International Journal Software Engineering and Computer Science (IJSECS)
6 (2), 2026, 417-433
Published Online August 2026 in IJSECS (http://www.journal.lembagakita.com/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v6i2.7199.

RESEARCH ARTICLE Open Access

Classification of Anemia Severity Using Random Forest and SHAP Analysis on Complete Blood Count Data

Tira Julia Indah Sari [1*], **Andreas Perdana** [2].

[1*,2] Department of Informatics Engineering, Universitas Dharma Wacana, Metro City, Lampung Province, Indonesia.

**Corresponding author*: tiraindahsariii@gmail.com.

Received: May 12, 2026; Accepted: May 25, 2026; Published: August 1, 2026.

Abstract: Anemia remains a major global public health challenge, requiring accurate severity classification to support early diagnosis and clinical decision-making, particularly in resource-limited settings. This study aimed to develop an accurate and interpretable multiclass model for classifying anemia severity according to the 2011 WHO hemoglobin thresholds. A Random Forest model combined with SHAP (SHapley Additive exPlanations) analysis was developed and evaluated using a Complete Blood Count (CBC) dataset comprising 364 samples categorized into four classes: Normal, Mild Anemia, Moderate Anemia, and Severe Anemia. Data preprocessing included Interquartile Range (IQR) capping, label encoding, and StandardScaler standardization. The model was optimized through Grid Search Cross-Validation using 432 hyperparameter combinations and five-fold Stratified K-Fold validation. On the test set, the model achieved an accuracy of 95.89%, macro-precision of 96.61%, macro-recall of 97.14%, and macro-F1-score of 96.85%. The Moderate and Severe Anemia classes achieved F1-scores of 100%, although the result for Severe Anemia should be interpreted cautiously because of the limited number of test samples. Mild Anemia was the most challenging class, particularly for samples near the hemoglobin classification thresholds. SHAP analysis identified HGB, PCV, and RBC as the most influential features. However, the SHAP results should be interpreted in light of the correlations among these variables and the use of HGB as the basis for WHO severity labeling. The findings indicate that Random Forest combined with SHAP can accurately reproduce anemia severity classifications based on the applied labeling criteria while providing interpretable predictions. Further external and clinical validation is required before the model can be considered for use in a clinical decision support system.

Keywords: Anemia Severity Classification; Random Forest; SHAP; Explainable Artificial Intelligence; Complete Blood Count.

1. Introduction

Anemia is a medical condition characterized by low hemoglobin levels in the blood, which reduce the ability of red blood cells to deliver oxygen effectively throughout the body. It is recognized as a major global health problem affecting approximately one-third of the world's population (Nugraha, 2023). In Indonesia, the prevalence of anemia remains a major public health concern, particularly among vulnerable groups such as pregnant women, adolescents, and children. According to Indonesia's Basic Health Research, the prevalence of anemia increased sharply over a five-year period, from 37.01% in 2013 to 48.9% in 2018 (Nugraha, 2023).

The effects of anemia are not limited to physiological disorders but are also associated with adverse pregnancy and neonatal outcomes, including low birth weight, premature birth, intrauterine growth restriction, and neonatal mortality (Farhan & Dhanny, 2021). The high prevalence and serious consequences of anemia make early detection and severity classification essential for supporting effective clinical decision-making, especially given the uneven distribution of hematology specialists across Indonesia.

Anemia is commonly diagnosed through a Complete Blood Count (CBC), which includes various hematological parameters, such as hemoglobin (HGB) concentration, Mean Corpuscular Volume (MCV), red blood cell (RBC) count, hematocrit or Packed Cell Volume (PCV), and other related parameters (Rivera *et al.*, 2023). A study involving Filipino women reported that several CBC parameters, including hematocrit, Mean Corpuscular Hemoglobin (MCH), and MCV, demonstrated strong discriminative ability in detecting anemia (Rivera *et al.*, 2023). However, interpreting CBC results for anemia classification often requires expert clinical evaluation, while limited access to specialists may reduce diagnostic efficiency in clinical practice (Karagül Yıldız *et al.*, 2021). Machine learning-based computational methods may help address this limitation by supporting a more consistent and measurable classification process. Machine learning has shown considerable potential in anemia classification and hematological diagnosis using CBC parameters (Vohra *et al.*, 2022). Among the algorithms previously evaluated, including Support Vector Machine (SVM), Naive Bayes, and Extreme Gradient Boosting (XGBoost), Random Forest demonstrated the best performance for anemia classification (Zemariam *et al.*, 2024). Random Forest can process multidimensional data, reduce the risk of overfitting through an ensemble of decision trees, and achieve high accuracy when applied to structured tabular data (Breiman, 2001). Previous studies have reported strong performance from this algorithm. Zam and Irwan (2025), for example, reported an accuracy of 98%, while Gómez *et al.* (2025) reported an accuracy of up to 99.82%.

Despite the high predictive accuracy reported in previous studies, two technical gaps remain insufficiently addressed. First, many studies have focused primarily on binary classification between normal and anemic conditions without providing a detailed classification of anemia severity. Binary classification cannot distinguish among mild, moderate, and severe anemia, although these categories represent different levels of severity and may require different levels of clinical attention. Second, ensemble algorithms such as Random Forest are often regarded as black-box models because their prediction processes are difficult to interpret. This limitation reduces transparency when explaining how individual features contribute to prediction outcomes (Sadeghi *et al.*, 2024; Srinivasu *et al.*, 2022). SHAP (SHapley Additive exPlanations) is an Explainable Artificial Intelligence (XAI) method that estimates the contribution of each feature to a model prediction based on Shapley values (Lundberg & Lee, 2017). SHAP can be used to identify features that influence the model's overall behavior and to explain their contributions to individual predictions. Amalia (2025) reported that SHAP provided interpretable explanations for decision-making in anemia classification models. Its use may therefore help clarify the prediction process of Random Forest while retaining the model's ability to classify structured clinical data.

This study uses a publicly available secondary CBC dataset obtained from Mendeley Data, consisting of 364 samples (Ahmad *et al.*, 2026). The dataset was selected because it was derived from real-world clinical data and contains the CBC parameters required for the analysis. Its original labels consisted of two categories, namely Normal and Anemia. To support multiclass classification, the original labels were reclassified using the 2011 WHO hemoglobin thresholds into four categories: Normal, Mild Anemia, Moderate Anemia, and Severe Anemia (World Health Organization, 2011). Random Forest was then applied to classify these four severity categories using the available CBC parameters, while SHAP was used to examine the contribution patterns of the features included in the model. This combination was selected to balance predictive performance with the need for a technically interpretable classification process. However, because hemoglobin is used as both a predictor and the basis for determining the WHO severity labels, the contribution of HGB and related hematological features must be interpreted according to the labeling procedure used in this study.

This study aims to develop a Random Forest model with SHAP analysis for classifying anemia severity using CBC data. The intended outcome is a model that achieves reliable computational performance while allowing the contribution patterns of individual features to be interpreted. This research is an exploratory computational study intended to serve as an initial prototype for a clinical decision support system that may be further developed through appropriate external and clinical validation. The proposed model is not intended to replace clinical diagnosis but to support clinical decision-making through computational analysis. Specifically, this study aims to reclassify anemia severity into multiple classes based on the 2011 WHO standards, develop a Random Forest classification model using CBC data, analyze the contributions of CBC features using SHAP, and evaluate model performance using accuracy, precision, recall, and F1-score.

2. Related Work

2.1 State-of-the-Art Research

Recent studies have demonstrated the growing effectiveness of machine learning methods for anemia classification and prediction. Zemariam *et al*. (2024) evaluated eight supervised learning algorithms for classifying and predicting anemia among adolescent girls in Ethiopia, with Random Forest demonstrating the best performance among the tested algorithms. Karagül Yıldız *et al*. (2021) classified multiple types of anemia using several artificial intelligence methods and reported that ensemble decision tree models achieved higher classification accuracy than individual classifiers. Gómez *et al*. (2025) developed a multiclass anemia classification system based on Complete Blood Count (CBC) parameters to distinguish among non-anemia, microcytic anemia, normocytic anemia, and macrocytic anemia. Using several machine learning algorithms on a Kaggle dataset containing 1,421 instances, the Random Forest model achieved an accuracy of 99.82%. Machine learning models for anemia detection have also achieved high accuracy across several classifier configurations, including logistic regression, decision trees, Random Forest, and Support Vector Machine, further demonstrating the suitability of ensemble-based methods for blood parameter data (Ramzan *et al*., 2024).

