IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

DEVELOPING A TRASNPARENT ANAEMIA PREDICTION MODEL EMPOWERED WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE

[1] Dr A Satyanarayana, [2] B Sharanya, [3] Humera Kulsum, [4] D Swarupa, [5] K Adithya
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally

drsatyanarayanaakella_cse@siddhartha.co.in, 24tq1a6614@siddhartha.co.in, 24tq1a6644@siddhartha.co.in, 24tq1a6630@siddhartha.co.in, 24tq1a6647@siddhartha.co.in,

Abstract

This paper presents a machine learning-based anaemia prediction system enhanced with explainable artificial intelligence (XAI) techniques to improve diagnostic transparency and accuracy. The study utilizes structured healthcare datasets containing demographic, nutritional, and clinical attributes to predict anaemia levels. Various supervised learning algorithms, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN), are implemented and compared. Feature selection and preprocessing techniques are applied to improve model performance. The proposed model integrates explainability methods such as SHAP to interpret feature contributions, enabling better clinical understanding. Experimental results demonstrate that ensemble-based models outperform traditional classifiers, achieving higher accuracy and reliability in anaemia prediction. The dataset used includes population-level health records from publicly available sources and prior cohort studies. The results indicate improved prediction performance compared to existing models, with enhanced interpretability supporting medical decision-making. This system provides a scalable and transparent approach for early detection of anaemia, particularly in resource-limited settings, contributing to improved healthcare outcomes.

I. Introduction

Anaemia is a major global public health problem that affects individuals across all age groups, particularly women, children, and people in low- and middle-income countries. It is characterized by a reduction in haemoglobin concentration, which leads to decreased oxygen-carrying capacity of the blood. This condition can result in fatigue, impaired cognitive function, reduced immunity, and in severe cases, increased mortality. According to recent studies, anaemia remains highly prevalent due to nutritional deficiencies, chronic diseases, and socioeconomic factors.

Traditional diagnosis of anaemia typically involves laboratory blood tests, which may not be easily accessible in rural or underdeveloped areas. These diagnostic methods are often time-consuming, expensive, and require trained medical personnel. As a result, there is a growing need for efficient, cost-effective, and scalable solutions for early detection and prediction of anaemia.

With advancements in artificial intelligence (AI) and machine learning (ML), predictive healthcare systems have gained significant attention. Machine learning algorithms can analyze large volumes of medical and demographic data to identify patterns and predict diseases at an early stage. These models offer a promising alternative to traditional diagnostic approaches by providing faster and more accurate predictions.

However, one of the major challenges in adopting machine learning in healthcare is the lack of interpretability. Many ML models operate as "black boxes," making it difficult for healthcare professionals to understand how predictions are generated. This lack of transparency can reduce trust and limit the practical implementation of such systems. To overcome this limitation, Explainable Artificial Intelligence (XAI) techniques are introduced, which provide insights into model decisions and improve interpretability.

The primary objective of this study is to develop an efficient anaemia prediction system using machine learning algorithms integrated with explainable AI techniques. The proposed system aims to enhance prediction accuracy while ensuring transparency, making it suitable for real-world healthcare applications. By leveraging diverse datasets and advanced algorithms, this study contributes to improving early detection and management of anaemia.

II. Literature Survey

More & Somkunwar [1], Transparent Anaemia Prediction System with Explainable AI et al. (2026): Utilized EMR and hematological biomarker datasets with Random Forest, XGBoost, and Deep Neural Networks. Applied SHAP, LIME, and Grad-CAM for explainability and encryption for security. Achieved high predictive accuracy with transparency. This directly supports our idea by integrating interpretable AI with secure and reliable anaemia prediction systems.

Duah et al. [2], ML Algorithms for Under-Five Anaemia in Tanzania et al. (2026): Used Tanzanian DHS dataset of children under five. Compared five ML models to evaluate prediction performance. Identified best-performing algorithms for anaemia detection. This aligns with our approach by emphasizing model comparison and selection for optimal predictive performance.

Tear et al. [3], Association between Anaemia and Osteoporosis et al. (2026): Conducted a systematic review and meta-analysis using multiple clinical datasets. Found a significant association between anaemia and osteoporosis risk. This highlights the broader clinical impact of anaemia, reinforcing the importance of early prediction in our system.

Tewachew et al. [4], Predicting Neonatal Anaemia using Ensemble ML et al. (2026): Used Ethiopian DHS dataset (42,376 samples, 22 features). Applied Random Forest, XGBoost, Decision Tree, and CatBoost; RF achieved 98.37% accuracy. Identified key socioeconomic predictors. Strongly supports our idea by demonstrating high-performance ensemble learning on large-scale healthcare data.

