Original Article

Check for updates

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 2, No. 2, 2024

Advanced Anemia Classification Using Comprehensive Hematological Profiles and Explainable Machine Learning Approaches

Teuku Rizky Noviandy [1], Ghifari Maulana Idroes [2], Rivansyah Suhendra [3], Tedy Kurniawan Bakri [4] and Rinaldi Idroes [5,*]

[1] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.)
[2] Department of Nuclear Engineering and Engineering Physics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia; ghifarimaulana145@gmail.com (G.M.I.)
[3] Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, Aceh Barat 23681, Indonesia; rivansyahsuhendra@utu.ac.id (R.S.)
[4] Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; tedykbakri@usk.ac.id (T.K.B.)
[5] School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

* Correspondence: rinaldi.idroes@usk.ac.id

Article History

Received 13 September 2024
Revised 16 November 2024
Accepted 21 November 2024
Available Online 29 November 2024

Keywords:
Hematological analysis
Data imbalance
Predictive algorithms
Clinical diagnostics
Health informatics

Abstract

Anemia is a common health issue with serious clinical effects, making timely and accurate diagnosis essential to prevent complications. This study explores the use of machine learning (ML) methods to classify anemia and its subtypes using detailed hematological data. Six ML models were tested: Gradient Boosting, Random Forest, Naive Bayes, Logistic Regression, Support Vector Machine, and K-Nearest Neighbors. The dataset was preprocessed using feature standardization and the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Gradient Boosting delivered the highest accuracy, sensitivity, and F1-score, establishing itself as the top-performing model. SHapley Additive exPlanations (SHAP) analysis was applied to enhance model interpretability, identifying key predictive features. This study highlights the potential of explainable ML to develop efficient, accurate, and scalable tools for anemia diagnosis, fostering improved healthcare outcomes globally.

1. Introduction

Anemia is one of the most prevalent global health disorders, characterized by reduced quantity or functional quality of red blood cells or hemoglobin [1]. This condition significantly impairs the body's capacity to transport sufficient oxygen to tissues, leading to various physiological and functional challenges [2]. According to the World Health Organization (WHO), anemia affects over 1.62 billion people worldwide, with a disproportionately high prevalence among vulnerable populations, including children, pregnant women, and the elderly [3]. This widespread occurrence of anemia and its potentially severe consequences highlights the pressing need for effective diagnostic and management strategies.

The clinical consequences of anemia vary in severity, ranging from reduced cognitive and physical performance to life-threatening complications such as organ dysfunction and cardiovascular strain [4, 5]. In children, anemia can lead to impaired growth,

developmental delays, and poor academic performance [6], while in adults, it is associated with fatigue, reduced work productivity, and decreased quality of life [7]. Among pregnant women, anemia increases the risk of maternal mortality, preterm delivery, and low birth weight [8, 9]. These wide-ranging effects underscore the critical importance of early detection and intervention.

Traditionally, anemia diagnosis relies on a combination of clinical evaluation and laboratory testing. Standard diagnostic protocols measure hematological parameters such as hemoglobin concentration, hematocrit or packed cell volume, red blood cell count and mean corpuscular volume [10]. These markers are supplemented by additional biomarkers, including mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, and red cell distribution width [11]. However, the conventional approach to anemia diagnosis is often time-consuming, labor-intensive, and susceptible to variability arising from human interpretation and laboratory discrepancies. Additionally, the cost and complexity of comprehensive testing can pose barriers to widespread implementation, particularly in resource-limited settings.

Recent machine learning (ML) advances are creating new ways to improve healthcare diagnostics [12–14]. ML, a subset of artificial intelligence (AI), leverages computational algorithms to identify patterns in complex, high-dimensional datasets, offering unparalleled potential for improving diagnostic accuracy and efficiency [15, 16]. Unlike traditional statistical methods, ML models excel at uncovering subtle, non-linear relationships within data, enabling the development of predictive tools that can significantly enhance clinical decision-making [17].

For the anemia classification task, ML offers unique advantages. Hematological data, rich in diverse and interrelated parameters, provides an ideal foundation for ML models to identify diagnostic patterns. Previous studies have demonstrated the feasibility of ML in anemia classification. For instance, Airlangga [18] highlights ML potential with a decision tree classifier, achieving 94.17% accuracy in anemia classification using complete blood count data. Similarly, advanced approaches integrating spatial attention mechanisms, like the AlexNet Multiple Spatial Attention model achieving 99.58% accuracy, underscore the ability of ML to revolutionize anemia detection [19].

Despite its promise, the application of ML in anemia diagnosis remains underexplored. Challenges such as data heterogeneity, limited availability of labeled datasets, and the need for explainability in clinical practice have impeded widespread adoption. Among these challenges, explainability is particularly critical to foster trust among clinicians and ensure ML models provide actionable insights aligned with medical expertise [20]. Clinicians require more than accurate predictions; they need to understand the reasoning behind model outputs to ensure alignment with medical knowledge and to guide clinical decisions confidently [21].

To address this, tools such as SHAP (SHapley Additive exPlanations) have emerged as leading solutions in explainable AI [22]. SHAP stands out because it offers consistent, locally interpretable explanations grounded in cooperative game theory. Unlike other methods, SHAP assigns precise feature attributions for individual predictions, ensuring robustness and fairness. For example, SHAP can clarify how specific blood parameters contribute to classifying anemia types, enabling clinicians to trace and validate the model's diagnostic logic [23]. Compared to alternative techniques like LIME, which relies on approximations that may vary across runs, SHAP provides more stable and mathematically consistent attributions, which is vital in the sensitive healthcare domain [24]. Adopting explainability methods like SHAP can bridge the gap between ML's diagnostic potential and its practical integration into clinical workflows, ultimately unlocking its full potential in anemia diagnosis.

This research aims to address these challenges by developing and evaluating ML models for anemia classification using comprehensive hematological profiles. Key blood parameters, such as RBC count, HGB, PCV, and MCV, are utilized to construct predictive models. These models are designed to enhance diagnostic accuracy and prioritize explainability, ensuring they can provide transparent insights into the reasoning behind their predictions. To achieve this, the study incorporates SHAP to interpret model outputs, offering clinicians detailed and consistent feature attributions for individual predictions. This focus on explainability is essential to foster trust and facilitate seamless integration into existing clinical workflows.

The study seeks to identify the most effective approaches for diagnosing anemia and its subtypes by comparing multiple ML algorithms and evaluating their diagnostic accuracy and interpretability. The findings are anticipated to contribute significantly to integrating ML into anemia diagnostics, aiming to improve global healthcare outcomes. By enabling timely, accurate, and interpretable anemia detection, this research has the potential to drive innovation in diagnostic methodologies, reduce healthcare disparities, and ultimately enhance patient care worldwide.

Table 1. Overview of hematological features used in the dataset for anemia classification.

No.	Feature	Description
1	Age	Current age of the patient (in years)
2	Gender	Gender of the patient (Male/Female)
3	HGB	Hemoglobin level (g/dL)
4	MCV	Mean cell volume (fL)
5	MCH	Mean cell hemoglobin (pg)
6	MCHC	Mean cell hemoglobin concentration (g/dL)
7	RDW	Red cell distribution width (%)
8	RBC	Red blood cell count (M/uL)
9	WBC	White blood cell count (ths/uL)
10	PLT	Platelet count (10^3/uL)
11	PCV	Packed cell volume (%)

2. Materials and Methods

2.1. Dataset

The dataset for this study was retrieved from Mendeley Data, as provided by Vohra et al. [25]. It includes a detailed account of anemia prevalence, severity, and association with age and gender. The dataset is based on complete blood count (CBC) parameters obtained from tests conducted using a Hematology analyzer at Eureka Diagnostic Center, Lucknow, India. All CBC tests followed standard operating protocols defined for the Hematology analyzer.

During the data collection period from September 2020 to December 2020, 1,000 CBC investigations were performed at the diagnostic center, which typically conducts 4–8 CBC tests daily. From these, 400 patient samples were randomly selected. After excluding individuals such as pregnant women, children under 10 years of age, and individuals with incomplete records, the final dataset included 364 adult patients aged 15. The sample consisted of both male and female participants. The dataset contains 11 features recorded during the CBC tests. Table 1 outlines these features, along with their descriptions.

2.2. Class Label Engineering

The labeling process for anemia classification involved engineering a binary class label to differentiate between anemic (Class 1) and non-anemic (Class 0) cases. This was achieved using several hematological parameters, including hemoglobin (HGB), red blood cell count (RBC), packed cell volume (PCV), and additional indices such as mean cell volume (MCV), mean cell hemoglobin (MCH), and mean cell hemoglobin concentration (MCHC).

Patients were labeled anemic (Class 1) if their HGB levels were below 13 g/dL for males or below 12 g/dL for females. This condition was further refined with additional criteria. If the RBC count was lower than 4.5 million/µL for males or 4.2 million/µL for females, or if the PCV fell below 40% for males or 36% for females, the patient was also labeled anemic. Furthermore, MCV, MCH, or MCHC abnormalities were incorporated to distinguish anemia subtypes, such as microcytic or macrocytic anemia, and to identify underlying conditions like iron deficiency or hemolytic anemia.

In contrast, patients were labeled as non-anemic (Class 0) if all their hematological values, including HGB, RBC, and PCV, fell within the normal range for their gender. Additionally, non-anemic cases showed no abnormalities in secondary indices such as MCV, MCH, or MCHC, ensuring that only healthy profiles were included in this category.

The labeling process resulted in 254 patients being categorized as anemic and 110 as non-anemic. This distribution reflects the prevalence of anemia within the dataset and provides a clear foundation for ML model development.

2.3. Dataset Preprocessing

Several preprocessing steps were undertaken to prepare the dataset for ML analysis to ensure effective and unbiased model training and evaluation. First, the data was standardized by scaling all numerical features to have a mean of zero and a standard deviation of one. Standardization ensures that features are on a comparable scale, which is particularly important for algorithms sensitive to feature magnitude, such as distance-based or gradient-based models [26].

After standardization, the dataset was split into training and testing subsets, with 80% of the data allocated to the training set and 20% reserved for testing. This split allowed the models to learn patterns from most data while preserving an independent subset for evaluating model performance on unseen data [27].

