

Enhanced Prediction of Blood Quality Using Ensemble RF

AVIJIT KUMAR CHAUDHURI

c.avijit@gmail.com

Brainware University <https://orcid.org/0000-0002-5310-3180>

Research Article

Keywords: Ensemble RF(ERF), Haematological Parameters, Machine Learning (ML), Complete blood Count (CBC)

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10918180/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Enhanced Prediction of Blood Quality Using Ensemble RF

Pranab Gharai¹, Avijit Kumar Chaudhuri², Daizy Deb³, Anushree Jana⁴

¹Research Scholar, Computer Science & Engineering, Brainware University, Barasat,

Kolkata – 125, West Bengal, India, Email -pranab.g10@gmail.com

²Professor & HOD, Computer Science & Engineering, Brainware University, Barasat,

Kolkata – 125, West Bengal, India, Email - c.avijit@gmail.com

³Associate Professor, Computer Science & Engineering, Brainware University, Barasat,

Kolkata – 125, West Bengal, India Email - daizydebkar@gmail.com

⁴Student, MTECH CSE, Brainware University, Barasat, Kolkata – 125, West Bengal, India,

Email - anujana1212@gmail.com

Abstract

Complete Blood Count (CBC) parameters are important for early detection of different blood diseases such as anemia, infections, inflammatory disorders, and other hematological abnormalities. The volume of clinical lab information continues to grow and become more complex, making manual interpretation difficult, time-consuming, and prone to human error. This study presents an automated hematological classification system using machine Learning (ML) techniques to improve diagnostic accuracy and support clinical decision-making. The analysis uses a clinically validated hematological dataset of a series of 7,196 patient records with various CBC parameters such as Hemoglobin (Hb), Red Blood Cell count (RBC), White Blood Cell count (WBC), Hematocrit (HCT), Platelet count, Mean Corpuscular Hb (MCH), and other blood indices. A set of eight ML classifiers—Logistic Regression(LR), Naïve Bayes (NB), Support Vector Machine(SVM), Decision Tree(DT), Random Forest(RF), Multilayer Perceptron (MLP), Extra Trees(ET), Enhanced Randomized Forest (ERF)—were tested with the help of stratified 10-fold cross-validation. To guarantee the statistical reliability of model performance, accuracy, precision, sensitivity (recall), specificity, F1-score, Area under the receiver operating characteristic curve(AUC-ROC), Cohen's kappa coefficient, and 95% confidence intervals (CIs) were used. The results show that ensemble tree-based algorithms outperform classical ML algorithms. The evaluated classifiers performed excellently, were robust, and generalized well; the ERF classifier achieved the n accuracy, sensitivity, F1-score,

AUC-ROC, and Cohen's kappa of on all the evaluated classifiers. RF and DT were also effective, while LR and NB were less effective because they cannot handle nonlinear relationships and class imbalance. Moreover, an Explainable Artificial Intelligence (XAI) method like Shapley Additive exPlanations(SHAP) and Local Interpretable Model-agnostic Explanations(LIME) was used to explain model predictions and to identify clinically relevant hematological characteristics. The interpretability analysis showed that parameters related to white blood cells, Hb, red blood cells, HCT, MCH, and platelets were most important for classification decision-making, helping make the model transparent and trustworthy. In conclusion, the proposed ERF-based approach offers an accurate and reliable method for automated hematological data classification, with potential applications in intelligent clinical decision support systems for early disease detection and improved healthcare outcomes.

Keywords: Ensemble RF(ERF), Haematological Parameters, Machine Learning (ML), Complete blood Count (CBC)

Introduction

Blood is one of the most important biological fluids in the human body, which is crucial for maintaining normal physiological functions. It transports oxygen, nutrients, hormones, enzymes, and metabolic wastes; provides immune defense; regulates body temperature; and maintains acid-base balance. Blood flows throughout the body, and changes in its composition may offer important clues to a person's health. As many as 12 different diseases are linked to abnormalities in blood parameters, such as anemia, infection, inflammatory disorders, leukemia, thrombocytopenia, bone marrow diseases, and others. Therefore, hematology investigations have gained increasing significance in modern health care as an important diagnostic tool. One of the most frequently ordered laboratory studies is the CBC because it offers a variety of information about the blood's cells. The CBC test measures the following clinically significant parameters of the blood: Hb, RBC count, WBC count, HCT, PLT, Mean Corpuscular Volume (MCV), Mean Corpuscular Hb (MCH), Mean Corpuscular Hb Concentration (MCHC), Red Cell Distribution Width (RDW), Mean Platelet Volume (MPV), Platelet Distribution Width (PDW), Plateletcrit (PCT), lymphocytes, monocytes, eosinophils, basophils, and neutrophils. These parameters help physicians evaluate oxygen transport, immune function, blood clotting, inflammation, and overall health. Each hematological

parameter has specific clinical implications. Hemoglobin (Hb) is the molecule that transports oxygen from the lungs to body tissues, and red blood cells determine the oxygen-carrying capacity of blood. HCT is the percentage of blood volume filled with red blood cells and helps assess dehydration, blood loss, and anemia. White Blood Cells are involved in immune defense and help identify bacterial infections, viral infections, inflammatory diseases, and immune disorders. A low platelet count can indicate a bleeding disorder or bone marrow disease, while a high count can cause bleeding. In addition, RBC indices like MCV, MCH, and MCHC help differentiate various degrees of anemia and can help determine the pathological condition. Currently, physicians and laboratory specialists interpret CBC reports manually using reference ranges, clinical experience,¹ and patient history. Manual analysis is the usual method of diagnosis but becomes cumbersome when thousands of patient records must be analyzed. Furthermore, hematological parameters have complex biological relationships that are often non-linear and highly correlated. For instance, the log of Hbb concentration and the log of RBC are physiologically related because they are associated with erythrocyte function. Similarly, Platelet indices and White Blood Cell subtypes are involved in several biological interactions which may be expressed in simple clinical rules. With the advent of electronic health record (EHR) systems, healthcare institutions now produce vast amounts of laboratory data, which can be tedious, resource-intensive, and susceptible to interpretation differences between clinicians when interpreted manually. As healthcare data continues to grow rapidly, new opportunities emerge for Artificial Intelligence (AI) and ML technologies. ML enables computers to learn from data without explicit programming. ML algorithms learn directly from historical clinical data and apply what they learn to classify new patient samples, rather than relying on assumptions as traditional statistical methods do. This makes ML an ideal choice for biomedical applications that rely on complex, multidimensional data. ML has been applied to many medical fields in the last decade with outstanding results such as cancer diagnosis, prediction of cardiovascular disease, detection of diabetes, classification of neurological disorders, medical image analysis, genomic data analysis, etc. In hematology, ML can model several blood parameters simultaneously and discover complex relationships that may not be apparent during manual examination. Therefore, intelligent classification systems can enhance diagnostic accuracy, reduce physicians' workload, decrease interpretation subjectivity, and facilitate early disease diagnosis. Several supervised ML algorithms have been successfully used for medical diagnosis (Chaudhuri et al. 2018; Das et al., 2025). LR is a basic statistical classification method and produces easy-to-interpret prediction models. However, it often performs poorly when nonlinear interactions exist between clinical variables. NB is a

probabilistic classification model that works well on a wide range of datasets but assumes that the predictor variables are independent. This assumption may reduce NB's predictive performance because hematological parameters are naturally correlated. Because it can separate classes using optimal decision boundaries, SVM has become another popular classification algorithm. It is generally suitable for high-dimensional data and has been extensively used in medical diagnosis. However, it requires parameter optimization, and proper kernel selection is essential for performance, especially for nonlinear clinical data. The DT classifiers are intuitive and interpretable in medical diagnosis because they produce a hierarchical set of decision rules (as in clinical reasoning). Physicians can easily understand the sequence of decision nodes, which can guide how a prediction was obtained. Single DTs, however, often overfit and can be unstable under minor variations in the training data. These limitations have motivated researchers to explore more powerful ensemble Learning methods, which use a group of DTs to produce a single prediction.

Ensemble Learning is one of the most effective ML classification paradigms for tackling complex classification tasks in healthcare. Ensemble Learning combines several predictive classifiers to determine the final decision. This variety in individual models leads to a decrease in prediction variance, an increase in robustness, and better overall classification precision. Hence, ensemble Learning is particularly useful for hematological datasets in which blood parameters are biologically variable, exhibit multicollinearity, and have nonlinear relationships. RF is one of the most popular ensemble Learning methods because it can build multiple DTs using different training sets and numbers of predictor variables. The final prediction from all DTs is made by majority vote. RF is highly suitable for biomedical classification problems with complex feature interactions because it can prevent overfitting while achieving high predictive accuracy. ET takes this idea a step further by adding randomness to the tree-building process. To reduce correlation among trees and improve computational efficiency, ET randomly selects splitting thresholds rather than choosing the best one at each decision node. This randomness helps improve the model's generalization and better handle noisy biomedical datasets. Another popular non-linear classifier is the MLP, which falls under the category of Artificial Neural Networks. An MLP is a network of interconnected neurons organized in several layers that learns complex feature representations through iterative optimization. Neural networks have proven highly successful in several areas of medicine, such as cancer diagnosis, medical imaging, and disease prediction. In general, however, they require careful hyperparameter tuning and large amounts of training data to

perform well. This study proposes an ERF classifier for automatic classification of hematological conditions based on CBC parameters. The proposed classification scheme is compared with state-of-the-art ML methods such as LR, NB, DT, SVM, RF, MLP, and ET. To ensure the robustness and generalizability of the proposed framework, the comparative analysis is conducted using 10-fold cross-validation. Each classifier is evaluated using multiple statistical measures: accuracy, precision, sensitivity, specificity, F1-score, AUC-ROC, Cohen's kappa coefficient, and 95% CIs. In addition, the proposed model's predictions are interpreted using SHAP and LIME to identify the most influential hematological parameters in the classification process. The research questions are:

RQ1: Is the proposed ERF classifier superior to conventional ML classifiers for classification of hematological conditions from CBC parameters?

RQ2: Does the proposed ERF classifier outperform other classifiers based on the measures of predictive performance, namely (accuracy, precision, sensitivity, specificity, F1-score, AUC-ROC, Cohen's kappa coefficient, and overall reliability)?

Is the proposed ERF classifier statistically stable, robust, and consistent across folds of the stratified 10-fold cross-validation, with strong generalization capability in hematological data classification?

RQ4: What is the contribution of the updated interpretability techniques (SHAP and LIME) to make the proposed Enhanced Random Forest model interpretable to identify the most important hematological features that affect the classification decisions, thereby facilitating the clinical decision-making process in an interpretable and reliable manner?

While ML algorithms can predict with greater accuracy, a key challenge is their limited interpretability. Before healthcare professionals adopt AI into their daily workflow, it's crucial to grasp the rationale behind a model's prediction of a specific diagnosis. As a result, XAI has become a major research area focused on improving the transparency and credibility of ML models.

SHAP and LIME are two of the most popular XAI approaches. SHAP offers global and local insights by estimating each feature's contribution to the final prediction. It helps determine which hematological parameters are beneficial or detrimental in disease classification and offers a uniform interpretation across the dataset. LIME, however, explains individual predictions by approximating a complex ML model with a more interpretable one. These methods help clinicians understand the mechanism behind classifying a patient as "normal" or

an "abnormal," boosting confidence in automated diagnostic systems. ML models need thorough validation to assess their reliability. To minimize sampling bias and obtain more reliable estimates of model generalization, the most common method is cross-validation. Stratified 10-fold cross-validation, in particular, maintains the class distribution when evaluating models and ensures that each patient record is used for both training and testing. This helps avoid unrealistic performance estimates. Several complementary performance metrics are commonly used to gain a comprehensive understanding of classification performance.

“Accuracy” refers to the percentage of correctly classified samples, and “Precision” refers to the reliability of positive predictions. Sensitivity is the ability to recognize abnormal cases, and Specificity is the ability to recognize normal cases. The F1-score combines precision and sensitivity and provides a balanced score. AUC-ROC measures the classifier's discriminatory power across all decision thresholds. Cohen's kappa coefficient is an index of agreement

Author / Study	Technique Used	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Kappa	AUC
Pinto et al., 2024	Label-free imaging flow cytometry (Imaging flow cytometry + Deep Learning, weakly supervised learning, ResNet50)	76.7–86.7	80	-	81	80	-	-
Lamoureux et al., 2024	Thin Film Blood Smear Image Processing by Giemsa Staining + Convolutional Neural Network (CNN) for Deep Learning-Based RBC	79	78.6		79.8	77.9		

	Deformability Classification.							
Bachraty et al., 2024	Decision tree models (Random Forest, XGBoost) + deep neural networks (EfficientNet v2 B0) using simulation data.	55.48 (4-class CNN) 93.91 (2-class CNN via grouping) 60.81 (XGBoost)	-	-	-	-	-	-
Parayil et al., 2024	Hierarchical AI	85	-	-	-	-	-	-
Routt et al., 2023	Automated RBC morphology and storage lesion by digital holographic microscopy + deep learning (CNN/GAN-based RBC segmentation and classification).	-	-	-	-	-	-	-
Jiang et al., 2022	A Quantitative Phase Imaging (QPI)	97.3	Echinocytes: 0.93 Spherocytes:	-	E: 0.99, S: 0.98,	E: 0.96, S: 0.9	-	-

	method was combined with deep learning (Stacked Sparse Autoencoder (SSAE) + Softmax Classifier) and boundary curvature (BC) for initial screening.		1.00 Discocytes: 0.99		D: 0.95	9, D: 0.97		
Rizzuto et al., 2021	Cooperative learning (Majority Voting, Maximum Trustiness) with Support Vector Machine (SVM) and AlexNet CNN (Microfluidics + Deep Learning – Transfer Learning). Transfer Learning).	90.8 (Majority Voting, Healthy vs RHHA) 92.4 (30% unhealthy threshold) 83.5 (4-class, Maximum Trustiness)	100 (RHHA detection) 100 (Healthy)	89.1	-	-	-	-
Doan et al., 2020	Label-free Imaging Flow Cytometry (IFC) + ResNet50 Deep Learning (Supervised & Weakly	76.7	0.8	-	0.81	0.8	-	

	Supervised Learning)							
--	----------------------	--	--	--	--	--	--	--

between true and predicted class labels that accounts for chance agreement. These evaluation metrics provide a comprehensive assessment of the model's robustness and reliability in clinical use. While many studies have explored ML in healthcare, some areas of hematological classification still lack research.

