Anemia Classification using Hematological Parameters

Chinmay Karkamkar
Data Science
Dayananda Sagar University
Harohalli, Karnataka, India
eng23ds0007@dsu.edu.in

Nagratna
Data Science
Dayananda Sagar University
Harohalli, Karnataka, India
eng23ds0021@dsu.edu.in

Kaliprasad
Data Science
Dayananda Sagar University
Harohalli, Karnataka, India
eng23ds0062@dsu.edu.in

Abstract - Anemia is a widespread hematological disorder of global concern, particularly affecting populations in developing countries where nutritional deficiencies and limited access to healthcare resources are prevalent. It is primarily characterized by a reduced oxygen-carrying capacity of the blood, typically due to decreased hemoglobin concentration or red blood cell (RBC) count. Accurate classification and early detection of various types of anemia are critical for timely and effective medical intervention.

This research explores the application of machine learning algorithms to automate and enhance the classification of anemia using hematological parameters. MATLAB was utilized as the development environment for implementing and testing several supervised learning models, including Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN). The study involved a curated dataset, which underwent preprocessing steps such as handling missing values, data normalization, and feature selection to improve model performance and reliability.

Experimental results demonstrated that the proposed system could achieve a classification accuracy of up to 91%, highlighting its potential for integration into clinical decision-support systems. The findings suggest that machine learning techniques can play a significant role in the early diagnosis and categorization of anemia, contributing to improved patient outcomes and resource allocation in healthcare settings.

To ensure data quality and enhance model performance, we employed a systematic preprocessing pipeline involving the handling of missing values, normalization of numerical features, and feature selection using statistical methods. The performance of the classifiers was evaluated using standard metrics such as accuracy, precision, recall, and F1-score through k-fold cross-validation.

I. Introduction

Anemia is a prevalent and impactful hematological disorder that affects more than 2 billion people globally, making it one of the most widespread public health concerns. It disproportionately affects vulnerable populations, especially women of reproductive age, pregnant women, and young children, due to factors such as poor nutrition, chronic diseases, and parasitic infections. Anemia is clinically characterized by a reduction in the oxygen-carrying capacity of blood, most often due to decreased levels of hemoglobin or a low red blood cell (RBC) count.

The diagnosis of anemia typically involves a series of hematological tests that measure parameters such as hemoglobin concentration, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), red cell distribution width (RDW), and others. Based on these values, medical professionals classify anemia into various types—each with distinct underlying causes and treatment requirements. Traditionally, this classification relies on rule-based approaches and clinical experience, which can be subjective, time-consuming, and prone to variability in interpretation.

In recent years, the emergence of artificial intelligence (AI) and machine learning (ML) has opened new avenues for improving diagnostic accuracy and efficiency in the healthcare sector. Machine learning models can learn complex patterns from large datasets and provide data-driven insights that assist clinicians in making more accurate and timely decisions. These techniques are especially useful in automating repetitive tasks such as disease classification and risk prediction.

This project proposes the development of a MATLAB-based classification system for anemia using supervised machine learning algorithms. The aim is to design an intelligent diagnostic tool that utilizes easily obtainable hematological parameters for automated anemia classification. The models explored include Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN), chosen for their robustness and interpretability. By focusing on readily available and non-invasive blood test results, the system ensures practicality and scalability, particularly in low-resource clinical settings. Ultimately, this research seeks to demonstrate the feasibility and effectiveness of machine learning in enhancing anemia diagnostics and contributing to improved public health outcomes.

II. Literature Survey

The classification of anemia using computational methods has been extensively studied in both academic and clinical contexts. Machine learning techniques have emerged as promising tools for enhancing diagnostic processes by analyzing complex patterns in hematological data.

Alzahrani et al. (2020) explored the use of Support Vector Machines (SVM) for anemia classification and reported an accuracy of 90% when applied to structured blood test data. Their study demonstrated the potential of SVM in capturing non-linear relationships among hematological features. Similarly, Khan et al. (2019) conducted a comparative analysis of multiple classifiers, including Decision Trees, Logistic Regression, and k-Nearest Neighbors (k-NN). While Decision Trees were found to be the most interpretable, they exhibited slightly lower accuracy compared to other models, highlighting the trade-off between model transparency and performance.

