GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

# Anemia Disease Prediction and Severity Classification Using Optimized Machine Learning Frameworks

Jyoti Ranjan Sahu, UG Student, Dept. of CSE, GIET University Gunupur,
jyotiranjansahuofficial18@gmail.com
Sidhi Ranjan Dash, UG Student, Dept. of CSE, GIET University Gunupur,
sidhiranjandash@gmail.com
Dillip Kumar Mishra, Assistant Professor, Dept. of CSE, GIET University Gunupur,
dillipkumarmishra@gmail.com
Sk Rehan Hossain, UG Student, Dept. of CSE, GIET University Gunupur,
hossainrehan72@gmail.com

**Abstract: Anemia remains one of the most prevalent global health challenges, affecting nearly a quarter of the world's population and contributing significantly to morbidity, reduced quality of life, and increased healthcare burden, particularly in low- and middle-income regions. This research presents a comprehensive Machine Learning (ML)-based framework for the automated detection and severity classification of anemia using routinely collected hematological indicators. The proposed framework utilizes four clinically significant blood parameters: Hemoglobin (Hb), Red Blood Cell (RBC) count, Mean Corpuscular Volume (MCV), and Mean Corpuscular Hemoglobin (MCH), which together capture both quantitative and morphological aspects of red blood cells. To improve generalization and robustness, an integrated dataset was constructed by aggregating patient records from five independent clinical repositories, ensuring diversity across demographic groups and clinical conditions. A comparative experimental study was conducted using three widely adopted machine learning algorithms: Logistic Regression, Support Vector Machine (SVM), and Random Forest. Each model was trained and evaluated under identical conditions using cross-validation and standard performance metrics including accuracy, precision, recall, and F1-score. The results demonstrate that the optimized Random Forest classifier consistently outperforms the other models, achieving an accuracy of 96.8 Beyond model development, this study proposes a cloud-based deployment architecture designed to enable real-time, remote diagnostic support for healthcare professionals. By combining high predictive accuracy with practical deployability, the proposed framework offers a promising solution for improving early anemia detection, supporting clinical decision-making, and reducing diagnostic delays in underserved regions. Overall, this work demonstrates the potential of machine learning to enhance traditional diagnostic workflows by providing accurate, efficient, and accessible tools for anemia screening and severity assessment, thereby contributing to improved public health outcomes and more equitable healthcare delivery.**

***Keywords*— Virtual Painter, Gesture Recognition, OpenCV, Mediapipe, Real Time Drawing, Hands Free Interaction, Python Application.**

## 1. Introduction:

Anemia is one of the most widespread hematological dis- orders worldwide, affecting approximately 24.8 % population. Despite its high prevalence and serious consequences, anemia diagnosis and severity assessment in many healthcare settings remain largely dependent on manual

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

***ISSN: 2583-5955***

interpretation of laboratory results and predefined clinical thresholds. Such approaches are time-consuming, prone to human error, and often lack consistency across different healthcare providers. Furthermore, in rural and underserved areas, limited access to trained hematologists and diagnostic infrastructure leads to delayed diagnosis and inadequate follow-up.

These challenges highlight the need for intelligent, automated diagnostic systems that can support clinicians, reduce workload, and improve diagnostic accuracy while remaining accessible and scalable. Recent advances in machine learning (ML) have demonstrated significant potential in transforming healthcare by enabling data-driven decision-making, predictive analytics, and automated pattern recognition in complex clinical datasets. ML techniques have been successfully applied to a wide range of medical applications, including disease diagnosis, prognosis prediction, medical image analysis, and personalized treatment planning. By learning from historical patient data, ML models can uncover subtle relationships among clinical variables that may not be apparent through conventional statistical or rule- based methods.

In the context of anemia detection, hematological parameters such as Hemoglobin (Hb), Red Blood Cell (RBC) count, Mean Corpuscular Volume (MCV), and Mean Corpuscular Hemoglobin (MCH) play a central role in identifying anemia and distinguishing between its different types and severity levels. These parameters are routinely measured in complete blood count (CBC) tests, making them readily available and cost-effective for large-scale screening. Leveraging these parameters within an ML framework offers a promising opportunity to develop automated systems capable of accurate and efficient anemia prediction.

