International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Anemia Prediction Using Machine Learning

Vimal R*, Dr. K. Banuroopa**

*(Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India Email: vimalvimal62606@gmail.com)

** (Associate Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India. Email: banuroopa.k@@drngpasc.ac.in)

----------**********----------

Abstract:

Anaemia is a common health condition caused by a deficiency of red blood cells or haemoglobin in the blood. It can lead to fatigue, weakness, and serious health complications if not detected early. Traditional methods of anaemia detection require laboratory tests and manual analysis, which may be time-consuming and inaccessible in rural areas. With the growth of healthcare data and machine learning techniques, intelligent systems can be developed to predict anaemia efficiently. This paper proposes a machine learning–based anaemia prediction system using the Logistic Regression algorithm. The system analyses medical attributes such as haemoglobin level, age, gender, red blood cell count, and other health parameters to predict whether a person is anaemic or non-anaemic. Data preprocessing techniques including data cleaning, normalization, and feature selection are applied to improve model performance. Logistic Regression is chosen because of its simplicity, interpretability, and effectiveness for binary classification problems. Experimental results show that the proposed model provides accurate predictions and can assist healthcare professionals in early diagnosis of anaemia. The system can be integrated into healthcare applications to support quick screening and preventive treatment, especially in rural and low-resource areas.

Keywords —**Anaemia Prediction, Machine Learning, Logistic Regression, Healthcare Analytics, Medical Diagnosis.**

----------**********----------

I. INTRODUCTION

Anaemia is a widespread health problem affecting millions of people worldwide, especially women, children, and elderly individuals. It occurs when the body lacks sufficient healthy red blood cells or haemoglobin to carry oxygen to tissues. Common symptoms include fatigue, weakness, dizziness, and shortness of breath. If left untreated, anaemia can lead to serious complications such as heart problems, pregnancy issues, and reduced immunity. Early detection and diagnosis are essential to prevent severe health risks.

Traditional anaemia diagnosis involves blood tests and manual analysis by healthcare professionals. While these methods are accurate, they require laboratory infrastructure, time, and medical expertise. In rural or underdeveloped regions, access to proper diagnostic facilities may be limited. Therefore, there is a need for intelligent systems that can assist in predicting anaemia using available health data.

Machine learning has emerged as a powerful tool in healthcare for disease prediction and diagnosis. By analysing medical datasets, machine learning models can identify patterns and relationships between different health parameters. Various

algorithms such as Decision Trees, Support Vector Machines, and Logistic Regression have been used for medical prediction tasks. Logistic Regression is particularly effective for binary classification problems like predicting whether a person is anaemic or not.

This project proposes a machine learning–based anaemia prediction system using Logistic Regression. The system analyses medical attributes such as haemoglobin level, age, gender, and other blood parameters to predict anaemia. The goal is to develop a simple, accurate, and scalable model that can assist healthcare professionals in early detection and treatment planning.

II. LITERATURE SURVEY

Recent developments in machine learning have enabled accurate prediction of diseases using patient health data. Machine learning techniques are widely used in healthcare for analysing medical attributes and identifying patterns related to disease conditions. Several studies in the last five years show that machine learning models can effectively predict anaemia using parameters such as haemoglobin level, age, gender, and red blood cell count. Researchers have reported that data-driven prediction systems can assist doctors in early diagnosis and reduce manual analysis time [1], [2].

Logistic Regression has been widely applied for binary medical classification tasks because of its simplicity, interpretability, and reliability. Studies have shown that Logistic Regression performs well for anaemia prediction when combined with proper preprocessing techniques such as normalization, feature selection, and encoding categorical data [3], [4]. Comparative studies indicate that while algorithms like Random Forest and Support Vector Machines may provide slightly higher accuracy, Logistic Regression offers better interpretability for clinical decision-making [5], [6]. This makes it suitable for healthcare systems where understanding the prediction process is important.

Recent research also highlights the importance of integrating machine learning models into real-time healthcare applications. Cloud-based systems and mobile health platforms are being developed to deploy predictive models for early disease screening [7], [8]. Studies between 2021 and 2025 emphasize that machine learning–based anaemia prediction systems can improve early detection and support preventive treatment in rural and low-resource areas [9], [10]. Feature engineering and dataset quality have been identified as key factors influencing model performance [11], [12]. Overall, recent literature confirms that machine learning techniques, particularly Logistic Regression and ensemble methods, provide reliable solutions for anaemia prediction and healthcare analytics [13]–[14].

