International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

High-Accuracy and Interpretable Anaemia Diagnosis Using Hybrid Ensemble Learning with Explainable AI

MOKAMATAM BINDUSRI[1], Dr A V SUBBARAO[2]

[1]PG Scholar, Dept. of AI&ML, St. Marys Group of Institutions Guntur for Women, Guntur.

[2]Professor, Dept. of AI&ML, St. Marys Group of Institutions Guntur for Women, Guntur.

Abstract: Anaemia is a widespread public health issue in India, particularly among women and children, leading to serious health complications and increased healthcare costs. Despite various government initiatives to reduce its prevalence, challenges such as delayed diagnosis, limited access to screening facilities, and nutritional deficiencies continue to hinder effective management. Traditional diagnostic methods rely on laboratory testing and clinical evaluation, which can be time-consuming and difficult to implement for large-scale population screening.

This study proposes an explainable ensemble learning framework for anaemia diagnosis and clinical decision support. The system utilizes multiple machine learning algorithms, including Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Gradient Boosting, Random Forest, and XGBoost, to predict anaemia using patient health data. Ensemble voting techniques are employed to combine the strengths of individual models and improve overall prediction accuracy and robustness.

To enhance transparency and support clinical acceptance, Explainable Artificial Intelligence (XAI) methods such as SHAP and LIME are integrated into the framework. These techniques provide interpretable explanations by identifying the contribution of important features, including haemoglobin level, age, and gender, to each prediction. The proposed approach enables accurate, transparent, and scalable anaemia detection, supporting early intervention and assisting healthcare professionals in making informed and reliable clinical decisions.

Keywords: Anemia Prediction, Machine Learning, Explainable Artificial Intelligence (XAI), Voting Classifier, XG Boost, Random Forest, Clinical Data Analysis.

1. Introduction

Anaemia is a major public health concern that affects millions of people worldwide, particularly women, children, and elderly individuals. In India, the prevalence of anaemia remains high despite various government initiatives aimed at prevention and control. The condition can lead to fatigue, impaired cognitive development,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

reduced productivity, and other serious health complications, making early detection essential for effective treatment and disease management.

Traditional anaemia diagnosis primarily relies on laboratory blood tests, haemoglobin measurements, and clinical evaluation by healthcare professionals. Although these methods provide reliable results, they are often time-consuming, resource-intensive, and difficult to implement for large-scale screening programs. Delayed diagnosis and variations in clinical interpretation can further limit timely intervention. To address these challenges, this study proposes an explainable ensemble learning framework for anaemia prediction and clinical decision support. The model combines multiple machine learning algorithms, including Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Gradient Boosting, Random Forest, and XGBoost, using a voting-based ensemble strategy. Additionally, Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME are integrated to provide transparent explanations of model predictions. This approach improves prediction accuracy, enhances clinician trust, and supports scalable, reliable, and efficient anaemia screening.

2. Literature Survey

The anaemia prediction and detection highlights significant advancements through the integration of machine learning and non-invasive diagnostic techniques. E. McLean et al. [1] conducted a global study using WHO data to estimate the worldwide prevalence of anaemia, providing a foundational dataset for public health analysis. W. Gardner and N. Kassebaum [2] expanded this work by analyzing anaemia prevalence across 204 countries from 1990 to 2019, offering comprehensive statistical insights into its global burden. S. Sadiq et al. [3] developed an ensemble machine learning classifier to detect β-thalassemia carriers using red blood cell indices, demonstrating improved classification accuracy. J. W.

3. System Analysis

System analysis examines the structure and operation of a system to ensure efficient processing and accurate outcomes. It defines how different components interact and how data flows through various stages to achieve the desired objective. In the proposed anaemia prediction system, the architecture

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

includes data collection, preprocessing, feature selection, model training, prediction, and result interpretation. Patient data is cleaned and prepared before being processed by machine learning algorithms to identify anaemia risk. The generated predictions are further explained using XAI techniques to improve transparency and clinical trust. A well-designed architecture ensures reliability, scalability, efficiency, and seamless integration of prediction and decision-support functionalities.