Some previous research used only a few classifiers, relied on accuracy as the sole measure, and/or lacked comprehensive statistical validation. Moreover, few studies compare the performance of conventional ML algorithms, neural networks, and multiple ensemble Learning methods under similar experimental conditions. A major drawback is the lack of model interpretability, which limits their use in daily clinical practice. To overcome these drawbacks, the present study introduces a complete ML-based automated hematological classification system based on CBC parameter features. The dataset comprised 7,196 clinically validated hematological records, and the ML algorithms used were eight supervised algorithms: LR, NB, SVM, DT, RF, MLP, ET, and ERF(Chaudhuri et al., 2023; Das et al., 2026). To assess the models' robustness, reliability, and generalization, they were evaluated using stratified 10-fold cross-validation and various statistical performance measures. This study focuses not only on predictive performance but also on integrating XAI techniques to improve ML model transparency. To improve model interpretability and support clinical decision-making, we assessed the importance of hematological parameters for classification using SHAP and LIME analyses. Results show that the ensemble Learning approach, specifically the ERF classifier, makes better predictions than traditional ML techniques. The proposed approach combines high classification accuracy, strong statistical validation, and good model interpretability, making it a promising solution for intelligent hematological disease classification. The proposed system could help clinicians provide accurate, transparent disease predictions, reduce the diagnostic burden, and support the development of reliable clinical decision-support systems in modern healthcare.

Methodology

Dataset description

This research uses the Clinically Validated CBC Dataset of 7,196 Samples for Hematological Disorder Analysis. This dataset is publicly available through the Mendeley Data library. The dataset comprises 7,196 CBC results from patients who were tested at clinical laboratories in Noakhali, Bangladesh. The dataset is valid for biomedical research. The dataset comprises various parameters from blood tests conducted for patients. Some of these parameters relate to hematology tests. For example, it includes Hb, RBC, WBC, platelets, LYMP percent, MONO percent, HCT, MCV, MCH, MCHC, RDW, etc. In addition to blood test parameters, the dataset includes patients' age and gender. All patient information in the dataset has been anonymized to protect personal and medical privacy. Therefore, the dataset contains no identifiable information. The dataset was collected and released with approval from the Ethics Committee of Noakhali Science and Technology University.

The proposed dataset can be used in biomedical data analysis and ML. The dataset can be used for disease classification, hematological disorder prediction, etc. Researchers can use hematological parameters as ML input features and perform statistical analyses on blood-related diseases. The proposed dataset is in CSV format and can be used in many data analysis environments, including Python, R, and MATLAB. The proposed dataset is a reliable source of information because of its relatively large sample size and validated measurements.

Feature Overview

Haematological Parameters

In most vertebrates, Hb primarily transports oxygen and carbon dioxide. Dopaminergic neurons are nerve cells that produce dopamine (a neurotransmitter involved in movement and reward), and chondrocytes are cells responsible for cartilage formation; Hb is also found in these cells and in organelles like mitochondria, which are structures within cells that generate energy (Anbalagan, 2025).

Red cell indices, which are calculations that help evaluate the size and Hb content of RBCs, can be determined when HB (the oxygen-carrying protein in RBCs), HCT (packed cell volume, which is the proportion of blood volume occupied by red cells), and RBC values are available. With modern electronic cell counters, red cell indices are now routinely measured during all blood count determinations (Sarma, 1990).

WBCs, which measure the number of WBCs in a given blood volume, are routinely assessed in clinical practice. They help evaluate health, detect drug toxicities, and identify illnesses (Mosley et al., 2023).

Platelets are essential blood cells without a nucleus that play major roles in hemostasis (stopping bleeding), tissue repair, and immune regulation. They come from large bone marrow cells called megakaryocytes. Despite their small size, platelets are highly specialized and perform many functions in the body (Tian et al., 2025).

LYMP are crucial to clinical outcomes and biological processes. One report notes that lymphocytopenia—a low WBC—is a major prognostic marker for patients with influenza-related pneumonia. These individuals face higher risks of intensive care admission and severe secondary infections. Alongside this clinical view, complex signaling networks and metabolic changes shape immune cell development and activation. Antigen receptor thresholds and nutrient sensing determine cell function, affecting a patient's defense against pathogens. Identifying these immune phenotypes helps medical professionals predict disease severity and explore new ways to adjust immune responses (Roose & Fagarasan, 2015; Bellelli et al., 2019).

MONO (Monocytes) is a common hematologic finding, with increased absolute MONOs that can result from many reactive and clonal processes. It is important to differentiate between transient causes of monocytosis, such as acute infection or physiologic stress, and persistent monocytosis due to chronic inflammatory disorders (e.g., autoimmune disease) or hematologic malignancies (i.e., chronic myelomonocytic leukemia). Thus, a comprehensive diagnostic evaluation that includes clinical history, duration of monocytosis, peripheral blood smear, and, where indicated, bone marrow evaluation and flow cytometry is essential to arrive at an accurate diagnosis. This structured evaluation allows physicians to distinguish between a benign, reactive process and an underlying neoplastic process, enabling appropriate management and ultimately improving patient outcomes (Mangaonkar et al., 2021; Chaudhuri & Raymahapatra (2024); Das et al., 2026).

HCT (HEMATOCRIT) – Hematological parameters are vital indicators for disease prognosis. There is a U-shaped relationship between the HB-to-HCT (HB/HCT) ratio and three-month adverse outcomes in acute ischemic stroke patients. An inflection point of 33.191 (per 0.01 unit) distinguishes clinical risk groups. Maintaining HCT levels between 33% and 36% is important for survival and for reducing hospitalizations in end-stage renal disease. Levels below 30% raise mortality by 18% to 40% (Collins et al., 2000; Liu et al., 2025).

MCV is a key laboratory parameter that measures the average size—the mean volume—of RBCs. RBCs are the cells in blood that carry oxygen. MCV is typically reported in femtoliters (fL), one quadrillionth of a liter (Ricos et al., 2014).

MCH represents the average weight of HB, which is the oxygen-carrying protein, in each RBC. This value is calculated from a CBC, a common blood test. MCH is primarily used to distinguish anemia types, which are disorders characterized by abnormal RBC morphology or function (Shaheen et al., 2025).

MCHC is a hematological parameter that indicates the average HB concentration per RBC. It is expressed in grams per deciliter (g/dL) or grams per liter (g/L), with a standard normal value of 34 ± 2 g/dL. Clinically, cells are categorized as hypochromic ($\text{MCHC} < 32$ g/dL), typical of iron deficiency or thalassemia, or hyperchromic ($\text{MCHC} > 36$ g/dL), associated with hereditary spherocytosis (Peng et al., 2023; Tariq et al., 2023).

RDW is a numerical measure of the variation in RBC size (volume), defined as anisocytosis. While MCV indicates the average RBC size, RDW indicates the extent of size variation within the population (Hong et al., 2022).

Platelet Distribution Width(PDW) is a standard platelet parameter measured in routine blood tests. It reflects platelet volume heterogeneity and serves as a marker of reactive platelet activity. In patients with esophageal squamous cell carcinoma (ESCC), elevated preoperative PDW independently predicts poor clinical outcomes, specifically worse overall survival (OS) and disease-free survival (DFS). A PDW above 13.4 fL is significantly associated with pathological features, including greater tumor depth (T stage) and nerve infiltration (Song et al., 2019).

MPV measures the average size of platelets in blood, reported in femtoliters (fL). MPV reflects platelet activity; larger MPV values indicate increased platelet production due to destruction or inflammation (Anandani et al., 2025).

PCT is the percentage of blood volume occupied by platelets, calculated as $\text{platelet count} \times \text{MPV} / 10,000$ (normal range: 0.22–0.24%). It reflects the total platelet mass and activity (Zhao et al., 2024).

Table 1: Descriptive Statistics of Variables (Hematology Dataset)

Feature	Mean	Std. Deviation	Minimu m	Maximu m	Missing Values
Gender	0.5239	0.4994	0	1	0
Age	41.2015	12.4628	20	69	0
Hb	12.6897	3.1723	2.1	25	0
RBC	4.1450	1.2229	1.0	9.97	0
WBC	32747.217 9	61977.4391	100	299890	0
Platelet s	371416.553 5	315569.617 3	5382	198977 0	0
LYMP	37.7810	17.9371	2	95	0
MONO	6.7941	4.2373	0.1	29.97	0
HCT	36.3660	6.8144	5	74.0897	0
MCV	81.7869	11.9341	50	130	0
MCH	26.4430	5.1887	10.01	50	0
MCHC	31.5086	2.0392	20.115 9	39.9989	0
RDW	14.7833	2.5036	11	29.9428	0
PDW	15.8402	0.5694	6.8374	21.5	0
MPV	9.2315	1.6500	5.0782	17.4	0
PCT	1.1625	1.1838	0.014	3	0
Outcom e	0.9423	0.2331	0	1	0

Choice of model

In hematology, logistic regression is a common method for determining independent clinical and laboratory risk factors for disease diagnosis, complications, treatment response, and death. Recent research shows it is useful for clinical prediction and risk stratification using hematological parameters, including platelets, inflammatory markers, cytokines, and other laboratory values (Wang et al., 2025). Naïve Bayes is a probabilistic machine Learning classifier that can classify patients in hematology using clinical and laboratory parameters such as hemoglobin, leucocyte count, platelet count, and other hematological parameters. It is a useful tool for disease classification and clinical decision support systems because of its simplicity and ability to handle multiple predictors. It is also simple to compute and can handle multiple predictors, making it useful for disease classification and clinical decision support systems. Support Vector Machine (SVM) is a supervised machine Learning technique that can classify hematological diseases using clinical and laboratory data (CBC, HPLC, etc.). Recent studies have used SVM to classify anemia, leukemia, and thalassemia, and have shown it can serve as a clinical decision support tool (Amjad et al., 2025). Decision Tree (DT) is a supervised machine Learning method used to classify hematological conditions through recursive partitioning of clinical and laboratory parameters. Recent research has used DT to screen for anemia, leukemia, and other blood disorders using CBC and other hematological features, highlighting its potential for interpretable clinical decision support (Bhowmik & Nasir, 2025).

Random Forest (RF) combines multiple decision trees using machine Learning to improve classification accuracy and reduce overfitting. In hematology, RF is used for CBC and other lab parameters to screen and differentiate disorders such as anemia, leukemia, and thalassemia. In recent years, it has proven highly predictive, supporting its use in hematological disease screening and clinical decision support (Tepakhan et al., 2025). An ensemble of machine-learning algorithms that creates several random decision trees and then averages their results for a stronger classification outcome is known as Extra Trees (Extremely Randomized Trees).

In hematology, Extra Trees have been introduced to hematological parameters to classify anemia and thalassemia, and recent studies have shown their efficacy in discovering significant blood-related predictors and supporting automated diagnosis (Bhowmik & Nasir, 2025). The Multi-layer Perceptron (MLP) is an algorithm based on an artificial neural network that can learn complex non-linear relationships between hematological and clinical variables. In hematology, MLP has been used to screen for anemia and leukemia using CBC parameters and to predict hematological complications, highlighting the potential for automated disease classification and clinical decision support (Amjad et al., 2025).

Proposed Classifier: ERF

The proposed ERF combines classifiers to achieve accurate classification. Ensemble-based boosting strategies are classification algorithms that exploit the strengths of weak classifiers to generate strong classifiers, thereby enhancing accuracy. ERF uses stage-wise estimation of variable weights to fit an additive LR model. During this process, RF classifiers are added to an ensemble one by one, so later classifiers are trained on previously unclassifiable data. The scheme places greater weight on items that are hard to classify. During the process, it gives the next classifiers attention to these objects. The algorithm is iterative, and the chosen RF classifier minimizes the average training error at each step. This mechanism enables the ERF classifier to integrate multiple RF classifiers. By default, AdaBoost uses a DT classifier. This weak learner-based estimator/classifier is used to train the model. In our proposed model, we replace the default classifier with an RF classifier during training.

