JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.11019

COMPARISON OF GAUSSIAN NAIVE BAYES AND RANDOM FOREST FOR ANEMIA CLASSIFICATION USING HEMATOLOGICAL PARAMETERS

Baik Budi[1*], Selvy Sandrika Ramadhani[1], Refki Budiman[1], Queen Hesti Ramadhamy[1], Riko Novendra[1], Baharuddin[1], Tiara Mahardika[2]

[1]Universitas Andalas; Jalan Dr. Mohammad Hatta, Kota Padang
[2]Politeknik Negeri Padang; Jl. Kampus, Limau Manis, Kec. Pauh, Kota Padang

Keywords:
Anemia;
Naïve Bayes;
Random Forest;
Classification;
Hematology;

Correspondent Email:
baikbudi@eng.unand.ac.id

Anemia is a global health problem affecting approximately 1.92 billion people, or 24% of the population, according to the WHO. Accurate early detection is crucial for data-driven healthcare. This study evaluates two machine learning algorithms, Gaussian Naive Bayes (GNB) and Random Forest (RF). The classification is based on four key hematological parameters: Hemoglobin (Hb), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), and Mean Corpuscular Hemoglobin Concentration (MCHC). GNB relies on Bayes' Theorem with Gaussian distribution assumptions, whereas RF is a decision-tree-based ensemble method capable of capturing non-linear patterns without specific distributional assumptions. Evaluated using 5-fold cross-validation and standard metrics (accuracy, precision, recall, F1-score), results showed that RF outperformed GNB. RF achieved 94.2% accuracy (CV 94.9% ± 1.1%), compared to GNB's 90.5% (CV 90.1% ± 1.3%). RF feature importance confirmed Hb as the dominant predictor (score 0.562), aligning with its strong correlation to anemia ($r = -0.80$). Although not surpassing larger-scale studies, these results remain highly competitive. Ultimately, this research provides evidence to support the development of automated, data-driven clinical decision support systems for anemia detection.

1. INTRODUCTION

Anemia is a global health problem that remains a major concern due to its high prevalence in various countries, especially in developing nations [1]. According to the latest report by the World Health Organization (WHO), approximately 1.92 billion people worldwide suffer from anemia, with the greatest burden borne by reproductive-age women and children [2], [3]. Anemia occurs when hemoglobin levels in the blood fall below normal values, impairing the blood's ability to transport oxygen throughout the body [1]. This condition can cause various impacts such as fatigue, decreased concentration, impaired cognitive development, and an increased risk of complications in pregnant women and children [1].

Clinical diagnosis of anemia is generally performed through laboratory examinations by analyzing four main hematological parameters: (1) Hemoglobin (Hb), the oxygen-carrying protein in erythrocytes with normal values of 12–16 g/dL in women and 13.5–17.5 g/dL in men; (2) Mean Corpuscular Volume (MCV), which reflects the average volume of erythrocytes in femtoliters (fL) with a normal range of 80–100 fL; (3) Mean Corpuscular

Hemoglobin (MCH), which represents the average hemoglobin mass per erythrocyte with a normal value of 27–33 pg; and (4) Mean Corpuscular Hemoglobin Concentration (MCHC), which represents the average concentration of hemoglobin per unit volume of erythrocytes with a normal value of 32–36 g/dL. The combination of these four parameters forms a hematological profile that can distinguish types of clinical anemia. It has been widely used as input features in various machine learning-based anemia classification studies [4], [5].

With the development of information technology, the use of machine learning in healthcare is increasing, particularly to assist with automated, data-driven disease diagnosis [6]. Various classification algorithms have been applied to a range of medical diagnosis problems, from heart disease [7], [8] and cancer [9] to chronic kidney disease and diabetes [10], showing promising results in supporting clinical decision-making systems. Among these various algorithms, Gaussian Naive Bayes (GNB) and Random Forest (RF) are two of the most widely used approaches due to their complementary characteristics.

GNB is known for its advantage in Gaussian-distribution-based probability estimation, which is computationally efficient and fast, making it suitable for medical datasets with limited dimensionality [11]. However, GNB has a fundamental weakness in its assumption of conditional independence among features, which is rarely met in real-world data, including hematological data [7], [12]. Various GNB variants have been proposed to mitigate the impact of this assumption violation, for instance, through attribute weighting, local learning, and ensemble approaches [12][13]. Meanwhile, Random Forest is a decision-tree-based ensemble method that aggregates predictions from multiple trees via majority voting [10][11]. RF can naturally handle feature interactions, is robust to noise, does not require specific assumptions about the data distribution, and yields clinically useful feature importance [14], [15], [16], [17]. These characteristics have led to the widespread adoption of RF in various healthcare decision support systems [7], [10].

Based on these issues, this study uses an anemia dataset from Kaggle [18] to build and evaluate classification models with GNB and Random Forest, individually and in comparison. This research is expected to contribute to the development of decision support systems in healthcare, specifically for faster, more accurate, and efficient anemia detection, while simultaneously providing empirical evidence regarding the trade-off between interpretability and accuracy in both approaches.

1.1 Research Gap

Most previous studies in anemia classification have tended to use a single method, evaluate many algorithms simultaneously without an in-depth discussion of their respective characteristics [5], or focus on non-tabular data modalities such as images [19]. GNB offers the advantage of fast, efficient probability estimation. Still, its performance can degrade when the assumption of feature independence is violated, particularly in hematological data, where parameters (Hb, MCV, MCH, MCHC) are physiologically correlated [12], [20]. On the other hand, Random Forest is well-suited to capturing non-linear patterns and feature interactions. Yet, it has not always been directly and systematically compared with GNB on the same tabular hematological anemia dataset, using consistent evaluation protocols (5-fold cross-validation, identical metrics) for both models. Therefore, this research focuses on a comprehensive evaluation and comparison of GNB and RF individually on the same anemia dataset to determine the strengths and weaknesses of each method for anemia classification based on four standard hematological parameters.

1.2 Research Contribution

The main contributions of this research can be summarized as follows: (1) Systematically comparing the performance of GNB and Random Forest on the same hematological anemia dataset with identical evaluation protocols (metrics and 5-fold cross-validation scheme), ensuring a fair comparison. (2) Presenting an integrated analysis of feature importance from the RF model and correlation analysis of hematological parameters to anemia status, confirming Hb as the dominant predictor consistent with hematological physiology

literature. (3) Providing a quantitative comparative analysis against several previous machine learning-based anemia studies, including a discussion on the differences in dataset characteristics that affect accuracy gaps. (4) Identifying methodological limitations (sample size, single data source, lack of statistical significance testing, and potential class imbalance) as a basis for recommending more rigorous future research.

2. RELATED WORK

Several previous studies have demonstrated that machine learning can be effectively used in disease classification, including anemia [11]. The Naive Bayes algorithm has proven effective in classifying numerical medical data; this probabilistic approach is well-suited to medical datasets because it integrates clinical prior information into the model [6], [21]. Rahman et al. [5] compared several machine learning and ensemble learning algorithms on 1,000 anemia data samples with 8 attributes, reporting Logistic Regression as the best model with 95% accuracy. Mugdha et al. [4] developed regression and classification models based on CBC data from Bangabandhu Sheik Mujib Medical University to estimate hemoglobin levels and predict anemia severity. Chandralekha et al. [22] evaluated Random Forest, ExtraTrees, Gradient Boosting, XGBoost, and a hybrid model for anemia prediction, with the hybrid model achieving the highest accuracy of 90.83%. Other studies have utilized palpebral conjunctival images as non-invasive features for anemia detection [19] XGBoost and LightGBM to classify anemia subtypes using limited, imbalanced data [23]. Gómez et al. [21] built an anemia classification system using Linear Discriminant Analysis, Decision Tree, and Random Forest based on blood count data, achieving competitive results for multi-class subclassification. Naïve Bayes is also used to predict milk quality with an accuracy 85% [24].

Random Forest has also been widely used in general medical data classification due to its ability to handle high-dimensional data, robustness to noise, and lack of reliance on distributional assumptions [14], [15]. In heart disease prediction, RF was reported to achieve 90.16% accuracy using the Cleveland Heart Disease dataset [7]. In contrast, a comparative study of six algorithms (AdaBoost, Random Forest, Decision Tree, KNN, Naive Bayes, and Perceptron) on an imbalanced cardiovascular disease dataset showed that AdaBoost had the highest AUC and F1-score, with Random Forest being a competitive candidate [8]. Several other studies have consistently reported that Random Forest outperforms Naive Bayes across various imbalanced binary classification domains, although the performance gap varies with dataset characteristics [16], [17]

Regarding the handling of class imbalance, which is commonly found in clinical data, oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) are widely used to balance class distributions before model training [25], [26]. However, the application of resampling techniques to medical data must be done with caution, as it can generate synthetic samples that do not reflect real clinical conditions. In this study, the level of class imbalance is considered mild (56.4%:43.6 %), so the models were trained directly on the original distribution without resampling. However, the evaluation still used minority-class-sensitive metrics, such as recall and F1 Score. The reliability of performance estimates in this study and most of the referenced studies above was validated using a k-fold cross-validation scheme, which has been shown to yield more stable estimates of model generalization than a single holdout split [27].

2.1. Gaussian Naïve Bayes

Naive Bayes is a probabilistic classifier that operates by applying Bayes' Theorem. This algorithm relies on a strong (naive) fundamental assumption that every predictor feature is conditionally independent of the others given the target class. The general mathematical formulation of Bayes' Theorem is expressed in Equation (1):

$$P(C_k|X) = \frac{P(X|C_k).P(C_k)}{P(X)} \quad (1)$$

Where:

$P(C_k|X)$ = The posterior probability of class C_k given the input feature vector X.

P(X | C_k) = The likelihood probability of feature X given that class C_k is known.

$P(C_k)$ = The prior probability of class C_k.
$P(X)$ = The prior probability of the predictors (normalization constant).

For continuous data (such as age or blood pressure in the heart disease dataset), the probability distribution is calculated using the Gaussian Naive Bayes probability density function, formulated in Equation (2)

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{x_i - \mu_k)^2}{2\pi\sigma_k^2}\right) \quad (2)$$

Where μ_k is the mean and σ_k^2 is the variance of the feature x_i associated with class C_k.

2.2. Random Forest

Random Forest is an ensemble method that constructs multiple decision trees during training and combines their predictions via majority voting. Each tree is trained using a random subset of the training data (bootstrap sampling) and a random subset of features, thereby generating inter-tree diversity that reduces the risk of overfitting [14]. The final prediction of the RF model is expressed as equation 3:

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (3)$$

where $\hat{y}$ is the final predicted class label, $h_1(x)$is the prediction of the *t-th* tree, and *T* is the total number of trees. Feature importance is calculated based on the mean decrease in Gini impurity (Equation 4):

$$I(f) = \frac{1}{T}\sum_{t=1}^{T}\sum_{v}\Delta Gini(f, v, t) \quad (4)$$

where *I(f)* is the importance of feature f, and $\Delta Gini(f, v, t)$is the decrease in Gini impurity at node *v* that utilizes feature *f* in the *t-th* tree.

The RF model was trained on the four original hematological parameters (Hb, MCV, MCH, and MCHC) as input features, with n_estimators = 100 and the Gini impurity criterion. The value of n_estimators = 100 was selected because it is the default setting in the scikit-learn library and has been shown to provide a good trade-off between accuracy and computational efficiency for medium-scale datasets [15]. Hyperparameter tuning (such as grid search or random search) was not performed, which constitutes one of the methodological limitations of this study.

3. METHODOLOGY

This study uses the Anemia Dataset from Kaggle, published by Biswaranjan Rao. The dataset comprises 1,421 patient samples featuring four primary hematological parameters: Hemoglobin (Hb), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), and Mean Corpuscular Hemoglobin Concentration (MCHC), along with a single class label indicating anemia status (positive/negative). This public dataset does not include comprehensive clinical metadata (such as age, gender, or sampling location); this limitation is discussed further in the study's limitations section. The overall research workflow comprises the following stages: data collection, data preprocessing, independent model training (GNB and RF), performance evaluation via 5-fold cross-validation, and comparative analysis.

3.1 Data Preprocessing

The data preprocessing phase is a critical step to ensure data quality and model reliability. This stage encompasses three primary procedures:

1. Handling Missing Values
   Missing data within each hematological parameter were addressed using median imputation. The median was selected over the mean because it is inherently more robust to outliers and skewed distributions, which are frequently encountered in real-world clinical datasets. This ensures that the central tendency of the data remains unaffected by extreme abnormal values.
2. Data Normalization
   A Min-Max Scaling method was applied to transform all feature values into a standardized range of 0 to 1. This step is essential because the four hematological parameters possess significantly different units of measurement and numerical ranges (e.g., Hb is measured in g/dL, while MCV is measured in fL). Standardizing the scales ensures that parameters with larger numerical magnitudes do not disproportionately influence the learning process, particularly for distance-based or probabilistically sensitive algorithms.
3. **Correlation Analysis**
   A statistical correlation analysis was conducted to assess the relationship between

each independent hematological parameter and the target class label (anemia status). This analysis serves as a foundational step for interpreting individual feature contributions, assessing predictive power, and identifying underlying physiological relationships before model training.

Furthermore, strict methodological protocols were maintained to ensure the validity of the evaluation. The dataset was partitioned into training and testing subsets using an 80:20 ratio. Crucially, the scaler-fitting process was performed *exclusively* on the training subset. The resulting transformation parameters were then applied to transform both the training and testing sets. This approach strictly prevents data leakage—the inadvertent introduction of test-set information into the model training phase—thereby ensuring that the model's performance evaluation reflects its true generalization capability on unseen data.

3.2 Model Evaluation

The reliability of both models was evaluated using accuracy, precision, recall, and F1-score, with an 80:20 train-test split. The models were further validated using 5-fold cross-validation to ensure generalization and mitigate bias from a single data split [27]. Particular attention was given to the recall metric due to the clinical context of anemia screening, where false negatives (undetected anemic patients) carry significantly higher risks and consequences than false positives. As a methodological note for future development, the performance comparison between models should ideally be accompanied by paired statistical significance testing (such as McNemar's test or a paired t-test across the fold results) to ensure that the difference in accuracy between GNB and RF is not merely due to random variation. In this study, such statistical tests were not conducted and are therefore recommended for future research (see the Research Limitations section).

4. RESULT AND DISCUSSION

This section presents the experimental results for applying the GNB and Random Forest models to the Kaggle anemia dataset. The dataset comprises 1,421 patient records featuring four primary hematological parameters: Hb, MCV, MCH, and MCHC, alongside a class label indicating the anemia status (positive/negative).

4.1 Preprocessing Data

The anemia dataset used in this study exhibits an imbalanced class distribution, although the imbalance is considered mild. Based on the data distribution graph, 620 samples are anemic (43.6%), while 801 are non-anemic (56.4%). This imbalance ratio of approximately 1.29:1 is relatively moderate compared to other medical datasets, which frequently present much more extreme ratios [25], [26]. Nevertheless, it still warrants consideration to ensure that the classification models do not develop a bias toward the majority class, as shown in Figure 1 below.

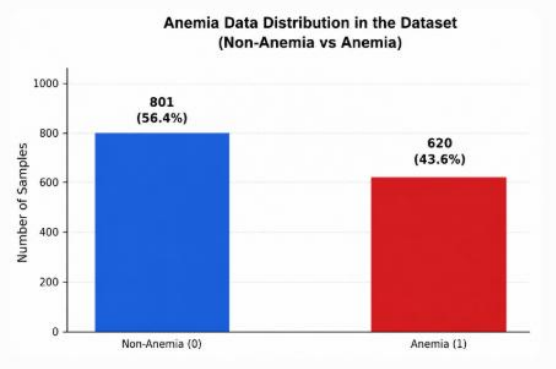

Figure 1. Dataset Distribution

Figure 2 below illustrates the distribution of Hemoglobin (Hb) levels according to anemia status. Patients with anemia tend to have lower Hb levels than the non-anemic group, indicating that Hb is the most discriminative parameter for anemia classification. This observation aligns with the clinical definition of anemia, which is fundamentally based on Hb level thresholds.

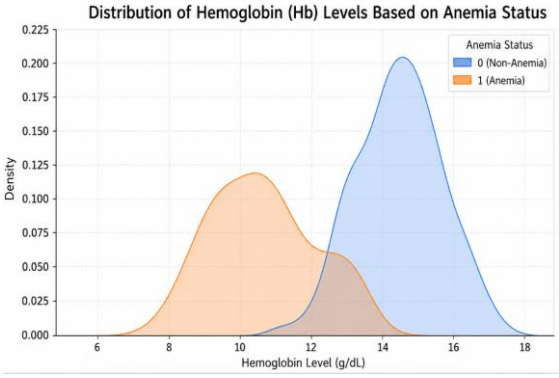

Figure 2. Distribution HB based on Anemia Status

The results of the correlation analysis between hematological parameters and anemia status reveal that Hb has the strongest negative correlation (r = −0.80), indicating a robust, dominant relationship compared with the other parameters. Meanwhile, the MCH (r = −0.03), MCV (r = −0.02), and MCHC (r = +0.05) parameters demonstrate very weak correlations with the class label. These findings are consistent with the clinical definition of anemia, which relies on Hb thresholds [20], and align with the feature-importance results of the RF model presented in the subsequent section. The correlation of hematological parameters to anemia status is shown in Figure 3.

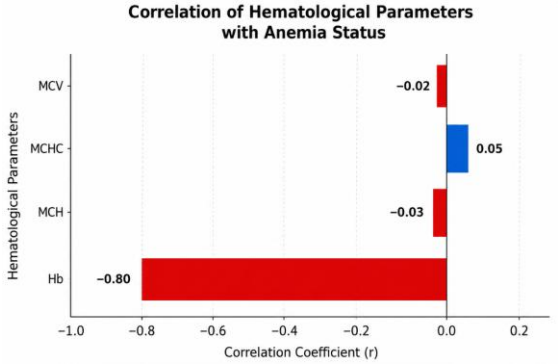

Figure 3. hematological parameters to anemia status

4.2 Classification using Gaussian Naïve Bayes

The Gaussian Naive Bayes model was trained on the preprocessed data using an 80% training set and a 20% test set. The model inputs consisted of the four original hematological parameters: Hb, MCV, MCH, and MCHC. Model validation was conducted using 5-fold cross-validation. Table 1 summarizes the performance evaluation results of the GNB model.

Table 1. GNB Performance Table

Matric Evaluation	Value
Accuracy	90,5%
Precision	86,5%
Recall	92,7%
F1-Score	89,5%
CV Accuracy (mean ± std)	90,1% ± 1,3%

Based on the evaluation results, the GNB model achieved an accuracy of 90.5% on the test set, with a precision of 86.5%, a recall of 92.7%, and an F1-score of 89.5%. The mean cross-validation accuracy reached 90.1% (±1.3%), indicating the model's consistency across various data subsets. The high recall value (92.7%) demonstrates the model's strong capability to identify actual anemic patients (true positives), thereby minimizing the risk of missed diagnoses that could endanger patients—a characteristic highly relevant for initial clinical screening scenarios. The confusion matrix for GNB is shown in Figure 4 below.

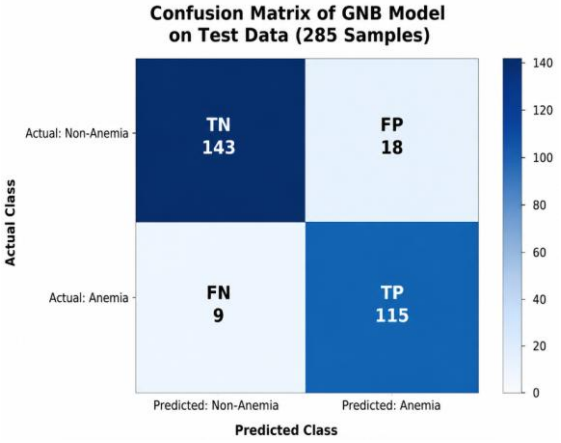

Figure 4. Gaussian Naïve Bayes Confusion Matrix

Analysis of the confusion matrix reveals that out of 285 test set samples, the model correctly classified 115 of the 124 anemia cases (true positives) and 143 of the 161 non-anemia cases (true negatives). There were 18 non-anemia cases incorrectly classified as anemia (false positives) and 9 anemia cases misclassified as non-anemia (false negatives). These misclassifications predominantly occurred in samples where hematological parameter values fell within borderline ranges.

The primary limitation of GNB lies in its assumption of feature independence when applied to hematological data. In reality, parameters such as Hb, MCV, MCH, and MCHC are clinically correlated; therefore, this assumption can diminish the accuracy of probability estimates in borderline cases—

aligning with the literature on the structural weaknesses of GNB. Nevertheless, GNB still delivers robust performance alongside high computational efficiency, making it a suitable candidate for initial clinical screening systems on devices with limited computational resources.

4.3 Classification using Random Forest

The Random Forest model was trained using the four original hematological parameters (Hb, MCV, MCH, and MCHC) as input features, shown in Figure 5, utilizing hyperparameters of 100 trees (n_estimators = 100) and the Gini impurity criterion. The feature importance analysis of the RF model yielded the following results: Hb exhibited the highest feature importance score of 0.562, followed by MCHC at 0.178, MCH at 0.145, and MCV at 0.115.

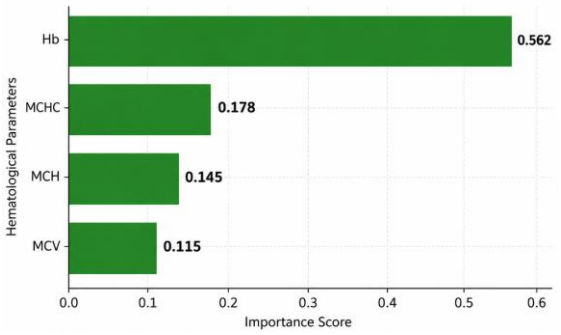

Figure 5. Feature Importance Model Random Forest

The feature importance results confirm the dominance of the Hb parameter as the primary predictor of anemia, consistent with the correlation analysis findings from the preprocessing stage ($r = -0.80$) and with the clinical definition of anemia based on Hb thresholds. Although the MCV, MCH, and MCHC parameters contribute less, they remain informative because RF can capture non-linear interactions among them. This capability represents a significant advantage of RF over GNB when dealing with datasets that contain correlated features. The matrix performance from RF is shown in Table 2 below.

Table 2. Random Forest Performance Table

Metric Evaluation	Score
Accuracy	94,2%
Precision	93,8%
Recall	95,1%
F1-Score	94,4%
CV Accuracy (mean ± std)	94,9% ± 1,1%

Based on the evaluation results, the Random Forest model achieved an accuracy of 94.2% on the test set, with a precision 93.8%, a recall 95.1%, and an F1-score 94.4%. The mean cross-validation accuracy was 94.9% (±1.1%), indicating greater model consistency than GNB across various data subsets. The confusion matrix is shown in Figure 6 below.

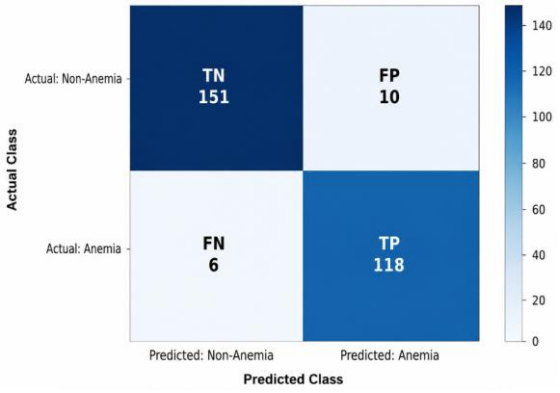

Figure 6. Random Confusion Matrix

Analysis of the confusion matrix for the Random Forest model reveals that out of 285 test set samples, the model correctly classified 118 of the 124 anemia cases (true positives) and 151 of the 161 non-anemia cases (true negatives). There were 10 non-anemia cases incorrectly classified as anemia (false positives) and 6 anemia cases misclassified as non-anemia (false negatives). Compared to GNB, RF demonstrated superior performance across all evaluation metrics. This superiority can be attributed to RF's ability to capture non-linear interactions among hematological parameters through its ensemble mechanism and bootstrap sampling. Furthermore, its inherent robustness to data noise yields more stable classifications, particularly in borderline cases that frequently lead to misclassifications in the GNB model.

4.4 Comparison of GNB and Random Forest Performance

Random Forest outperformed GNB across all evaluation metrics. RF achieved an accuracy of 94.2%, which is 3.7 percentage points higher than GNB (90.5%). The largest margin was observed in precision, where RF (93.8%) significantly outperformed GNB (86.5%) by 7.3 percentage points, indicating that RF is more accurate in classifying positive anemia cases while generating fewer false positives. Furthermore, the recall value for RF (95.1%) was higher than that of GNB (92.7%), demonstrating a superior capability in detecting true anemia cases. The consistency of the RF model was also evidenced by a lower cross-validation standard deviation (±1.1% compared to ±1.3% for GNB), which descriptively indicates reduced inter-fold performance variability—although the statistical significance of this difference has not yet been formally tested. Table 3 compares the performance of GNB and Random Forest.

Table 3. Performance comparison between GNB and Random Forest.

Metric Evaluation	GNB	Random Forest
Accuracy	90,5%	94,2%
Precision	86,5%	93,8%
Recall	92,7%	95,1%
F1-Score	89,5%	94,4%
CV Accuracy (mean ± std)	90,1% ± 1,3%	94,9% ± 1,1%

Nevertheless, GNB maintains distinct advantages in terms of computational speed and probabilistic interpretability. GNB can provide class probability estimates that can be used directly as diagnostic confidence indicators, which is highly valuable in clinical contexts that demand decision transparency. On the other hand, while RF excels in classification accuracy and model stability, it is inherently more of a "black-box" model. Consequently, its interpretation relies on feature importance scores or supplementary explainability methods such as SHAP.

To contextualize the findings of this study within the broader literature, Table 4 summarizes the comparison of anemia classification accuracy between the present study and several previous studies that utilized machine learning approaches based on hematological or Complete Blood Count (CBC) parameters.

Table 4. Comparison with other researchers on Anemia Prediction

Researcher	Model	Data Size	Accuracy
This research (2026)	Random Forest	1.421 sample	94,2%
This research (2026)	Gaussian Naive Bayes	1.421 sample	90,5%
Rahman [5] (2024)	Logistic Regression	1.000 sampel, 8 atribut	95,0%
Chandralekha [22] (2024)	Model Hybrid (ensemble)	Data klinis campuran	90,83%
Asʿad [11] (2023)	Gaussian Naive Bayes	Dataset medis umum	92,3%

This comparison indicates that the two models proposed in this study yield competitive results, although they have not yet achieved the highest performance across all benchmarks. The RF model in this study (94.2%) outperformed the GNB model in Asʿad's study [9] (92.3%) and performed competitively with the hybrid model in Chandralekha et al. [22] (90.83%). The GNB model in this study (90.5%) also performed slightly below the GNB results reported in Asʿad's study [11](92.3%), likely due to differences in dataset characteristics and class distributions. Overall, these findings reinforce a consistent pattern in the literature: Random Forest systematically outperforms Naive Bayes across various binary classification domains based on correlated numerical features. Simultaneously, it affirms that absolute accuracy heavily depends on the characteristics of the specific dataset used, necessitating careful interpretation of cross-study comparisons.

4.5 Research Limitation

Several methodological limitations in this study must be transparently acknowledged to ensure a balanced interpretation of the results and to guide future research directions:

a. Single data source without external validation: The dataset originates from a single public source (Kaggle) and lacks comprehensive clinical metadata (such as age, gender, and comorbidities). Consequently, the model's generalizability to real-world clinical populations, particularly in Indonesia, cannot be guaranteed without validation using external data.
b. Lack of statistical significance testing: The performance difference between GNB and RF is reported descriptively (mean ± standard deviation across a 5-fold CV) but has not been formally tested for statistical significance (e.g., using a paired McNemar's test or a paired t-test on the results of each fold). Thus, the claim regarding RF's superiority remains indicative rather than statistically conclusive.
c. Absence of hyperparameter optimization: The RF model parameters (n_estimators = 100, Gini criterion) utilized default scikit-learn values without any systematic tuning (such as grid search, random search, or Bayesian optimization). Therefore, the reported performance of the RF model could potentially be further improved.
d. Minimal handling of class imbalance: Although the class ratio (56.4%: 43.6%) is considered mild, this study did not implement resampling techniques, such as SMOTE, or class weighting. This omission could potentially affect the model's sensitivity toward the minority class in borderline cases.
e. Limited feature scope: The study utilized only four standard hematological parameters. Additional parameters that have been clinically proven to be relevant in the literature, such as ferritin levels, reticulocyte count, or red cell distribution width (RDW), were not included.
f. Model comparison limited to two algorithms: This research did not incorporate other comparative algorithms (such as SVM, XGBoost, or Logistic Regression) that have proven competitive in similar anemia studies. As a result, the assertion that RF is the superior approach applies exclusively to GNB, rather than across the entire spectrum of machine learning algorithms.

5. CONCLUSION

This study successfully developed and evaluated two machine learning models, namely Gaussian Naive Bayes (GNB) and Random Forest (RF), both individually and in comparison, for anemia classification using four hematological parameters. Several principal conclusions can be drawn as follows:

a. The GNB model achieved an accuracy of 90.5% with a recall of 92.7% on the hematological parameter-based anemia dataset. GNB demonstrated the capacity to provide robust probabilistic estimates and exhibited a notably high capability for anemia detection (true positives). However, its precision (86.5%) indicated more false positives than RF. GNB is well-suited for rapid initial screening due to its computational efficiency.
b. The Random Forest model achieved an accuracy of 94.2% with a recall of 95.1% and an F1-score of 94.4%, outperforming GNB across all evaluation metrics with an accuracy margin of 3.7 percentage points. RF proved more robust at capturing non-linear patterns among hematological parameters through its ensemble mechanism. The feature importance analysis of the RF model confirmed the dominance of the Hb parameter as the primary predictor of anemia (score 0.562), followed by MCHC, MCH, and MCV—a finding consistent with the correlation analysis ($r = -0.80$) and the literature on hematological physiology.
c. A direct comparison between GNB and RF demonstrated that RF delivered descriptively higher and more stable performance (CV accuracy ±1.1% vs. ±1.3%) on this hematological anemia dataset. However, GNB retains practical utility in clinical contexts due to its ability to generate directly interpretable probability estimations. Meanwhile, the superiority of RF must be interpreted with

the caveat that formal statistical significance testing has not yet been conducted in this study.

d. Comparisons with previous studies indicated that the results of both models are competitive, yet they have not achieved the highest performance among existing machine learning-based anemia studies. This suggests room for improvement through hyperparameter optimization, the inclusion of additional clinical features, and the application of advanced ensemble techniques such as stacking.

For future research, it is recommended to: (1) conduct external validation using real-world clinical data from healthcare facilities in Indonesia; (2) apply statistical significance testing (e.g., McNemar's test) to substantiate the model comparison claims; (3) systematically perform hyperparameter optimization; (4) explore class imbalance handling techniques such as SMOTE and class weighting; (5) incorporate other comparative algorithms, such as SVM, XGBoost, and Logistic Regression; (6) consider the addition of other hematological parameters like ferritin levels and reticulocyte counts; and (7) develop web-based interfaces or mobile applications equipped with explainable AI methods (e.g., SHAP) to facilitate the transparent and accountable implementation of these models in clinical practice.

REFERENCES

[1] K. Abbotts and P. Sriskandarajah, "Anemia of Chronic Disease: Pathophysiology, Diagnosis and Management," *Hematol. Rep.*, vol. 18, no. 4, p. 48, Jul. 2026, doi: 10.3390/hematolrep18040048.

[2] World Health Organization, "Anaemia," 2024, *World Health Organization, Geneva, Switzerland*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/anaemia

[3] G. A. Stevens *et al.*, "National, regional, and global estimates of anaemia by severity in women and children for 2000–19: a pooled analysis of population-representative data," *Lancet Glob. Health*, vol. 10, no. 5, pp. e627–e639, 2022, doi: 10.1016/S2214-109X(22)00084-5.

[4] A. G. Mugdha, F. T. Pinki, and S. K. Talukdhar, "Hemoglobin Estimation and Anemia Severity Prediction Using Machine Learning Algorithms," in *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, 2023, pp. 1–6. doi: 10.1109/STI59863.2023.10464565.

[5] Md. M. Rahman, M. U. Mojumdar, H. A. Shifa, N. R. Chakraborty, N. P. Stenin, and Md. A. Hasan, "Anemia Disease Prediction using Machine Learning Techniques and Performance Analysis," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2024, pp. 1276–1282. doi: 10.23919/INDIACom61295.2024.10498962.

[6] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/NEJMra1814259.

[7] A. Badhoutiya, R. P. Verma, A. Shrivastava, K. Laxminarayanamma, A. L. N. Rao, and A. K. Khan, "Random Forest Classification in Healthcare Decision Support for Disease Diagnosis," in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, 2023, pp. 1–7. doi: 10.1109/ICAIIHI57871.2023.10489244.

[8] M. Maydanchi *et al.*, "Comparative Study of Decision Tree, AdaBoost, Random Forest, Naïve Bayes, KNN, and Perceptron for Heart Disease Prediction," in *SoutheastCon 2023*, 2023, pp. 204–208. doi: 10.1109/SoutheastCon51012.2023.10115189.

[9] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," in *2019 International Engineering Conference (IEC)*, 2019, pp. 165–170. doi: 10.1109/IEC47844.2019.8950650.

[10] M. Maindola *et al.*, "Utilizing Random Forests for High-Accuracy Classification in Medical Diagnostics," in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 2024, pp. 1679–1685. doi: 10.1109/IC3I61595.2024.10828609.

[11] I. As'ad, "Advancing Healthcare Diagnostics: A Study on Gaussian Naive Bayes Classification of Blood Samples," *International Journal of Artificial Intelligence in Medical Issues*, vol. 1, no. 2, pp. 115–123, Nov. 2023, doi: 10.56705/ijaimi.v1i2.120.

[12] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," in *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, 2017, pp. 209–212. doi: 10.1109/AISP.2017.8324083.

[13] F.-J. Yang, "An Implementation of Naive Bayes Classifier," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.

[14] X. Hu, M. Cui, and B. Chen, "Feature selection based on random forest and application in correlation analysis of symptom and disease," in *2009 IEEE International Symposium on IT in Medicine & Education*, 2009, pp. 120–124. doi: 10.1109/ITIME.2009.5236450.

[15] D. M. Reif, A. A. Motsinger, B. A. McKinney, J. E. Crowe, and J. H. Moore, "Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types," in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006, pp. 1–8. doi: 10.1109/CIBCB.2006.330987.

[16] K. Swarupa, V. H. Sree, S. Nookambika, Y. K. S. Kishore, and U. R. Teja, "Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm," in *2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)*, 2021, pp. 48–51. doi: 10.1109/ICISSGT52025.2021.00021.

[17] S. Paul, P. Ranjan, S. Kumar, and A. Kumar, "Disease Predictor Using Random Forest Classifier," in *2022 International Conference for Advancement in Technology (ICONAT)*, 2022, pp. 1–4. doi: 10.1109/ICONAT53423.2022.9726023.

[18] B. R. Rao, "Anemia Dataset," 2022, *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset

[19] N. K. Mangalapu, L. T K, G. J, G. Kalyani, and B. K S, "Revolutionizing Anaemia Detection: Leveraging Eye Condition Data and Machine Learning," in *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, 2025, pp. 1–6. doi: 10.1109/IEMENTech65115.2025.10959598.

[20] D. E. Alarcón-Yaquetto *et al.*, "Hematological Parameters and Iron Status in Adult Men and Women Using Altitude Adjusted and Unadjusted Hemoglobin Values for Anemia Diagnosis in Cusco, Peru (3400 MASL)," *Physiologia*, vol. 2, no. 1, pp. 1–19, 2022, doi: 10.3390/physiologia2010001.

[21] 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, 2025, doi: 10.3390/informatics12010019.

[22] Sathya, G. S. Priyatharsini, A. Kumar. V, I. Vasudevan, and M. Umapathy, "Machine Learning Models for Predicting Anemia: Evaluation and Performance Insights," in *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, 2024, pp. 1–7. doi: 10.1109/ICICEC62498.2024.10808776.

[23] A. M, J. R, and Sasirekha. E, "Data to Diagnosis Machine Learning Models for Accurate Anemia Classifications," in *2025 International Conference on Data Science and Business Systems (ICDSBS)*, 2025, pp. 1–6. doi: 10.1109/ICDSBS63635.2025.11031746.

[24] Baik Budi and Rizky Sarhans, "Hybrid Intelligence For Precision Milk Quality Assessment Integrating Naïve Bayes And K-Means Clustering," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 14, no. 2, Apr. 2026, doi: 10.23960/jitet.v14i2.9357.

[25] Mohanarathinam, and J. Velusamy, "Optimized SMOTE for Imbalanced Data Handling in Machine Learning," in *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 2025, pp. 349–354. doi: 10.1109/InCACCT65424.2025.11011340.

[26] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding," in *2006 8th international Conference on Signal Processing*, 2006. doi: 10.1109/ICOSP.2006.345752.

[27] K. Pal and Biraj. V Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 83–87. doi: 10.1109/ICCMC48092.2020.ICCMC-00016.