ISSN 2579-7069

A Comparative Analysis of Machine Learning Classifier of Anemia Diagnosis Based on Complete Blood Count (CBC) Data

Nadya Awalia Putri[1,*], Bayu Priya Mukti[2]

[1,2]*Magister of Computer Science, Amikom Purwokerto University*

(Received March 7, 2025; Revised July 5, 2025; Accepted October 3, 2025; Available online December 28, 2025)

Abstract

Anemia is a prevalent hematological condition that requires accurate and timely diagnosis to ensure effective treatment. This study aims to compare the performance of several machine learning algorithms Random Forest, Support Vector Machine (SVM), Naive Bayes, and XGBoost in classifying different types of anemia based on Complete Blood Count (CBC) data. The dataset includes three diagnostic categories: Healthy, Normocytic hypochromic anemia, and Normocytic normochromic anemia. After preprocessing and normalization, each model was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The results show that XGBoost achieved the highest overall performance with 99% accuracy and a perfect AUC of 1.00, followed closely by SVM and Naive Bayes. Naive Bayes showed lower performance, particularly in identifying normocytic normochromic anemia. These findings suggest that machine learning, especially ensemble-based models, holds strong potential in supporting clinical diagnosis of anemia using CBC data.

Keywords: Anemia classification, Complete Blood Count (CBC), Machine Learning, XGBoost, SVM, Naive Bayes, Medical Diagnosis

1. Introduction

Anemia is one of the most common blood disorders, estimated to affect approximately 1.6 billion people worldwide [1]. According to the World Health Organization (WHO), anemia is a condition in which the number of red blood cells and the blood's ability to carry oxygen are insufficient to meet the body's physiological needs [2]. It is also defined as a decrease in red cell mass, hemoglobin, and hematocrit levels in the blood. Normal values for hemoglobin and hematocrit vary based on age and gender. If these values fall below the normal thresholds for a specific age and sex group, the individual is considered anemic.

Hemoglobin levels are influenced by various factors such as age, sex, pregnancy status, health condition, as well as genetic and environmental factors. For instance, normal Hb levels in newborns range between 170–210 g/L, decrease to approximately 100 g/L by 6–9 months of age, and then gradually rise to 110 g/L for children up to 59 months, 115 g/L for children up to 11 years old, 120 g/L for adult females, and around 130 g/L for adult males [2].

Common symptoms of anemia include fatigue, palpitations, headache, shortness of breath, and pallor of the conjunctiva and palms. Although these symptoms may serve as initial indicators, they have low sensitivity and moderate specificity, and therefore cannot be relied upon for definitive diagnosis [3]. These symptoms function more as early warning signs, particularly in contexts with limited access to laboratory testing.

A study by Kassebaum et al. [4], which analyzed data from 189 countries and included multiple age groups and both sexes using WHO data from the 2010 Global Burden of Disease Study, estimated the global prevalence of anemia at 32.9%. This condition is most commonly found in children under five years of age and in women, with Iron Deficiency Anemia (IDA) being the most frequently encountered type. Anemia has a multifactorial etiology, meaning it can be caused by various factors.

*Corresponding author: Nadya Awalia Putri (nadya@amikompurwokerto.ac.id)
DOI: https://doi.org/10.47738/ijiis.v8i4.286

In developing countries, iron deficiency remains the primary cause. In Colombia, for example, anemia is strongly associated with malnutrition due to poverty. In 2022, Del Castillo et al. [5] was reported that 18.3 million people were living in poverty and 6.9 million in extreme poverty, despite a 3.1% decrease compared to the previous year.

Medically, anemia can be classified based on its cause and the morphology of red blood cells. Microcytic anemia is commonly caused by iron deficiency, normocytic anemia is often associated with inflammation or chronic disease, while macrocytic anemia is typically caused by vitamin B12 or folate deficiency. Anemia can also be categorized by its severity as mild, moderate, or severe [6].

According to WHO reports, the global prevalence of anemia remains high, with around 29.9% of women aged 15–49 and 39.8% of children aged 6 to 59 months being affected. In Africa, the prevalence among children in the same age group is as high as 60.2% [7]. If left undiagnosed and untreated, anemia can lead to serious complications.

Given the widespread impact of anemia on public health and quality of life, early detection and timely intervention are essential. However, traditional diagnostic methods still pose significant challenges, especially for low-income populations. Limited access to quality healthcare services and the high cost of laboratory testing often prevent individuals from obtaining timely diagnoses [8]. This creates a vicious cycle where untreated anemia leads to worsening complications and a marked decline in patients' quality of life.

In the digital era, machine learning has emerged as a powerful approach for developing clinical decision-support systems, including for anemia detection and diagnosis. Several studies have demonstrated that machine learning–based predictive models can enhance both the efficiency and accuracy of medical diagnoses while also optimizing resource allocation through early warning systems that enable faster clinical response [9]. By reducing subjectivity in clinical interpretation, these technologies offer more consistent and reliable diagnostic outcomes.

For instance, a study by Pullakhandam and McRoy [10] demonstrated that machine learning algorithms could accurately differentiate IDA from non-IDA using CBC data from NHANES (2003–2020). The models incorporated Explainable AI (XAI) techniques to interpret the contribution of hematological features, along with data balancing strategies such as random oversampling. The findings support the potential of machine learning for non-invasive anemia diagnosis, although further development is still needed to improve model interpretability and clinical applicability [11].

Based on this context, this study aims to compare the performance of several machine learning algorithms in classifying anemia types based on Complete Blood Count (CBC) data. The classification is focused on four categories: Healthy, Normocytic hypochromic anemia and Normocytic normochromic anemia. This approach seeks to develop an intelligent system that not only accelerates the diagnostic process but also supports more precise and personalized clinical decision-making, thereby potentially improving patient outcomes.

By adopting a comparative approach using four algorithms, Support Vector Machine (SVM), Naïve Bayes and XGBoost, this research aims to identify the most effective model for anemia classification based on CBC data. Additionally, this study contributes to expanding the application of artificial intelligence in healthcare, particularly in areas with limited access to medical services.

2. Literature Review

A study conducted by Airlangga utilized CBC data to evaluate the performance of several machine learning algorithms for anemia classification. The models tested included Decision Tree, Extra Tree, Random Forest, ExtraTrees Regressor, XGBoost, LightGBM, and CatBoost. Among these seven models, the Decision Tree algorithm achieved the best performance, with a balanced accuracy of 94.17%, outperforming more complex ensemble models such as XGBoost and Random Forest. The study highlights that simpler models can produce competitive results when supported by effective feature selection and appropriate data preprocessing.

These findings align with a study by Abdul-Jabbar et al. [12], who performed a comparative analysis of twelve classification algorithms to diagnose anemia using two CBC datasets, one international dataset and a newly developed dataset collected from several hospitals in Iraq. The results showed that LogitBoost, Random Forest, XGBoost, and

Multilayer Perceptron achieved the highest accuracies, with XGBoost emerging as the top-performing algorithm. This study emphasized that algorithm selection should be aligned with the characteristics of the dataset to achieve more systematic and comprehensive results. It also suggested exploring ensemble approaches to combine the strengths of different models while addressing their limitations. Overall, the study supports the use of machine learning and digital analytics tools in healthcare for generating valuable insights from large-scale datasets, ultimately improving diagnostic accuracy and advancing medical understanding.

Furthermore, Végh et al. [13] conducted a comparative evaluation of nine machine learning classification models, including Ensemble methods, Decision Tree, SVM, K-Nearest Neighbor (KNN), Naïve Bayes, Neural Networks, Kernel methods, and Discriminant Analysis, using CBC data. All models were optimized using Bayesian hyperparameter tuning, trained on 90% of the data and tested on the remaining 10%, with a 10-fold cross-validation for robustness. The results showed that Ensemble (bagging) models achieved a validation accuracy of 99.22% and test accuracy of 100%, followed closely by Tree-based models (99.05% / 100%). Further analysis using permutation feature importance revealed that hemoglobin was the most influential feature for anemia classification.

In addition, Kabootarizadeh et al. [14] proposed an Artificial Neural Network (ANN)-based model to accurately and efficiently differentiate between IDA and β-thalassemia trait (β-TT). The study utilized CBC test data from 268 patients and tested various ANN configurations using MATLAB to determine the optimal architecture. The results showed that a multilayer ANN with four input features and 70 neurons achieved high performance, with an accuracy of 92.5%, sensitivity of 93.13%, and specificity of 92.33%.

3. Methodology

3.1. Dataset Description

The dataset used in this study was obtained from an open-source platform on Kaggle, titled “Anemia Types Classification,” which is accessible at the following link: https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification. This dataset contains 1,281 patient records based on CBC test results, which have been labeled according to anemia diagnoses made by professional medical personnel. The data were collected from various healthcare facilities and include commonly used hematological laboratory parameters to analyze the blood condition of patients, as summarized in Table 1.

Table 1. Dataset Description

Attribute	*Description*
WBC	White Blood Cell count ($10^3/\mu L$)
LYMp	Lymphocyte percentage (%)
NEUTp	Neutrophil percentage (%)
LYMn	Absolute lymphocyte count ($10^3/\mu L$)
NEUTn	Absolute neutrophil count ($10^3/\mu L$)
RBC	Red Blood Cell count ($10^6/\mu L$)
HGB	Hemoglobin concentration (g/dL)
HCT	Hematocrit (%)
MCV	Mean Corpuscular Volume (fL)
MCH	Mean Corpuscular Hemoglobin (pg)
MCHC	Mean Corpuscular Hemoglobin Concentration (g/dL)
PLT	Platelet count ($10^3/\mu L$)
PDW	Platelet Distribution Width (%)
PCT	Plateletcrit (%)
Diagnosis	Target variable representing anemia type

3.2. Pre- Processing

In this study, the anemia classification data underwent several preprocessing steps to ensure that the dataset met quality and consistency standards before being applied to machine learning models. One of the key steps in this process was normalizing the numerical features using the Min-Max Scaling method [15], [16] as shown in equation (1):

$$Normalized\ z = \frac{z - \min(z)}{\max(z) - \min(z)} \quad (1)$$

Normalization was applied to several numerical attributes such as RBC, HGB, MCV, MCH, and other hematological parameters to ensure that each feature was scaled within the range of 0 to 1. The purpose of this normalization was to standardize feature scales, allowing the model to be trained more optimally and fairly, without bias toward any particular feature due to unit discrepancies.

The next step involved checking for any potential missing values. Based on the evaluation, all attributes in the dataset were found to be complete, with no missing entries. A summary of the missing value inspection is presented in Table 2.

Table 2. Missing Value Summary

Feature	*Number of Missing Values*
RBC	0
HGB	0
HCT	0
MCV	0
MCH	0
MCHC	0
RDW	0
WBC	0
PLT	0
MPV	0
PDW	0
PCT	0
Diagnosis	0

Before selecting the features for model training, a correlation analysis was conducted using the Pearson Correlation Coefficient to assess the linear relationship between each input feature and the anemia diagnosis label. The analysis revealed that features such as HGB, RBC, and MCV had relatively strong correlations with the anemia types, indicating their high predictive potential for classification tasks.

To enhance model performance and mitigate the risk of overfitting, a feature selection process was applied to the CBC data, using two primary methods. First, the Pearson Correlation Coefficient was utilized to measure the strength of the linear relationship between each numerical CBC attribute and the target variable (anemia type), selecting features with higher correlation values as more relevant for prediction. Additionally, the SelectKBest method with the f_classif scoring function was employed to automatically select the top K features based on their statistical significance in relation to the target variable. As a result, the most significant features for model training were Hemoglobin (HGB), Red Blood Cell Count (RBC), Mean Corpuscular Volume (MCV), Hematocrit (HCT), and MCHC. By retaining only these informative features, the training process became more computationally efficient and led to improved accuracy in the model's predictions [17], [18].

After the preprocessing stage was completed, the dataset was divided into two main subsets: a training set and a testing set. Specifically, 70% of the data was allocated for training the machine learning models, while the remaining 30% was used to evaluate model performance on unseen data. Proper dataset partitioning is essential to ensure that the trained models can generalize well to new, independent data. In this study, several machine learning algorithms were developed and their performance was compared using various evaluation metrics, including True Positive Rate (TPR), True Negative Rate (TNR), accuracy, recall (sensitivity), precision, specificity, and the F1-score.

3.3. Exploratory Data Analysis (EDA)

Before building the machine learning models, a feature selection step was carried out to identify the most influential attributes in the dataset [19]. The process began with a correlation analysis to evaluate relationships among the numerical features. The dataset, provided in CSV format, was uploaded and processed using Google Colab as the data analysis platform.

Various visualization techniques, such as heatmaps and scatter plots, were used to examine the strength of correlations between features and to determine their relevance to the classification task. These visual outputs were saved in TIF format for further analysis. Additionally, EDA was conducted using the Python programming language to gain deeper insights into data distributions and patterns. This stage was essential to ensure that the selected features were truly relevant and had a significant contribution to the development of an effective classification model.

3.4. Machine Learning Model Development

SVM is a supervised classification algorithm that aims to separate data points into distinct classes by identifying the optimal hyperplane that maximizes the margin between classes. In this study, SVM was applied to classify anemia types based on CBC parameters [20]. SVM is particularly effective in handling high-dimensional datasets and typically performs well after feature normalization, which ensures that all features contribute equally to the decision boundary [19]. To handle data that is not linearly separable, SVM uses kernel functions such as the linear kernel or Radial Basis Function (RBF) to map data into higher-dimensional space where separation becomes feasible. In the context of this research, the implementation of SVM aims to evaluate the algorithm's ability to accurately classify patients into four anemia categories: Healthy, Normocytic hypochromic anemia and Normocytic normochromic anemia, based on their hematological attributes.

Naïve Bayes is a probabilistic classification algorithm based on Bayes' Theorem, which assumes that all features are conditionally independent given the target class. Despite its simplicity, Naïve Bayes remains highly effective and is often used as a baseline model in medical classification tasks due to its fast-processing speed and scalability particularly when working with small to medium-sized datasets [21]. One of the key strengths of Naïve Bayes lies in its computational efficiency and ability to produce reliable results even with limited data. It supports both categorical and numerical attributes, as long as the input features are properly encoded and normalized. However, its main limitation is the strong assumption of feature independence, which may reduce performance when there are strong correlations between clinical features. In this study, Naïve Bayes is employed as one of the comparative models to assess its effectiveness in classifying anemia types: Healthy, Normocytic hypochromic anemia, and Normocytic normochromic anemiabased on hematological parameters from the CBC dataset. It also serves as a baseline reference to compare against more advanced models.

XGBoost is an advanced ensemble learning algorithm that constructs a series of decision trees in a sequential manner, where each subsequent tree attempts to correct the errors of the previous ones using gradient descent optimization. This algorithm integrates both L1 (Lasso) and L2 (Ridge) regularization to reduce the risk of overfitting and improve the model's ability to generalize [22]. XGBoost is widely recognized for its high speed, robust performance, and ability to handle missing values, class imbalance, and high-dimensional datasets efficiently making it well-suited for clinical applications. In this study, XGBoost was implemented as one of the comparative machine learning models to evaluate its performance in classifying anemia types namely Healthy, Normocytic hypochromic anemia and Normocytic normochromic anemia, based on numerical features from the CBC dataset. Its results are compared with those of other algorithms, including SVM, and Naive Bayes, in order to identify the most effective model for multi-class anemia classification.

3.5. Model Performance

After the Machine Learning (ML) models were developed, their performance was evaluated using the 30% testing dataset. Several key evaluation metrics were employed in this study, including precision, recall (sensitivity), specificity, F1-*Score*, error rate, and accuracy. These metrics were derived from the confusion matrix and supported by visual tools such as ROC curves, all implemented using Python. The confusion matrix, as shown in Table 3, provides a detailed comparison between the predicted labels and the actual classes in the test dataset. This enabled a comprehensive performance analysis of each model, illustrating how well each algorithm classified the four types of anemia: Healthy, Normocytic hypochromic anemia and Normocytic normochromic anemia.

ISSN 2579-7069

To increase the robustness of the results, 5-fold cross-validation was applied during model training and testing. This technique evaluates the model across different data partitions, helping to mitigate overfitting and improve generalizability. After hyperparameter tuning, each model was re-evaluated using the test dataset, and their final performance was compared using accuracy and other relevant metrics.

Table 3. Confusion Matrix Structure

	Predicted Class 0	*Predicted Class 1*	*Predicted Class 2*
Actual Class 0	True Positives (TP)	FN, etc.	...
Actual Class 1	...	TP	...
Actual Class 2	...	...	TP

In a multi-class classification setting, the confusion matrix is a tool used to evaluate the performance of the model by presenting the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) for each class. True Positives (TP) represent the instances that were correctly predicted as belonging to a specific class. False Positives (FP) occur when instances are incorrectly predicted as belonging to a class, even though they actually belong to another. True Negatives (TN) are the instances that were correctly identified as not belonging to a specific class, while False Negatives (FN) refer to instances that actually belong to a class but were incorrectly predicted as belonging to another. These values are essential for calculating various performance metrics, such as accuracy, precision, recall, and F1-score, for each class individually.

Based on the generated confusion matrix, several evaluation metrics were calculated to measure the performance of the classification models.

Accuracy represents the proportion of all predictions that the model classified correctly. It is calculated using the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision also known as the positive predictive value, precision measures the proportion of correctly predicted positive cases out of all predicted positive cases:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall quantifies the model’s ability to correctly identify actual positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Specificity also referred to as the TNR, specificity measures the proportion of actual negative cases that are correctly predicted by the model. It serves as a complement to recall:

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

F1-score is the harmonic mean of precision and recall. It offers a balanced metric that accounts for both false positives and false negatives, and is particularly useful in imbalanced classification tasks. The score ranges from 0 (worst) to 1 (best) [23]:

$$F1\ Score = \frac{2 * Precision * Recall}{2 * TP + FP + FN} \tag{6}$$

4. Results and Discussion

4.1. Correlation Analysis

To better understand the interrelationships among variables in the CBC dataset and assess the potential for multicollinearity, a Pearson Correlation Coefficient analysis was conducted. The results of this analysis are visualized in a correlation heatmap, as shown in Figure 1. In this heatmap, a color gradient is used to represent the strength and direction of linear relationships between the variables. Dark red indicates a strong positive correlation, approaching +1, while blue represents a strong negative correlation, closer to -1. White or lighter shades are used to signify weak or negligible correlations, providing a clear visual indication of the relationships between the variables in the dataset.

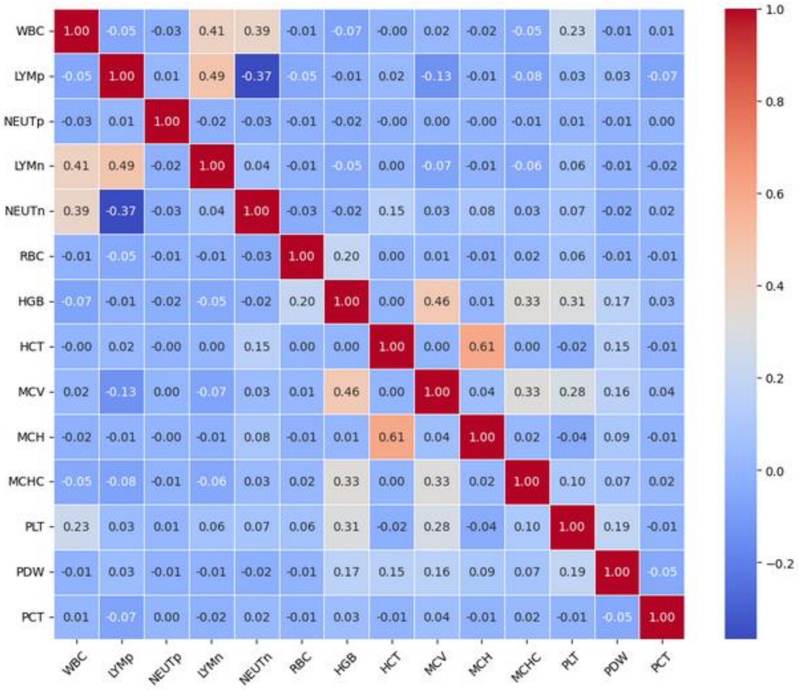

Figure 1. Heatmap Analysis of the Dataset

The analysis revealed several notable patterns within the CBC dataset. HGB and HCT exhibit a very strong positive correlation of 0.92, reflecting their physiological connection as key components of red blood cells. Additionally, Mean Corpuscular Volume (MCV) shows a positive correlation with both Mean Corpuscular Hemoglobin (MCH) and HCT, with a correlation value of 0.61 for each, suggesting that these variables share overlapping information related to red blood cell size and hemoglobin content. Furthermore, moderate correlations were observed between Red Blood Cell count (RBC) and both HGB (0.46) and HCT (0.46), reinforcing their interdependence in clinical assessments of red blood cell health.

In contrast, features such as White Blood Cell count (WBC), lymphocyte percentages (LYMp and LYMn), and platelet indices (PDW and PCT) show minimal correlation with most other features (typically between -0.1 and 0.2), indicating that they may provide independent or complementary information. Interestingly, no single feature dominates with high correlations across the board, which reflects the dataset's heterogeneity and diverse hematological dimensions. This diversity is advantageous for machine learning classification, as it minimizes the risk of multicollinearity and enhances the potential for uncovering meaningful patterns. The findings from this correlation analysis informed the feature selection process. Although some variables were highly correlated (e.g., HGB and HCT), all features were retained for model development due to their potentially distinct contributions, especially when using non-linear classifiers such as XGBoost.

4.2. Machine Learning Models

To evaluate the performance of the SVM classifier in diagnosing anemia, several performance metrics were computed, including class-specific precision, recall, and F1-score. These metrics provide insight into how well the model distinguishes among the three diagnostic categories: *Healthy*, *Normocytic hypochromic anemia*, and *Normocytic normochromic anemia*. The detailed results are presented in Table 4.

ISSN 2579-7069

Table 4. SVM Model Performance Metrics

Class	***Precision***	***Recall***	***F1-Score***	***Support***
Healthy (0)	0.95	0.93	0.94	67
Normocytic hypochromic anemia (1)	0.92	0.96	0.94	94
Normocytic normochromic anemia (2)	0.92	0.89	0.91	54
Accuracy	—	—	**0.93**	**215**
Macro Average	0.93	0.92	0.93	215
Weighted Average	0.93	0.93	0.93	215

As shown in the table above, the SVM classifier achieved a high overall accuracy of 93%, with balanced precision and recall across all classes. These metrics suggest that the model is effective in distinguishing between different anemia types. To further analyze the prediction behavior and observe misclassification patterns, a confusion matrix was constructed as illustrated in the next section.

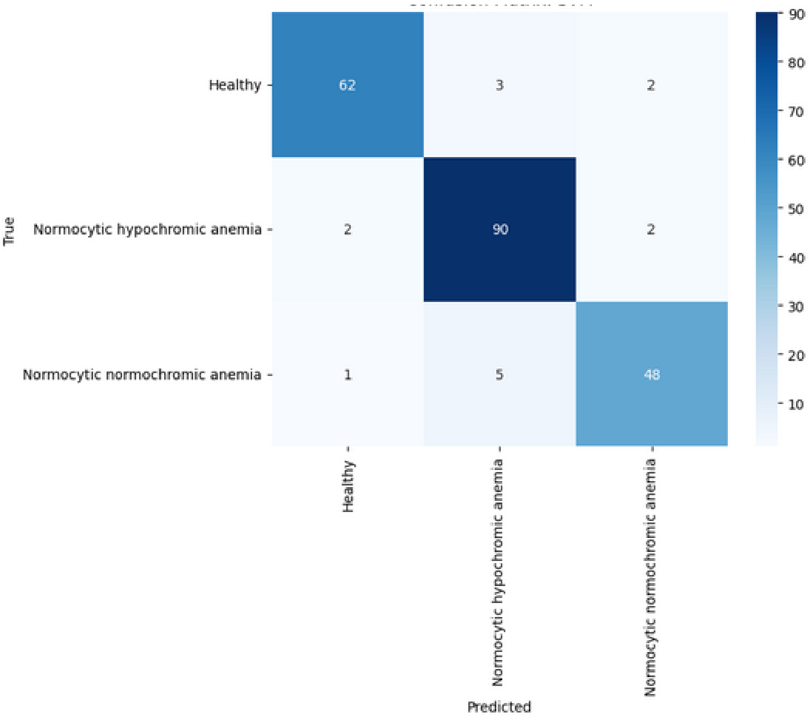

Figure 2. Confusion Matrix of the SVM Model

Figure 2 displays the confusion matrix for the SVM classifier in diagnosing anemia based on CBC data, comparing predicted class labels to the actual class labels. This matrix allows for a detailed assessment of the classification performance across different anemia categories. For the Healthy class (class 0), out of 67 actual healthy instances, 62 were correctly classified, while 3 were misclassified as Normocytic hypochromic anemia and 2 as Normocytic normochromic anemia, indicating strong performance with minimal false positives. For Normocytic hypochromic anemia (class 1), this class was the most accurately predicted, with 90 out of 94 cases correctly classified. Only 2 samples were incorrectly labeled as healthy, and 2 as Normocytic normochromic anemia, reflecting excellent recall and precision. For Normocytic normochromic anemia (class 2), out of 54 actual cases, 48 were correctly identified, while 5 were misclassified as Normocytic hypochromic anemia, and 1 as healthy. Although the performance was still strong, this class exhibited slightly more misclassifications compared to the other categories.The confusion matrix shows that the SVM model performs exceptionally well in classifying anemia types, particularly Normocytic hypochromic anemia, which had the highest number of correct predictions. Misclassifications were minimal and relatively balanced across classes, suggesting that the model generalizes well to unseen data.These results confirm the reliability of SVM in multiclass anemia diagnosis based on CBC features and highlight its potential as a clinical decision support tool.

To assess the classification performance of the Naive Bayes model across three diagnostic categories Healthy, Normocytic hypochromic anemia, and Normocytic normochromic anemia, standard evaluation metrics such as precision, recall, and F1-score were used. Table 5 summarizes these results based on the test dataset.

Table 5. Naive Bayes Model Performance Metrics

Class	*Precision*	*Recall*	*F1-Score*	*Support*
Healthy (0)	0.71	0.87	0.78	67
Normocytic hypochromic anemia (1)	0.82	0.72	0.77	94
Normocytic normochromic anemia (2)	0.50	0.46	0.48	54
Accuracy			**0.70**	**215**
Macro Average	0.68	0.68	0.68	215
Weighted Average	0.70	0.70	0.70	215

As shown in the table, the Naive Bayes classifier achieved an overall accuracy of 70.2%. The model demonstrated strong performance in identifying healthy individuals and those with normocytic hypochromic anemia. However, its ability to detect normocytic normochromic anemia was more limited, as reflected by the lower precision and recall for that class. To better understand where the model tends to misclassify, a confusion matrix is presented in the next section.