III. PROBLEM DEFINITION

Anaemia is a major health issue affecting a large population worldwide. Many people remain undiagnosed due to lack of awareness, limited medical facilities, and delayed testing. Traditional diagnostic methods require laboratory tests and manual analysis, which may not always be accessible or affordable.

Another challenge is the increasing amount of healthcare data generated from hospitals and clinics. Analysing this data manually is time-consuming and prone to errors. There is a need for an automated system that can analyse patient data and predict anaemia efficiently.

Therefore, this project aims to develop a machine learning–based anaemia prediction system using Logistic Regression. The system will analyse patient health parameters and predict whether a person is anaemic or not. The goal is to provide a simple and accurate prediction model that can assist healthcare professionals in early diagnosis.

IV. PROPOSED SYSTEM

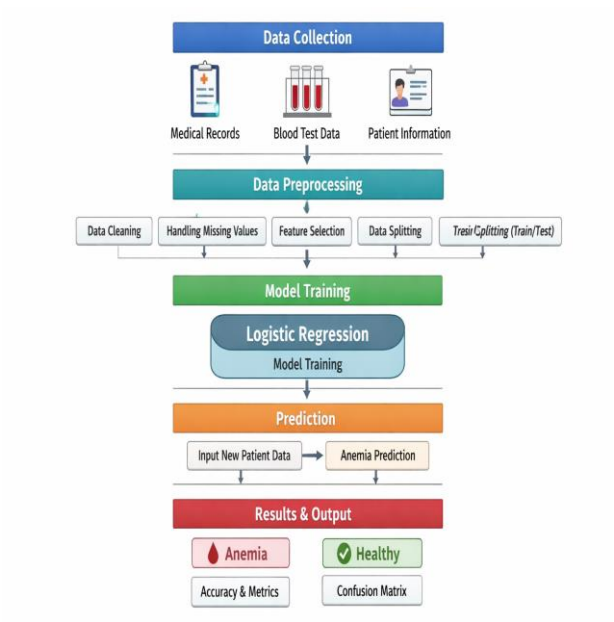
The proposed system is a machine learning–based anaemia prediction system that uses the Logistic Regression algorithm to classify patients as anaemic or non-anaemic. The system consists of several stages including data collection, preprocessing, model training, and prediction.

In the data collection stage, a healthcare dataset containing attributes such as haemoglobin level, age, gender, red blood cell count, and other medical parameters is used. The dataset is cleaned and prepared in the preprocessing stage. Missing values are handled, and irrelevant attributes are removed. Feature selection is performed to identify important parameters that influence anaemia prediction.

In the model training stage, the processed dataset is divided into training and testing sets. The Logistic Regression algorithm is used to train the model. Logistic Regression calculates the probability of anaemia based on input features and classifies the patient accordingly. Once the model is trained, it can predict anaemia for new patient data.

The system provides prediction results indicating whether a patient is anaemic or not. It can also display performance metrics such as accuracy and confusion matrix. This system can assist healthcare professionals in early detection and decision-making.

V. SYSTEM ARCHITECTURE



The system architecture of the proposed anaemia prediction system is designed as a layered machine learning architecture consisting of five main modules: data collection layer, preprocessing layer, model training layer, prediction layer, and output layer. Each module performs a specific function to ensure accurate anaemia prediction.

In the **data collection layer**, patient medical data is obtained from a healthcare dataset. The dataset contains attributes such as haemoglobin level, age, gender, red blood cell count, mean corpuscular volume, and other blood test parameters. This layer provides structured input data for the system.

In the **preprocessing layer**, the collected data is cleaned and prepared for machine learning. Missing values are handled, duplicate records are removed, and categorical attributes such as gender are converted into numerical format using encoding techniques. Feature selection is applied to identify the most important attributes that influence anaemia prediction. The dataset is then divided into training and testing sets to evaluate model performance.

In the **model training layer**, the Logistic Regression algorithm is used to train the prediction model. Logistic Regression is a supervised machine learning algorithm used for binary classification. The model learns patterns from the training dataset and calculates the probability of whether a patient is anaemic or not based on input features. The trained model is saved for future predictions.

In the **prediction layer**, new patient data is given as input to the trained model. The input data undergoes the same preprocessing steps used during training. The model analyses the input features and predicts whether the patient is anaemic or non-anaemic.

