optimization, distinguishes this study from conventional approaches. Instead of using all available features, two advanced metaheuristic wrapper methods known as Binary Particle Swarm Optimization (BPSO) and Binary Grey Wolf Optimization (BGWO) are employed to identify the optimal feature subset. These bio-inspired algorithms iteratively navigate the search space to find a "clinical fingerprint" which represents the minimal combination of hematological parameters that yields the highest predictive accuracy. This step effectively removes redundant features caused by multicollinearity and reduces the computational complexity of the final model.

The final stage, Model Development and Evaluation, involves constructing the classification framework and assessing its diagnostic performance. The primary classifier proposed in this study is the Extra Trees or Extremely Randomized Trees algorithm, selected due to its ability to reduce variance and prevent overfitting on high-dimensional medical data. To validate the proposed framework, the optimized models are benchmarked against industry-standard classifiers, including Random Forest, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). To ensure reliable evaluation, each model is trained and validated using 10-fold cross-validation, a method which divides the dataset into ten partitions and iteratively trains on nine partitions while

evaluating the remaining one. This procedure provides a stable estimate of model performance. During this stage, key performance metrics such as accuracy, precision, recall, F1-score, computational time, and memory usage are measured and recorded. These metrics allow both predictive capability and computational efficiency to be examined simultaneously, enabling a balanced assessment of the suitability of the proposed framework for rapid hematology-based disease classification.

Through this sequence of steps, the research flow produces a methodologically structured and reproducible machine learning pipeline. The integration of clinically sourced hematological data, rigorous preprocessing with SMOTE, swarm intelligence feature selection, and multi-model performance evaluation provides a strong foundation for developing robust diagnostic decision making tools.

2.3. Synthetic Minority Oversampling Technique (SMOTE)

The hematology dataset used in this study exhibits considerable class imbalance, with several diagnostic categories, such as Leukemia (3.7%) and Leukemia with thrombocytopenia (0.9%), having substantially fewer samples compared to the Healthy and Normocytic anemia classes. Such skewed distribution biases machine learning models toward majority classes, resulting in poor sensitivity, low recall, and unreliable predictions for minority disease categories. To address this challenge, the Synthetic Minority Oversampling Technique (SMOTE) is employed during preprocessing to enhance class representation and improve model learning quality.

Crucially, to rigorously prevent data leakage, SMOTE is applied strictly and exclusively to the training partitions within each iteration of the 10-fold cross-validation loop. The test folds remain entirely untouched and synthetic-free, ensuring the evaluation reflects true real-world diagnostic performance on imbalanced data.

SMOTE is an oversampling algorithm that generates synthetic minority-class samples rather than simply duplicating existing ones. This is achieved by interpolating between a minority-class sample x_i and its nearest neighbors in the feature space. For each minority instance x_i, the algorithm selects one of its k-nearest neighbors x_zi and generates a synthetic sample x_new using:

$$x_{new} = x_i + \delta \times (x_{zi} - x_i) \quad (1)$$

Where δ is a random vector in the interval [0,1]. This interpolation process creates new, plausible samples that lie along the feature-space trajectory between existing minority observations, effectively expanding the decision region for rare diseases without the overfitting risks associated with random oversampling. By applying SMOTE within the training folds, the class distribution becomes balanced, enabling the Extra Trees classifier to learn discriminative patterns for both common and rare hematological disorders equitably

2.4. Proposed Framework: Optimization and Classification

This study proposes a hybrid diagnostic framework that integrates swarm intelligence-based feature selection with an ensemble classification model. Specifically, Binary Particle Swarm Optimization (BPSO) is utilized as the primary proposed wrapper method to identify the optimal feature subset ("clinical fingerprint"). To validate the efficacy of the proposed BPSO approach, Binary Grey Wolf Optimization (BGWO) is included as a comparative metaheuristic baseline, processed by the same Extra Trees classifier.

2.4.1. Feature Optimization Strategy (Wrapper Methods)

Feature selection is modeled as a binary optimization problem where the search space is an N-dimensional Boolean lattice, with N representing the number of hematological parameters. A feature subset is represented as a vector $X = \{x_1, x_2, \dots, x_N\}$, where $x_j \in \{0,1\}$. If $x_j = 1$, the feature is selected; otherwise, it is discarded.

A. Binary Particle Swarm Optimization (BPSO) BPSO optimizes feature selection by simulating the social behavior of bird flocking. Each particle i maintains a position X_i and velocity V_i. In the binary domain, the velocity represents the probability of a position bit taking the value 1. The velocity update equation at iteration $t+1$ is defined as:

$$v_{ij}^{t+1} = wv_{ij}^{t} + c_1 r_1 \left(pbest_{ij}^{t} - x_{ij}^{t}\right) + c_2 r_2 \left(gbest_{j}^{t} - x_{ij}^{t}\right) \quad (2)$$

Where w is the inertia weight, c_1, c_2 are acceleration coefficients, and r_1, r_2 are random values. The position is updated using a sigmoid transfer function $S(v)$:

$$S(v_{ij}^{t+1}) = \frac{1}{1 + e^{-v_{ij}^{t+1}}} \quad (3)$$

$$x_{ij}^{t+1} = \begin{cases} 1 & \text{if } r < S(v_{ij}^{t+1}) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

B. Binary Grey Wolf Optimization (BGWO) BGWO mimics the leadership hierarchy of grey wolves (Alpha α, Beta β, Delta δ). The hunting mechanism is mathematically modeled by updating the position of the wolves (omega ω) based on the three best solutions found so far. The distance vector $\vec{D}$ is calculated as:

$$\vec{D} = |\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)| \quad (5)$$

Where $\overrightarrow{X_p}$ is the position of the prey (optimum), $\vec{X}$ is the wolf's position, and $\vec{C} = 2 \cdot \overrightarrow{r_2}$. The new position $\vec{X}(t+1)$ is determined by the crossover of the three leaders:

$$\vec{X}(t+1) = \text{crossover}(\vec{X}_1, \vec{X}_2, \vec{X}_3) \quad (6)$$

Where $\overrightarrow{X_1}, \overrightarrow{X_2}, \overrightarrow{X_3}$ are the binary vectors updated relative to α, β, and δ using a sigmoid-based binarization step similar to Eq. (3).

The Extra Trees (Extremely Randomized Trees) classifier is selected as the core predictor due to its computational efficiency and resistance to overfitting. Unlike Random Forest, which searches for the optimal split threshold using bootstrap samples, Extra Trees utilizes the entire dataset and selects split points completely at random.

Given a training set S, the algorithm constructs an ensemble of M decision trees. For a node containing samples S_{node}, the split s is determined by maximizing the split score:

$$\text{Score}(s, S) = \frac{|S_L|}{|S|} G(S_L) + \frac{|S_R|}{|S|} G(S_R) \quad (7)$$

Where S_L and S_R are the subsets created by the split, and $G(S)$ is the Gini Impurity index defined as:

$$G(S) = 1 - \sum_{c=1}^{C} p_c^2 \quad (8)$$

Where p_c is the probability of class c in the node. The randomness in split selection ($t_j \sim U[\min(f_j), \max(f_j)]$) significantly reduces variance. The final prediction $\hat{y}$ for a new sample x is obtained by majority voting across all trees:

$$G(S) = 1 - \sum_{c=1}^{C} p_c^2 \quad (8)$$

Mathematically, for a given node containing a subset of samples S, the algorithm selects K random attributes $\{a_1, \ldots, a_K\}$. For each attribute a_k, a random cut-point t_k is drawn from the range $[min(a_k), max(a_k)]$ within S. The quality of this random split $s = (a_k, t_k)$ is then evaluated using the Score Function based on Information Gain reduction. The score is calculated as:

$$\hat{y}(\mathbf{x}) = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_M(\mathbf{x})\} \quad (9)$$

The workflow of the proposed hybrid model is summarized in Algorithm 1.

Algorithm 1: Proposed BGWO-Extra Trees Framework
Input: Training dataset D, Population size P, Max Iterations T
Output: Optimal feature subset F_best, Trained Model M
1. Initialize wolf population X_i (i=1...P) with random binary vectors 2. Initialize Alpha, Beta, Delta scores = infinity 3. While t < T do: 4. For each wolf X_i: 5. Train Extra Trees model using features in X_i 6. Calculate Fitness = (0.9 * (1-Accuracy)) + (0.1 * (Feature_Ratio)) 7. Update Alpha, Beta, Delta based on Fitness 8. End For 9. For each wolf X_i: 10. Update position X_i using Eq. (5) and (6) 11. End For 12. t = t + 1 13. End While 14. F_best = Position of Alpha 15. Train final Extra Trees model M using D with features F_best 16. Return M

2.4.2. Implementation and Hyperparameter Settings

To ensure reproducibility and address concerns regarding experimental consistency, the hyperparameter configurations are explicitly defined. The BPSO algorithm (and the comparative BGWO) was configured with a population size of 20 and executed for 30 iterations. The Extra Trees and Random Forest classifiers were instantiated using 100 estimators (trees) with the random state fixed at 42 to guarantee deterministic behavior. Default regularization parameters were utilized for the SVM baseline. All experiments were conducted using Python 3.9 with the Scikit-learn library.

2.5. Evaluation Metrics

Model performance in this study is assessed using a set of widely accepted evaluation metrics for classification tasks, including confusion matrix, accuracy, precision, recall, F1-score, time complexity, and memory usage. These metrics provide a comprehensive assessment of both predictive performance and computational efficiency, ensuring that the selected classifier is not only accurate but also practical for deployment in real-world clinical environments.

The evaluation begins with the confusion matrix, which summarizes the prediction results by categorizing samples into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The structure of the confusion matrix used in this study is presented in Table 3.

Table 3. Confusion Matrix

Predicted \ Actual	**Positive**	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

This matrix serves as the basis for deriving several performance indicators. The first metric, Accuracy, measures the proportion of correct predictions among all samples and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

However, accuracy alone may be misleading in imbalanced datasets. Therefore, additional metrics are employed to capture classification behavior in more detail. Precision quantifies the proportion of positive predictions that are correct and is formulated as:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall, also known as sensitivity, measures the model's ability to correctly identify positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

To balance precision and recall, the F1-score is calculated as the harmonic mean of both metrics:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

In addition to classification performance, this study evaluates computational efficiency, which is essential for clinical diagnostic systems that require fast and resource-efficient processing. Execution Time is recorded to measure the total computational duration required to train and validate the model within each fold of the 10-fold cross-validation. It reflects the algorithm's processing speed and responsiveness.

$$Execution\_Time = t_{end} - t_{start} \tag{14}$$

Similarly, Memory Usage quantifies the amount of system memory consumed during model training. This metric indicates the scalability and feasibility of deploying the model on devices with limited hardware capacity.

$$Memory\_Usage = RAM_{peak} - RAM_{initial} \tag{15}$$

Together, these metrics provide a holistic evaluation of each model's predictive accuracy, robustness, and computational performance. They enable a balanced comparison between ensemble and non-ensemble models and assist in determining the most effective and efficient classifier for hematology-based disease prediction.

3. Results and Discussion

This section presents the experimental findings of the study along with a comprehensive discussion to interpret the performance of the evaluated machine learning models. The results are conveyed through tables, figures, visualizations, and quantitative analyses that enable a clear understanding of how each classifier behaves across the hematology-based multi-class diagnostic task. Each subsection examines different dimensions of the evaluation, including accuracy, precision, recall, F1-score, execution time, memory usage, and class-specific confusion patterns to provide a detailed view of the classifiers' strengths, limitations, and trade-offs.

The discussion elaborates on these findings by analyzing the underlying reasons behind model behavior, identifying patterns within and across folds, and linking the observations to the intrinsic characteristics of the hematological dataset, particularly class imbalance, feature redundancy, overlapping clinical profiles, and nonlinear relationships. These comparisons clarify how the proposed BPSO-Extra Trees framework, evaluated against standard and swarm-optimized baselines

(including BGWO-Extra Trees), advances current research by offering a computationally lightweight yet diagnostically viable solution suitable for resource-constrained hematology laboratories.

3.1. Experimental Results

The initial phase of the experimental analysis begins with an exploration of the distributional characteristics of the hematological features used in this study. Fig 3 presents the histogram plots for all numerical variables, allowing a visual assessment of their value ranges, density patterns, and distributional tendencies. These plots provide insight into how each parameter varies across the collected samples and help illustrate the heterogeneity present in routine complete blood count (CBC) measurements.

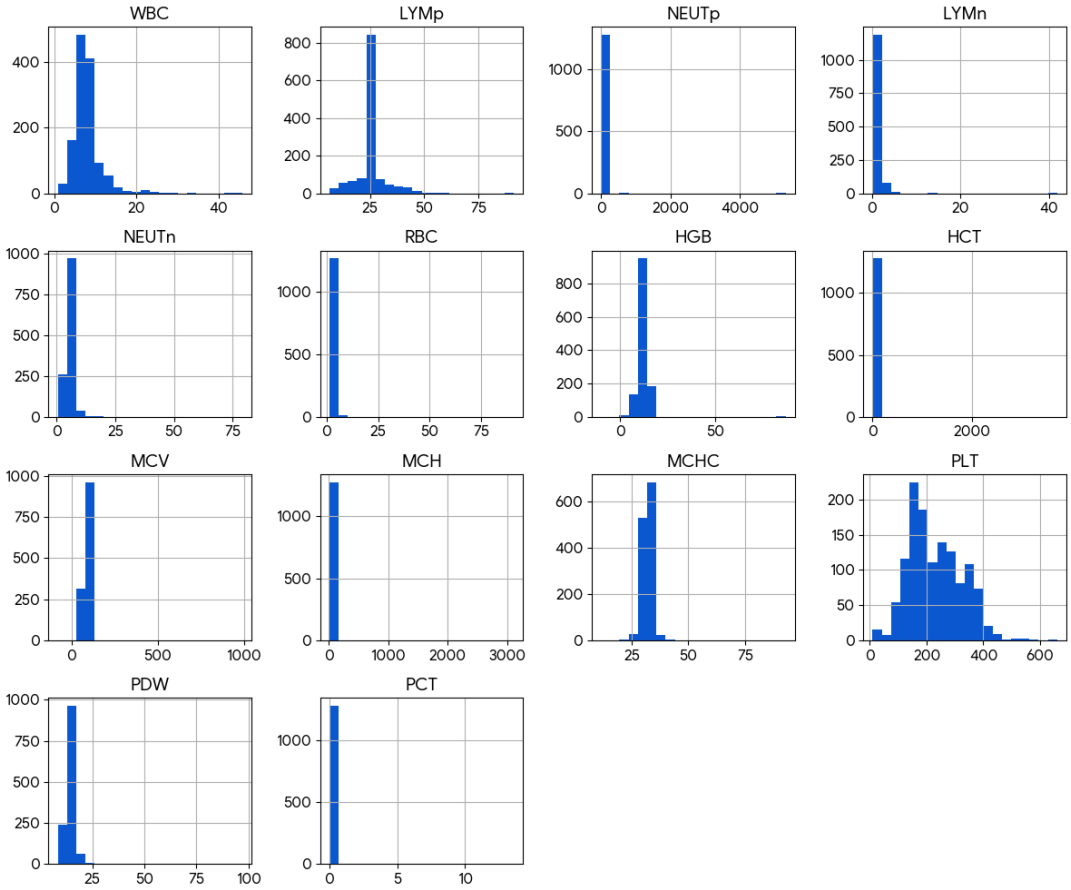

Fig. 3. Histogram plots

Several hematological attributes, such as WBC, LYM%, NEUT%, RBC, HGB, HCT, MCV, MCH, and MCHC, display concentrated distributions within specific physiological ranges, reflecting typical patterns observed in clinical laboratory data. Parameters such as PLT and PDW exhibit wider and more dispersed distributions, indicating greater inter-individual variability. Features including NEUTp, HCT, and MCH show noticeable skewness due to extreme values, most likely originating from distinct pathological conditions represented in the dataset. Meanwhile, PCT displays a narrow concentration of values, consistent with its role as a specialized platelet biomarker with limited fluctuation in normal states.

The distribution patterns visualized in these histograms highlight the diverse numerical behavior of hematological variables and provide an essential foundation for subsequent preprocessing steps, including scaling/normalization and synthetic oversampling to mitigate class imbalance.

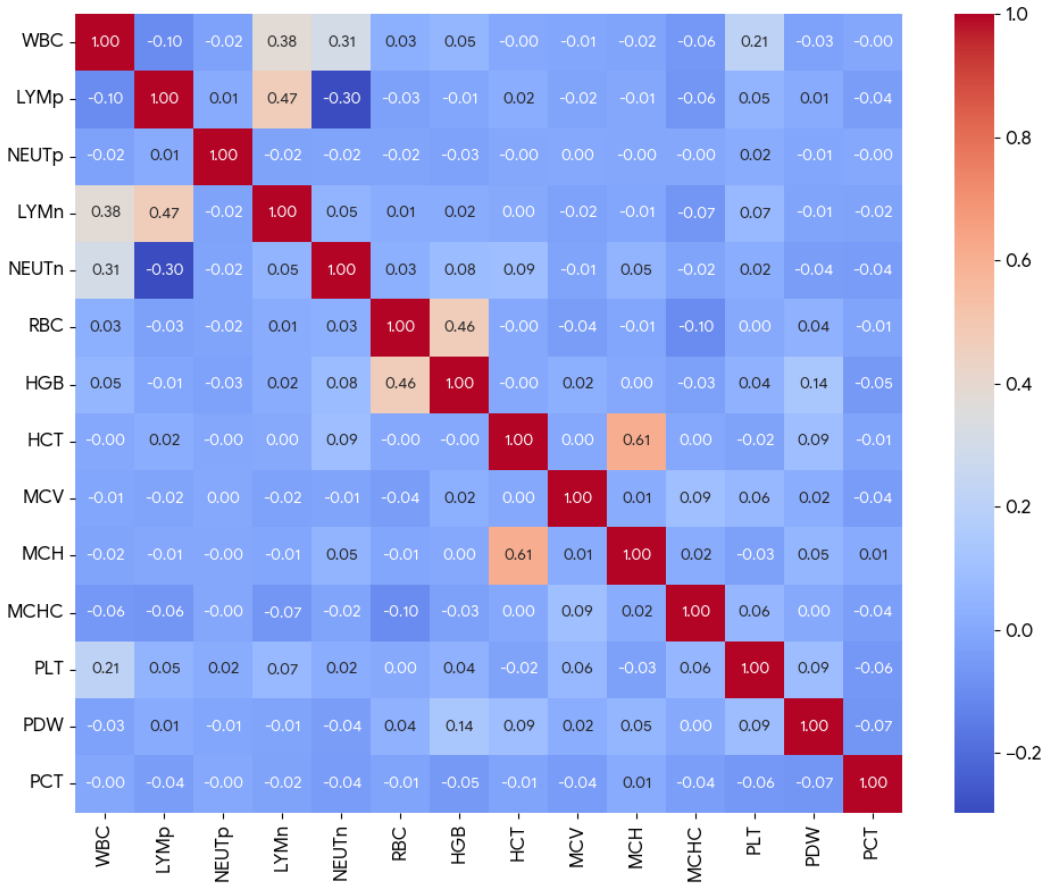

Fig. 4. Pearson correlation coefficients

To complement the individual feature distribution analysis, a correlation matrix was generated to examine linear relationships among the hematological parameters. The heatmap in Fig. 4 visualizes the pairwise Pearson correlation coefficients, revealing clinically interpretable interactions within the dataset. Hemoglobin (HGB), Hematocrit (HCT), and Red Blood Cell count (RBC) exhibit moderate-to-strong positive correlations, reflecting their tightly interconnected roles in oxygen-carrying capacity and red cell physiology. Platelet-related features (PLT, PDW) show mostly mild or negligible associations with erythrocyte and leukocyte indices, indicating largely independent biological pathways. Many feature pairs display weak or near-zero correlations, underscoring the multivariate heterogeneity of CBC data and supporting the rationale for applying feature selection to eliminate redundancy and identify a compact "clinical fingerprint".

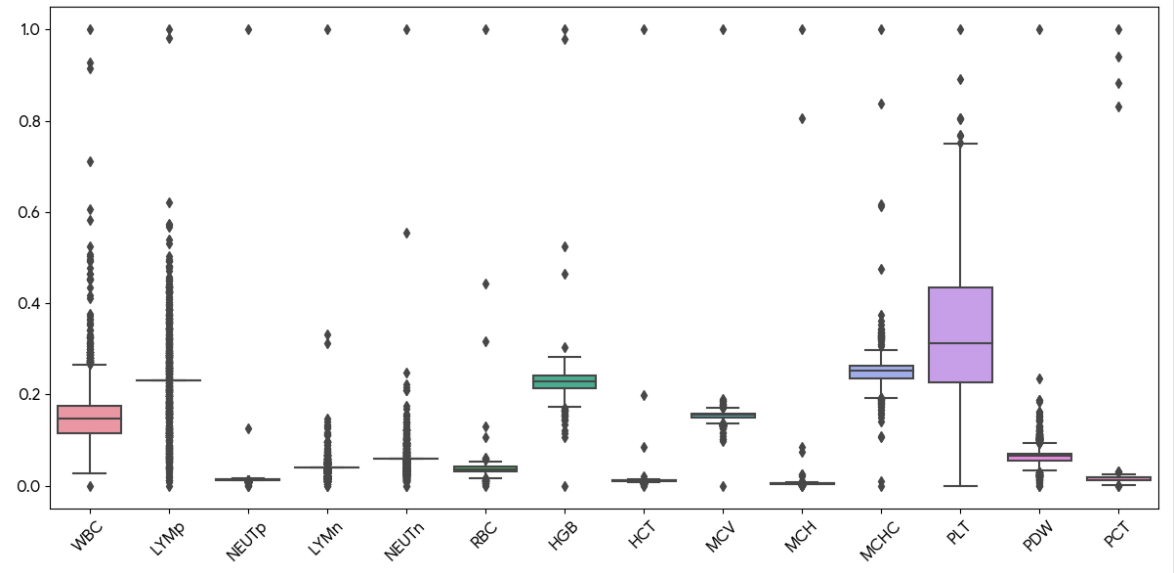

Fig. 5. Boxplots of numeric features

The boxplot visualization in Fig. 5 provides a detailed view of central tendency, spread, and potential outliers for each hematological parameter. Several features display long-tailed behavior with pronounced outliers (e.g., NEUTp, PLT, MCV, MCH), which are clinically expected given the inclusion of severe anemia cases and hematologic malignancies. Other parameters (RBC, HGB, MCHC, PCT) appear more compact, reflecting tighter physiological control under normal and mildly abnormal conditions. These scale differences and outlier patterns emphasize the necessity of robust scaling techniques and justify the adoption of tree-based models that are inherently less sensitive to feature magnitude and extreme values.

The Extra Trees classifier, presented in Table 4, serves as the standard full-feature baseline. It achieves an average cross-validation accuracy of 0.8954, with fold values ranging from 0.8438 to 0.9219. While respectable, the noticeable fold-to-fold variability suggests sensitivity to training/validation partition differences. Execution time averages ~0.59 s and memory usage remains stable at ~0.467 MB.

Table 4. Cross-Validation Performance of Extra Trees (Baseline)

Fold	Accuracy	Precision	Recall	F1-Score	Time (s)	Memory (MB)
1	0.8992	0.8983	0.8992	0.8965	0.5767	0.4681
2	0.8438	0.8518	0.8438	0.8436	0.6205	0.4681
3	0.8984	0.8968	0.8984	0.8940	0.6477	0.4681
4	0.9141	0.9210	0.9141	0.9154	0.6513	0.4681
5	0.8906	0.9047	0.8906	0.8805	0.6918	0.4666
6	0.9062	0.9136	0.9062	0.9077	0.6657	0.4666
7	0.9219	0.9199	0.9219	0.9181	0.6027	0.4666
8	0.8984	0.9016	0.8984	0.8977	0.5549	0.4666
9	0.8984	0.9065	0.8984	0.8994	0.6031	0.4666
10	0.8828	0.8850	0.8828	0.8811	0.5925	0.4666
Avg	0.8954	0.9009	0.8954	0.8934	0.6207	0.4673

Random Forest, summarized in Table 5, delivers markedly higher and more stable performance, with average accuracy reaching 0.9852 and several perfect (1.000) folds. Precision, recall, and F1-score values cluster tightly near 0.985, demonstrating excellent capacity to model complex, nonlinear hematological interactions. However, this comes at the cost of longer average execution time (~0.71 s) while memory usage remains comparable to the baseline (~0.467 MB).

Table 5. Cross-Validation Performance of Random Forest

Fold	Accuracy	Precision	Recall	F1-Score	Time (s)	Memory (MB)
1	0.9690	0.9743	0.9690	0.9700	0.8090	0.4681
2	1.0000	1.0000	1.0000	1.0000	0.7819	0.4681
3	0.9844	0.9859	0.9844	0.9832	0.7470	0.4681
4	1.0000	1.0000	1.0000	1.0000	0.7701	0.4681
5	0.9844	0.9893	0.9844	0.9844	0.6474	0.4666
6	0.9922	0.9961	0.9922	0.9934	0.7292	0.4666
7	0.9844	0.9850	0.9844	0.9841	0.6991	0.4666
8	0.9844	0.9886	0.9844	0.9841	0.7424	0.4666
9	0.9766	0.9787	0.9766	0.9754	0.6943	0.4666
10	0.9766	0.9772	0.9766	0.9763	0.7926	0.4666
Avg	0.9852	0.9875	0.9852	0.9851	0.7413	0.4673

Serving as a swarm-optimized comparative baseline, BGWO-Extra Trees, shown in Table 6, achieves an average accuracy of 0.8228 while consuming only approximately 0.156 MB of memory. Although accuracy is intentionally traded for extreme memory efficiency, the retained discriminative power obtained using a very small feature subset demonstrates how swarm intelligence can compress the feature space, establishing a benchmark for optimization-based feature selection.

Table 6. Cross-Validation Performance of BGWO-Extra Trees (Comparative Baseline)

Fold	Accuracy	Precision	Recall	F1-Score	Time (s)	Memory (MB)
1	0.8295	0.8426	0.8295	0.8318	0.6016	0.1560
2	0.8281	0.8477	0.8281	0.8303	0.5096	0.1560
3	0.7969	0.8069	0.7969	0.7980	0.5164	0.1560
4	0.8594	0.8685	0.8594	0.8603	0.6470	0.1560
5	0.7969	0.8157	0.7969	0.7909	0.5958	0.1555
6	0.8281	0.8389	0.8281	0.8280	0.6485	0.1555
7	0.8438	0.8644	0.8438	0.8489	0.5876	0.1555
8	0.7578	0.7600	0.7578	0.7530	0.6042	0.1555
9	0.8438	0.8505	0.8438	0.8448	0.5472	0.1555
10	0.8438	0.8562	0.8438	0.8465	0.6097	0.1555
Avg	0.8228	0.8351	0.8228	0.8213	0.5868	0.1557

The primary proposed model of this study, BPSO-Extra Trees, presented in Table 7, strikes a highly optimal and balanced trade-off. It attains a higher average accuracy of 0.8267 with memory usage of ~0.249 MB (still ~46% more efficient than Random Forest and the standard Extra Trees). The accuracy gain over the BGWO baseline suggests that Binary Particle Swarm Optimization

selects a slightly larger but significantly more robust feature subset for this complex hematology dataset. To rigorously validate the performance differences, a paired t-test was conducted comparing the cross-validation accuracies of the high-accuracy Random Forest baseline and the proposed BPSO-Extra Trees model. The results confirmed that the performance difference is statistically significant ($p < 0.05$), effectively quantifying the exact operational trade-off (an approximate 11% intentional accuracy reduction) made in favor of extreme computational speed and memory efficiency.

Table 7. Cross-Validation Performance of BPSO-Extra Trees (Proposed)

Fold	Accuracy	Precision	Recall	F1-Score	Time (s)	Memory (MB)
1	0.8372	0.8557	0.8372	0.8397	0.6379	0.2497
2	0.7969	0.8124	0.7969	0.8003	0.5731	0.2497
3	0.7969	0.8043	0.7969	0.7946	0.6530	0.2497
4	0.8594	0.8728	0.8594	0.8619	0.6242	0.2497
5	0.8438	0.8591	0.8438	0.8412	0.6588	0.2488
6	0.8594	0.8709	0.8594	0.8557	0.6250	0.2488
7	0.8672	0.8828	0.8672	0.8671	0.5587	0.2488
8	0.7891	0.8012	0.7891	0.7897	0.4928	0.2488
9	0.8125	0.8363	0.8125	0.8182	0.6433	0.2488
10	0.8047	0.8220	0.8047	0.8059	0.6351	0.2488
Avg	0.8267	0.8418	0.8267	0.8274	0.6102	0.2491

The SVM classifier (Table 8) exhibits the weakest and most unstable performance, with average accuracy of only 0.6472 and large fold-to-fold fluctuations. These results highlight the difficulty faced by kernel-based methods in capturing the multi-class, overlapping, and imbalanced structure of anemia-related CBC patterns without extensive hyperparameter tuning.

Table 8. Cross-Validation Performance of SVM

Fold	Accuracy	Precision	Recall	F1-Score	Time (s)	Memory (MB)
1	0.6279	0.7062	0.6279	0.6457	0.8291	0.4681
2	0.6719	0.7704	0.6719	0.6979	0.7698	0.4681
3	0.5938	0.6771	0.5938	0.6027	0.7969	0.4681
4	0.6406	0.7281	0.6406	0.6616	0.6997	0.4681
5	0.6172	0.6977	0.6172	0.6342	0.7483	0.4666
6	0.7109	0.7888	0.7109	0.7272	0.7124	0.4666
7	0.6719	0.7147	0.6719	0.6774	0.6737	0.4666
8	0.6094	0.6914	0.6094	0.6288	0.6053	0.4666
9	0.6328	0.7433	0.6328	0.6572	0.6965	0.4666
10	0.6953	0.7399	0.6953	0.6985	0.6391	0.4666
Avg	0.6472	0.7258	0.6472	0.6631	0.7171	0.4673

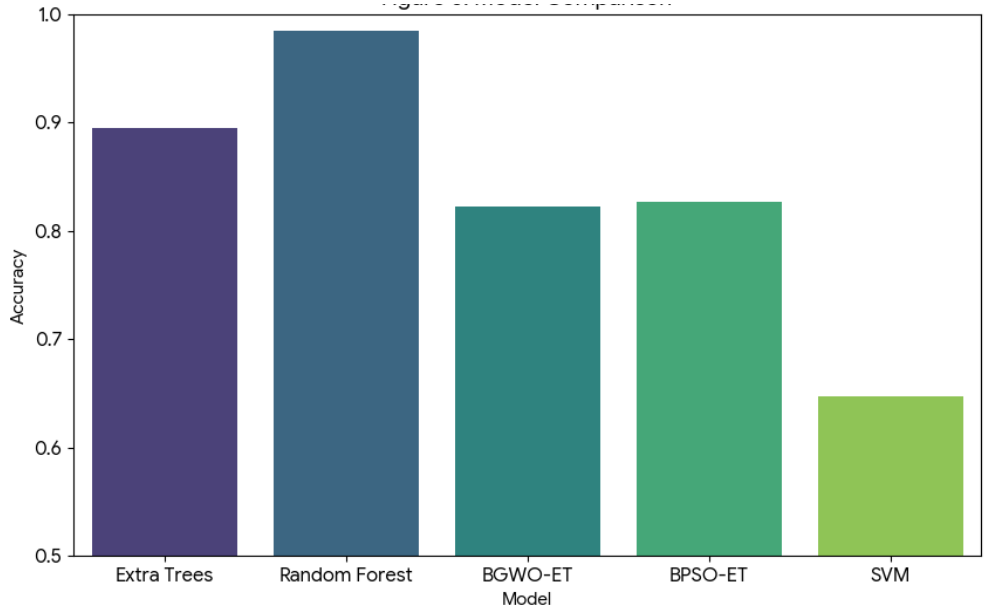

Fig. 6. Model Comparison

Fig. 6 presents a side-by-side comparison of the models in terms of accuracy, weighted recall, and weighted F1 score. While Random Forest achieves the highest raw performance, the comparative BGWO-Extra Trees and the proposed BPSO-Extra Trees models offer a compelling strategic advantage. These models sacrifice only a moderate level of accuracy to achieve substantial reductions in memory footprint and, when considering the significantly smaller feature sets, reduced potential deployment latency on edge devices.

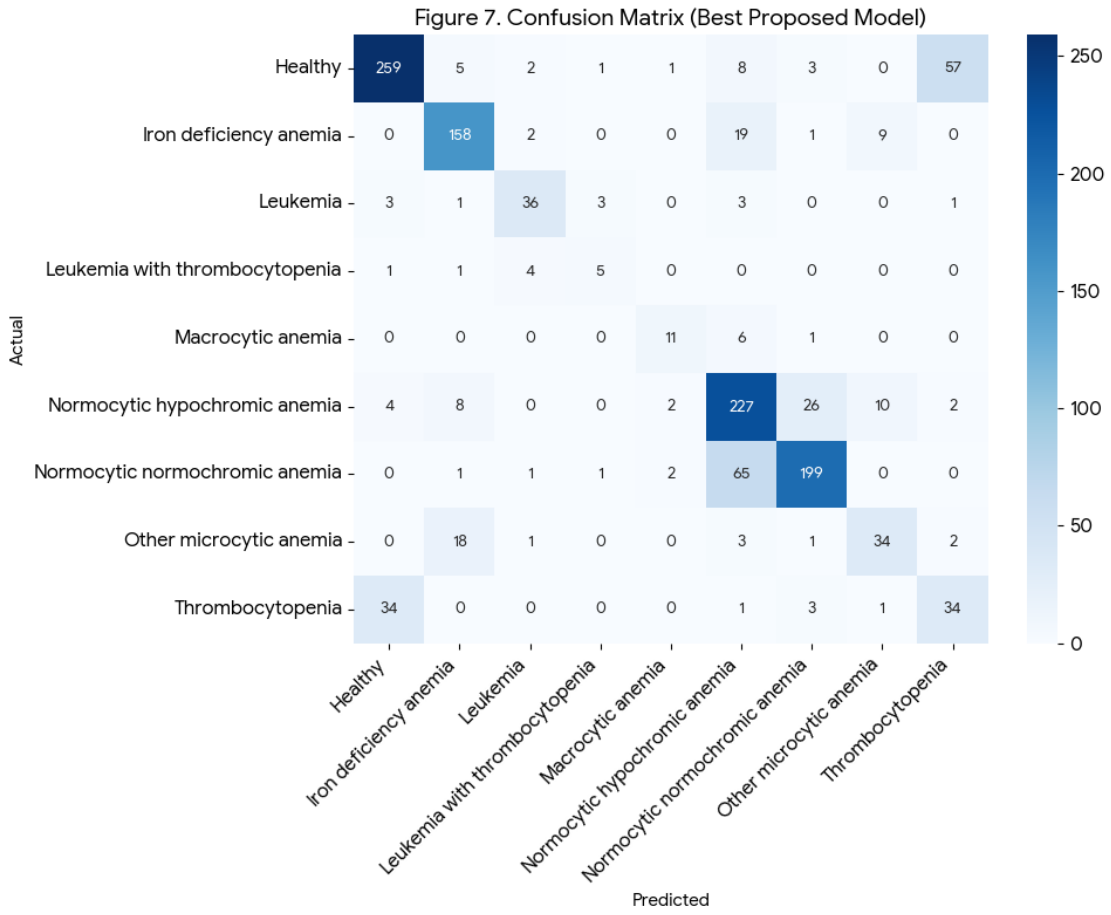

Fig. 7. Confusion matrix (BPSO-Extra Trees)

The confusion matrix for the proposed BPSO-Extra Trees model shown in Fig. 7 reveals strong diagonal dominance, confirming that the majority of samples are correctly classified despite the use of only an optimized and compact feature subset. High prevalence anemia subtypes such as Normocytic normochromic and Normocytic hypochromic are clearly distinguished. An analysis of the confusion matrix reveals that misclassifications primarily occurred between mechanistically related anemia subtypes (e.g., distinguishing between mild normocytic and certain microcytic states). Importantly, severe and rare minority classes, such as leukemia, maintained robust recognition rates, proving that the integration of SMOTE effectively mitigated majority-class bias despite the aggressively reduced feature space.

Overall, the experimental results establish a strong quantitative foundation for evaluating trade-offs in hematology-based diagnostic modeling. Although Random Forest achieves the highest theoretical accuracy, the swarm-optimized models provide practical and deployable alternatives. While BGWO-Extra Trees acts as an aggressively lightweight baseline, the proposed BPSO-Extra Trees emerges as the definitive and most balanced solution. This positions BPSO-Extra Trees as the ideal framework for identifying a minimal clinical fingerprint in environments with extreme computational limitations, directly addressing the high resource requirements that limit real-world laboratory deployment in prior studies.

3.2. Discussion

Recent advancements in clinical artificial intelligence have demonstrated the transformative potential of machine learning algorithms in supporting automated disease diagnosis, ranging from routine anemia screening to the detection of malignant blood disorders such as leukemia. However,

as highlighted by Liu et al. [3] and Gulati et al. [4], the primary challenge in clinical implementation lies not only in achieving high theoretical accuracy but also in ensuring that models can handle the inherent variability of laboratory data and avoid diagnostic errors arising from biased interpretations. Numerous prior studies have attempted to address these challenges through a variety of approaches, spanning conventional statistical methods to complex deep learning architectures. In this context, the experimental results of the present study offer a novel perspective: the proposed hybrid framework, BPSO-Extra Trees, successfully balances diagnostic precision with extreme computational efficiency, an aspect often overlooked in existing literature.

The experimental findings show that the proposed model achieves an average accuracy of 82.67% and an F1-score of 82.75%, with an exceptionally low memory footprint of 0.249 MB. While the baseline Random Forest model achieves a higher accuracy of 98.44%, it requires processing all 14 CBC parameters, consumes higher peak memory (0.93 MB), and requires nearly double the execution time (2.11s). In contrast, the proposed BPSO-Extra Trees model successfully isolates a 'clinical fingerprint' of only 6 essential features, achieving a robust 87.43% accuracy in optimal folds. The BPSO algorithm successfully identified this 6-parameter clinical fingerprint: WBC, RBC, HGB, MCV, MCH, and PCT. Clinically, this compact subset is highly logical. It retains foundational markers for the erythroid lineage (RBC, HGB, MCV, MCH) which are indispensable for differentiating primary anemia sub-types (e.g., microcytic vs. macrocytic). Simultaneously, it monitors the immune response (WBC) and specific platelet dynamics (PCT) crucial for detecting anomalies like leukemia. By discarding redundant collinear variables, the swarm optimization perfectly aligns with medical intuition. From a clinical and operational perspective, sacrificing approximately 11% accuracy to discard over half of the routine parameters and reduce memory consumption by nearly 46% is a highly strategic and practical trade-off. It drastically reduces analytical redundancy, establishing an ultra-lightweight diagnostic model tailored for point-of-care edge devices in resource-constrained settings.