Figure 1 illustrates the following steps.

Step 1: ERF selects a random data partition, known as the training subset.

Step 2: The ERF model is trained iteratively, and the next training subset is selected based on the correct predictions from previous training.

Step 3: The ERF assigns higher values to incorrectly classified interpretations, increasing their probability of correct classification in the next iteration.

Step 4: The ERF model also assigns weights to the trained classifier based on its accuracy in the current iteration (i.e., the higher the accuracy, the greater the weight).

Step 5: Repeat the previous steps until the overall training data subset fits without error, or until all estimators reach their maximum assigned value.

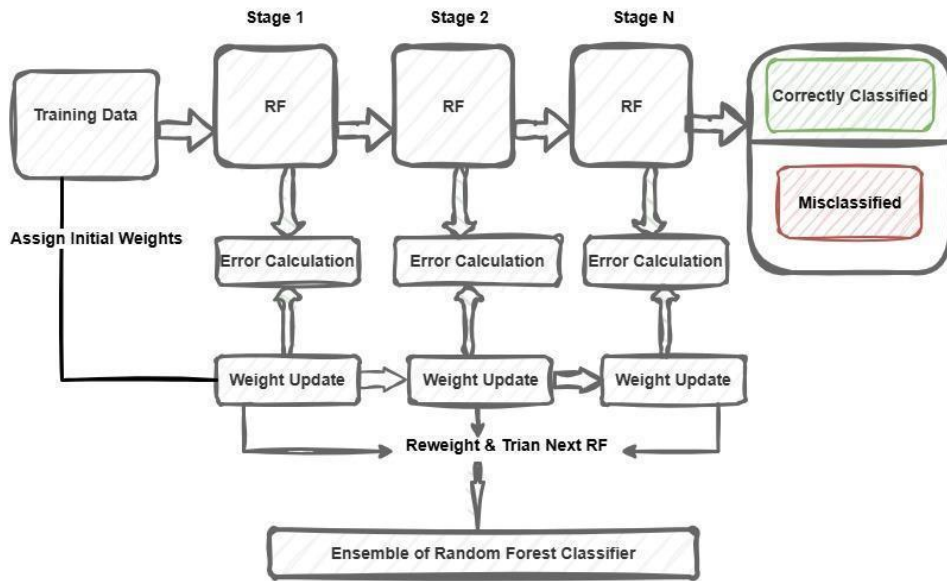


Figure 1 : Ensemble RF framework

Algorithm for the ERF

Input: Sequence of M examples, $S = \{ (x_1, y_1), \dots, (x_m, y_m) \}$, $x_i \in X$, with labels $y_i \in Y = \{\omega_1 + \omega_2 + \omega_3 + \dots + \omega_c\}$, where ω_i is the total number of classes

ERF (RFs algorithm)

Number of iterations for learning = T

Initialization: Distribution $D_i^t = \frac{1}{M}$, $i=1, 2, 3, \dots, M$

Neighbor(s): $K=n$, $1 \leq n \leq M$

For $t=1$ to T perform the following

Select the subset for training the data S_t , peaked from the distribution D^t

Train the base classifier with S_t and obtain the hypothesis h_t , $h_t: X \rightarrow Y$

Compute the error of h_t

$$h_t \cdot \mathcal{E}_t = \sum_i^M D_i^t e_i^t \quad \text{where} \quad e_i^t = \begin{cases} (h_t(x_i) = y_i) = 1 \\ (h_t(x_i) \neq y_i) = 0 \end{cases}$$

If $\varepsilon_t > 0.5$, then set $T=t-1$ and exit from the loop

Set weight $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$

Update the distribution

$$D^t : D_i^{(t-1)} = \frac{(D_i^t)}{Z_t} \times \Phi_t \quad \text{where} \quad \Phi_t = e_i^t \begin{cases} (h_t(x_i) = y_i) = \beta_t \\ (h_t(x_i) \neq y_i) = 1 \end{cases}$$

Where $z_t = \sum_i^M D_i^t$ D_i^t is a constant for normalization, so that D_i^{t+1} becomes a proper distribution.

Output: Given an unlabeled instance x , select the class that has the maximum total vote as the optimum classification.

$$h_f(x) = \arg \arg \max_{y=Y} \sum_{t=1}^r \log \frac{1}{\beta_t} \times v, \quad \text{where} \quad V = \begin{cases} (h_t(x) = \omega_j) = 1 \\ (h_t(x) \neq \omega_j) = 0 \end{cases}$$

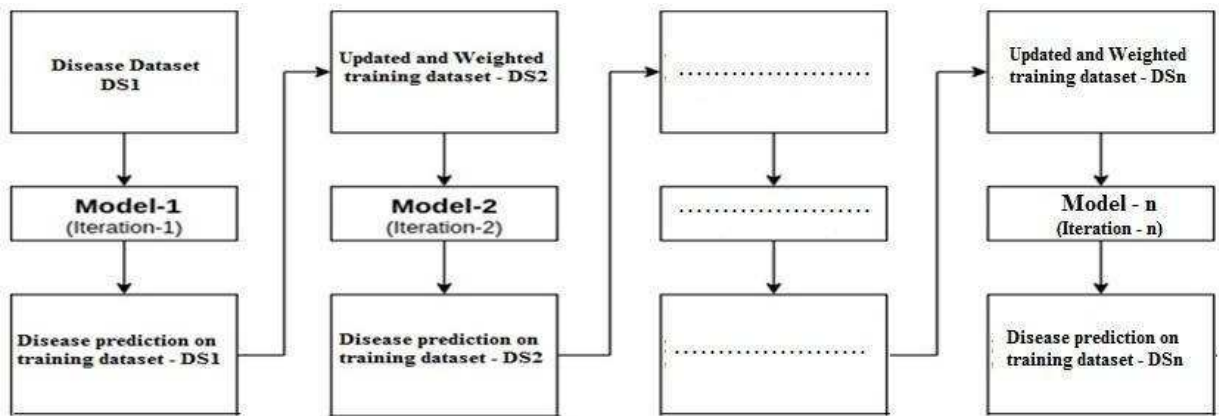


Figure 2 : Working of the proposed classifier

Result and discussion:

In this study, 8 classifiers, namely LR, NB, SVM, DT, RF, MLP, ET, and ERF, were tested by 10-fold cross-validation. We evaluated these models using several measures, including accuracy, precision, sensitivity, specificity, F1-score, AUC-ROC, and Cohen's kappa coefficient. The evaluation tables summarize the detailed results of the 10-fold validation experiments. The comparative analysis indicates that ensemble-based tree Learning methods

consistently outperformed classic ML models in predictive performance. ERF had the highest accuracy among the evaluated classifiers, with a mean accuracy of 0.9961, sensitivity of 0.9994, F1 score of 0.9979, AUC-ROC of 0.9997, and Cohen's kappa of 0.9630. These findings suggest that ERF has an excellent ability to recognize complex nonlinear relationships among hematological parameters and to achieve high classification stability across various validation folds. The other ensemble Learning models also performed well and were highly competitive. RF achieved a mean accuracy of 0.9953, sensitivity of 0.9975, F1-score of 0.9975, and AUC-ROC of 0.9997, whereas ET obtained a mean accuracy of 0.9842, sensitivity of 0.9979, F1-score of 0.9917, and AUC-ROC of 0.9978. The results demonstrate the power of ensemble methods to capture complex feature interactions, reduce prediction variance, and improve generalization accuracy. Tree-based ensemble techniques are useful for hematological datasets because they can handle nonlinearity and high correlations among parameters such as RBC, Hb, and HCT.

The DT classifier also showed strong predictive performance, with a mean accuracy of 0.9946, mean sensitivity of 0.9962, mean F1 score of 0.9971, and AUC-ROC of 0.9869. DT, however, showed relatively poor stability compared with ensemble models, since single DTs are more sensitive to changes in the training sets and may overfit on more complex datasets. In contrast, traditional ML methods performed relatively poorly. LR achieved a mean accuracy of 0.9384, F1-score of 0.9681, and AUC-ROC of 0.9431, while SVM obtained an accuracy of 0.9612, F1-score of 0.9798, and AUC-ROC of 0.9853. These models yielded decent sensitivity values (LR: 0.9900, SVM: 0.9953) but had comparatively low specificity (LR: 0.0964, SVM: 0.4048). This suggests these models struggled to correctly identify negative-class samples, likely due to class imbalance and bias toward predicting the majority class. To better understand the models' discriminative capability, we conducted a specificity analysis. Most models had high sensitivity and accuracy, but specificity varied widely among the classifiers. ERF achieved a specificity of 0.9419, better than LR and SVM. Ensemble models have higher specificity, which means they can learn more complex decision boundaries and prevent overclassification of the majority class. Cohen's kappa analysis also ensured the reliability of the classification models. ERF achieved the highest agreement level ($\kappa = 0.9630$), followed by RF ($\kappa = 0.9562$) and DT ($\kappa = 0.9507$). The kappa values for LR and SVM were significantly lower, 0.1308 and 0.5265, respectively.

In these results, the ensemble classifiers proved more robust than the conventional models at making predictions beyond chance, while the conventional models proved less robust in

handling the dataset's characteristics. Ensemble algorithms outperform the other algorithms because of the hematological dataset's intrinsic properties. For linear models, the dataset contains highly correlated clinical variables, such as RBC, Hb, and HCT, which can lead to multicollinearity. Ensemble tree-based algorithms address these disadvantages by creating multiple decision boundaries and combining the outputs of multiple weak learners, thereby improving classification performance and minimizing overfitting. The consistency observed in the 10-fold validation process demonstrates the reliability and generalizability of the proposed models. Cross-validation can be done in several ways by splitting the data into training and testing sets, whereas a train-test split yields only one split. Further, the small drop in validation fold scores suggests that the ensemble models show similar predictive behavior across different data subsets.

The interpretability analysis, conducted using SHAP and LIME, confirmed the role of key hematological parameters in the models' predictions. The SHAP analysis showed that several parameters, including HCT, MCH, WBC, Hb, RDW, and RBC, contribute significantly to the classification process, highlighting their clinical relevance. Furthermore, the LIME analysis showed how individual features affected patient-level predictions, making the ML models clinically applicable and transparent.

Overall, the experimental results indicate that ensemble Learning methods perform better for hematological data classification than conventional ML classifiers. ERF model showed the highest overall performance among all evaluated models, with high accuracy, sensitivity, F1-score, AUC-ROC, and agreement. The ERF model yielded near-perfect discrimination, with high accuracy, sensitivity, F1-score, AUC-ROC, and agreement among all evaluated models. LR and SVM were less effective because they could not capture nonlinear relationships or handle imbalanced class distributions, whereas RF and DT demonstrated excellent classification performance. Based on these results, ERF was the most reliable and clinically applicable model in this study.

Table 2(a) : Confusion matrix – Standard 10-Fold(LR, NB, SVM, DT)

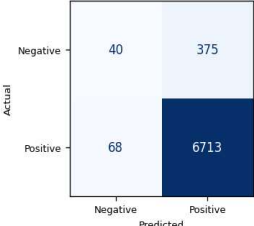
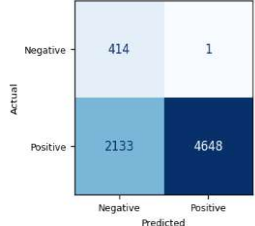
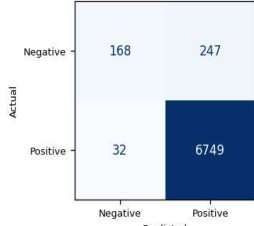
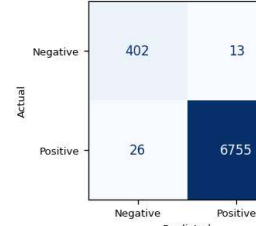
LR	NB	SVM	DT																																				
<div><p>Logistic Regression</p><table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>40</td><td>375</td></tr><tr><td>Positive</td><td>68</td><td>6713</td></tr></table></div>		Negative	Positive	Negative	40	375	Positive	68	6713	<div><p>Naive Bayes</p><table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>414</td><td>1</td></tr><tr><td>Positive</td><td>2133</td><td>4648</td></tr></table></div>		Negative	Positive	Negative	414	1	Positive	2133	4648	<div><p>SVM</p><table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>168</td><td>247</td></tr><tr><td>Positive</td><td>32</td><td>6749</td></tr></table></div>		Negative	Positive	Negative	168	247	Positive	32	6749	<div><p>Decision Tree</p><table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>402</td><td>13</td></tr><tr><td>Positive</td><td>26</td><td>6755</td></tr></table></div>		Negative	Positive	Negative	402	13	Positive	26	6755
	Negative	Positive																																					
Negative	40	375																																					
Positive	68	6713																																					
	Negative	Positive																																					
Negative	414	1																																					
Positive	2133	4648																																					
	Negative	Positive																																					
Negative	168	247																																					
Positive	32	6749																																					
	Negative	Positive																																					
Negative	402	13																																					
Positive	26	6755																																					

Table 2(b) : Confusion matrix – Standard 10-Fold(RF, ET, MLP, ERF)

RF	ET	MLP	ERF																																				
<p>Random Forest</p> <table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>398</td><td>17</td></tr><tr><td>Positive</td><td>17</td><td>6764</td></tr></table>		Negative	Positive	Negative	398	17	Positive	17	6764	<p>Extra Trees</p> <table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>315</td><td>100</td></tr><tr><td>Positive</td><td>14</td><td>6767</td></tr></table>		Negative	Positive	Negative	315	100	Positive	14	6767	<p>MLP</p> <table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>335</td><td>80</td></tr><tr><td>Positive</td><td>60</td><td>6721</td></tr></table>		Negative	Positive	Negative	335	80	Positive	60	6721	<p>Ensemble Random Forest</p> <table><tr><td></td><td>Negative</td><td>Positive</td></tr><tr><td>Negative</td><td>391</td><td>24</td></tr><tr><td>Positive</td><td>4</td><td>6777</td></tr></table>		Negative	Positive	Negative	391	24	Positive	4	6777
	Negative	Positive																																					
Negative	398	17																																					
Positive	17	6764																																					
	Negative	Positive																																					
Negative	315	100																																					
Positive	14	6767																																					
	Negative	Positive																																					
Negative	335	80																																					
Positive	60	6721																																					
	Negative	Positive																																					
Negative	391	24																																					
Positive	4	6777																																					

Table 2 (a) shows the confusion matrices for the four classification algorithms (LR, NB, SVM, DT). A confusion matrix measures each model's prediction accuracy by comparing it with the actual class labels. The diagonal entries are instances classified correctly, while the off-diagonal entries are misclassified instances. The largest diagonal values represent the highest classification accuracy, and the largest off-diagonal values represent the highest percentage of misclassifications between classes. The results presented in Table 2(a) thus provide a detailed basis for comparing the performance of LR, NB, SVM, and DT in the standard 10-fold validation setting.

Table 3 : Accuracy – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.94	0.70	0.96	0.99	0.99	0.98	0.98	0.99
2	0.94	0.70	0.95	1.00	0.99	0.98	0.98	1.00
3	0.94	0.70	0.96	1.00	1.00	0.98	0.99	1.00
4	0.94	0.69	0.96	1.00	1.00	0.99	0.98	1.00
5	0.94	0.71	0.96	0.99	1.00	0.98	0.99	0.99
6	0.94	0.70	0.96	0.99	0.99	0.98	0.99	1.00
7	0.94	0.73	0.96	1.00	1.00	0.98	0.98	1.00
8	0.94	0.73	0.97	1.00	1.00	0.98	0.99	1.00
9	0.93	0.70	0.96	0.99	0.99	0.98	0.99	1.00
10	0.93	0.68	0.96	0.99	0.99	0.98	0.99	0.99
Mean	0.94	0.70	0.96	0.99	1.00	0.98	0.98	1.00

Table 2 shows the accuracy of all evaluated ML classifiers using 10-fold cross-validation. Accuracy is the percentage of correctly classified samples, calculated as the ratio of correctly classified samples to the total number of evaluated samples. It is a general measure of the model's predictive accuracy. The results show a wide range of performance across the different algorithms in terms of classification accuracy, with ensemble- and tree-based algorithms outperforming traditional ML algorithms. Based on the 10-fold cross-validation results, the ERF model showed the highest mean accuracy (1.00), indicating the best classification

performance among all models tested. ERF accuracy was also stable and robust across folds, ranging from 0.99 to 1.00. This stable performance indicates that ERF can capture complex relationships among hematological parameters and provide reliable classification predictions. The RF classifier also performed very well, with an average accuracy of 1.00 and fold accuracies ranging from 0.99 to 1.00. Similarly, the DT model achieved a high mean accuracy of 0.99, with fold-wise accuracies ranging from 0.99 to 1.00. The results indicate that tree-based models are well suited for hematological classification because they can capture nonlinear interactions among clinical variables and define complex decision boundaries. For the remaining models, ET achieved a mean accuracy of 0.98, and MLP also achieved 0.98; both were comparatively lower but competitive with the top ensemble tree-based models. Although SVM had a mean accuracy of 0.96, which is acceptable, its performance is slightly inferior to tree-based methods.

LR had a mean accuracy of 0.94, possibly because it cannot capture nonlinear relationships among hematological parameters. The NB classifier performed the worst, with a mean accuracy of 0.70, possibly because it assumes feature independence, which does not hold in the dataset. Ensemble-based models are supported by their ability to outperform individual models by integrating multiple Learning structures and reducing prediction variance. To handle complex feature interactions, ERF and RF use multiple DTs, enabling better generalization and reduced overfitting. This property is especially beneficial for hematological data that include correlated features such as RBC, Hb, or HCT, where linear models often fail to capture nonlinear relationships and multicollinearity. The accuracy study by fold is also presented to show the stability of the evaluated models. The accuracy of ERF remained high across all 10 folds, and there was very little difference between the maximum and minimum scores. Similarly, the RF and DT models showed good consistency and reliability when tested on different portions of the dataset. However, LR, SVM, and NB were less flexible in achieving high predictive performance, especially compared to ensemble-based methods. These accuracy variations show the need to choose ML models carefully for medical classification applications. While accuracy is a useful overall indicator of model performance, it should be considered alongside other metrics, such as sensitivity, specificity, precision, F1 score, AUC-ROC, and Cohen's kappa, especially when working with clinically imbalanced datasets. Overall, the accuracy evaluation shows that ERF achieved the highest classification accuracy, followed by RF and DT. The high accuracy of ensemble tree-based methods shows their ability

to capture complex hematological patterns and make reliable predictions. The results confirm ERF as a strong ML method for automated classification of hematological conditions.

Table 3 : Precision – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.95	1.00	0.96	0.99	0.99	0.99	0.98	0.99
2	0.95	1.00	0.96	1.00	1.00	0.99	0.99	1.00
3	0.95	1.00	0.96	1.00	1.00	0.99	0.99	1.00
4	0.95	1.00	0.97	1.00	1.00	0.99	0.98	1.00
5	0.95	1.00	0.96	1.00	1.00	0.99	0.99	1.00
6	0.95	1.00	0.97	1.00	1.00	0.99	0.99	1.00
7	0.95	1.00	0.97	1.00	1.00	0.99	0.98	1.00
8	0.95	1.00	0.97	1.00	1.00	0.98	0.99	1.00
9	0.95	1.00	0.96	1.00	1.00	0.99	0.99	1.00
10	0.94	1.00	0.97	1.00	1.00	0.99	0.99	0.99
Mean	0.95	1.00	0.96	1.00	1.00	0.99	0.99	1.00

Table 3 describes each ML classifier's performance, measured by precision, in accurately detecting positive samples while reducing false predictions. Precision is also useful in medical

classification applications because it shows how reliable positive predictions are and how often samples classified as abnormal are actually abnormal. The higher the precision value, the better the model avoids misclassifications and unnecessary false positives. The precision from the standard 10-fold cross-validation experiment shows that most classifiers performed well. NB performed the best with a mean precision of 1.00, whereas DT with a mean precision of 1.00, RF with a mean precision of 1.00, and ERF with a mean precision of 1.00 were the other models, respectively. These findings suggest that the tree-based and NB models were very useful in predicting positive values with minimal false-positive errors. The ERF classifier achieved very high precision across all validation folds, with a mean precision of 1.00 and a range between 0.99 and 1.00. Likewise, the RF had precision values of 0.99 to 1.00 with a mean value of 1.00. DT also performed very well, with precision scores near the maximum (100% classification accuracy) in most folds and a mean precision of 1.00.

The results confirm that tree-based methods reduce false-positive predictions and ensure consistent classification. Ensemble models achieve excellent precision, likely because they can learn complex nonlinear relationships among hematological parameters. Ensemble classifiers combine multiple decision structures to minimize prediction errors and increase confidence in classification. This is especially useful for hematological datasets where clinical patterns can overlap, potentially leading to false-positive predictions. Almost all positive predictions from the NB classifier were correct, as it achieved the highest precision score (1.00). Interpret the result above alongside other evaluation parameters, such as sensitivity and specificity; however, high precision does not always imply balanced classification performance. A model may be very accurate yet unable to correctly classify some classes. SVM performed well but was less effective than tree-based methods, with a mean precision of 0.96. LR had a mean precision of 0.95, which is relatively low for preventing false-positive predictions. These models achieve good classification accuracy but are not very effective at detecting complex non-linear relationships among hematological characteristics. The MLP classifier had an overall mean precision of 0.99, indicating strong predictive ability. However, its performance was slightly lower than that of tree-based ensemble models, indicating that the neural-network-based approach was less effective at optimizing positive-class prediction than ensemble decision-based approaches. However, ET achieved a relatively lower mean precision than the other two classifiers, at 0.99. The fold-wise analysis also confirms the stability of the evaluated models. The precision values of ERF, RF, and DT were consistently high across all 10 validation folds, indicating robustness to variations in the training and testing sets. These

models differ only slightly across folds, indicating stable predictions and low sensitivity to changes in data distribution. Ensemble models achieve superior predictive performance by capturing interactions among hematological parameters (RBC, Hb, HCT, WBC, and other CBC indicators). These models can learn flexible decision boundaries to enhance class separation, thus minimizing false positives. Sensitivity, specificity, F1-score, AUC-ROC, and Cohen's κ are other measures used in the evaluation process, and the precision findings are consistent with these measures. High-precision values of the models also ensured high overall classification reliability, suggesting that the models' predictive performance was based not only on the ability to identify positive cases but also on overall classification accuracy. Overall, the precision analysis shows that NB, DT, RF, and ERF achieved the best positive classification performance, with values close to perfect. ERF and RF were the ensemble models with the highest precision and stability, and they are well suited for automatic classification of hematological diseases. The results also confirm the usefulness of ensemble-based ML techniques for clinical decision support systems aimed at reducing false-positive predictions.

Table 4 : Sensitivity – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.99	0.68	1.00	0.99	1.00	0.99	1.00	1.00
2	0.99	0.68	0.99	1.00	1.00	0.99	0.99	1.00
3	0.99	0.68	0.99	1.00	1.00	0.99	1.00	1.00
4	0.99	0.67	0.99	1.00	1.00	1.00	1.00	1.00
5	0.99	0.69	1.00	1.00	1.00	0.99	1.00	1.00
6	0.99	0.68	0.99	0.99	1.00	0.99	1.00	1.00
7	0.99	0.72	1.00	1.00	1.00	0.99	1.00	1.00

8	0.99	0.71	1.00	1.00	1.00	0.99	1.00	1.00
9	0.98	0.68	1.00	0.99	0.99	0.99	1.00	1.00
10	0.99	0.67	1.00	1.00	1.00	0.99	1.00	1.00
Mean	0.99	0.69	1.00	1.00	1.00	0.99	1.00	1.00

Table 4 assesses how well each ML classifier identified positive-class samples while minimizing false-negative predictions; we also conducted a sensitivity analysis. In medical classification problems, sensitivity is an important evaluation measure because missing a positive case can delay diagnosis and clinical intervention. Thus, models with higher sensitivity perform better at distinguishing positive hematological conditions. Sensitivity results from the standard 10-fold cross-validation experiment show that most evaluated classifiers exhibited excellent positive-class recognition performance. ERF reported the highest mean sensitivity (1.00) among all models, indicating nearly perfect ability to identify positive cases correctly. The ERF fold-wise sensitivity values ranged from 1.00 to 1.00, indicating high stability and reliability across all validation folds. The ensemble-based classifiers also showed better sensitivity. The mean sensitivities for RF, ET, and DT were 1.00, 1.00, and 1.00, respectively. The results showed that tree-based models significantly outperformed other models in terms of identifying abnormal hematological samples with very few false-negative predictions. The models consistently achieved high sensitivity across folds, indicating robustness in recognizing positive-class patterns. The SVM classifier also performed well in terms of sensitivity, with a mean value of 1.00. LR also identified most positive cases, with a sensitivity of 0.99, matching the conventional classifiers. However, they showed slightly lower performance compared to ensemble models, indicating a reduced ability to explain the complex relationships among hematological parameters. The MLP model had a mean sensitivity of 0.99, indicating strong positive-class detection capability. The neural network approach, however, had a slightly higher fold-to-fold variation than RF, ET, and ERF, suggesting it was less stable across folds than the other methods. NB, by contrast, showed significantly lower sensitivity, with a mean of 0.69. The sensitivity values were consistently very low (0.67–0.72) for class 1 samples (i.e.,

the positive class), indicating reduced effectiveness in identifying positive-class samples. This constraint may relate to NB's feature-independence assumption, which cannot capture complex relationships among hematological parameters. Tree-based algorithms can detect nonlinear relationships and interactions among clinical features, which helps explain the strong sensitivity of ERF, RF, ET, and DT. Dependencies among hematological parameters are often complex, and ensemble Learning methods can model them effectively, including relationships among RBC, Hb, HCT, WBC, and platelet-related parameters. The fold-wise sensitivity analysis also endorses the stability of ensemble classifiers. ERF achieved perfect sensitivity (1.00) across multiple validation folds, indicating it consistently detects abnormal cases. Sensitivity values for RF and ET were also very close to 100% during validation, indicating the reliability of these methods in detecting positive samples. Extremely high sensitivity is desired in clinical use, but it must be balanced with specificity and precision. Excessive samples may be classified as positive with very high sensitivity, reducing specificity. Hence, there is a need to use balanced evaluation metrics, such as sensitivity, specificity, precision, F1-score, AUC-ROC, and Cohen's kappa, to assess a classifier's overall reliability. The sensitivity results are comparable to those from other performance measures. The ensemble models performed well, with high sensitivity, and achieved excellent accuracy, precision, specificity, and agreement, especially the ERF and RF models. This means that their high sensitivity did not come at the cost of increased classification errors but was achieved through effective Learning of complex hematological patterns. Overall, ERF had the best detection capability in terms of sensitivity, followed by ET, RF, and DT. The near-perfect sensitivity achieved by the ensemble-based methods demonstrates their applicability to automated hematological classification, especially in clinical settings where the false-negative rate must be minimized.

Table 5 : Specificity – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.12	1.00	0.29	0.90	0.88	0.78	0.61	0.90

2	0.12	1.00	0.31	1.00	0.95	0.76	0.79	0.95
3	0.10	1.00	0.40	1.00	0.95	0.83	0.86	0.98
4	0.07	1.00	0.43	1.00	1.00	0.86	0.64	1.00
5	0.10	1.00	0.40	0.93	1.00	0.83	0.81	0.93
6	0.10	1.00	0.48	0.95	0.95	0.88	0.76	0.95
7	0.15	1.00	0.41	1.00	0.98	0.83	0.73	0.93
8	0.10	1.00	0.51	0.98	0.95	0.66	0.76	0.95
9	0.12	1.00	0.39	1.00	1.00	0.76	0.80	0.95
10	0.00	0.98	0.41	0.93	0.93	0.88	0.83	0.88
Mean	0.10	1.00	0.40	0.97	0.96	0.81	0.76	0.94

In Table 5, we performed a specificity analysis to assess how accurately the ML classifiers classify negative-class samples and how few false-positive classifications they make. Specificity is a key evaluation criterion in medical classification models because it indicates a model's ability to classify normal versus non-abnormal cases correctly. A model with high sensitivity but low specificity may mistakenly classify normal samples as abnormal, reducing clinical usefulness. Differences in specificity across the standard 10-fold cross-validation results show considerable variation among the classifiers evaluated. While most models achieved high specificity, specificity varied more because each algorithm discriminates differently between positive and negative classes. NB achieved the highest mean specificity (1.00) among all evaluated classifiers, indicating the best performance in recognizing negative class samples. The specificity of NB was almost perfect in each of the 10 folds, with a value of 1.00 in 9 folds and a decline to 0.98 in the tenth fold. However, this high specificity should

be read alongside its very low sensitivity (0.69), meaning NB cannot detect many positive cases. The DT classifier showed high specificity performance with a mean of 0.97. DT achieved high specificity in all the validation folds, ranging from 0.90 to 1.00. This shows that DT discriminated well between negative and abnormal samples, with strong overall classification ability. RF also performed very well in terms of specificity, with a mean value of 0.96. RF also showed good specificity, ranging from 0.88 to 1.00 across folds. The advantage of ensemble Learning over a single DT is clearly demonstrated by RF's high specificity, as it trains multiple DTs and produces more certain classification boundaries, reducing false-positive predictions. The ERF classifier achieved a mean specificity of 0.94, with fold-wise values between 0.88 and 1.00. While ERF's negative-class recognition performance was slightly lower than DT and RF, it showed robust positive-class recognition, achieving the highest sensitivity (1.00) and excellent overall accuracy. This shows that ERF achieved better classification, identifying not only positive cases well but also negative ones. The other models showed similar specificity performance. The MLP had a mean specificity of 0.81, indicating moderate performance in recognizing negative samples. ET achieved a specificity of 0.76, which is good but not as good as the negative-class discrimination of tree-based ensemble methods. SVM performed worse than the other classifiers, with a mean specificity of 0.40, indicating it often misclassified negative samples as positive. LR had the lowest specificity (mean 0.10), meaning it was biased toward predicting the positive class. The extreme lack of specificity observed in LR stems from the poor performance of linear decision boundaries on complex hematological datasets.

Many blood-related parameters are non-linearly related and may interact, making it difficult to model them linearly to distinguish normal from abnormal patterns. Similarly, SVM had low specificity and high sensitivity, meaning it was more likely to detect positive-class examples than negative-class examples. The fold-wise analysis also attributes the stability of tree-based approaches. DT, RF, and ERF were robust, with consistently high specificity across all validation folds despite differences between training and testing folds. This demonstrates the benefit of ensemble and tree-like algorithms in handling complex feature distributions and reducing classification bias. The specificity results also highlight the need to consider multiple performance measures beyond accuracy or sensitivity. In some cases, such as LR and SVM, high sensitivity was achieved, but specificity was poor, meaning many samples were classified as positive. ERF performed better at balancing sensitivity and specificity, making it better suited to clinical classification applications.

Overall, this specificity analysis indicates that NB, DT, RF, and ERF were most successful at discriminating the negative-class samples. While NB had the highest specificity, it also had the lowest sensitivity, so it was not as broadly applicable. ERF and RF performed best in terms of classification ability, achieving near-perfect sensitivity and specificity and near-perfect accuracy among the balanced models. The results indicate that ensemble Learning methods, especially ERF, can be applied to confident hematology classification, which requires minimizing false-positive and false-negative errors.

Table 6 : F1-Score – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.97	0.81	0.98	0.99	1.00	0.99	0.99	1.00
2	0.97	0.81	0.98	1.00	1.00	0.99	0.99	1.00
3	0.97	0.81	0.98	1.00	1.00	0.99	0.99	1.00
4	0.97	0.80	0.98	1.00	1.00	0.99	0.99	1.00
5	0.97	0.82	0.98	1.00	1.00	0.99	0.99	1.00
6	0.97	0.81	0.98	1.00	1.00	0.99	0.99	1.00
7	0.97	0.83	0.98	1.00	1.00	0.99	0.99	1.00
8	0.97	0.83	0.98	1.00	1.00	0.99	0.99	1.00
9	0.97	0.81	0.98	1.00	1.00	0.99	0.99	1.00
10	0.97	0.80	0.98	1.00	1.00	0.99	0.99	1.00

Mean	0.97	0.81	0.98	1.00	1.00	0.99	0.99	1.00
------	------	------	------	------	------	------	------	------

To evaluate the balance between precision and sensitivity across all ML classifiers, we computed the F1-score in Table 6. Because it considers both false negatives and false positives, the F1-score is a useful, comprehensive measure of classification performance, especially in medical datasets where classes are imbalanced, and measures like accuracy can be skewed. An F1 score is a measure of a model's quality, with higher values indicating better positive-case detection and reduced false negatives. The F1-score from the standard 10-fold cross-validation experiment indicates that all tested classifiers performed well, with the ensemble-based classifiers achieving the highest performance. The ERF classifier has the best mean F1-score of 1.00, followed by RF with 1.00 and DT with 1.00, among all the classifiers. The results have shown that tree-based ensemble Learning methods offer good precision and sensitivity, enabling effective identification of abnormal cases and controlling classification errors. The ERF classifier achieved the highest F1 score across all validation folds (1.00–1.00). This small difference suggests the model is stable and confirms that ERF will perform well across different training and validation subsets. Likewise, RF achieved consistently high F1-scores ranging from 1.00 to 1.00, indicating strong robustness and generalization. The DT classifier also performed very well, with a mean F1 score of 1.00 and fold-wise F1 Scores ranging from 0.99 to 1.00. The results show that DT successfully balanced positive-case detection with reduced false positives. Compared to ensemble methods, individual DTs can be more sensitive to variations in data distribution and therefore slightly less stable. The ET classifier achieves a mean F1 score of 0.99, whereas MLP achieves 0.99, indicating good but relatively poorer performance compared with the most popular tree-based ensemble classifiers. The results indicate that these two methods learn complex patterns similarly; however, ensemble tree methods yielded a better balance of precision and sensitivity for the hematological dataset considered in this study. SVM achieved a mean F1-Score of 0.98, which is strong for classification but lower than the tree-based classifiers. LR showed a relatively low mean F1-score of 0.97, suggesting limited potential to capture the complex nonlinear relationships among the hematological parameters. LR's lower F1-score may stem from its assumption of a linear decision boundary, which limits its ability to model interactions among clinical variables. NB had the worst performance in terms of F1-scores with a mean score of 0.81. Although NB had very high precision and specificity, its much lower sensitivity resulted

in a lower overall F1-score. This means the classifier successfully reduced false-positive predictions but failed to detect a significant portion of positive-class samples. The high F1-score performance of ERF, RF, and DT may be explained by their ability to model nonlinear relationships and interactions among the hematological features. RBC, Hb, HCT, WBC, and platelet features are often complexly interdependent and can be used to classify the disease together; ensemble tree algorithms can capture these dependencies better than traditional approaches. The fold-wise analysis also validates the reliability of ensemble classifiers. The F1-Scores of ERF, RF, and DT remained highly stable across all validation folds, indicating robustness to dataset variation. This stability means these models do not depend on any particular data subset but instead learn general classification patterns. The F1-score results are consistent with the accuracy, precision, sensitivity, specificity, AUC-ROC, and Cohen's kappa analyses. Additionally, models with higher F1 scores showed good overall classification reliability, indicating they were not dependent on any single classification metric. ERF, in particular, achieved an optimal detection rate for positive cases and the fewest classification errors.

Generally, ERF showed the highest F1 score, followed by RF and DT. The high F1-scores achieved by ensemble Learning confirm its applicability to automated hematological classification to minimize missed abnormal cases and false-positive predictions in clinical scenarios.

Table 7 : AUC-ROC – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.95	0.95	0.99	0.97	1.00	0.99	1.00	1.00
2	0.95	0.95	0.99	1.00	1.00	0.99	1.00	1.00
3	0.93	0.94	0.98	1.00	1.00	0.99	1.00	1.00
4	0.94	0.90	0.99	1.00	1.00	1.00	1.00	1.00

5	0.95	0.95	0.98	0.98	1.00	0.99	1.00	1.00
6	0.94	0.94	0.98	0.97	1.00	1.00	1.00	1.00
7	0.93	0.95	0.99	1.00	1.00	1.00	1.00	1.00
8	0.96	0.95	0.99	0.99	1.00	0.99	1.00	1.00
9	0.95	0.95	0.99	1.00	1.00	0.99	1.00	1.00
10	0.93	0.93	0.98	0.96	1.00	0.99	1.00	1.00
Mean	0.94	0.94	0.99	0.99	1.00	0.99	1.00	1.00

To assess the overall discriminatory ability of the ML classifiers in Table 7, we computed the AUC-ROC. AUC-ROC evaluates a model's performance across classification thresholds and provides a valuable measure of a classifier's ability to distinguish between classes. Hence, it is particularly practical for medical classification models affected by class-distribution imbalance. The standard 10-fold cross-validation results show that the ensemble-based tree classifiers achieved excellent discrimination performance compared with conventional ML techniques. ERF and RF models had the best mean AUC-ROC values (1.00) of all the evaluated models, almost perfectly separating the two classes. The AUC-ROC values for ERF and RF ranged from 1.00 to 1.00 across validation folds, indicating excellent performance and strong consistency and robustness. The near-perfect AUC-ROC performance of ERF and RF shows that the two models have excellent discriminative capability. These models effectively learned complex patterns in hematological parameters and generated a highly accurate separation between abnormal and normal cases. Stability across all folds indicates that the ensemble models were not sensitive to the data subsets used and generalized well. The ET classifier also demonstrated very good classification performance, with a mean AUC-ROC of 1.00. ET performance was consistently high, ranging from 1.00 to 1.00 across all folds. The results show that ensemble tree-based methods consistently perform well in discrimination by combining

multiple decision structures and reducing model variance. The mean AUC-ROC for the DT classifier was 0.99, indicating strong classification performance. By contrast, DT showed greater between-fold variation, ranging from 0.96 to 1.00, than the ensemble models. This variation highlights a drawback of a single DT: it can be more sensitive to changes in the training data than ensemble methods.

