International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04838
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Anemia Prediction Using Machine Learning Algorithms: A Comparative Analysis

[1]Hari Prakash G
Department of Biotechnology
KIT Kalaignarkarunanidhi Institute of Technology
Coimbatore, India

[2]Shanmugabharath V
Department of Biotechnology
KIT Kalaignarkarunanidhi Institute of Technology
Coimbatore, India

[3]Gokulakrishnan M
Department of Biotechnology
KIT Kalaignarkarunanidhi Institute of Technology
Coimbatore, India

Ms. Kamali M L (Corresponding Author)
Department of Biotechnology
KIT Kalaignarkarunanidhi Institute of Technology
Coimbatore, India

Abstract - A prevalent public health concern that impacts billions of individuals globally is anemia, especially in nations with low and middle incomes. Timely intervention and better patient outcomes depend on early diagnosis and detection. Using clinical blood test data, this study attempts to create and assess different machine learning (ML) models for anaemia prediction. A dataset of 1,421 patient records was analyzed using five supervised machine learning algorithms: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). Accuracy and ROC-AUC scores were used to assess performance. The models were interpreted using SHapley Additive exPlanations (SHAP). The findings showed that ensemble models outperformed other models with 100% accuracy and AUC. These results show how ML can be used in clinical diagnostic support systems.

Keywords: Anemia prediction, machine learning, SHAP, Random Forest, Gradient Boosting

1. INTRODUCTION

A major global health concern is anemia, a disorder marked by a lack of red blood cells or hemoglobin, either in quantity or quality. According to the World Health Organization (WHO), anemia affects 1.62 billion people worldwide, with the largest frequency occurring in children and women who are of reproductive age. Anaemia can cause extreme exhaustion, developmental delays, pregnancy complications, and elevated morbidity if it is not identified or treated.[1].

Traditionally, anemia diagnosis involves a series of laboratory tests, including complete blood counts and specific biomarkers like hemoglobin concentration, red blood cell indices, and iron levels. Despite their effectiveness, these tests take a lot of time and need to be interpreted by professionals. Delays in diagnosis could lead to worse clinical outcomes in areas with inadequate healthcare infrastructure.[2]. In order to enhance clinical decision-making, we employed artificial intelligence (AI) in this study, specifically machine learning (ML). ML models are highly accurate at predicting disease states and can identify patterns in vast amounts of medical data. They could provide prompt, reasonably priced, and easily accessible diagnostic assistance if properly implemented.[3]. This study explores the classification of individuals as either anemic or non-anemic based on standard haematological data using machine learning approaches. Five machine learning models' performances are compared, and SHapley Additive exPlanations (SHAP) is used for model explainability. Our goal is to support early anaemia diagnosis and contribute to the ongoing digital transformation in healthcare by developing interpretable and accurate models.[4][5].

2. Materials and Procedures

2.1 Description of the Dataset

This dataset is taken from Kaggle. The dataset used in this study contains 1,421 anonymized patient samples collected through routine clinical blood tests. The dataset includes several hematological parameters such as:

- Hemoglobin
- Mean Corpuscular Volume (MCV)
- Mean Corpuscular Hemoglobin (MCH)
- Red Cell Distribution Width (RDW)

The target variable is a binary label showing whether anemia is present (1) or not (0)

2.2 Preprocessing of Data

For any machine learning to be reliable and of high quality, data preparation is essential. First, the dataset was examined for any missing, null, or inconsistent values. Invalid entries were removed and encoded appropriately. Numerical features were standardised using z-score normalisation to maintain

uniform scales across features. The final cleaned dataset was split into two subsets:

- Training set: 75% of the data (1,065 samples)
- Testing set: 25% of the data (356 samples)

This preprocessing ensured that the model was trained on consistent, normalised data, thereby minimising the risks of overfitting and data leakage.

2.3 Machine Learning Models

The following models were used for the supervised classification.

- **Random Forest Classifier**

This ensemble method creates a decision tree "forest" by selecting random subsets of the training data and characteristics. Each tree votes on the outcome, and the majority vote determines the final forecast. We chose this model because of its robustness on structured healthcare data and capacity to minimise overfitting. Each class receives a probability score as the output, which is subsequently used to designate anaemia.[6]

- **Gradient Boosting Classifier**

Gradient Boosting constructs trees in a sequential manner, in contrast to Random Forest. Every new tree attempts to fix the mistakes of its previous ones. Complex non-linear relationships can be handled well by this model. Here, we employed it to record minute changes in haematological parameters that could point to anaemia. This model shows a more accurate prediction with less bias and variance as the end result.[7]

- **Logistic Regression**

A probabilistic interpretation of binary classification is offered by logistic regression as a baseline model. We used this because of its ease of use and interpretability; it is frequently utilised in medical diagnostics. We benchmarked the performance of more intricate models using it. A probability value that is thresholded to assign a class label (anaemia or not) is produced by the model.[8]

- **Support Vector Machine (SVM)**

In a high-dimensional space, SVM determines the best hyperplane to divide the two classes. When data cannot be separated linearly, it is especially helpful. To capture non-linear patterns, we employed the RBF kernel. Along with a decision function that indicates the classification's level of confidence, the SVM model should produce a class label.[9]

- **K-Nearest Neighbours (KNN)**

To categorise a sample, KNN uses the majority vote from its "k" nearest neighbours in the training set. It is an easy-to-understand instance-based learning algorithm. In order to compare performance with both simple and ensemble models, we included KNN. This is used because its decision boundaries can be used to visualise the class label that the model returns based on proximity.[10]

All models were implemented using the scikit-learn library with default hyperparameters. The purpose was to compare their performance on the same dataset with minimal tuning.

2.4 Metrics for Evaluation

Models were assessed using:

- **Accuracy**: The percentage of accurate forecasts
- **ROC-AUC**: Measures the ability of the model to distinguish between classes
- **Confusion Matrix**: Displays true/false positives and negatives
- **SHAP Analysis**: Assesses feature contributions to individual predictions

3. RESULTS

3.1 Model Performance

To assess the categorisation performance of each model, the following metrics were calculated:

- **Accuracy**: The proportion of accurately forecasted observations to all observations.
- **Precision**: The ratio of correctly predicted positive observations to total predicted positives.
- **Recall** (Sensitivity): The ratio of correctly predicted positives to all actual positives.
- **F1 Score**: Precision and Recall weighted average.
- **AUC Score**: Area under the ROC curve; measures separability of classes.

Random Forest and Gradient Boosting both achieved perfect performance in both metrics, while Logistic Regression showed strong generalization. SVM and KNN, although slightly less accurate, still demonstrated reasonable discriminatory power.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 04 Issue: 07 | July – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM04838

Table -1: Model Performance

Model	Precision	Recall	F1 Score	AUC Score	Accuracy
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000
Gradient Boosting	1.0000	1.0000	1.0000	1.0000	1.0000
Logistic Regression	0.9789	1.0000	0.9893	0.9871	0.9888
Support Vector Machine (SVM)	0.8695	0.9772	0.9204	0.9504	0.9157
K-Nearest Neighbors (KNN)	0.8431	0.8863	0.8641	0.9210	0.8652

