iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** V **Month of publication:** May 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.82778

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue V May 2026- Available at www.ijraset.com*

# Anemia Prediction Using Machine Learning and NLP

K. Tulasi Krishna Kumar[1], K. Appanna[2]
[1]*Associate Professor & Training & Placement Officer,* [2]*MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India*

***Abstract:*** ***Anemia is a common health disorder caused by a deficiency of hemoglobin or red blood cells, leading to fatigue, weakness, and serious complications if not detected early. Early diagnosis is essential for effective treatment and prevention. This project presents a machine learning-based system designed to predict anemia using important patient health parameters such as hemoglobin levels, red blood cell count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). The dataset is carefully pre-processed to handle missing values and improve overall data quality, which helps enhance the accuracy of the prediction models. Various machine learning algorithms, including Random Forest, Logistic Regression, and Decision Tree, are applied to analyze the clinical data and classify whether an individual is anemic or not. These models are evaluated using performance metrics such as accuracy and precision, with Random Forest providing better results due to its ability to handle complex relationships among features. The system is further integrated into a user-friendly web interface that allows users to input medical data and receive instant predictions. Overall, this project demonstrates the significant role of artificial intelligence in healthcare by enabling early detection of anemia, reducing manual effort, and supporting medical professionals in making accurate and timely decisions.***

***Keywords: Random Forest Classifier, Logistic Regression, Decision tree, XGBoost classifier, LSTM (Long Short -Term Memory), Support Vector Machines (SVM), MinMaxScaler, LabelEncoder***

## I. INTRODUCTION

Anemia is a common and clinically significant health condition that occurs due to a deficiency of hemoglobin or a reduced number of red blood cells (RBCs), leading to insufficient oxygen supply to body tissues. This results in symptoms such as fatigue, weakness, dizziness, and reduced physical performance. If anemia remains undetected or untreated for a prolonged period, it may lead to serious health complications affecting the heart, brain, and overall immunity. Therefore, early and accurate diagnosis plays a crucial role in ensuring timely medical intervention and effective treatment outcomes.This project focuses on developing a machine learning-based anemia prediction system that enables early detection using key clinical parameters such as hemoglobin levels, RBC count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). These features are widely used in medical diagnosis and provide important insights into a patient's blood health condition. By analyzing these parameters, the system identifies patterns that help in classifying whether a patient is anemic or not.The dataset used in this project undergoes careful preprocessing to enhance its quality and reliability. This includes handling missing values, removing inconsistencies, and applying feature scaling techniques to ensure uniformity across all input variables. Proper data preprocessing significantly improves model performance and ensures more accurate and stable predictions.Several machine learning algorithms, including Random Forest, Logistic Regression, and Decision Tree classifiers, are implemented and evaluated to determine the most effective model. Each algorithm is assessed using performance metrics such as accuracy, precision, recall, and F1-score. Among these, the Random Forest algorithm delivers superior performance due to its ensemble learning approach, which combines multiple decision trees to handle complex and non-linear relationships within the data while reducing the risk of overfitting.Furthermore, the trained model is integrated into a user-friendly web application using the Flask framework. This interface allows users to input patient data easily and obtain real-time prediction results in a simple and interactive manner. The system demonstrates how artificial intelligence and machine learning can be effectively applied in healthcare to improve early disease detection, minimize manual diagnostic effort, enhance accuracy, and support healthcare professionals in making faster and more informed clinical decisions.

## II. LITERATURE SURVEY

The literature on anemia detection highlights the growing use of machine learning techniques to improve early diagnosis and clinical decision-making.

©IJRASET: All Rights are Reserved 

Traditional methods rely on laboratory tests and manual interpretation, which can be time-consuming and prone to human error. Recent studies have explored data-driven approaches using algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, and Random Forest to analyse patient health parameters like hemoglobin, red blood cell count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). Researchers have found that ensemble methods, particularly Random Forest, often achieve higher accuracy due to their ability to handle complex and non-linear relationships in medical data. Several works also emphasize the importance of data pre-processing, including handling missing values, normalization, and feature selection, to improve model performance. Additionally, some studies have integrated these predictive models into web-based or clinical decision support systems, enabling real-time anemia detection and assisting healthcare professionals. [9] Despite these advancements, challenges such as limited dataset availability, data imbalance, and model generalization remain areas of ongoing research. Overall, the existing literature demonstrates that machine learning has significant potential to enhance anemia diagnosis by providing faster, more accurate, and scalable solutions compared to traditional approaches.