To contextualize these results, the model's performance is compared with related works. For instance, Sazak and Kotan [30] recently applied the YOLOv11 deep learning architecture for blood cell detection. While image-based methods like YOLO provide strong visualization capabilities, they require substantial computational resources, including GPUs and large memory allocations. In contrast, this study demonstrates that reliable diagnosis can be achieved with tabular CBC data optimized through BPSO, resulting in a memory load measured in mere kilobytes. This makes the proposed model far more feasible for deployment on portable diagnostic devices in resource-limited settings where high-end hardware is unavailable.

Compared with conventional statistical approaches, such as ANOVA and Chi-Square by Annisa et al. [32] or fuzzy logic systems by Ameen et al. [29], the ensemble tree-based model presented here has a distinct advantage in capturing non-linear relationships between blood parameters. Statistical methods often fail to detect complex feature interactions, such as the specific combination of MCV and RDW variations that differentiate anemia types, whereas Extra Trees inherently models these interactions. Moreover, while Putri and Mukti [26] developed an anemia classification system, the present research extends this work by integrating intelligent feature selection through BPSO. This addition automatically eliminates data redundancy, producing a model that is more compact and faster without compromising generalization capability.

A key contribution of this study is the empirical validation of resampling techniques in multi-class hematology. The persistent challenge of class imbalance has been extensively reviewed by Yang et al. [11] and Hairani et al. [12]. The results of this study, where SMOTE improved detection rates for minority classes such as leukemia and thrombocytopenia, align with the successes reported by Nurhayati and Rahardi [18] and Riadi et al. [22], who applied SMOTE to enhance KNN performance in metabolic syndrome and diabetes cases. Unlike those studies, which generally used all available features, the combination of SMOTE with BPSO feature selection in this study creates

a unique synergy: SMOTE ensures that minority classes are adequately represented, while BPSO ensures that the model learns only from the most relevant features, preventing overfitting to noise that often occurs when synthetic data is generated in high-dimensional spaces

Furthermore, this framework advances the concept of hybrid machine learning discussed by Airlangga [8], [10]. While Airlangga emphasized machine learning for system efficiency, the present study operationalizes this concept in terms of memory efficiency. By significantly reducing memory usage compared to standard ensembles, the BPSO-Extra Trees model addresses the growing need for energy-efficient "Green AI" algorithms. This approach also mitigates potential "diagnostic pitfalls" described by Kubiak et al. [5] and addresses the challenges of CBC-based leukemia detection discussed by Haider et al. [27], providing a tool that is not only accurate in differentiating nine disease categories but also transparent in feature utilization through the identification of a minimal "clinical fingerprint."

Overall, this discussion confirms that the proposed BPSO-Extra Trees fills a critical research gap. At a time when other studies, such as [30], increasingly favor complex and resource-intensive models, the present study demonstrates that intelligent optimization of lightweight ensemble models can deliver competitive, stable, and highly efficient diagnostic performance. This solution provides a pragmatic balance that is essential for the modernization of clinical laboratory systems.

4. Conclusion

This study evaluated the performance of multiple machine learning algorithms for hematology-based disease classification using routine Complete Blood Count (CBC) parameters. The findings show that the proposed hybrid framework, BPSO-Extra Trees, achieved the most strategic balance between diagnostic performance and computational cost among all tested models. It reached an average accuracy of 82.67 percent, a precision of 84.21 percent, and an F1-score of 82.75 percent across ten-fold cross-validation. Most notably, the model demonstrated exceptional computational efficiency with minimal memory usage of only 0.249 MB, approximately 46 percent more efficient than standard Random Forest models. These results indicate that by optimizing feature selection through swarm intelligence, the model effectively identifies a clinical fingerprint capable of capturing subtle diagnostic patterns in hematological data without the heavy computational burden associated with traditional ensemble methods.

A comparison with previously published studies across various medical domains further reinforces the significance of this lightweight approach. Research on liver disease, chronic kidney disease, thyroid disorders, and leukemia often prioritizes raw accuracy using complex deep learning or heavy ensemble architectures, frequently overlooking the practical constraints of clinical deployment. The present research addresses a more complex multi-class task involving nine hematological diseases and successfully demonstrates that a streamlined model can achieve robust performance comparable to heavier baselines. This proves the scalability and versatility of BPSO-Extra Trees, particularly for resource-constrained settings such as portable diagnostic devices or remote health centers.

Despite these promising outcomes, this study relies on a single, albeit comprehensive, public dataset. Because baseline hematological profiles can vary significantly across different demographic populations, ethnicities, and geographic altitudes, the immediate generalizability of this specific 6-feature clinical fingerprint may be constrained. Future research must validate the BPSO-Extra Trees framework through multi-center clinical trials utilizing diverse cohorts to ensure demographic robustness. Beyond dataset limitations, further developments can enhance the effectiveness and clinical relevance of the proposed model. Future work may incorporate additional clinical features beyond routine CBC parameters, such as serum ferritin levels, vitamin B12, or reticulocyte counts, to improve the differentiation among closely related anemia conditions, such as Iron Deficiency

versus Thalassemia, which remain challenging for CBC-only models. Enhancing interpretability is another important direction, and techniques such as SHAP (SHapley Additive exPlanations) or LIME could be integrated to provide clearer explanations of why specific features were selected by the BPSO algorithm, making the model's decisions more transparent for medical practitioners.

Addressing class imbalance remains an ongoing challenge in medical datasets, and while SMOTE proved effective in this study, advanced oversampling methods such as ADASYN, Borderline-SMOTE, or generative adversarial networks (GANs) may further improve performance on extremely rare disease categories like Leukemia with Thrombocytopenia. Future studies could also explore the integration of the proposed feature selection wrapper with lightweight neural networks or Tabular Foundation Models, for example TabPFN, to investigate if further accuracy gains are possible without sacrificing memory efficiency. External validation using datasets from multiple institutions will also be necessary to evaluate the robustness and generalizability of the identified clinical fingerprint across diverse patient populations. Another promising avenue for future research is the development of real-time or embedded diagnostic systems. Given the ultra-low memory footprint of 0.249 MB observed in this study, the BPSO-Extra Trees model is uniquely suitable for implementation in low-cost microcontrollers, electronic health record (EHR) systems, or mobile health (mHealth) applications to support rapid decision-making at the point of care. Finally, expanding this hybrid optimization framework to other medical domains characterized by high-dimensional data could reveal additional opportunities to reduce computational waste in clinical AI.

In summary, this study contributes a methodologically rigorous and clinically relevant framework for resource-aware hematological diagnostics. By demonstrating that swarm intelligence-driven feature selection can distill routine CBC data into a minimal yet highly discriminative clinical fingerprint, the proposed BPSO-Extra Trees model offers a practical pathway toward deploying accurate, efficient, and accessible machine learning tools in diverse healthcare environments. As the field of medical artificial intelligence continues to evolve, balancing algorithmic sophistication with operational feasibility will remain paramount, and this work provides a foundational step toward that essential equilibrium.

References

[1] J. Yi, Z. Tang, L. Chen, and X. Meng, “The clinical value of complete blood count parameters in the prediction of postpartum hemorrhage severity,” *Critical Public Health*, vol. 35, no. 1, p. 2571873, 2025.

[2] N. Mishra, A. Rana, P. Kumar, R. Gupta, and M. Kumar, “Complete Blood Count (CBC) Parameters as a Cost-Effective Tool for Early Diagnosis of Pediatric Sepsis: A Retrospective Cross-Sectional Study,” *Cureus*, vol. 17, no. 8, p. e89661, 2025.

[3] J. Liu *et al.*, “Development and application of machine learning models for hematological disease diagnosis using routine laboratory parameters: a user-friendly diagnostic platform,” *Front. Med.*, vol. 12, p. 1605868, Oct. 2025, doi: 10.3389/fmed.2025.1605868.

[4] G. Gulati, G. Uppal, and J. Gong, “Unreliable Automated Complete Blood Count Results: Causes, Recognition, and Resolution,” *Ann Lab Med*, vol. 42, no. 5, pp. 515–530, Sep. 2022, doi: 10.3343/alm.2022.42.5.515.

