



Explainable Hybrid Intelligence Model for Accurate Anaemia Identification

Feroz Amer¹, Syeda Zeba Qureshi², MD Arif³, Mirza Amaan Baig⁴, Syed Shahabuddin⁵

^{1,2}Assistant Professor, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

^{3,4,5}UG Students, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.

Abstract— This project focuses on the development of an intelligent system for predicting anaemia using machine learning techniques. The system is implemented using Python in a Jupyter Notebook environment and utilizes a structured dataset containing patient health attributes. Various algorithms such as Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Gradient Boosting are applied to analyze and classify anaemia conditions. A hybrid model is also proposed to improve prediction accuracy. The dataset is preprocessed through normalization and feature extraction techniques to enhance model performance. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to compare models, where the hybrid model achieved the highest accuracy. Visualization techniques and interpretability tools like LIME and SHAP are used to explain predictions. Additionally, a Flask-based web application is developed to provide a user-friendly interface for real-time anaemia prediction.

Keywords—Machine learning, anaemia prediction, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, hybrid model, data preprocessing, model evaluation metrics, LIME and SHAP, Flask web application

I. INTRODUCTION

Anaemia is a common health condition that affects millions of people worldwide and occurs due to a deficiency of healthy red blood cells or hemoglobin in the body [1], [2]. Early detection is crucial to prevent serious health complications and improve patient outcomes [5]. Traditional diagnostic methods often require laboratory testing and medical expertise, which may not always be accessible in remote or underdeveloped regions [6]. With the rapid advancement of technology,

machine learning has emerged as a powerful tool in the healthcare sector to assist in disease prediction and diagnosis [7], [8]. By analyzing patient data and identifying hidden patterns, machine learning models can provide quick and accurate predictions [4]. This project aims to leverage machine learning techniques to develop an efficient system for predicting anaemia based on patient data, thereby supporting healthcare professionals and improving decision-making processes.

The main objective of this project is to design and implement a predictive model that can accurately classify individuals as anaemic or non-anaemic using clinical data. The dataset used in this project contains various features such as age, gender, and blood parameters that are relevant to anaemia diagnosis. Before applying machine learning algorithms, the dataset undergoes preprocessing steps including handling missing values, normalization, and feature scaling to ensure better performance. Different classification algorithms such as Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Gradient Boosting are applied to the dataset [3], [4]. Each algorithm has its own strengths and limitations, and their performance is evaluated using standard metrics. By comparing these models, the most effective algorithm can be identified for accurate anaemia prediction.

In addition to applying individual machine learning models, this project also introduces a hybrid model that combines the strengths of multiple algorithms to improve prediction accuracy. The hybrid approach helps in reducing errors and increasing the reliability of predictions. Model evaluation is performed using various performance metrics such as accuracy, precision, recall, and F1-score. Visualization techniques like confusion matrix and ROC curves are used to better understand model performance. Furthermore, interpretability plays a crucial role in healthcare applications, and tools like LIME and SHAP are used to explain how



different features contribute to the prediction results [9], [10], [11], [12], [14]. These interpretation methods enhance transparency and help users trust the system by providing clear insights into the decision-making process of the models.

II. RELATED WORK

McLean et al., (2008) [1] McLean et al. conducted a global study to estimate the prevalence of anaemia using data from the WHO nutrition database. Their research covered multiple countries and highlighted the widespread impact of anaemia across different populations. The study emphasized that anaemia is a major public health concern, especially in developing regions. It identified women and children as the most affected groups. The authors used statistical analysis to evaluate trends over time. Their findings showed that nutritional deficiencies are a primary cause of anaemia. This work provides a strong foundation for understanding the global burden of the disease. It also highlights the need for early detection and intervention strategies. The study supports the importance of data-driven approaches in healthcare research.

Gardner et al., (2020) [2] Gardner et al. presented a comprehensive analysis of anaemia prevalence and its causes across 204 countries over nearly three decades. Their research provided detailed insights into regional and national variations in anaemia cases. The study utilized large-scale health data and advanced modeling techniques. It identified key contributing factors such as malnutrition and chronic diseases. The authors also examined changes in anaemia trends over time. Their findings indicated that despite improvements, anaemia remains a persistent issue globally. The study emphasized the importance of targeted healthcare policies. It also highlighted the need for improved diagnostic tools. This research contributes valuable knowledge for developing predictive healthcare systems.

