Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00325-2

Original Article

Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification

Pradip R. Bhakare [1], Chandrakant M. Jadhao [1], Jayesh K. Kokate [1]

1. *Department of Electronics and Telecommunication Engineering, Mauli Group of Institution's, College of Engineering and Technology, Shegaon, IND*

Received: July 18, 2026 | Review began: August 18, 2026 | Review ended: September 04, 2026 | Published: October 15, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Anaemia is a common public health problem that can benefit from accurate and timely identification to inform effective medical practice. This paper experimentally investigates the performance of machine learning models for the binary classification of anaemia based on complete blood count (CBC) information and relevant haematological variables. We created a comprehensive dataset of 2,785 CBC results obtained from publicly available sources, including haemoglobin, red blood cell count, packed cell volume, and red cell indices. After data preprocessing and correlation analysis for missing values and multicollinearity, we used principal component analysis for dimensionality reduction. We trained Logistic Regression, Random Forest, and XGBoost models on an 80:20 split for model training and testing and evaluated them using stratified 5-fold cross-validation. Accuracy, precision, recall, F1-score, and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) were used as performance metrics. The findings demonstrate that ensemble models perform better than Logistic Regression, with XGBoost performing best: accuracy of 98.92%, F1-score of 0.9906, and ROC-AUC of 1.00. The high recall of the models indicates their excellent potential for anaemia detection, while their precision indicates their suitability for accurate decision-making in anaemia diagnosis.

Categories: AI applications, AI/ML-based decision support systems, Artificial Intelligence and Machine Learning in the Cloud

Keywords: complete blood count, binary classification, ai-ml, healthcare analytics, invasive method

Introduction

Anaemia is a common haematological disorder affecting approximately 1.62 billion people worldwide and is characterised by reduced haemoglobin concentration or red blood cell count, leading to diminished oxygen-carrying capacity of the blood. The World Health Organization (WHO) provides standardised haemoglobin thresholds for the diagnosis and assessment of anaemia severity, facilitating uniform clinical evaluation [1]. According to the National Family Health Survey (NFHS-5), anaemia remains a significant public health concern in India, particularly among women and children [2]. Early and accurate diagnosis is essential because different forms of anaemia require distinct therapeutic interventions. Conventional diagnosis relies on Complete Blood Count (CBC) analysis performed by trained professionals. This process can be time-consuming and resource-intensive, particularly in low-resource healthcare settings.

Some of the recent developments in the field of machine learning (ML) are providing new possibilities in the automated diagnostics of anaemia. ML algorithms enable the identification of hidden patterns in CBC variables that may not be readily apparent to physicians [3]. The binary classification that divides patients into anaemic and non-anaemic categories is the basis of diagnosis and has attracted a lot of attention lately [4]. Yet the success of ML-based diagnostic systems is

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

greatly influenced by how accurately their performance is evaluated. Although accuracy is commonly used for model evaluation, it can provide misleading results when the dataset is imbalanced or when different types of classification errors have different consequences [5].

As a result, clinically relevant metrics, like sensitivity, specificity, precision, F1 score, and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), are now used more often to evaluate the models' performance. Additionally, stratified k-fold cross-validation has become a popular way to validate the models and ensure their robustness across different patient groups. Nevertheless, despite recent progress in the field, practical implementation of ML models in a clinical setting continues to face obstacles because of the absence of uniform evaluation approaches.

To address these limitations, the present study emphasises rigorous performance assessment using stratified k-fold cross-validation alongside clinically relevant evaluation metrics. The primary contributions of this work are as follows:

1. Design of a large, combined CBC dataset for binary anaemia classification.

2. Comparative evaluation of baseline and ensemble ML models using stratified cross-validation.

3. Clinical interpretation of precision-recall trade-offs for anaemia screening and diagnosis.

Related work

Traditional ML methods used for anaemia classification include the Random Forest (RF) method, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Logistic Regression (LR), and Decision Trees. In particular, the RF technique, which is the result of the combination of many decision trees, performed exceptionally well. As Gómez et al. indicated, the accuracy of the classification using the RF method was 100% [6]; however, Aditya et al. managed to reach 99.2% of accuracy when using the RF technique, significantly outperforming SVM (82.1%) and LR (75.1%). The high efficiency of the RF is provided by its ability to account for non-linear dependencies and provide feature importance ranking [7].

KNN is another simple ML algorithm, which has shown good performance. Faradila et al. indicated that KNN achieved an accuracy rate of 94.72% with high precision and recall rates for anaemia detection. The performance of KNN heavily relies on choosing proper values for k and using the right distance measurement technique with Euclidean distance being the most popular [8].

Variable performance has been obtained from the application of SVMs. The results presented by Abdul-Jabbar et al. suggest that SVM can achieve an accuracy rate of 99.71% depending on the dataset and hyperparameters. Artificial neural network techniques have also shown their capability in detecting anaemia [9].

Ensemble learning algorithms that use predictions generated by multiple base learners have proven to be very effective for classification of anaemia. The ensemble learning algorithms in the gradient boosting family such as XGBoost, LightGBM, CatBoost, and Gradient Boosting yield the best results in the literature. In their comparison of six different ML algorithms, Noviandy et al. found out that Gradient Boosting + SMOTE yields the best performance in terms of 97.98% accuracy, precision, sensitivity, and F1-score as well as 98.48% specificity and 0.987 ROC-AUC [10]. Of all the ensemble algorithms, XGBoost has been found to be one of the most effective algorithms. Similarly, Çakmak reported that LightGBM was the top-performing algorithm (98.38%), followed by CatBoost (98.37%) and XGBoost (97.72%) for hematological disorder classification [11].

Furthermore, some other recent studies have widened the scope of research in the field of anaemia by considering various parameters like its prevalence, risk factors, and modeling. In this context, Aravindalochanan et al. examined the prevalence and risk factors of anaemia, stressing the need for appropriate screening measures for the disease [12]. On the other hand, Sankar et al. have reviewed anaemia in various perspectives such as aplastic anaemia [13]. Likewise, Hirok et al. proposed the prediction of multiclass anaemia with the use of ML algorithms based on CBC information [14]. Similarly, Amjad et al. suggested efficient ML techniques for screening anaemia along with leukemia using CBC reports [15].

In dealing with issues related to imbalanced datasets, Desuky et al. came up with Hybrid Nature Inspired Imbalance Handling Algorithm (HNIHA), which greatly increased the accuracy in detecting anaemia [16]. Furthermore, Diaz-Del-Pino et al. also showed the use of artificial intelligence in helping clinicians in detecting haematological diseases [17]. Hybrid

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

learning models and ensemble learning models have also shown great potential, with Airlangga noting increased accuracy in classifying anaemia using ensemble ML models in CBC [18]. Extending these predictive techniques to routine hematological testing allows earlier detection of systemic chronic conditions. These publicly available data sets, such as the Kaggle anaemia prediction data set [19] and Mendeley CBC data sets, have been crucial in providing reproducible research and benchmarking of ML algorithms used for anaemia detection [20,21].

Hybrid and stacking methods are considered the state-of-the-art approach to anaemia classification. Chandralekha et al. introduced a hybrid algorithm that combines the RF, ExtraTrees, Gradient Boosting, and XGBoost methods, and attained 90.83% accuracy, 95.93% precision, 88.74% recall, and 92.17% F1-Score [22].

Earlier work investigated optical character recognition (OCR)-assisted extraction and multiclass anaemia classification from heterogeneous CBC reports [23]. Similarly, a recent review of AI-based anaemia detection and classification studies highlighted persistent challenges related to dataset diversity, external validation, model interpretability, and the limited investigation of OCR-based extraction from CBC reports [24]. These limitations motivate the present study, which focuses on harmonised CBC datasets, comparative evaluation of multiple ML classifiers, and model-level explainability using Shapley Additive Explanations (SHAP).