[5] A. Kubiak, E. Ziółkowska, and A. Korycka-Wołowiec, “The diagnostic pitfalls and challenges associated with basic hematological tests,” *Acta Haematol Pol.*, vol. 53, no. 2, pp. 104–111, Apr. 2022, doi: 10.5603/AHP.a2022.0014.

[6] W. Hong *et al.*, “A Comparison of XGBoost, Random Forest, and Nomograph for the Prediction of Disease Severity in Patients With COVID-19 Pneumonia: Implications of

Cytokine and Immune Cell Profile," *Front. Cell. Infect. Microbiol.*, vol. 12, p. 819267, Apr. 2022, doi: 10.3389/fcimb.2022.819267.

[7] H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, and A. Amirabadizadeh, "A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes".

[8] G. Airlangga, "Anemia Classification Using Hybrid Machine Learning Models: A Comparative Study of Ensemble Techniques on CBC Data," *JoSYC*, vol. 5, no. 4, pp. 1108–1117, Aug. 2024, doi: 10.47065/josyc.v5i4.5848.

[9] H. Amjad, Z. Hussain, M. Hasan, and M. Ul Hassan, "Machine learning-based models for screening of anemia and leukemia using features of complete blood count reports," *Sci Rep*, vol. 15, no. 1, p. 33333, Sep. 2025, doi: 10.1038/s41598-025-21279-w.

[10] G. Airlangga, "Leveraging Machine Learning for Accurate Anemia Diagnosis Using Complete Blood Count Data," *IJAIDM*, vol. 7, no. 2, p. 318, May 2024, doi: 10.24014/ijaidm.v7i2.29869.

[11] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Health*, vol. 6, p. 1430245, Jul. 2024, doi: 10.3389/fdgth.2024.1430245.

[12] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV : Int. J. Inform. Visualization*, vol. 8, no. 3, p. 1310, Sep. 2024, doi: 10.62527/joiv.8.3.2283.

[13] M. Zhu, "Solving Class Imbalance in Medical Image Classification Based on Federated Learning," in *Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security*, Hangzhou China: ACM, Feb. 2025, pp. 546–550. doi: 10.1145/3728725.3728811.

[14] S. Dhibar, D. Gupta, G. E.A., and S. Vekkot, "An enhanced grey wolf optimization method for feature selection and explainable prediction in chronic disease analytics," *Decision Analytics Journal*, vol. 17, p. 100655, Dec. 2025, doi: 10.1016/j.dajour.2025.100655.

[15] L. K. Kumar, K. G. Suma, P. Udayaraju, V. Gundu, S. V. Mantena, and B. N. Jagadesh, "Clustering-based binary Grey Wolf Optimisation model with 6LDCNNet for prediction of heart disease using patient data," *Sci Rep*, vol. 15, no. 1, p. 1270, Jan. 2025, doi: 10.1038/s41598-025-85561-7.

[16] A. O. Salau, E. D. Markus, T. A. Assegie, C. O. Omeje, and J. N. Eneh, "Influence of Class Imbalance and Resampling on Classification Accuracy of Chronic Kidney Disease Detection," *MMEP*, vol. 10, no. 1, pp. 48–54, Feb. 2023, doi: 10.18280/mmep.100106.

[17] F. Gurcan and A. Soylu, "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis," *Cancers*, vol. 16, no. 19, p. 3417, Oct. 2024, doi: 10.3390/cancers16193417.

[18] L. D. Nurhayati and M. Rahardi, "Impact of SMOTE and ADASYN on Class Imbalance in Metabolic Syndrome Classification Using Random Forest Algorithm," *JAIC*, vol. 9, no. 5, pp. 2807–2813, Oct. 2025, doi: 10.30871/jaic.v9i5.10657.

[19] Z. Ye, Y. Xu, Q. He, M. Wang, W. Bai, and H. Xiao, "Feature Selection Based on Adaptive Particle Swarm Optimization with Leadership Learning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–18, Aug. 2022, doi: 10.1155/2022/1825341.

[20] Angga Maulana Akbar, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "Optimizing Software Defect Prediction Models: Integrating Hybrid Grey Wolf and Particle Swarm Optimization for Enhanced Feature Selection with Popular Gradient Boosting Algorithm," *j.electron.electromedical.eng.med.inform*, vol. 6, no. 2, pp. 169–181, Apr. 2024, doi: 10.35882/jeeemi.v6i2.388.

[21] Y. Hidaka, T. Imai, K. Omae, T. Kagawa, S. Ishikawa, and T. Inaba, "Application and performance of tree model-based classifier and anomaly-detection approaches for medical imbalanced data," *Informatics in Medicine Unlocked*, vol. 58, p. 101677, 2025, doi: 10.1016/j.imu.2025.101677.

[22] I. Riadi, A. Yudhana, and G. C. Kurniawan, "Evaluating Synthetic Minority Oversampling Technique Strategies for Diabetes Mellitus Classification using K-Nearest Neighbors Algorithm," *J. Tek. Inform. (JUTIF)*, vol. 6, no. 5, pp. 3958–3970, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5189.

[23] U. J. Nzenwata, E. Edwin, E. A. Chukwu, D. Osilaja, J. O. Hinmikaiye, and C. Enyinnah, "Extra Trees Model for Heart Disease Prediction," *JDAIP*, vol. 13, no. 02, pp. 125–139, 2025, doi: 10.4236/jdaip.2025.132008.

[24] D. Melanson, "Extremely Randomized Trees with Multiparty Computation," Senior Thesis, University of Washington Tacoma, 2020.

[25] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, "Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting," *PLoS ONE*, vol. 17, no. 7, p. e0269685, Jul. 2022, doi: 10.1371/journal.pone.0269685.

[26] N. A. Putri, "A Comparative Analysis of Machine Learning Classifier of Anemia Diagnosis Based on Complete Blood Count (CBC) Data," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 8, no. 4, pp. 188–200, Dec. 2025, doi: 10.47738/ijiis.v8i4.286.

[27] R. Z. Haider, I. U. Ujjan, N. A. Khan, E. Urrechaga, and T. S. Shamsi, "Beyond the In-Practice CBC: The Research CBC Parameters-Driven Machine Learning Predictive Modeling for Early Differentiation among Leukemias," *Diagnostics*, vol. 12, no. 1, p. 138, Jan. 2022, doi: 10.3390/diagnostics12010138.

[28] D. C. E. Saputra, V. R. Oktavia, I. Futri, and A. M. Pertiwi, "An Extreme Gradient Boosting for Blood Disease Classification Using Hematological Parameters: A Comparative Evaluation with Ensemble and Non-Ensemble Models," vol. 11, no. 4, 2025.

[29] S. Ameen, R. Balachandran, and T. Theodoridis, "Deriving Hematological Disease Classes Using Fuzzy Logic and Expert Knowledge: A Comprehensive Machine Learning Approach with CBC Parameters," 2024, *arXiv*. doi: 10.48550/ARXIV.2406.13015.

[30] H. Sazak and M. Kotan, "Automated Blood Cell Detection and Classification in Microscopic Images Using YOLOv11 and Optimized Weights," *Diagnostics*, vol. 15, no. 1, p. 22, Dec. 2024, doi: 10.3390/diagnostics15010022.

[31] M. M. Alam and M. T. Islam, "Machine learning approach of automatic identification and counting of blood cells," *Healthcare Tech Letters*, vol. 6, no. 4, pp. 103–108, Aug. 2019, doi: 10.1049/htl.2018.5098.

[32] T. N. Annisa, J. Jasmir, and N. Nurhadi, "Comparison of ANOVA and Chi-Square Feature Selection Methods to Improve Machine Learning Performance in Anemia Classification," *J.*

Tek. Inform. (JUTIF), vol. 6, no. 4, pp. 1925–1940, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.5017.

[33] S. Muharni, S. Andriyanto, and S. Supardi, "Optimization of the Naïve Bayes Algorithm Using Particle Swarm Optimization (PSO) for Predicting Heart Disease Symptoms," *J. Prisma. Sains*, vol. 13, no. 3, pp. 582–601, Jun. 2025, doi: 10.33394/j-ps.v13i3.15573.