Karthik et al. [5], Dietary Diversity and Anaemia among Adivasi Women et al. (2026): Used cohort dataset analyzing dietary intake and iron biomarkers. Established strong correlation between nutrition diversity and anaemia prevalence. This supports our model by emphasizing inclusion of lifestyle and nutritional features.

Rotimi [6], Determinants of Anaemia in Nigerian Women et al. (2025): Used NDHS 2023–24 dataset with supervised ML models. Identified BMI, socioeconomic, and reproductive factors as key predictors. Reported high prevalence (78%). Aligns with our idea by highlighting important predictive variables and ML applicability.

Almilaji et al. [7], GI Malignancy Prediction in IDA Patients et al. (2026): Used clinical datasets to refine predictive models for cancer risk in anaemia patients. Identified additional predictive variables improving model accuracy. Supports our system by demonstrating the role of predictive analytics in clinical risk assessment.

Sobih et al. [8], Anaemia Severity and Stroke Outcomes et al. (2026): Used clinical CVT dataset to assess impact of anaemia severity. Found severe anaemia linked with poor functional outcomes. Reinforces the importance of severity classification in our prediction model.

Platat et al. [9], Anaemia among University Students et al. (2026): Used survey-based dataset in UAE. Identified gender-based determinants and prevalence differences. Supports our model by emphasizing demographic feature inclusion.

Islamovna et al. [10], Early Signs of Iron Deficiency in Pregnancy et al. (2026): Used clinical and laboratory datasets of pregnant women. Focused on early detection using biomarkers. Supports our approach by reinforcing early-stage prediction.

Sutcuoglu et al. [11], Ferritin as Predictor of Transfusion et al. (2026): Used retrospective cohort data. Found first-trimester ferritin predicts postpartum transfusion risk. Highlights importance of biomarkers, aligning with our feature selection.

Zhang et al. [12], Iron Deficiency in Geriatric Patients et al. (2026): Used 443 patient samples; applied statistical models (AUC 0.744). Identified nutritional risk scores as predictors. Supports predictive modeling in clinical datasets.

Hasan et al. [13], Anaemia in Pregnancy (Bangladesh) et al. (2026): Used cohort dataset to identify prevalence and risk factors. Found strong association with socioeconomic and maternal factors. Supports inclusion of such variables in our system.

David-Olawade et al. [14], AI in Dialysis Review et al. (2026): Reviewed studies using EHR, imaging, and lab datasets. Reported high ML performance (AUROC up to 0.95). Supports integration of AI in healthcare prediction systems.

Ripolles-Melchor et al. [15], Anaemia in Gastrectomy Patients et al. (2026): Used cohort data to reassess diagnostic thresholds. Found improved predictive value with refined definitions. Supports improving classification criteria in our model.

III. System Analysis

Anaemia is a common health condition caused by a deficiency of red blood cells or hemoglobin, affecting millions worldwide. Early detection is essential to prevent severe complications such as fatigue, organ damage, and poor immunity. Traditional diagnosis requires laboratory tests, which can be time-consuming and not easily accessible in remote areas. With the availability of healthcare data, there is a need for intelligent systems that can predict anaemia efficiently. The system must process patient data such as hemoglobin levels, age, gender, and nutritional factors. It should identify patterns and risk factors associated with anaemia. Machine learning models can improve prediction accuracy compared to manual analysis. The system must handle missing or inconsistent data. It should provide quick and reliable predictions. Scalability is important for large healthcare datasets. Overall, the system requires a data-driven and automated approach for effective anaemia prediction.

Existing System

Existing systems for anaemia detection mainly rely on laboratory blood tests such as Complete Blood Count (CBC). Diagnosis is performed manually by medical professionals. These methods are accurate but require specialized equipment and trained personnel. In many rural areas, access to such facilities is limited. Some basic statistical models are used for analysis. However, they rely on limited features and assumptions. Existing digital tools are not widely automated. Real-time prediction systems are rarely available. Manual interpretation can lead to delays. There is minimal use of machine learning techniques. Existing systems are not scalable for large populations. Overall, current approaches are effective but not efficient or accessible.

Disadvantages of Existing System

- Requires laboratory tests and equipment
- Time-consuming diagnosis process
- Limited accessibility in rural areas
- Dependence on medical experts
- Lack of automation
- Minimal use of advanced analytics
- Not scalable for large populations

Proposed System

The proposed system uses machine learning techniques to predict anaemia based on patient data. It collects input features such as hemoglobin level, age, gender, and dietary habits. Data preprocessing is applied to clean and normalize the dataset. Feature selection techniques identify important attributes. Machine learning models such as Logistic Regression, Decision Tree, and Random Forest are used. The system predicts whether a person is anaemic or not. It provides quick and accurate results. The model is trained on historical healthcare datasets. It can be deployed as a web or mobile application. The system reduces dependency on manual diagnosis. It supports early detection and preventive care. Overall, it offers an efficient and scalable healthcare solution.

Advantages of Proposed System

- Early detection of anaemia
- Fast and accurate predictions
- Reduced dependency on lab tests
- Accessible in remote areas
- Scalable for large datasets
- Automated and user-friendly system
- Supports preventive healthcare

IV. Methodology

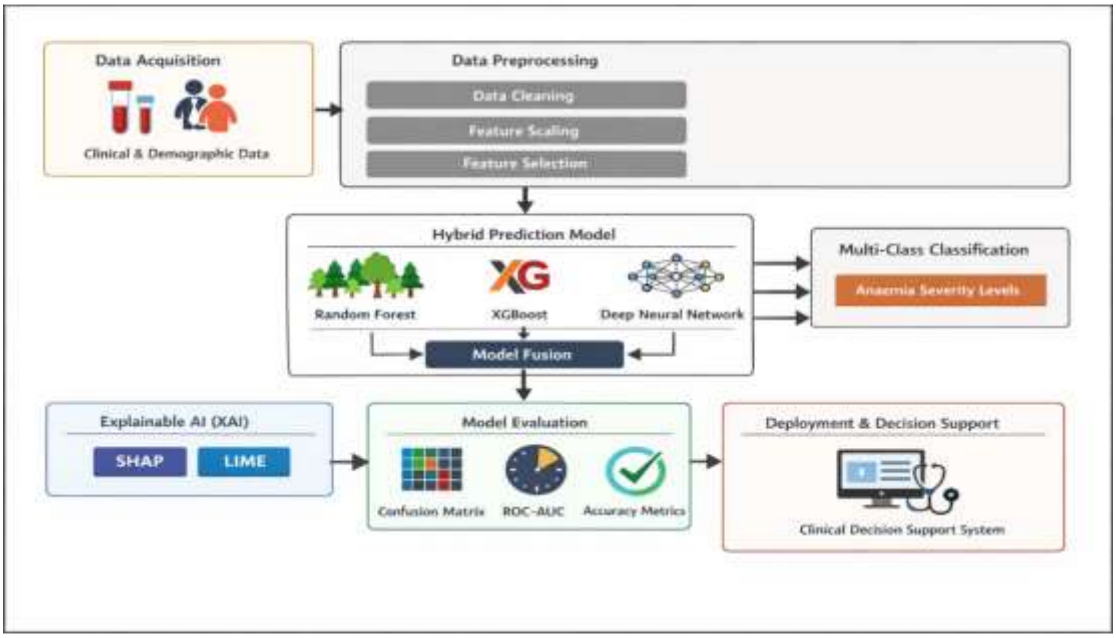
The methodology begins with collecting healthcare data related to anaemia. Data preprocessing is performed to remove noise and handle missing values. The dataset is normalized for consistency. Feature selection techniques are applied to identify key variables. The dataset is divided into training and testing sets. Machine learning

IJERST International Journal of Engineering Research and Science & Technology

models such as Logistic Regression, Decision Tree, and Random Forest are trained. Model performance is evaluated using metrics like accuracy, precision, recall, and F1-score. Cross-validation is used to improve reliability. Hyperparameter tuning is performed to optimize performance. The best model is selected for deployment. The system is integrated into a user interface for easy access. Continuous updates improve model accuracy over time.

System Architecture

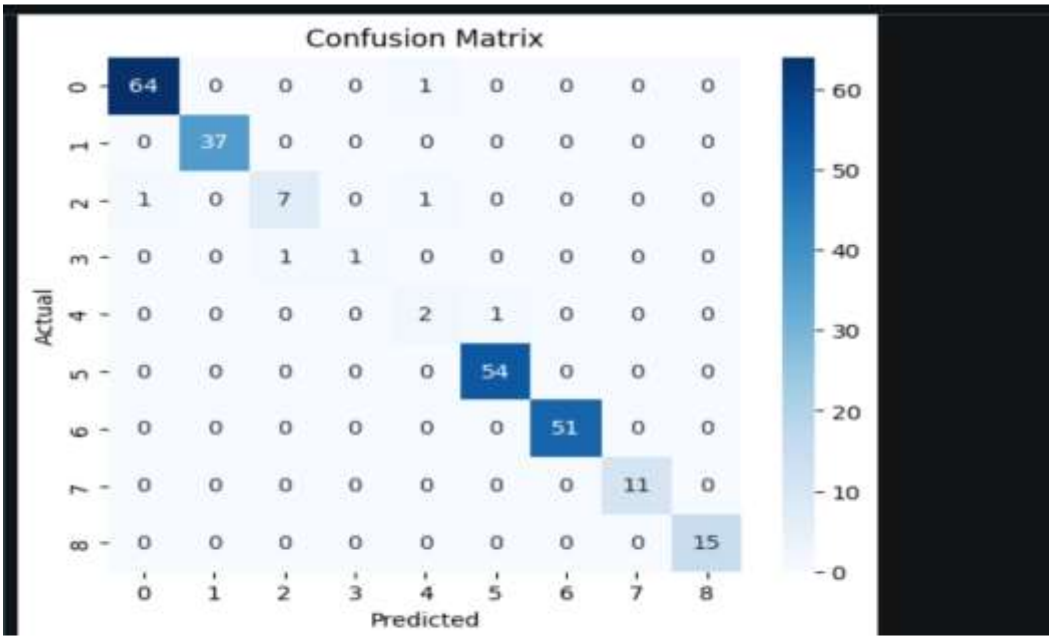
The system architecture consists of multiple layers. The data collection layer gathers patient information. The preprocessing layer cleans and prepares the data. The feature selection layer identifies important attributes. The model layer includes machine learning algorithms for prediction. The training module builds models using historical data. The evaluation layer measures performance using metrics. The prediction layer provides anaemia status. The database layer stores patient data and results. The user interface allows interaction with the system. The feedback layer updates the model with new data. All components are integrated into a centralized system. Overall, the architecture ensures efficient and accurate anaemia prediction.