Predictive ML applications have shown high efficacy across diverse clinical diagnostic domains. Research such as "Artificial Intelligence for Predicting Progression and Personalizing Healthcare to Alzheimer's Disease Patients" demonstrates AI's strength in personalising complex patient management [25].

These findings demonstrate that combining complementary ML models can significantly enhance predictive performance compared with individual classifiers.

Despite substantial progress, existing ML-based approaches for anaemia classification predominantly emphasise accuracy while overlooking standardised and clinically meaningful evaluation strategies. Comparative analyses of Logistic Regression, RF, and XGBoost using ROC-AUC and stratified cross-validation under realistic class imbalance scenarios remain limited. This gap hinders reproducibility, clinical interpretability, and the broader adoption of ML-based diagnostic systems in healthcare settings.

Materials And Methods

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

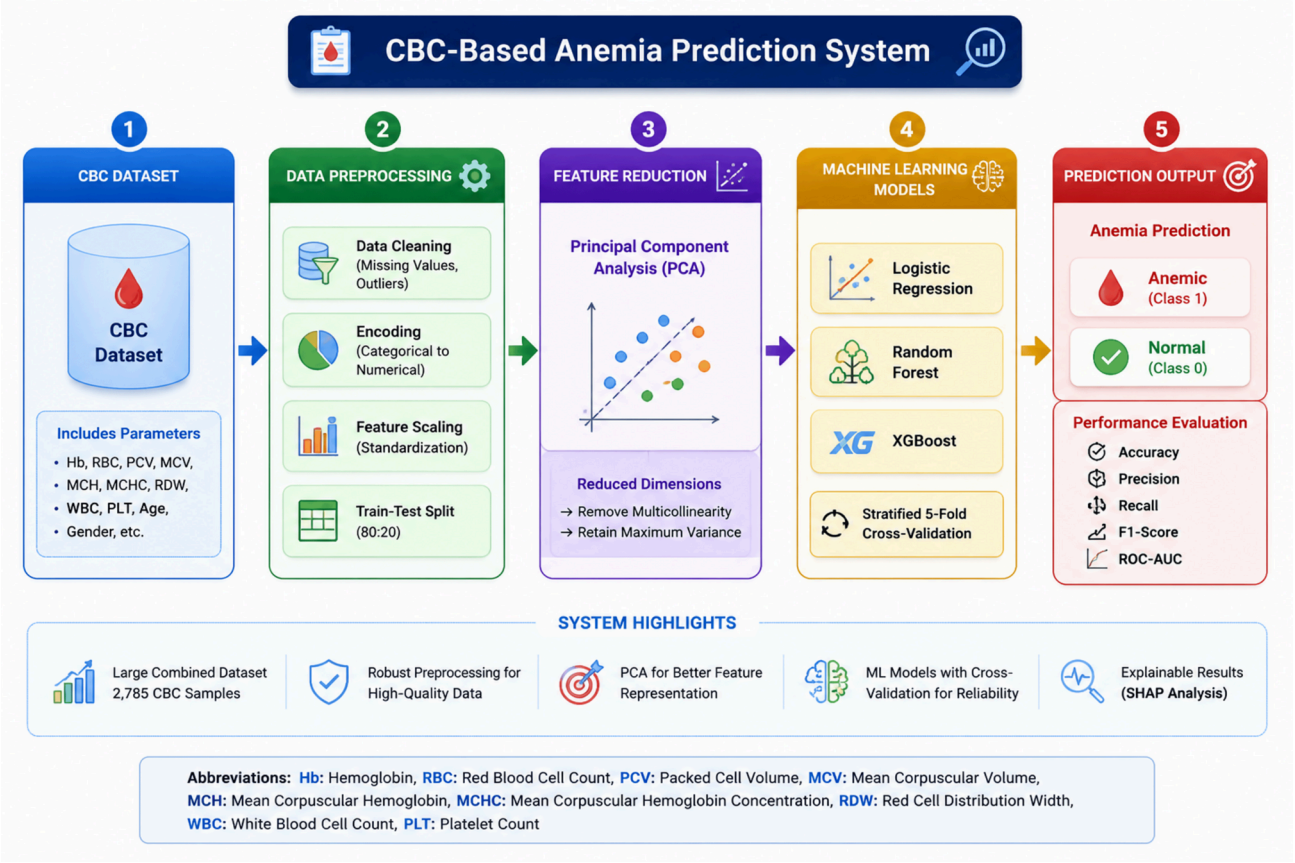


FIGURE 1: Anaemia Prediction System Showing Data Preprocessing, Feature Reduction, Machine Learning Model Development, SHAP-Based Explainability, and Performance Evaluation.

SHAP = Shapley Additive Explanations.

Figure *1* illustrates the overall workflow of the proposed CBC-based anaemia prediction system, including data preprocessing, feature reduction, ML model development, explainability, and performance evaluation.

Dataset description

The dataset used in this study was constructed by integrating three publicly available CBC datasets obtained from Kaggle and Mendeley Data repositories [19-21]. The Kaggle dataset comprised 1,421 CBC records, while the two Mendeley datasets contained 364 and 1,000 CBC records, respectively. These datasets were selected because they provide haematological parameters commonly used for anaemia diagnosis and collectively represent a diverse set of patient records.

Before integration, all three datasets were independently examined to identify the CBC parameters common to every source. Only the shared variables were retained to ensure a consistent feature space across datasets. The final feature set included Age, Gender, Haemoglobin (Hb), Red Blood Cell Count (RBC), Packed Cell Volume (PCV/HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Haemoglobin (MCH), Mean Corpuscular Haemoglobin Concentration (MCHC), and the binary outcome class (Anaemic/Non-anaemic). Dataset-specific attributes that were unavailable in one or more repositories were excluded from the analysis to maintain uniformity.

A harmonisation procedure was subsequently performed to standardise the integrated dataset. Feature names and measurement formats were unified across the three data sources (for example, PCV and HCT were treated as equivalent hematocrit measurements), and data types were converted into a consistent numerical format. Categorical variables were

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

encoded uniformly, with Gender represented as Female = 0 and Male = 1, and the outcome variable encoded as Non-anaemic = 0 and Anaemic = 1. This standardisation ensured compatibility among the datasets and eliminated inconsistencies arising from different repository conventions. After integration, the final dataset contained 2,785 CBC samples, which served as the basis for model development and evaluation.

To ensure data quality, the integrated dataset was examined for duplicate records using combinations of demographic information and CBC parameters. No exact duplicate entries were found in the integrated dataset. Missing values and inconsistent records were addressed during preprocessing using appropriate statistical imputation techniques, including mean, median, or mode imputation depending on the nature and distribution of each feature. Numerical variables were validated for clinically acceptable ranges, and categorical values were verified for consistency. The resulting harmonised dataset was subsequently used for feature analysis and dimensionality reduction. PCA was initially considered as an exploratory technique but was not included in the final experimental pipeline for the development of ML models.