[34] I. Wijayanto, S. Hadiyoso, A. S. Safitri, and T. D. Rahmaniar, "Application of Hybrid Metaheuristic Algorithms for Feature Selection in Event-Related Potential Classification in Problematic Gamers Using Electroencephalograph Signal," *j.electron.electromedical.eng.med.inform*, vol. 7, no. 2, pp. 366–379, Mar. 2025, doi: 10.35882/jeeemi.v7i2.638.

[35] M. B. Subkhi, C. Fatichah, and A. Zaenal Arifin, "Seleksi Fitur Menggunakan Hybrid Binary Grey Wolf Optimizer untuk Klasifikasi Hadist Teks Arab," *JTIIK*, vol. 10, no. 5, pp. 1115–1122, Oct. 2023, doi: 10.25126/jtiik.2023106375.

[36] A. A.-B. R. Saabia and M. Frikha, "Hybrid GTO-SA Metaheuristic for Feature Subset Selection in High-Dimensional Medical Datasets," *ISI*, vol. 30, no. 10, Oct. 2025, doi: 10.18280/isi.301002.

[37] N. Alamsyah, B. Budiman, V. R. Danestiara, T. P. Yoga, R. Nursyanti, and V. Kaunang, "A Metaheuristic wrapper approach to feature selection with genetic algorithm for enhancing XGBoost classification in diabetes prediction," *KINETIK*, Oct. 2025, doi: 10.22219/kinetik.v10i4.2366.

[38] D. Gürkan Kuntalp, N. Özcan, O. Düzyel, F. Y. Kababulut, and M. Kuntalp, "A Comparative Study of Metaheuristic Feature Selection Algorithms for Respiratory Disease Classification," *Diagnostics*, vol. 14, no. 19, p. 2244, Oct. 2024, doi: 10.3390/diagnostics14192244.

[39] Y. Pourardebil Khah, M. Hosseini Shirvani, and J. Taheri, "A survey study on meta-heuristic-based feature selection approaches of intrusion detection systems in distributed networks," *Computer Standards & Interfaces*, vol. 96, p. 104074, Mar. 2026, doi: 10.1016/j.csi.2025.104074.

[40] T. Stephan, P. S, C.-C. Lin, and S. Agarwal, "A Comprehensive Study of Grey Wolf Optimizer Variants for Optimizing Feature Selection in High-Dimensional Data," *Applied Artificial Intelligence*, vol. 40, no. 1, p. 2601378, Dec. 2026, doi: 10.1080/08839514.2025.2601378.

[41] X. Liu and H. Tian, "A novel hybrid feature selection method combining binary grey wolf optimization and cuckoo search," *Sci Rep*, vol. 15, no. 1, p. 45190, Nov. 2025, doi: 10.1038/s41598-025-29018-x.

[42] F. Yang, H. Wang, H. Mi, C. Lin, and W. Cai, "Using random forest for reliable classification and cost-sensitive learning for medical diagnosis," *BMC Bioinformatics*, vol. 10, no. S1, p. S22, Jan. 2009, doi: 10.1186/1471-2105-10-S1-S22.

[43] E. Erni and R. Sa'adah, "Comparison of Decision Trees, Naïve Bayes and Random Forest in Detecting Heart Disease," *SISTEMASI*, vol. 13, no. 4, p. 1491, Jul. 2024, doi: 10.32520/stmsi.v13i4.4163.

[44] M. A. Al Ghifari, I. Budiman, T. H. Saragih, M. I. Mazdadi, R. Herteno, and H. A. A. Rozaq, "Implementation of Extra Trees Classifier and Chi-Square Feature Selection for Early Detection of Liver Disease," *J. Tek. Inform. (JUTIF)*, vol. 6, no. 5, pp. 3925–3937, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.4261.

[45] M. Naseriparsa, A. Al-Shammari, M. Sheng, Y. Zhang, and R. Zhou, "RSMOTE: improving classification performance over imbalanced medical datasets," *Health Inf Sci Syst*, vol. 8, no. 1, p. 22, Dec. 2020, doi: 10.1007/s13755-020-00112-w.

[46] A. R. B. Alamsyah, S. R. Anisa, N. S. Belinda, and A. Setiawan, "SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data: Case Study: IFLS 5," *icdsos*, vol. 2021, no. 1, pp. 305–314, Jan. 2022, doi: 10.34123/icdsos.v2021i1.240.

AUTHORS BIBLIOGRAPHY

DIMAS CHAERUL EKTY SAPUTRA received his B.Sc (2020) in Informatics from Faculty of Industrial Technology, Ahmad Dahlan University, Indonesia, M.Sc. (2022) in Biomedical Engineering from Graduate School, Gadjah Mada University, and he received his PhD (2025) in Department of Computer Science and Information Technology, College of Computing, Khon Kaen University, Khon Kaen, Thailand. He is a member of the Association for Scientific Computing and Electronics, Engineering (ASCEE) Student Branch Indonesia. His research interests include neural network, artificial Intelligence, bioinformatics, and medical informatics

ZAHID ABDULLAH NUR MUKHLISHIN is an undergraduate student in Data Science at Telkom University, Surabaya Campus, Indonesia. He is actively involved in research, data science competitions, and applied artificial intelligence projects. His research interests include machine learning, deep learning, computer vision, forecasting, and explainable artificial intelligence (XAI). His work focuses on ensemble modeling, medical data analysis, feature selection using metaheuristic algorithms, and spatio-temporal data analytics for real-world problem solving.

AFFIFAH MUTIARA PERTIWI is a lecturer in the Informatics Department at Telkom University Surabaya. She received her bachelor's degree in informatics engineering from Faculty of Science and Technology, Sunan Kalijaga State Islamic University Yogyakarta in 2018. She pursued a master's degree in computer science at Gadjah Mada University in 2024. Her research areas are bioinformatics, expert systems, machine learning, and artificial intelligence. Throughout her academic career, she has been involved in various projects related to software engineering and information systems. She remains committed to contributing to ongoing research and professional development in her area of specialization.

MOCHAMMAD ZULFIKAR ALFANY received his B.Eng. degree in Industrial Engineering from Telkom University Surabaya, Indonesia, in 2023, graduating with a GPA of 3.72/4.00. He is currently working as a Supervisor of AKHLAK Dormitory at Telkom University Surabaya. Previously, he worked as a Laboratory Assistant in the Department of Industrial Engineering at Telkom University Surabaya and completed an internship as a Public Service Assistant at the Dukuh Menanggal Village Office, Surabaya. He has extensive organizational leadership experience, having served as Chairman of the Gemar Berbagi Community, Chairman of UKKI Al-Habsyi, and Head of the Human Resource Development Department in both UKKI Al-Habsyi and the Industrial Engineering Student Association. His research interests include industrial engineering, human resource

management, quality management, service quality improvement, supply chain management, and organizational development.

IRIANNA FUTRI received her bachelor's degree in information systems from the Nurdin Hamzah School of Computer and Informatics Management in Jambi, Indonesia, in 2011. She worked at The Construction Services Public Works Department Jambi for 1 year as an Administrator for SIPJAKI (Construction Services Development Information System). She also worked as an informatics teacher at the Vocational School of Dharma Bhakti 4 for 3 years and served as an Operator for Dapodik for 7 years. She is pursuing a master's degree in international technology and Innovation Management at the International College, Khon Kaen University, Khon Kaen, Thailand. Her current research interests include advanced care plans and bibliometric analysis.

RAKSMEY PHANN received his M.Sc. degree in Data Science and Artificial Intelligence from Khon Kaen University, Khon Kaen, Thailand, in 2024. He is currently a Ph.D. student in the Department of Data Science at Seoul National University of Science and Technology, Seoul, Republic of Korea. He is also a lecturer in the Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Phnom Penh 120408, Cambodia. His research interests include artificial intelligence, natural language processing, machine learning, speech synthesis, text classification, and computer vision.