## III. CHALLENGES

One of the major challenges in anemia prediction using machine learning is the availability and quality of medical data, as datasets are often limited, incomplete, or contain missing and inconsistent values that can affect model performance. Another key issue is data imbalance, where the number of anemic and non-anemic cases may not be evenly distributed, leading to biased predictions. Feature selection is also challenging, as identifying the most relevant clinical parameters without introducing redundancy is crucial for achieving accurate results. Additionally, different machine learning models may produce varying outcomes, making it difficult to select the most reliable algorithm without extensive evaluation. There are also concerns related to model generalization, as a model trained on one dataset may not perform well on data from different populations or regions. Integration into real-world healthcare systems presents further difficulties, including ensuring user-friendly interfaces, maintaining patient data privacy, and meeting medical standards. [8] Overall, these challenges highlight the need for robust data preprocessing, careful model selection, and continuous validation to develop an effective and reliable anemia prediction system.uncertainties of human behaviour and climate projections, which can hinder the accuracy of demand forecasting. Furthermore, many organizations find it difficult to transition from traditional physical servers to scalable cloud technology, often lacking the in-house expertise required to manage the complex AI and machine learning systems necessary for real-time, automated decision-making.

## IV. PROPOSED METHODOLOGY

The proposed methodology involves collecting a dataset with key health parameters such as hemoglobin, RBC count, MCV, and MCH, followed by data preprocessing to handle missing values and improve quality. The processed data is then split into training and testing sets, and multiple machine learning algorithms like Logistic Regression, Decision Tree, and Random Forest are applied. [7] These models are evaluated using metrics such as accuracy and precision, with Random Forest selected as the best-performing model. Finally, the chosen model is integrated into a user-friendly web application to provide instant anemia predictions based on user input.

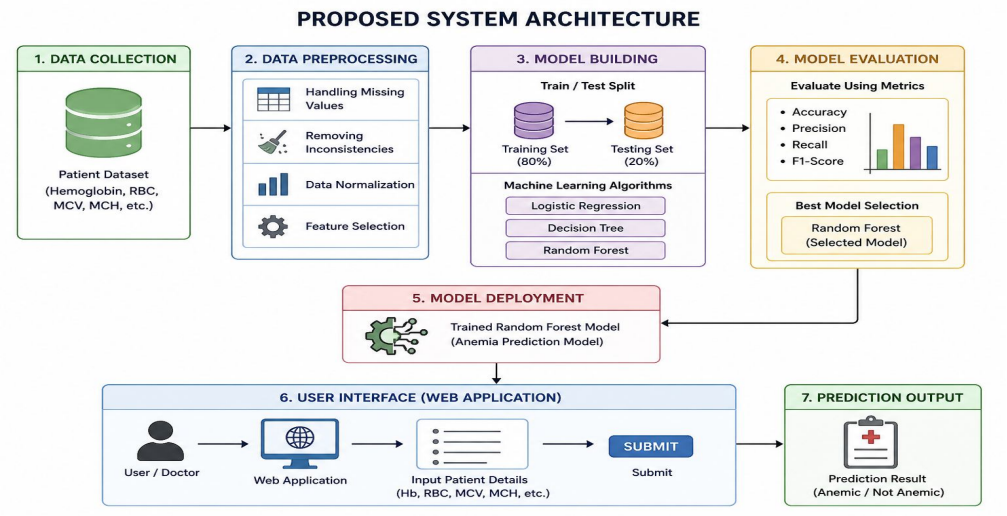



Fig.1Flow chart of the proposed methodology

## V. ALGORITHMS AND TECHNIQUES

### A. *Random Forest Classifier*

Random Forest Classifier is an advanced **ensemble machine learning algorithm** that combines the predictions of multiple decision trees to produce a more accurate and stable result. It is widely used for both classification and regression problems due to its high accuracy and ability to handle complex datasets.

Working Principle:

Random Forest works on the principle of "wisdom of the crowd." Instead of relying on a single decision tree, it builds a large number of decision trees during training. Each tree is trained on a different random subset of the dataset using a method called bootstrap sampling (bagging). During prediction, each tree gives its output, and the final result is determined by majority voting (for classification).