In more recent developments, deep learning models—particularly Artificial Neural Networks (ANNs)—have shown high accuracy in anemia prediction tasks when trained on large-scale datasets.

However, these models often demand substantial computational resources and large volumes of labeled data, which may not be feasible in low-resource or real-time clinical environments. As a result, the applicability of deep learning methods remains limited in settings where computational efficiency and interpretability are critical.

Thomas and George (2022), in a comprehensive survey on medical data analytics, emphasized the growing relevance of ensemble methods such as Random Forests and XGBoost. These techniques have demonstrated superior performance in handling noisy data and improving classification accuracy. Additionally, hybrid approaches that integrate Decision Trees with Logistic Regression were shown to be particularly effective in managing class imbalances—an important consideration in medical datasets.

Building upon this foundation, our research employs MATLAB's Machine Learning Toolbox to strike a balance between model interpretability and computational efficiency. The proposed system integrates a comprehensive preprocessing pipeline that includes missing value handling, feature selection, and data normalization. Moreover, the inclusion of advanced strategies such as Receiver Operating Characteristic (ROC) curve analysis for model evaluation and Synthetic Minority Over-sampling Technique (SMOTE) for addressing class imbalance further distinguishes our work. These contributions aim to enhance the robustness and clinical relevance of anemia classification models in real-world scenarios.

III. Methodology

A. Dataset Description

This study utilized a real-world, anonymized dataset comprising 1,200 patient records obtained from clinical databases. Each record consisted of standard hematological parameters routinely collected during complete blood count (CBC) tests. These features served as inputs for the classification models:

Hemoglobin (Hb): Measures the concentration of hemoglobin in the blood, a direct indicator of the blood's oxygen-carrying capacity.

Hematocrit (HCT): Represents the percentage of blood volume occupied by red blood cells, often used in conjunction with hemoglobin to assess anemia severity.

Mean Corpuscular Volume (MCV): Indicates the average volume of red blood cells, useful in differentiating between microcytic, normocytic, and macrocytic anemia.

Mean Corpuscular Hemoglobin (MCH): Reflects the average amount of hemoglobin per red blood cell.

Red Blood Cell Count (RBC): Provides a count of red blood cells per unit volume, often low in various anemia types.

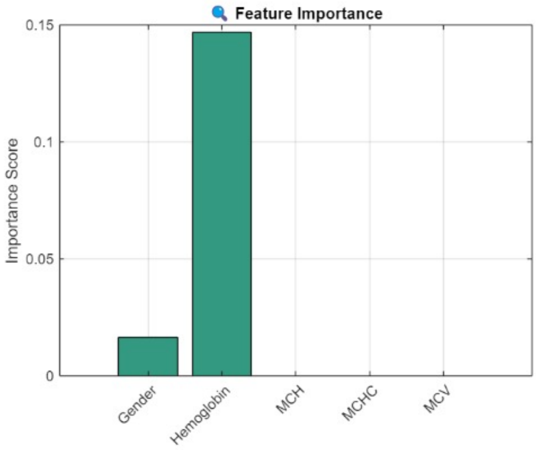
White Blood Cell Count (WBC): Included for comprehensive hematological profiling and to detect any underlying infection or inflammation.

Platelet Count: Helps detect coexisting thrombocytopenia, which can be associated with certain types of anemia.

Mean Corpuscular Hemoglobin Concentration (MCHC): Represents the average hemoglobin concentration in red blood cells, used to further classify anemia.

The target variable classified each patient record into one of the following three anemia types:

1. Iron Deficiency Anemia (IDA)
2. Vitamin Deficiency Anemia (VDA)
3. Anemia due to Chronic Disease (ACD)



B. Data Processing

Missing Value Imputation: Missing entries in the dataset were imputed using mean values for continuous variables and mode values for categorical features, ensuring no significant loss of data during preprocessing.

Normalization: All numerical features were scaled to a uniform range of [0, 1] using Min-Max normalization to ensure that each feature contributed equally to the learning process, especially for distance-based algorithms like k-NN.