ID	Sample Datasets Description with Outcome Class								
	Age	Gender	Hb	RBC	PCV	MCV	MCH	MCHC	Prediction
2509	28	0	9.6	5.66	34	60.1	17	28.2	1
2510	41	0	13.8	4.78	44.5	93.1	28.9	31	0
2511	40	1	13.4	4.65	41.6	89.5	28.8	32.2	0
2512	76	0	11.3	4.24	36.7	86.6	26.7	30.8	1
2513	20	1	11.5	4.14	36.9	89.1	27.8	31.2	1
2514	24	0	12.7	4.29	40.1	93.5	29.6	31.7	1
2515	28	1	12.4	4.98	42.3	84.9	24.9	29.3	0
2516	14	0	13.9	4.97	43.8	88.1	28	31.7	0

TABLE 1: Datasets Parameters for Training Different ML Models

ML = Machine Learning; RBC = Red Blood Cell; PCV = Packed Cell Volume; MCV = Mean Corpuscular Volume; MCH = Mean Corpuscular Haemoglobin; MCHC = Mean Corpuscular Haemoglobin Concentration

Table *1* presents representative records from the integrated dataset. For binary anaemia classification, eight primary predictors Hb, RBC, PCV, MCV, MCH, MCHC, Age, and Gender were used as input variables. Correlation analysis confirmed that these features exhibited the strongest association with the target variable and were therefore selected for subsequent model training and evaluation. The final integrated dataset exhibited class imbalance between anaemic and non-anaemic samples, which was addressed during model training through class-weighted learning within the stratified five-fold cross-validation framework.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

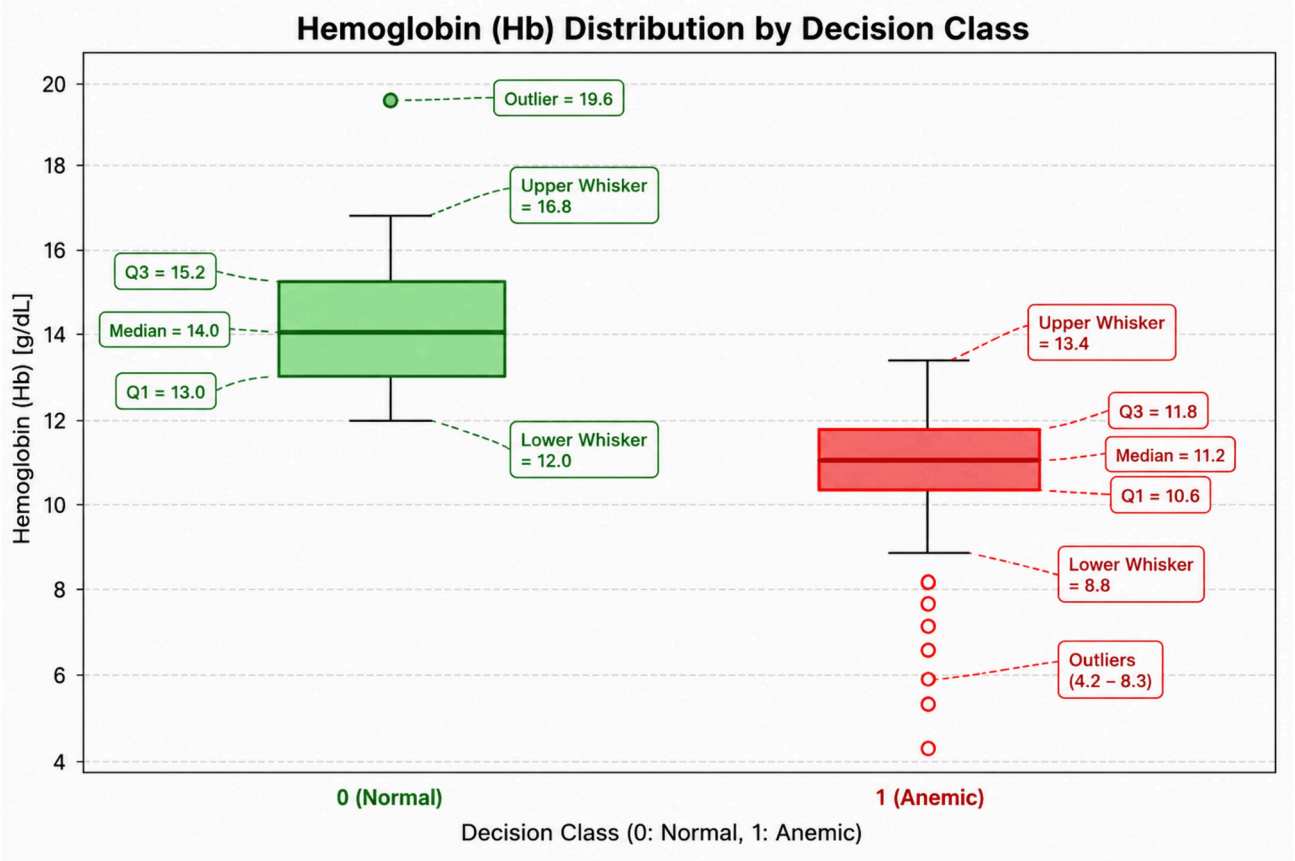


FIGURE 2: Box Plot Showing the Distribution of Hemoglobin (Hb) Levels in Normal (Class 0) and Anaemic (Class 1) Subjects, Highlighting Median Values, Interquartile Ranges, Whiskers, and Outliers.

Figure *2* illustrates the distribution of Hb values between the decision classes. Normal individuals (Class 0) have higher median Hb values with smaller dispersion, whereas anaemic individuals (Class 1) have significantly lower Hb values with larger dispersion, thus emphasising the distinctiveness of the two classes.

Preprocessing

The dataset was examined for missing values, inconsistent entries, and outliers. Missing numerical values were handled using appropriate statistical imputation methods based on feature distributions. Categorical variables were standardised and encoded numerically, with Female = 0, Male = 1, Non-anaemic = 0, and Anaemic = 1.

The numerical features were inspected for extreme values using descriptive statistical analysis and box plots. Correlation analysis was performed to identify relationships among CBC parameters and to assess multicollinearity. Since several hematological variables exhibited strong correlations, PCA was employed as a dimensionality reduction technique to transform correlated variables into a smaller set of orthogonal principal components while preserving most of the variance.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