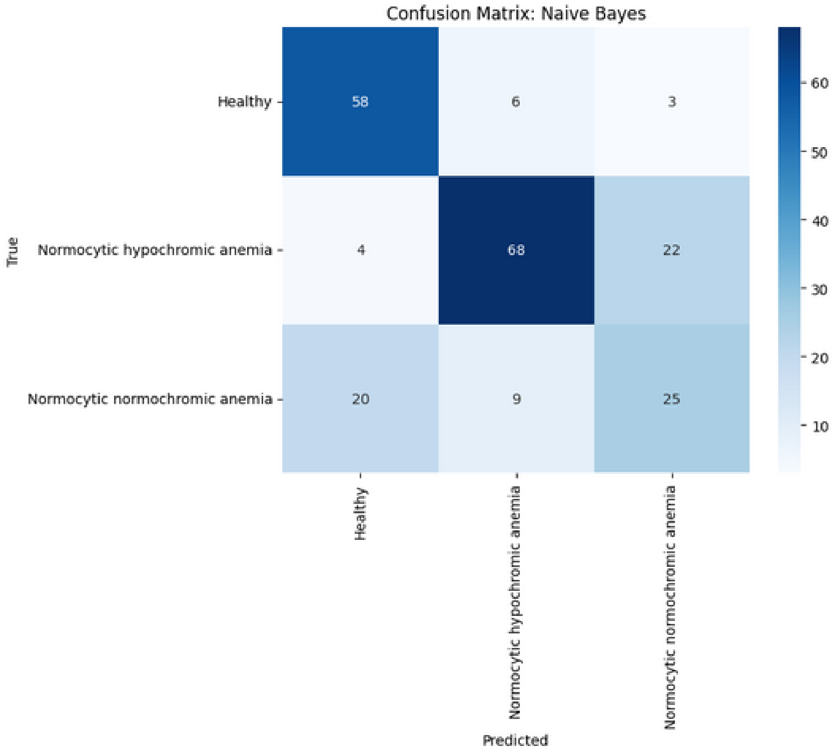

Figure 3. Confusion Matrix of the Naive Bayes Model

Figure 3 presents the confusion matrix for the Naive Bayes classifier, showcasing the model's performance on individual classes in the test dataset. For the Healthy class (Class 0), out of 67 actual healthy cases, 58 were correctly classified, while 6 were misclassified as Normocytic hypochromic anemia and 3 as Normocytic normochromic anemia. This indicates strong recall for healthy individuals, though there is some overlap with anemic categories. For Normocytic hypochromic anemia (Class 1), of the 94 true cases, 68 were correctly predicted, but 22 were misclassified as Normocytic normochromic anemia and 4 as healthy. The classifier frequently confuses Class 1 and Class 2, likely due to overlapping CBC features between these anemia types. Finally, for Normocytic normochromic anemia (Class 2), the model showed the weakest performance, correctly predicting only 25 out of 54 cases, while 20 were misclassified as healthy and 9 as Normocytic hypochromic anemia. The high misclassification rate suggests that this class is more challenging for the model to distinguish, possibly due to less distinctive or overlapping patterns in the data.The confusion matrix highlights that the Naive Bayes classifier performs reasonably well in identifying healthy individuals and normocytic hypochromic anemia cases. However, it struggles significantly to accurately detect normocytic normochromic anemia, with more than half of its cases misclassified. This outcome reflects Naive Bayes' limitations in modeling complex or overlapping distributions, especially in multi-class medical diagnosis tasks involving subtle feature variations. Further refinement through feature selection, data balancing, or the use of more flexible classifiers (e.g., XGBoost or SVM) may improve performance on underrepresented or harder-to-classify anemia types.

To assess the classification performance of the XGBoost model, the precision, recall, and F1-score were calculated for each class label: *Healthy*, *Normocytic hypochromic anemia*, and *Normocytic normochromic anemia*. These metrics offer a comprehensive view of the model's predictive capabilities. Table 6 summarizes the classification performance on the test dataset.

Table 6. XGBoost Model Performance Metrics

Class	*Precision*	*Recall*	*F1-Score*	*Support*
Healthy (0)	0.97	1.00	0.99	67
Normocytic hypochromic anemia (1)	1.00	1.00	1.00	94
Normocytic normochromic anemia (2)	1.00	0.96	0.98	54
Accuracy	—	—	**0.99**	**215**
Macro Average	0.99	0.99	0.99	215
Weighted Average	0.99	0.99	0.99	215

The results in Table 6 indicate that XGBoost achieved an exceptionally high overall accuracy of 99%. It performed with near-perfect precision and recall for all classes, particularly excelling in classifying normocytic hypochromic anemia. Minor misclassifications were observed in the normocytic normochromic anemia class, which slightly lowered the recall to 0.96. A confusion matrix is provided in the following section to further explore the distribution of prediction errors.

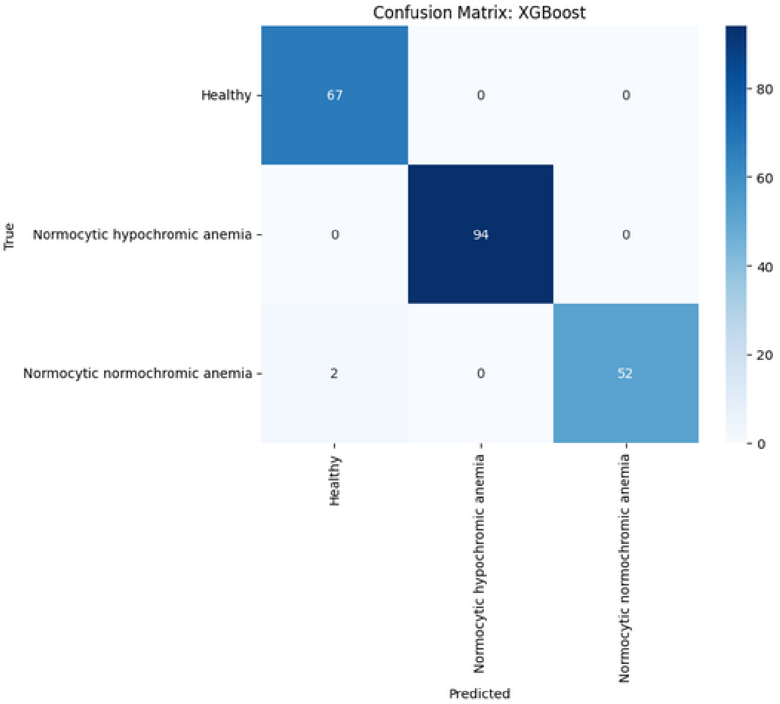

Figure 4. Confusion Matrix of the XG-Boost Model

Figure 4 presents the confusion matrix for the XGBoost classifier on the test dataset, highlighting its performance in predicting each anemia category. For the Healthy class (Class 0), all 67 instances were correctly classified, with no misclassifications, reflecting perfect recall and precision for this category. For Normocytic hypochromic anemia (Class 1), the model also achieved perfect classification, correctly identifying all 94 cases without any errors, demonstrating exceptional capability in distinguishing this anemia type. For Normocytic normochromic anemia (Class 2), out of 54 actual cases, 52 were correctly predicted, while 2 were misclassified as healthy. There were no misclassifications into Class 1, indicating that the model can differentiate normochromic anemia fairly well from hypochromic types, although there is minor confusion with the healthy class.

The confusion matrix confirms the exceptional performance of the XGBoost model, with only 2 misclassifications out of 215 test samples, resulting in an overall accuracy of 99%. Notably, the model made zero errors in classifying both healthy individuals and normocytic hypochromic anemia. It also demonstrated very high precision and recall across all classes. While there was slight overlap between the normocytic normochromic anemia and healthy classes, this was minimal and acceptable. These results indicate that XGBoost is highly effective in anemia classification based on CBC data, showcasing its strong potential as a decision support tool for clinical diagnostics.

To provide a clearer overview of the overall classification performance, Table 7 summarizes the key evaluation metrics accuracy, precision, recall, and F1-score for each machine learning model used in this study. These metrics are computed using macro averages to ensure fair representation across all classes, especially in a multi-class classification setting.

Table 7. Summary of performance metrics across all four models.

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Random Forest	0.9702	0.96	0.84	0.89
SVM	0.9624	0.97	0.79	0.85
Naive Bayes	0.9048	0.71	0.78	0.74
XGBoost	0.9709	0.96	0.85	0.89

From the summary above, it is evident that the XGBoost and Random Forest classifiers achieved the highest overall performance, with both models attaining accuracy scores above 97% and balanced precision–recall values. SVM also performed well, although with slightly lower recall. In contrast, Naive Bayes showed the lowest performance across all metrics, particularly in precision, suggesting limitations in capturing complex data patterns compared to tree-based or margin-based models. These results highlight the suitability of ensemble-based methods like XGBoost for anemia classification using CBC data.

To further evaluate the discriminative power of each classifier in this multiclass anemia diagnosis task, the Receiver Operating Characteristic (ROC) curve was plotted using macro-averaged scores. ROC curves illustrate the trade-off between sensitivity (TPR) and specificity (1 - false positive rate), providing a visual assessment of model performance across all classification thresholds. The Area Under the Curve (AUC) is also calculated to quantify the overall performance, where a higher AUC indicates better classification capability.

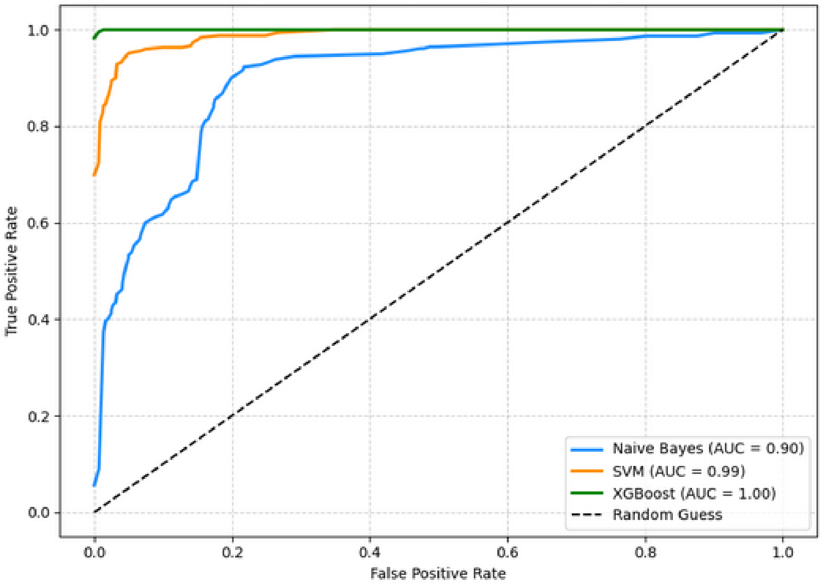

Figure 5. ROC Curve

From the ROC curve in Figure 5, XGBoost shows perfect classification with an AUC of 1.00, followed by SVM with a strong AUC of 0.99, and Naive Bayes with a moderate AUC of 0.90. All models perform well above the random classifier baseline (AUC = 0.5). This analysis highlights XGBoost as the most accurate and reliable model, with SVM also performing robustly. While Naive Bayes is less optimal, it remains a viable option when model simplicity and speed are prioritized.

5. Conclusion

This study evaluated the performance of four machine learning algorithms, Random Forest, SVM, Naive Bayes, and XGBoost for classifying anemia types using CBC data. The results showed that XGBoost performed best, achieving the highest accuracy (99%) and perfect AUC (1.00), followed by SVM and Random Forest, which also showed strong performance. In contrast, Naive Bayes had lower accuracy and struggled to distinguish normocytic normochromic anemia. Overall, the findings confirm that machine learning, particularly ensemble methods like XGBoost, can effectively support automated and accurate anemia diagnosis based on blood test data.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, N.A.P. and B.P.M; Methodology, N.A.P. and B.P.M.; Software, N.A.P.; Validation, N.A.P.; Formal Analysis, N.A.P.; Investigation, B.P.M.; Resources, N.A.P.; Data Curation, B.P.M.; Writing Original Draft Preparation, N.A.P.; Writing Review and Editing, N.A.P. and B.P.M.; Visualization, B.P.M. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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.

References

[1] T. K. Yıldız, N. Yurtay, and B. Öneç, "Classifying Anemia Types Using Artificial Learning Methods," *Eng. Sci. Technol. Int. J.*, vol. 24, no. 1, pp. 50–70, 2021, doi: 10.1016/j.jestch.2020.12.003.

[2] M. N. Garcia-Casal, O. Dary, M. E. Jefferds, and S.-R. Pasricha, "Diagnosing Anemia: Challenges Selecting Methods, Addressing Underlying Causes, and Implementing Actions at The Public Health Level," *Ann. N. Y. Acad. Sci.*, vol. 1524, pp. 37–50, 2023, doi: 10.1111/nyas.14996.

[3] J. G. Gómez, C. Parra Urueta, D. S. Álvarez, V. Hernández Riaño, and G. Ramirez-Gonzalez, "Anemia Classification System Using Machine Learning," *Informatics*, vol. 12, no. 1, p. 19, 2025, doi: 10.3390/informatics12010019.

[4] N. J. Kassebaum, R. Jasrasaria, M. Naghavi, S. K. Wulf, N. Johns, R. Lozano, M. Regan, D. Weatherall, D. P. Chou, T. P. Eisele, S. R. Flaxman, R. L. Pullan, S. J. Brooker, and C. J. Murray, "A Systematic Analysis of Global Anemia Burden from 1990 to 2010," *Blood*, vol. 123, no. 5, pp. 615–624, 2014, doi: 10.1182/blood-2013-06-508325.