Key Features:

- Handles large datasets efficiently.
- Works well with both numerical and categorical data.
- Reduces overfitting compared to a single decision tree.
- Provides feature importance ranking.

Application in This Project:

In the anemia prediction system, Random Forest analyzes medical parameters such as:

- Hemoglobin level
- RBC count
- MCV (Mean Corpuscular Volume)
- MCH (Mean Corpuscular Hemoglobin)

It captures complex relationships between these features and produces highly accurate predictions of whether a person is anemic or not.

Advantages:

- High accuracy and robustness
- Handles missing data effectively
- Resistant to overfitting
- Works well with non-linear data

Disadvantages:

- Computationally expensive
- Less interpretable compared to simpler models
- Requires more memory

### B. *Logistic Regression*

Logistic Regression is a supervised statistical learning algorithm used for binary classification problems. Despite its name, it is used for classification rather than regression.

Working Principle:

Logistic Regression uses the sigmoid function (logistic function) to map predicted values between 0 and 1. The output is interpreted as probability:

- If probability $\geq 0.5 \rightarrow$ Class = 1 (Anemic)
- If probability $< 0.5 \rightarrow$ Class = 0 (Not Anemic)

Mathematically:

- It calculates a weighted sum of input features
- Then applies the sigmoid function to convert it into probability

Key Features:

- Simple and fast algorithm
- Works well for linearly separable data
- Provides probabilistic interpretation of results

Application in This Project:
Logistic Regression is used as a baseline model to compare performance with advanced models like Random Forest. It helps understand how each medical feature contributes to anemia prediction.

Advantages:
- Easy to implement
- Fast training and prediction
- Highly interpretable
- Works well for small datasets

Disadvantages:
- Assumes linear relationship between variables
- Not suitable for complex datasets
- Sensitive to outliers

*C. Decision Tree Classifier*

Decision Tree Classifier is a supervised learning algorithm that builds a tree-like structure for decision-making. It is one of the most intuitive machine learning models.
Working Principle:
The algorithm splits the dataset into smaller subsets based on feature values. It selects the best feature for splitting using criteria like:
- Gini Index
- Entropy (Information Gain)

Each internal node represents a decision, each branch represents an outcome, and each leaf node represents the final prediction.

Example in Medical Context:
- If Hemoglobin < threshold → likely Anemia
- Else check RBC count → further classification

Key Features:
- Easy to understand and visualize
- Handles both numerical and categorical data
- No need for feature scaling
-

Application in This Project:
Decision Tree helps in understanding rule-based classification of anemia. It makes the system interpretable and shows how predictions are made step by step.

Advantages:
- Simple and interpretable
- No data normalization required
- Works well with small datasets

Disadvantages:
- Prone to overfitting
- Sensitive to small changes in data
- Can become complex with large datasets

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

*D. Flask Frame work*

Flask is a lightweight and flexible Python web framework used to develop web applications. It acts as the backbone of the anemia prediction system's backend.

Working Principle:

Flask works by routing user requests to Python functions. When a user enters medical data in the web interface:

1. The data is sent to the Flask backend
2. Flask processes the input
3. The trained ML model predicts the result
4. The result is sent back to the user interface

Key Features:

- Lightweight and easy to use
- Supports REST API development
- Highly scalable with extensions
- Integrates easily with machine learning models

Application in This Project:

Flask connects the machine learning model with the user interface, allowing:

- Real-time anemia prediction
- User input handling
- Result display on web page

Advantages:

- Simple architecture
- Fast development
- Easy integration with Python ML libraries

Disadvantages:

- Not suitable for very large enterprise systems without extensions
- Requires additional security configuration

*E. Data preprocessing*

Data preprocessing is a crucial step in machine learning that improves data quality and model performance.

(a) Handling Missing Values:

Medical datasets often contain missing values due to incomplete records. These are handled by:

- Removing missing rows (if small in number)
- Filling values using mean/median/imputation techniques

This ensures that the model receives complete and consistent data.

(b) Feature Scaling (Standardization/Normalization):

Different features like hemoglobin and RBC count may have different ranges. Scaling ensures all features contribute equally.

- Standardization: Converts data to mean = 0 and standard deviation = 1
- Normalization: Scales data between 0 and 1

This improves model accuracy and convergence speed.

(c) Train-Test Split:

The dataset is divided into:

- Training set: Used to train the model
- Testing set: Used to evaluate performance

Common split ratio: 80% training, 20% testing.

This helps in checking how well the model generalizes to unseen data.

*F. Model Evaluation*

Evaluation metrics are used to measure the performance of classification models.

(a) Accuracy: Measures the percentage of correctly predicted results.

(b) Precision: Measures how many predicted positive cases are actually positive.

c) Recall: Measures how many actual positive cases were correctly identified.

(d) F1-Score: Harmonic mean of precision and recall, used when dataset is imbalanced.

Importance in This Project:

These metrics help compare models like:

- Random Forest
- Logistic Regression
- Decision Tree

The best-performing model is selected based on highest accuracy and balanced F1-score.

These algorithms and techniques work together to build an efficient, accurate, and user-friendly anemia prediction system. [5]

## VI. ARCHITECTURE

The architecture of the anemia prediction system is designed as a multi-layered framework that integrates data processing, machine learning, and user interaction components. It begins with the data layer, where patient medical data such as hemoglobin, RBC, MCV, and MCH is collected and stored. This data is passed to the preprocessing layer, where missing values are handled, features are normalized, and relevant attributes are selected to ensure high-quality input. [2] The processed data is then fed into the model layer, where machine learning algorithms like Random Forest, Logistic Regression, and Decision Tree are trained and used for prediction. The best-performing model, Random Forest, is selected and deployed. Above this, the application layer is built using the Flask web framework, which connects the frontend user interface with the backend prediction model. Users or doctors can input patient details through the web interface, and the system processes this input and sends it to the trained model. Finally, the output layer displays the prediction result (Anemic or Not Anemic) in an easy-to-understand format. This structured architecture ensures efficient data flow, accurate predictions, and a seamless user experience.

## VII. OUTPUT

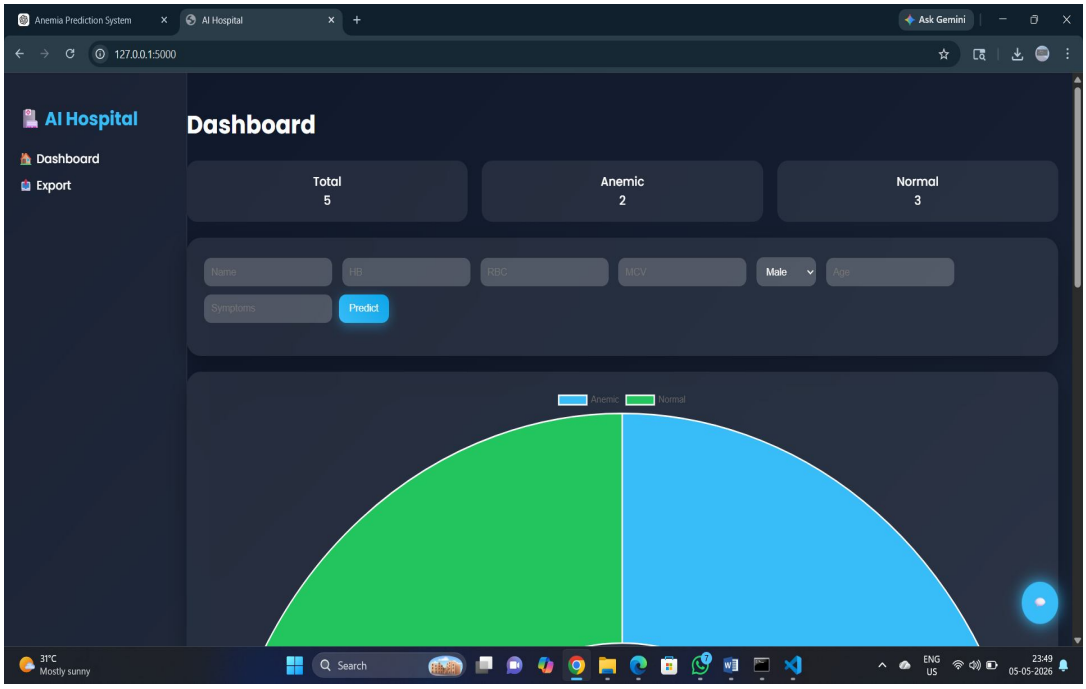



FIG1: Home Screen

The dashboard of the anemia prediction system is designed to provide a clear, interactive, and user-friendly interface for both users and healthcare professionals. It serves as the main control panel where users can easily navigate through different sections such as patient data entry, prediction results, and analytics. The dashboard includes input fields for key medical parameters like hemoglobin (Hb), red blood cell (RBC) count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH), allowing users to enter patient details efficiently. Once the data is submitted, the system processes the input and displays the prediction result (Anemic or Not Anemic) instantly in a visually appealing format. Additionally, the dashboard may include graphical representations such as charts and reports to help users understand trends and model performance. Built using modern web technologies like HTML, CSS, JavaScript, and integrated with a Flask backend, the dashboard ensures smooth interaction, fast response, and an enhanced user experience, making the system both practical and effective for real-world use.




FIG2: Sample Prediction Graph

The sample prediction graph displayed on the dashboard provides a clear visual representation of the anemia classification results. It is designed as a donut chart that compares the proportion of “Anemic” and “Normal” cases based on the processed data. Each segment is color-coded for easy understanding, where one colour represents anemic cases and the other represents normal cases, along with a legend for clarity. This type of visualization helps users quickly interpret the model’s output without needing technical knowledge. The graph enhances the overall user experience by making the results more intuitive and visually appealing, allowing doctors or users to analyse patterns and distributions at a glance. Integrated into the dashboard, this graphical representation supports better decision-making by presenting prediction outcomes in a simple, interactive, and informative manner.

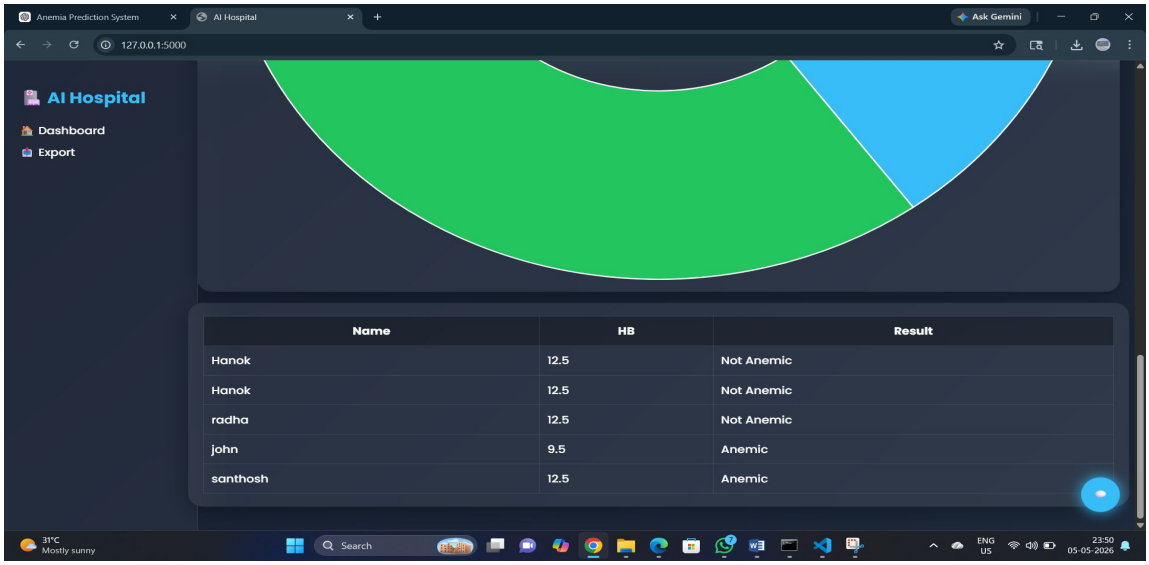



FIG3: Previous Patient Data

The dashboard also includes a tabular representation of prediction results, which provides a structured and detailed view of patient data along with the model's output. This table displays important information such as patient name, hemoglobin (HB) values, and the corresponding prediction result (Anemic or Not Anemic), making it easier to understand individual patient records in a clear format. It allows users to compare multiple patient entries simultaneously within a single interface, which is highly useful for monitoring, tracking, and analyzing trends in patient health conditions over time. The table is designed with a clean, responsive, and user-friendly layout, ensuring better readability across different devices and screen sizes. Proper alignment, spacing, and formatting enhance the overall usability of the dashboard, allowing even non-technical users to interpret the results easily. By combining both graphical visualization (such as donut charts or pie charts) and tabular representation, the dashboard delivers a more complete and meaningful overview of the prediction outcomes.This dual representation not only improves data interpretation but also supports better clinical understanding, as healthcare professionals can quickly analyze both individual values and overall patterns in the dataset. Additionally, this feature enhances data management capabilities and can be further extended to include functionalities such as exporting reports in PDF/Excel format, filtering patient records, and maintaining historical patient data for long-term analysis and medical reference.

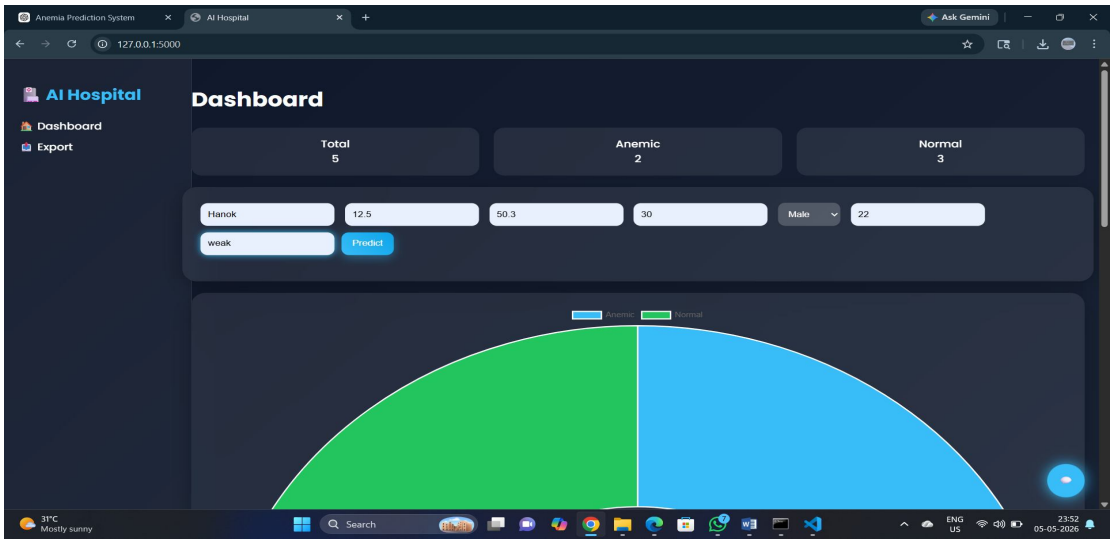



FIG4: New Patient Data Input

The dashboard interface of the anemia prediction system is designed to provide a comprehensive and interactive environment for users to analyse and manage patient data efficiently. At the top, summary cards display key statistics such as the total number of records, number of anemic cases, and normal cases, giving a quick overview of the dataset. Below this, an input form allows users to enter patient details including name, hemoglobin level, MCV, MCH, gender, age, and symptoms, enabling real-time prediction through a "Predict" button. The system processes this input using the trained machine learning model and updates the results instantly.

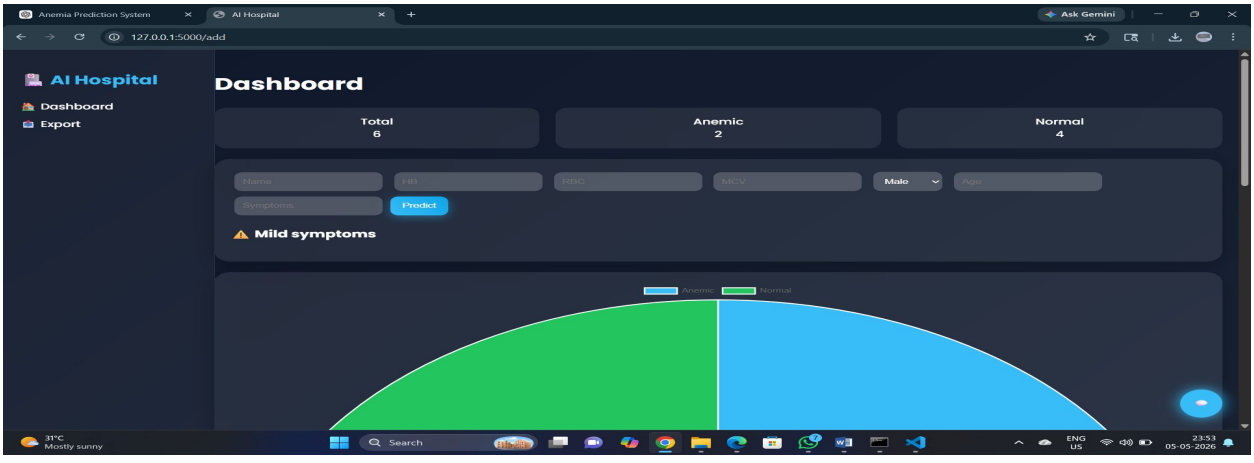



FIG5: Final Output

## VIII. CONCLUSION

The anemia prediction system demonstrates the practical application of machine learning in the healthcare domain for early detection and decision support. It uses key clinical parameters such as hemoglobin, RBC count, MCV, and MCH to analyze patient data and predict whether a person is anemic or not.In this project, multiple machine learning algorithms such as Random Forest, Logistic Regression, and Decision Tree are implemented and evaluated. Among these, Random Forest provides the highest accuracy due to its ensemble learning technique, while Logistic Regression and Decision Tree help in comparison and interpretability. This combination ensures both accuracy and better understanding of the prediction process.The system is deployed using the Flask web framework, which connects the machine learning model with a user-friendly web interface. This allows users to enter medical values easily and receive instant prediction results in real time. The simplicity of the interface makes the system accessible to both medical professionals and general users.This project helps in reducing manual workload, saving time, and improving the efficiency of medical diagnosis. It also ensures faster and more consistent results compared to traditional methods. However, the system's performance depends on the quality, size, and balance of the dataset, and it may require further improvement for real-world clinical deployment.Overall, the project highlights the importance of artificial intelligence in healthcare and shows how machine learning can be used to build fast, scalable, and cost-effective solutions for early disease prediction and medical decision support.

## REFERENCES

[1] [S. Rajaraman, S. Candemir, and G. Thoma, "Visualizing and explaining deep learning predictions for pneumonia detection in chest radiographs," Proc. IEEE Int. Conf. Comput. Vis. Workshops, pp. 1–9, 2018.
[2] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778, 2016.
[4] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.
[5] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," Annu. Rev. Biomed. Eng., vol. 19, pp. 221–248, 2017.
[6] J. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017.
[7] M. Abadi et al., "Tensor Flow: Large-scale machine learning on heterogeneous systems," 2016. [Online]. Available:
[8] https://www.tensorflow.orgF. Chollet, "Keras: Deep learning library for Python," 2015. [Online]. Available: https://keras.io
[9] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, "Densely connected convolutional networks," Proc. IEEE CVPR, pp. 4700–4708, 2017.
[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.
[11] S. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[12] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE CVPR, pp. 1–9, 2015.
[13] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Brief. Bioinform., vol. 19, no. 6, pp. 1236–1246, 2018.
[14] T. Rahman et al., "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," Comput. Biol. Med., vol. 132, pp. 104319, 2021.
[15] P. Kaur, G. Singh, and P. Kaur, "Intelligent diagnosis of anemia using machine learning techniques," Proc. Int. Conf. Comput. Intell. Data Sci., pp. 1–6, 2020.
[16] S. R. Dubey, S. K. Singh, and R. K. Singh, "Multimodal deep learning for medical diagnosis: Areview," IEEE Access, vol. 9, pp. 121001–121020, 2021.
[17] H. Greenspan, B. van Ginneken, and R. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1153–1159, 2016.
[18] J. Arevalo, F. González, R. Ramos-Pollán, J. Oliveira, and M. Guevara Lopez, "Representation learning for mammography mass lesion classification with deep learning," Comput. Methods Programs Biomed. vol. 127, pp. 248–257, 2016.
[19] M. Anthimopoulos et al., "Lung pattern classification for interstitial lung diseases using a deep CNN," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1207–1216, 2016.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

## BIBLOGRAPHY

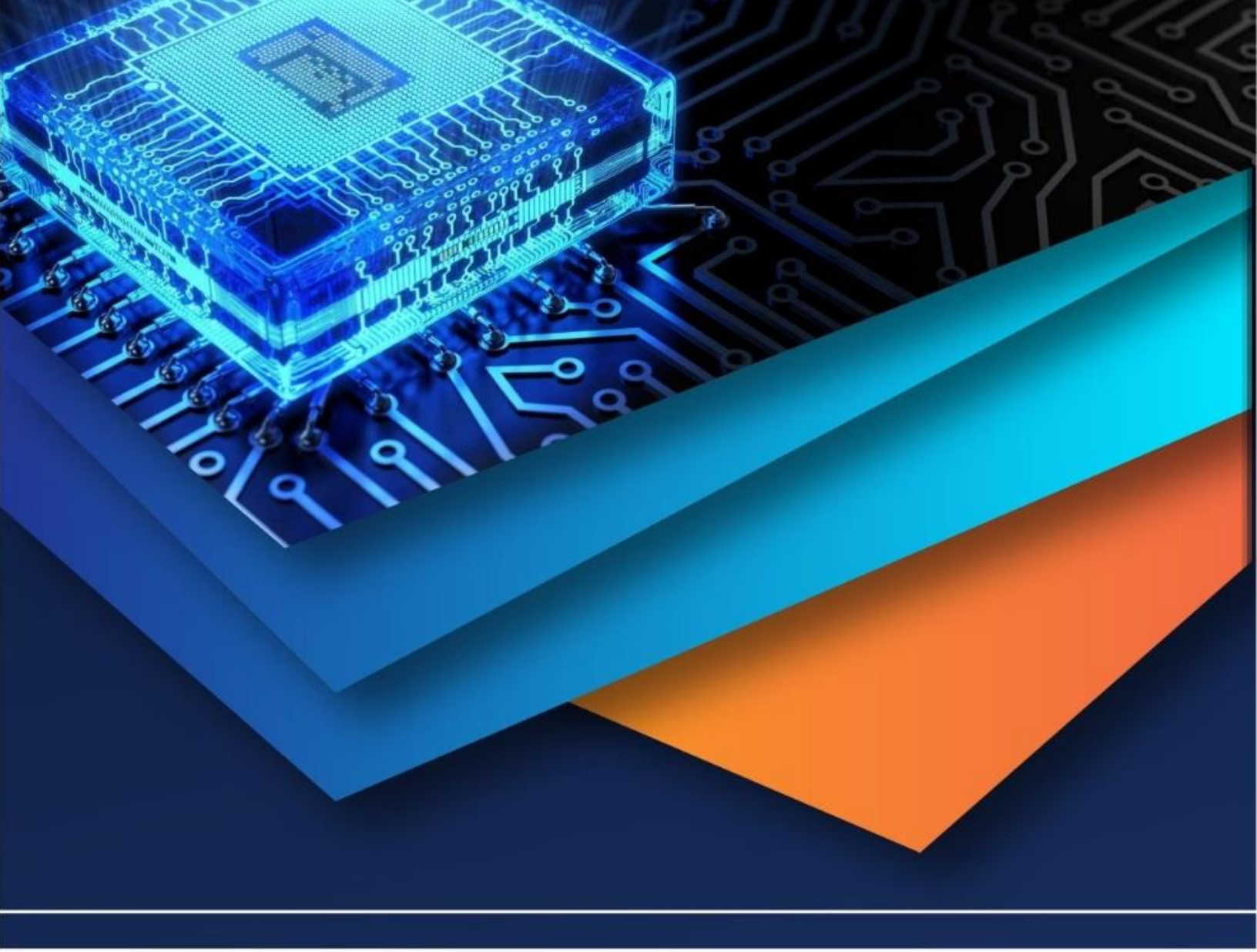


K. Tulasi Krishna Kumar is a Ratified Associate Professor affiliated with Andhra University and currently serves as the Placement Officer at SVPEC. With over 17 years of distinguished experience, he has played a pivotal role in training and placing students across IT, ITES, and core industry sectors, having mentored more than 16,000 students and 750 faculty members. An accomplished academician, authored eight books and successfully guided over 85 undergraduate and postgraduate project teams. He has also contributed extensively to research, with guidance leading to more than 70 publications in reputed international journals. A Certified Campus Recruitment Trainer (JNTUA), he holds an M. Tech in Computer Science and is presently pursuing his Ph.D. in CSE.



K. Appanna is currently pursuing her final semester of Master of Computer Applications (MCA) at Sanketika Vidya Parishad Engineering College, which is accredited with an ‘A’ grade by NAAC, affiliated to Andhra University, and approved by AICTE. With a keen interest in Machine Learning and Artificial Intelligence, he has undertaken her postgraduate project titled “Anemia prediction using machine learning and NLP” The project has been successfully carried out under the guidance of K. Tulasi Krishna Kumar, Assistant Professor, SVPEC.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)