3.2 AUC - ROC Curve Analysis

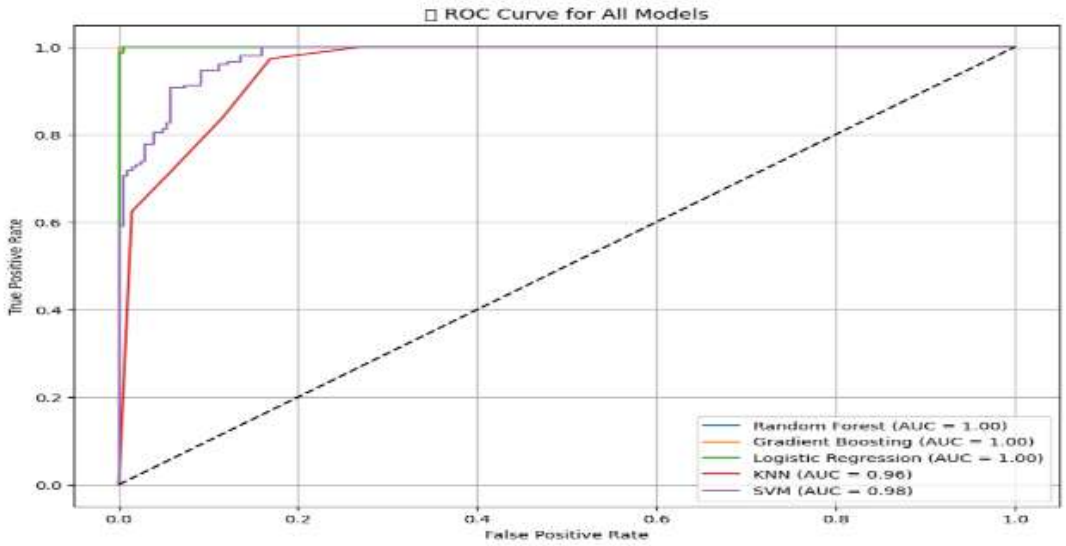


Figure 1: *ROC curves for all five models showing comparative classification ability*

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04838
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

3.3 Confusion Matrix

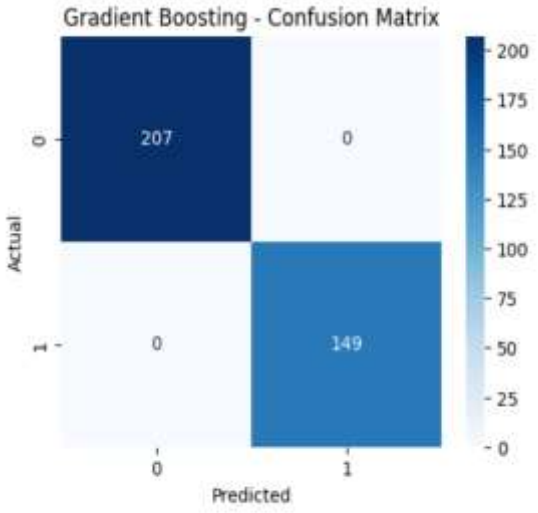


Figure 2: *Confusion matrix for the Random Forest model indicating no misclassifications*

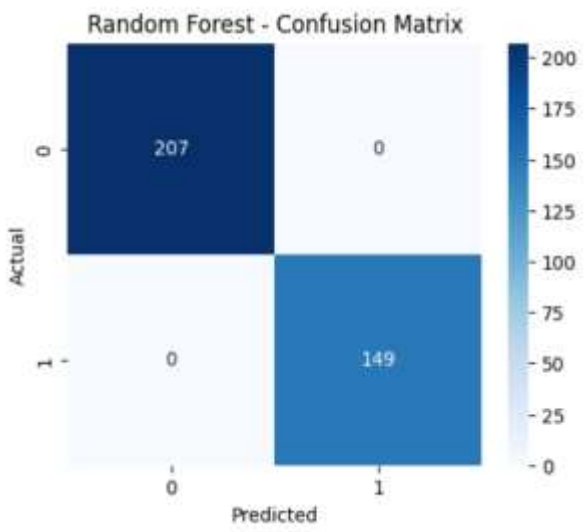


Figure 3: *Confusion matrix for the Gradient Boosting model indicating no misclassifications*

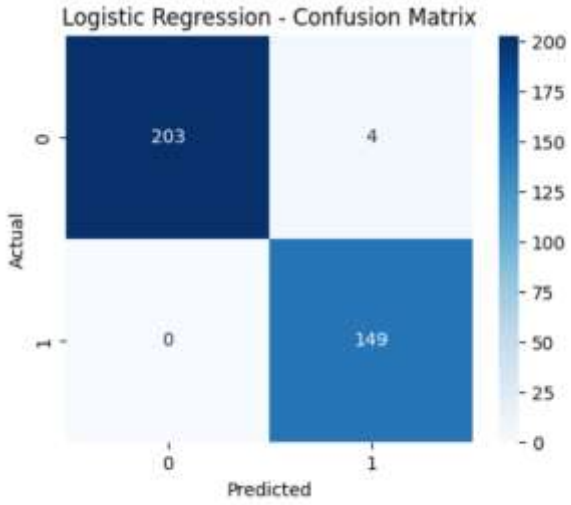


Figure 4: *Confusion matrix for the Logistic Regression model indicating no misclassifications*

