iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84443

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 14 Issue VII July 2026- Available at www.ijraset.com

Machine Learning-Based Anemia Prediction: A Comparative Analysis of Random Forest, Lasso, and Ridge Algorithms

Deepika Kolluru[1], Gangadhar Doma[2]
[1]*Student, Master of Computer Applications (MCA), AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Visakhapatnam, Andhra Pradesh, India*
[2]*Associate Professor, Department of MCA, AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Visakhapatnam, Andhra Pradesh, India*

Abstract: ***Anemia is one of the most common hematological disorders affecting millions of people worldwide and is characterized by reduced hemoglobin concentration or insufficient healthy red blood cells. Early diagnosis plays a significant role in preventing severe health complications and improving patient care. Machine Learning (ML) has emerged as an effective approach for developing intelligent clinical decision-support systems capable of analyzing medical data with high accuracy.***
This paper presents a machine learning-based anemia prediction system using Complete Blood Count (CBC) parameters. Three supervised machine learning algorithms, namely Random Forest, Logistic Regression with L1 (Lasso) regularization, and Logistic Regression with L2 (Ridge) regularization, were implemented and evaluated using the same dataset and preprocessing pipeline. The proposed model utilizes five important hematological parameters: Gender, Hemoglobin (Hb), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), and Mean Corpuscular Volume (MCV) for predicting whether a patient is likely to have anemia. A graphical user interface (GUI) developed using Python Tkinter enables users to enter patient blood parameters and receive immediate prediction results. Experimental evaluation was performed using Accuracy, Precision, Recall, and F1-Score. Among the evaluated algorithms, the Random Forest classifier achieved the highest accuracy of 100.00%, outperforming Lasso (99.30%) and Ridge (98.95%). Based on the comparative study, Random Forest was selected as the final deployment model. The proposed system provides a fast, accurate, and user-friendly approach for anemia prediction and demonstrates the effectiveness of ensemble learning techniques in healthcare applications.
Keywords: Machine Learning, Anemia Prediction, Random Forest, Logistic Regression, Lasso Regression, Ridge Regression, CBC Parameters, Healthcare Analytics, Python, Tkinter.

I. INTRODUCTION

Anemia is a major public health concern affecting individuals across all age groups. According to the World Health Organization (WHO), anemia is characterized by a reduction in hemoglobin concentration below the normal range required to meet the body's oxygen demand. It commonly results from nutritional deficiencies, chronic diseases, blood loss, or inherited disorders. Early diagnosis is essential because untreated anemia can lead to fatigue, impaired cognitive function, reduced immunity, pregnancy-related complications, and other severe health conditions. The integration of machine learning techniques into healthcare has enabled faster and more reliable disease prediction, supporting clinicians in making informed decisions. Traditionally, anemia diagnosis depends on laboratory investigations and interpretation of Complete Blood Count (CBC) reports by healthcare professionals. Although laboratory methods provide accurate measurements, manual interpretation may become time-consuming when processing large numbers of patient records. The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has enabled the development of intelligent systems capable of assisting clinicians by automatically identifying disease patterns from clinical data. Machine learning algorithms analyze relationships among hematological parameters and learn predictive patterns from historical datasets. Such models can provide rapid and reliable predictions while reducing manual effort. These systems act as decision-support tools and assist medical professionals in identifying patients who may require further clinical examination. The proposed work presents a comparative analysis of three supervised machine learning algorithms: Random Forest, Logistic Regression with L1 (Lasso) regularization, and Logistic Regression with L2 (Ridge) regularization. The models were trained and evaluated using the same preprocessing pipeline to ensure a fair comparison.

A Python Tkinter-based graphical user interface (GUI) was also developed to provide an interactive platform where users can enter patient details and obtain immediate anemia prediction results. Experimental analysis demonstrated that Random Forest achieved the highest predictive performance among the evaluated models and was selected as the final prediction model for deployment.

II. RELATED WORK

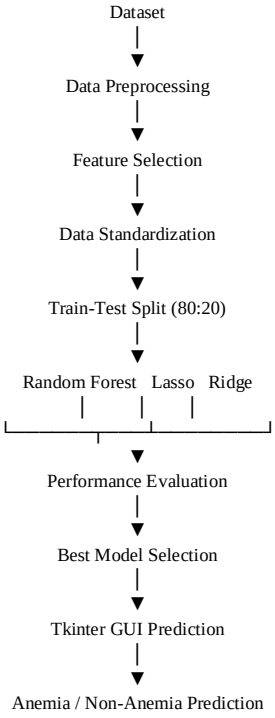
Several researchers have investigated the application of machine learning techniques for anemia prediction using hematological parameters obtained from Complete Blood Count (CBC) tests. Traditional classification algorithms such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), Naïve Bayes, Artificial Neural Networks, and ensemble learning techniques have demonstrated promising performance in predicting anemia and related blood disorders. Recent studies indicate that ensemble learning methods generally achieve higher predictive accuracy for structured medical datasets because they combine the predictions of multiple decision trees and effectively reduce overfitting. Logistic Regression remains one of the most widely used classification techniques due to its simplicity, computational efficiency, and interpretability. Regularization methods such as Lasso (L1) and Ridge (L2) further improve Logistic Regression by reducing overfitting and enhancing model generalization. Despite these developments, many existing anemia prediction systems focus primarily on prediction accuracy and provide limited comparison among multiple algorithms. Furthermore, only a few studies integrate an interactive graphical user interface that allows healthcare professionals or end users to obtain real-time prediction results. The proposed work addresses these limitations by implementing and comparing three supervised machine learning algorithms under identical preprocessing conditions. The developed system also includes a user-friendly Tkinter-based graphical interface that facilitates practical deployment for anemia prediction.

III. PROPOSED METHODOLOGY

The proposed anemia prediction framework consists of the following stages:

1) Data Collection
2) Data Preprocessing
3) Feature Selection
4) Data Standardization
5) Train-Test Split (80:20)
6) Model Training
7) Model Evaluation
8) GUI-Based Prediction

Figure 1. Workflow of the Proposed Anemia Prediction System



The dataset contains hematological parameters commonly obtained from Complete Blood Count (CBC) reports. During preprocessing, missing values were handled where necessary, categorical variables such as Gender were encoded, and numerical features were standardized using the StandardScaler technique to improve model performance.

Five clinical parameters were selected as input features:

- Gender
- Hemoglobin (Hb)
- Mean Corpuscular Hemoglobin (MCH)
- Mean Corpuscular Hemoglobin Concentration (MCHC)
- Mean Corpuscular Volume (MCV)

Three supervised machine learning algorithms were trained using the same dataset and preprocessing pipeline:

- Random Forest
- Logistic Regression with L1 (Lasso) Regularization
- Logistic Regression with L2 (Ridge) Regularization

Each model was evaluated using Accuracy, Precision, Recall, and F1-Score. Comparative analysis was performed to determine the most suitable algorithm for anemia prediction. Finally, the best-performing model was integrated into a Python Tkinter-based graphical user interface to enable real-time prediction based on patient input.

IV. MACHINE LEARNING ALGORITHMS

To identify the most suitable model for anemia prediction, three supervised machine learning algorithms were implemented and evaluated under identical preprocessing conditions. All models were trained using the same training dataset and tested using the same testing dataset to ensure a fair comparison.

A. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree is trained using a randomly selected subset of the training data and features. During prediction, every tree independently predicts the output class, and the final prediction is determined through majority voting.

Random Forest effectively captures complex nonlinear relationships among hematological parameters, making it highly suitable for medical diagnosis and classification problems.

Advantages

- High prediction accuracy
- Handles nonlinear relationships effectively
- Robust against overfitting
- Works well with clinical datasets
- Handles feature interactions efficiently

B. Lasso (Logistic Regression with L1 Regularization)

Lasso uses Logistic Regression with L1 regularization. The L1 penalty encourages sparsity by reducing less important feature coefficients toward zero, thereby simplifying the model and improving interpretability.

Lasso is particularly useful when feature selection is required because it automatically reduces the influence of less significant variables while maintaining prediction performance.

Advantages

- Performs automatic feature selection
- Reduces overfitting
- Produces simpler and interpretable models
- Efficient for binary classification problems

C. Ridge (Logistic Regression with L2 Regularization)

Ridge uses Logistic Regression with L2 regularization. Unlike Lasso, Ridge shrinks model coefficients without eliminating them completely. This improves model stability and reduces variance while retaining information from all input features.

Ridge performs well when all clinical features contribute to prediction and helps improve model generalization.

Advantages

- Reduces overfitting
- Improves model stability
- Handles multicollinearity effectively
- Maintains contributions from all features

V. GRAPHICAL USER INTERFACE (GUI)

A graphical user interface was developed using Python Tkinter to simplify interaction with the prediction model. The GUI enables healthcare professionals and users to enter patient information without requiring programming knowledge.

The interface accepts the following clinical parameters:

1) Gender
2) Hemoglobin (Hb)
3) Mean Corpuscular Hemoglobin (MCH)
4) Mean Corpuscular Hemoglobin Concentration (MCHC)
5) Mean Corpuscular Volume (MCV)

Input validation is performed to ensure that only valid numerical values are accepted for hematological parameters. After entering the patient details, the user clicks the **Predict** button. The trained machine learning model processes the input and displays one of the following outcomes:

- Patient has Anemia
- Patient does not have Anemia

The GUI provides a simple, intuitive, and efficient mechanism for real-time anemia prediction. The GUI also validates user input by ensuring that hematological parameters contain valid numerical values before prediction.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed anemia prediction system was evaluated using three supervised machine learning algorithms under identical preprocessing conditions. Performance evaluation was conducted using Accuracy, Precision, Recall, and F1-Score.

The obtained results are summarized in Table 1.

Table 1. Performance Comparison of Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	100.00%	1.000	1.000	1.000
Lasso (L1)	99.30%	0.992	0.994	0.993
Ridge (L2)	98.95%	0.988	0.991	0.989

The experimental results demonstrate that ensemble learning through Random Forest provides superior predictive performance compared to regularized Logistic Regression models for the selected anemia dataset.

The comparative analysis indicates that all three algorithms achieved excellent classification performance on the anemia dataset. However, the Random Forest classifier consistently outperformed both Logistic Regression models with L1 and L2 regularization.

The superior performance of Random Forest can be attributed to its ensemble learning strategy, which combines multiple decision trees to improve prediction accuracy while reducing overfitting. The algorithm effectively captures complex nonlinear relationships among hematological parameters that may not be identified by linear classification models.

Although Lasso and Ridge also achieved excellent predictive performance, Random Forest demonstrated the highest Accuracy, Precision, Recall, and F1-Score, making it the most suitable algorithm for deployment in the proposed anemia prediction system.

VII. CONCLUSION

This paper presented a machine learning-based anemia prediction system using hematological parameters obtained from Complete Blood Count (CBC) reports. Three supervised machine learning algorithms—Random Forest, Logistic Regression with L1 (Lasso) regularization, and Logistic Regression with L2 (Ridge) regularization—were implemented and compared using the same preprocessing pipeline and evaluation methodology.

Experimental results demonstrated that all three algorithms achieved high predictive performance. Among them, the Random Forest classifier achieved the highest accuracy of **100.00%**, outperforming both Lasso (**99.30%**) and Ridge (**98.95%**). Based on this comparative evaluation, Random Forest was selected as the final prediction model.

The developed Python Tkinter-based graphical user interface allows users to enter patient hematological parameters and obtain immediate prediction results. The proposed system provides an efficient, accurate, and user-friendly decision-support tool that can assist healthcare professionals in the early identification of anemia. The developed system demonstrates the practical application of machine learning in healthcare and has the potential to assist clinicians in early anemia screening.

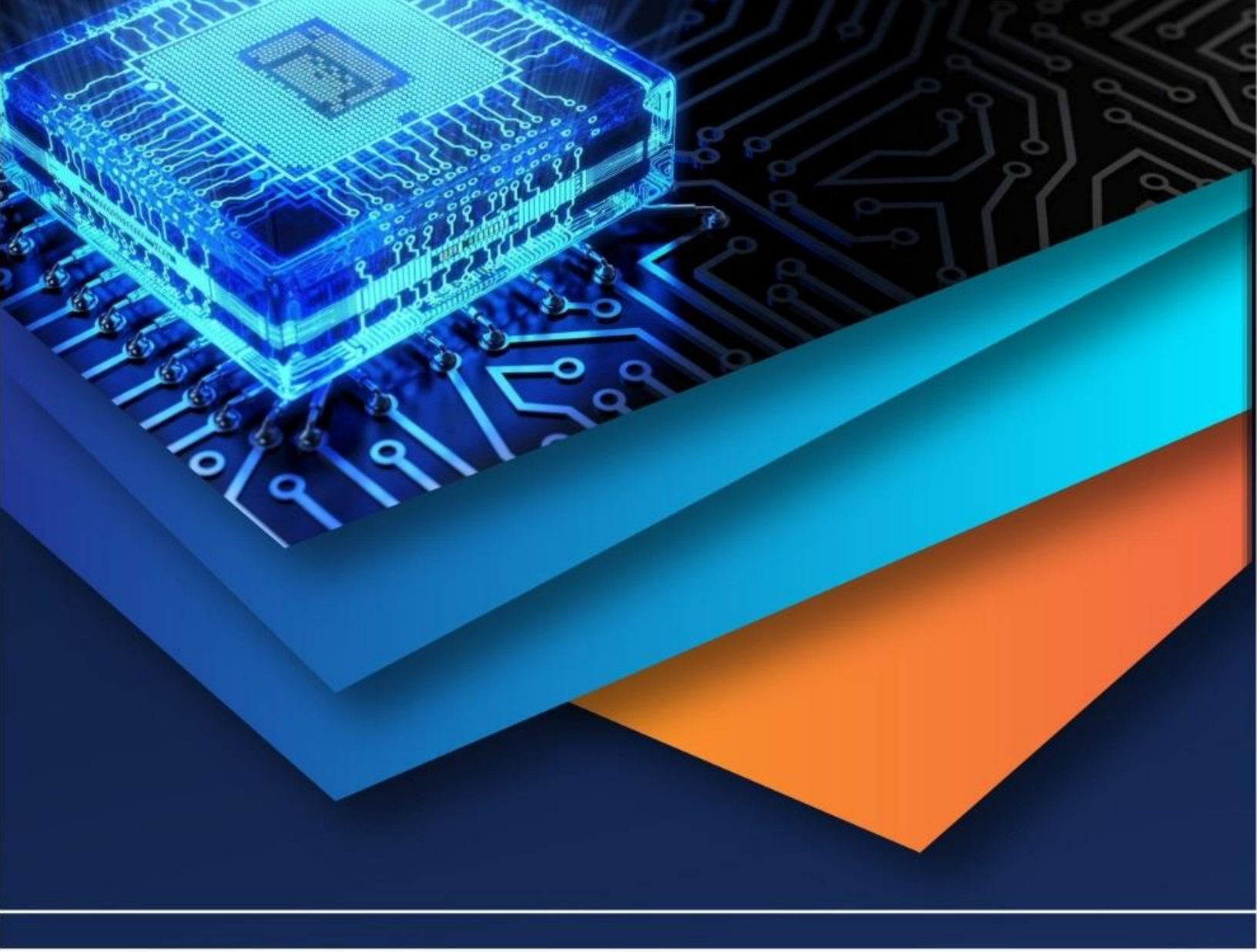
VIII. FUTURE SCOPE

The proposed anemia prediction system can be extended by incorporating larger and more diverse clinical datasets collected from multiple healthcare institutions. Future work may include evaluating advanced machine learning algorithms such as Support Vector Machines (SVM), Gradient Boosting, XGBoost, and deep learning models to further improve prediction performance.

The current desktop application can also be enhanced into a web-based or mobile healthcare platform to enable remote access for clinicians and patients. Integration with hospital information systems and electronic health records (EHRs) can further improve clinical usability and support real-time decision making.

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[3] World Health Organization (WHO), "Anaemia," Geneva, Switzerland, 2023.
[4] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., John Wiley & Sons, 2013.
[5] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[6] W. McKinney, Python for Data Analysis, 3rd ed., O'Reilly Media, 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY