iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73251

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Anemia Detection Using CBC Data: A Comparative Study

Prerna Samtani[1], Suryansh Saxena[2], Juhi Janjua[3]
[1, 2] *Computer Department, Thadomal Shahani Engineering College, Bandra*
[3]*Thadomal Shahani Engineering College, P.G. Kher Marg, 32nd Road, Marg, Off Linking Rd, TPS III, Bandra West, Mumbai, Maharashtra 400050*

Abstract: ***Over 1.6 billion people worldwide suffer from anemia, a common hematologic disorder that goes undiagnosed in situations with limited resources due to a lack of diagnostic tools. This study presents a machine learning approach to the low-cost, non-invasive diagnostic method of detecting anemia from complete blood count (CBC) data. Using high-quality CBC data, we compare the predictability of three supervised models: Random Forest, Support Vector Machine (SVM), and Logistic Regression. GridSearchCV was used for extensive preprocessing, hyperparameter tuning, and stratified cross-validation techniques. Among the models compared, Random Forest had the highest accuracy of 99.48%, outperforming the SVM model (23.81%) and the Logistic Regression model (57.23%). SHAP (SHapley Additive exPlanations), which has a strong correlation with clinical relevance, was used to select the most contributing features influencing predictions in order to enhance model interpretability. Our results show that interpretability and ensemble learning can work well together as a diagnostic support system for the early identification of anemia in clinical settings.***

Keywords: Anemia Detection, Random Forest, SHapley Additive exPlanations (SHAP), Complete Blood Count (CBC), Medical Diagnosis

I. INTRODUCTION

Anemia is a clinical condition characterized by inadequate hemoglobin or red blood cells in the blood and is also a significant worldwide public health issue which impacts over 1.6 billion people worldwide. According to the World Health Organization, anemia is disproportionately distributed in the at-risk populations such as children and pregnant women due to nutritional deficiency, chronic disease, and hereditary factors especially. Clinical symptomatology of the condition includes weaknesses, dizziness, and cognitive impairment; in resource-poor settings however, anemia typically presents as an asymptomatic disease state owing to limited healthcare systems and diagnostic workflows which lack the necessary laboratory or intrusive testing.

Diagnosis of anemia has focused on hemoglobin testing through complete blood count (CBC) tests - which although reliable, are increasingly resource-intensive and are difficult to obtain in remote or low-resourced populations. Recently, machine learning (ML) has also emerged as a prominent innovation in the healthcare sector to detect minor patterns from clinical findings that facilitate early-stage diagnosis and clinical decision-making. However, there is still a scepticism in defining the best algorithm for clinical purposes, as selecting the best algorithm requires a critical balance of accuracy, interpretability, and generalization.

This research aims to create an accurate and comparative model for anemia prediction from Complete Blood Count (CBC) data. In a semi-controlled setting, we compare and evaluate the performance of three commonly used supervised learning models: Logistic Regression, Support Vector Machine (SVM) and Random Forest; on a highly pre-processed and curated array of CBC values. After an extensive learning, validating, and training opportunity, as well as hyper-parameter tuning using Grid Search methods. We demonstrated that the Random Forest classifier provides the most efficient model and highest accuracy at 99% as contrasted with Logistic Regression (57%) and SVM (23%). Together the results indicate that both model selection and hyper-parameter tuning are significant contributors to striving for greater predictive accuracy towards clinical utility.

The principal contributions of the current research are:

1) An end-to-end preprocessing pipeline for machine learning data sets involving CBCs.
2) A comparison of three basic machine learning algorithms, reported on in terms of accuracy, confusion matrix and classification metrics.
3) Illustration and clear performance improvement due to hyper-parameter optimization using Grid Search and cross-validation methods.

By leveraging standard ML approaches and enhancing them with optimization strategies, our proposed methodology offers a scalable and accurate tool for aiding anemia diagnosis using non-invasive and accessible clinical data.

Section II summarizes the relevant literature on the detection of anemia and machine learning techniques in medicine, Section III outlines the methods for the study, which included data pre-processing, feature selection and model development, Section IV discusses the experimental findings of training and evaluating the three machine learning models, as well as performance comparison between the machine learning models and Section V concludes with a summary, limitations and future work.

II. LITERATURE REVIEW

In recent years, artificial intelligence (AI) and machine learning (ML) are utilized as a growing tool in medical diagnostics, primarily hematological disorders such as anemia. Many studies describe the use of ML algorithms to predict anemia using complete blood count (CBC) parameters, with varying results and limitations.