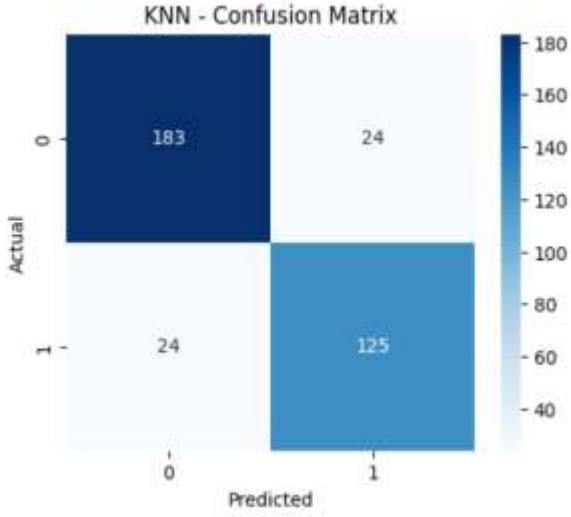
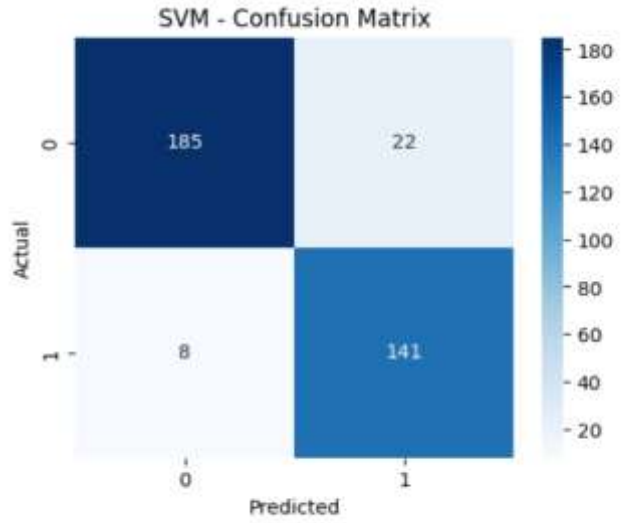


Figure 5: *Confusion matrix for KNN model indicating no misclassifications*

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04838
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata



The confusion matrix of the Support Vector Machine (SVM) model for anemia prediction reveals the classification performance across two classes: anemic (1) and non-anemic (0). The matrix shows that out of 207 actual non-anemic cases, the model correctly predicted 185 as non-anemic (true negatives) and misclassified 22 as anemic (false positives). For the anemic class, the model correctly identified 141 out of 149 cases (true positives), with only 8 instances misclassified as non-anemic (false negatives). These results indicate that the SVM model achieved high classification accuracy, with relatively low misclassification rates. The model exhibits strong generalisation capability in distinguishing between anemic and non-anemic patients, demonstrating its effectiveness as a diagnostic tool in medical data analysis.