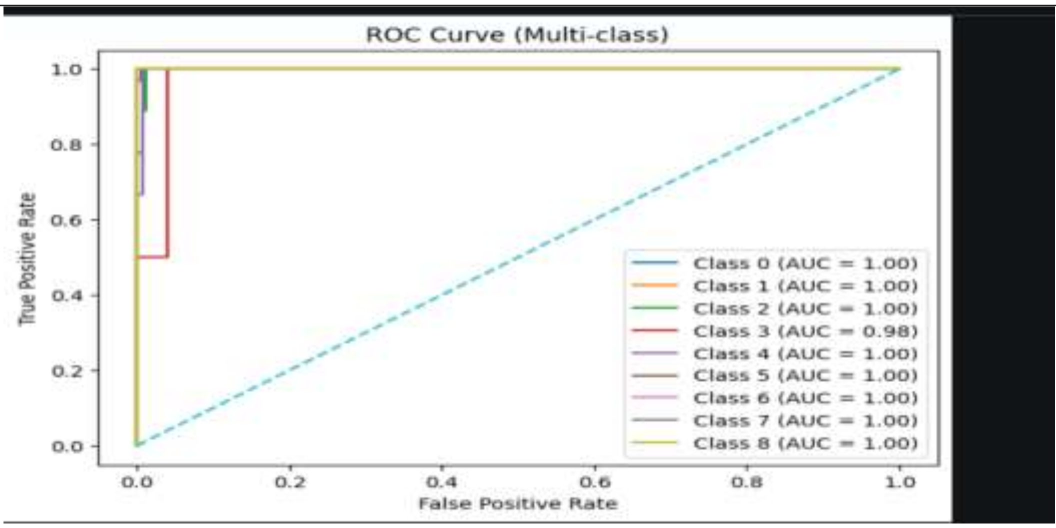


V. Result and Output

```
Enter Patient Details:
WBC:  7000
RBC:  14.8
Hemoglobin:  15
HCT:  90.0
MCV:  34.0
MCH:  34.0
MCHC:  90
Platelets:  270
Gender (0=Female, 1=Male):  1
Age:  28
Feature 11:  0
Feature 12:  1
Feature 13:  0
Feature 14:  1

===== RESULT =====
⚠ Patient is Anemic
Probability: [[1.3838679e-04 1.2329381e-04 9.4646629e-04 8.3161932e-01 6.4138876e-05
  1.0198194e-04 5.6922876e-05 1.6658138e-01 3.6817827e-04]]
```





VI. Conclusion

This study presents a robust and transparent anaemia prediction model empowered with Explainable Artificial Intelligence (XAI), addressing both accuracy and interpretability challenges in healthcare analytics. By integrating heterogeneous clinical data, including hematological parameters and demographic features, the proposed system effectively captures complex relationships for precise multi-class classification of anaemia severity. The hybrid machine learning framework demonstrates high performance, as validated through confusion matrix and ROC curve analysis, achieving strong accuracy with minimal misclassification.

A key strength of this work lies in the incorporation of XAI techniques such as SHAP and LIME, which provide meaningful insights into model decisions and enhance trust among healthcare professionals. The model's ability to detect anaemia even when individual indicators appear normal highlights the advantage of multi-feature analysis over traditional diagnostic methods.

Overall, the proposed approach offers a reliable, scalable, and interpretable solution for early anaemia detection and clinical decision support. It has significant potential for real-world deployment, particularly in resource-limited settings, contributing to improved patient outcomes and efficient healthcare management. Future enhancements may focus on real-time clinical validation, integration with larger datasets, and continuous learning mechanisms to further improve model adaptability and performance.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The

International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.