Many previous studies showed that machine learning models are effective for anemia classification. For example, in [1], the authors applied different supervised learning algorithms to CBC data to predict anemia, including Support Vector Machines (SVM), Logistic Regression (LR) and Random Forest (RF); RF showed better performance when using various classifiers and ensemble learning. Despite the use of RF as a learning algorithm, they only suggest a screening tool and again raise limitations in interpretations and robustness as significant factors in the clinical setting.

RF is particularly good in biomedical diagnostics considering it is an ensemble learning strategy that attempts to moderate the number of predictive variables and appropriate high-dimensional data; as well, finding improved performance to avoid overfitting using bootstrapped aggregation. In [2], the authors used RF as a screening tool for pediatric anemia and demonstrated consistent accuracy above 90%. In a similar study [3], the authors also referred to RF for differentiating and amplifying classifications of the subtypes of anemia and showed better performance over the single-decision-tree models.

However, many of the models did not have proper hyperparameter tuning, and also did not have any methods to interpret the model to provide explanations, such as SHapley Additive exPlanations (SHAP), which presents as an important part for interpretable clinical care.

Support Vector Machines have also been suggested as a method for anemia prediction tasks. Support Vector Machines can provide effective classification boundaries and perform especially well in high dimensional feature space, as mentioned in reference [1].

The generally simple named Logistic Regression, remains a common baseline method applied for binary classification in healthcare. In complex datasets with multiplicative outcomes, like CBCs, this is more likely to happen, giving LR limitations to its flexibility as a model - and despite its interpretability and statistical model mathematical properties - due to this and general overfit, will lead to lower predictive performance limits to magnitude than more explainable models.

Alternative approaches exist too, including hybrid approaches (differences in modelling strategies) and use of deep learning to improve performance in medical diagnosis. In studies such as [4] and [5] deep learning and ensemble of models was beginning to develop the group of models that could potentially model non-linear relationships between CBC features.That said, these models often require annotated data, computational resources, and remain non-linear, with interpretability being a critical requirement in hospitals.

Generally, the literature shows a growing curiosity to use ML in the anemia context, with the common methods being Random Forest and SVM. However, the challenges have been a lack of explainability, lack of sufficient tuning of parameters, and performances lacking consistency across datasets indicating that more work remains to make these second-stage learning models understood and clinically viable.

III.PROPOSED METHODOLOGY

This section describes the systematic method used for the prediction of anemia using machine learning techniques on complete blood count (CBC) data. A step-by-step overview of the workflow is illustrated in Fig. 1.

A. Dataset Description

The dataset used in this study, titled diagnosed_cbc_data.csv [6] includes hematological features obtained from CBC reports and a corresponding diagnosis that identifies either the nature of anemia or the lack of anemia. The features include quantities such as Hemoglobin, RBC, Hematocrit, MCH, MCV, and MCHC, among others. The variable of interest, that is, "Diagnosis," is introduced as a categorical label that was encoded to a number using the label encoding method. Correspondence was maintained between the original diagnosis names and their respective integer labels.

B. Dataset Preprocessing

The following steps were taken as part of the preprocessing stage to standardize and improve consistency, which improves the quality of all observations of the data set:

1) Special characters were removed and column names were made lowercase.
2) The dependent variable (y) and independent variables (X) were split into separate dataframes.
3) Target labels were encoded using LabelEncoder from scikit-learn [7].
4) Data was split into a train and test set with stratification maintaining the same label distribution in train and test using a 70/30% ratio, the random_state=42 parameter was employed to create reproducible data set to allow for complete transparency.

C. Model Implementation

Three of the classification algorithms were compared:

1) Logistic Regression – a custom implementation built from scratch as a linear model suitable for both binary and multiclass classification tasks [8].
2) Support Vector Machine (SVM) – implemented with a linear kernel to be used as a more powerful margin-based classifier [9].
3) Random Forest – a model built from scratch, consisting of several Decision Trees that were trained on arbitrary subsets of features and samples [10].

With a 70/30 train-test split, each model was trained on the training set and then assessed on the test set.