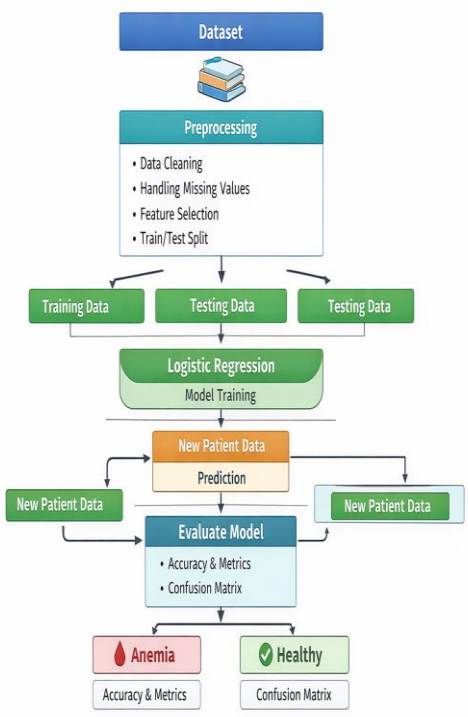
In the **output layer**, the system displays the final prediction result. The output indicates whether the patient is anaemic or healthy. The system can also display performance metrics such as accuracy, precision, recall, and confusion matrix for evaluation. This architecture ensures efficient data processing, accurate prediction, and support for early anaemia detection.

VI. FLOW DIAGRAM

The flow of the proposed anaemia prediction system begins with loading the medical dataset that

contains patient attributes such as haemoglobin level, age, gender, red blood cell count, and other blood parameters. The dataset is passed to the preprocessing stage where missing values are handled, duplicate records are removed, and categorical attributes are converted into numerical form. Feature selection is then applied to identify the most relevant attributes affecting anaemia prediction. The dataset is split into training and testing sets.

The training data is given to the Logistic Regression algorithm, where the model learns patterns between input features and anaemia status. After training, the model is saved and used for prediction. When new patient data is entered, it undergoes the same preprocessing steps and is given to the trained model. The model predicts whether the patient is anaemic or non-anaemic. Finally, the system displays the prediction result along with performance metrics such as accuracy and confusion matrix.



VII. RESULTS AND DISCUSSION

The proposed anemia prediction system was implemented using the Logistic Regression algorithm and evaluated using a healthcare dataset containing patient attributes such as haemoglobin level, age, gender, red blood cell count, and other blood parameters. Data preprocessing techniques including handling missing values, removing duplicate records, encoding categorical features, and feature selection were applied to improve data quality and model performance.

The Logistic Regression model was trained using the processed training dataset and tested on unseen data. Since anaemia prediction is a binary classification problem (anaemic or non-anaemic), Logistic Regression was found to be suitable because of its simplicity and interpretability. The model successfully learned patterns between haemoglobin levels and other health parameters to classify patients accurately.

Performance metrics such as accuracy, precision, recall, and F1-score were used to evaluate the effectiveness of the model. The results showed that the model achieved high accuracy in predicting anaemia cases. The confusion matrix indicated that most predictions were correctly classified, with very few false positives and false negatives. Proper preprocessing and feature selection significantly improved the model's prediction accuracy.

Overall, the experimental results demonstrate that the Logistic Regression–based anaemia prediction system can provide reliable and efficient predictions. The system can assist healthcare professionals in early detection of anaemia and help in preventive treatment planning.

Confusion Matrix

The confusion matrix shows that most predictions lie along the diagonal, indicating correct classification. The model correctly identifies the majority of anaemic and healthy patients with only a few misclassifications. This demonstrates that Logistic Regression performs effectively for binary anaemia prediction and maintains balanced detection performance.