The class distribution in the dataset revealed an imbalance, with 254 cases labeled as anemic (Class 1) and 110 cases labeled as non-anemic (Class 0). To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set [28, 29]. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, effectively balancing the class distribution. The process is mathematically represented as shown in Equation 1:

$$x_{\text{synthetic}} = x_i + \lambda \cdot (x_{\text{nn}} - x_i) \quad (1)$$

where x_i represents an existing minority class sample, x_{nn} is a randomly chosen sample from the k-nearest neighbors of x_i, and λ is a random value uniformly distributed in the range $[0,1]$. This formula generates a

synthetic sample $x_{\text{synthetic}}$ by interpolating between x_i and x_{nn}, helping to balance the dataset by adding new, plausible samples to the minority class.

Applying SMOTE prevents the model from becoming biased toward the majority class and improves its ability to correctly identify the minority class. This step is crucial for ensuring the model performs well across both classes, especially in medical applications where identifying minority cases can significantly impact patient outcomes.

2.4. Model Training

To classify anemia, we trained six ML models: Random Forest, Gradient Boosting, Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models were chosen due to their diverse learning approaches, which allowed us to compare their performance and identify the best classifier for anemia detection.

- Random Forest was selected for its ability to handle high-dimensional data and robustness against overfitting through ensemble learning.
- Gradient Boosting was included for its strong predictive capabilities, especially for complex-pattern datasets.
- Naive Bayes was chosen due to its simplicity and efficiency, particularly for probabilistic learning in classification tasks.
- Logistic Regression provided a baseline linear model, offering interpretability and suitability for binary classification problems.
- SVM was used for its capacity to create complex decision boundaries through kernel functions, making it effective for non-linear relationships.
- KNN was included for its instance-based learning approach, which relies on the similarity of data points and is straightforward to implement.

All models were trained using 5-fold cross-validation (CV) [30]. This method involves dividing the dataset into five subsets, training the model on four subsets, and validating it on the remaining subset. The process is repeated five times, with each subset used as the validation set once. This approach provides a more reliable assessment of model performance by reducing the risk of overfitting and ensuring that a specific train-test split does not bias the evaluation.

The models were trained with default hyperparameters to establish a baseline performance. A fixed random state of 42 was used across all models to ensure reproducibility, allowing consistent results in randomness, such as in data splitting or the initialization of model components. By combining diverse models, a rigorous evaluation method, and consistent training conditions, we aimed to identify the most effective approach for anemia classification.

2.5. Model Performance Assessment

The performance of the models was assessed using several evaluation metrics: accuracy, precision, sensitivity (recall), specificity, and F1-score [31, 32], calculated using Equation 2-6, respectively. These metrics provide a comprehensive understanding of model performance, balancing overall correctness (accuracy), the ability to identify true positives (sensitivity), the ability to avoid false positives (specificity), and the harmonic mean of precision and sensitivity (F1-score). These metrics are particularly important in medical contexts, where false negatives can have critical consequences and a balanced evaluation is required.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$F1\text{-}Score = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \quad (6)$$

Where TP is True Positives, representing the cases correctly predicted as positive; TN is True Negatives, the cases correctly predicted as negative; FP is False Positives, which are negative cases incorrectly predicted as positive; and FN is False Negatives, the positive cases incorrectly predicted as negative [33].

To further analyze the impact of data balancing, we compared the performance of models trained on datasets with and without SMOTE oversampling. This comparison aimed to determine whether SMOTE improved the models' ability to handle the imbalanced dataset and provide better classification outcomes.

Additionally, the models were ranked based on their performance across the evaluation metrics. We calculated a "payoff" value for each model to identify the best model. The payoff is derived from the aggregated rank of a model across all metrics, where the model with the lowest payoff value is considered the best [34]. This method ensures a balanced evaluation by considering

Table 2. Performance metrics of machine learning models trained on the SMOTE-augmented dataset.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Random Forest	97.47	97.47	97.47	97.47	97.47
Gradient Boosting	97.98	97.98	97.98	98.48	97.98
Naive Bayes	92.17	92.35	92.17	95.45	92.16
Logistic Regression	94.95	95.06	94.95	97.47	94.95
SVM	95.45	95.57	95.45	97.98	95.45
KNN	93.43	94.2	93.43	100	93.4

Table 3. Performance metrics of machine learning models trained on the original dataset.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Random Forest	95.88	95.88	95.88	93.55	95.88
Gradient Boosting	97.25	97.25	97.25	94.62	97.24
Naive Bayes	89.69	90.32	89.69	90.32	89.84
Logistic Regression	92.78	92.77	92.78	88.17	92.77
SVM	92.44	92.65	92.44	91.4	92.5
KNN	90.38	90.52	90.38	87.1	90.43

multiple performance aspects, providing a robust approach to select the optimal model for anemia classification.

$$Payoff = \sum_{i=1}^{n} Rank_{Metric_i} \quad (7)$$

where $Rank_{Metric_i}$ represents the rank of the model for the i-th evaluation metric, and n is the total number of metrics considered.

2.6. Model Explainability

To ensure the ML models' interpretability, SHAP was employed to analyze each feature's contribution to the predictions. This approach enables transparency by quantifying how specific hematological parameters influence the classification of anemia [22]. Global feature importance was summarized using SHAP values, visualized through bar plots and bee swarm plots to provide complementary perspectives.

The bar plot highlights each feature's mean absolute SHAP values, ranking them by their overall contribution to the model's predictions. This concise visualization identifies the key hematological parameters most strongly influencing the model's output, making it especially valuable for medical professionals seeking straightforward insights. On the other hand, the bee swarm plot reveals the distribution of SHAP values for individual cases, showcasing how features impact predictions differently across the dataset. By visualizing variability and interactions at a granular level, the bee swarm plot provides a deeper understanding of complex patterns, complementing the broader insights from the bar plot. These visualizations enhance interpretability, ensuring the model's decisions are understandable and actionable.

3. Results and Discussion

We trained six ML models to classify anemia. To evaluate the impact of oversampling, we trained and tested the models on both the SMOTE-augmented and original datasets. The models' performance on the SMOTE-augmented dataset is summarized in Table 2, while their performance on the original dataset is shown in Table 3.

These tables show that applying SMOTE improves the models' sensitivity and precision, particularly for classifiers like KNN and Logistic Regression, which are sensitive to imbalanced datasets. The increase in F1-Score and specificity across most models after SMOTE augmentation suggests a more balanced classification of both anemic and non-anemic cases. For instance, the Random Forest model achieved an accuracy of 97.47% with SMOTE compared to 95.88% without it. Similarly, Gradient Boosting demonstrated improved metrics across the board with SMOTE, achieving the highest accuracy and sensitivity (97.98% and 97.98%, respectively) among all models.

In contrast, models trained on the original dataset generally exhibited slightly lower sensitivity, indicating a bias toward the majority class. This was particularly noticeable in models like Naive Bayes and KNN, which rely heavily on balanced data for optimal performance. The SMOTE-augmented dataset also allowed for improved specificity, reflecting the models' ability to correctly classify non-anemic cases, which is critical in ensuring reliable diagnostic outcomes. Gradient Boosting, in particular, achieved consistent performance improvements with SMOTE, making it a strong candidate for anemia classification.

Because the application of SMOTE significantly improved model performance, we further evaluated the models by

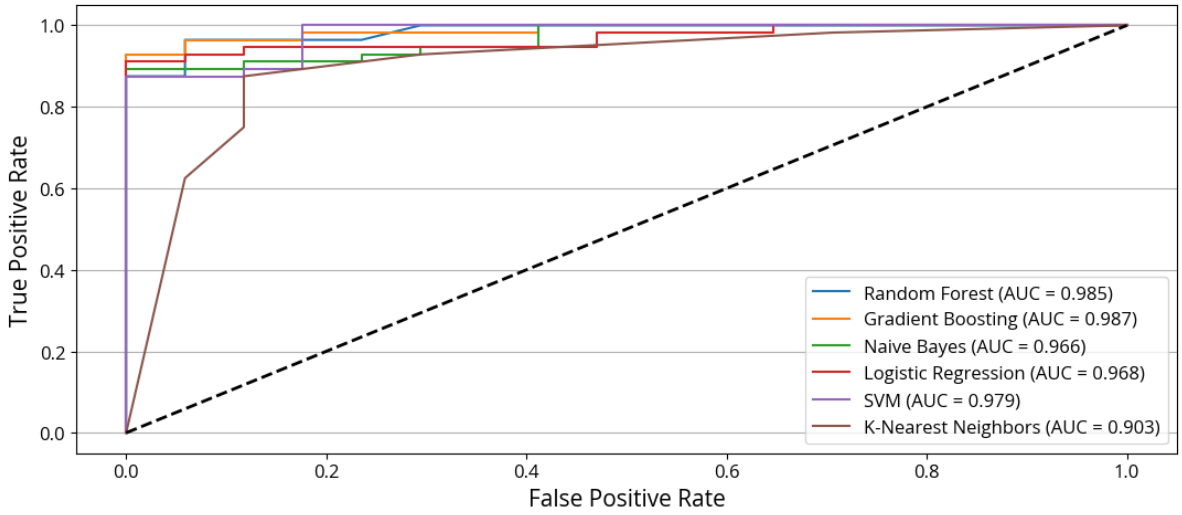

Figure 1. ROC curves comparing the classification performance of machine learning models.

Table 4. Ranking and payoff values of machine learning models with SMOTE based on aggregated performance metrics.

Model	Rank					Payoff
	Accuracy	Precision	Sensitivity	Specificity	F1-Score	
Gradient Boosting	1	1	1	2	1	6
Random Forest	2	2	2	4	2	12
SVM	5	5	5	3	5	23
Logistic Regression	6	6	6	4	6	28
KNN	7	7	7	1	7	29
Naive Bayesian	10	10	10	5	10	45

visualizing their Receiver Operating Characteristic (ROC) curves, as shown in Figure 1. The ROC curves provide a detailed view of the trade-off between the True Positive Rate (sensitivity) and False Positive Rate (1-specificity) across various thresholds, enabling a comparative analysis of the models' classification capabilities. The Area Under the Curve (AUC) summarizes each model's ability to distinguish between anemic and non-anemic cases.