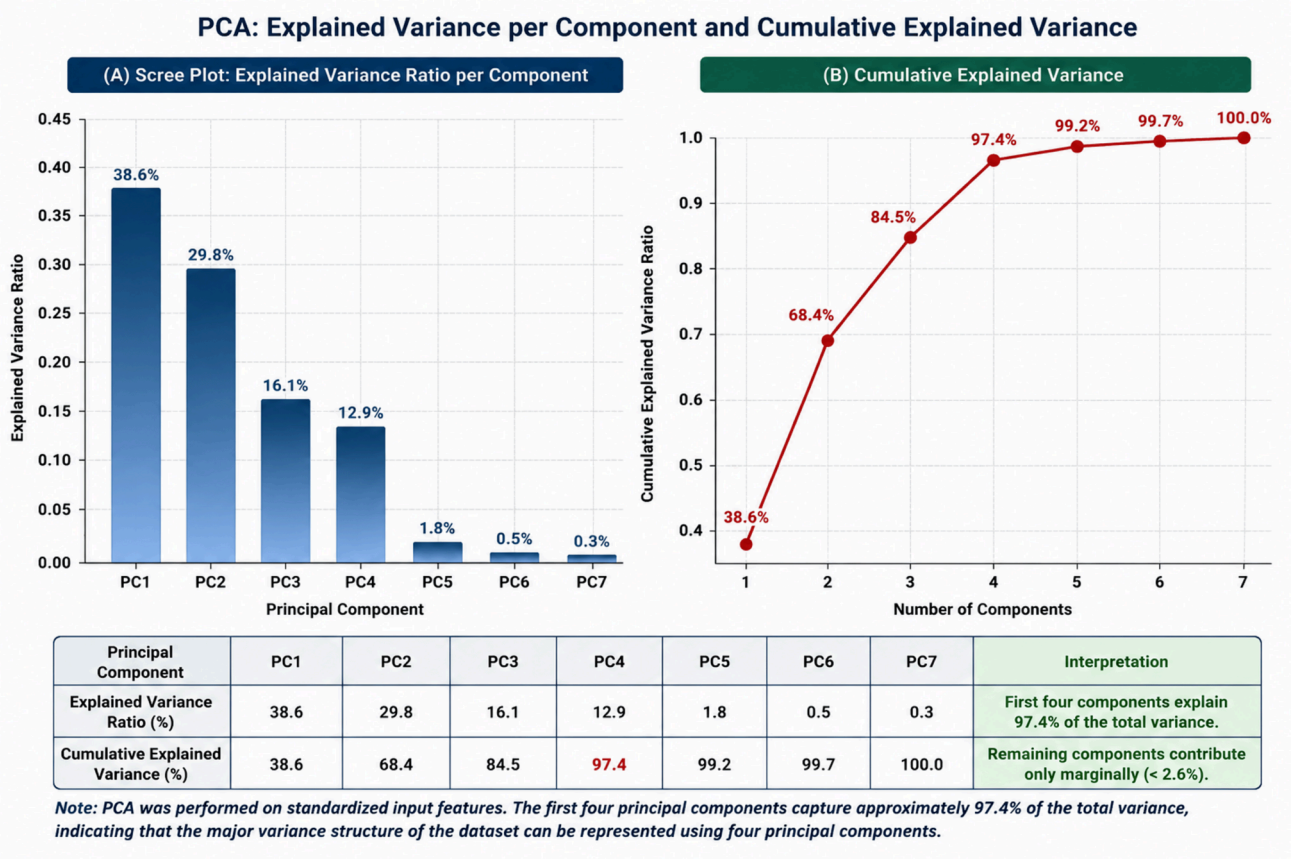


FIGURE 3: Scree Plot and Cumulative Explained Variance of the Principal Components. The First Four Principal Components Explain Approximately 97.4% of the Total Variance, Indicating That the Major Variance Structure of the Dataset Can Be Represented Using Four Principal Components.

PCA was performed to examine the variance structure of the input features shown in Figure *3*. PC1, PC2, PC3, and PC4 explained 38.6%, 29.8%, 16.1%, and 12.9% of the variance, respectively. The cumulative explained variance reached 97.4% after the fourth component, while the remaining components contributed only marginally. Thus, the first four principal components captured the majority of the variance present in the PCA input feature space.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

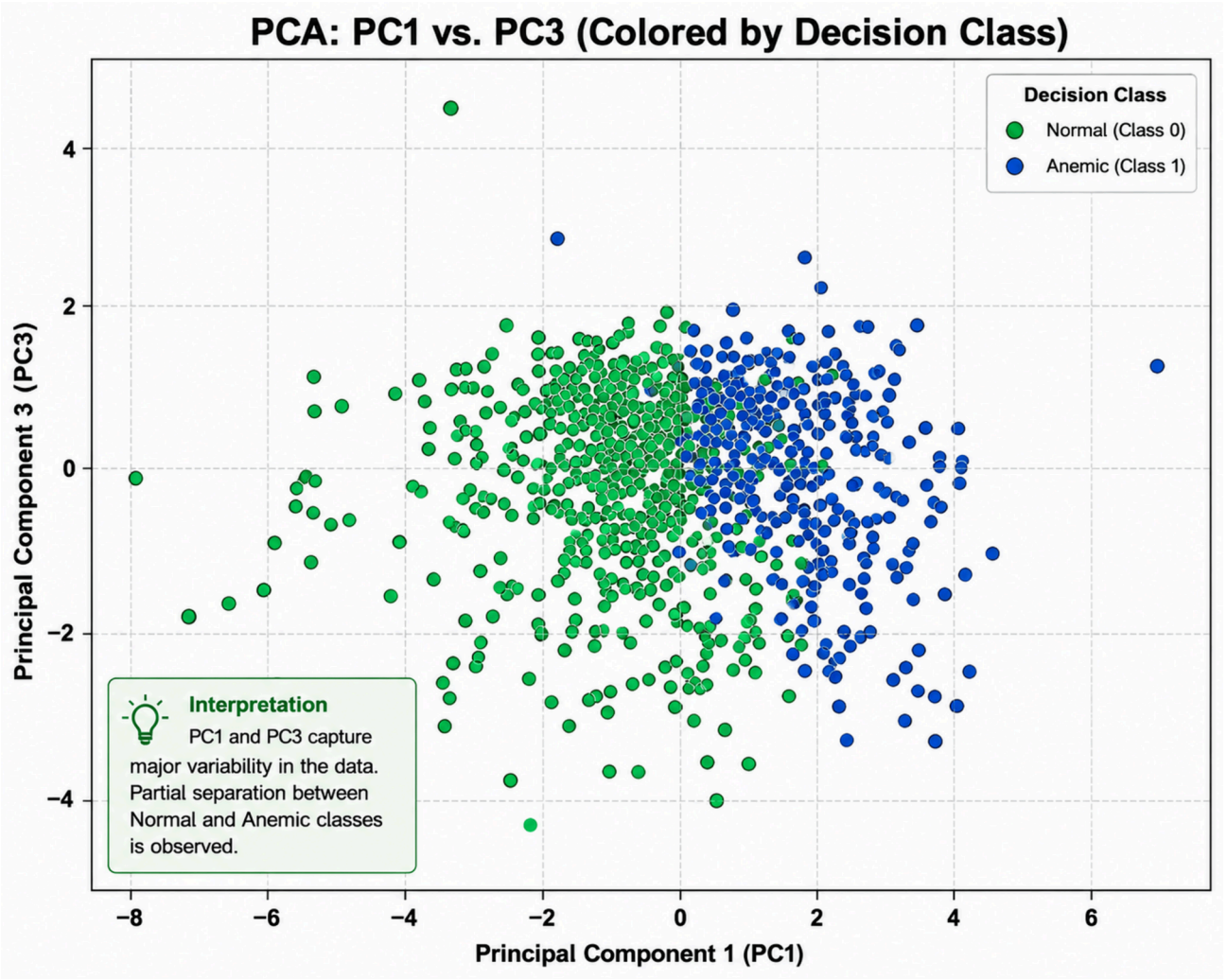


FIGURE 4: Principal Component Analysis (PCA) Scatter Plot Showing the Distribution of Normal (Class 0) and Anaemic (Class 1) Samples Along Principal Components PC1 and PC3, Demonstrating Partial Class Separation in the Reduced Feature Space.

Figure *4* shows the separation between anaemic and normal classes in reduced PCA space, which helps in evaluating the discriminability of features, checking for possible overlaps between classes.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