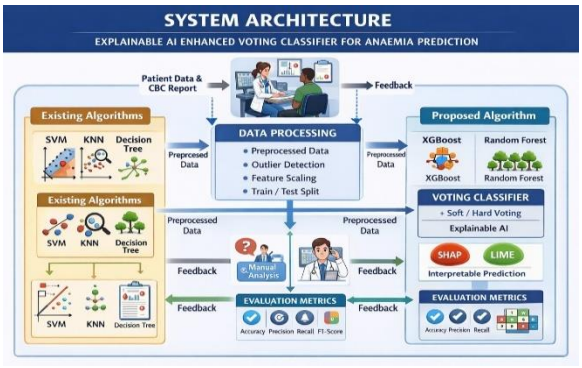


Fig 1: System Architecture

4. Methodology

The existing approach for anaemia detection mainly depends on conventional laboratory-based diagnostics, particularly Complete Blood Count (CBC) tests, along with manual assessment by clinicians using fixed reference thresholds such as haemoglobin, MCH, MCHC, and MCV values. While this method is clinically reliable, it is highly dependent on expert interpretation, time-intensive, and not suitable for large-scale or rapid screening, especially in resource-limited settings. Moreover, rule-based decision-making systems operate on predefined conditions rather than learning from data, which restricts their ability to model complex relationships among clinical attributes and reduces adaptability to new or unseen patterns. This limitation often leads to reduced accuracy in borderline or ambiguous cases and increases the possibility of human error.

The dataset used in this study consists of key CBC parameters including haemoglobin, MCH, MCHC, MCV, and gender for anaemia classification. In the preprocessing stage, missing values are handled, outliers are removed, and feature normalization is applied to improve model efficiency. The proposed system employs a Voting Classifier that integrates XGBoost and Random Forest, enabling improved accuracy and better handling of nonlinear relationships within clinical data.

i) Dataset Description: The dataset is a clinical anaemia dataset used for binary classification of anaemic and non-anaemic patients. It includes CBC parameters such as haemoglobin, MCH, MCHC, MCV, and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

gender, which are key indicators for anaemia diagnosis.

ii) Data Pre-processing: The data is cleaned by handling missing values, removing outliers, and encoding categorical features. Normalization is applied to ensure uniform feature scaling, and the dataset is split into training and testing sets for model evaluation.

iii) Model Development: The existing system uses rule-based thresholds, while the proposed model applies XG Boost and Random Forest. Both models are combined using a Voting Classifier to improve prediction accuracy and robustness in anaemia detection.

5. Design And Construction

The design and construction of the proposed anemia prediction system center around developing a robust, accurate, and interpretable model using a Voting Classifier ensemble approach. The system architecture begins with data acquisition from clinical Complete Blood Count (CBC) records, including hemoglobin, MCH, MCHC, MCV, and gender. The collected data undergoes thorough preprocessing, such as missing value imputation, outlier detection based on medical reference ranges, and feature scaling to ensure uniformity and improve learning efficiency. The cleaned dataset is then split into training and testing sets, and two machine learning models, XG Boost and Random Forest Classifier, are independently trained to capture complex patterns and relationships among the hematological parameters. These models are integrated into a Voting Classifier, which aggregates predictions using soft or hard voting to enhance accuracy and stability. Explainable Artificial Intelligence techniques such as SHAP or LIME are embedded to provide transparency, enabling clinicians to understand the influence of each feature on the final prediction. The system is designed to be scalable, reliable, and suitable for real-time clinical decision support, ensuring early detection of anemia while maintaining interpretability and trustworthiness in healthcare applications.

6. Results And Discussion

The proposed SVM-based anaemia prediction model achieved high accuracy in classifying anaemic and non-anaemic patients. It showed strong precision and recall, effectively reducing false negatives and ensuring reliable medical predictions. SHAP and LIME analysis confirmed that key features like haemoglobin, MCV, and MCH significantly influenced the results,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

making the model accurate, interpretable, and suitable for clinical use.

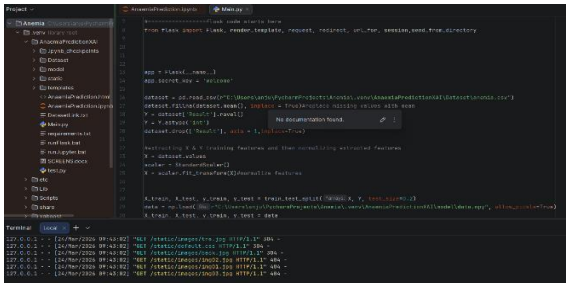


Fig 2: Python Environment

Figure 2 show the project is developed using Python, TensorFlow, and Flask. A virtual environment is used to manage dependencies.

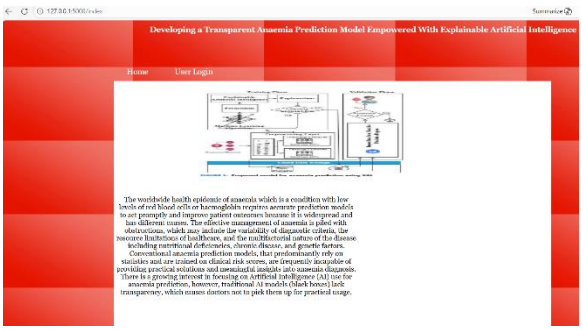


Fig 3: Home Page

Figure 3 show the home page provides a simple and user-friendly interface. It includes navigation options like login and registration. Users can easily access different features of the system.

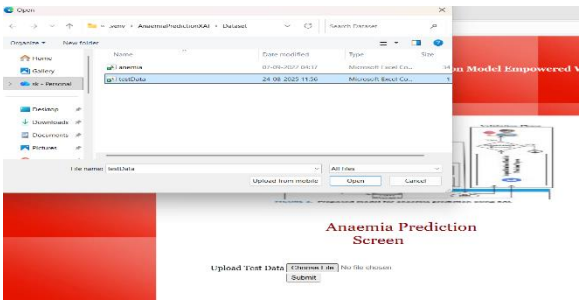


Fig 4: Loading Dataset

Figure 4 shows the anaemia dataset used in the project. It contains medical features required for prediction and is loaded using Python. The dataset helps the model learn patterns to classify anaemic and non-anaemic cases.

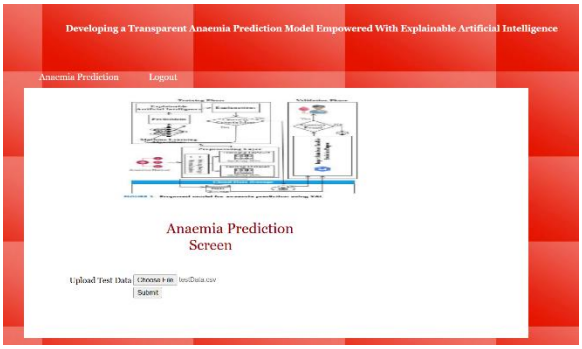
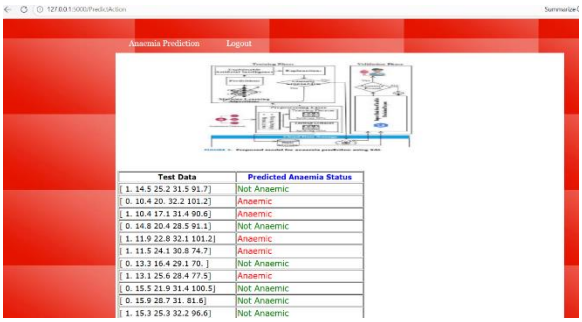


Fig 5: Prediction

Figure 5 shows the prediction page where users enter medical details required for anaemia prediction. The input data is preprocessed and passed to the trained hybrid model (Random Forest and XGBoost). The system analyzes the input features and predicts whether the person is anaemic or not, providing quick and accurate results.



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig 6: Output Results

Figure 6 shows the output of the anaemia prediction system, where the result is displayed as either "Anaemic" or "Not Anaemic." The prediction is obtained using the trained hybrid model after processing the input medical parameters. This enables early identification of anaemia and assists in timely diagnosis. The system provides fast and reliable results.