Figure 6: *Confusion matrix for SVM model indicating no misclassifications*

3.4 Feature Importance

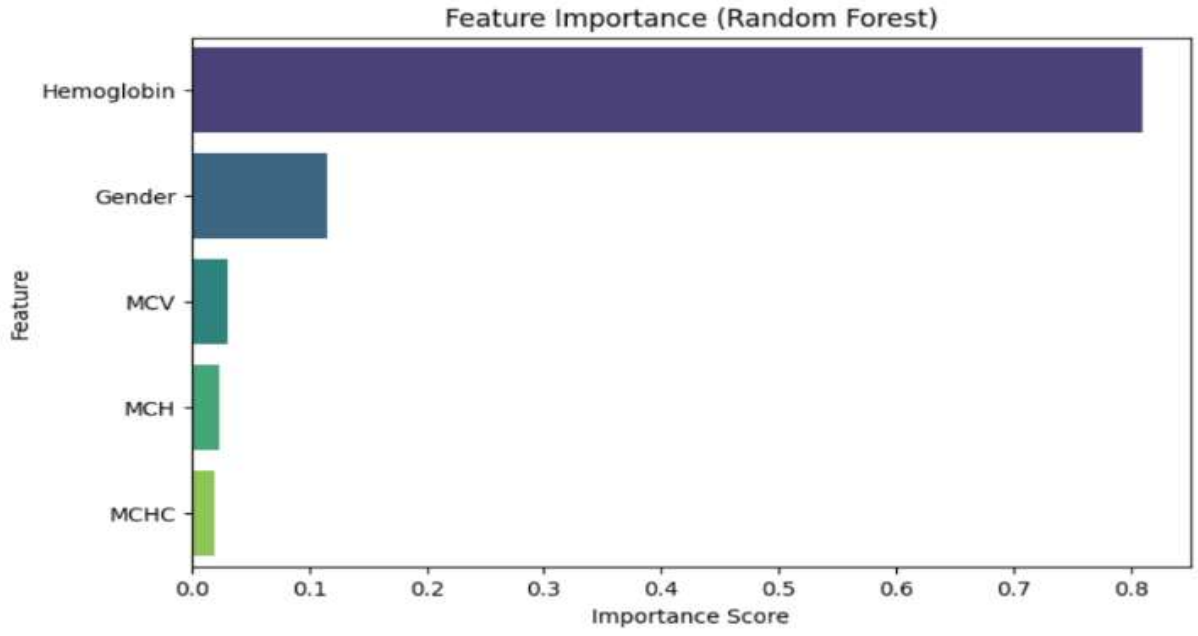


Figure 7*: Relevance of features in the Random Forest model*

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04838
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

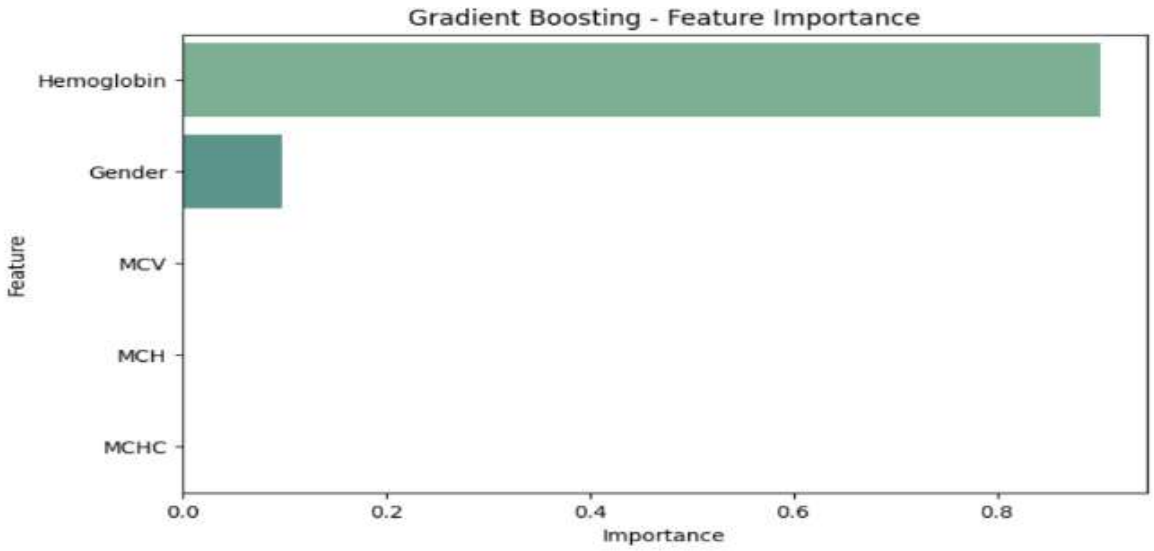


Figure 8: *Feature importance in the Gradient Boost model*

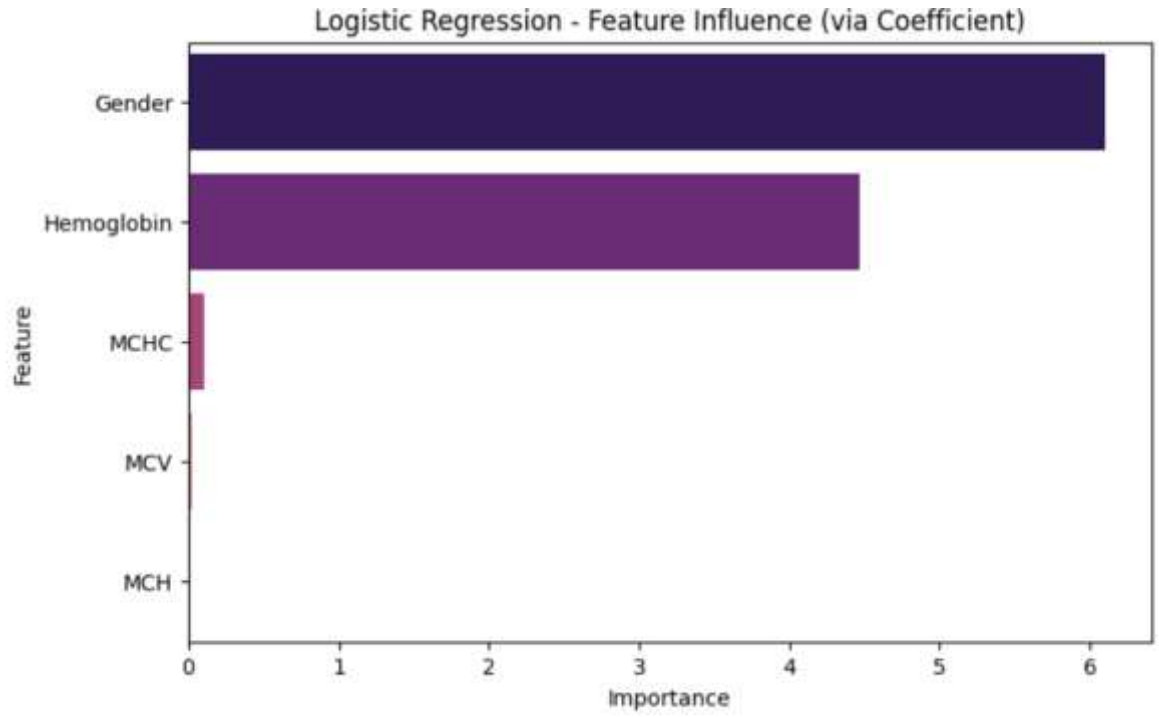


Figure 9: *Feature importance in the Logistic Regression model*

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 04 Issue: 07 | July – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM04838

SHAP Feature Importance

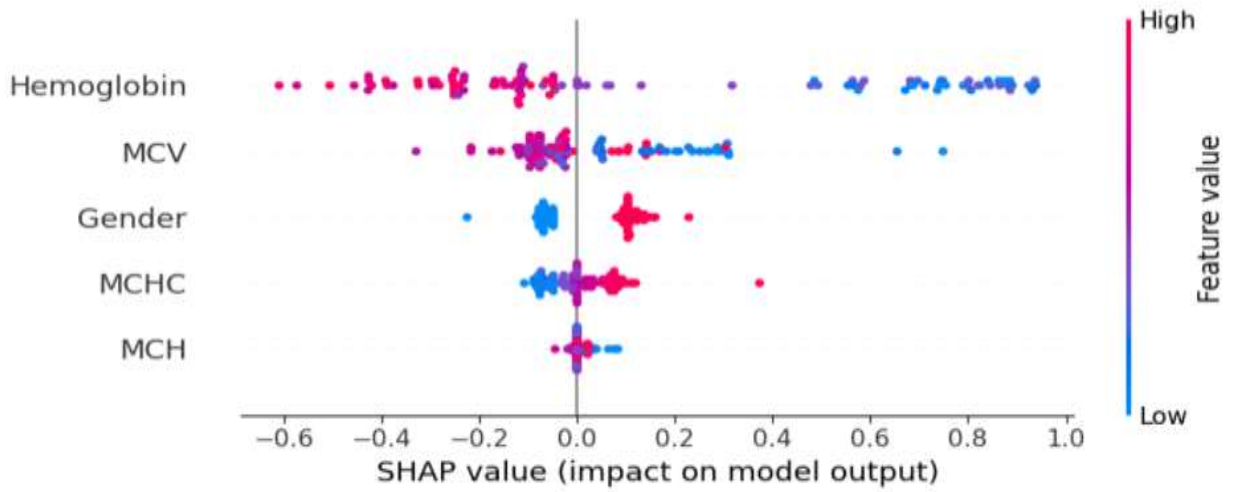


Figure 10: *SHAP summary plot highlighting hemoglobin, MCV, MCH, and RDW as top predictive features (KNN)*

.

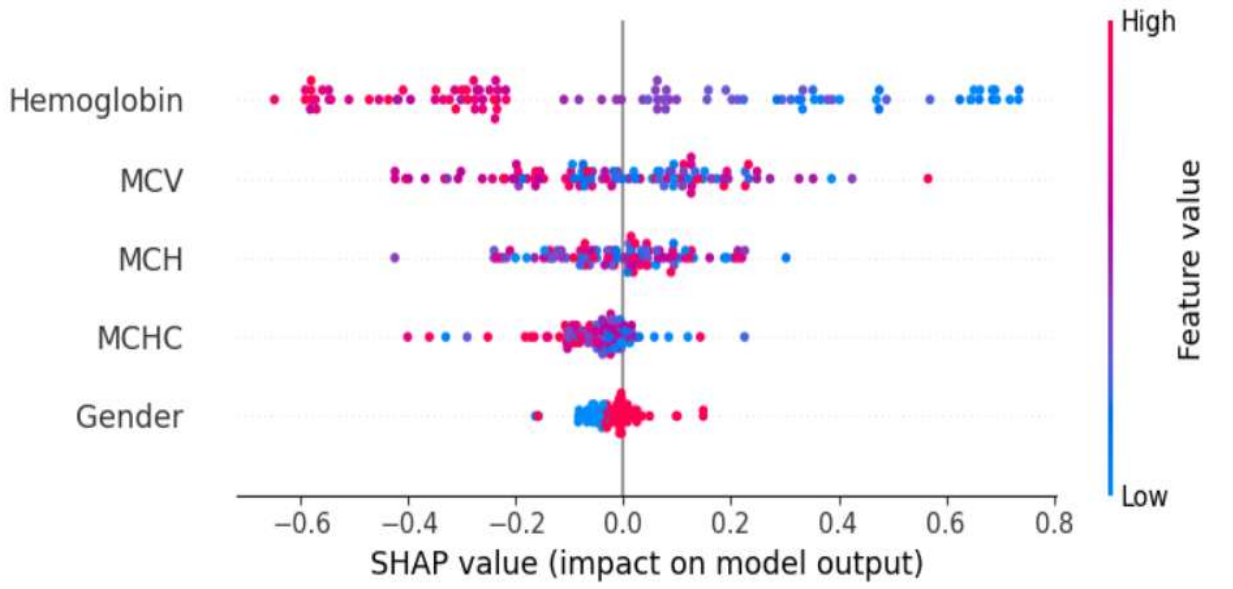


Figure 11*: SHAP summary plot highlighting hemoglobin, MCV, MCH, and RDW as top predictive features.(SVM)*

SHAP (SHapley Additive exPlanations) The contribution of each feature to the model's predictions for anemia was interpreted using analysis. Out of all the input variables, Red Cell Distribution Width (RDW), Mean Corpuscular Volume (MCV), Mean Corpuscular Haemoglobin (MCH), and hemoglobin emerged as the most impactful features. SHAP summary plots visually reinforced the dominant role of Hemoglobin across all machine learning models, highlighting its critical influence on prediction outcomes. This observation not only aligns with established clinical understanding where hemoglobin is a primary indicator of anemia, but also

enhances the interpretability and trustworthiness of the AI system in a healthcare context.[7]

4. DISCUSSION

The findings demonstrate that ensemble models such as Random Forest and Gradient Boosting perform better when it comes to forecasting anemia from blood data. These models achieved 100% accuracy and AUC, suggesting high robustness and generalisation in the dataset used. SHAP analysis provided critical insights into feature importance, increasing clinical knowledge that red blood cell and hemoglobin indices are important markers of anemia. The discrepancy between accuracy and AUC in models like SVM reflects their threshold-dependent nature, where AUC evaluates ranking performance independent of the classification threshold. This underscores the importance of using multiple metrics for model assessment in medical applications.

5. CONCLUSION

Based on commonly available haematological parameters, according to this study, machine learning (ML) models have the potential to completely transform the early detection and classification of anaemia. We implemented and evaluated five supervised machine learning (ML) algorithms using a dataset of 1,421 patient records: Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Gradient Boosting and Logistic Regression. Among these, ensemble-based techniques like Random Forest and Gradient Boosting showed flawless classification performance, achieving 100% F1-score, AUC, precision, accuracy and recall. These findings support the models' resilience and flexibility in managing structured clinical data. A more straightforward linear model, logistic regression, also demonstrated exceptional performance, confirming its ongoing applicability in medical settings. SVM and KNN, on the other hand, performed fairly well but had somewhat lower metrics, indicating that they might need more fine-tuning or be more appropriate for different kinds of datasets. In this study, we used SHapley Additive exPlanations (SHAP), which improved the models' interpretability, is one of its key advantages. Haemoglobin, MCV, MCH, and RDW were the most significant features across models, according to SHAP analysis. These results strongly support the reliability of ML-powered diagnostic tools since they are in line with clinical experience. In conclusion, a strong, precise, and understandable method for anaemia screening is provided by combining machine learning models, particularly ensemble approaches, with explainability tools such as SHAP. These tools have the potential to significantly close diagnostic gaps as healthcare systems around the world transition to data-driven technologies, particularly in settings with limited resources. To further improve clinical utility, future research should concentrate on real-time model deployment, electronic health record integration, and the investigation of multiclass anaemia classification.

ACKNOWLEDGMENT

The research and model development were conducted as part of a real-world machine learning application project.

REFERENCES

[1] M. M. Usta, M. Çakmak, and D. Ekmekçi, “Anemia Types Prediction Using Ensemble Learning”.
[2] W. Tepakhan, W. Srisintorn, T. Penglong, and P. Saelue, “Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms,” *Sci Rep*, vol. 15, no. 1, May 2025, doi: 10.1038/s41598-025-01458-5.
[3] P. Dhakal, “Prediction of Anemia using Machine Learning Algorithms,” *IJCSIT*, vol. 15, no. 1, pp. 15–30, Feb. 2023, doi: 10.5121/ijcsit.2023.15102.
[4] J. G. Gómez, C. Parra Urueta, D. S. Álvarez, V. Hernández Riaño, and G. Ramirez-Gonzalez, “Anemia Classification System Using Machine Learning,” *Informatics*, vol. 12, no. 1, p. 19, Feb. 2025, doi: 10.3390/informatics12010019.
[5] Y. Li, “Analyzing the Application of Machine Learning in Anemia Prediction,” *ITM Web Conf.*, vol. 70, p. 04006, 2025, doi: 10.1051/itmconf/20257004006.
[6] P. Dhakal, “Prediction of Anemia using Machine Learning Algorithms,” *IJCSIT*, vol. 15, no. 1, pp. 15–30, Feb. 2023, doi: 10.5121/ijcsit.2023.15102.
[7] S. Pullakhandam and S. McRoy, “Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning,” *BioMedInformatics*, vol. 4, no. 1, pp. 661–672, Mar. 2024, doi: 10.3390/biomedinformatics4010036.
[8] K. Animut and G. Berhanu, “Determinants of anemia status among pregnant women in ethiopia: using 2016 ethiopian demographic and health survey data; application of ordinal logistic regression models,” *BMC Pregnancy Childbirth*, vol. 22, no. 1, Aug. 2022, doi: 10.1186/s12884-022-04990-8.
[9] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, “A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia,” *Healthcare*, vol. 11, no. 5, p. 697, Feb. 2023, doi: 10.3390/healthcare11050697.
[10] Z. Faradila, A. Homaidi, and J. D. Prasetyo, “Classification of Anaemia Status Using The K-Nearest Neighbor Algorithm”.