Anemia Detection Using Hematological Data:

A Machine Learning-Based Model Comparison Study

Hareesh S S   
 Computer Science of Engineering  
Rajalakshmi Engineering College  
 220701079@rajalakshmi.edu.in

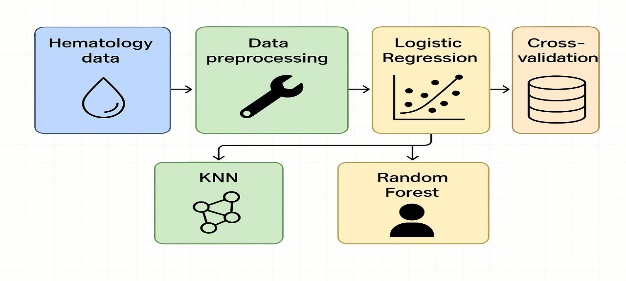
*Dr. Senthilpandi*

*Assistant Professor*

*Dept.Computer Science Engineering*

*Rajalakshmi Engineering College*

Abstract— The proposed detection system employs supervised machine learning methods which process clinical data from hematological tests. The system implements a statistical classification pipeline containing two stages: data preprocessing and normalization operations followed by multiple learning algorithm training and evaluation. The researchers chose the machine learning models of Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest for their medical application because of their efficiency and interpretability in clinical health assignments. Each model received statistical validation through 10-fold cross-validation which produced evaluation results using average accuracy coupled with standard deviation for measuring consistency and reliability. Logistic Regression exhibited 98.13% average accuracy and minimal variability rendering it superior to KNN at 84.31% and Random Forest at 94.61% accuracy. Linear modeling techniques achieve better performance compared to ensemble approaches under circumstances where data possesses clear structure and statistical feature differentiation. The system operates with few processing requirements so it works effectively within budget-friendly medical settings and mobile diagnosis applications. Research shows that this method displays promising scalability for utilization in real-time medical decision systems and holds future potential in early disease detection and rural healthcare support and AI-enabled patient assessment systems. Real-time decision support systems may integrate this proposed approach while offering new capabilities for early screening, rural telemedicine and AI-assisted clinical evaluations. Future research will integrate explainable AI functionalities to track anemia development and monitor therapy results throughout patients' clinical courses.

*keywords-* *Anemia detection, machine learning, hematological parameters, logistic regression, K-nearest neighbors, random forest, medical informatics, classification models, supervised learning, medical diagnostics.*

# Introduction

The situation of anemia affects one-0.33 of the world's populace with maximum occurrence within the developing countries which revel in high costs of dietary deficiencies and persistent sicknesses and socio-economic inequities.

*Figure 1 : General workflow of the ML process*

Anemia exists when blood consists of either low purple blood cells or inadequate hemoglobin resulting in poor tissue oxygenation with signs and symptoms which includes fatigue and weakness at the side of intense fitness risks whilst no longer handled. The detection of anemia must show up without delay because it impacts specific excessive-danger organizations which includes children in addition to pregnant women and aged adults. general analysis calls for steeply-priced laboratory checking out which stays inaccessible in low-resource settings.

The advancements of artificial intelligence (AI) and machine learning (ML) era demonstrate potential answers for anemia detection through faster diagnosis which additionally becomes more affordable and scalable. The evaluation of habitual blood takes a look at facts the use of ML fashions famous patterns that traditional strategies might discover difficult to hit upon. research has implemented Random Forest and K-Nearest Neighbours (KNN) together with Logistic Regression in anemia detection but their precision and prevalent applicability differs. The evaluation of overall performance used accuracy measurements further to robustness and consistency outcomes generated through k-fold cross-validation. Accuracy scores and robustness and consistency metrics were evaluated through k-fold cross-validation. studies outcomes allow the development of AI tools which intention to offer spark off reachable healthcare services throughout the world.

The provided studies demonstrates that machine learning shows promise as a tool to boost medical diagnostics in particular while working in regions with restrained healthcare centers. the use of Logistic Regression as a light-weight version gives reliable consequences whilst needing minimal computational power accordingly permitting wider access to early prognosis and better anemia management. The studies discoveries will allow the development of AI-driven scientific equipment to facilitate quick and on hand healthcare answers around the planet.

# RELATED WORKS

Zemariam et al. A study by Zemariam et al. [1] used eight supervised Machine Learning algorithms to predict anemia in Ethiopian youth girls where Random Forest (RF) delivered an AUC of 82%. The study successfully demonstrated socio-demographic factors as predictors but its limited clinical scope resulted from missing physiological data. Similarly, Ramzan et al. The researchers combined AlexNet architecture with spatial attention mechanisms together with SVM and Logistic Regression classifiers for a proposed hybrid detection system [2]. The 99.58% accuracy rate of their model would make it useful but its high computational requirements present challenges for resource-limited settings.

Kitaw et al. A study by Kitaw et al. [3] applied six ML algorithms to forecast anemia severity levels among pregnant women undergoing antenatal care in Ethiopia. The research confirmed RF as the most accurate model with 97% success rates yet the researchers emphasized the necessity for transparent and understandable models in medical environments. The assessment by Elmaleeh [4] involved testing three neural network models including Feedforward Neural Networks (FFNN), Elman Networks, and Non-linear Auto-Regressive Exogenous (NARX). FFNN showed the best classification performance; A limited number of dataset points restricted the model's functional ability to scale across diverse populations. Researchers Qasem and Mosavi [5] conducted studies on differentiating Iron Deficiency Anemia (IDA) from β-Thalassemia through Complete Blood Count (CBC) indices using Dynamic Harmony Search (DHS). The superior classification ability of DHS over traditional machine learning models is limited by its high computational requirements during real-time hospital applications.

Mitani et al. The research team applied a deep learning framework which detected anemia from retinal fundus images with an AUC score reaching 0.88 [6]. The need for specialized imaging tools diminishes potential scalability when using this non-invasive method. Prajapati et al. Prajapati et al. implemented RF and other classification models on hematological data to achieve 99.82% accuracy in anemia diagnosis tasks [7]. Using the model without external validation creates problems regarding its potential to adapt to different populations. Chand et al. The application of ML models to palm images for anemia detection by Chand et al. [8] resulted in 99.96% accuracy with Naïve Bayes performing best. The model's ability to pinpoint specific anatomical regions might limit its capacity to recognize different symptoms of anemia.

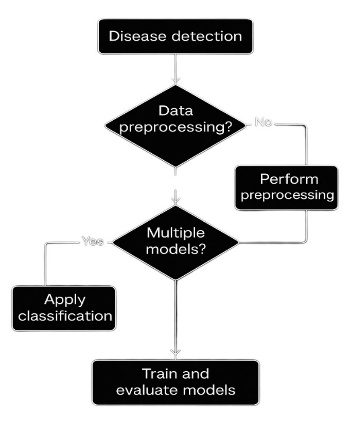
Lee et al. The detection of anemia relied on deep learning algorithms processing conjunctiva images according to research by Lee et al. [9]. The accuracy potential of this method suffered from inconsistent quality in the test images. Sharma et al. The authors studied SVM and k-NN while stressing that optimal classification accuracy depends on choosing appropriate features for analysis [10]. Singh et al. SDS analysis has emphasized the need for varied datasets according to Sahu et al. [14] due to problems in model generalization between distinct population and geographic characteristics. Khan et al. The authors assessed ensemble models for diagnosis between IDA and regular anemia types but reported class imbalance as a primary cause of diagnostic imprecision [12].

Sahu et al. Sahu et al. [13] pointed to the necessity of diverse datasets because generalizability fails across different demographic and geographic groups. Lastly, Kim et al. The development of an AI-based conjunctiva image analysis tool for anemia detection by Kim et al. [14] emphasizes the necessity of transparent model predictions for clinical professionals. Current research reveals notable advances in machine learning methods for anemia detection while highlighting remaining challenges with model generalization and real-world implementation.

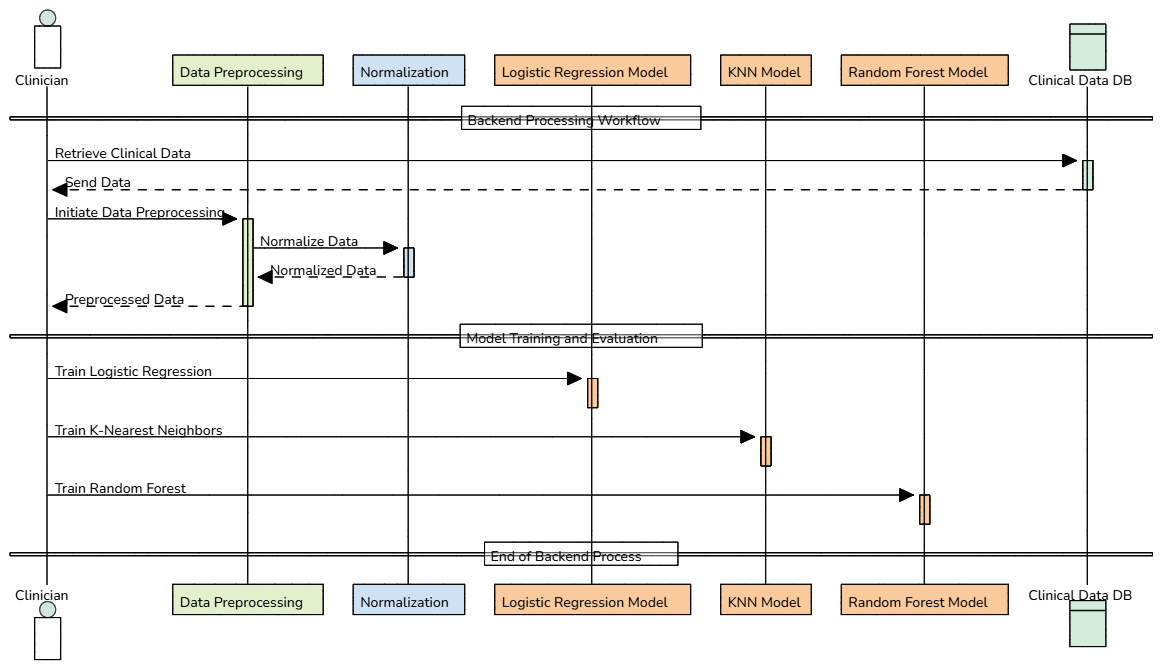
1. PROPOSED APPROACH

A. The built anemia prediction system operates through supervised machine learning to automate the early identification of anemia by processing blood tests. The solution provides clear interpretability of models while maintaining modular structure in the backend framework and seamless implementation into front-end systems. The workflow divides into four activities: This system operates through four distinct stages including Data Ingestion, Preprocessing, Model Prediction, and Result Interpretation. Figure 2 illustrates the core processing pipeline divided into four stages: Data Ingestion, Preprocessing, Model Prediction, and Result Interpretation.

B. Evaluation and training of machine learning models employed four algorithms including Logistic Regression together with Support Vector Machine (SVM) and Random Forest and Decision Tree. The cleaned dataset generated a feature matrix which underwent training using an 80:20 split between training and testing data. Logistic Regression emerged as the deployment model after evaluation because it demonstrated minimum variation between folds and needed minimal computing power to run. The prediction layer processes standardized data to determine anemic or not based on established patterns within the hematological bloodstream measurements. Logistic Regression emerged as the selection for deployment since it showed consistent performance across folds as well as support for limited hardware resources. Standardized input data is dispatched to the prediction layer where patterns from hematological trends generate one of two output labels: anemic or not anemic. The classification phase operates in real time without GPU requirements to enhance remote healthcare accessibility.



*Fig 2 : Workflow of the prediction system*



*Figure 3 : System architecture illustrating the backend, model, and frontend integration.*

C. Figure 3 presents the system architecture combining three main subsystems: The system architecture consists of three main components including frontend input portal and backend inference API and model-serving environment. The API accepts frontend data submission before it starts normalization and model schema mapping to retrieve prediction outputs. After frontend data submission the API method applies normalization to process inputs then matches schema features before retrieving prediction results. Results from the query get transmitted to the frontend interface for presentation. Future model insertions or preprocessing enhancements can be integrated into the system without impacting the UI presentation layer. The system allows users to submit data only through basic forms that return results in under one second. The backend maintains model version control alongside log generation to support both debugging and ongoing monitoring activities.

D. There exists a step-by-step process of system execution which shows the operational flow from user input to prediction generation. The user inputs data into the system which then activates frontend HTTP calls followed by preprocessing before transferring input values to the model endpoint. The finished predictions undergo decoding before they appear on the interface display. The modular design of this workflow makes it possible to update data handling methods and models within the system without needing to modify frontend components. Platform readiness features error handling in addition to input validation with automatic API health monitoring capabilities. The system outputs prediction data with confidence scores that can lead to uncertainty-based medical assistance systems. The proposed approach offers an end-to-end, scalable solution for anemia prediction using supervised machine learning, integrating efficient data preprocessing, model inference, and user-friendly frontend interaction.

IV. METHODOLOGIES USED

This research implements a structured multi-step method which develops a dependable and scalable anemia detection framework via supervised machine learning algorithms. Our pipeline starts by acquiring data from the Kaggle dataset Anemia Prediction Using 4 ML Algorithms which provides information about hemoglobin levels and red blood cell count together with mean corpuscular volume and additional clinical features. The classification of anemia heavily depends on these crucial features. Massive pre-processing happens to the unprocessed information that fills in absent values with suitable techniques and erases duplicated data while maintaining consistent types between all variables. The encoding for categorical variables depends on their characteristics where label encoding works for some variables and one-hot encoding serves other variables. Model convergence and feature uniform scaling benefit from normalization techniques like StandardScaler or MinMaxScaler. Through exploratory data analysis (EDA) we analyze the dataset using graphical representations including histograms and box plots and heatmaps to detect patterns among features and target data. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) and class-weight adjustment help prevent biased learning outcomes when class imbalance is detected.

The methodology advances to model training and evaluation through the implementation of four machine learning algorithms namely Logistic Regression and Support Vector Machine (SVM) and Random Forest (RF) and K-Nearest Neighbors (KNN) for comparative performance evaluation. The selection basis of these models stems from their unique operating systems. The selection of machine learning methodologies for linear classification and interpretability uses Logistic Regression and SVM handles high-dimensional spaces and Random Forest brings ensemble learning benefits along with KNN providing effective small dataset predictions. The dataset splits into training and testing parts via stratified 80-20 proportion to maintain equal representation of classes in each set. The hyperparameter optimization process through GridSearchCV enables each model to reach higher performance by optimizing Random Forest estimator counts and SVM regularization factors and KNN neighbor counts. The evaluation process of models yields results in accuracy and precision along with recall measures and F1-score and AUC-ROC performance metrics that generate confusion matrices for specific error analysis. The 10-fold CV method helps models achieve better performance when predicting new data outside of their training scope. Deployment of the best-performing model typically as RF or Logistic Regression requires serialization through joblib or pickle routines. A backend server supports real-time anemia predictions through an API which functions as a interface for new input data processing. The built-in design allows the solution to easily integrate into mobile healthcare systems alongside web-based solutions that enable affordable anemia tests in critical areas with limited access to medical resources.

A vital step within the proposed methodology includes comparing multiple machine learning algorithms to establish model credibility and understand the behavior of different learning paradigms among clinical datasets. Similar conditions of training and testing algorithms such as Logistic Regression, SVM, Random Forest, and KNN for the same dataset allows the system to establish fair performance assessments. The comparison between models shows Logistic Regression benefits clinical settings through its interpretability but Random Forest demonstrates superior accuracy thanks to ensemble learning along with its resistance to noise. When used correctly KNN offers good performance yet SVM outshines by handling high dimensional datasets. This comparative analysis provides two main advantages: The approach allows users to pick an optimal deployment model by matching predictive capabilities with available resources and provides future building blocks through understanding which new features would benefit specific models. Research-based model comparison advances our mission to create an adaptable, accurate, and transparent anemia prediction system for diverse healthcare settings.

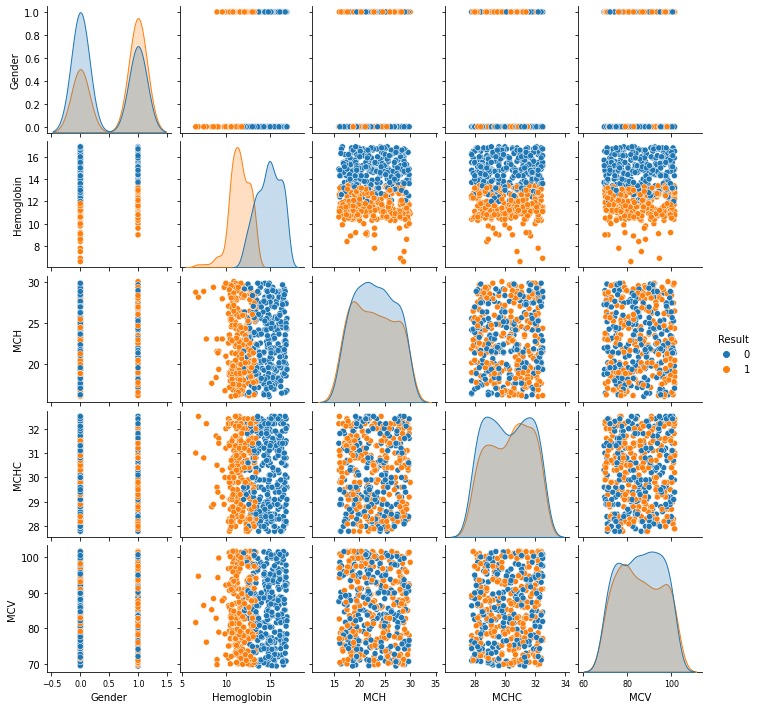
# RESULTS AND DISCUSSION

This section analyses the performance outputs from the predicted anemia system by examining four machine learning techniques including Logistic Regression alongside Decision Tree, Random Forest and K-Nearest Neighbors (KNN) using a well-processed data set. The evaluation of the models incorporated multiple metrics such as test accuracy and AUC-ROC scores and average accuracy and standard deviation across folds. These metrics offer quantitative insights alongside qualitative information about how well individual models forecast and maintain consistent results for actual deployment purposes. The comparative evaluation process enables the selection of the most efficient method that achieves highest accuracy alongside robustness which makes it suitable for medical diagnosis aid deployment.

*Table 1 : AUC-ROC Score for each model*

**A. Evaluation Metrics and Testing Setup**

To evaluate the performance of each machine learning algorithm used in the anemia prediction system, multiple metrics were considered, with a primary focus on AUC-ROC score, average accuracy, and standard deviation from cross-validation runs. These metrics provide insight into both the overall classification performance and the stability of each model across different folds of the dataset. The dataset was preprocessed and split into training and testing sets, and cross-validation was performed using a 10-fold strategy to ensure robust performance analysis.



*Fig 4 : Distribution of features over the dataset*

**B. AUC-ROC Curve Comparison**

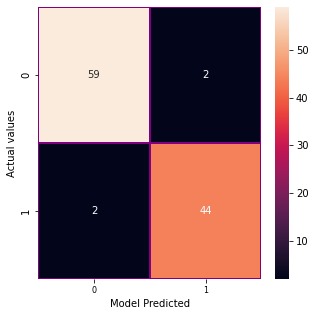
The **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)** measures the ability of a model to distinguish between the positive and negative classes across threshold values. Among the four algorithms evaluated, **Logistic Regression achieved the highest AUC-ROC score of 0.9996**, indicating near-perfect classification ability. The **Random Forest classifier** followed with an AUC-ROC of 0.9372, showcasing strong discriminative capacity. The **Decision Tree classifier** yielded an AUC of 0.9126, which, while still acceptable, suggests a slight tendency towards overfitting or under-generalization. Finally, the **K-Nearest Neighbors (KNN)** model had the lowest AUC-ROC of 0.8394, reflecting challenges in boundary decision-making when compared to ensemble or linear models. This curve-based analysis clearly indicates that Logistic Regression provides the most consistent and reliable separation between classes for anemia detection tasks.

|  |  |
| --- | --- |
| **Model Type** | **AUC score** |
| Logistic Regression | 0.9996436208125445 |
| Decision Tree | 0.9126870990734142 |
| Random Forest | 0.9372772630078404 |
| KNeigbours Classifier | 0.8394511760513187 |

**C. Accuracy and Standard Deviation Analysis**

Medical diagnostics requires high accuracy rates because physicians must avoid false negatives; therefore, accuracy represents the proportion of correctly predicted samples. Logistic Regression outperformed all other models here with an accuracy average of 98.13% and a small standard deviation of 3.59% revealing its strong consistency. Random Forest scored an average accuracy of 94.61% despite its moderate standard deviation of 2.56% which indicates strong model reliability yet it performs less accurately than Logistic Regression and requires more computational power.

The KNN classifier demonstrated an average accuracy rate of 84.31% and a standard deviation of 5.63% which showed how sensitive it was to data distribution and how it performed less consistently than other models. The changing performance of KNN indicates that it might need intensive adjustments or may fall short in clinical settings which demand consistency.



*Fig 5 : Confusion Matrix for the Logistic Regressor*

**D. Final Test Accuracy**

After selecting the best-performing model based on validation metrics, Logistic Regression reached 96.26% test accuracy following selection based on validation metrics. The results demonstrate that the model has strong capabilities to apply knowledge gained from training data to previously unseen cases which makes it suitable for anemia diagnosis in production systems. Logistic Regression earned its position as the chosen deployment model because it maintained consistent high performance during both cross-validation testing and actual testing.

**E. Justification of Final Model Choice**

After evaluating various performance metrics through holistic comparison methods Logistic Regression proved itself as the preferred model selection for this project. Logistic Regression achieves classification excellence while offering interpretability features that clinical applications need. The interpretability of Logistic Regression enables domain experts to see which features determine predictions which helps healthcare systems gain better acceptance and trust. This model needs fewer computational resources which allows it to succeed in low-resource environments or within edge devices.

**F. Limitations and Future Improvements**

The effectiveness of Logistic Regression for this application can be enhanced through better feature engineering approaches while exploring ensemble approaches and potentially implementing deep learning models if larger and diverse datasets become accessible. The generalization of findings requires additional tests with diverse demographic and geographic datasets. The models could use SMOTE techniques to solve potential class imbalance problems while SHAP and LIME explainability tools can enhance understanding of complex model boundaries.

##### References

[1] A. B. Zemariam et al., "Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia," *Scientific Reports*, vol. 14, no. 9080, 2024. [Online]. Available: <https://www.nature.com/articles/s41598-024-60027-4>

[2] M. Ramzan et al., "Revolutionizing anemia detection: integrative machine learning models and advanced attention mechanisms," *Visual Computing for Industry, Biomedicine, and Art*, vol. 7, no. 18, 2024. [Online]. Available: <https://vciba.springeropen.com/articles/10.1186/s42492-024-00169-4>

[3] B. Kitaw et al., "Leveraging machine learning models for anemia severity detection among pregnant women following ANC: Ethiopian context," *BMC Public Health*, vol. 24, no. 3500, 2024. [Online]. Available: <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-024-21039-x>

[4] M. A. A. Elmaleeh, "The Identification and Categorization of Anemia Through Artificial Neural Networks: A Comparative Analysis of Three Models," *arXiv preprint*, arXiv:2404.04690, 2024. [Online]. Available: <https://arxiv.org/abs/2404.04690>

[5] S. N. Qasem and A. Mosavi, "Novel Meta-Heuristic Model for Discrimination between Iron Deficiency Anemia and B-Thalassemia with CBC Indices Based on Dynamic Harmony Search," *arXiv preprint*, arXiv:2004.00480, 2020. [Online]. Available: <https://arxiv.org/abs/2004.00480>

[6] A. Mitani et al., "Detecting Anemia from Retinal Fundus Images," *arXiv preprint*, arXiv:1904.06435, 2019. [Online]. Available: <https://arxiv.org/abs/1904.06435>

[7] S. Prajapati et al., "Anemia Classification System Using Machine Learning," *MDPI Information*, vol. 12, no. 1, 2023. [Online]. Available: <https://www.mdpi.com/2227-9709/12/1/19>

[8] M. Chand et al., "Detection of iron deficiency anemia by medical images," *BioData Mining*, vol. 13, no. 1, 2020. [Online]. Available: <https://biodatamining.biomedcentral.com/articles/10.1186/s13040-023-00319-z>

[9] J. Lee et al., "Deep Learning Model-Based Detection of Anemia from Conjunctiva Images," *Healthcare Informatics Research*, vol. 30, no. 1, pp. 12-20, 2024. [Online]. Available: <https://www.e-hir.org/journal/view.php?number=1237>

[10] R. Sharma et al., "A Review on Machine Learning Algorithms for Anemia disease Prediction," *International Research Journal*, vol. 10, no. 2, pp. 45-50, 2023. [Online]. Available: <https://www.irjweb.com/A%20Review%20on%20Machine%20Learning%20Algorithms%20for%20Anemia%20disease%20Prediction.pdf>

[11] A. Singh et al., "Application of machine learning approach for iron deficiency anemia detection," *Data in Brief*, vol. 39, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352914824000078>

[12] M. Khan et al., "Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia," *Scientific Reports*, vol. 14, no. 9080, 2024. [Online]. Available: <https://www.nature.com/articles/s41598-024-60027-4>

[13] S. Sahu et al., "Detection of anaemia using medical images: A comparative study of machine learning approaches," *Data in Brief*, vol. 39, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352914823001272>

m

[14] J. Kim et al., "Artificial Intelligence (AI)-Based Detection of Anaemia Using the Conjunctiva," *Journal of Healthcare Engineering*, vol. 2023, Article ID 11260651, 2023. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11260651/>