CBC-based classification has also been applied to specific anemia subtypes. Pullakhandam and McRoy (2024) reported that gradient boosting combined with feature selection achieved a recall of 0.98 for iron deficiency anemia detection. Their SHAP analysis identified hemoglobin level, age, and red cell distribution width as the most influential predictors. Saputra *et al*. (2025) developed SmartForest, an ensemble model combining Random Forest and XGBoost, optimized using Sea Horse Optimization (SHO), with SHAP used to explain the classification of Beta-Thalassemia Trait (BTT) and Iron Deficiency Anemia (IDA). The model achieved an accuracy of 98.82%. In another application of Explainable Artificial Intelligence (XAI), Darshan *et al*. (2025) applied SHAP, LIME, and ELI5 to distinguish iron deficiency anemia from aplastic anemia, while Amalia (2025) used SHAP to provide interpretable feature attribution in anemia classification models.

2.2 Comparison with Previous Research

Table 1 compares previous studies with the present study in terms of methods, datasets or topics, main findings, and limitations. The comparison indicates that previous research has generally focused on binary anemia detection, anemia subtype classification, or multiclass morphological classification. Few studies have combined Random Forest and SHAP to classify anemia severity into multiple categories based on the 2011 WHO hemoglobin thresholds.

Table 1. Comparison of Previous Research with the Present Study

Researcher and Year	Method	Dataset/Topic	Key Findings	Limitations
Zam and Irwan (2025)	SVM and Random Forest	CBC-based binary anemia classification	Random Forest achieved 98% accuracy	Binary classification only; no XAI
Gómez *et al*. (2025)	Random Forest, Decision Tree, and LDA	Morphological anemia classification using 1,421 samples	Random Forest achieved 99.82% accuracy	Focused on morphological classification rather than anemia severity; no XAI
Saputra *et al*. (2025)	SmartForest: Random Forest, XGBoost, SHO, and SHAP	IDA versus BTT using 676 Indonesian samples	Achieved 98.82% accuracy; RBC was identified as a key feature	Binary classification only
Darshan *et al*. (2025)	Multiple machine learning algorithms and SHAP	IDA versus aplastic anemia using 500 samples	LightGBM achieved 96% accuracy; PLT, WBC, and RBC were identified as key features	Binary classification only
Amalia (2025)	Random Forest and SHAP	CBC-based anemia classification	SHAP explained the model predictions; HGB and MCV were the dominant features	Binary classification only
Vohra *et al*. (2022)	SVM, Random Forest, MLP, DT, and NB	Multiclass anemia severity classification using 364 samples	MLP achieved the best performance	No XAI analysis

Present study (2026)	Random Forest and SHAP	Multiclass anemia severity classification based on the 2011 WHO standards using 364 samples	Achieved 95.89% accuracy and a macro-F1-score of 96.85%	Small dataset, limited pediatric samples, and no clinical validation

Abbreviations: SVM = Support Vector Machine; RF = Random Forest; XGBoost = Extreme Gradient Boosting; SHO = Sea Horse Optimization; SHAP = SHapley Additive exPlanations; MLP = Multilayer Perceptron; DT = Decision Tree; NB = Naive Bayes; CBC = Complete Blood Count; IDA = Iron Deficiency Anemia; BTT = Beta-Thalassemia Trait; XAI = Explainable Artificial Intelligence; LDA = Linear Discriminant Analysis; PLT = platelet count; WBC = white blood cell count; RBC = red blood cell count; HGB = hemoglobin; MCV = Mean Corpuscular Volume.

2.3 Research Positioning

As shown in Table 1, this study addresses two gaps in previous research. Several studies focused on binary classification, including the classification of normal and anemic conditions or the differentiation of specific anemia subtypes (Amalia, 2025; Darshan *et al.*, 2025; Saputra *et al.*, 2025; Zam & Irwan, 2025). Although Gómez *et al.* (2025) used a multiclass approach, their study classified anemia according to morphological type rather than severity. Therefore, these studies did not classify anemia into the Normal, Mild, Moderate, and Severe categories specified by the WHO hemoglobin thresholds. Vohra *et al.* (2022) conducted multiclass severity classification using the same dataset, but their study did not include SHAP or another XAI method. The present study combines multiclass anemia severity classification based on the 2011 WHO standards with SHAP-based explanations. Random Forest is used to classify anemia severity, while SHAP is applied to examine the contribution of each CBC feature to the model's predictions. This approach is intended to produce accurate classifications accompanied by technically interpretable feature contributions.

2.4 Review of Technologies, Frameworks, and Algorithms

A Complete Blood Count is a standard laboratory test used to support the diagnosis of anemia. Rivera *et al.* (2023) demonstrated that CBC parameters such as hematocrit, with an Area Under the Curve (AUC) of 0.96; Mean Corpuscular Hemoglobin (MCH), with an AUC of 0.81; Mean Corpuscular Hemoglobin Concentration (MCHC), with an AUC of 0.80; and MCV, with an AUC of 0.77, had strong discriminatory ability for iron deficiency anemia. The WHO classification of anemia severity uses hemoglobin cutoff values adjusted according to age and sex, as presented in Table 2. The original WHO thresholds were reported in g/L and converted to g/dL in this study to improve clinical readability (World Health Organization, 2011).

Table 2. WHO (2011) Hemoglobin Thresholds for Anemia Severity Classification

Population Group	No Anemia (g/dL)	Mild Anemia (g/dL)	Moderate Anemia (g/dL)	Severe Anemia (g/dL)
Children aged 5–11 years	≥ 11.5	11.0–11.4	8.0–10.9	< 8.0
Children aged 12–14 years	≥ 12.0	11.0–11.9	8.0–10.9	< 8.0
Women aged ≥ 15 years	≥ 12.0	11.0–11.9	8.0–10.9	< 8.0
Men aged ≥ 15 years	≥ 13.0	11.0–12.9	8.0–10.9	< 8.0

Source: World Health Organization (2011).

Random Forest, introduced by Breiman (2001), is an ensemble learning algorithm that constructs multiple decision trees using bootstrap samples and combines their predictions through majority voting. Its main advantages include the ability to process high-dimensional data, reduce overfitting compared with a single decision tree, and estimate feature importance. These characteristics make Random Forest suitable for structured CBC data containing multiple related hematological parameters. SHAP is an XAI method based on Shapley values from cooperative game theory. It provides consistent feature-attribution explanations by estimating the contribution of each feature to an individual model prediction (Lundberg & Lee, 2017). The relevance of XAI in clinical applications was discussed by Sadeghi *et al.* (2024), who stated that explainability is necessary for AI-assisted decisions to gain ethical acceptance among healthcare professionals. Chaddad *et al.* (2023) identified SHAP as one of the widely used XAI methods in healthcare applications. A review by Aziz *et al.* (2024) also identified SHAP as a commonly adopted post hoc explanation method across various clinical domains, supporting its suitability for the present study. This study was implemented in Python 3.11 using Google Colaboratory, scikit-learn version 1.6.1, and SHAP version 0.51.0. In the proposed analytical process, CBC parameters serve as model inputs, Random Forest classifies anemia severity, and SHAP explains the contribution of each feature to the resulting predictions.

3. Methodology

3.1 Dataset and Research Design

This study employed a quantitative experimental design based on secondary data. The Complete Blood Count (CBC) dataset was obtained from Mendeley Data (Ahmad *et al.*, 2026) and consisted of 364 samples with 12 columns: Age, Sex, RBC, PCV, MCV, MCH, MCHC, RDW, TLC, PLT/mm³, HGB, and the binary target label Anemia. The dataset was selected because it originated from real-world clinical data and had previously been used as a reference in a similar study (Vohra *et al.*, 2022). The age distribution was dominated by adults aged 15 years or older, who accounted for 96.2% of the dataset, or 350 samples. The remaining data comprised 13 samples from individuals aged 12–14 years and one sample from an individual aged 11 years.

3.2 Tools and Technologies

The experiment was conducted in Google Colaboratory using Python 3.11. The main libraries used in this study were Pandas version 2.2.2 for tabular data manipulation and scikit-learn version 1.6.1 for RandomForestClassifier, GridSearchCV, StratifiedKFold, StandardScaler, LabelEncoder, and model evaluation metrics. SHAP version 0.51.0 was used for TreeExplainer and feature contribution analysis, while Matplotlib and Seaborn were used to visualize the data and model results. The libraries, versions, and main functions are summarized in Table 3.