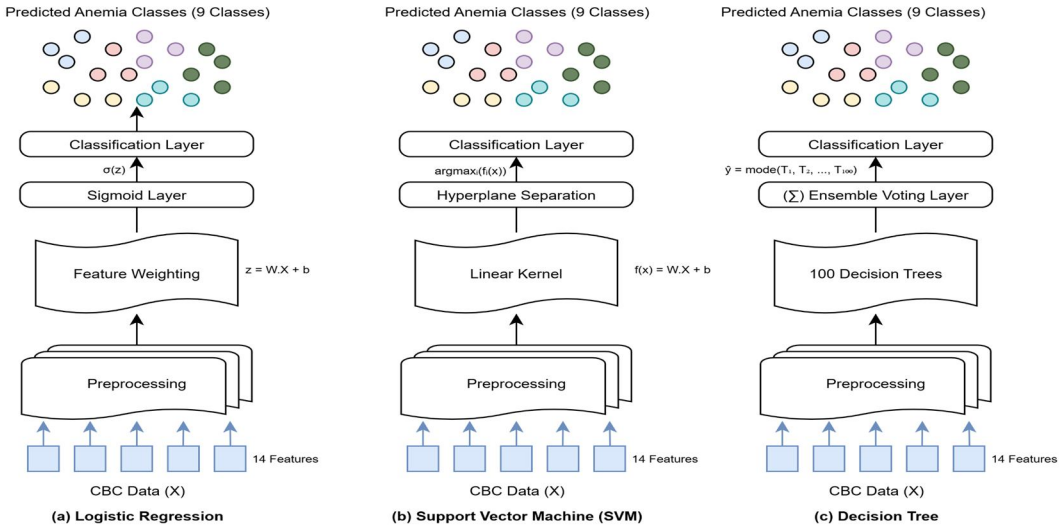


Fig. 1 Comparative model architectures for anemia classification using CBC data

D. Hyperparameter Optimization

Hyperparameter tuning was performed with GridSearchCV for each algorithm. Hyperparameter optimization was executed in order to select the highest-performing model and improve performance. As 3-Fold Stratified Cross-Validation was used to evaluate model performance, stratification ensured the same distribution of class in each fold. Thus, this was a reliable way to estimate model performance.

A grid of hyperparameter combinations were included the following:

For Random Forest models: n_estimators, max_depth, min_samples_split, min_samples_leaf, and n_features_per_tree.

For Logistic Regression and SVM models: C for regularization strength, various penalty types, and kernel choices for SVM.

E. Evaluation Metrics

The models that were best from each algorithm, after tuning; we then evaluated using:

Accuracy Score: The accuracy score is the proportion of correct predictions to all predictions. It is calculated using the formula shown in equation 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: The ratio of correctly predicted positive observations to the total predicted positive observations. It is calculated using the formula shown in equation 2:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: The ratio of correctly predicted positive observations to all actual positive observations. It is calculated using the formula shown in equation 3:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score: The F1 Score is the harmonic mean of precision and recall. An F1 Score is useful when the distribution of the data is uneven. It is calculated using the formula in equation 4:

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Confusion Matrix: A confusion matrix allows you to see all correct and incorrect predictions.
These metrics were calculated on the test dataset, therefore avoiding any biased evaluations.

F. Final Model Selection

After the evaluation Random Forest was rated as the top performer with 99% test accuracy, Logistic Regression was second with 57%, and SVM was third with 23%. The combination of custom implementation and hyperparameter tuning offered the model a significant boost in prediction accuracy; therefore, Random Forest was declared the best algorithm for detecting anemia in this study.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section introduces and evaluates the three machine learning models Random Forest, Support Vector Machine(SVM), and Logistic Regression — that were used to discover anemia. A clean and stratified CBC dataset was used for training and testing each model. To improve predictive performance, hyperparameter tuning was applied to also optimize Random Forest. Evaluation was conducted through metrics like Accuracy, Precision, Recall, and F1-score. The interpretability of the model was estimated through SHAP (SHapley Additive exPlanations) [11], and results were supported through visualization with the confusion matrix and Pearson correlation heatmap.

A. Hyperparameter Configuration

All the models were trained with hyperparameters optimized with GridSearchCV, employing a Stratified K-Fold cross-validation approach with 3 splits and random_state = 42 to ensure reproducibility. The practice was done to preserve the class distribution among non-anemia and anemia cases balanced in training and validation sets to eliminate class imbalance bias in evaluation.
The final parameters used in the best Random Forest model are displayed in Table I that specifies the optimal depth, number of estimators, sampling of features, and the criteria for split collectively providing the model's great generalization performance.

TABLE I
CHOSEN PARAMETERS

Parameter	Value
n_estimators	51
max_depth	15
min_samples_split	4
min_samples_leaf	3
n_features_per_tree	9
random_state	42

These resulted in the highest cross-validation accuracy and were subsequently compared to unseen data from the test set.