From the figure, Gradient Boosting exhibited the highest AUC value (0.987), indicating superior classification performance compared to the other models. Random Forest closely followed with an AUC of 0.985, confirming its effectiveness as a classifier. Logistic Regression and SVM also performed well, achieving AUC values of 0.968 and 0.979, respectively. Naive Bayes had a slightly lower AUC of 0.966, while KNN showed the least effective performance among the models with an AUC of 0.903.

The diagonal dashed line represents the performance of a random classifier with an AUC of 0.5, serving as a baseline for comparison. All models substantially outperformed this baseline, validating their predictive reliability. The steep curves observed for Gradient Boosting, Random Forest, and SVM indicate higher sensitivity and specificity across different thresholds, reinforcing their robustness as classifiers for anemia detection. This visualization highlights SMOTE's efficacy in enhancing the models' discriminatory power, particularly for Gradient Boosting and Random Forest, which emerged as the most reliable classifiers in this study.

To determine the overall performance of the models, we ranked them based on each evaluation metric. The ranks for all metrics were summed to calculate the "payoff" value for each model, as shown in Table 4. The model with the lowest payoff value is considered the best-performing, consistently ranking higher across the evaluation metrics. From the table, Gradient Boosting with SMOTE achieved the lowest payoff value of 6, indicating its superior performance across all metrics, with top rankings in accuracy, precision, sensitivity, and F1-score and a second-place ranking in specificity. Random Forest with SMOTE followed as the second-best model, with a payoff value of 12, maintaining high rankings across all metrics. SVM with SMOTE achieved a payoff of 23, placing it third overall. The Logistic Regression and KNN models with SMOTE showed moderate performance with payoff values of 28 and 29, respectively. At the same time, Naive Bayes with SMOTE

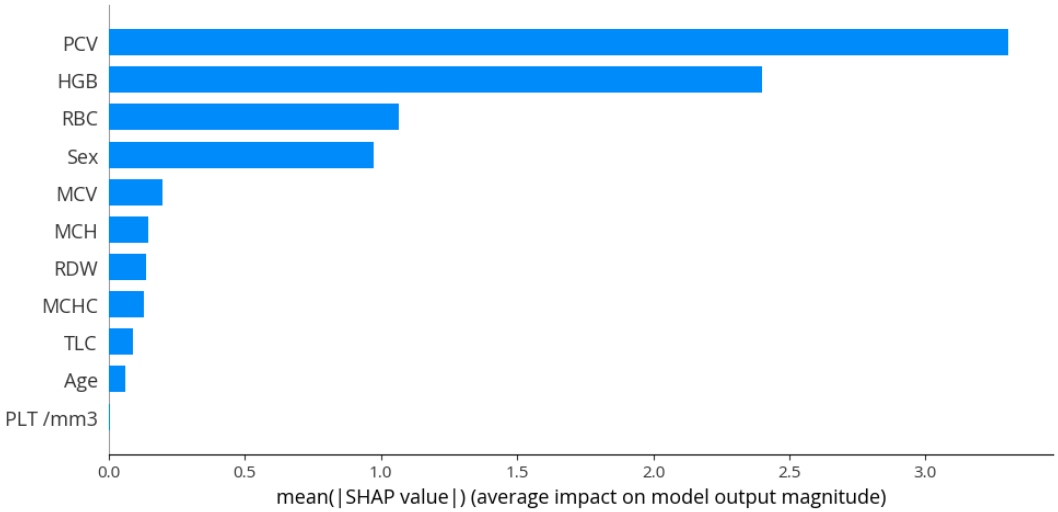

Figure 2. SHAP summary bar plot showing the average impact of each feature on the Gradient Boosting model's predictions for anemia classification.

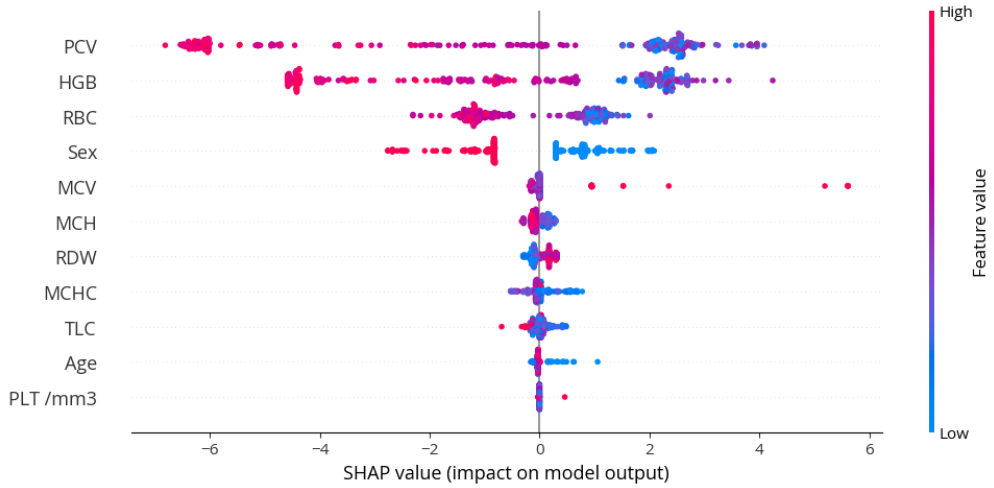

Figure 3. SHAP bee swarm plot illustrating the distribution of feature impacts on individual predictions in the Gradient Boosting model for anemia classification.

ranked the lowest with a payoff value of 45 due to consistently lower metric rankings.

Because Gradient Boosting emerged as the best-performing model in this study, we proceeded with its explainability using SHAP. The summary of feature importance for the Gradient Boosting model, visualized through SHAP values, is shown in Figure 2. From the results, PCV was the most influential feature in the Gradient Boosting model, contributing significantly to the classification of anemia. This aligns with the clinical understanding of anemia, as PCV directly reflects the proportion of red blood cells in the blood. HGB and red RBC were also highly impactful, further supporting their well-established roles in diagnosing anemia. Gender was the fourth most important feature, indicating its influence in setting thresholds for hematological parameters such as HGB and RBC. Other features, such as MCV and MCH, had moderate contributions, consistent with their relevance in identifying anemia subtypes. Meanwhile, PLT and age showed minimal impact on the model's predictions, suggesting that they are less critical for the classification task in this dataset. This analysis underscores the model's alignment with clinical expectations, enhancing confidence in its utility for anemia detection.

In addition to the summary bar plot, Figure 3 presents the bee swarm plot, which provides a detailed view of the distribution of SHAP values for individual features in the Gradient Boosting model. Each dot represents a single prediction, with its position indicating the impact of a feature on the model's output and the color denoting the feature's value. The bee swarm plot highlights the variability in how features like PCV, HGB, and RBC influence predictions across the dataset. For example,

higher PCV values are associated with a strong negative impact on the likelihood of being classified as frail, as indicated by the clustering of blue dots on the negative SHAP value axis. Conversely, lower PCV values (red dots) contribute positively, pushing predictions toward anemia classification. Similar patterns are observed for HGB and RBC, further affirming their critical role in anemia detection. Other features, such as MCV and MCH, exhibit more dispersed contributions, reflecting their secondary but still notable influence on the model's decisions. Overall, the bee swarm plot complements the summary bar chart by showcasing the heterogeneity in feature impacts, providing deeper insights into the model's behavior at the individual prediction level.

The overall results indicate that Gradient Boosting with SMOTE outperformed other models in classification accuracy, precision, sensitivity, specificity, and F1-score, achieving the lowest payoff values. Gradient Boosting emerged as the best model, benefiting from its iterative nature of building weak learners and its ability to capture complex, non-linear relationships in the data. Random Forest followed closely due to its ensemble learning approach, which effectively reduces overfitting and handles feature interactions. On the other hand, models like Naive Bayes and KNN performed relatively poorly, with higher payoff values and lower AUC scores. Naive Bayes struggled due to its simplistic feature independence assumption, which is unrealistic for hematological data with interdependent parameters. KNN's performance was hindered by its sensitivity to class imbalance and reliance on distance metrics, which can be distorted in high-dimensional spaces.

The difference in model performance highlights the importance of understanding the strengths and limitations of each algorithm when applied to medical datasets. Complex ensemble methods like Gradient Boosting and Random Forest excel at handling the inherent variability and interdependence of hematological features, as evidenced by their superior performance metrics. Moreover, applying SHAP explainability to these models further demonstrated their robustness, providing insight into how key features such as PCV, HGB, and RBC drive the predictions. Simpler models, on the other hand, may not adequately capture these intricate patterns, resulting in reduced predictive performance.

Additionally, using SMOTE proved crucial in balancing the dataset and improving model performance, particularly for classifiers sensitive to imbalanced data, such as KNN and Logistic Regression. SMOTE enhanced the sensitivity of these models and helped ensure a more equitable classification of anemic and non-anemic cases. Without SMOTE, models exhibited reduced sensitivity, favoring the majority class and potentially leading to missed anemia diagnoses. SHAP analysis reinforced these findings by highlighting how feature contributions became more balanced and aligned with clinical expectations in SMOTE-augmented models, emphasizing the importance of data preprocessing and model explainability in medical applications.

This study has several important implications. First, it demonstrates the value of leveraging advanced ML techniques and data preprocessing methods, such as SMOTE, in addressing the challenges of imbalanced medical datasets. By identifying Gradient Boosting and Random Forest as the most effective models, this research provides a foundation for implementing automated anemia diagnostic systems in clinical settings. These systems could support healthcare professionals by offering accurate, fast, and scalable solutions for anemia classification, ultimately improving early detection and patient outcomes. Furthermore, this study underscores the importance of rigorous model evaluation, including multiple metrics and payoff-based rankings, to ensure that the chosen classifier is accurate but also balanced, and robust.

4. Conclusions

This study demonstrated the potential of ML models, particularly Gradient Boosting, in advancing anemia classification using comprehensive hematological profiles. Gradient Boosting emerged as the best-performing model, achieving the highest accuracy (97.98%), sensitivity, and F1-score, confirming the effectiveness of ensemble methods in capturing complex, non-linear patterns in medical data. The application of SMOTE oversampling significantly enhanced the models' ability to handle class imbalance, improving sensitivity and specificity, especially for imbalanced datasets. These findings underscore the transformative potential of ML in anemia diagnostics, offering accurate, scalable, and reliable tools for early detection and classification, ultimately paving the way for improved patient outcomes and healthcare efficiency.

Author Contributions: Conceptualization, T.R.N., G.M.I. and R.I.; methodology, T.R.N. and R.I.; software, T.R.N. and R.S.; validation, T.K.B. and R.I.; formal analysis, T.R.N. and G.M.I.; investigation, T.R.N. and R.S.; resources, R.I.; data curation, T.K.B. and R.I.; writing—original draft preparation, T.R.N., G.M.I. and R.S.; writing—review and editing, T.K.B. and R.I.; visualization, G.M.I.; supervision, R.I.; project administration, R.I.; funding acquisition, R.I. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data utilized in this study is publicly available and can be accessed at https://data.mendeley.com/datasets/dy9mfjchm7/1.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Garcia-Casal, M. N., Dary, O., Jefferds, M. E., and Pasricha, S. (2023). Diagnosing Anemia: Challenges Selecting Methods, Addressing Underlying Causes, and Implementing Actions at the Public Health Level, *Annals of the New York Academy of Sciences*, Vol. 1524, No. 1, 37–50. doi:10.1111/nyas.14996.
2. Simon, G. I., Craswell, A., Thom, O., Chew, M. S., Anstey, C. M., and Fung, Y. L. (2019). Impacts of Aging on Anemia Tolerance, Transfusion Thresholds, and Patient Blood Management, *Transfusion Medicine Reviews*, Vol. 33, No. 3, 154–161. doi:10.1016/j.tmrv.2019.03.001.
3. Shah, S. A., Soomro, U., Ali, O., Tariq, Y., Waleed, M. S., Guntipalli, P., and Younus, N. (2023). The Prevalence of Anemia in Working Women, *Cureus*. doi:10.7759/cureus.44104.
4. He, W., Ruan, Y., Yuan, C., Luan, X., and He, J. (2020). Hemoglobin, Anemia, and Poststroke Cognitive Impairment: A Cohort Study, *International Journal of Geriatric Psychiatry*, Vol. 35, No. 5, 564–571. doi:10.1002/gps.5272.
5. Wiciński, M., Liczner, G., Cadelski, K., Kołnierzak, T., Nowaczewska, M., and Malinowski, B. (2020). Anemia of Chronic Diseases: Wider Diagnostics—Better Treatment?, *Nutrients*, Vol. 12, No. 6, 1784. doi:10.3390/nu12061784.
6. Samson, K. L. I., Fischer, J. A. J., and Roche, M. L. (2022). Iron Status, Anemia, and Iron Interventions and Their Associations with Cognitive and Academic Performance in Adolescents: A Systematic Review, *Nutrients*, Vol. 14, No. 1, 224. doi:10.3390/nu14010224.
7. van Haalen, H., Jackson, J., Spinowitz, B., Milligan, G., and Moon, R. (2020). Impact of Chronic Kidney Disease and Anemia on Health-Related Quality of Life and Work Productivity: Analysis of Multinational Real-World Data, *BMC Nephrology*, Vol. 21, No. 1, 88. doi:10.1186/s12882-020-01746-4.
8. Noviandy, T. R., Nainggolan, S. I., Raihan, R., Firmansyah, I., and Idroes, R. (2023). Maternal Health Risk Detection Using Light Gradient Boosting Machine Approach, *Infolitika Journal of Data Science*, Vol. 1, No. 2, 48–55. doi:10.60084/ijds.v1i2.123.
9. Kabir, M. A., Rahman, M. M., and Khan, M. N. (2022). Maternal Anemia and Risk of Adverse Maternal Health and Birth Outcomes in Bangladesh: A Nationwide Population-Based Survey, *PLOS ONE*, Vol. 17, No. 12, e0277654. doi:10.1371/journal.pone.0277654.
10. Hemoglobinometry, A., Red, C., Histogram, E., and Width, R. C. D. (2015). Principles and Practice of Clinical Hematology, *Linne & Ringsrud's Clinical Laboratory Science-E-Book: The Basics and Routine Techniques*, Vol. 2, 291.
11. Said, A. S., Spinella, P. C., Hartman, M. E., Steffen, K. M., Jackups, R., Holubkov, R., Wallendorf, M., and Doctor, A. (2017). RBC Distribution Width: Biomarker for Red Cell Dysfunction and Critical Illness Outcome?, *Pediatric Critical Care Medicine*, Vol. 18, No. 2, 134–142. doi:10.1097/PCC.0000000000001017.
12. Solomon, D. D., Khan, S., Garg, S., Gupta, G., Almjally, A., Alabduallah, B. I., Alsagri, H. S., Ibrahim, M. M., and Abdallah, A. M. A. (2023). Hybrid Majority Voting: Prediction and Classification Model for Obesity, *Diagnostics*, Vol. 13, No. 15, 2610. doi:10.3390/diagnostics13152610.
13. Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., Noviandy, T. R., Sasmita, N. R., Subianto, M., Earlia, N., Niode, N. J., and Idroes, R. (2023). Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification, Heca Journal of Applied Sciences, Vol. 1, No. 2, 54–61. doi:10.60084/hjas.v1i2.85.
14. Noviandy, T. R., Alfanshury, M. H., Abidin, T. F., and Riza, H. (2023). Enhancing Glioma Grading Performance: A Comparative Study on Feature Selection Techniques and Ensemble Machine Learning, *2023 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, IEEE, 406–411. doi:10.1109/IC3INA60834.2023.10285778.
15. Noviandy, T. R., Nisa, K., Idroes, G. M., Hardi, I., and Sasmita, N. R. (2024). Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM, *Journal of Computing Theories and Applications*, Vol. 2, No. 2, 138–147. doi:10.62411/jcta.10129.
16. Rufo, D. D., Debelee, T. G., Ibenthal, A., and Negera, W. G. (2021). Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM), *Diagnostics*, Vol. 11, No. 9, 1714. doi:10.3390/diagnostics11091714.
17. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.
18. Airlangga, G. (2024). Leveraging Machine Learning for Accurate Anemia Diagnosis Using Complete Blood Count Data, *Indonesian Journal of Artificial Intelligence and Data Mining*, Vol. 7, No. 2, 318. doi:10.24014/ijaidm.v7i2.29869.
19. Ramzan, M., Sheng, J., Saeed, M. U., Wang, B., and Duraihem, F. Z. (2024). Revolutionizing Anemia Detection: Integrative Machine Learning Models and Advanced Attention Mechanisms, *Visual Computing for Industry, Biomedicine, and Art*, Vol. 7, No. 1, 18. doi:10.1186/s42492-024-00169-4.
20. Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., and Mooney, C. (2021). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review, *Applied Sciences*, Vol. 11, No. 11, 5088. doi:10.3390/app11115088.
21. Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., and Moosa, M. (2023). The Enlightening Role of Explainable Artificial Intelligence in Medical & Healthcare Domains: A Systematic Literature Review, *Computers in Biology and Medicine*, Vol. 166, 107555. doi:10.1016/j.compbiomed.2023.107555.
22. Lundberg, S. M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, Vol. 30.
23. Nohara, Y., Matsumoto, K., Soejima, H., and Nakashima, N. (2022). Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital, *Computer Methods and Programs in Biomedicine*, Vol. 214, 106584. doi:10.1016/j.cmpb.2021.106584.
24. Gramegna, A., and Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk, *Frontiers in Artificial Intelligence*, Vol. 4. doi:10.3389/frai.2021.752558.
25. Vohra, R., Pahareeya, J., and Hussain, A. (2021). Complete Blood Count Anemia Diagnosis, *Mendeley Data*. doi:10.17632/dy9mfjchm7.1.
26. Gunda, T., Hackett, S., Kraus, L., Downs, C., Jones, R., McNalley, C., Bolen, M., and Walker, A. (2020). A Machine Learning Evaluation of Maintenance Records for Common Failure Modes in PV Inverters, *IEEE Access*, Vol. 8, 211610–211620. doi:10.1109/ACCESS.2020.3039182.
27. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

28. Chawla, N. V, Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique, *Journal of Artificial Intelligence Research*, Vol. 16, 321–357.

29. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques, *Indatu Journal of Management and Accounting*, Vol. 1, No. 1, 29–35. doi:10.60084/ijma.v1i1.78.

30. Berrar, D. (2019). Cross-Validation, *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 542–545. doi:10.1016/B978-0-12-809633-8.20349-X.

31. Noviandy, T. R., Zahriah, Z., Yandri, E., Jalil, Z., Yusuf, M., Yusof, N. I. S. M., Lala, A., and Idroes, R. (2024). Machine Learning for Early Detection of Dropout Risks and Academic Excellence: A Stacked Classifier Approach, *Journal of Educational Management and Learning*, Vol. 2, No. 1, 28–34. doi:10.60084/jeml.v2i1.191.

32. Pratyusha, M., and Kanimozhi, K. V. (2022). Heart Disease Prediction Using Decision Tree in Comparison with K-Nearest Neighbor to Improve Accuracy, *Advances in Parallel Computing*, Vol. 0, No. 41, 231–236. doi:10.3233/APC220031.

33. Idroes, G. M., Noviandy, T. R., Maulana, A., Zahriah, Z., Suhendrayatna, S., Suhartono, E., Khairan, K., Kusumo, F., Helwani, Z., and Abd Rahman, S. (2023). Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring, *Leuser Journal of Environmental Studies*, Vol. 1, No. 2, 62–68. doi:10.60084/ljes.v1i2.99.

34. Magazzino, C., Madaleno, M., Waqas, M., and Leogrande, A. (2024). Exploring the Determinants of Methane Emissions from a Worldwide Perspective Using Panel Data and Machine Learning Analyses, *Environmental Pollution*, Vol. 348, 123807. doi:10.1016/j.envpol.2024.123807.