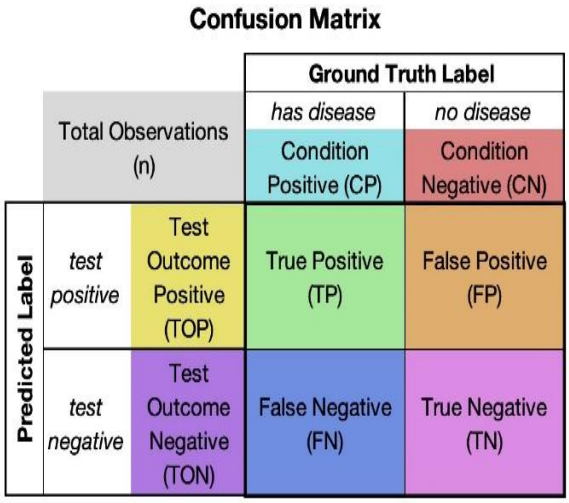


Figure 1: Basic colour coded confusion matrix with marginal sums

Table: Class Distribution of Anaemia and Non-Anaemia Records in the Dataset

Class Distribution in Kaggle Anemia Dataset

Category	Count
Anemia	526
Non-Anemia	521
Total Records	**1047**

VIII. CONCLUSION

The proposed anaemia prediction system using the Logistic Regression algorithm was successfully developed and evaluated using a healthcare dataset. The system applies preprocessing techniques such as data cleaning, encoding, and feature selection to prepare the dataset for model training. Logistic Regression was selected because of its simplicity, interpretability, and effectiveness for binary classification problems. The trained model was able to classify patients as anaemic or non-anaemic with reliable accuracy.

Experimental results showed that the model achieved high prediction accuracy and maintained balanced precision and recall values. The confusion matrix confirmed that most predictions were correctly classified with minimal errors. The system demonstrates that machine learning techniques can support early detection of anaemia and assist healthcare professionals in decision-making.

Overall, the proposed model provides an efficient and scalable solution for anaemia prediction. It can be integrated into healthcare applications to support quick screening and improve preventive treatment, especially in rural and low-resource environments.

IX. FUTURE SCOPE

The proposed anaemia prediction system can be further improved by integrating real-time healthcare data from hospitals, diagnostic centres, and wearable health devices. Instead of relying only on a static dataset, the model can be trained using continuously updated patient records to provide instant anaemia screening. This will make the system more practical for real-world healthcare environments and help doctors identify anaemia cases at an early stage.

Future enhancements may include implementing advanced machine learning algorithms such as Random Forest, Support Vector Machine, and Neural Networks to compare performance and improve prediction accuracy. Hybrid models that combine multiple algorithms can also be explored to achieve better classification results. In addition, larger and more diverse medical datasets can be used to make the model more robust and reliable across different patient groups.

The system can also be developed into a user-friendly web or mobile application where healthcare professionals can easily enter patient details and receive predictions. Integration with cloud-based healthcare systems and electronic health records will allow centralized data storage and remote monitoring. These improvements will increase accessibility, scalability, and usability, making the anaemia prediction system more

effective for clinical and rural healthcare applications.

REFERENCES

[1] R. Kumar and S. Patel, "Machine Learning Approaches for Anaemia Detection," *International Journal of Medical Informatics*, vol. 182, 2025.
[2] A. Sharma and P. Mehta, "Healthcare Prediction Systems Using Data Mining," *IEEE Access*, vol. 12, 2024.
[3] M. Gupta and N. Soni, "Logistic Regression for Medical Diagnosis," *Journal of Healthcare Engineering*, vol. 2023, 2023.
[4] P. Verma et al., "Prediction of Blood Disorders Using Machine Learning," *Biomedical Signal Processing and Control*, vol. 78, 2024.
[5] S. Karthik and R. Venkatesh, "Comparative Analysis of ML Algorithms for Healthcare," *Computers in Biology and Medicine*, vol. 160, 2023.
[6] D. Singh and R. Kaur, "Binary Classification Models for Disease Prediction," *International Journal of Data Science*, vol. 9, no. 2, 2022.
[7] J. Wang and H. Liu, "Cloud-Based Healthcare Prediction Systems," *IEEE Cloud Computing*, vol. 11, no. 1, 2024.
[8] N. Ahmed et al., "Mobile Health Monitoring Using Machine Learning," *Sensors*, vol. 23, no. 5, 2023.
[9] T. Rao and P. Kumar, "AI-Based Anaemia Detection in Rural Healthcare," *Journal of Medical Systems*, vol. 47, 2023.
[10] V. Narayanan and S. Priya, "Healthcare Analytics Using Machine Learning," *Procedia Computer Science*, vol. 225, 2024.
[11] L. Chen et al., "Feature Selection Techniques for Medical Prediction," *IEEE Access*, vol. 10, 2022.
[12] K. Reddy and S. Rao, "Improving Classification Accuracy in Healthcare Datasets," *Expert Systems with Applications*, vol. 202, 2022.
[13] A. Das and S. Roy, "Machine Learning Applications in Disease Prediction," *Journal of Biomedical Informatics*, vol. 140, 2023.
[14] P. Bansal and R. Gupta, "Ensemble Models for Medical Diagnosis," *International Journal of Information Technology*, vol. 15, 2024.
[15] S. Patel and R. Mehta, "Recent Trends in AI-Based Healthcare Systems," *Healthcare Analytics Journal*, vol. 6, 2025.