Table 3. Libraries and Versions Used in the Study

Library	Version	Main Functions
Python	3.11	Programming language
Pandas	2.2.2	Tabular data manipulation
scikit-learn	1.6.1	Random Forest modeling, GridSearchCV, preprocessing, and evaluation
SHAP	0.51.0	TreeExplainer and feature contribution analysis
Matplotlib and Seaborn	—	Data and model result visualization

3.3 Rule-Based Labeling

The original binary target labels were converted into four anemia severity categories using rule-based labeling based on the HGB thresholds established by the World Health Organization (2011). The labeling procedure was individually adjusted according to the age group and sex of each sample. This procedure produced 168 Normal samples (46.2%), 88 Mild Anemia samples (24.2%), 92 Moderate Anemia samples (25.3%), and 16 Severe Anemia samples (4.4%). Figure 1 presents the resulting distribution of anemia severity classes, while Figure 2 shows the distribution of HGB values across the four classes and the WHO hemoglobin thresholds used as the basis for labeling.

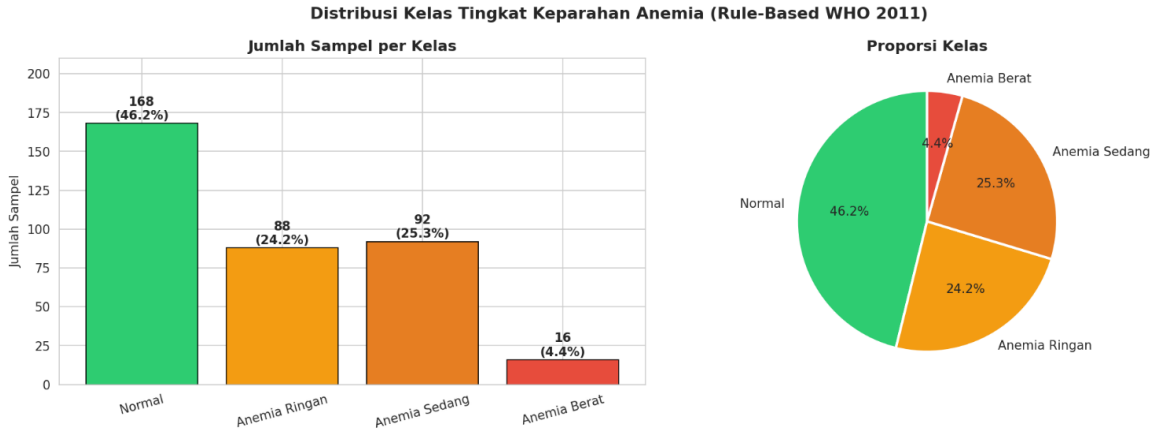

Figure 1. Distribution of anemia severity classes based on rule-based labeling according to the WHO standards (2011).

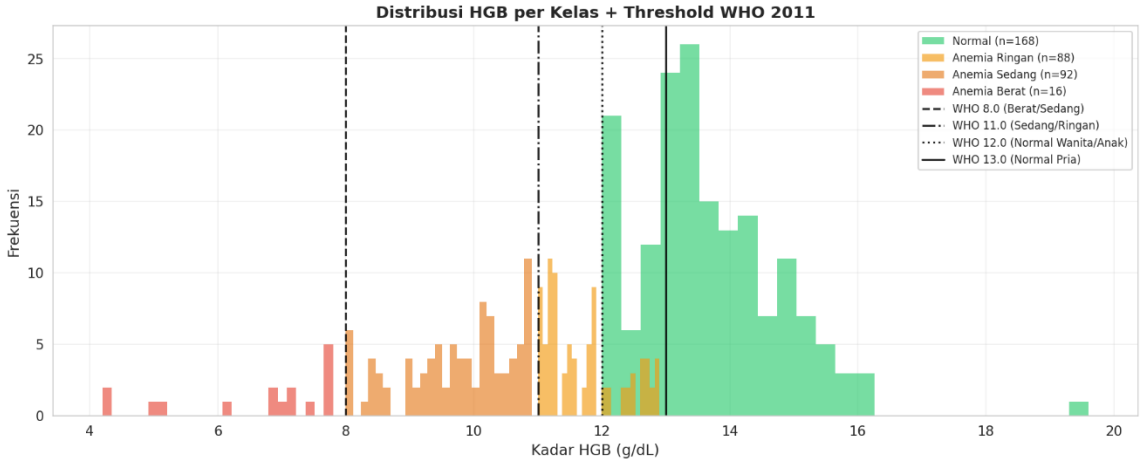

Figure 2. Distribution of HGB levels across anemia severity classes based on the WHO hemoglobin thresholds (2011).

3.4 Data Preprocessing

Data preprocessing was conducted in three main stages. The first stage involved checking for missing values. No missing values were identified in any of the features; therefore, no imputation procedure was required. In the second stage, outliers were handled using Interquartile Range (IQR) capping, also referred to in this study as winsorizing, on nine features: Age, RBC, PCV, MCV, MCH, MCHC, RDW, TLC, and PLT/mm³. HGB was excluded from the outlier-handling procedure because it served as the basis for determining the target labels. Modifying its values could create inconsistencies between the feature values and the assigned anemia severity categories. IQR capping was selected because it allowed all 364 samples to be retained without deleting observations identified as outliers. The third stage involved encoding and feature standardization. The Sex feature was encoded as Female = 0 and Male = 1. The numerical features were then standardized using StandardScaler, whose parameters were fitted only to the training data and subsequently applied to the testing data. This procedure was intended to prevent information from the testing data from influencing the transformation of the training data. Although Random Forest is generally insensitive to differences in feature scale, standardization was retained to ensure consistency in the preprocessing procedure. Figure 3 presents the Pearson correlation heatmap of the CBC features after preprocessing. The heatmap indicates strong correlations among HGB, PCV, and RBC, with correlation coefficients greater than 0.70. These relationships were considered when interpreting the SHAP feature contributions because correlated features may share related predictive information.

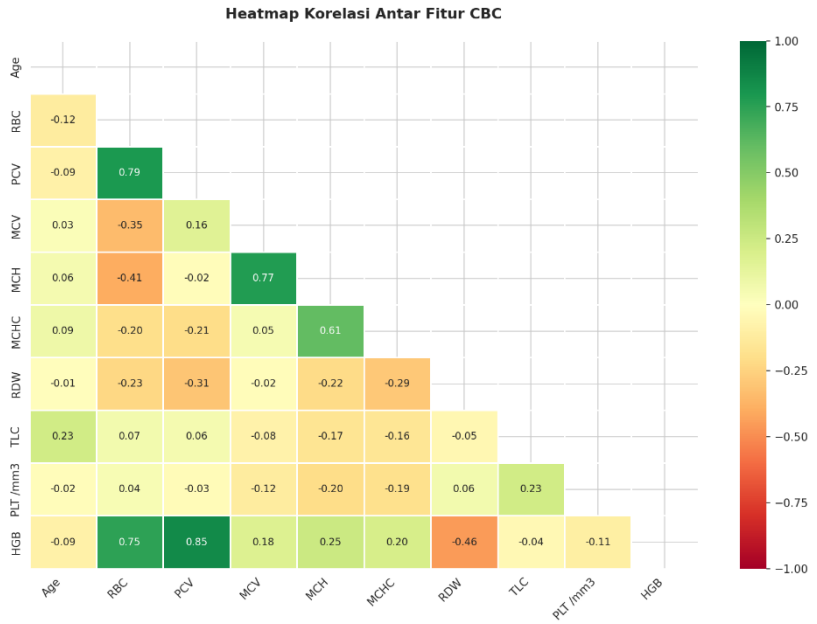

Figure 3. Pearson correlation heatmap of the CBC features after preprocessing.