B. Performance Metrics

We used a collection of four widely accepted metrics, Accuracy, Precision, Recall, and F1-Score, to quantitatively compare the performance of the three classification models. So we can make a fair assessment of each model's sensitivity to varying types of errors, efficacy overall, and reliability per class. Together they provide an understanding of classification per model and multi-class anemia overall.

The performance metrics shown in Table II presents a comparison of the models' statistical predictive power, particularly considering the multi-class configuration such as anemia diagnosis.

TABLE II
COMPARISON OF PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.5723	0.58	0.57	0.56
Support Vector Machine	0.2381	0.23	0.24	0.22
Random Forest	0.9948	0.9949	0.9948	0.9948

C. Visual Results

To complement the quantitative results, we present a set of visualizations that give insight into the model's behavior and data characteristics.

To assess model interpretability, SHAP (SHapley Additive exPlanations) [11] was applied to the best-performing model. Figure 2 illustrates the contribution of each CBC feature to the final predictions across 100 test samples.

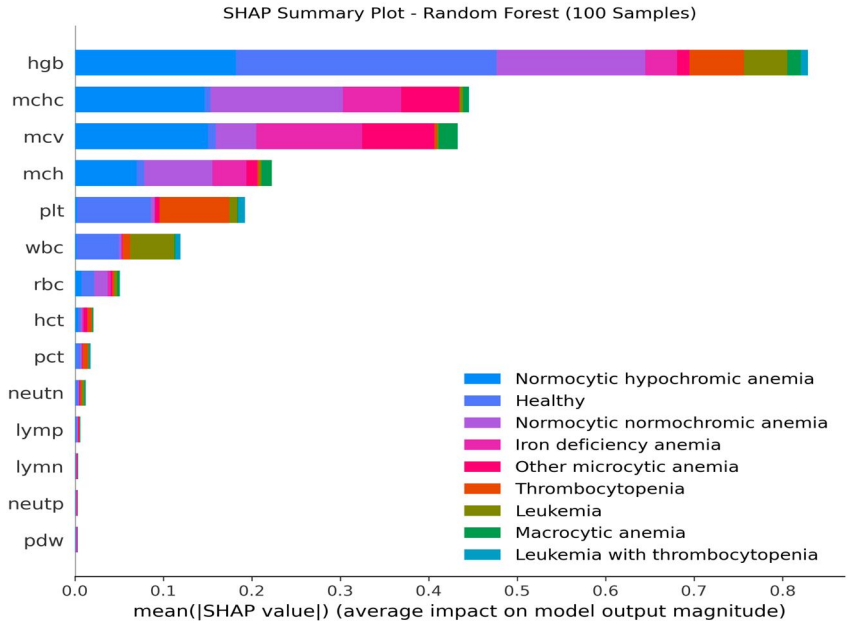


Fig. 2 SHAP summary plot for Random Forest model. Features are ranked by their average impact on model output.

The top contributing features include *Hemoglobin, Mean Corpuscular Hemoglobin Concentration, Mean Corpuscular Volume, Mean Corpuscular Hemoglobin,* and *Platelets,* confirming the clinical relevance of the model's decision-making process. This level of transparency is crucial for trust and adoption in healthcare contexts.

The confusion matrix in Figure 3 demonstrates the classification performance of the Random Forest model across all nine anemia categories. The matrix presents the distribution of predictions compared to the actual diagnoses in the test dataset.