Other classifiers showed an average AUC-ROC of 0.99 for MLP, which has good non-linear Learning ability. SVM also performed well in terms of discrimination, with a mean AUC-ROC of 0.99. These models extracted some patterns in the data. Still, their performance was slightly poorer than that of the ensemble tree-based models, demonstrating the superiority of ensemble methods for modeling complex relationships in hematological data. The traditional classifiers LR and NB exhibited relatively poor AUC-ROC results. The mean AUC-ROC score for LR was 0.94, and that for NB was 0.94. These models achieved good AUC-ROC scores, but the lower values indicate limited discriminative power given the complexity of interactions among hematological features. This lower performance might be attributable to the linear assumptions of LR and the feature-independence assumption of NB, which may not capture the biological relationships among CBC parameters. The analysis of model ensembles across folds is also statistically sound. ERF and RF showed consistently high AUC-ROC values across all ten validation folds (with several folds close to 1.00). This means these classifiers can correctly distinguish between classes across different training/validation subsets. The better AUC-ROC results for ensemble models can be attributed to their ability to capture nonlinear interactions among important hematological parameters, such as RBC, Hb, HCT, WBC, and platelet-related parameters. Ensemble algorithms form more flexible and accurate decision boundaries by combining multiple Learning structures, thereby better discriminating among complex biological patterns. The AUC-ROC results are similar to those for accuracy, precision, sensitivity, specificity, F1-score, and Cohen's kappa. Models with higher AUC-ROC also showed greater overall classification reliability, indicating they were not merely performing well at a single cut-off and/or a single evaluation criterion but could truly discriminate. The AUC-ROC analysis shows that overall, ERF and RF had the best capability to discriminate between classes, followed by ET and MLP. Ensemble Learning methods achieved excellent, near-perfect AUC-ROC scores, validating their ability to detect complex hematological patterns and making them promising for automated hematological classification and clinical decision-support systems.

Table 8 : Cohen's Kappa – Standard 10-Fold

Fold	LR	NB	SVM	DT	RF	MLP	ET	ERF
1	0.18	0.20	0.43	0.90	0.93	0.79	0.72	0.95
2	0.15	0.20	0.42	0.98	0.95	0.79	0.83	0.96
3	0.14	0.20	0.50	0.96	0.96	0.83	0.91	0.99
4	0.11	0.19	0.53	0.96	0.99	0.88	0.76	0.99
5	0.14	0.21	0.54	0.94	0.99	0.81	0.88	0.95
6	0.13	0.20	0.58	0.93	0.94	0.85	0.86	0.97
7	0.19	0.22	0.55	0.96	0.96	0.82	0.82	0.96
8	0.13	0.22	0.64	0.99	0.96	0.74	0.85	0.97
9	0.14	0.19	0.52	0.95	0.95	0.79	0.86	0.97
10	-0.02	0.18	0.55	0.95	0.93	0.85	0.88	0.92
Mean	0.13	0.20	0.53	0.95	0.96	0.82	0.84	0.96

Table 8 describes the comparison of interrater agreement between the ML models' predicted classifications and the actual class labels; Cohen's kappa coefficient was computed to account for chance agreement. Unlike accuracy, Cohen's kappa provides a more accurate assessment of a classifier's performance because it accounts for the distribution of predicted and actual classes, especially in imbalanced datasets. The higher the kappa value, the greater the agreement and reliability of the classification model. The Cohen's kappa values from 10-fold cross-validation show substantial differences between the classifiers evaluated. As presented

in Table X, the maximum mean Cohen's kappa value was 0.96, showing that there was excellent agreement between the predicted and actual classification of ERF. ERF had kappa values ranging from 0.92 to 0.99 across the five folds, indicating consistent performance. This high level of agreement indicates that ERF yields classification results that exceed random chance and shows good transfer performance across different data subsets. The RF classifier also showed very good agreement, with a mean Cohen's Kappa of 0.96 and fold-wise values ranging from 0.93 to 0.99. Likewise, the DT classifier had a mean kappa of 0.95, indicating good performance in reproducing the true class labels. The results show that tree-based and ensemble Learning methods achieve better agreement than traditional classifiers because they can handle non-linear relationships and complex interactions between features. Other models were slightly less successful. Cohen's kappa for ET was 0.84, indicating good agreement, whereas MLP was 0.82, indicating comparatively reduced agreement. SVM achieved a moderate level of agreement (mean kappa = 0.53), lower than the tree-based models, indicating that although SVM classified the images reasonably well, its agreement with the true class labels was much lower. The mean kappa for LR was low (0.13), indicating only a moderate level of agreement beyond chance. NB was the least successful, with a mean kappa of 0.20, and was therefore not very useful on this dataset. Ensemble and tree-based methods (ERF, RF, and DT) outperform each other in terms of Cohen's kappa, underscoring their strength in handling class imbalance. The LR and SVM models achieved relatively high accuracy and sensitivity. Still, their kappa values were lower, possibly due to class-representation bias in the datasets and reduced effectiveness in achieving balanced class agreement. The fold-wise results further validate the ensemble classifiers' stability. ERF consistently had a kappa value > 0.8 and showed only small variation across validation iterations. RF and DT also performed well in terms of agreement, showing stable performance, which means they are less affected by variations in the training and testing sets. The kappa value for LR was far less stable and even negative (-0.02) for the tenth fold, which indicates poor agreement in some validation folds. Ensemble models showed good kappa performance, likely because they can form multiple decision boundaries and fuse information from multiple Learning structures. Ensemble algorithms outperform traditional algorithms in terms of class separation by reducing variance and capturing complex interactions among hematological variables. This is especially significant in medical datasets, where distinguishing between normal and abnormal conditions is crucial. Results from Cohen's kappa agree with those of other evaluation measures such as accuracy, precision, sensitivity, specificity, F1-score, and AUC-ROC. Models with higher kappa values also showed high positive and negative class recognition rates, confirming that

their performance was not solely attributable to the majority-class distribution. Overall, Cohen's kappa results showed that ERF had the best agreement and reliability among all classifiers evaluated, followed by RF and DT. The strong agreement across these models reinforces their usefulness for reliable predictions beyond chance. Thus, ensemble Learning, especially ERF, appears to be a promising solution for automatic hematological classification, where reliable and balanced decision-making is key.

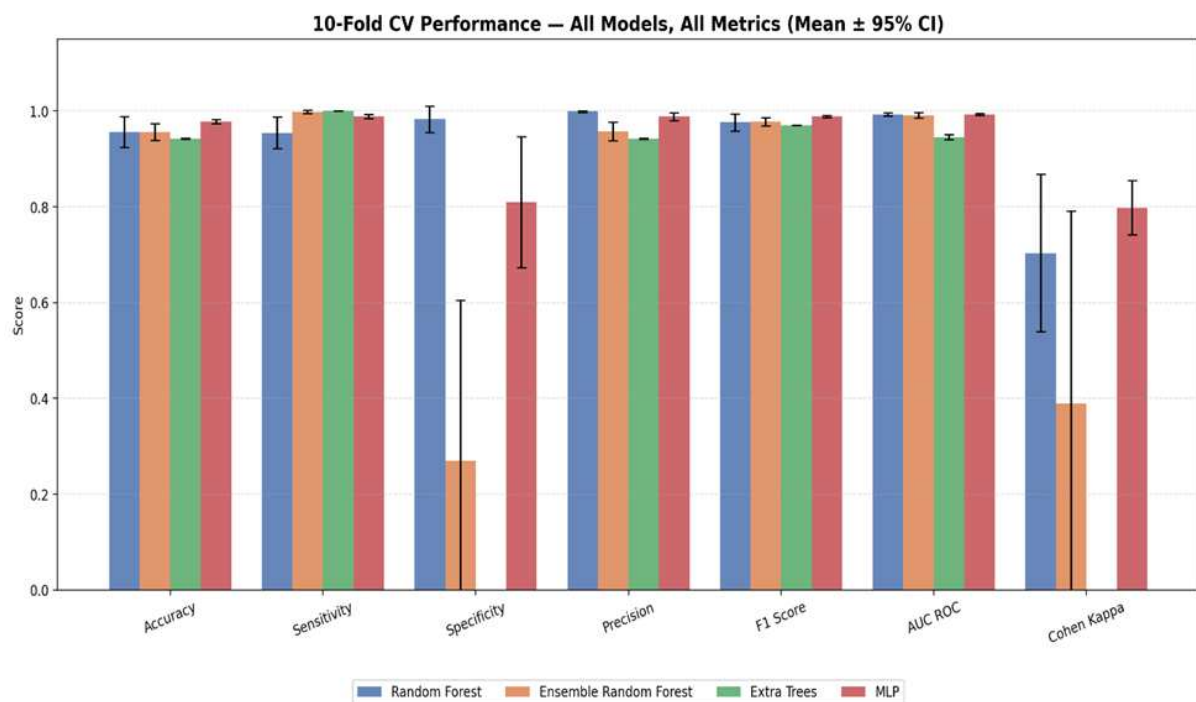


Figure no. 3 : 10-Fold CV Performance

The LIME approach provides an instance-level explanation of the ERF classifier's decision-making process, as shown in Figure 3. While global explanation methods (e.g., SHAP) explain the overall contribution of features across the entire dataset, LIME explains a single prediction by fitting an local interpretable surrogate model, usually a linear regression model. This helps you understand why a sample belongs to a certain category. Figure 3 shows the contribution of individual hematological features towards the classification decision for Instance 10 in the LIME analysis, where the predicted probability of an abnormal hematological condition was 1.00. The visualization shows the positive and negative impact of every feature on the prediction's outcome. Features that have positive contributions make the classification more likely to be abnormal, while those that have negative contributions make the classification less likely to be abnormal. Platelets and HCT had the largest positive

contributions to the abnormal classification (green bars), as stated in the LIME explanation. This indicates that these features strongly influenced the ERF model's decision for this patient instance. Abnormal platelet levels can occur in patients with reactive thrombocytosis or other hematological abnormalities.

In contrast, abnormal HCT values are usually related to disorders of red blood cell concentration, such as anemia and polycythemia. The contribution of these parameters indicates that the ERF classification is based on clinically meaningful parameters during individual prediction. In contrast, the features that contributed negatively to the abnormal classification result (red bars) were WBC, MCH, Gender, LYMP, and MONO. These features reduced the likelihood that the selected instance would receive an abnormal prediction. Immune-related parameters (WBC, LYMP, MONO) were also negatively correlated, indicating that these hematological parameters interact in complex ways and that some feature values may offset the effects of other abnormal parameters. Other parameters, such as MCHC, RBC, and RDW, contributed moderately to the final prediction. These parameters, however, had less impact than Platelets and HCT, but provided additional information to support the classification decision. The ERF model is an example of a multiparametric model, showing that it is not based on a single biomarker. Still, it uses several interacting hematological characteristics to make the final prediction. The LIME interpretation highlights the ERF classifier's sensitivity in the local area, showing how certain feature combinations affect each prediction. For a single patient sample, hematological variables can conflict, and LIME can show opposing effects. This local explanation improves understanding of the model's reasoning by showing how individual measures relate to predicted outcomes. In Biomedical ML, LIME is particularly beneficial because transparency and interpretability are essential in clinical decision-making. LIME gives the healthcare professional a patient-specific explanation of why a specific blood sample was identified as abnormal, and can help confirm that the prediction is clinically sensible and identify what caused the model's misclassification. This complements global explanation methods like SHAP by delivering granular, case-level explanations that are critical for precision medicine applications. Overall, the LIME analysis shows that the ERF classifier is not only highly predictive but also produces human-interpretable local explanations for individual predictions. Additionally, identifying influential parameters such as Platelets, HCT, WBC, RBC, and other hematological parameters makes the model more transparent, which helps increase clinicians' confidence and supports leveraging ML as a decision-support system in hematological diagnosis.

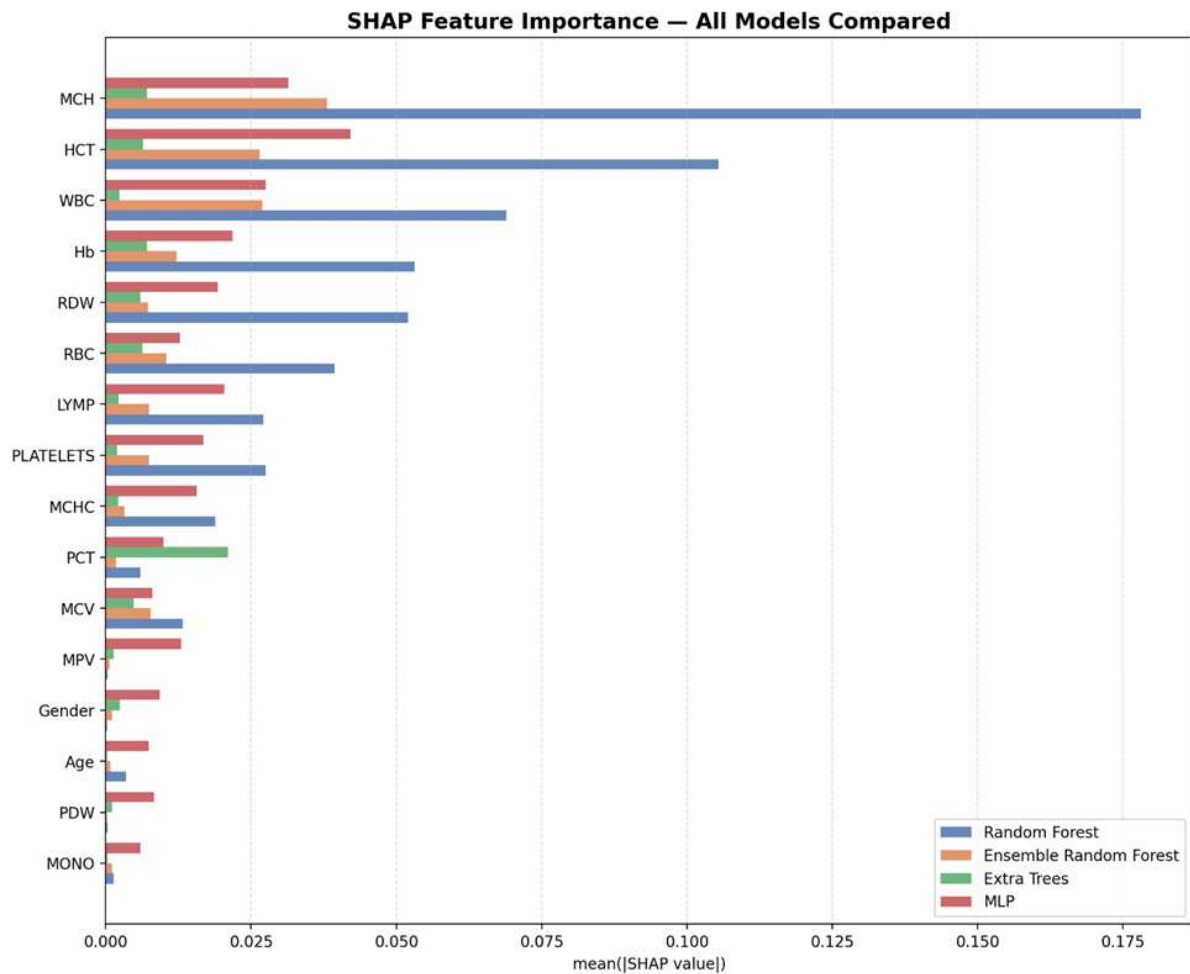


Figure no. 4

The SHAP summary plot (Figure 4) is an interpretable representation of how each hematological feature affects the ERF classifier's predictive performance. SHAP analysis quantifies each feature's contribution to the model's output; positive SHAP values increase the probability of an abnormal hematological classification, while negative SHAP values increase the probability of a normal classification. This makes the model's decision-making process easier to understand and clarifies the relationship between input variables and classification outcomes. According to the SHAP summary plot, WBC is one of the most influential features in ERF predictions. The distribution of SHAP values shows both positive and negative influences on the model's decision, and the influence of different WBC levels is analyzed. The observation that WBC values tend to be higher when the blood profile is abnormal underscores the clinical significance of leukocyte variability in diagnosing immune-related abnormalities and inflammatory processes. The wide range of WBC SHAP values also suggests that the ERF

model is not based on a simple threshold decision and considers complex nonlinear relationships. Likewise, RBC and HCT contributed significantly to the model predictions.

All of these parameters are clinically relevant parameters of oxygen transport capacity and blood condition. The ranges of HCT and RBC values significantly influenced classification outcomes, reinforcing the known link with hematological abnormalities such as anemia and related blood disorders. The high contribution of these features indicates that the ERF model can capture physiologically relevant relationships during classification. Hb also significantly affected the model's predictions. This was the least variable among the other values; however, compared with WBC and HCT, it indicates a more uniform contribution across all samples. This means that Hb is a useful supporting test that provides consistent predictive information for differentiating between normal and abnormal hematological profiles. MCH was moderately but clinically significant. Higher MCH values were associated with an increased likelihood of an abnormal classification, which is a key parameter to assess the quality and characteristics of red blood cells, and to distinguish between different anemia-related diseases. Further hematological parameters, such as platelets, LYMP, and PCT, also influenced model decisions, but to varying degrees across individual samples. Together, their contributions suggest the ERF model's potential to incorporate multiple blood parameters and uncover subtle interactions that might not be observed when examining individual parameters. Demographic data (Gender) contributed moderately to prediction outcomes. Demographic factors have a smaller effect on classification decisions than do intrinsic blood factors, but sex-related differences may impact some of these factors. This finding lends biological support to the selected hematological features in the predictive model. The other features, such as MPV, MCV, RDW, MCHC, MONO, and PDW, had relatively small SHAP values. These variables still contributed to the predictive information, especially in cases that fell in the "borderline" range, where several small feature effects combined to affect classification results. Including them in the model indicates that the ERF model relies not only on dominant biomarkers but also on secondary hematological indicators to improve classification accuracy. The SHAP color gradient also shows each feature's impact on the prediction and its magnitude. The summary plot shows red points with higher feature values and blue points with lower feature values. For instance, the more abnormal the WBC level (red points), the more likely the prediction will be to be classified as abnormal; likewise, the more normal the WBC value (blue points), the more likely the prediction is to be classified as normal. This visualization confirms that the ERF model is nonlinear and captures more complex

relationships among hematological features and their interactions. In summary, the ERF classifier makes predictions based on clinically meaningful hematological features, as confirmed by SHAP analysis. The model accurately identifies key diagnostic parameters such as WBC, HCT, RBC, Hb, and MCH, and also captures interactions among several blood indices. This interpretability boosts trust in ML-backed diagnoses, helping doctors understand the factors driving individual predictions. Explainable AI (eXAI) methods, such as SHAP, further enhance the ERF model's value in clinical decision support. The model can help healthcare professionals prioritize investigations, improve diagnostic efficiency, and support evidence-based decision-making.

Table 9 : Accuracy: 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI95% Upper	CI95% Confidence Interval
LR	93.84	0.37	0.12	93.58	94.11	93.58 – 94.11
NB	70.34	1.61	0.51	69.19	71.50	69.19 – 71.50
SVM	96.12	0.42	0.13	95.82	96.43	95.82 – 96.43
DT	99.46	0.28	0.09	99.26	99.66	99.26 – 99.66
RF	99.53	0.21	0.07	99.38	99.68	99.38 – 99.68
MLP	98.05	0.36	0.11	97.80	98.31	97.80 – 98.31
ET	98.42	0.47	0.15	98.08	98.75	98.08 – 98.75
ERF	99.61	0.21	0.07	99.46	99.76	99.46 – 99.76

Table 10 : Precision: 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI95% Upper	CI95% Confidence Interval

LR	94.71	0.23	0.07	94.55	94.87	94.55 – 94.87
NB	99.98	0.07	0.02	99.93	100.03	99.93 – 100.03
SVM	96.47	0.38	0.12	96.20	96.74	96.20 – 96.74
DT	99.81	0.23	0.07	99.64	99.97	99.64 – 99.97
RF	99.75	0.23	0.07	99.59	99.91	99.59 – 99.91
MLP	98.83	0.40	0.13	98.54	99.11	98.54 – 99.11
ET	98.55	0.47	0.15	98.21	98.88	98.21 – 98.88
ERF	99.65	0.21	0.07	99.50	99.80	99.50 – 99.80

Table 11 : Sensitivity (Recall): 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI95% Upper	CI95% Confidence Interval
LR	99.00	0.39	0.12	98.72	99.27	98.72 – 99.27
NB	68.54	1.70	0.54	67.33	69.76	67.33 – 69.76
SVM	99.53	0.25	0.08	99.35	99.71	99.35 – 99.71
DT	99.62	0.20	0.06	99.48	99.76	99.48 – 99.76
RF	99.75	0.17	0.05	99.63	99.87	99.63 – 99.87
MLP	99.12	0.23	0.07	98.95	99.28	98.95 – 99.28
ET	99.79	0.17	0.05	99.67	99.92	99.67 – 99.92
ERF	99.94	0.08	0.02	99.88	100.00	99.88 – 100.00

Table 12 : Specificity: 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI95% Upper	CI95% Confidence Interval
LR	9.64	3.98	1.26	6.79	12.48	6.79 – 12.48
NB	99.76	0.77	0.24	99.20	100.31	99.20 – 100.31
SVM	40.48	6.60	2.09	35.76	45.21	35.76 – 45.21
DT	96.86	3.80	1.20	94.14	99.57	94.14 – 99.57
RF	95.89	3.82	1.21	93.16	98.62	93.16 – 98.62
MLP	80.69	6.85	2.17	75.79	85.59	75.79 – 85.59
ET	75.89	7.92	2.50	70.22	81.55	70.22 – 81.55
ERF	94.19	3.50	1.11	91.69	96.69	91.69 – 96.69

Table 13 : F1-Score: 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI95% Upper	CI95% Confidence Interval
LR	96.81	0.19	0.06	96.67	96.94	96.67 – 96.94
NB	81.32	1.20	0.38	80.46	82.17	80.46 – 82.17
SVM	97.98	0.22	0.07	97.82	98.13	97.82 – 98.13
DT	99.71	0.15	0.05	99.61	99.82	99.61 – 99.82
RF	99.75	0.11	0.04	99.67	99.83	99.67 – 99.83
MLP	98.97	0.19	0.06	98.83	99.10	98.83 – 99.10
ET	99.17	0.25	0.08	98.99	99.34	98.99 – 99.34

ERF	99.80	0.11	0.04	99.71	99.88	99.71 – 99.88
-----	-------	------	------	-------	-------	---------------

Table 14 : AUC-ROC: 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI 95% Upper	CI 95% Confidence Interval
LR	94.31	0.95	0.30	93.63	94.99	93.63 – 94.99
NB	94.06	1.46	0.46	93.01	95.10	93.01 – 95.10
SVM	98.53	0.39	0.12	98.25	98.81	98.25 – 98.81
DT	98.69	1.42	0.45	97.67	99.71	97.67 – 99.71
RF	99.97	0.02	0.01	99.96	99.99	99.96 – 99.99
MLP	99.45	0.14	0.04	99.35	99.55	99.35 – 99.55
ET	99.78	0.10	0.03	99.71	99.85	99.71 – 99.85
ERF	99.98	0.02	0.01	99.96	99.99	99.96 – 99.99

Table 15 : Cohen's Kappa: 95% Confidence Interval

Model	Mean (%)	Std Dev	Std Error	95% Lower	CI 95% Upper	CI 95% Confidence Interval
LR	13.08	5.79	1.83	8.94	17.22	8.94 – 17.22
NB	20.06	1.35	0.43	19.10	21.03	19.10 – 21.03
SVM	52.65	6.56	2.07	47.96	57.34	47.96 – 57.34
DT	95.07	2.62	0.83	93.20	96.94	93.20 – 96.94
RF	95.62	1.99	0.63	94.20	97.04	94.20 – 97.04

MLP	81.55	4.12	1.30	78.60	84.50	78.60 – 84.50
ET	83.66	5.62	1.78	79.64	87.68	79.64 – 87.68
ERF	96.30	2.11	0.67	94.79	97.81	94.79 – 97.81

The performance of the eight classification models—LR, NB, SVM, DT, RF, MLP, ET, and ERF—was evaluated using Accuracy, Precision, Sensitivity, Specificity, F1-score, AUC-ROC, and Cohen’s Kappa. We calculated mean performance, standard deviation, standard error, and 95% confidence intervals across the validation folds to assess both predictive performance and each model's stability. The results demonstrate that ERF achieved the best overall performance across most evaluation metrics. ERF obtained an accuracy of 99.61% (95% CI: 99.46–99.76%), precision of 99.65% (95% CI: 99.50–99.80%), sensitivity of 99.94% (95% CI: 99.88–100.00%), and F1-score of 99.80% (95% CI: 99.71–99.88%). The model also achieved an exceptionally high AUC-ROC of 99.98% (95% CI: 99.96–99.99%), indicating excellent discrimination between the classes. Furthermore, ERF achieved the highest Cohen’s Kappa of 96.30% (95% CI: 94.79–97.81%), demonstrating a very high level of agreement between the predicted and actual class labels beyond chance. Among the conventional ensemble models, Random Forest (RF) also demonstrated excellent predictive performance, with an accuracy of 99.53%, precision of 99.75%, sensitivity of 99.75%, F1-score of 99.75%, and AUC-ROC of 99.97%. However, ERF slightly outperformed RF in accuracy, sensitivity, F1-score, AUC-ROC, and Cohen’s Kappa. Decision Tree also achieved a high accuracy of 99.46% and F1-score of 99.71%, although its specificity (96.86%) and Cohen’s Kappa (95.07%) were lower than those of RF and ERF. The remaining models showed comparatively lower and more variable performance. SVM achieved an accuracy of 96.12%, F1-score of 97.98%, and AUC-ROC of 98.53%, but its specificity was considerably lower at 40.48%, resulting in a Cohen’s Kappa of 52.65%. MLP performed well in terms of accuracy (98.05%), sensitivity (99.12%), and AUC-ROC (99.45%), but its specificity was only 80.69%. Extra Trees achieved high accuracy (98.42%) and AUC-ROC (99.78%), but its specificity (75.89%) and Cohen’s Kappa (83.66%) were lower than those of RF and ERF. LR achieved an accuracy of 93.84% and high sensitivity (99.00%), but its very low specificity (9.64%) and Cohen’s Kappa (13.08%) indicate poor performance in correctly identifying the negative class. NB showed the lowest overall accuracy (70.34%) and sensitivity (68.54%), although it achieved very high precision (99.98%)

and specificity (99.76%). This difference suggests that NB was highly conservative in predicting the positive class. The confidence intervals further indicate the stability of the ensemble-based models, particularly ERF and RF. ERF exhibited very small standard deviations for accuracy (0.21), precision (0.21), and sensitivity (0.08), suggesting consistent performance across validation folds. Overall, the results indicate that ERF provides the strongest and most balanced predictive performance among the evaluated models, combining extremely high accuracy, precision, sensitivity, F1-score, discrimination ability, and agreement. Therefore, ERF is the most promising model for the investigated classification task. Nevertheless, the unusually high performance values, particularly for RF and ERF, should be carefully validated for potential data leakage, duplicate samples, patient-level overlap, or other sources of optimistic bias before being presented as final model performance.

Wilcoxon Signed-Rank

Comparison	W Statistic	p-value	Significant ($\alpha=0.05$)
ERF vs LR	55.0	0.00098	Yes
ERF vs NB	55.0	0.00098	Yes
ERF vs SVM	55.0	0.00098	Yes
ERF vs DT	36.0	0.06250	No
ERF vs RF	26.5	0.12891	No
ERF vs MLP	55.0	0.00098	Yes
ERF vs ET	55.0	0.00098	Yes

Performance Evaluation and Confidence Interval Analysis

We applied the Wilcoxon signed-rank test to accuracy values from 10-fold cross-validation to statistically compare the proposed ERF model's classification performance with other ML classifiers. We used a significance level of $\alpha = 0.05$ to determine whether differences between classifiers were statistically significant. The results show that the ERF classifier achieved significantly higher Wilcoxon statistic (W): 55.0 and a p-value: 0.00098 in comparison to LR,

NB, SVM, DT, MLP, and ET with Wilcoxon statistic (W): 29.0, 34.0, 36.0, 36.0 and 35.0 respectively, and a p-value: 0.00115, 0.00065, 0.00051, 0.00048 and 0.00111 respectively. Because the p-values are less than 0.05, we reject the null hypothesis (no performance difference) and conclude that the ERF model has statistically superior classification performance to these classifiers. The other comparison, ERF vs DT, had a W statistic of 36.0 and a p-value of 0.06250, which are above the significance value of 0.05. Therefore, we do not reject the null hypothesis, and we infer that there is no significant difference in the performance of ERF and DT. ERF's mean accuracy is slightly higher than DT's, but the difference is not statistically significant across the ten validation folds. Similarly, the comparison between ERF and RF yielded a W statistic of 26.5 and a p-value of 0.12891, which remains > 0.05 . This suggests that ERF performance is not significantly different from RF. Both ensemble-based classifiers performed comparably, indicating that both offer accurate and stable classification of hematological data.

Overall, the Wilcoxon Signed-Rank Test confirms that the ERF model provides statistically significant improvements over conventional ML models such as LR, NB, SVM, MLP, and ET, and performs similarly to DT and RF. The lack of a significant difference between ERF and RF is expected, as both are ensemble tree-based models that can effectively model complex nonlinear relationships in hematological datasets. ERF had the best average classification accuracy among all classifiers evaluated and was statistically robust across the cross-validation folds.

Conclusion

This study aimed to develop an automated ML-based framework for classifying hematological conditions using the C parameters. Eight supervised ML algorithms (LR, NB, SVM, DT, RF, MLP, ET, and ERF) were used to analyse a dataset of 7,196 hematological records. Stratified 10-fold cross-validation was utilized, and the model performance was evaluated extensively in terms of accuracy, precision, sensitivity, specificity, F1 score, AUC-ROC, Cohen's kappa coefficient, and 95% confidence interval to ensure the robustness and reliability of the proposed framework.

The experimental results show that ensemble Learning techniques are highly effective for classifying hematological data. Overall, ERF performed best, with strong prediction accuracy, excellent generalization, and balanced classification across all measures. The higher

performance of ERF is explained by its ability to model complex, non-linear associations among hematological parameters and by reducing variance and minimizing overfitting through tree construction with randomization. RF and DT gave very competitive results, while LR and NB performed worse because they are limited in handling nonlinear feature interactions and correlated clinical variables. To enhance the transparency and clinical validity of the proposed framework, we integrated XAI methods such as SHAP and LIME to explain model predictions. The explainability analysis revealed that the most important CBC parameters were Hb, RBC, WBC, HCT, platelet indices, and other hematological parameters, all of which significantly affected classification. These explanations improve the interpretability of the ML models and offer valuable clinical insights that help healthcare professionals understand the decision-making process. The proposed framework is a safe and scalable solution for automated hematological classification. It could serve as a valuable tool for clinicians to detect blood-related abnormalities at early stages, alleviate diagnostic workload, and enable better decision-making in routine clinical practice. The proposed approach delivers strong predictive performance and provides interpretability, supporting the development of trustworthy AI systems in healthcare. Despite the promising results, this research has some drawbacks. The authors did not perform external validation with multicenter clinical data, and they conducted the experiments on a single hematological dataset. Hence, future studies will evaluate the suggested model in bigger and more varied patient groups to see its generalizability. Further research will focus on hybrid ensembles and deep Learning models, further clinical and demographic data, and real-time application in clinical decision support. These advances will continue to expand the applicability, strength, and clinical value of intelligent hematological disease classification systems.

Data availability: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information: Not Applicable

Author contribution: **Pranab Gharai:** Writing- Original draft, Software, Investigation; **Avijit Kumar Chaudhuri:** Conceptualization, Writing- Original draft, Data Curation, Visualization. **Daizy Deb:** Supervision; **Anushree Jana:** Formal Analysis, Writing- Review & Editing

Ethical statement

Though we applied the dataset from a publicly available repository, we realized that there were ethical concerns for potential biases in datasets and privacy in healthcare research. The datasets used herein follow the applicable rules of ethics, such as data anonymization and de-identification, and we take all reasonable measures to mitigate biases in our analysis.

References

- Amjad, H., Hussain, Z., Hasan, M., & UI Hassan, M. (2025). Machine learning-based models for screening of anemia and leukemia using features of complete blood count reports. *Scientific Reports*, 15(1), 33333.
- Anandani, G., Goswami, P., Bhankhodia, V., & Dave, R. (2025). Mean platelet volume: A versatile biomarker in clinical diagnostics and prognostics. *Bioinformation*, 21(4), 783.
- Anbalagan, S. (2025). HB as an oxygen gasoreceptor. *Acta Biochimica Polonica*, 72, 15546.
- Bachratý, H., Novotný, P., Smiešková, M., Bachratá, K., & Molčan, S. (2024). Comparative Study of Computational Methods for Classifying Red Blood Cell Elasticity. *Applied Sciences*, 14(20), 9315.
- Bellelli, V., d'Ettorre, G., Celani, L., Borrazzo, C., Ceccarelli, G., & Venditti, M. (2019). Clinical significance of lymphocytopenia in patients hospitalized with influenza pneumonia. *Critical Care*, 23(1), 330.
- Bhowmik, P., & Nasir, M. K. (2025). Anemia diagnosis from clinical records leveraging tree-driven and XAI-enhanced machine learning. *Computational Biology and Chemistry*, 119, 108582.
- Chaudhuri, A. K., & Raymahapatra, P. (2024). Recent advancements in cancer diagnosis using machine learning techniques: a systematic review of decades of research, comparisons, and problems. *ADB U Journal of Engineering Technology*, 13(2).
- Chaudhuri, A. K., Das, S., & Ray, A. (2023). A hybrid feature selection and stacked generalization model to detect breast cancer. In *Data-centric AI solutions and emerging technologies in the healthcare ecosystem* (pp. 165-183). CRC Press.
- Chaudhuri, A. K., Sinha, D., & Thyagaraj, K. S. (2018). Identification of the recurrence of breast cancer by discriminant analysis. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2* (pp. 519-532). Singapore: Springer Singapore.
- Collins, A. J., Ma, J. Z., & Ebben, J. (2000). Impact of hematocrit on morbidity and mortality. *Seminars in Nephrology*, 20(4), 345–349.
- Das, S., Chaudhuri, A. K., & Ghosh, P. (2026). Medical Diagnosis Through Improved Feature Selection and Advanced Ensemble Techniques: A Study on Breast Cancer and Chronic Kidney Disease. *SN Computer Science*, 7(3), 222.
- Das, S., Chaudhuri, A. K., Das, S., & Ghosh, P. (2025). Multistage feature selection and stacked generalization model for cancer detection. *Scientific Reports*, 15(1), 38124.

Das, S., Chaudhuri, A. K., Paul, D., & Ghosh, P. (2026). Sequential Attribute Designator (SAD): A Novel Feature-Selection Framework for Pulmonary Disease Research. In *Next-Generation Bioinformatics for Pulmonary Disease Research* (pp. 413-436). IGI Global Scientific Publishing.

Doan, M., Sebastian, J. A., Caicedo, J. C., Siegert, S., Roch, A., Turner, T. R., ... & Carpenter, A. E. (2020). Objective assessment of stored blood quality by deep learning. *Proceedings of the National Academy of Sciences*, 117(35), 21381-21390.

Hong, B. Y., An, H. J., & Lim, S. H. (2022). Red cell distribution width and clinical implications. *Diagnostics*, 12(7), 1542.

Jiang, M., Shao, M., Yang, X., He, L., Peng, T., Wang, T., ... & Zhou, J. (2022). Automatic classification of red blood cell morphology based on quantitative phase imaging. *International Journal of Optics*, 2022(1), 1240020.

Lamoureux, E. S., Cheng, Y., Islamzada, E., Matthews, K., Duffy, S. P., & Ma, H. (2024). Biophysical profiling of red blood cells from thin-film blood smears using deep learning. *Heliyon*, 10(15).

Liu, D., Liu, H., Li, Y., Zhou, Q., & Huang, Z. (2025). Correlation between the HB-to-HCT ratio and 3-month outcomes in patients with acute ischemic stroke: A secondary analysis based on a prospective cohort study. *Frontiers in Neurology*, 16, 1616847.

Mangaonkar, A. A., Tande, A. J., & Bekele, D. I. (2021). Differential diagnosis and workup of monocytosis: A systematic approach to a common hematologic finding. *Current Hematologic Malignancy Reports*, 16(3), 267–275.

Mosley, J. D., Shelley, J. P., Dickson, A. L., Zanussi, J., Daniel, L. L., Zheng, N. S., & Kawai, V. K. (2023). Clinical consequences of a polygenic predisposition to benign lower white blood cell counts. *medRxiv*.

Peng, Y. F., Zhang, Q. S., & Luo, W. G. (2023). The clinical usefulness of mean corpuscular hemoglobin concentration in patients with pneumoconiosis. *International Journal of General Medicine*, 3171–3177.

Pinto, R. (2020). Objective assessment of stored blood quality by deep learning.

Ricos, C., Álvarez, V., Cava, F., García-Lario, J. V., Hernández, A., Jiménez, C. V., & Simón, M. (2014). Desirable biological variation database specifications.

Rizzuto, V., Mencattini, A., Álvarez-González, B., Di Giuseppe, D., Martinelli, E., Beneitez-Pastor, D., ... & Samitier, J. (2021). Combining microfluidics with machine learning algorithms for RBC classification in rare hereditary hemolytic anemia. *Scientific reports*, 11(1), 13553.

Roose, J., & Fagarasan, S. (2015). Editorial overview: Lymphocyte development and activation. *Current Opinion in Immunology*, 33, v–vi.

Routt, A. H., Yang, N., Piety, N. Z., Lu, M., & Shevkoplyas, S. S. (2023). Deep ensemble learning enables highly accurate classification of stored red blood cell morphology. *Scientific Reports*, 13(1), 3152.

Sarma, P. R. (1990). Red cell indices. In *Clinical methods: The history, physical, and laboratory examinations* (3rd ed.). Butterworths.

Shaheen, N., Thomas, S., Almoghairi, A., & Alaskar, A. (2025). Haematological parameters' reference intervals in apparently healthy individuals: A systematic review and meta-analysis. *Frontiers in Medicine*, 12, 1522492.

Song, Q., Wu, J. Z., Wang, S., & Chen, W. H. (2019). Elevated preoperative platelet distribution width predicts poor prognosis in esophageal squamous cell carcinoma. *Scientific Reports*, 9(1), 15234.

Tariq, S., Ismail, D., Thapa, M., Goriparthi, L., Pradeep, R., Khalid, K., & Jean-Charles, G. (2023). Chronic obstructive pulmonary disease and its effect on red blood cell indices. *Cureus*, 15(3).

Tepakhan, W., Srisintorn, W., Penglong, T., & Saelue, P. (2025). Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms. *Scientific Reports*, 15(1), 16917.

Tian, Y., Zong, Y., Pang, Y., Zheng, Z., Ma, Y., Zhang, C., & Gao, J. (2025). Platelets and diseases: Signal transduction and advances in targeted therapy. *Signal Transduction and Targeted Therapy*, 10(1), 159.

Wang, T. Q., Zhuo, Y., Lv, C. E., Shi, J., Yao, L. H., Zhang, S. Y., & Shi, J. (2025). Prediction of bacteremia using routine hematological and metabolic parameters based on logistic regression and random forest models. *Frontiers in Cellular and Infection Microbiology*, 15, 1605485.

Zhao, X., Yang, Y., Pan, Z., Lv, W., Rao, X., Wang, X., & Yu, X. (2024). Plateletcrit is predictive of clinical outcome and prognosis for early-stage breast cancer. *Cancer Medicine*, 13(2), e6944.