3.5 Research Workflow

The research workflow consisted of ten stages. The first stage was a literature review conducted to establish the theoretical and methodological basis of the study. The second stage involved collecting the CBC dataset from Mendeley Data, followed by data preprocessing in the third stage, including the inspection of missing values and the handling of outliers. In the fourth stage, the original binary labels were converted into four severity categories through rule-based labeling based on the 2011 WHO hemoglobin thresholds. Exploratory Data Analysis (EDA) was then conducted in the fifth stage to examine the characteristics of the dataset through descriptive statistics, class distributions, and correlation analysis. In the sixth stage, the dataset was divided into training and testing subsets using stratified sampling with an 80:20 ratio. The seventh stage involved training the Random Forest classifier and optimizing its hyperparameters through Grid Search Cross-Validation. Model performance was evaluated in the eighth stage, followed by SHAP analysis in the ninth stage to interpret the contributions of the CBC features to the model predictions. The final stage involved documenting and visualizing the experimental results. The complete research workflow is illustrated in Figure 4.

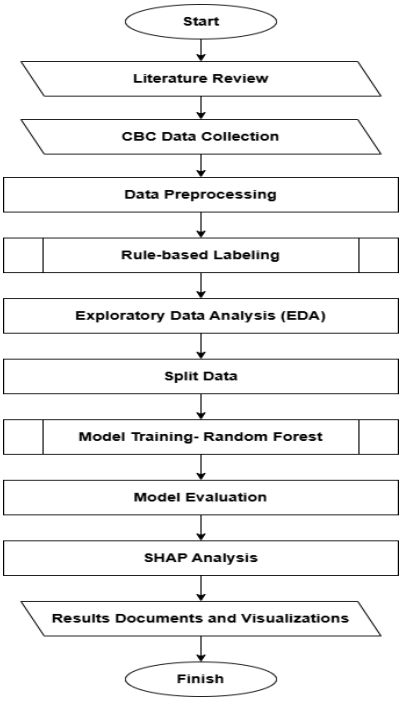

Figure 4. Research workflow.

3.6 Data Split and Model Training

The data were divided using a stratified train-test split with an 80:20 ratio and random_state=42, resulting in 291 training samples and 73 testing samples. Stratified sampling was selected to preserve the proportional distribution of each class in both subsets, particularly the minority Severe Anemia class. The resulting class distributions in the training and testing datasets are presented in Table 4.

Table 4. Class Distribution in the Training and Testing Datasets

Class	Training Samples	Testing Samples	Percentage
Normal	134	34	46.2%
Mild Anemia	70	18	24.2%
Moderate Anemia	74	18	25.3%
Severe Anemia	13	3	4.4%
Total	291	73	100%

The Random Forest classifier was selected because of its ability to process high-dimensional tabular medical data, its lower tendency to overfit than an individual decision tree, and its strong performance in previous anemia classification studies. The model was optimized using GridSearchCV with five-fold StratifiedKFold cross-validation. Macro-F1 was used as the optimization metric because it assigns equal importance to all classes, including the minority Severe Anemia class. The hyperparameter search covered 432 combinations. The candidate values consisted of n_estimators = {50, 100, 200}, max_depth = {None, 10, 20, 30}, min_samples_split = {2, 5, 10}, min_samples_leaf = {1, 2, 4}, max_features = {sqrt, log2}, and class_weight = {None, balanced}. Evaluating the 432 combinations through five-fold cross-validation resulted in 2,160

model-fitting processes during hyperparameter selection. The class_weight='balanced' configuration was included as an alternative for reducing the effect of class imbalance without generating synthetic data through oversampling techniques.

3.7 Model Evaluation and SHAP Analysis

The final model was evaluated using the 73 testing samples. The evaluation metrics included accuracy, macro-precision, macro-recall, and macro-F1-score, accompanied by a confusion matrix and a classification report for each class. Macro-averaged metrics were selected because they assign equal weight to every class rather than allowing the majority class to dominate the overall evaluation results. SHAP analysis was performed using TreeExplainer on all 73 testing samples. The analysis included global feature importance calculated using mean absolute SHAP values, multiclass SHAP bar plots, class-specific beeswarm summary plots, dependence plots for the most influential features, and waterfall plots for individual predictions. Global feature importance was used to rank the overall contributions of the CBC features, while the class-specific plots were used to examine how those contributions varied across anemia severity categories. The waterfall plots illustrated how the SHAP values and the expected model output combined to reconstruct the output for an individual prediction. SHAP analysis was restricted to the testing data so that the explanations represented model behavior on observations that were not used to train the classifier.

4. Result and Discussion

4.1 Results

4.1.1 Preprocessing Results

The missing-value inspection confirmed that all 364 samples were complete, with no missing values identified in any of the variables. The IQR capping procedure identified 129 outlying values, which were handled without removing any samples from the dataset, as reported in Table 5. Among the features with detected outliers, MCV had the highest number at 25, followed by MCHC at 22 and MCH at 21. The wide variation observed in these erythrocyte indices may reflect differences in red blood cell morphology, including patterns associated with hypochromic, microcytic, normocytic, or macrocytic conditions. However, because the dataset did not provide independent clinical verification for every outlying value, these observations should be interpreted as potentially meaningful clinical variation rather than definitive evidence that all outliers represented valid clinical conditions.

Table 5. Results of Outlier Handling Using the IQR Capping Method

Feature	Lower Limit	Upper Limit	Outliers
RBC	2.250	6.330	9
PCV	19.475	55.275	6
MCV	70.125	104.925	25
MCH	20.150	36.550	21
MCHC	25.800	37.800	22
RDW	9.850	19.850	13
TLC	−0.557	16.662	19
PLT/mm³	−19.875	441.125	14
Total			129

The negative lower IQR limits for TLC and PLT/mm³ were mathematical results of the IQR calculation and should not be interpreted as physiologically valid laboratory values. The IQR capping procedure was applied only to observations falling outside the calculated boundaries.

4.1.2 Hyperparameter Tuning Results

Grid Search Cross-Validation evaluated 432 hyperparameter combinations and identified the optimal configuration presented in Table 6. Five-fold cross-validation of the best-performing model produced an accuracy of 91.40% with a standard deviation of 2.91%, a macro-F1-score of 91.48% with a standard deviation of 4.41%, macro-precision of 93.09% with a standard deviation of 2.50%, and macro-recall of 90.67% with a standard deviation of 5.54%. The variation across folds was relatively limited, although macro-recall and macro-F1 showed greater variability than accuracy and macro-precision. This pattern indicates that model performance remained reasonably consistent across the five validation folds, while sensitivity to individual classes varied to some extent.

Table 6. Optimal Hyperparameters for the Random Forest Model

Hyperparameter	Optimal Value	Description
n_estimators	100	Number of decision trees
max_depth	None	No predefined maximum tree depth
min_samples_split	10	Minimum number of samples required to split an internal node
min_samples_leaf	2	Minimum number of samples required at a leaf node
max_features	sqrt	Number of features considered at each split
class_weight	balanced	Adjustment for class imbalance

4.1.3 Model Evaluation Results

Evaluation on the test set of 73 samples produced the results reported in Tables 7 and 8. The model achieved an accuracy of 95.89%, macro-precision of 96.61%, macro-recall of 97.14%, and macro-F1-score of 96.85%. The test-set accuracy was higher than the mean cross-validation accuracy of 91.40%. This difference may be associated with sampling variation because the test set contained only 73 observations. The relatively small cross-validation standard deviation indicates reasonably consistent performance across folds, but it does not independently rule out overfitting. Figure 5 presents the confusion matrix in count and row-normalized formats, allowing the classification errors and class-specific prediction patterns to be examined.

Table 7. Summary of Model Evaluation Results: Cross-Validation and Testing

Evaluation Metric	Five-Fold Cross-Validation	Test Set (n = 73)
Accuracy	91.40% ± 2.91%	95.89%
Macro-Precision	93.09% ± 2.50%	96.61%
Macro-Recall	90.67% ± 5.54%	97.14%
Macro-F1-Score	91.48% ± 4.41%	96.85%

Table 8. Classification Report by Class on the Test Set (n = 73)

Class	Precision	Recall	F1-Score	Support
Normal	96.97%	94.12%	95.52%	34
Mild Anemia	89.47%	94.44%	91.89%	18
Moderate Anemia	100.00%	100.00%	100.00%	18
Severe Anemia	100.00%	100.00%	100.00%	3
Macro-Average	96.61%	97.14%	96.85%	73

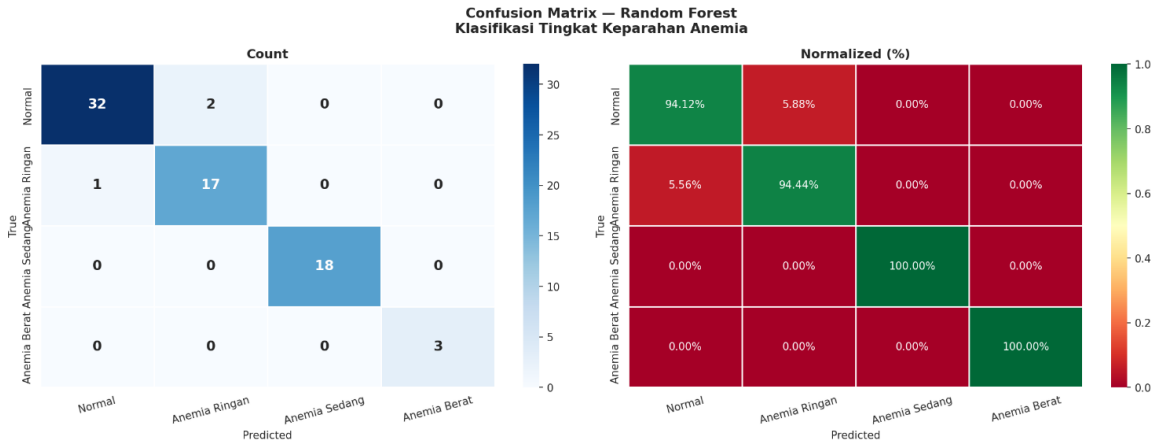

Figure 5. Confusion matrices for the Random Forest model in count format (left) and row-normalized format (right).

The Mild Anemia class recorded the lowest F1-score at 91.89%. Samples in this class occupy a narrow HGB interval adjacent to the threshold for the Normal class, as illustrated in Figure 2. Although the HGB ranges used for rule-based labeling do not overlap, samples near the classification boundary may have similar values for other CBC parameters, which can make the model's approximation of the labeling rules more difficult. The Moderate and Severe Anemia classes achieved F1-scores of 100% on the test set. Their HGB values and related hematological characteristics may have produced clearer separation in the available test data. Nevertheless, the perfect result for Severe Anemia should be interpreted cautiously because only three test samples belonged to this class, substantially limiting the statistical reliability and generalizability of the result. Vohra *et al.* (2022) provide a relevant point of comparison because their study used the same CBC dataset and a similar multiclass classification task. The Random Forest model in the present study achieved higher

performance than the Random Forest results reported in that study. The difference may be associated with the more extensive hyperparameter search involving 432 combinations and the inclusion of class_weight='balanced' to account for class imbalance. However, the contribution of these methodological choices cannot be established independently without a controlled comparison using identical data splits and evaluation procedures.

4.1.4 SHAP Analysis Results: Global Feature Importance

SHAP analysis was conducted using TreeExplainer on all 73 test samples. The resulting global feature rankings are reported in Table 9. HGB was ranked first, with a mean absolute SHAP value of 0.176493. This result was expected because hemoglobin concentration is the primary parameter used by the World Health Organization (2011) to define anemia and determine its severity, and HGB was also used to construct the multiclass target labels in this study. Its dominant SHAP contribution therefore reflects both its clinical role and its direct methodological relationship with the target variable. HGB was retained as a predictive feature because CBC results provide hemoglobin concentration together with the other hematological parameters. Excluding HGB would represent a different research question, namely whether anemia severity could be estimated indirectly from non-HGB parameters. Nevertheless, using HGB both to construct the labels and to train the classifier introduces methodological circularity. The performance results and SHAP rankings should therefore be interpreted primarily as evidence that the model learned the applied WHO-based labeling pattern rather than as independent confirmation that HGB is a newly identified predictor.

Table 9. Global Feature Rankings Based on Mean Absolute SHAP Values

Rank	Feature	Mean Absolute SHAP Value	Clinical Interpretation
1	HGB (Hemoglobin)	0.176493	Primary criterion in the 2011 WHO anemia classification and the main parameter used to construct the target labels
2	PCV (Packed Cell Volume/Hematocrit)	0.062447	Proportion of blood volume occupied by red blood cells; Rivera *et al*. (2023) reported an AUC of 0.96 for iron deficiency anemia discrimination
3	RBC (Red Blood Cell Count)	0.036827	Red blood cell count may differ between normal and anemic conditions and has demonstrated discriminatory value in anemia assessment (Rivera *et al*., 2023)
4	MCV (Mean Corpuscular Volume)	0.015530	Average red blood cell volume and a morphological indicator for microcytic, normocytic, and macrocytic patterns (Nugraha, 2023; Rivera *et al*., 2023)
5	MCH (Mean Corpuscular Hemoglobin)	0.015087	Average hemoglobin content per red blood cell; Rivera *et al*. (2023) reported an AUC of 0.81
6	RDW (Red Cell Distribution Width)	0.013433	Measure of variation in red blood cell size that may assist in differentiating anemia types (Nugraha, 2023)
7	MCHC (Mean Corpuscular Hemoglobin Concentration)	0.012281	Average hemoglobin concentration in red blood cells; Rivera *et al*. (2023) reported an AUC of 0.80
8	Age	0.006274	Used to determine the applicable WHO hemoglobin threshold for children and adults (World Health Organization, 2011)
9	PLT (Platelet Count)	0.005198	May provide supporting information in conditions involving bone marrow abnormalities, including aplastic anemia (Darshan *et al*., 2025)
10	TLC (Total Leukocyte Count)	0.003510	Represents white blood cell count and may provide supporting information on concurrent hematological or infectious conditions

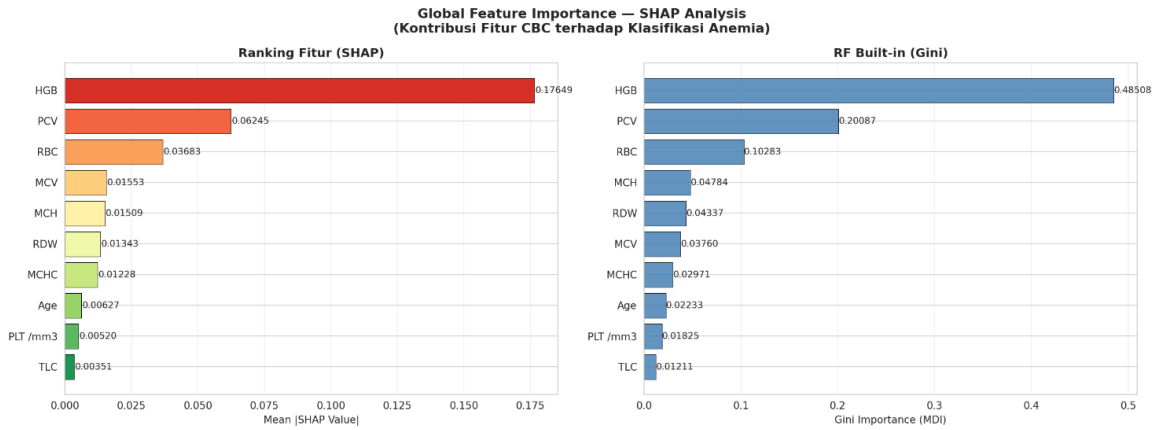

Figure 6. Comparison of global feature importance based on mean absolute SHAP values (left) and Random Forest Gini importance (right).

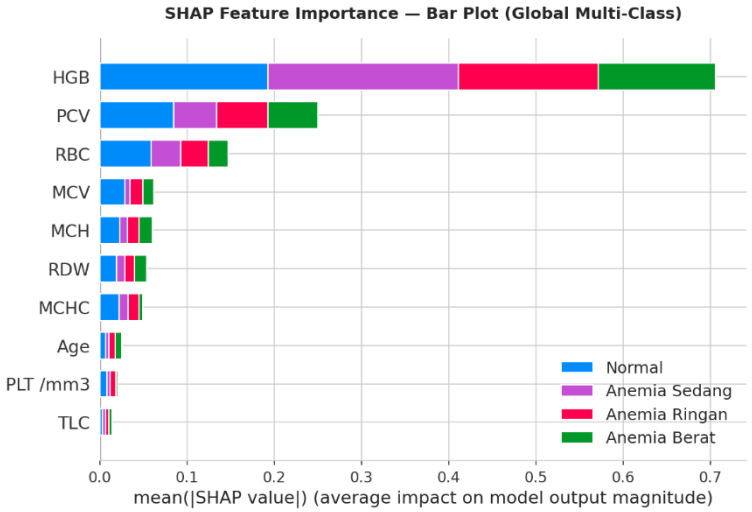

Figure 7. Global SHAP bar plot showing feature contributions across the four anemia severity classes.