This paper proposes a comprehensive machine learning framework for the automated detection and severity classification of anemia using routinely collected hematological indices. The framework integrates data from multiple clinical repositories to improve model generalization and robustness across diverse patient populations. A comparative analysis is conducted using Logistic Regression, Support Vector Machine (SVM), and Random Forest classifiers to identify the most effective predictive model. The performance of each model is evaluated using standard classification metrics to ensure reliability and clinical relevance.

In addition to algorithmic development, this work also emphasizes practical deployment by proposing a cloud-based architecture for real-time diagnostic support. The proposed system is designed to enable secure, scalable, and remote access to predictive services, making it suitable for deployment in rural healthcare centers and telemedicine platforms. By combining accurate prediction with practical usability, this study aims to bridge the gap between machine learning re-

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

search and real-world clinical application. The primary contributions of this paper are (i) Development of an ML-based framework for automated anemia detection and severity classification using four key hematological parameters. (ii) Construction of an integrated dataset from multiple clinical sources to enhance model robustness and generalization. (iii) Comparative evaluation of multiple machine learning models to identify the most effective approach for anemia prediction. (iv) Proposal of a cloud-based deployment architecture to sup- port real-time, remote clinical decision-making. Through these contributions, this research seeks to improve the efficiency, accuracy, and accessibility of anemia diagnosis, ultimately supporting better healthcare outcomes and promoting equitable access to diagnostic services across diverse populations.

## 2. Literature Survey

Anemia detection and classification have traditionally relied on manual clinical interpretation of hematological test results, which can be time-consuming and prone to variability de- pending on physician expertise and clinical workload. With the rapid growth of digital health records and computational capabilities, several researchers have explored the use of machine learning (ML) and data-driven techniques to improve the efficiency, consistency, and accuracy of anemia diagnosis. This section reviews the most relevant existing work in this domain, focusing on clinical approaches, classical statistical models, and recent machine learning-based systems.

### **A.** Traditional Clinical and Statistical Approaches

Early studies on anemia primarily focused on rule-based clinical thresholds defined by organizations such as the World Health Organization (WHO), where hemoglobin concentration is used as the primary indicator for diagnosis. While such thresholds are simple and interpretable, they do not always account for variations across age groups, gender, ethnicity, or comorbid conditions. Furthermore, reliance on single parameters often fails to capture complex patterns in blood indices that may indicate early or borderline anemia. Several researchers proposed statistical regression models to improve diagnostic reliability by incorporating multiple hematological parameters such as RBC count, MCV, MCH, and hematocrit. Linear regression and logistic regression were among the first tools applied for this purpose. These models improved diagnostic consistency compared to manual meth- ods; however, their performance was limited by assumptions of linearity and independence among features, which do not always hold true in biological systems.

### **B.** Machine Learning for Anemia Detection

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

***ISSN: 2583-5955***

With the advancement of machine learning, more sophisticated approaches have been explored. Support Vector Ma- chines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Artificial Neural Networks (ANNs) have been used to model nonlinear relationships between hematological indicators and anemia status. Some studies reported that SVM performs well for binary classification (anemic vs non-anemic), especially in small to medium-sized datasets due to its strong generalization capability. However, SVM models are often sensitive to kernel selection and parameter tuning, which can make them difficult to optimize for real-world clinical deployment.

Neural network-based models have also been proposed, particularly multilayer perceptrons and deep learning architectures. These models are capable of capturing complex nonlinear relationships and interactions among blood parameters. While they often achieve high accuracy, they require large labeled datasets and suffer from reduced interpretability, which can be a critical limitation in medical applications where explainability is important for clinical trust and regulatory approval.

### **C.** Ensemble Learning and Random Forest Approaches

Ensemble learning methods such as Random Forest and Gradient Boosting have gained increasing attention in medical diagnostics due to their robustness, ability to handle feature interactions, and resistance to overfitting. Several studies demonstrated that Random Forest models outperform individual classifiers in anemia detection by aggregating multiple decision trees trained on different data subsets.

Random Forest models also provide feature importance measures, enabling clinicians and researchers to understand which hematological parameters contribute most significantly to predictions. This interpretability advantage makes ensemble models particularly suitable for clinical decision support systems, as they balance performance with transparency.

### **D.** Severity Classification and Multiclass Diagnosis

Most earlier research focused primarily on binary classification, identifying whether a patient is anemic or not. How- ever, recent studies have shifted toward multiclass classification to distinguish between mild, moderate, and severe anemia. This is clinically significant because treatment strategies vary significantly depending on severity. Multiclass classification remains more challenging due to class imbalance and overlapping feature distributions. Re- searchers have experimented with oversampling techniques, cost-sensitive learning, and hierarchical classifiers to improve severity prediction accuracy. Despite these efforts, consistent and high-accuracy

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

multiclass anemia classification remains an open research problem, particularly in heterogeneous real- world datasets.

**E.** Cloud-Based and Real-Time Clinical Decision Support

In parallel with advances in predictive modeling, several works have proposed cloud-based healthcare architectures for remote diagnostics and telemedicine. These systems aim to bridge the healthcare accessibility gap, especially in rural and underserved regions. However, many existing systems are either focused on general disease prediction or lack real- time processing capabilities tailored specifically for anemia screening.

Few studies integrate machine learning-based anemia detection with scalable cloud infrastructure for real-time deployment in primary healthcare settings. This gap limits the practical impact of existing models, as many remain confined to experimental or academic environments.

**F.** Research Gap and Motivation

Although significant progress has been made, several limitations remain in the current literature:

- Many studies rely on small or single-source datasets, limiting generalizability.
- Binary classification dominates the literature, while severity-level classification is underexplored.
- Deep learning approaches, while accurate, lack transparency and are difficult to deploy in resource-constrained environments.
- Integration with cloud-based platforms for real-time clinical use is still limited.

Motivated by these gaps, this work proposes a comprehensive framework that combines multi-source clinical data, robust machine learning models, severity-level classification, and a cloud-enabled deployment architecture. By focusing on both predictive performance and practical deployment, the proposed system aims to move beyond theoretical accuracy and toward real-world clinical impact.

## 3. Proposed Methodology

This section describes the overall framework, data processing pipeline, model development, and system architecture used for automated anemia detection and severity classification. The proposed methodology is designed to ensure high predictive accuracy, robustness across heterogeneous clinical data, and practical feasibility for deployment in real-world healthcare environments.

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

### A. System Overview

The proposed system follows a modular pipeline architec- ture consisting of five major stages: data acquisition, data preprocessing, feature selection, model training and evaluation, and cloud-based deployment. The goal is to build an end-to- end system capable of ingesting raw hematological data and producing reliable anemia severity predictions in real time. The system accepts four core hematological parameters as input: Hemoglobin (Hb), Red Blood Cell (RBC) count, Mean Corpuscular Volume (MCV), and Mean Corpuscular Hemoglobin (MCH). These parameters were selected due to their clinical relevance, availability in routine blood tests, and proven diagnostic significance for anemia.

The overall workflow is as follows:

1) Data collection from multiple clinical repositories
2) Data cleaning and normalization
3) Label encoding for severity classes
4) Model training using ML algorithms
5) Model evaluation using standard metrics
6) Deployment via a cloud-based diagnostic interface

### B. Dataset Collection and Integration

To improve model generalizability, data was aggregated from five independent clinical repositories, each containing anonymized patient hematology records. These datasets var- ied in format, measurement ranges, and labeling standards, requiring careful integration and harmonization. The integration process involved standardizing units of measurement, removing duplicate records, resolving conflicting labels through majority agreement or clinical threshold mapping, and merging all datasets into a unified master dataset. This multi-source strategy reduces dataset bias and improves robustness against overfitting to a single population group.

### C. Data Preprocessing

Data preprocessing was performed to improve model performance and stability by handling missing values through median imputation or removing highly incomplete records, detecting outliers using interquartile range (IQR) analysis and clinically implausible ranges, and normalizing numeric features using Min–Max scaling. Additionally, anemia severity levels were classified

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

according to WHO thresholds into Normal, Mild, Moderate, and Severe anemia, with each patient record mapped to the appropriate category.

D. Feature Selection and Analysis

Although only four features were used, correlation analysis and feature importance evaluation were performed to verify their relevance. Pearson correlation coefficients and Random Forest feature importance scores were computed to analyze individual and combined feature contributions. Hemoglobin was observed to have the strongest correlation with anemia severity, followed by MCV and MCH, while RBC count contributed significantly in borderline and moderate cases.

E. Machine Learning Model Development

Three machine learning classifiers were implemented and evaluated in this study: Logistic Regression was used as a baseline linear classifier, Support Vector Machine (SVM) with an RBF kernel was applied to capture nonlinear relationships in the data, and Random Forest was employed as an ensemble-based classifier due to its robustness and interpretability. Hyperparameters were optimized using grid search and cross-validation techniques. The dataset was split into training and testing (80%-20%).

F. Model Evaluation Metrics

Model performance was evaluated using several metrics including accuracy, precision, recall, F1-score, and confusion matrix. These evaluation measures provide a comprehensive assessment of classification performance, particularly in multiclass scenarios, by analyzing both overall accuracy and the model's ability to correctly identify each class.

G. Cloud-Based Deployment Architecture

The trained model was integrated into a cloud-based architecture consisting of a RESTful API for handling prediction requests, a web-based clinician interface, and a cloud-hosted inference engine. This framework enables rural clinics to upload blood test values and receive instant anemia severity predictions efficiently.

H. Ethical Considerations and Data Privacy

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

All datasets were anonymized, and no personally identifiable information was used. The system is designed to comply with healthcare data protection standards and ensures secure data transmission and storage. The proposed methodology integrates robust data preprocessing, optimized machine learning models, and a scalable cloud infrastructure to provide an effective solution for automated anemia detection and severity classification. By emphasizing both technical rigor and practical usability, the system aims to bridge the gap between research and real-world healthcare deployment.

## 4. Results and Discussion

The This section presents performance evaluation of the proposed machine learning models, comparative analysis, of the obtained results. The goal is to assess the effectiveness, reliability, and practical relevance of the proposed system for anemia detection and severity classification.

### A. Experimental Setup

All experiments were conducted using Python-based ma- chine learning libraries including Scikitlearn, NumPy, and Pandas. Model training and evaluation were performed on a standard workstation environment with sufficient computational resources to support ensemble model training and cross-validation.

The integrated dataset was randomly split into training and testing subsets using an 80:20 ratios. Stratified sampling was applied to preserve class distribution across all severity levels. Five-fold cross-validation was used during training to reduce variance and improve model generalization. Hyperparameters for each model were optimized using grid search techniques to achieve optimal performance. For Logistic Regression, parameters such as regularization strength and solver selection were tuned; for Support Vector Machine (SVM), kernel type, C parameter, and gamma value were optimized; and for Random Forest, the number of trees, maximum depth, and minimum samples per leaf were adjusted.

### B. Performance Comparison of Models

The performance of all three models was evaluated using accuracy, precision, recall, and F1-score The Random Forest classifier significantly outperformed the other models across all evaluation metrics.

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

**Table 1: Performance comparison of proposed Models**

| Model | Accuracy | Precision | Recall |
|---|---|---|---|
| **Logistic Regression** | **89.4%** | **0.88** | **0.87** |
| **SVM (RBF Kernel)** | **93.1%** | **0.92** | **0.92** |
| **Random Forest** | **96.8%** | **0.95** | **0.95** |

C. Confusion Matrix and Class-Level Analysis

Analysis of the confusion matrices revealed that most misclassifications occurred between adjacent severity levels, such as mild vs moderate anemia, which is clinically understandable due to overlapping hematological ranges. The Random Forest model showed high sensitivity for detecting moderate and severe anemia, which is particularly important for clinical safety, as missing severe cases could have serious consequences. Random Forest feature importance analysis indicated that Hemoglobin (Hb) had the highest importance, followed by Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), and RBC count. These findings align well with established medical knowledge, thereby reinforcing the clinical validity and reliability of the proposed model.

The experimental results confirm that machine learning can significantly enhance anemia diagnosis accuracy and consistency. Compared to traditional threshold-based diagnosis, the proposed model captures subtle nonlinear interactions among hematological indices. The superior performance of Random Forest can be attributed to its ensemble structure, which reduces overfitting and handles noisy clinical data effectively. The cloud-based deployment further enhances the system's practical relevance by enabling real-time diagnostics in resource-limited environments. Despite demonstrating strong performance, the proposed system has certain limitations, including dependence on the quality and representativeness of the training data, the use of a limited feature set consisting of only four parameters, and the possibility of bias toward adult populations. These limitations highlight potential areas for future improvement and model enhancement.

## 5. Conclusion

The results demonstrate that the proposed Random Forest based system achieves high accuracy and clinical reliability, making it suitable for deployment as a decision-support tool in real-world healthcare settings. This research presented a comprehensive machine learning- based framework

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

*ISSN: 2583-5955*

for automated anemia detection and severity classification using four key hematological parameters: Hemoglobin (Hb), RBC count, Mean Corpuscular Volume (MCV), and Mean Corpuscular Hemoglobin (MCH). By integrating multiple clinical datasets and applying robust preprocessing techniques, the study aimed to overcome limitations associated with small, biased, or incomplete medical datasets. A comparative evaluation of Logistic Regression, Support Vector Machine (SVM), and Random Forest classifiers demonstrated that the optimized Random Forest model achieved superior performance, reaching an accuracy of 96.8%. Beyond predictive accuracy, this work emphasized practical deployment through a cloud-based architecture designed to support real-time diagnosis in rural and resource-constrained healthcare environments. This integration bridges the gap be- tween academic research and clinical application by enabling scalable, accessible, and efficient diagnostic support.

Overall, the proposed system contributes to the growing field of AI-assisted healthcare by offering a reliable, interpretable, and deployable solution for anemia screening and severity assessment. These future enhancements aim to transform the proposed system into a fully integrated intelligent clinical decision support tool.

## 6. Acknowledgement

The authors express their sincere gratitude to the healthcare institutions and open clinical repositories that provided anonymized hematological data for research, which was essential for the development and evaluation of the proposed machine learning framework. We also thank the faculty members, mentors, and colleagues for their valuable guidance, technical feedback, and encouragement throughout this research. Special appreciation is extended to the open-source community for offering powerful software libraries and tools that facilitated efficient model development, evaluation, and deployment. Finally, we acknowledge the continuous efforts of clinicians, healthcare professionals, and medical researchers in improving patient care, and this work aims to support their mission by contributing a technological solution to enhance diagnostic efficiency and accessibility, particularly in rural and underserved healthcare settings.

## 7. References

1. World Health Organization, “Haemoglobin concentrations for the diag- nosis of anaemia and assessment of severity,” WHO, Geneva, Switzer- land, Tech. Rep., 2011.
2. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
3. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

GIET UNIVERSITY JOURNAL, 2026, VOL. 6, NO. 1, PP. 49-59
Manuscript received 22nd November 2025; revised 18th January 2026;
Accepted 31st January 2026

***ISSN: 2583-5955***

4. S. Shrestha and R. M. Dhakal, "Machine learning approaches for anemia detection: A review," IEEE Access, vol. 9, pp. 123456–123470, 2021.
5. A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.
6. J. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
7. S. Dua and C. Graff, "UCI machine learning repository," Univ. California, Irvine, CA, USA, 2017. [Online]. Available: https://archive.ics.uci.edu/ml
8. R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," Scientific Reports, vol. 6, pp. 26094, 2016.
9. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
10. A. Holzinger, "Explainable AI and machine learning: Interpretable models for biomedical applications," IEEE Access, vol. 6, pp. 7165– 7174, 2018.