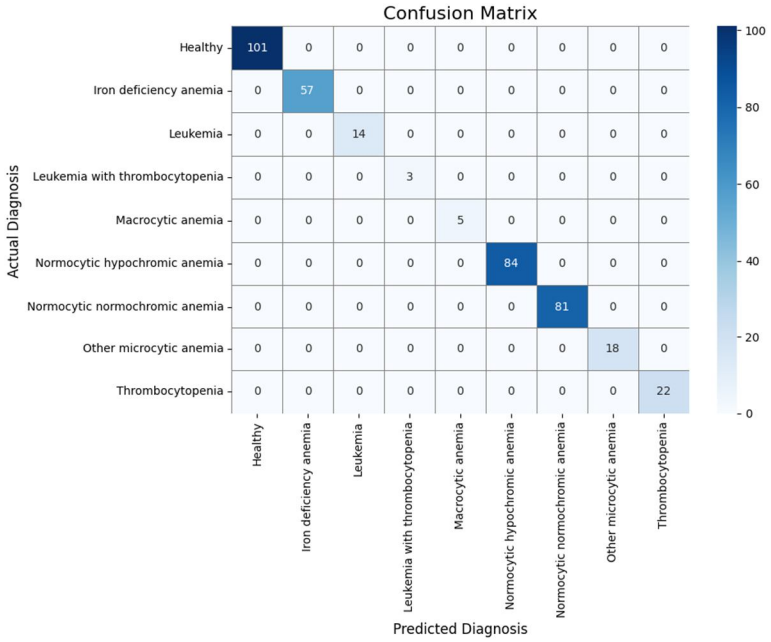


Fig. 3 Confusion Matrix for Random Forest predictions on the CBC dataset. Rows represent actual diagnoses, while columns represent predicted diagnoses.

To explore internal relationships within the dataset, a Pearson correlation matrix was generated among the CBC features. Figure 4 highlights how various hematological parameters relate to one another.

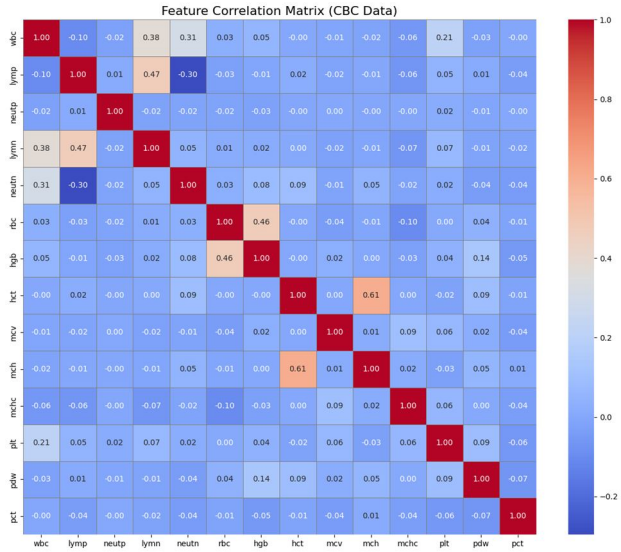


Fig. 4 Correlation matrix of CBC features. Strong positive and negative correlations are indicated by darker color intensities.

Notably, *Hemoglobin*, *Hematocrit*, and *RBC count* exhibit strong positive correlations, which aligns with clinical expectations. These correlations suggest that the model could exploit redundancies among key features during classification.

V. CONCLUSION

In this study, Random Forest, Logistic Regression, and Support Vector Machine (SVM) were evaluated to predict anemia with CBC data. After hyperparameter tuning to refine models through Grid Search, Random Forest achieved a high accuracy of 99.48% compared to Logistic Regression (57%) and SVM (23%) accuracy. These findings strongly suggest that ensemble methods are far superior at acquiring complex patterns in medical data when compared to Logistic Regression and SVM.