[5] L. Del Castillo, N. Cardona-Castro, D. R. Whelan, M. M. Restrepo, J. P. Cuello, L. F. Contreras, M. C. Flórez, J. P. Builes, H. Serrano-Coll, M. Arboleda, and J. S. Leon, "Prevalence and Risk Factors of Anemia in The Mother–Child Population from a Region of The Colombian Caribbean," *BMC Public Health*, vol. 23, p. 1533, 2023, doi: 10.1186/s12889-023-16475-0.

[6] J. M. Pineda, "Predictive Models in Health Based on Machine Learning," *Adv. Med. Eng. Interdiscip. Res.*, vol. 2, no. 4, pp. 1–9, 2024, doi: 10.32629/ameir.v2i4.2813.

[7] World Health Organization, “Anaemia in Women and Children,” Global Health Observatory (GHO) Data, 2024. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children.

[8] M. D. Cappellini and I. Motta, “Anemia in Clinical Practice Definition and Classification: Does hemoglobin change with aging?,” *Semin. Hematol.*, vol. 52, no. 4, pp. 261–269, 2015, doi: 10.1053/j.seminhematol.2015.07.006.

[9] S. E. Calle-Pesántez and J. M. Pallo-Chiguano, “Capítulo 3. Inteligencia Artificial en la Comunicación Científica,” *Esp. Monogr. Comun. Soc.*, no. 23, pp. 59–81, 2024, doi: 10.52495/c3.emcs.23.ti12.

[10] S. Pullakhandam and S. McRoy, “Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning,” *BioMedInformatics*, vol. 4, no. 1, pp. 661–672, 2024, doi: 10.3390/biomedinformatics4010036.

[11] G. Airlangga, “Leveraging Machine Learning for Accurate Anemia Diagnosis Using Complete Blood Count Data,” *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 2, pp. 318–326, 2024, doi: 10.24014/ijaidm.v7i2.29869.

[12] S. S. Abdul-Jabbar, A. K. Farhan, and A. S. Luchinin, “A Comparative Study of Anemia Classification Algorithms for International and Newly CBC Datasets,” *Int. J. Online Biomed. Eng. (iJOE)*, vol. 19, no. 6, pp. 141–157, 2023, doi: 10.3991/ijoe.v19i06.38157.

[13] L. Végh, O. Takáč, K. Czakóová, D. Dancsa, and M. Nagy, “Evaluating Optimizable Machine Learning Models for Anemia Type Prediction from Complete Blood Count Data,” *Int. J. Adv. Nat. Sci. Eng. Res.*, vol. 8, no. 7, pp. 108–119, 2024. [Online]. Available: https://as-proceeding.com/index.php/ijanser/article/view/1973.

[14] L. Kabootarizadeh, A. Jamshidnezhad, and Z. Koohmareh, “Differential Diagnosis of Iron-Deficiency Anemia From B-Thalassemia Trait Using an Intelligent Model in Comparison with Discriminant Indexes,” *Acta Inform. Med.*, vol. 27, no. 2, pp. 78–84, 2019, doi: 10.5455/aim.2019.27.78-84.

[15] H. Benhar, A. Idri, and J. L. Fernández-Alemán, “Data Preprocessing for Heart Disease Classification: A Systematic Literature Review,” *Comput. Methods Programs Biomed.*, vol. 195, p. 105635, 2020, doi: 10.1016/j.cmpb.2020.105635.

[16] Y. Kumar, A. Koul, R. Singla, and others, “Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review, Synthesizing Framework and Future Research Agenda,” *J. Ambient Intell. Human Comput.*, vol. 14, pp. 8459–8486, 2023, doi: 10.1007/s12652-021-03612-z.

[17] Y.-C. Wang, H.-M. Song, J.-S. Wang, X.-R. Ma, Y.-W. Song, and Y.-L. Qi, “Multi-Strategy Fusion Binary SHO Guided by Pearson Correlation Coefficient for Feature Selection with Cancer Gene Expression Data,” *Egypt. Inform. J.*, vol. 29, p. 100639, 2025, doi: 10.1016/j.eij.2025.100639.

[18] K. M. M. Uddin, A. A. Mamun, A. Chakrabarti, and R. Mostafiz, “An Ensemble Machine Learning-Based Approach to Predict Thyroid Disease Using Hybrid Feature Selection,” *Biomed. Anal.*, vol. 1, no. 3, pp. 229–239, 2024, doi: 10.1016/j.bioana.2024.08.001.

[19] F. Palace-Berl, S. D. Jorge, K. F. Pasqualoto, A. K. Ferreira, D. A. Maria, R. R. Zorzi, L. de Sá Bortolozzo, J. Â. Lindoso, and L. C. Tavares, “5-Nitro-2-Furfuriliden Derivatives as Potential Anti-Trypanosoma Cruzi Agents: Design, Synthesis, Bioactivity Evaluation, Cytotoxicity And Exploratory Data Analysis,” *Bioorg. Med. Chem.*, vol. 21, no. 17, pp. 5395–5406, 2013, doi: 10.1016/j.bmc.2013.06.017.

[20] E. Dritsas and M. Trigka, “Data-Driven Machine-Learning Methods for Diabetes Risk Prediction,” *Sensors*, vol. 22, no. 14, p. 5304, 2022, doi: 10.3390/s22145304.

[21] H. Habehh and S. Gohel, “Machine Learning in Healthcare,” *Curr. Genomics*, vol. 22, no. 4, pp. 291–300, 2021, doi: 10.2174/1389202922666210705124359.

[22] A. Sirag and N. Mohamed Nor, “Out-of-Pocket Health Expenditure and Poverty: Evidence from A Dynamic Panel Threshold Analysis,” *Healthcare*, vol. 9, no. 5, p. 536, 2021, doi: 10.3390/healthcare9050536.

[23] Ş. K. Çorbacıoğlu and G. Aksel, “Receiver Operating Characteristic Curve Analysis in Diagnostic Accuracy Studies: A Guide to Interpreting The Area Under The Curve Value,” *Turk. J. Emerg. Med.*, vol. 23, no. 4, pp. 195–198, 2023, doi: 10.4103/tjem.tjem_182_23.