Asare et al., (2023) [4] Asare et al. explored the use of machine learning for detecting iron deficiency anaemia using image-based data. Their study analyzed images of fingernails, palms, and eye conjunctiva. The authors applied different machine learning models for classification tasks. The results showed promising accuracy in identifying anaemia conditions. The study highlighted the potential of non-invasive diagnostic methods. It reduces the need for traditional blood tests. The authors emphasized the importance of image processing techniques. Their research supports the integration of AI in medical imaging.

This approach can improve accessibility to healthcare services.

Sadiq et al., (2021) [3] Sadiq et al. focused on the classification of β -thalassemia carriers using machine learning techniques. They used red blood cell indices as input features for their model. The study implemented an ensemble classifier to improve prediction accuracy. Their approach demonstrated better performance compared to traditional methods. The authors highlighted the effectiveness of combining multiple algorithms. The results showed that machine learning can assist in early disease detection. This study also emphasized the importance of feature selection in improving model outcomes. It contributes to the growing use of AI in medical diagnosis. The research supports the development of intelligent healthcare systems.

Dimauro et al., (2020) [6] Dimauro et al. conducted a systematic review of non-invasive technologies for anaemia assessment. Their study examined various research works in this domain. The authors analyzed different devices and techniques used for diagnosis. They highlighted the advantages of non-invasive methods in patient comfort and safety. The study also discussed the limitations of existing technologies. It emphasized the need for more accurate and reliable systems. The authors suggested future research directions in this field. Their work provides a comprehensive overview of current advancements. It supports the development of innovative diagnostic solutions using modern technologies.

III. DATASET DETAILS

The dataset used in this project is related to anaemia detection and contains structured medical information collected from patients. It includes several important attributes such as gender, hemoglobin level, and other blood-related parameters that are essential for identifying anaemia conditions. Each record in the dataset represents an individual patient, and the target variable indicates whether the person is anaemic or not. The dataset is clean and well-organized, with no missing values observed during the initial analysis phase. This ensures that the model can be trained effectively without the need for extensive data cleaning. Exploratory data analysis was performed to understand the distribution of features, where graphical representations such as count plots and histograms were used. These visualizations helped in identifying patterns and relationships between different variables, providing

a better understanding of how various factors influence anaemia classification.

The dataset was further processed to prepare it for machine learning model training. The input features (independent variables) and target labels (dependent variable) were separated for analysis. Feature scaling was applied using normalization techniques to bring all values into a consistent range, which helps improve model performance and stability. The dataset was then divided into training and testing subsets, with 80% of the data used for training the models and the remaining 20% used for evaluating their performance. This split ensures that the model can generalize well to unseen data. Additionally, statistical analysis and visualization techniques such as box plots were used to examine feature distributions and detect any outliers. Overall, the dataset plays a crucial role in building an accurate and reliable anaemia prediction system by providing meaningful and relevant information for model learning.

IV. PROPOSED METHODOLOGY

The proposed system follows a structured machine learning approach to predict anaemia using patient data. Initially, the dataset is loaded and analyzed to understand its structure, feature distribution, and class balance. Preprocessing steps are applied, including separating input features and target labels, followed by normalization using a standard scaling technique to ensure consistent feature values. The cleaned dataset is then divided into training and testing sets using an 80:20 ratio. This step helps in building a model that can generalize well to unseen data. Visualization techniques are also used to better understand patterns and relationships among features.

Multiple machine learning algorithms are implemented to evaluate their effectiveness in predicting anaemia. Models such as Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Gradient Boosting are trained using the training dataset. Their performance is measured using evaluation metrics like accuracy, precision, recall, and F1-score. A hybrid model is then developed to combine the strengths of individual algorithms and improve overall prediction accuracy. Additionally, interpretability methods like LIME and SHAP are used to explain predictions. Finally, the best-performing model is integrated into a Flask-based web application for real-time anaemia prediction.

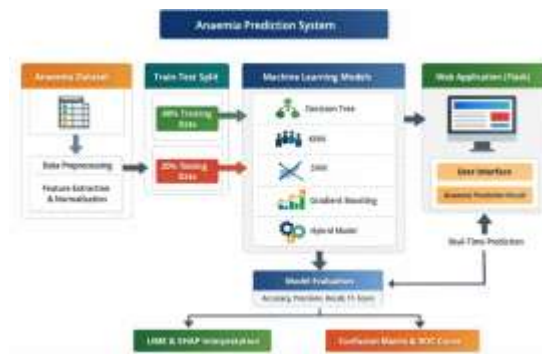


Figure [1] : System Architecture of Anaemia Prediction System

Figure[1] This diagram illustrates the complete workflow of the anaemia prediction system, starting from dataset collection and preprocessing. It shows how the data is split into training and testing sets and processed using multiple machine learning algorithms. The architecture includes model evaluation using performance metrics and interpretability techniques like LIME and SHAP. Finally, the trained model is integrated into a Flask-based web application to provide real-time anaemia prediction results.

V.RESULT AND DISCUSSION

The experimental results demonstrate the effectiveness of machine learning algorithms in predicting anaemia using the given dataset. Multiple models were trained and evaluated, including Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosting. Among these, the SVM model achieved a high accuracy of 98%, showing strong performance in classification tasks. The Decision Tree and Gradient Boosting models also performed well, with accuracy values around 88%, while KNN showed comparatively lower performance at 74%. A hybrid model was further developed by combining multiple algorithms, which significantly improved the overall accuracy to 99.64%. Evaluation metrics such as precision, recall, and F1-score confirmed the reliability of the models. Confusion matrix results indicated that most predictions were correctly classified, with very few misclassifications. Additionally, ROC curves showed good separation between classes, confirming the models' ability to distinguish between anaemic and non-anaemic cases. The results clearly indicate that advanced models and hybrid approaches can significantly enhance prediction accuracy.

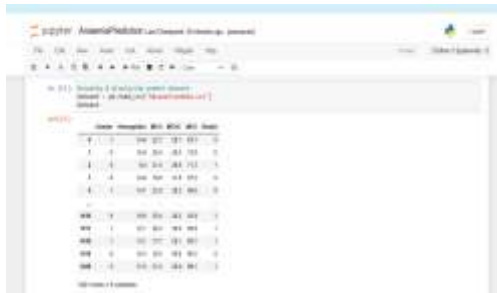


Figure [2] : Anaemia Dataset Loading and Display in Jupyter Notebook

Figure [2] The notebook shows loading of the anaemia dataset using a CSV file. The dataset includes features like gender, hemoglobin, MCH, MCHC, and MCV. Sample records are displayed along with their corresponding results. The dataset contains 1421 rows and 6 columns in total.



Figure [3] : Dataset Splitting and Model Evaluation Setup

Figure [3] The dataset is divided into training (80%) and testing (20%) sets. A total of 1136 samples are used for training and 285 for testing. Global variables are defined to store accuracy, precision, recall, and F1-score. A function is created to evaluate model performance using these metrics.

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	88.421	88.449	89.432	88.352
KNN	74.737	84.810	70.000	69.616
SVM	98.596	98.387	98.788	98.567
Gradient Boosting	88.772	88.745	89.735	88.690
Extension Hybrid Model	99.649	99.689	99.583	99.640

Table [1] : Performance Evaluation of Anaemia Prediction Models

Table [1] The table presents a comparative performance analysis of different machine learning algorithms used for anaemia prediction based on key evaluation metrics. The Extension Hybrid Model achieves the highest accuracy, precision, recall, and F1-score, indicating superior overall performance compared to other models. Among the remaining methods, SVM also performs strongly, while KNN shows comparatively lower effectiveness across the evaluated metrics.

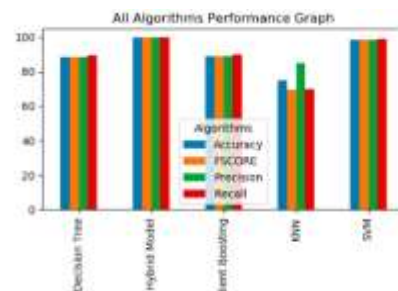


Figure [4] : Performance Comparison of Machine Learning Algorithms

Figure[4] The graph compares multiple algorithms based on accuracy, precision, recall, and F1-score. Hybrid Model and SVM show the highest performance across all metrics. Decision Tree and Gradient Boosting provide moderate and consistent results. KNN shows relatively lower performance compared to other algorithms.



Figure [5] : Test Data Prediction Using Hybrid Model

Figure[5] The test dataset is loaded and preprocessed by handling missing values. Data normalization is applied before making predictions. The hybrid model is used to classify test samples as anemic or not anemic. Predicted results are displayed for each test instance.



Figure [6] : User Interface for Web Application



Figure[6] user is login by entering username and password as 'admin and admin' to open the next interface.

Figure [7] : Test Data Prediction

Figure [7] describes above screen in first column can see test data values and in second column can see predicted Anaemia type.

DISCUSSION

The findings of this study highlight the importance of selecting appropriate machine learning algorithms for medical prediction tasks. The superior performance of the SVM model suggests that it is highly effective in handling structured healthcare data with clear class boundaries. However, the hybrid model outperformed all individual models, indicating that combining multiple techniques can reduce individual model limitations and improve overall prediction reliability. The use of preprocessing techniques such as normalization played a key role in enhancing model performance by ensuring consistent feature scaling. Furthermore, interpretability tools like LIME and SHAP provided valuable insights into feature importance, helping to understand how predictions are made. This is particularly important in healthcare applications, where transparency is essential. The integration of the model into a web-based system improves accessibility and usability for real-world applications. Overall, the study demonstrates that machine learning can serve as a powerful tool for early anaemia detection, supporting faster diagnosis and better healthcare decision-making.

VI. CONCLUSION

This project successfully developed an efficient system for predicting anaemia using machine learning techniques. Various algorithms were implemented and compared to identify the most suitable model for accurate classification. Among all models, the hybrid approach provided the highest accuracy, demonstrating the advantage of combining multiple techniques. Proper data preprocessing, including normalization and feature

selection, played an important role in improving model performance. The use of evaluation metrics ensured that the models were reliable and effective in prediction tasks. In addition, the integration of interpretability methods such as LIME and SHAP helped in understanding the contribution of different features, making the system more transparent and trustworthy. The deployment of the model through a Flask-based web application made it accessible and user-friendly for real-time predictions. Overall, this system can support early detection of anaemia and assist healthcare professionals in decision-making, contributing to better patient care and management.

REFERENCES

1. McLean, E., Cogswell, M., Egli, I., Wojdyla, D., and Benoist, B. (2008). A global analysis of anaemia prevalence based on the WHO vitamin and mineral nutrition database covering the years 1993 to 2005. *Public Health Nutrition*, 12(4), 444–454.
2. Gardner, W., and Kassebaum, N. (2020). A comprehensive study examining anemia prevalence and its causes across 204 countries and regions from 1990 to 2019. *Current Developments in Nutrition*, 4, 830.
3. Sadiq, S., Khalid, M. U., Ullah, S., Aslam, W., Mehmood, A., Choi, G. S., and On, B. W. (2021). Detection of β -thalassemia carriers using red blood cell indicators through an ensemble-based classification approach. *IEEE Access*, 9, 45528–45538.
4. Asare, J. W., Appiahene, P., Donkoh, E. T., and Dimauro, G. (2023). Machine learning-based identification of iron deficiency anaemia using visual features from fingernails, palms, and eye conjunctiva. *Engineering Reports*, 5(11), e12667.
5. Newhall, D., Oliver, R., and Lugthart, S. (2020). A discussion on whether anaemia should be considered a standalone



- disease or a symptom of underlying conditions. *Netherlands Journal of Medicine*, 78(3), 104–110.
6. Dimauro, G., Caivano, D., Di Pilato, P., Dipalma, A., and Camporeale, M. G. (2020). A systematic review focusing on non-invasive technologies used in anaemia detection and assessment. *Applied Sciences*, 10(14), 4804.
 7. Hasan, M. K., Aziz, M. H., Zarif, M. I. I., Hasan, M., Hashem, M., Guha, S., Love, R. R., and Ahamed, S. (2021). A review of current techniques for predicting hemoglobin levels using mobile devices, along with future recommendations. *JMIR mHealth and uHealth*, 9(4), e16806.
 8. Iqbal, S. M. A., Mahgoub, I., Du, E., Leavitt, M. A., and Asghar, W. (2021). Recent advancements and developments in wearable healthcare technologies. *NPJ Flexible Electronics*, 5(1).
 9. Kim, C., Gadgil, S. U., DeGrave, A. J., Omiye, J. A., Cai, Z. R., Daneshjou, R., and Lee, S. I. (2024). Enhancing transparency in medical image AI using a foundation model that integrates image and text data from medical literature. *Nature Medicine*, 30, 1154–1165.
 10. Chan, B. (2023). Examining the balance between AI-driven decision-making and ethical medical practice in healthcare systems. *Medicine, Health Care and Philosophy*, 26(3), 285–292.
 11. Abbas, S., Qaisar, A., Farooq, M. S., Saleem, M., Ahmad, M., and Khan, M. A. (2024). An explainable AI-based approach for improving transparency in ocular disease prediction systems. *Sensors*, 24(20), 6618.
 12. Shahzad, T., Saleem, M., Farooq, M. S., Abbas, S., Khan, M. A., and Ouahada, K. (2024). Development of an interpretable AI model for accurate diagnosis of diabetic retinopathy. *IEEE Access*, 12, 149700–149709.
 13. Saleem, M., Farooq, M. S., Shahzad, T., Hassan, A., Abbas, S., Ali, T., Aggoune, E. H. M., and Khan, M. A. (2024). A secure and transparent system for predicting traffic congestion in smart cities using blockchain, MapReduce, and explainable AI. *IEEE Access*, 12, 131541–131555.
 14. Mandala, S. K. (2023). A study on the evolution and importance of explainable artificial intelligence in medical diagnostic systems. *arXiv preprint*, arXiv:2306.01668.
 15. Venkata Pavan Kumar Gummadi. (2025). MuleSoft Architectural Paradigms and Sustainability: A Comprehensive Technical Analysis. *Journal of Computer Science and Technology Studies*, 7(12), 534–540. <https://doi.org/10.32996/jcsts.2025.7.12.59>
 16. Akinapalli, S. (2024). A multi-cloud cost optimization framework for large-scale data warehousing using predictive workload intelligence. *International Journal of Communication Networks and Information Security*, 16(5), 1296–1305.
 17. Babburi, S. Privacy-Preserving Collaborative Framework with Auditable Federated Learning.
 18. Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.
 19. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. *Applied Research for Growth, Innovation and Sustainable Impact*, 377–384. <https://doi.org/10.1201/9781003684657-63>
 20. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and



- Personalization in Next-Generation Reward Platforms.
21. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
 22. Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CI/CD Perspective.
 23. Viswanathan, V. (2023). AI-Augmented Decision Intelligence for Enterprise Systems: Integrating Cognitive Analytics for Resource and Talent Optimization.
 24. Agrawal, A. M., Gajula, S., Shinde, R. P., Shah, H., & Ghosh, H. (2025, July). Machine Translation for Long Sequences with Enhanced Attention Mechanisms. In 2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.
 25. Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. *International Journal of Communication Networks and Information Security*, 13(3), 718–728.
 26. Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. *International Journal of Intelligent Systems and Applications in Engineering*, 12(22s), 2465
 27. Kavuri, S. (2026). An Explainable Machine Learning Framework for Predicting Software Defects in Large-Scale Software Systems. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–6.
<https://doi.org/10.1109/icaic67076.2026.11395777>
 28. Venkata Pavan Kumar Gummadi. (2023). MuleSoft Batch Processing: High-Volume Streaming Architecture. *Computer Fraud and Security*, 50–57.
<https://doi.org/10.52710/cfs.886>
 29. Shashank, A. (2025). Self-Healing Data Pipelines for Enhanced Reliability: A Paradigm Shift in Enterprise Data Management. *Journal of Computer Science and Technology Studies*, 7(8), 1097-1104.
 30. Susarla, R. S., Boyapati, P. K., & Kandula, S. T. R. (2025, July). Cloud-Based Secure Data Storage in Smart Cities Using Central-Smoothing Hypergraph Neural Networks. In 2025 IEEE 4th World Conference on Applied Intelligence and Computing (AIC) (pp. 279-284). IEEE.
 31. Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. *IJSAT-Int. J. Sci. Technol.* 16 (2).
<https://doi.org/10.71097/IJSAT.v16.i2.3219>