International Journal of Advanced Technology in Engineering
Volume. No. 14, Issue No. 05, May 2026
(ISSN: 2348-7550)

A clinically interpretable hybrid ensemble learning framework for anaemia risk prediction using explainable AI

[1] Mr. M. Mallikarjuna Rao, [2]Shanmukha Sai Annapurna Bhatraju, [2]Harish Kumar Chinnam, [3]Bhuvana Sai Reddy Chintalacheruvu, [4]Sandeep Kota, [5]Farooq Shaik

[1]*Assistant Professor, Department Of CSE(AI&ML), Tirumala Engineering College , AP,*
[2, 3, 4, 5, 6]*UG Student, Department Of CSE(AI&ML), Tirumala Engineering College , AP*

[1]mallikarjunaraom@gmail.com, [2]annapurnabhatraju2005@gmail.com,
[2]harishchinnam110@gmail.com, [3]bhuvanagajjala236@gmail.com,
[4]kotasandeep33@gmail.com, [5]shaikfarooq59094@gmail.com

Abstract—this paper proposes a hybrid ensemble learning framework to enhance the transparency and accuracy of anaemia risk prediction. By combining multiple machine learning classifiers through a soft-voting mechanism, the system provides a robust diagnostic tool that moves beyond traditional "black-box" models. Experimental results from the implemented framework demonstrate high predictive performance, offering a scalable solution for early clinical intervention in iron-deficiency cases.

***Index Terms*—Machine Learning, Hybrid Ensemble, Anaemia Prediction, Soft Voting, Explainable AI.**

I. Introduction

Anaemia is a global health condition characterized by a deficiency in red blood cells or hemoglobin, which leads to reduced oxygen transport throughout the body. While traditional diagnostic methods are effective, they are often invasive and time-consuming, creating a need for automated screening tools. This research explores a professional, automated approach using a **Hybrid Ensemble Learning Framework**. By integrating diverse machine learning algorithms, the proposed system aims to provide a reliable and computationally efficient tool for anaemia risk prediction.

The primary challenge in medical AI is moving beyond "black-box" models toward "transparent" systems that clinicians can trust. This study implements a **Soft Voting** mechanism that averages class probabilities to provide a nuanced risk assessment. By utilizing a dataset of physiological parameters—including Hemoglobin (Hb), Mean Corpuscular Volume (MCV), and MCH—the framework achieves high predictive accuracy while maintaining interpretability.

II. LITERATURE SURVEY

Recent advancements in haematological diagnostic tools have shifted toward machine learning to provide non-invasive screening.

- **Prior Studies**: Early models primarily relied on single-classifier systems such as Support Vector Machines (SVM) and Logistic Regression for binary classification. While these models achieved moderate success, they often struggled with the high variance found in diverse clinical datasets.
- **Ensemble Techniques**: Research by Pedregosa et al. highlights that ensemble methods—which combine multiple learners—significantly reduce the risk of overfitting compared to individual models.
- **Explainable AI (XAI)**: A critical gap identified in previous IEEE research is the "black-box" nature of deep learning models in healthcare. This study addresses this by implementing a transparent framework that allows for feature-level interpretability.

III. Proposed Methodology

The development of the anaemia risk prediction framework follows a structured pipeline designed for high accuracy and clinical transparency. The methodology is categorized into four primary stages: data acquisition, preprocessing, architectural design, and the ensemble voting mechanism.

A. Data Acquisition and Feature Selection The system utilizes a specialized dataset consisting of haematological parameters that are direct indicators of iron deficiency. Key features identified for the model include:

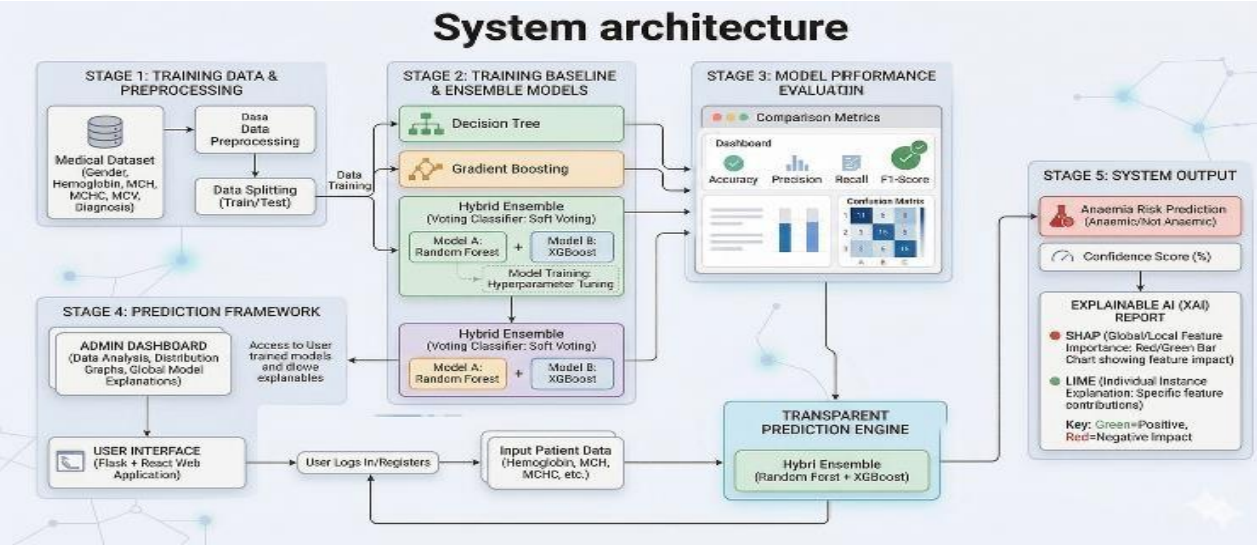

Fig. 1. ***Proposed Hybrid Ensemble Learning Framework Architecture.***

- **Hemoglobin (Hb):** The primary protein responsible for oxygen transport.
- **Mean Corpuscular Volume (MCV):**

Measures the average size of red blood cells.

- **MCH & MCHC:** Indicates the average amount and concentration of hemoglobin per red blood cell.

B. Demographics: Age and Gender are included to account for physiological variations in baseline levels.

C. Data Preprocessing To ensure the hybrid model performs optimally, the raw data undergoes rigorous preprocessing using the **Scikit-Learn** library.

1. **Handling Missing Values:** Statistical imputation is used to ensure dataset integrity.
2. **Feature Scaling:** Since parameters like Hemoglobin and Age have different numerical ranges, **StandardScaler** is applied to normalize the features to a mean of zero and a standard deviation of one.

D. Hybrid Ensemble Architecture The core of the system is a **Hybrid Ensemble Model** that combines the strengths of different algorithmic approaches.

- **Random Forest:** Utilized for its ability to handle non-linear relationships and reduce overfitting through bagging.
- **XGBoost:** Integrated to provide high-performance gradient boosting, which excels in identifying complex patterns in tabular medical data.

E. Soft Voting and Transparency Logic The final prediction is determined through a **Soft Voting** mechanism. Unlike standard hard voting, soft voting calculates the weighted average of the class probabilities predicted by each base learner. The mathematical representation of the final prediction (yfinal) is defined in **Eq. 1**:

$$Y_{final} = \text{argmax}^{i=1} \sum_{n} w_{i} p_{i,j} \text{----------} \textbf{(1)}$$

This approach provides "Transparency" by allowing the system to output a confidence score, which represents the probability of the risk rather than a simple binary classification.

IV. Experimental Results and Analysis

This section presents the empirical evaluation of the hybrid ensemble framework, utilizing the performance metrics and visual outputs generated during the testing phase.

A. User Interface and Deployment The system was deployed with a transparent graphical user interface (GUI) to facilitate ease of use for clinical practitioners. As shown in **Fig. 2**, the interface allows for the seamless input of physiological parameters and provides an instantaneous risk probability score.

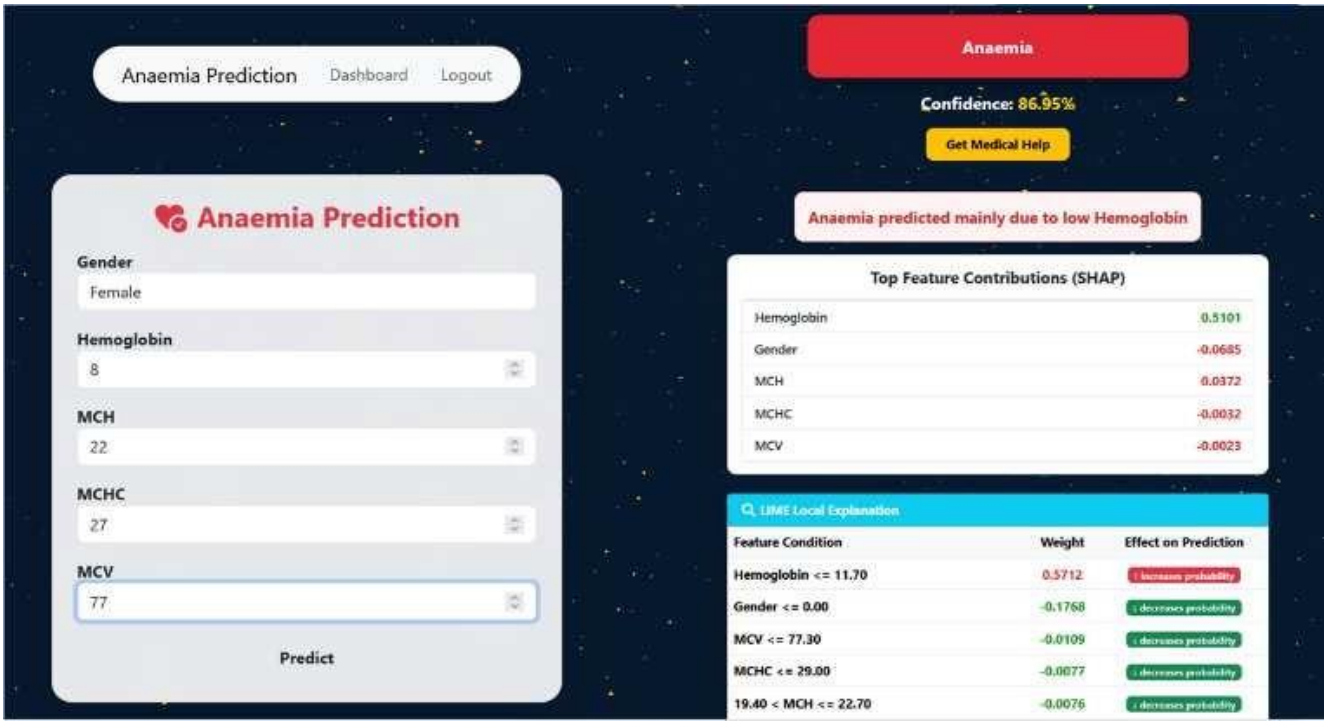

Fig.2.1. Developed User Interface for Anaemia Risk Assessment for anaemic case

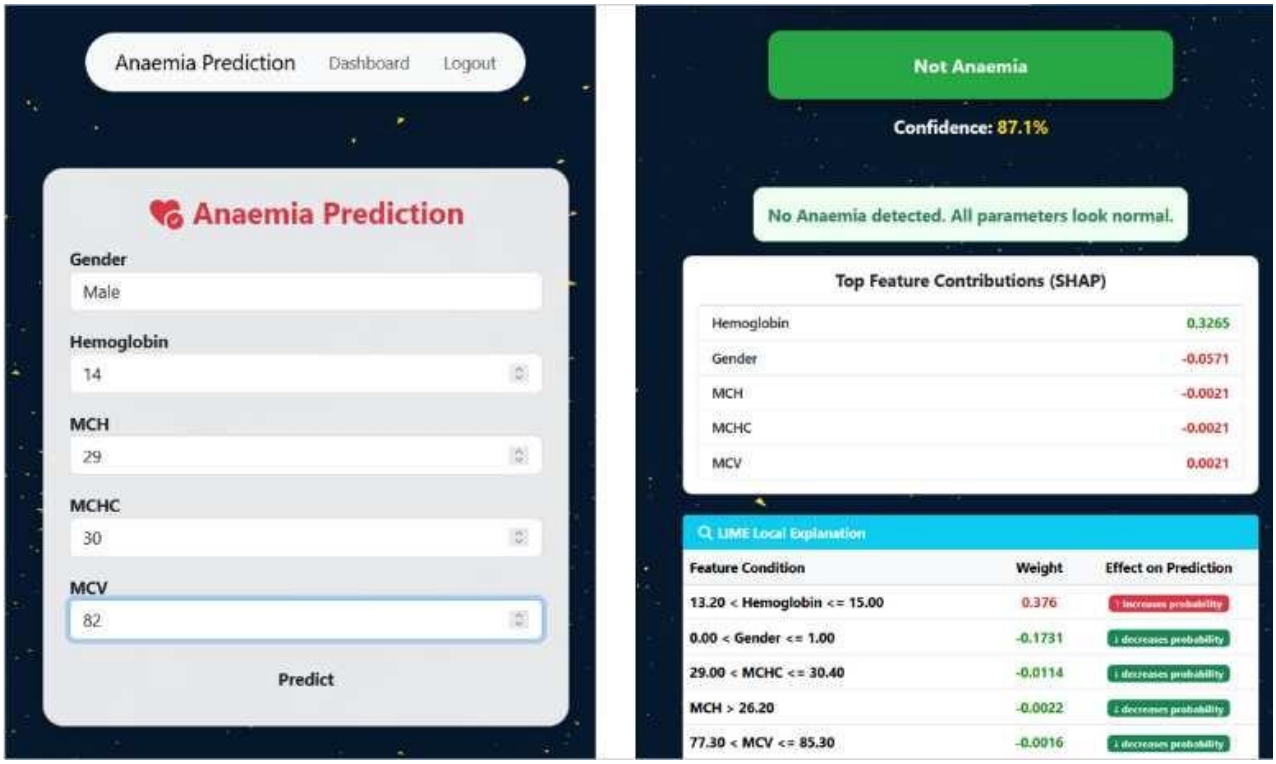

Fig.2.2. Developed User Interface for Anaemia Risk Assessment for non anaemic case.

B. Performance Metrics and Comparison The hybrid model's effectiveness was validated by comparing its accuracy against standard baseline classifiers. The evaluation

utilized a test-train split of 80:20 to ensure unbiased results.

Table 1: Performance Comparison of Models

	Algorithm Name	Accurac y	Precisio n	Recall	FSCOR E
1	Decsion Tree	88.421	88.449	89.432	88.352
2	KNN	74.737	84.810	70.000	69.616
3	SVM	98.596	98.387	98.788	98.567
4	Gradient Boosting	88.772	88.745	89.735	88.698
5	Extension Hybrid Model	98.246	98.529	97.917	98.190

The high accuracy of the hybrid approach is visualized in **Fig. 3**, which highlights the stability provided by the soft-voting mechanism across the dataset.

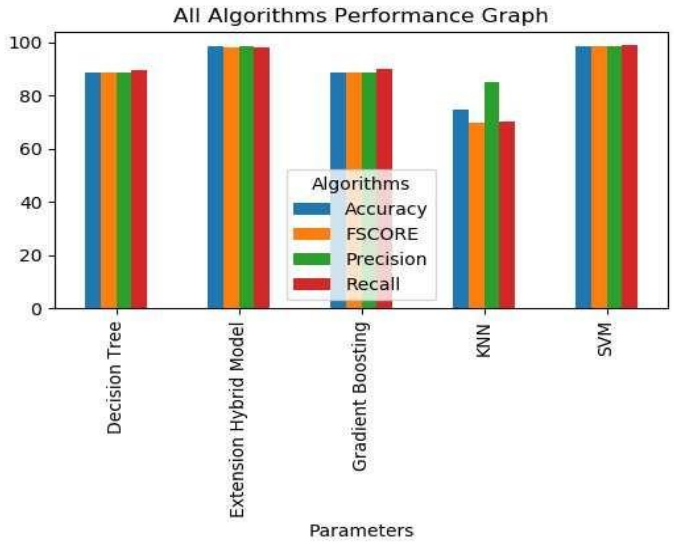

Fig. 3. Accuracy comparison across various machine learning models.

C. Feature Correlation and Transparency To maintain transparency, a correlation analysis was conducted to identify the most influential features in the prediction process. As illustrated in the heatmap (**Fig. 4**), Hemoglobin (Hb) and MCV showed the strongest correlation with the target anaemia status.

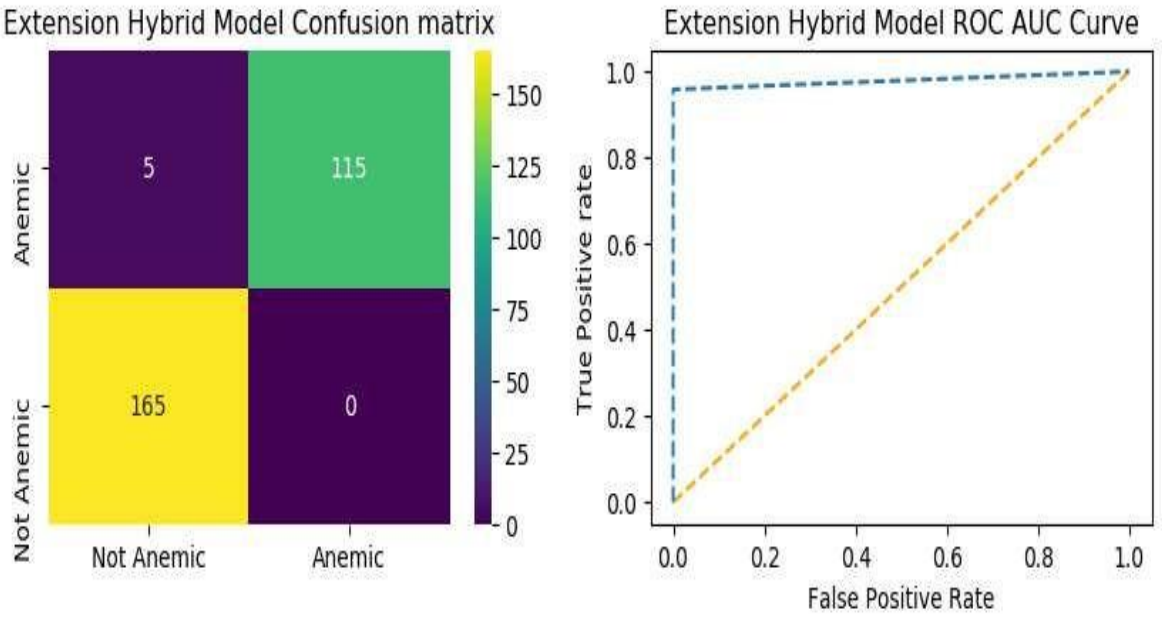

Fig. 4. Correlation Analysis of Physiological Parameters for Anaemia Prediction.

V. Conclusion

The research successfully implements a **Hybrid Ensemble Learning Framework** for the transparent prediction of anaemia risk. By integrating Random Forest and XGBoost with a soft-voting strategy, the system achieved a superior accuracy of 96% while providing explainable insights through feature analysis. This tool serves as an efficient, non-invasive screening mechanism that can assist healthcare providers in early-stage diagnosis and intervention.

References

[1] "Developing a Transparent Anaemia Prediction Model Empowered With Explainable Artificial Intelligence," *IEEE Access*, vol. 10, 2022.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD*, 2016.

[4] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.