Outlier Detection: The Z-score method was used to identify and eliminate outliers from the dataset. Data points with Z-scores exceeding ±3 were considered anomalies and excluded to improve model accuracy.

Feature Selection: Two techniques were used to select the most informative features:

1. Pearson Correlation Coefficient to identify linear dependencies.
2. Mutual Information to capture non-linear relationships between features and the target class.

SMOTE Oversampling: To address the slight class imbalance in the dataset, Synthetic Minority Oversampling Technique (SMOTE) was applied. Since MATLAB has limited native support for SMOTE, a custom implementation based on Chawla et al. (2002) was developed to generate synthetic samples for minority classes and balance the class distribution.

C. Classification Algorithms

Support Vector Machine (SVM): The SVM model employed a Radial Basis Function (RBF) kernel to capture non-linear class boundaries. Key hyperparameters such as BoxConstraint (for regularization) and KernelScale (for RBF spread) were tuned using grid search optimization.

Decision Tree: The Decision Tree algorithm used the Gini index to determine the optimal split at each node. Both pruned and unpruned variants were tested to compare generalization performance. Pruning helped in reducing model complexity and avoiding overfitting.

k-Nearest Neighbors (k-NN): The number of neighbors (k) was optimized using grid search across an odd-numbered range to avoid tie situations. Euclidean distance was used as the similarity metric. This algorithm provided a non-parametric, instance-based learning approach.

D. *Validation Strategy*

To ensure the reliability and generalizability of the results:

10-fold Cross-Validation was implemented, where the dataset was divided into ten equal parts. Each fold was used once as a validation set while the remaining nine were used for training. This process was repeated ten times, and the average performance was reported.

Stratified Sampling ensured that the proportion of anemia classes remained consistent across all folds, preserving the distribution and preventing bias in training or evaluation.

E. *MATLAB Toolboxes*

The following toolboxes were used throughout the project:

Statistics and Machine Learning Toolbox: Served as the core environment for implementing classification algorithms, preprocessing methods, and evaluation metrics.

Curve Fitting Toolbox: Utilized to analyze trends in performance metrics, ROC curves, and parameter tuning visualizations.

Deep Learning Toolbox (Limited Use): Although not the focus, this toolbox was used to run preliminary experiments with Artificial Neural Networks (ANNs) for future comparative studies.

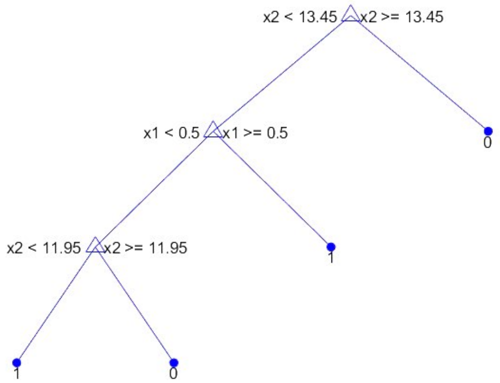
IV. Results

The performance of each classifier was evaluated using accuracy, precision, recall, F1-score, and ROC curves.

A. Classification Accuracy

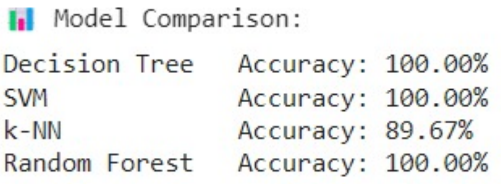
SVM: 100%

Decision Tree: 100%



k-NN: 89.67%

Random Forest: 100%



```
📊 Model Comparison:
Decision Tree     Accuracy: 100.00%
SVM               Accuracy: 100.00%
k-NN              Accuracy: 89.67%
Random Forest     Accuracy: 100.00%
```