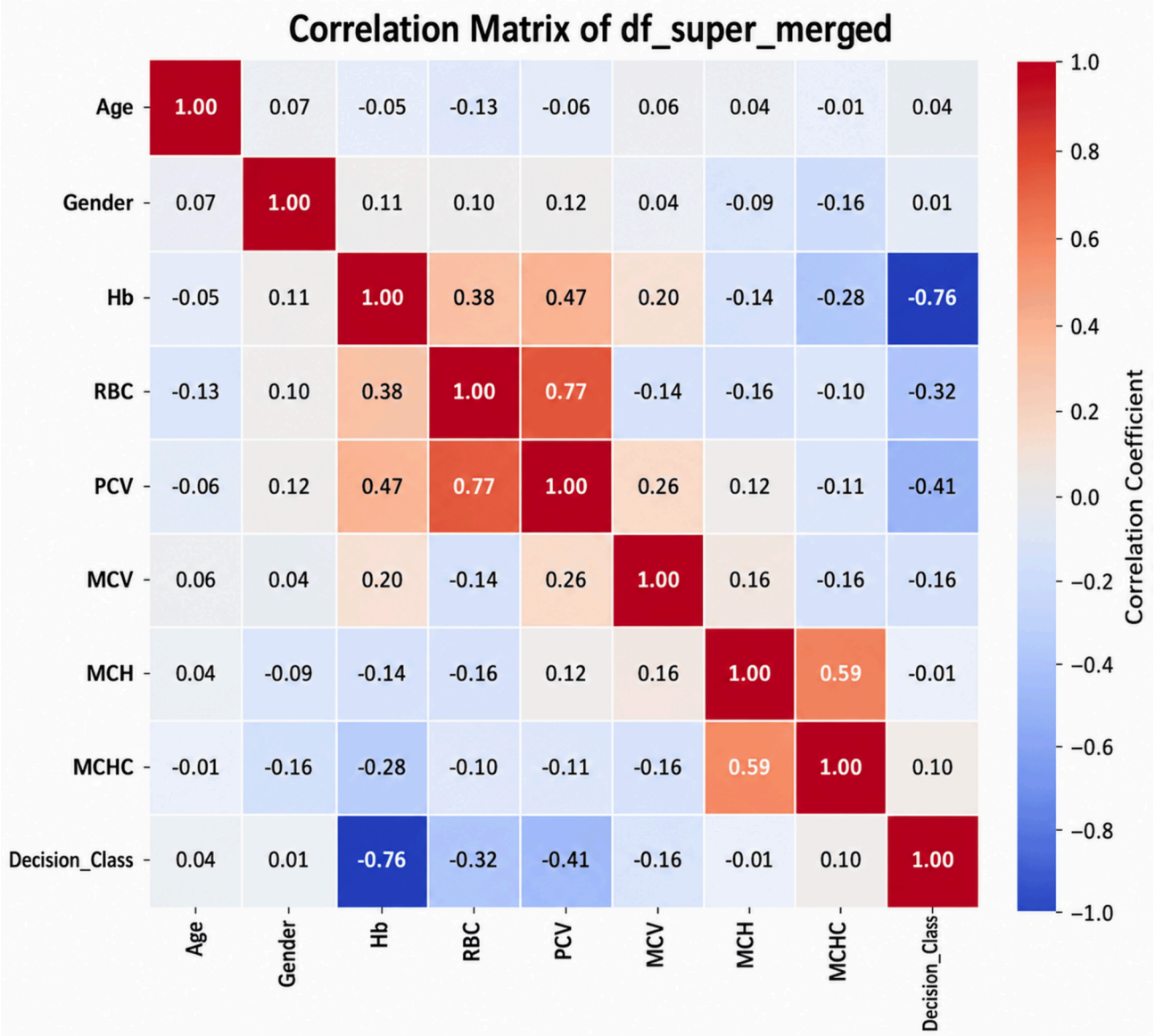


FIGURE 5: Pearson Correlation Matrix of Complete Blood Count (CBC) Parameters and Demographic Variables Showing Their Relationships with the Binary Anaemia Decision Class. Hemoglobin (Hb) Exhibits the Strongest Negative Correlation with Anaemia Status, Highlighting Its Importance in Disease Classification.

Figure *5* shows the heatmap describing the relationships between the demographic variables, the CBC variables, and whether an individual belongs to the anaemia group, highlighting the straight-line (Pearson) correlations in the data. Hb is particularly prominent with a strong negative correlation to the anaemia class, emphasising its importance in the diagnosis process. RBC and PCV are related to Hb and the class labels in a moderate way, indicating they complement Hb well. MCV, MCH, and MCHC are related in a more tenuous manner, indicating smaller, secondary roles. Age and sex are unrelated to the anaemia class. Overall, the results support a multi-feature approach to classification based on CBC data with low multicollinearity between features.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

Experimental setup

The experiments were implemented using Python 3 on the Google Colab cloud computing platform. Three supervised ML algorithms LR, RF, and XGBoost were selected based on their widespread use and demonstrated effectiveness in CBC-based anaemia classification.

The integrated dataset was randomly divided into 80% training data and 20% testing data while preserving the class distribution through stratified sampling. Model robustness and generalisation capability were evaluated using stratified five-fold cross-validation (k = 5).

Model performance was assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These complementary metrics enabled a comprehensive comparison of the predictive performance and potential clinical applicability of the three ML models. Model interpretability was further evaluated using SHAP to identify the contribution of individual features to the model predictions.

Models and algorithm

The following three supervised ML algorithms were employed for binary anaemia classification.

LR served as the baseline linear classifier and estimated the probability of anaemia using the logistic (sigmoid) function with L2 regularisation to reduce overfitting.

RF is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection. Final predictions are obtained through majority voting, improving robustness and reducing variance.

XGBoost is a gradient-boosting ensemble algorithm that sequentially builds decision trees to minimise the classification loss while incorporating regularisation to control model complexity. Its ability to capture non-linear relationships and interactions among CBC parameters makes it well suited for binary anaemia classification.

All models were implemented using the Scikit-learn and XGBoost Python libraries and trained using the same training and validation protocol to ensure a fair comparative evaluation.

Model interpretability with SHAP

To improve the interpretability of the proposed ML framework, SHAP was applied to the best-performing XGBoost classifier after model training. SHAP provides both global and local explanations by quantifying the contribution of each input feature to the model predictions based on cooperative game theory.

The SHAP analysis was implemented using the SHAP Python library (version 0.46.0) together with XGBoost in the Google Colab environment. Since XGBoost is a tree-based ensemble model, the TreeExplainer algorithm was employed to efficiently compute exact SHAP values. SHAP values were calculated on the independent test dataset (20% of the integrated dataset) after completion of model training using stratified five-fold cross-validation. Feature importance was determined using the mean absolute SHAP value (mean SHAP) across all test samples, where larger values indicate a greater overall contribution of a feature to the prediction of anaemia. The resulting feature rankings were visualised using the SHAP summary plot shown in Figure *8*.

AI/LLM Use Disclosure: OpenAI's ChatGPT was used to assist the authors with figure formatting, layout, labelling, and presentation. The underlying data, analyses, scientific content, and final figures were developed and verified by the authors. The authors reviewed all AI-assisted outputs and retain full responsibility for the accuracy, integrity, and scientific interpretation of the figures.

Results And Discussion

The comparative performance of LR, RF, and XGBoost is summarised in Table *2*. All three algorithms were assessed in terms of accuracy and ROC-AUC metrics using stratified five-fold cross-validation (k = 5). LR also reports a mean accuracy of 93.54% ± 0.0136, whereas RF and XGBoost report 98.74% ± 0.0032 and 98.92% ± 0.0043, reflecting accuracy

improvements of 5.20% and 5.38%, respectively. The small standard deviation values suggest that all models are quite stable. XGBoost also reports the largest mean anaemic F1-score of 0.9906 ± 0.0043, indicating its best robustness for anaemia detection with CBC data.

Model	Accuracy (%)	Precision (Normal) (%)	Recall (Normal) (%)	F1-score (Normal) (%)	Precision (Anaemic) (%)	Recall (Anaemic) (%)	F1-score (Anaemic) (%)	ROC-AUC
Logistic Regression	93.54	92.05	92.83	92.44	94.65	94.06	94.36	0.99 ± 0.0136
Random Forest	98.74	97.52	99.58	98.54	99.68	98.13	98.9	1.00 ± 0.0032
XGBoost	98.92	98.33	99.16	98.74	99.37	98.75	99.06	1.00 ± 0.0043

TABLE 2: Comparative Performance Evaluation of Different ML Models.