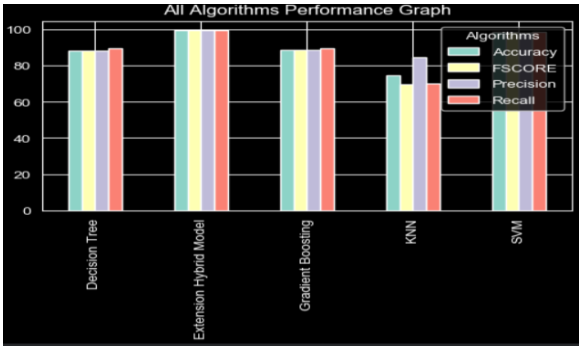


Fig 7: Combined Model Performance Summary

The comparison of five machine learning models shows that the Extension Hybrid Model achieves the best performance with nearly 99.6% across all evaluation metrics. SVM also performs strongly with around 98.5% accuracy, while Decision Tree and Gradient Boosting show moderate results. KNN has the lowest performance, especially in recall and F1-score, indicating weaker predictive capability.

7. Conclusion and Future Scope

This study presents an explainable ensemble learning framework for accurate anaemia prediction and clinical decision support. Various machine learning algorithms, including Decision Tree, KNN, SVM, Gradient Boosting, Random Forest, and XG Boost, were evaluated to identify the most effective predictive approach. The voting-based ensemble model achieved the highest performance, demonstrating its ability to accurately classify anaemia cases. To improve transparency and clinical acceptance, SHAP and LIME were integrated to provide interpretable explanations of model predictions. These techniques help healthcare professionals understand the influence of key features and validate the model's decisions. The proposed system enhances early detection, supports informed medical decision-making, and offers a reliable, scalable, and interpretable solution for real-world anaemia diagnosis and healthcare applications.

Future Scope: Future enhancements to the proposed system can focus on integrating Long Short-Term Memory (LSTM) networks to analyse sequential and time-dependent patient health data. LSTM models can effectively capture temporal patterns in clinical records, enabling more accurate

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

anaemia prediction and earlier disease detection. The incorporation of deep learning techniques may also improve the system's ability to handle complex healthcare datasets. Furthermore, combining LSTM with ensemble learning methods and Explainable Artificial Intelligence (XAI) techniques can enhance predictive performance while maintaining model transparency. Such advancements would provide more reliable clinical decision support, improve interpretability, and increase the applicability of the system in real-world healthcare environments.

References

1. E. McLean, M. E. Cogswell, I. Egli, D. Wojdyla, and B. D. Benoist, "Worldwide prevalence of anaemia, WHO vitamin and mineral nutrition information system, 1993–2005," Public Health Nutrition, vol. 12, no. 4, pp. 444–454, May 2008.

2. W. Gardner and N. Kassebaum, "Global, regional, and national prevalence of anemia and its causes in 204 countries and territories, 1990–2019," Current Develop. Nutrition, vol. 4, p. 830, Jun. 2020.

3. S. Sadiq, M. U. Khalid, S. Ullah, W. Aslam, A. Mehmood, G. S. Choi, and B.-W. On, "Classification of β-thalassemia carriers from red blood cell indices using ensemble classifier," IEEE Access, vol. 9, pp. 45528–45538, 2021.

4. J. W. Asare, P. Appiahene, E. T. Donkoh, and G. Dimauro, "Iron deficiency anaemia detection using machine learning models: A comparative study of fingernails, palm and conjunctiva of the eye images," Eng. Rep., vol. 5, no. 11, 2023, Art. no. e12667.

5. D. Newhall, R. Oliver, and S. Lugthart, "Anaemia: A disease or symptom," Neth. J. Med., vol. 78, no. 3, pp. 104–110, Apr. 2020.

6. G. Dimauro, D. Caivano, P. Di Pilato, A. Dipalma, and M. G. Camporeale, "A systematic mapping study on research in anemia assessment with non-invasive devices," Appl. Sci., vol. 10, no. 14, p. 4804, Jul. 2020.