Figure 6 shows that SHAP and Gini importance consistently ranked HGB, PCV, and RBC as the three most influential features, although the positions of MCV and MCH differed between the two methods. Figure 7 indicates that HGB contributed strongly across all classes, particularly to predictions for Moderate and Severe Anemia. Its contribution to Mild Anemia was smaller and more dispersed, which may reflect the narrower HGB range of this category and the similarity of other CBC characteristics near the boundary with the Normal class. PCV ranked second, with a mean absolute SHAP value of 0.062447. This ranking is consistent with Rivera *et al.* (2023), who reported an AUC of 0.96 for hematocrit in discriminating iron deficiency anemia. Age ranked eighth, with a mean absolute SHAP value of 0.006274. Although its contribution was smaller than that of the direct hematological measurements, age was relevant because it determined which WHO threshold was applied to each sample. Interpretation of these results should also account for the strong correlations among HGB, PCV, and RBC, which exceeded 0.70, as shown in Figure 3. Salih *et al.* (2025) noted that correlated features can share related predictive information, causing SHAP attribution to be distributed among them. Their individual SHAP values should therefore not be interpreted as fully independent effects.

4.1.5 SHAP Analysis Results: Summary Plots by Class

Figure 8 presents class-specific SHAP beeswarm plots displaying the direction, magnitude, and distribution of feature contributions across the four severity classes. The Normal class shows a relatively consistent pattern: high HGB values, shown in red, generally make positive contributions to the Normal prediction, whereas low HGB values, shown in blue, make negative contributions. This direction is consistent with the WHO hemoglobin thresholds used to assign the labels. The Mild Anemia class shows a more varied HGB contribution pattern. Values within the relatively narrow interval used for this class produce both positive and negative SHAP contributions depending on their relationship with other features and the decision paths formed by the model. This pattern may help explain why Mild Anemia had the lowest class-specific F1-score. The Moderate and Severe Anemia plots show clearer contributions from low HGB and RBC values toward their respective class predictions. These patterns correspond to the perfect F1-scores observed for these classes in

the test set, although the Severe Anemia pattern remains based on only three test samples and should not be generalized beyond the available data.

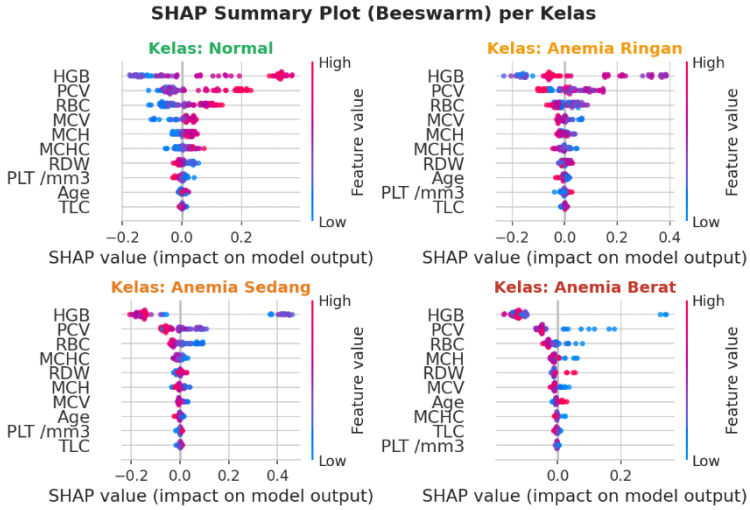

Figure 8. Class-specific SHAP beeswarm plots showing feature contributions to predictions for the four anemia severity classes.

4.1.6 SHAP Analysis Results: Dependence Plots

Figures 9 and 10 present SHAP dependence plots for the Normal and Mild Anemia classes. These classes were selected because the model produced its main classification errors near the boundary between them. For the Normal class, the HGB dependence plot shows a threshold-like pattern: values above the relevant WHO threshold generally contribute positively to the Normal prediction, whereas values below the threshold contribute negatively. This nonlinear pattern indicates that the Random Forest model formed decision rules that approximate the threshold structure of the labeling procedure. Although a basic logistic regression model assumes a linear relationship in the log-odds, nonlinear effects may also be represented in regression-based models through feature transformations and interaction terms. The Mild Anemia class displays a less uniform pattern of HGB contributions within its threshold interval, as shown in Figure 10. This variation may arise from the narrow HGB range assigned to Mild Anemia, interactions with age and sex thresholds, and the contribution of correlated CBC variables such as PCV and RBC. Dependence plots for Moderate and Severe Anemia are not included because the analysis focused on the classes associated with the observed classification errors. In addition, interpretation of the Severe Anemia dependence pattern would be limited by the small number of test samples.

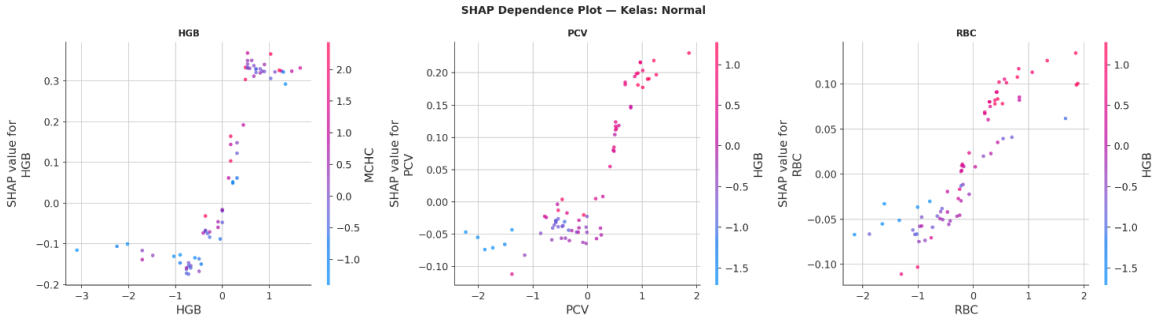

Figure 9. SHAP dependence plots for the Normal class based on HGB, PCV, and RBC.

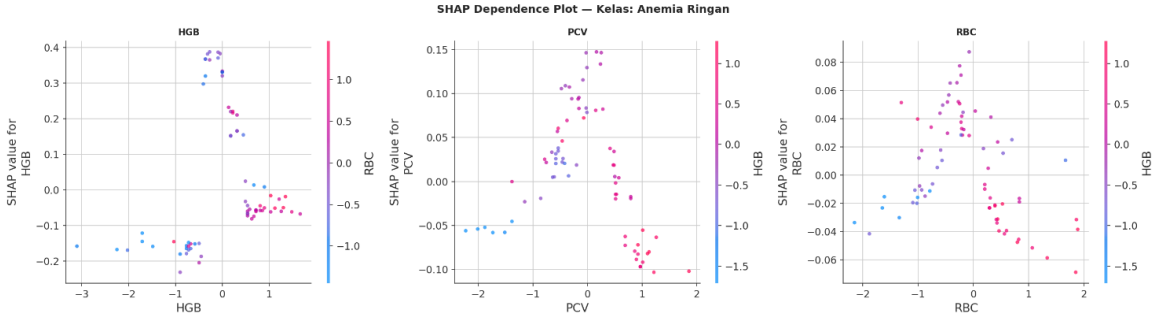

Figure 10. SHAP dependence plots for the Mild Anemia class based on HGB, PCV, and RBC.

4.1.7 SHAP Analysis Results: Local Accuracy

Figure 11 presents a SHAP waterfall plot for a representative sample from the Normal class. The plot shows how the base value and the SHAP contribution of each feature combine to reconstruct the model output for that individual observation, in accordance with the local accuracy property described by Lundberg and Lee (2017). The same additive property was observed in representative samples from the Mild, Moderate, and Severe Anemia classes. This result confirms that the generated SHAP explanations are mathematically consistent with the corresponding Random Forest outputs. It does not, however, independently verify that the model predictions are clinically correct. Rather, the plot explains how the model arrived at a particular output by showing the direction and magnitude of each feature contribution. This interpretation is consistent with previous studies that describe SHAP as a useful post hoc method for explaining model predictions in clinical applications (Chaddad *et al.*, 2023).

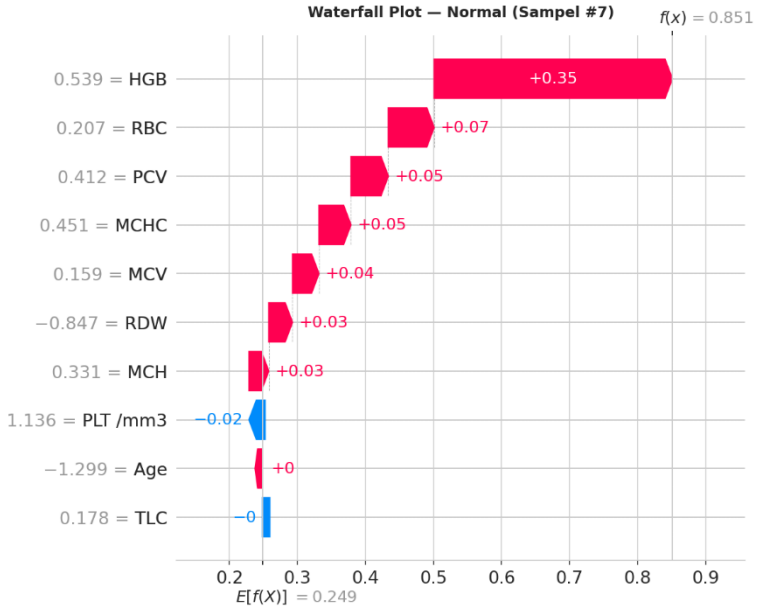

Figure 11. SHAP waterfall plot for a representative sample from the Normal class.

The combination of Random Forest and SHAP in a multiclass classification task based on the 2011 WHO thresholds produced high computational performance and feature-attribution patterns that were generally consistent with established hematological relationships. The agreement between the SHAP ranking and previous findings concerning HGB, PCV, and RBC supports the clinical plausibility of the model's behavior (Rivera *et al.*, 2023; World Health Organization, 2011). However, because HGB was also used to construct the severity labels, this agreement is partly determined by the study design and should not be interpreted as independent clinical validation. The results indicate that the model may be considered for further study as an initial component of a Clinical Decision Support System (CDSS), particularly for healthcare facilities with limited access to specialists (Sadeghi *et al.*, 2024; Srinivasu *et al.*, 2022). External validation and direct assessment by qualified healthcare professionals remain necessary before implementation in diagnostic practice.

4.2 Discussion

The proposed Random Forest model demonstrated strong performance in classifying anemia severity into four categories. Five-fold cross-validation produced a mean accuracy of 91.40% and a macro-F1-score of 91.48%, while evaluation on the test set produced an accuracy of 95.89% and a macro-F1-score of 96.85%. The standard deviation of cross-validation accuracy was 2.91%, indicating relatively limited variation across the five data partitions. Macro-recall and macro-F1 showed somewhat greater variation, suggesting that class-specific sensitivity was more affected by changes in fold composition. These results indicate that the model performed consistently within the available dataset, although external data would be required to assess whether the same performance can be maintained in other clinical populations.

The Mild Anemia class obtained the lowest F1-score at 91.89%. This result may be related to its relatively narrow HGB range and its position adjacent to the Normal threshold established by the World Health Organization (2011). As shown in Figures 2 and 10, samples near this boundary may have similar PCV, RBC, and other CBC characteristics even though their rule-based HGB labels do not overlap. Consequently, the Random Forest model may have greater difficulty approximating the boundary for Mild Anemia than for classes defined by lower and more widely separated HGB values. The Moderate and Severe Anemia classes achieved F1-scores of 100% in the test set, suggesting that the available samples were clearly separated by the learned

decision rules. Nevertheless, the Severe Anemia result was based on only three test samples and cannot provide a statistically reliable estimate of performance in a broader population.

Vohra *et al*. (2022) provide a relevant benchmark because they used the same 364-sample CBC dataset and a similar WHO-based multiclass classification task. Their study reported that the Multilayer Perceptron model achieved the best performance among the six evaluated algorithms, whereas the Random Forest model developed in the present study achieved a macro-F1-score of 96.85%. The difference in Random Forest performance may be associated with the broader hyperparameter search and the use of class_weight='balanced'; however, this interpretation remains tentative because direct comparison requires an identical train-test split, random seed, preprocessing procedure, and evaluation protocol.

Zemariam *et al*. (2024) similarly identified Random Forest as the best-performing model among eight algorithms for anemia prediction. Their study, however, used a binary classification task based on a large sociodemographic survey dataset containing 5,642 observations and reported an AUC of 0.82 as the primary metric. Direct numerical comparison is therefore inappropriate because the classification task, predictors, population, sample size, and evaluation measures differ from those used in the present study. Gómez *et al*. (2025) reported an accuracy of 99.82%, but their model classified morphological anemia categories rather than WHO-based severity levels. These differences show that performance values should be interpreted according to the target definition and evaluation design of each study rather than compared solely on the basis of accuracy.

Unlike studies that focused only on predictive performance or binary classification, the present study combines a WHO-based multiclass severity task with SHAP analysis. This combination provides class predictions together with estimates of how each CBC feature contributes to the model output. HGB was identified as the most influential feature because it directly forms the basis of the WHO diagnostic thresholds and was used to construct the target labels. PCV and RBC also made substantial contributions because of their physiological relationships with hemoglobin concentration and red blood cell characteristics. The feature rankings were broadly consistent with findings reported by Rivera *et al*. (2023). Evidence from a wider hematological setting also indicates that ensemble models applied to routine laboratory data frequently identify CBC indices such as HGB, MCV, and RBC among the dominant SHAP contributors across several disease categories (Liu *et al*., 2025).

The interpretation of these rankings requires two methodological qualifications. First, using HGB both as a predictor and as the basis for target construction creates circularity. The high importance of HGB and the strong classification results mainly demonstrate that Random Forest can approximate the WHO-based labeling rules from the available CBC data. They do not establish that the model predicts an independently assessed clinical diagnosis. Second, the strong correlations among HGB, PCV, and RBC suggest that some SHAP attribution may represent shared predictive information distributed among correlated variables rather than independent effects. The rankings should therefore be interpreted as explanations of model behavior, not as estimates of causal or isolated clinical influence.

From a practical perspective, combining Random Forest with SHAP in a WHO-based multiclass task may provide a preliminary basis for developing a Clinical Decision Support System, particularly in primary healthcare settings with limited access to hematology specialists. SHAP summary plots, dependence plots, and waterfall plots allow users to inspect the variables associated with the model's predictions rather than relying only on the predicted class. Such explanations may help healthcare personnel assess whether a prediction follows a clinically plausible pattern. Nevertheless, SHAP does not convert Random Forest into a fully transparent model, nor does it establish the clinical validity of its predictions. Its role is limited to providing post hoc explanations of the model output. Several limitations must be considered. First, the dataset was relatively small, especially for Severe Anemia, which was represented by only 16 samples overall and three samples in the test set. Second, the study relied on a single publicly available dataset, which may restrict the generalizability of the results to populations with different demographic, clinical, or laboratory characteristics. Third, 96.2% of the observations were obtained from individuals aged 15 years or older, limiting the reliability of the findings for children. Fourth, the target labels were constructed exclusively from WHO hemoglobin thresholds and were not supported by additional clinical indicators or independent annotations from physicians. Fifth, HGB served as both the principal labeling variable and a model input, making the classification task partly circular. Future studies should conduct external validation using larger multicenter datasets with more balanced representation across severity classes and age groups. Models trained without HGB should also be evaluated to determine whether other CBC parameters can estimate anemia severity independently. Additional experiments could compare class_weight='balanced' with resampling methods and examine alternative algorithms under the same data partitions and evaluation procedures. Direct assessment by hematologists and other qualified healthcare professionals is also required to evaluate clinical reliability, interpretability, and usability before real-world deployment is considered.

5. Conclusion and Recommendations

This study developed a multiclass model for anemia severity classification using Random Forest combined with SHAP analysis on Complete Blood Count data. The optimal hyperparameters were selected from 432 combinations through grid search with five-fold cross-validation. On the test set, the final model achieved an accuracy of 95.89% and a macro-F1-score of 96.85%. The Moderate and Severe Anemia classes each achieved an F1-score of 100% on the available test data. SHAP analysis identified HGB, with a mean absolute SHAP value of 0.176; PCV, with a value of 0.062; and RBC, with a value of 0.037, as the three most influential features. These rankings were generally consistent with the 2011 WHO hemoglobin criteria and the hematological relationships reported in previous studies. However, the dominant contribution of HGB was also directly affected by its use as the basis for constructing the target labels.

This study offers two main contributions. First, it combines Random Forest and SHAP in a multiclass anemia severity classification task based on the 2011 WHO criteria, an approach that remains limited in previous research. Second, the study shows that the model can produce feature-attribution patterns that are broadly consistent with established hematological knowledge. The SHAP explanations allow the contribution of each CBC parameter to be examined at both global and individual prediction levels. These results provide a preliminary basis for the further development of an interpretable Clinical Decision Support System (CDSS) for healthcare professionals. Nevertheless, the model should be regarded as an exploratory computational prototype rather than a clinically validated diagnostic system.

Several limitations must be acknowledged. The dataset was relatively small, comprising only 364 samples, with limited representation of Severe Anemia cases. This class contained 16 samples overall and only three samples in the test set. Consequently, the F1-score of 100% for Severe Anemia was obtained from a very small number of observations and has limited statistical reliability. In addition, adults accounted for 96.2% of the dataset, meaning that model performance among children aged 5–14 years could not be adequately evaluated. The use of a single publicly available dataset also limits the generalizability of the findings to populations with different demographic, clinical, and laboratory characteristics.

Methodological circularity represents another important limitation because HGB served both as the basis for constructing the anemia severity labels and as a predictive feature. This relationship limits the independent interpretation of HGB dominance in the SHAP analysis and indicates that the model primarily learned to reproduce the applied WHO-based labeling rules. Strong correlations among HGB, PCV, and RBC, with correlation coefficients above 0.70, may also have caused shared predictive information to be distributed across these variables in the SHAP results (Salih *et al.*, 2025). Their individual SHAP values should therefore not be interpreted as independent or causal clinical effects. Finally, this study remains exploratory and has not undergone external validation or direct clinical evaluation by qualified healthcare professionals.

Future studies should use larger and more diverse datasets with adequate representation of Severe Anemia cases to obtain more reliable estimates of class-specific performance. More proportional representation of children aged 5–14 years is also needed to evaluate model performance across different age groups and WHO threshold categories. External validation using multicenter data should be conducted to determine whether the model can maintain its performance across healthcare institutions, patient populations, and laboratory settings. Models developed without HGB should also be examined to determine whether the remaining CBC parameters can estimate anemia severity independently rather than reproduce labels directly derived from hemoglobin levels.

Further research may assess carefully selected feature-engineering methods, including clinically justified ratios, while considering whether the resulting variables retain a clear hematological meaning. Dimensionality-reduction methods such as Principal Component Analysis may also be evaluated, although their effects on clinical interpretability should be examined because transformed components may be less straightforward than the original CBC variables. Approaches for handling class imbalance should be compared under identical data partitions and evaluation procedures, including class_weight='balanced', SMOTE, and ADASYN. Any oversampling procedure must be applied only to the training folds to prevent data leakage. SHAP may also be compared with other explanation methods, such as LIME and ELI5, by evaluating the stability, consistency, and clinical plausibility of their explanations. Finally, external and prospective validation in collaboration with hematologists remains necessary to assess the model's clinical reliability, usability, and safety before its implementation in a practical CDSS can be considered (Srinivasu *et al.*, 2022).

References

Ahmad, M. H., Bormon, S., & Hossain, M. A. (2026). *Expert-labeled clinical dataset for anemia detection* (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/8sycpbhzn7.2

Amalia, I. (2025). Anemia classification with clinical feature engineering and SHAP interpretation. *Journal of Applied Informatics and Computing (JAIC), 9*(5), 2195–2201. https://doi.org/10.30871/jaic.v9i5.10912

Aziz, N. A., Manzoor, A., Mazhar Qureshi, M. D., Qureshi, M. A., & Rashwan, W. (2024). Unveiling explainable AI in healthcare: Current trends, challenges, and future directions [Preprint]. *medRxiv*. https://doi.org/10.1101/2024.08.10.24311735

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors, 23*(2), Article 634. https://doi.org/10.3390/s23020634

Darshan, B. S. D., Sampathila, N., Bairy, G. M., Prabhu, S., Belurkar, S., Chadaga, K., & Nandish, S. (2025). Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable artificial intelligence utilizing blood attributes. *Scientific Reports, 15*(1), Article 505. https://doi.org/10.1038/s41598-024-84120-w

Farhan, K., & Dhanny, D. R. (2021). Anemia ibu hamil dan efeknya pada bayi. *Muhammadiyah Journal of Midwifery, 2*(1), 27–33. https://doi.org/10.24853/myjm.2.1.27-33

Gómez, J. G., Parra Urueta, C., Álvarez, D. S., Hernández Riaño, V., & Ramirez-Gonzalez, G. (2025). Anemia classification system using machine learning. *Informatics, 12*(1), Article 19. https://doi.org/10.3390/informatics12010019

Karagül Yıldız, T., Yurtay, N., & Öneç, B. (2021). Classifying anemia types using artificial learning methods. *Engineering Science and Technology, an International Journal, 24*(1), 50–70. https://doi.org/10.1016/j.jestch.2020.12.003

Liu, J., Gou, Y., Yang, W., Wang, H., Zhang, J., Wu, S., Liu, S., Tao, T., Tang, Y., Yang, C., Chen, S., Wang, P., Feng, Y., Zhang, C., Liu, S., Peng, X., & Zhang, X. (2025). Development and application of machine learning models for hematological disease diagnosis using routine laboratory parameters: A user-friendly diagnostic platform. *Frontiers in Medicine, 12*, Article 1605868. https://doi.org/10.3389/fmed.2025.1605868

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

Nugraha, G. (2023). *Mengenal anemia: Patofisiologi, klasifikasi, dan diagnosis*. Penerbit BRIN. https://doi.org/10.55981/brin.906.c799

Pullakhandam, S., & McRoy, S. (2024). Classification and explanation of iron deficiency anemia from complete blood count data using machine learning. *BioMedInformatics, 4*(1), 661–672. https://doi.org/10.3390/biomedinformatics4010036

Ramzan, M., Sheng, J., Saeed, M. U., Wang, B., & Duraihem, F. Z. (2024). Revolutionizing anemia detection: Integrative machine learning models and advanced attention mechanisms. *Visual Computing for Industry, Biomedicine, and Art, 7*(1), Article 18. https://doi.org/10.1186/s42492-024-00169-4

Rivera, A. K. B., Latorre, A. A. E., Nakamura, K., & Seino, K. (2023). Using complete blood count parameters in the diagnosis of iron deficiency and iron deficiency anemia in Filipino women. *Journal of Rural Medicine, 18*(2), 79–86. https://doi.org/10.2185/jrm.2022-047

Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering, 118*, Article 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems, 7*(1), Article 2400304. https://doi.org/10.1002/aisy.202400304

Saputra, D. C. E., Sunat, K., Ratnaningsih, T., Ma'arif, A., & Suwarno, I. (2025). SmartForest: Optimizing anemia diagnosis with enhanced random forest and XGBoost integration. *Results in Engineering, 27*, Article 106442. https://doi.org/10.1016/j.rineng.2025.106442

Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., & Raut, R. (2022). From black box to explainable AI in healthcare: Existing tools and case studies. *Mobile Information Systems, 2022*, Article 8167821. https://doi.org/10.1155/2022/8167821

Vohra, R., Hussain, A., Dudyala, A. K., Pahareeya, J., & Khan, W. (2022). Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting. *PLOS ONE, 17*(7), Article e0269685. https://doi.org/10.1371/journal.pone.0269685

World Health Organization. (2011). *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity* (WHO/NMH/NHD/MNM/11.1). https://iris.who.int/handle/10665/85839

Zam, N. W., & Irwan, M. (2025). Klasifikasi machine learning untuk anemia menggunakan metode support vector machine dan random forest. *Jurnal Edukasi dan Sains Matematika (JES-MAT), 11*(1), 62–76. https://journal.uniku.ac.id/index.php/JESMath/article/view/11694

Zemariam, A. B., Yimer, A., Abebe, G. K., Wondie, W. T., Abate, B. B., Alamaw, A. W., Yilak, G., Melaku, T. M., & Ngusie, H. S. (2024). Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia. *Scientific Reports, 14*(1), Article 9080. https://doi.org/10.1038/s41598-024-60027-4.