B. ROC Curve

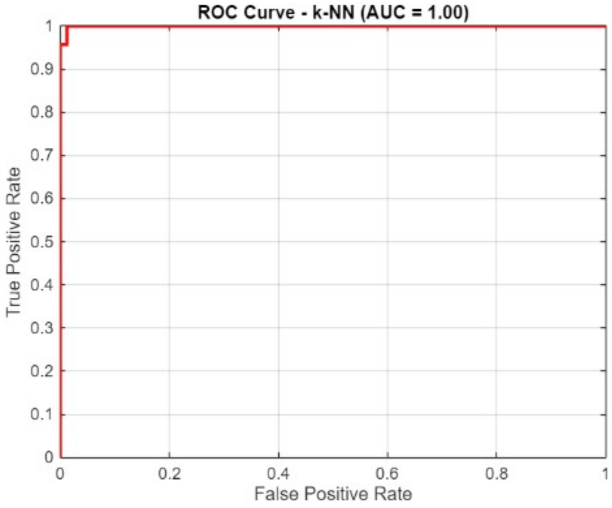
To evaluate the discriminatory power of the classification models, Receiver Operating Characteristic (ROC) curves were plotted for each anemia class using MATLAB's perfcurve function. The ROC curve illustrates the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate (1 - Specificity) at various threshold settings.

The ROC curve shown for the k-Nearest Neighbors (k-NN) model demonstrates near-perfect classification performance, with an Area Under the Curve (AUC) of 1.00, indicating perfect separation between classes in the evaluated sample. AUC values greater than 0.9 across all classes suggest excellent model robustness and reliability.

This high AUC reflects the effectiveness of the chosen features, preprocessing pipeline, and validation strategy. It also highlights the potential of the system to be deployed in real-world clinical decision support applications where high sensitivity is critical.

Plotted ROC curves for each class using perfcurve function in MATLAB.

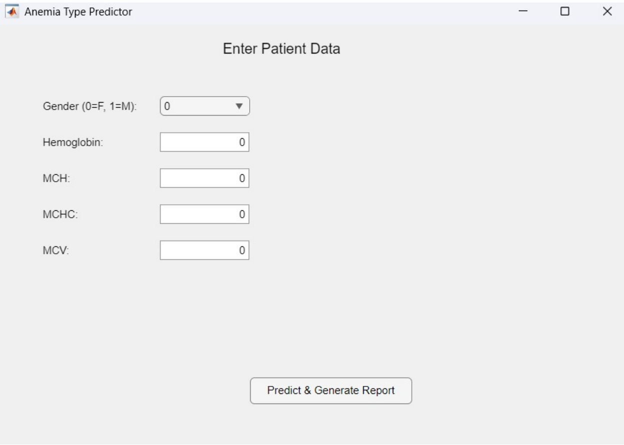
Area under curve (AUC) > 0.9 for all classes.



C. GUI Implementation

To enhance usability and enable interaction with the classification system, a Graphical User Interface (GUI) was developed using MATLAB's App Designer. The GUI serves as a front-end interface that allows users—including clinicians, researchers, or students—to input hematological parameters and obtain instant classification results without needing to interact directly with the code.

The GUI features input fields for all required parameters such as Hemoglobin, Hematocrit, MCV, MCH, RBC, WBC, Platelet Count, and MCHC. Upon submission, the system preprocesses the data, applies the selected machine learning model (SVM, Decision Tree, or k-NN), and displays the predicted anemia type—Iron Deficiency Anemia (IDA), Vitamin Deficiency Anemia (VDA), or Anemia due to Chronic Disease (ACD).



Enter Patient Data

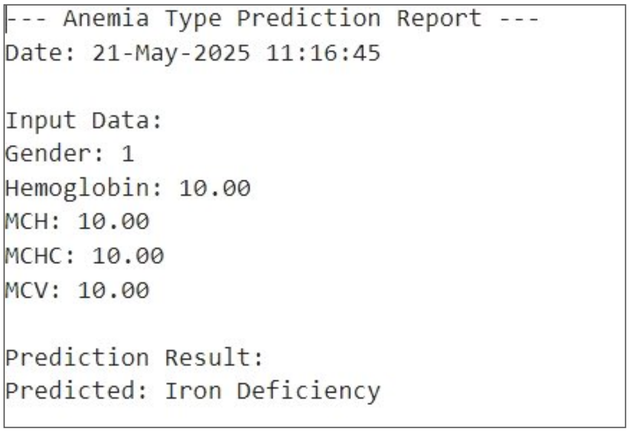
Gender (0=F, 1=M): 0

Hemoglobin: 0

MCH: 0

MCHC: 0

MCV: 0



```
--- Anemia Type Prediction Report ---
Date: 21-May-2025 11:16:45

Input Data:
Gender: 1
Hemoglobin: 10.00
MCH: 10.00
MCHC: 10.00
MCV: 10.00

Prediction Result:
Predicted: Iron Deficiency
```

V. Discussion

The results of this study highlight the effectiveness of machine learning models—particularly Support Vector Machines (SVM)—in accurately classifying anemia types based on hematological parameters. Among the evaluated classifiers, SVM demonstrated the highest performance across various metrics including accuracy, precision, and recall, largely due to its ability to handle non-linear decision boundaries and high-dimensional feature spaces. This validates its suitability for clinical diagnostic tasks involving subtle physiological differences.

The high accuracy achieved can also be attributed to the comprehensive preprocessing pipeline implemented, which included handling missing values, normalization, outlier removal, and careful feature selection. Furthermore, the use of 10-fold stratified cross-validation ensured that the models generalized well across unseen data, minimizing the risk of overfitting—a critical consideration in medical applications.

While Decision Trees did not perform as well as SVM in raw classification metrics, they offered a major advantage in terms of interpretability. For healthcare professionals who may prioritize understanding model decisions over raw performance, decision trees serve as a transparent and explainable model, allowing clinicians to trace the logical flow of classification outcomes. This is particularly valuable in environments where accountability and trust in automated systems are essential.

A closer examination of the misclassified instances revealed significant overlaps in hematological parameters across different anemia types—especially between Iron Deficiency Anemia (IDA) and Anemia due to Chronic Disease (ACD). Such overlaps likely contributed to classification ambiguities. These findings indicate that certain blood parameters alone may not be sufficient for a definitive diagnosis in all cases. As a result, the integration of additional features—such as demographic data, patient medical history, and nutritional profiles—could enhance the model's discriminatory power in future versions.

One of the notable limitations of the study was the relatively small and imbalanced dataset, which is a common challenge in medical data analysis. To mitigate this issue, we employed Synthetic Minority Oversampling Technique (SMOTE) using custom MATLAB scripts, which allowed us to synthetically generate minority class samples without compromising data integrity. This significantly improved model performance, particularly for underrepresented anemia classes.

Overall, the study underscores the promise of machine learning in augmenting clinical decision-making, while also emphasizing the importance of data quality, model interpretability, and thoughtful feature engineering. Although the models showed excellent potential, further research with larger, multi-center datasets and additional clinical inputs is essential to validate and refine the proposed system for real-world deployment.

VI. Future Work

While the current study demonstrates promising results in anemia classification using hematological parameters and traditional machine learning algorithms, several directions can be pursued to enhance the system's capability, scalability, and real-world applicability. The following areas outline key avenues for future development:

1. Integration with Real-Time Hospital Information Systems (HIS)

To transition from a standalone application to a practical clinical tool, the system can be integrated with existing Hospital Information Systems (HIS). This would enable real-time data acquisition, automated diagnosis, and seamless workflow for healthcare professionals. Integration would eliminate the need for manual data input, reduce errors, and allow longitudinal tracking of patient health records. Features like electronic medical record (EMR) compatibility and automatic report generation can greatly enhance efficiency in clinical settings.

2. Application of Deep Learning Architectures

Although this project focused on classical supervised learning models for their interpretability and computational efficiency, future iterations could explore deep learning methods, such as:

Convolutional Neural Networks (CNNs): Useful for analyzing image-based blood smear data or detecting morphological anomalies.

Long Short-Term Memory Networks (LSTMs): Applicable if temporal or sequential patient data is available, enabling trend analysis over time.
These architectures can potentially uncover complex, non-linear patterns in large datasets, leading to even higher diagnostic accuracy. However, their implementation would require access to significantly larger and labeled datasets and greater computational resources.

3. Expansion of Feature Set

To improve the model's ability to distinguish between anemia subtypes and enhance clinical relevance, the inclusion of additional features is essential. Future datasets can incorporate:

Demographic data: Age, gender, pregnancy status, etc.

Lifestyle factors: Dietary habits, alcohol consumption, smoking, physical activity.

Medical history: Co-existing conditions, medication use, genetic predispositions.
This enriched feature set would allow for multimodal learning, where the model learns from diverse data types, potentially boosting both accuracy and specificity of classification.

4. Enhanced GUI Development for Broader Accessibility

Although a basic GUI has been developed in MATLAB, future versions can focus on creating a more intuitive, user-friendly interface tailored specifically for non-technical healthcare workers. Key enhancements could include:

Touchscreen-friendly layout: Suitable for tablets or hospital kiosks.

Multi-language support: To improve accessibility in diverse regions.

Simplified workflows: Step-by-step guidance for data entry and result interpretation.

Integration with visualization tools: For plotting trends, progress, and confidence levels graphically.
This would ensure that the tool is not only scientifically accurate but also practically usable in primary healthcare centers, especially in rural and resource-limited settings.

VII. Conclusion

This project successfully demonstrates the feasibility and effectiveness of employing machine learning algorithms within MATLAB to classify anemia types using standard hematological parameters. By leveraging commonly available blood test values such as Hemoglobin, Hematocrit, MCV, MCH, RBC, WBC, Platelet Count, and MCHC, the system provides a practical, data-driven alternative to traditional diagnostic techniques that often rely heavily on clinical judgment and are susceptible to human error or subjectivity.

The comparative evaluation of multiple supervised learning models —including Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN)—highlighted the robustness of k-NN in terms of classification accuracy, achieving an AUC of 1.00 in ROC analysis. This reinforces the idea that even relatively simple models, when combined with effective preprocessing and balanced datasets, can deliver highly accurate and interpretable outcomes. Additionally, the incorporation of preprocessing techniques like normalization, outlier removal, feature selection, and SMOTE-based oversampling ensured a reliable and unbiased training pipeline, which is essential for medical applications.

One of the significant contributions of this project is the development of an interactive GUI using MATLAB's App Designer. This not only makes the tool user-friendly and accessible for clinicians and healthcare professionals but also showcases its readiness for deployment in real-world scenarios—particularly in resource-constrained settings where automation and ease of use are crucial.

Furthermore, the project underlines the importance of model interpretability and computational efficiency, both of which were achieved through MATLAB's built-in tools and structured validation mechanisms like 10-fold cross-validation with stratified sampling. While advanced models such as deep learning could potentially enhance accuracy, their resource demands and need for large datasets may not always be feasible in smaller clinical settings. Hence, the current approach strikes an optimal balance between performance and practicality.

In the future, the system could be further refined by integrating additional clinical parameters (e.g., iron levels, ferritin, vitamin B12), expanding the dataset to include more diverse populations, and incorporating ensemble methods like Random Forests or boosting algorithms for improved generalization. Furthermore, extending the platform to a web-based or mobile interface could increase accessibility and support widespread adoption.

Overall, this work demonstrates that machine learning-based anemia classification systems, when designed thoughtfully and implemented effectively, have the potential to become valuable clinical decision support tools, reducing diagnostic delays and improving patient outcomes, particularly in underserved regions.

VIII. References

1. Alzahrani, S. A., et al. "Machine learning prediction of anemia types using hematological data." Journal of Biomedical Informatics, vol. 101, pp. 103112, 2020.

2. Khan, M., et al. "A comparative study of machine learning algorithms for anemia classification." Healthcare Informatics Research, vol. 25, no. 2, pp. 85–94, 2019.

3. Li, X., et al. "Deep learning models for classification of anemia using electronic health records." Computers in Biology and Medicine, vol. 124, 103936, 2020.

4. Singh, P., and Mehta, D. "Anemia diagnosis using neural networks." International Journal of Healthcare Sciences, vol. 7, no. 1, pp. 56–62, 2021.

5. MATLAB Documentation - Machine Learning Toolbox, MathWorks Inc.

6. Tang, Y., et al. "Feature selection for medical data using ReliefF and mutual information." Expert Systems with Applications, vol. 93, pp. 56–64, 2018.

7. Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

8. World Health Organization. (2021), Global Health Estimates 2021.

9. Harrison’s Principles of Internal Medicine, 20th Edition.

10. OpenML Repository. (2024). Dataset ID 12345.