ML = Machine Learning; ROC-AUC = Receiver Operating Characteristic-Area Under the Curve

As illustrated in Table *2*, LR reports a mean AUC of 0.99 ± 0.0136, whereas RF and XGBoost report perfect discrimination with a mean AUC of 1.00 ± 0.0032 and 1.00 ± 0.0043, respectively (ΔAUC = +0.01).

Figure *6* shows the ROC-AUC analysis further confirmed the discriminative capability of the evaluated models. LR achieved a mean ROC-AUC of 0.99 ± 0.0136, whereas both RF(1.00 ± 0.0032) and XGBoost (1.00 ± 0.0043) demonstrated near-perfect discrimination between anaemic and non-anaemic samples. The small standard deviations across the five validation folds indicate that the models produced stable and consistent performance throughout the cross-validation process.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

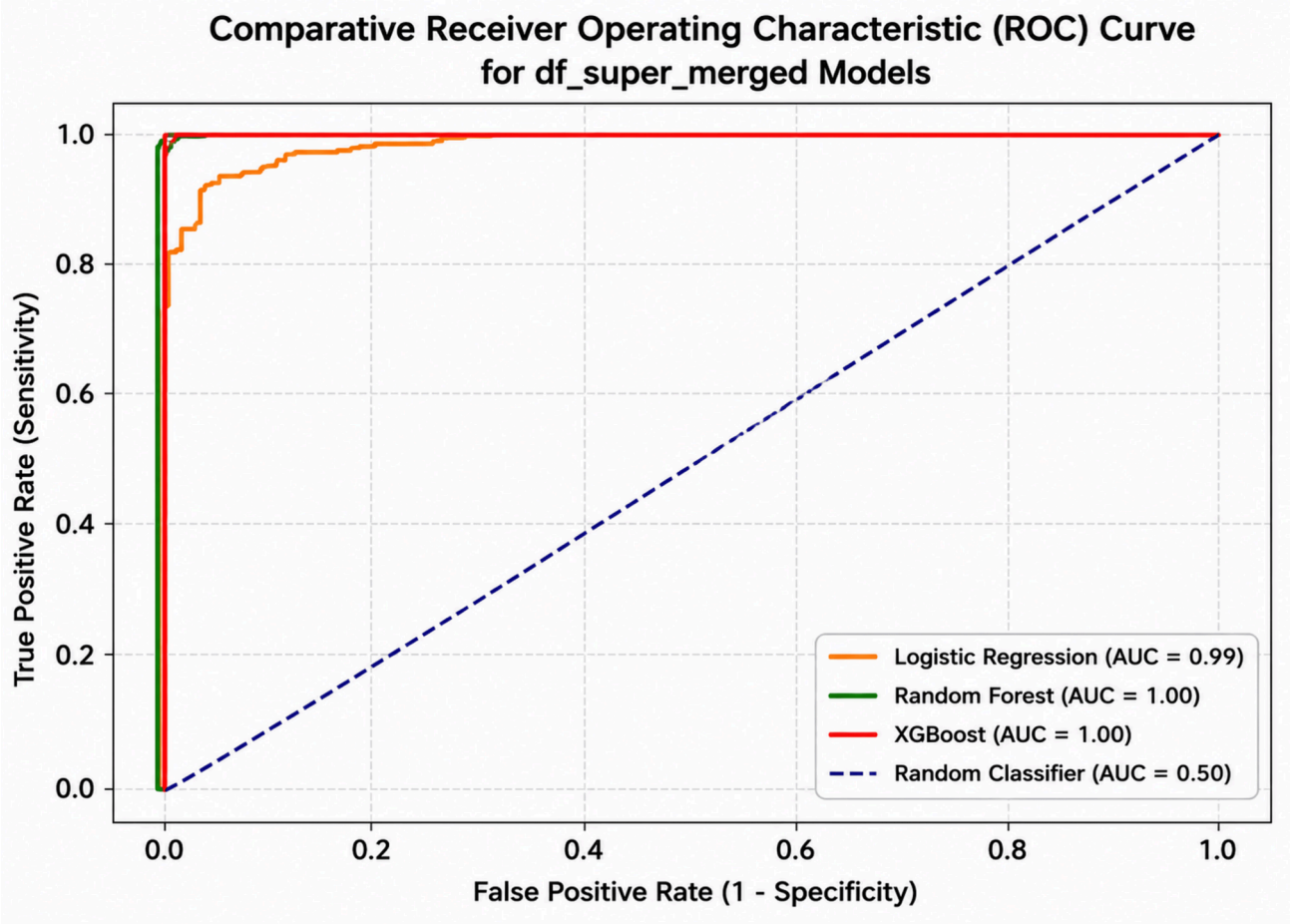


FIGURE 6: ROC Curve Analysis of Machine Learning Models for Binary Anaemia Classification Showing Near-Perfect Performance of Random Forest and XGBoost on the Integrated CBC Dataset.

ROC = Receiver Operating Characteristic; CBC = Complete Blood Count

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

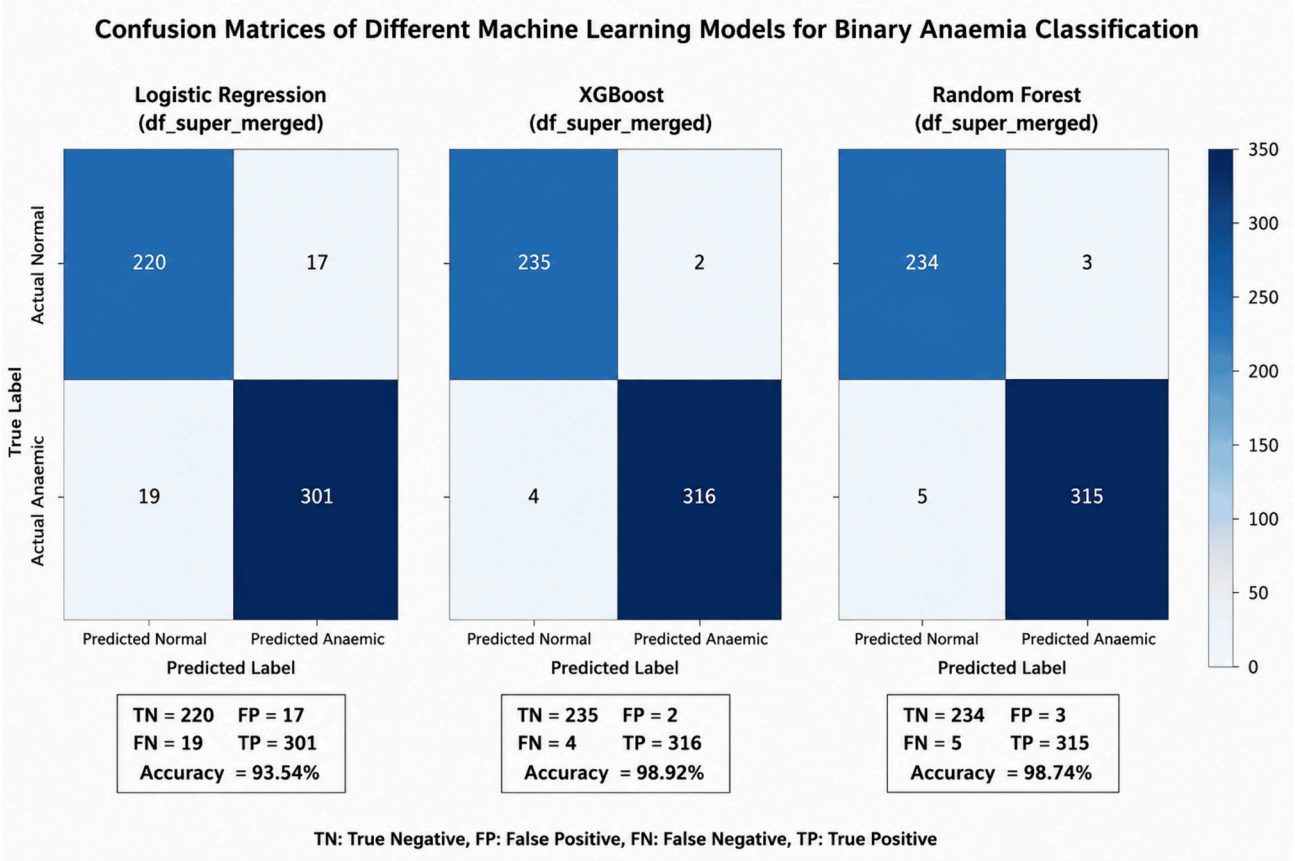


FIGURE 7: Confusion Matrices for Logistic Regression, XGBoost, and Random Forest Models Showing the Distribution of True Positives, True Negatives, False Positives, and False Negatives in Binary Anaemia Classification. XGBoost and Random Forest Demonstrated Superior Classification Performance with Minimal Misclassifications Compared to Logistic Regression.

Figure *7* shows the confusion matrices of the models that were evaluated. LR has more instances of both false positives and false negatives than the other models. RF and XGBoost significantly reduce the instances of misclassification, and they have excellent precision and recall. The high recall of the models is particularly important for the detection of anaemia because it will help avoid missing cases, and they also have high precision that will offer reliable support for decision-making.

Feature importance and explainability

Hb exhibited the highest influence on model predictions, followed by Gender, MCHC, and PCV, highlighting their relative importance in ML-based anaemia classification.

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

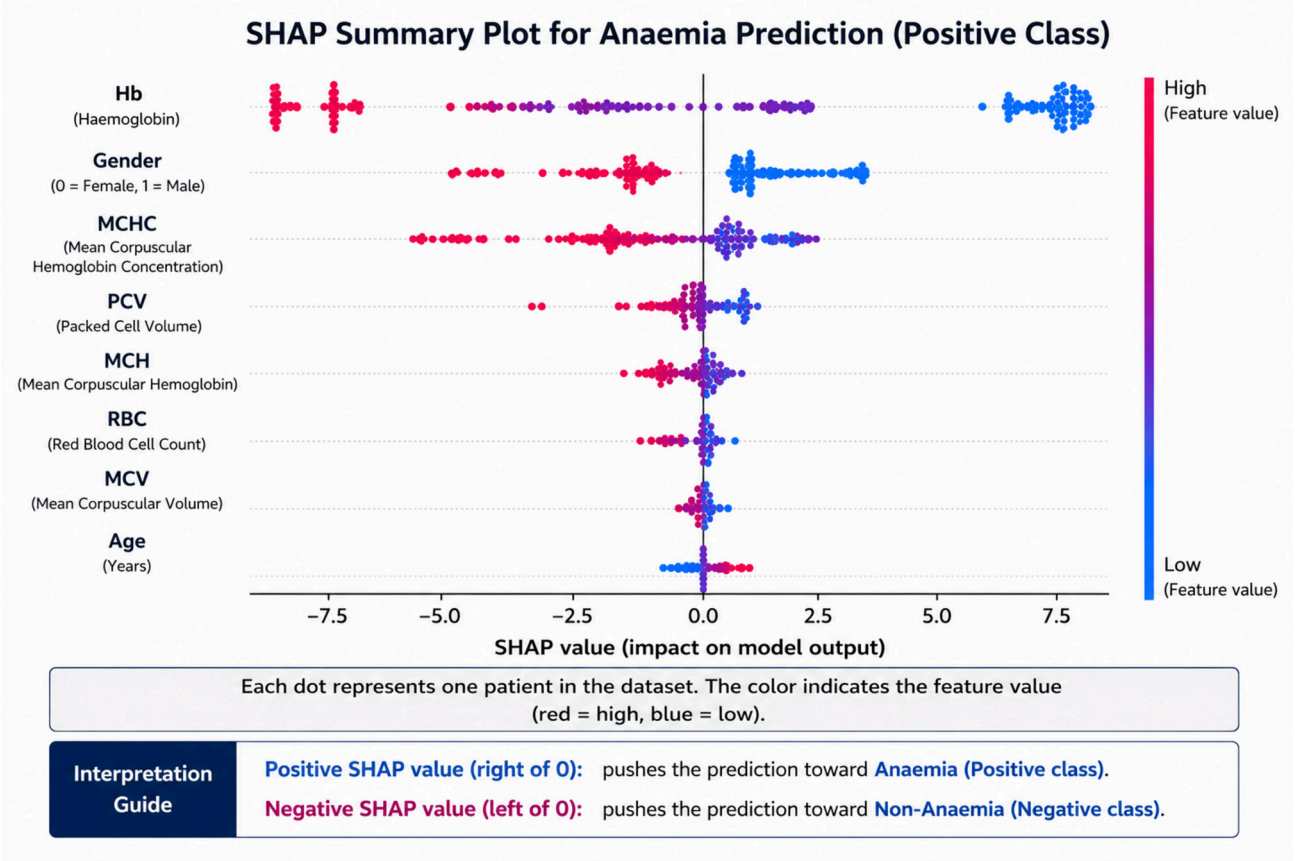


FIGURE 8: SHAP Summary Plot for Binary Anaemia Prediction Showing the Contribution of Complete Blood Count (CBC) Features to the Positive (Anaemic) Class.

To enhance the interpretability of the proposed model, SHAP analysis was performed on the best-performing XGBoost model to quantify the contribution of individual CBC parameters to the prediction of anaemia. As shown in Figure *8*, Hb was identified as the most influential feature, exhibiting the greatest overall impact on the model's predictions. This was followed by Gender, MCHC, and PCV, which also demonstrated substantial contributions to distinguishing anaemic from non-anaemic individuals. Other parameters, including RBC, MCV, MCH, and Age, showed comparatively lower contributions.

The feature-importance pattern presented in Figure *8* is consistent with established clinical knowledge, as haemoglobin is a primary indicator used in the diagnosis of anaemia, while other haematological indices provide complementary information regarding RBC characteristics. The SHAP analysis therefore provides an interpretable explanation of the model's decision-making process by illustrating the relative contribution of individual features to the classification outcome. This improved transparency can facilitate clinical interpretation and increase confidence in the potential use of the proposed XGBoost model as a clinical decision-support tool.

Conclusions

This study demonstrates that ML algorithms can accurately classify binary anaemia using routinely available CBC parameters, with XGBoost achieving the strongest overall predictive performance among the evaluated models. The model also demonstrated robust performance under stratified cross-validation. SHAP-based analysis provided model-level explanations of feature contributions, with haemoglobin and other erythrocyte-related parameters contributing

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

substantially to the predictions, consistent with their established relevance to anaemia assessment. The findings suggest that CBC-based machine learning models may have potential as cost-effective decision-support tools for anaemia screening, particularly in resource-constrained healthcare settings. However, the present study is limited by its reliance on secondary datasets and the absence of prospective and external clinical validation. Further evaluation using independent, diverse, and prospectively collected clinical data is therefore required before deployment in routine clinical practice.

Future research will focus on validating the proposed models using independent, prospective, and real-world clinical datasets to assess their generalisability across diverse patient populations and healthcare settings. The proposed framework can be extended from binary classification to multi-class classification of anaemia severity and can incorporate additional demographic, clinical, and laboratory parameters. Further development of advanced explainable AI techniques may provide deeper insights into model predictions and improve model transparency. These enhancements may support the future development of reliable ML-based decision-support tools for anaemia screening, particularly in resource-constrained healthcare settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Pradip R. Bhakare, Chandrakant M. Jadhao, Jayesh K. Kokate

Acquisition, analysis, or interpretation of data: Pradip R. Bhakare, Chandrakant M. Jadhao, Jayesh K. Kokate

Drafting of the manuscript: Pradip R. Bhakare, Jayesh K. Kokate

Critical review of the manuscript for important intellectual content: Pradip R. Bhakare, Chandrakant M. Jadhao, Jayesh K. Kokate

Supervision: Chandrakant M. Jadhao, Jayesh K. Kokate

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

Publicly available Kaggle and Mendeley CBC datasets [References 19–21] were integrated through preprocessing and feature standardisation. Dataset: https://www.kaggle.com/datasets/pradipbhakare/cbc-dataset-for-binary-anaemia-classification/settings; The Python code for preprocessing, feature engineering, model training, stratified 5-fold cross-validation, evaluation, and SHAP interpretation is available at: https://colab.research.google.com/drive/17C7V77-_DXYtQmNv0tJQ3q1-Yxp-ublr?usp=sharing

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

Acknowledgements

The authors gratefully acknowledge Dr. C. M. Jadhao, Principal and Professor, Department of Electronics and Telecommunication Engineering, Mauli Group of Institutions, College of Engineering and Technology, Shegaon, for his guidance, valuable insights, and continuous support throughout this research. The authors also thank Dr. J. K. Kokate for his constructive suggestions, technical expertise, and encouragement, which significantly contributed to the completion of this study. The authors further acknowledge the support provided by the Department of Electronics and Telecommunication Engineering, Mauli Group of Institutions, College of Engineering and Technology, Shegaon, in facilitating this research.

References

1. WHO global anaemia estimates: key findings, 2025. (2025). Accessed: August 15, 2025: https://www.who.int/publications/i/item/9789240113930.
2. International Institute for Population Sciences (IIPS) and ICF: National Family Health Survey (NFHS-5), 2019-21. India: Volume I. IIPS, Mumbai; 2021.
3. Pullakhandam S, McRoy S: Classification and explanation of iron deficiency anaemia from complete blood count data using machine learning. BioMedInformatics. 2024, 4:661-672. 10.3390/biomedinformatics4010036
4. Dhruva BSD, Sampathila N, Bairy GM, Prabhu S, Belurkar S, Chadaga K, Nandish S: Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable artificial Intelligence utilizing blood attributes. Scientific Reports. 2025, 15:505. 10.1038/s41598-024-84120-w
5. Shehab S, Shehab EA, Khawaga A: Anemia diagnosis and prediction based on machine learning. Kafr El-Sheikh Journal of Information Sciences. 2023, 4:1-9.
6. Gómez JG, Parra C, Álvarez DS, Hernández Riaño V, Ramírez-González G: Anemia classification system using machine learning. Informatics. 2025, 12:19. 10.3390/informatics12010019
7. Aditya MR, Sutanto T, Budiman H, Ridha MRN, Syapotro U, Azijah N: Machine learning models for classification of anemia from CBC results: random forest, SVM and logistic regression. INTI Journal. 2024, 2022:49.
8. Faradila Z, Homaidi A, Prasetyo JD: Classification of anaemia status using the k-nearest neighbor algorithm. G-Tech: Jurnal Teknologi Terapan. 2025, 9:436-444. 10.70609/gtech.v9i1.6377
9. Abdul-Jabbar SS, Farhan AK, Luchinin AS: Comparative study of anemia classification algorithms for international and newly CBC datasets. International Journal of Online and Biomedical Engineering (iJOE). 2023, 19:141-157. 10.3991/ijoe.v19i06.38157
10. Noviandy TR, Idroes GM, Suhendra R, Bakri TK, Idroes R: Advanced anemia classification using comprehensive hematological profiles and explainable machine learning approaches. Infolitika Journal of Data Science. 2024, 2:72-81. 10.60084/ijds.v2i2.237
11. Çakmak Y: Machine learning approaches for enhanced diagnosis of hematological disorders. Computational Systems and Artificial Intelligence. 2025, 1:8-14. 10.69882/adba.csai.2025072
12. Aravindalochanan V, Sinharoy S, Thangavel G, et al.: Prevalence and determinants of anaemia among pregnant women using biomass fuel in rural Tamil Nadu: a cross-sectional study. BMC Pregnancy and Childbirth. 2026, 26:555. 10.1186/s12884-026-08999-1
13. Sankar D, Oviya IR: Multidisciplinary approaches to study anaemia with special mention on aplastic anaemia (Review). International Journal of Molecular Medicine. 2024, 54:95. 10.3892/ijmm.2024.5419
14. Hirok MK, Rahman S, Parvin M: Anemia prediction and classification of all classes with and without anemia patients using a machine learning model. IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT). 2024, 1-6. 10.1109/compas60761.2024.10796730
15. Amjad H, Hussain Z, Hasan M, Ul Hassan M: Machine learning-based models for screening of anemia and leukemia using features of complete blood count reports. Scientific Reports. 2025, 15:33333. 10.1038/s41598-025-21279-w
16. Desuky AS, Hussain S: An improved hybrid approach for handling class imbalance problem. Arabian Journal for Science and Engineering. 2021, 46:3853-3864. 10.1007/s13369-021-05347-7

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2

17. Diaz-del-Pino S, Trelles-Martinez R, González-Fernández FA, Guil N: Artificial intelligence to assist specialists in the detection of haematological diseases. Heliyon. 2023, 9:e15940. 10.1016/j.heliyon.2023.e15940

18. Airlangga G: Anemia classification using hybrid machine learning models: a comparative study of ensemble techniques on CBC data. Journal of Computer System and Informatics (JoSYC). 2024, 5:1108-1117. 10.47065/josyc.v5i4.5848

19. Anemia Prediction Using 4 ML algorithms. (2026). Accessed: March 24, 2026: https://www.kaggle.com/code/biswaranjanrao/anemia-prediction-using-4-ml-algorithms.

20. CBC Dataset. Mendeley Data, V1. (2022). Accessed: November 22, 2022: https://data.mendeley.com/datasets/28s2bhdjfd/1.

21. Complete Blood Count Anemia Diagnosis. Mendeley Data, V1. (2021). Accessed: April 22, 2021: https://data.mendeley.com/datasets/dy9mfjchm7/1.

22. Chandralekha E, Sathya V, Priyatharsini GS, Kumar AV, Vasudevan I, Umapathy M: Machine learning models for predicting anemia: evaluation and performance insights. International Conference on Inventive Computation Technologies (ICICT). 2024, 1-7. 10.1109/icicec62498.2024.10808776

23. Bhakare PR, Jadhao CM, Kokate JK: Optical character recognition-based anaemia classification and prediction system using artificial intelligence: a comparative performance analysis. Journal of Technical Education. 2026, 49:85.

24. Bhakare PR, Jadhao CM, Kokate JK: Artificial intelligence-based recent anaemia detection and classification systems: a review. 2026 International Conference on Innovations in Communication, Geoscience and Robotics (ICICGR). 2026, 1-5.

25. Ali I, Manda VK: Artificial intelligence for predicting progression and personalizing healthcare to Alzheimer's disease patients. AI-Driven Personalized Healthcare Solutions. Chemingui H, Lamloumi M (ed): IGI Global Scientific Publishing, Hershey, PA; 2025. 155-190. 10.4018/979-8-3693-7858-8.ch006

How to cite this article:
Bhakare P R, Jadhao C M, Kokate J K (October 15, 2026) Experimental Evaluation of Machine Learning Models for Complete Blood Count (CBC)-Based Binary Anaemia Classification. Cureus J Comput Sci 3 : es44389-026-00325-2. DOI https://doi.org/10.7759/s44389-026-00325-2