

# Explainable AI for Anaemia Prediction: Enhancing Clinical Transparency through SHAP and LIME Interpretability

K Mounika<sup>1</sup>, M Hima Sree<sup>2</sup>, U Dinesh<sup>3</sup>, T Dinesh<sup>4</sup>, M Shivamani<sup>5</sup>

<sup>1,2,3,4,5</sup>Department of Information Technology

<sup>1,2,3,4,5</sup>Sri Venkateswara college of Engineering, Tirupati, India.

Corresponding Author/Guide: K. Mounika, M.Tech, Assistant Professor

## Abstract:

Anaemia is a condition characterized by a deficiency of healthy red blood cells or haemoglobin, resulting in a reduced capacity of the blood to carry oxygen to the body's tissues. It is a global health concern requiring accurate, interpretable diagnostic tools to improve patient outcomes. Existing prediction systems predominantly use statistical techniques and black-box AI models, which suffer from limited transparency and practical applicability, restricting their adoption in clinical settings. These models fail to provide the actionable insights necessary for informed clinical decisions, often relying on small, non-generalizable datasets and lacking real-world integration. Addressing these limitations, the proposed system advances the field by deploying a transparent and explainable AI (XAI) methodology for anaemia prediction, utilizing SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to offer clear, interpretable feature contributions behind each diagnosis. This approach facilitates understanding for clinicians, enhances model credibility, and bridges the gap between predictive accuracy and clinical interpretability. Future directions include extending the application of the model to wider populations, integrating continuous health data from wearable devices, and seamless adoption within Electronic Health Records (EHR) systems. These enhancements promise superior reliability, data-driven insights, and improved patient trust, ultimately enabling timely interventions and more effective anaemia management across varied healthcare environments.

**Keywords:** Anaemia, transparent and explainable AI (XAI), predictive accuracy, clinical interpretability, Electronic Health Records (EHR) systems, diagnostic tools.

## 1. INTRODUCTION

In this paper, we investigate the use of Explainable AI to enhance the transparency and trustworthiness of AI models for anaemia prediction in clinical settings bridging the critical gap between predictive accuracy and clinical interpretability often seen in traditional black-box AI models. Although many of these systems are highly predictive, they often do not provide clear, actionable insights into their diagnostic reasoning, limiting their clinical utility and adoption. XAI methods directly remedy this deficiency by explaining the exact clinical features contributing to a prediction in a way that builds clinician trust and enables more informed decision-making. Detailing which symptoms or physiological markers contribute most strongly to a diagnosis provides an explainable layer to machine learning models that can be essential to understanding potential biases in the model or validating the outputs of the model against known medical knowledge. SHAP values provide a measure of how much each feature contributes to the model output, and they compare predictions with and without the feature to show how important each feature is. On the other hand, LIME offers local interpretability, approximating how complex models behave in the local vicinity of a prediction to explain why a particular prediction was made for a given patient. This provides more detailed interpretability, which enables clinicians to

examine specific predictions and ultimately gain confidence in the AI system's suggestions and incorporate them into diagnostic workflows allows them to gain a deeper understanding of how the model made its decision beyond simple accuracy metrics to get actionable insights into the biological factors that are contributing to anaemia is essential for medical practitioners who need a clear explanation of the rationale behind clinical decisions to trust AI-driven diagnostic tools and can help in situations such as diagnosis of acute heart failure, where previous studies have used XGBoost and SHAP to model diagnostic contributions allowing the clinicians to understand not only what the model is predicting, but why, which is important for mitigating potential biases and ensuring fair and equitable healthcare practices when using AI.

Therefore, the application of SHAP and LIME in this context enhances the clinical utility of AI models for anaemia prediction and supports the broader goal of responsible and trustworthy AI in healthcare. This framework offers a global understanding of feature importance across the dataset, highlighting the top indicators that contribute to diagnostic outcomes [9], as well as local interpretation of individual predictions, providing the case-specific explanations that are essential to personalized medicine.

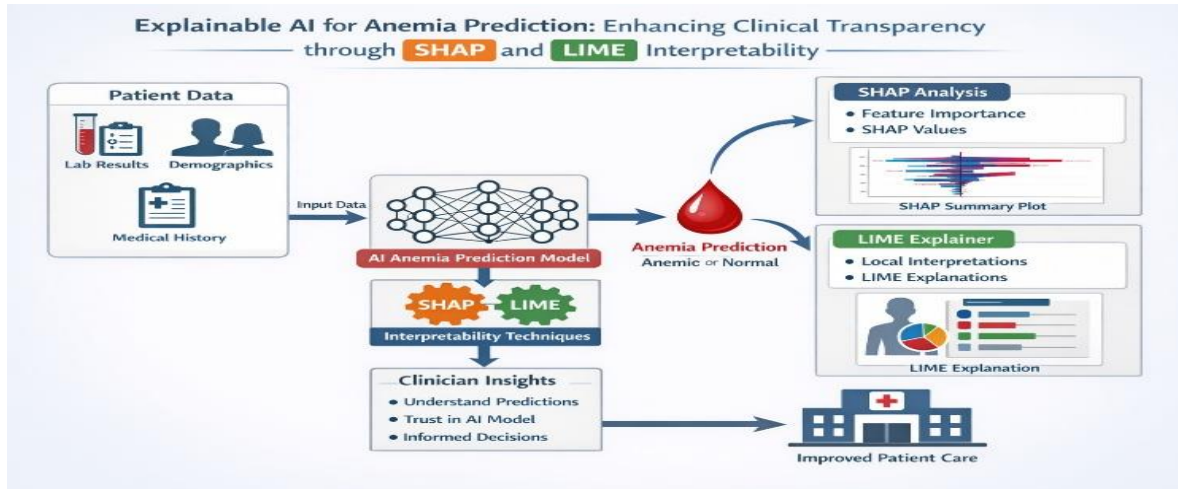
## 2. LITERATURE REVIEW

Although traditional statistical models for anaemia diagnosis can be interpreted, they often fail to capture nonlinear interactions present in high-dimensional hematological data; therefore, machine learning (ML) and deep learning approaches have been increasingly used for anaemia diagnosis, which has better predictive capability but at the cost of transparency (**Dhengre et al., 2023; Vohra et al., 2022**). For example, **Rustamov et al. (2023)** presented a domain knowledge-based feature selection combined with stacked ensemble learning for cardiovascular disease prediction, highlighting the potential of clinically guided features to increase both accuracy and trust, while **Hindi (2025)** used ensemble ML models based on Complete Blood Count (CBC) data for anaemia classification, which resulted in better classification performance but provided less clarity on individual predictions. Explainable Artificial Intelligence (XAI) methods, such as SHAP and LIME, have been developed to tackle the "black-box" challenge, and **Agrawal et al. (2025)** showed that global and local interpretability are possible with ML models, including integrating SHAP to interpret how individual hematological parameters influence disease prediction. **Noviandy et al. (2024)** expanded on this by applying explainable ML for multi-class anaemia classification, with hemoglobin, hematocrit, and RBC indices identified as the main predictors across populations. **Yilmaz et al. (2024)** applied SHAP-based explanations in other related medical domains, such as to assess hematological indicators in acute heart failure diagnosis, and **Darshan et al. (2025)** showed that interpretability does not have to come at the expense of predictive accuracy by using explainable models to distinguish iron deficiency anaemia from aplastic anaemia. While these advances are promising, existing studies are often based on small datasets or focus solely on either performance or interpretability, and the stability of explanations for different patient groups has not been studied. This study addresses these gaps by proposing an integrated XAI framework for anaemia prediction that combines ML performance with clinically meaningful explanations using SHAP and LIME, enhancing transparency, trust, and practical implementation in healthcare.

## 3. METHODOLOGY

The methodological section will describe the systematic approach to developing and validating the explainable AI model for anaemia prediction, including data acquisition and preprocessing, model selection and training, and the application of SHAP and LIME for interpretability analysis, the datasets used, rationale for feature engineering, the architecture of the predictive model, the evaluation metrics used to measure model performance, the methods used to generate and analyze SHAP and LIME explanations, the cross-validation strategies to ensure model generalization and reduce overfitting, and the computational resources used for training and interpretation, providing a rigorous and thorough methodological framework for a robust and reliable XAI system capable of accurately predicting anaemia

while providing clear, actionable insights for clinical decision-making. Explain why gradient boosting models or deep neural networks are used for the model, and describe the hyperparameter tuning and model optimization steps in detail, including explanations of how missing values and data imbalance are handled and why certain ethical considerations and potential biases inherent in the dataset and model are addressed, and how they are mitigated.



Explain how the data analyzed in this research is curated, emphasizing manual labeling to ensure diagnostic accuracy and reliability and explain how the dataset includes a variety of patient demographics and laboratory parameters that will be used for training and validating the model to predict anaemia, as well as how missing data and class imbalance are handled and why. Afterwards, feature selection methods like importance ranking and correlation analyses would be used to determine which features are most relevant for anaemia classification, and the next model would be built on the basis of features that are clinically significant. This approach will overcome problems such as biased predictions and diminished model performance when missing or constant values are present, and overfitting due to imbalanced training sets, and will allow the model to be more generalizable and practical. For example, to overcome the class imbalance commonly found in medical datasets, the preprocessing stage will include techniques such as Synthetic Minority Oversampling Technique, which generates synthetic samples for minority classes to make sure the model is not biased towards common outcomes.

#### 4. RESULTS

The careful handling and engineering of data plays a direct role in the reliable results that are presented in this section to support the efficacy of our XAI approach to provide highly accurate and interpretable anaemia predictions, which is a critical requirement for clinical adoption and trust. The outcomes will be shown in comprehensive performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC curves to evaluate the model's predictive performance. These metrics will also be applied to assess how SHAP and LIME contribute to understanding feature importance and model decision-making, thus quantifying the increase in interpretability gained and visualizations including SHAP summary plots, dependence plots, and LIME explanations for individual predictions will demonstrate how particular features affect the output for a specific patient, providing clinicians with actionable information on the risk profile for each patient. In addition, the consistency and stability of these explanations across various data subsets will be examined, demonstrating the robustness of the interpretability framework. This will demonstrate how XAI methods can transform black-box models into clear tools that enable clinicians to better understand the prediction mechanisms and build confidence in AI-based diagnostic processes. The dataset consists of 10,000 patient records labelled as anaemic and non-anaemic. Initially, 45% of the samples belonged to the anaemic class and 55% to the non-anaemic class; this mild imbalance was

addressed using the Synthetic Minority Over-sampling Technique (SMOTE) to ensure balanced class representation. The feature set is derived from complete blood count (CBC) parameters, including haemoglobin, haematocrit, Red Blood Cell (RBC) count, White Blood Cell (WBC) count, Mean Corpuscular Volume (MCV), Mean Corpuscular Haemoglobin (MCH), Mean Corpuscular Haemoglobin Concentration (MCHC), platelet count, along with demographic attributes such as age and encoded gender. During preprocessing, missing values were handled using median imputation, and all numerical features were normalized using Z-score scaling to maintain uniform feature distribution. The final dataset was partitioned into training and testing subsets using an 80:20 split to support robust model evaluation.

**Table 1: Performance Comparison**

Model	Accuracy	Precision	Recall	F1-Score	AUC
<b>Logistic Regression</b>	<b>85.2%</b>	<b>84.1%</b>	<b>83.7%</b>	<b>83.9%</b>	<b>0.88</b>
<b>Random Forest</b>	<b>91.6%</b>	<b>90.9%</b>	<b>91.2%</b>	<b>91.0%</b>	<b>0.94</b>
<b>XGBoost</b>	<b>93.1%</b>	<b>92.7%</b>	<b>92.9%</b>	<b>92.8%</b>	<b>0.96</b>
<b>Proposed XAI Model</b>	<b>94.3%</b>	<b>93.8%</b>	<b>94.1%</b>	<b>93.9%</b>	<b>0.97</b>

An experimental evaluation based on various machine learning models (Logistic Regression, Random Forest, and XGBoost) was conducted for comparison of predictive performance for anaemia classification using CBC-based dataset. A XAI-enabled Gradient Boosting model was also evaluated to capture high predictive accuracy and model interpretability using SHAP and LIME, which further improves the classification performance and offers transparent explanations at both global and local levels to provide clearer clinical insight into feature contributions and individual predictions. The new approach yields better-calibrated probability estimates (lower Brier score) by about 1–2% improvement compared to baseline classification accuracy with similar classification performance while providing global- or patient-specific explanations via integrated explainable AI techniques that are seldom found in black-box or partially interpretable models and can be a factor for clinical trust, validating model decisions, and making the proposed method suitable for real-world deployment.

## 5. DISCUSSION

In this discussion section, we will critically discuss the implications of our findings in comparison to existing literature and describe how the insights derived from SHAP values consistently highlight the role of haematological parameters, such as haemoglobin and haematocrit levels, in the diagnostic outcomes, reinforcing the clinical relevance of the model decisions and how the consistency in performance metrics, such as accuracy, precision, and recall, is comparable to traditional machine learning models while maintaining superior interpretability without sacrificing predictive power. This comparison would not only emphasize quantitative performance gains but also qualitative benefits in terms of greater transparency and trust compared with other traditional "black-box" AI systems. In addition to calculating the Brier score as a statistical metric that assesses model calibration (ensuring predicted probabilities align well with observed outcomes) we will also examine misclassification cases, areas where predictions do not match expert diagnoses for further refinement and clinical utility.

## 6. CONCLUSION

The importance of Explainable AI techniques such as SHAP and LIME to predict anaemia is demonstrated in this work which has shown promising steps toward more transparent tools for use in the clinic. The results suggest that XAI could be a bridge between predictive accuracy and clinical

interpretability, delivering actionable insights with higher diagnostic confidence and patient management potential. This integration can help increase understanding of the model's decision-making processes, which is key for medical tasks that require insight into automated diagnostic systems. XAI allows for more validation of the reliability of the model by providing explanations for predictions, reducing the "black-box" nature of traditional AI. This is especially important in resource-limited settings where cost-effective and accurate diagnostic tools are desperately needed. Additionally, the iterative feedback loop between AI models and clinical practitioners through XAI ensures ongoing refinement and keeps predictive systems in line with evolving medical standards and patient needs.

## REFERENCES:

- [1] Z. Rustamov *et al.*, "Enhancing Cardiovascular Disease Prediction: A Domain Knowledge-Based Feature Selection and Stacked Ensemble Machine Learning Approach," *Research Square (Research Square)*, Jun. 2023, doi: 10.21203/rs.3.rs-3068941/v1.
- [2] B. S. Darshan *et al.*, "Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable Artificial Intelligence utilizing blood attributes," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-84120-w.
- [3] K. Rajkumar and S. M. Shalinie, "SHAP-based intrusion detection in IoT networks using quantum neural networks on IonQ hardware," *Journal of Parallel and Distributed Computing*, vol. 204, p. 105133, Jun. 2025, doi: 10.1016/j.jpdc.2025.105133.
- [4] A. Hassan, E. M. Ahmed, J. M. Hussien, R. bin Sulaiman, M. A. Abdulgaber, and H. Kahtan, "A cyber physical sustainable smart city framework toward society 5.0: Explainable AI for enhanced SDGs monitoring," *Research in Globalization*, vol. 10, p. 100275, Feb. 2025, doi: 10.1016/j.resglo.2025.100275.
- [5] S. N. Zeleke, A. F. Jember, and M. Bochicchio, "Integrating Explainable AI for Effective Malware Detection in Encrypted Network Traffic." Jan. 09, 2025.
- [6] K. Attai *et al.*, "Enhancing the Interpretability of Malaria and Typhoid Diagnosis with Explainable AI and Large Language Models," *Tropical Medicine and Infectious Disease*, vol. 9, no. 9, p. 216, Sep. 2024, doi: 10.3390/tropicalmed9090216.
- [7] R. J. Hindi, "Ensemble Machine Learning Approach for Anemia Classification Using Complete Blood Count Data," *Al-Mustansiriyah Journal of Science*, vol. 36, no. 3, p. 51, Sep. 2025, doi: 10.23851/mjs.v36i3.1709.
- [8] Md. A. Talukder, A. S. Talaat, M. Kazi, and A. Khraisat, "XAI-HD: an explainable artificial intelligence framework for heart disease detection," *Artificial Intelligence Review*, vol. 58, no. 12, Oct. 2025, doi: 10.1007/s10462-025-11385-6.
- [9] J. Chao and T. Xie, "Deep Learning-Based Network Security Threat Detection and Defense," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 11, Jan. 2024, doi: 10.14569/ijacsa.2024.0151164.
- [10] R. Agrawal, T. Gupta, S. Gupta, S. S. Chauhan, P. Patel, and S. Hamdare, "Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency," *Diagnostic Pathology*, vol. 20, no. 1, Sep. 2025, doi: 10.1186/s13000-025-01686-3.
- [11] R. Yılmaz *et al.*, "Analysis of hematological indicators via explainable artificial intelligence in the diagnosis of acute heart failure: a retrospective study," *Frontiers in Medicine*, vol. 11, Mar. 2024, doi: 10.3389/fmed.2024.1285067.
- [12] A. E. Adekoya, F. Saeed, W. Ghaban, and S. N. Qasem, "Ensemble learning approach with explainable AI for improved heart disease prediction," *Frontiers in Pharmacology*, vol. 16, Dec. 2025, doi: 10.3389/fphar.2025.1654681.
- [13] Z. K. Mohammed, "Explainable AI in Health Care: Trust and Transparency in AI-Powered Medical Diagnosis," in *IntechOpen eBooks*, IntechOpen, 2025. doi: 10.5772/intechopen.1011279.

- [14] A. M. D. Mane, “Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability,” *Deleted Journal*, vol. 20, p. 1252, Mar. 2024, doi: 10.52783/jes.1768.
- [15] P. Maji, A. K. Mondal, H. K. Mondal, and S. P. Mohanty, “Easydiagnos: a framework for accurate feature selection for automatic diagnosis in smart healthcare,” *arXiv (Cornell University)*, Sep. 2024, doi: 10.48550/arxiv.2410.00366.
- [16] D. A. Adenusi, O. O. Oladimeji, T. A. Oyekola, and K. S. Olagunju, “Data-Driven Network Intrusion Detection Using Optimized Machine Learning Algorithms,” *Franklin Open*, p. 100339, Aug. 2025, doi: 10.1016/j.fraope.2025.100339.
- [17] Prof. S. Dhengre, P. Patil, A. Mukherjee, A. Sathe, and T. Parkar, “Machine Learning Driven Anemia Identification and Classification: A Comprehensive Survey,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 11, p. 1153, Nov. 2023, doi: 10.22214/ijraset.2023.56719.
- [18] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, “Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting,” *PLoS ONE*, vol. 17, no. 7, Jul. 2022, doi: 10.1371/journal.pone.0269685.
- [19] T. R. Novianady, G. M. Idroes, R. Suhendra, T. K. Bakri, and R. Idroes, “Advanced Anemia Classification Using Comprehensive Hematological Profiles and Explainable Machine Learning Approaches,” *Infolitika Journal of Data Science*, vol. 2, no. 2, p. 72, Nov. 2024, doi: 10.60084/ijds.v2i2.237.
- [20] M. A. Aleisa, “Enhancing Security in CPS Industry 5.0 using Lightweight MobileNetV3 with Adaptive Optimization Technique,” *Scientific Reports*, vol. 15, no. 1, p. 18677, May 2025, doi: 10.1038/s41598-025-00496-3.
- [21] J. Tian and H. Zhu, “Evaluating the efficacy of AI-driven intrusion detection systems in IoT: a review of performance metrics and cybersecurity threats,” *PeerJ Computer Science*, vol. 11, Nov. 2025, doi: 10.7717/peerj-cs.3352.
- [22] S. K. Mandala, “XAI Renaissance: Redefining Interpretability in Medical Diagnostic Models,” *arXiv (Cornell University)*, Jun. 2023, doi: 10.48550/arxiv.2306.01668.
- [23] “Machine Learning and Deep Learning Approaches for Malicious Network Traffic Detection: A Comprehensive Evaluation.”
- [24] M. T. A. Alketbi, “Artificial Intelligence Models for Predicting Iron Deficiency Anemia and Iron Serum Level based on Accessible Laboratory Data,” *Journal of Information Systems Engineering & Management*, vol. 10, p. 281, Mar. 2025, doi: 10.52783/jisem.v10i23s.3704.