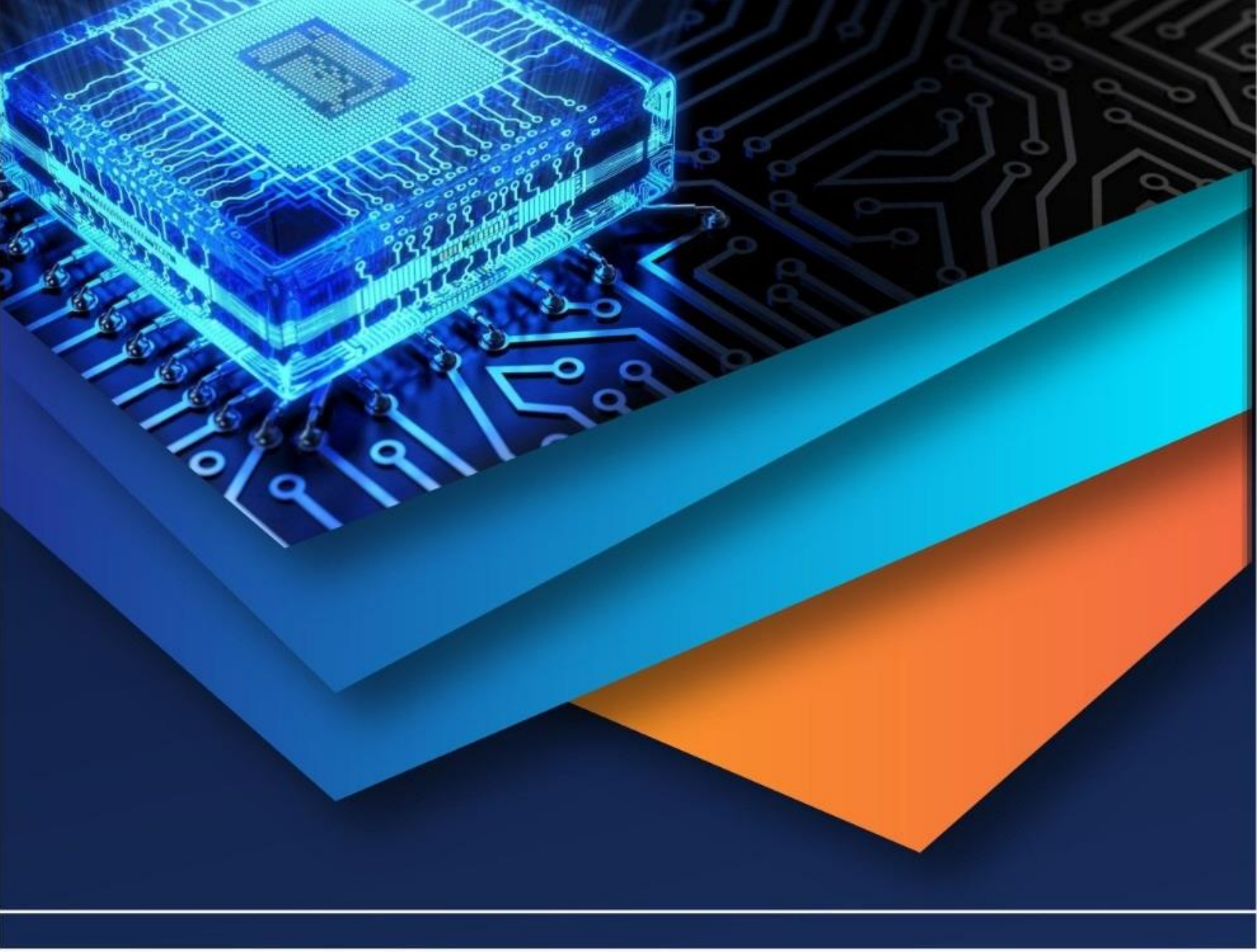
Random Forest emerged as the most appropriate model because it demonstrated high predictive accuracy, robustness to overfitting, and an ability to be compatible with explanation tools such as SHAP. It is the only algorithm that achieved performance and explanation. The latter is a high priority for clinical settings and relevance to clinical decision making.

VI. ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Engineering at Thadomal Shahani Engineering College, Mumbai, for their guidance and support in carrying out this research.We would like to acknowledge and thank Dr. Tanuja Sarode, Head of Computer Engineering Department, for her suggestions and guidance through this project. We would also like to thank Dr. G.T. Thampi, Principal of Thadomal Shahani Engineering College, for having the infrastructure and academic environment.

REFERENCES

[1] M. R. Aditya, T. Sutanto, H. Budiman, M. R. N. Ridha, U. Syapotro, and N. Azijah, “Machine learning models for classification of anemia from CBC results: Random Forest, SVM, and Logistic Regression,” J. Data Sci., vol. 2024, no. 49, 2024.
[2] Kitaw B, Asefa C, Legese F, et al. Leveraging machine learning models for anemia severity detection among pregnant women following ANC: Ethiopian context. BMC Public Health. 2024 Dec 18;24(1):3500.
[3] Gómez-Gómez J, Rico A, Guzmán JR, et al. Anemia Classification System Using Machine Learning. Informatics. 2025;12(1):19.
[4] Awaad AS, Elbarawy YM, Mancy H, Ghannam NE. Exploring CBC Data for Anemia Diagnosis: A Machine Learning and Ontology Perspective. BioMedInformatics. 2025;5(3):35.
[5] Shweta N, Pande SD. Prediction of Anemia using Various Ensemble Learning and Boosting Techniques. EAI Endorsed Transactions on Pervasive Health and Technology. 2023;10.4108/eetpht.9.4197.
[6] E. Aboelnaga. Anemia Types Classification [Online Kaggle dataset]. 2023. Available: https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification (accessed Jun. 21, 2025).
[7] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
[8] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. New York, NY, USA: Wiley, 2013.
[9] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[10] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[11] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Adv. Neural Inf. Process. Syst. 30 (NIPS), 2017, pp. 4765–4774.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY