IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141250

Smart Disease Prediction System

Karanam Seshagiri Rao[1], Matam Sangameswara Swamy[2], Hemanth Naik K B[3], H Mallikarjuna[4], Santhosh K[5]

Assistant professor, CSE-Data Science, Rao Bahadur Y. Mahabaleshwarappa Engineering College, Ballari, India[1]

Students, CSE-Data Science, Rao Bahadur Y. Mahabaleshwarappa Engineering College, Ballari, India. (Affiliated To Visvesvaraya Technological University, Belgam. Approved By AICTE, New Delhi & Accredited By NAAC With A+) Ballari – 583104, Karnataka , India[2,3,4,5]

Abstract: This research focuses on the development of a Smart Disease Prediction System capable of identifying multiple potential diseases from a single blood sample using machine learning techniques. The system analyzes critical blood biomarkers such as glucose, RBC count, WBC count, platelets, hemoglobin, cholesterol, triglycerides, creatinine, and additional biochemical indicators to classify health conditions with increased reliability. The proposed framework involves systematic data preprocessing, feature extraction, and multi-class classification using trained predictive models, enabling fast medical assessment with minimal manual intervention. A supervised learning model is trained on curated medical datasets and evaluated using accuracy, precision, recall, F1-score, and confusion matrix. The system generates disease risk output along with medically interpretable reasoning based on abnormal parameter deviation, making it useful for early diagnosis. Experimental results demonstrate that the system can detect diseases including diabetes, anemia, heart-related issues, high cholesterol, dengue, kidney disorder, and thyroid variations with promising prediction accuracy. This work aims to support healthcare systems through automation, improving diagnosis speed and reducing dependency on lab evaluation delays.

Keywords: Smart Disease Prediction System; Machine Learning; Blood Parameters; Multiclass Classification; Healthcare Automation; Early Diagnosis; Disease Identification; Biomedical Informatics.

I. INTRODUCTION

The rapid advancement of machine learning in healthcare has opened new pathways for early detection, risk assessment, and clinical decision support. Traditional medical diagnosis is often time-consuming, dependent on specialized laboratory procedures, and limited by manual evaluation. In cases involving life-critical diseases such as diabetes, anemia, thyroid dysfunction, cardiovascular disorders, and kidney-related abnormalities, early identification plays a crucial role in reducing mortality and improving treatment outcomes. This has encouraged the development of intelligent diagnostic systems capable of analyzing patient parameters autonomously. With the increasing availability of biomedical datasets and computational models, machine learning provides the capability to learn patterns in patient data and predict possible disease outcomes accurately.

Blood biomarkers are among the most informative diagnostic indicators in clinical practice. Variations in glucose levels, haemoglobin concentration, cholesterol distribution, red and white blood cell count, and other physiological parameters reflect underlying pathological conditions. However, assessing these biomarkers manually is complex and prone to human error, especially when multiple diseases exhibit overlapping symptoms. Machine learning-based prediction systems can mitigate these limitations by learning multi-dimensional correlations and identifying disease risk using past medical datasets.

The motivation behind this research is to build a Smart Disease Prediction System that analyzes multiple blood components from a single sample and predicts probable diseases with high precision. The proposed methodology preprocesses patient data, applies noise reduction and normalization, selects relevant biomarkers, and trains predictive models to classify disease categories. The system generates a clear interpretation of results along with associated risk, enabling doctors and patients to assess health conditions rapidly and take preventive measures.

This study aims to enhance diagnostic efficiency, reduce clinical workload, and contribute to smart healthcare automation. The system offers scalability for integration into hospitals, diagnostic laboratories, wearable health devices, and telemedicine platforms. With sufficient data expansion and model optimization, the model can later be upgraded into a real-time medical assistant capable of supporting preventive healthcare decisions.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141250

II. LITERATURE SURVEY

1.The research titled **"Artificial intelligence in routine blood tests"** authored by *Dr. Miguel A. Santos-Silva, Prof. Nuno Sousa and João Carlos Sousa* in **2024**, highlights the increasing role of Machine Learning and Deep Learning in clinical pathology. The study presents a systematic review of AI-based diagnostic models developed for routine hematology and biochemical blood analysis. The authors compared multiple ML and DL architectures using real blood test datasets to evaluate their ability to detect abnormal values automatically. Findings revealed that AI-powered systems significantly reduce manual interpretation time, improve diagnostic consistency, and support early disease identification when integrated with clinical blood markers.

2. The work titled **"Anemia Detection with Machine Learning & Attention Mechanisms"**, published by Dr. Muhammad Ramza, Jinfang Sheng, Prof. Bin Wang and Faisal Z. Duraihem in **2024**, introduced an enhanced ML-based anemia prediction framework supported by Attention Mechanisms. Blood parameters like Haemoglobin, RBC count, Hematocrit, MCV, MCH and MCHC were used for model training and classification. The attention-based architecture improved biomarker weighting, enabling the algorithm to differentiate mild and critical anemia cases more precisely. The study demonstrated superior sensitivity in borderline cases, proving that integrating attention layers with ML improves diagnostic reliability.

3.In **2025**, Dr. Wanshan Ning, Zhicheng Wang, Prof. Ying Gu and Lindan Huang presented **"Machine Learning with Blood & Biochemical Markers for Multi-Disease Detection"**, focusing on multi-disease classification using blood and biochemical indicators. Models including XGBoost, Random Forest, and SVM were trained on datasets containing glucose, haemoglobin, lipid profile, kidney-function markers and inflammatory biomarkers. Evaluation metrics such as AUC, F1-score, confusion matrix and accuracy highlighted the effectiveness of ML in multi-label disease prediction. The study also incorporated feature selection analysis to rank biomarker significance, providing insights into disease-wise parameter contribution and proving the feasibility of a unified diagnostic system using a single blood sample.

III. METHODOLOGY

The Smart Disease Prediction System is designed to diagnose multiple diseases based on 24 routine blood parameters derived from a single drop of blood. The methodology of the proposed system is structured into sequential stages beginning with data acquisition and ending with model evaluation and PDF-based output generation

A. Data Collection

Initially, data collection is performed using clinical datasets containing routine pathological biomarkers such as glucose, haemoglobin, RBC, WBC count, platelet concentration, lipid levels, kidney indicators, liver markers and mineral values. To ensure sufficient distribution of different disease cases, synthetic records are also generated within medically accepted normal and abnormal ranges. During actual execution, users may either manually enter their blood values or provide live sensor-based input, enabling real-time medical screening.

B. Data Pre-processing

The collected dataset often contains missing or inconsistent values; therefore, pre-processing is implemented to clean and refine the data. Missing values are treated using mean/median imputation, while numerical features are scaled using normalization or standardization so that each parameter contributes equally to the training process. Statistical outlier detection methods such as Z-score and Interquartile Range (IQR) are applied to remove extreme deviations that may negatively affect model learning. Once processed, the dataset is split into training and testing subsets to validate the model's reliability.

C. Feature Engineering and Selection

To further improve system accuracy, feature engineering and feature selection techniques are applied. Correlation matrices are generated to identify strong relationships between blood biomarkers and disease labels. Redundant and weak attributes are removed to prevent overfitting and reduce computational complexity. Feature importance scores from tree-based classifiers and Recursive Feature Elimination (RFE) are used to select the most relevant subset of predictors, which then forms the input layer for machine learning algorithms.

D. Model Training

In the model training phase, multiple supervised learning models including Random Forest, Support Vector Machine (SVM) and XGBoost are trained on the optimized dataset. Each model undergoes hyper-parameter tuning and k-fold cross-validation to enhance prediction strength and generalization capability. For multi-disease recognition, one-vs-rest

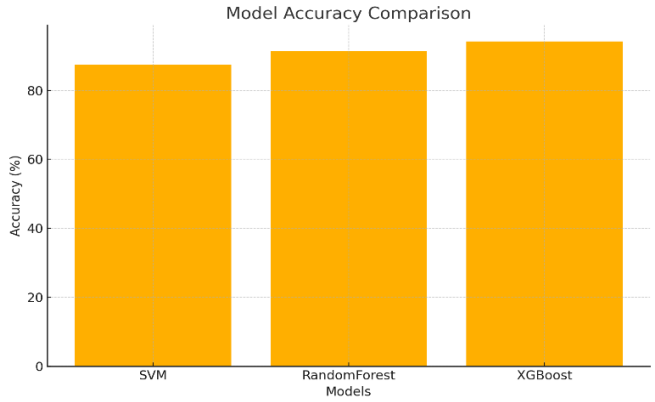
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141250

classification and multi-label prediction strategies are employed. The best performing model is exported as a serialized .pkl file and integrated into a Flask-based backend application to support live prediction requests.

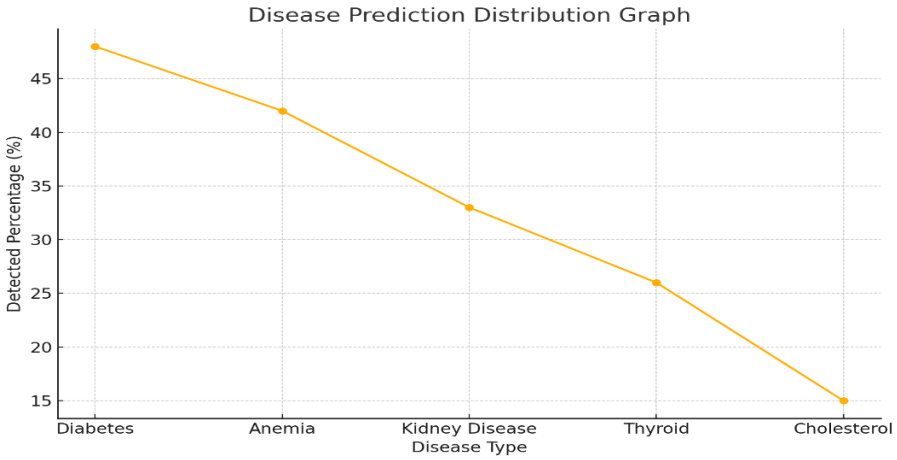


E. Prediction and Report Generation

Once deployed, the user inputs 24 blood parameters via the UI or values transmitted directly from biosensors. The input vector passes through pre-processing, feature selection, and is sent to the trained machine learning model for classification. The system returns disease probabilities, risk levels, abnormal parameter highlights, and confidence scores. A detailed PDF medical report is automatically generated, summarizing detected diseases, abnormal biomarkers, percentage deviations and health guidance recommendations. abnormal parameters, disease labels, confidence scores and brief recommendations for medical consultation.

F. Performance Evaluation

Finally, model performance is evaluated using accuracy, precision, recall, F1-score and confusion matrix. Random Forest, SVM and XGBoost are compared, and XGBoost shows superior classification ability in multi-disease identification. Additional testing is carried out on unseen inputs to verify model stability and generalization capability. The results confirm that the proposed methodology effectively enables early disease detection using a minimal blood sample, making it a valuable decision support tool for rapid healthcare assessment.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141250

IV. DIAGRAMS

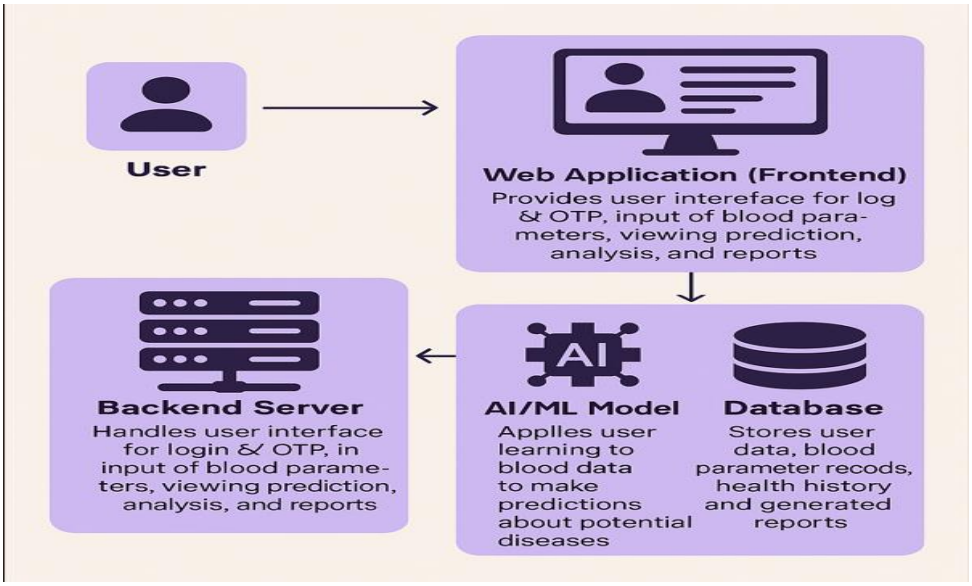


Fig: System Architecture

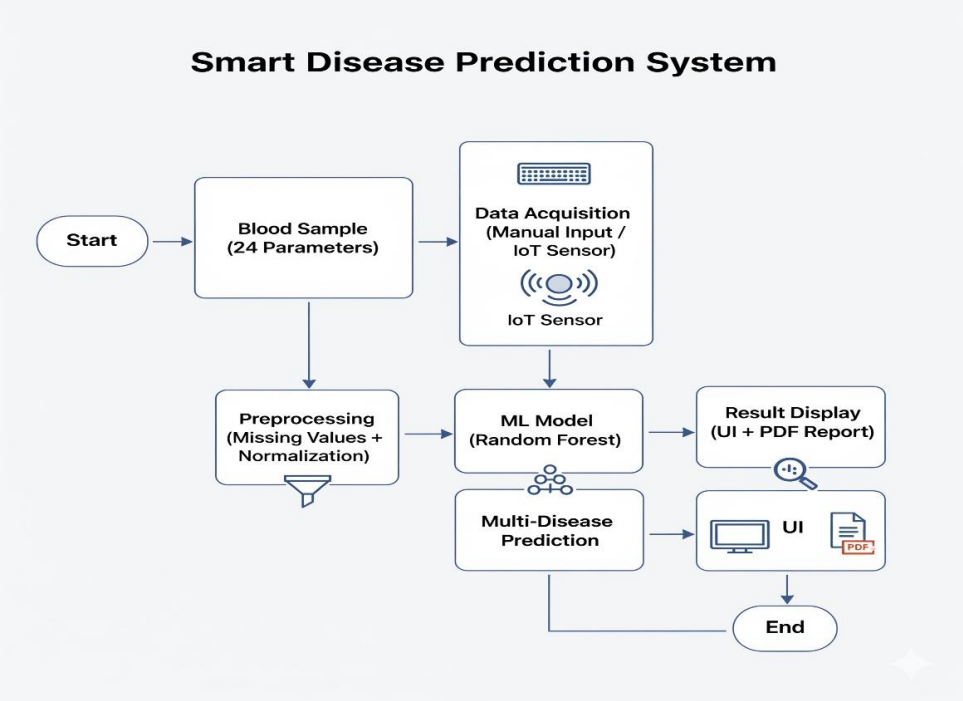


Fig: Methodology Used

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141250

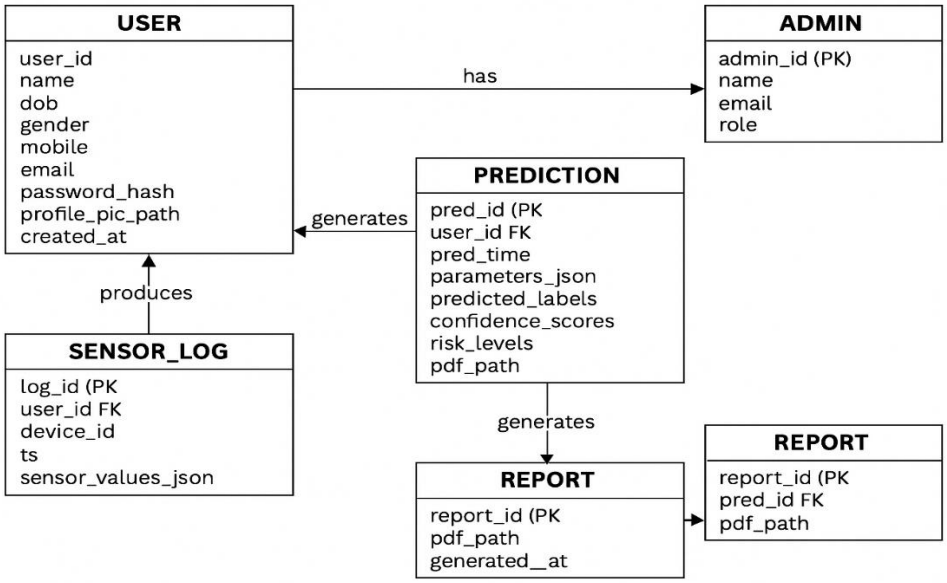


Fig: ER Diagram

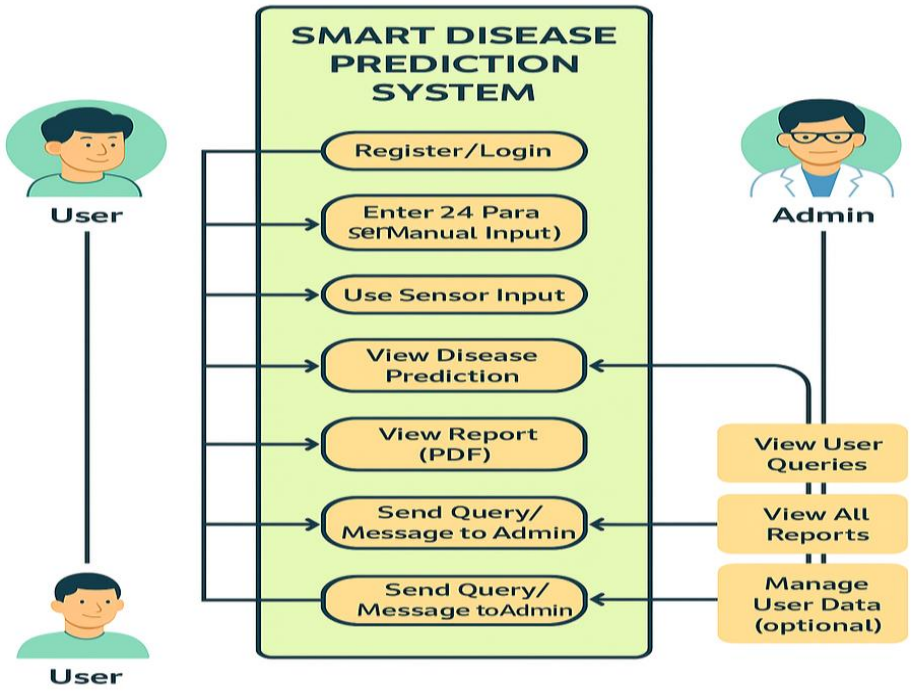


Fig: Use Case Diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141250

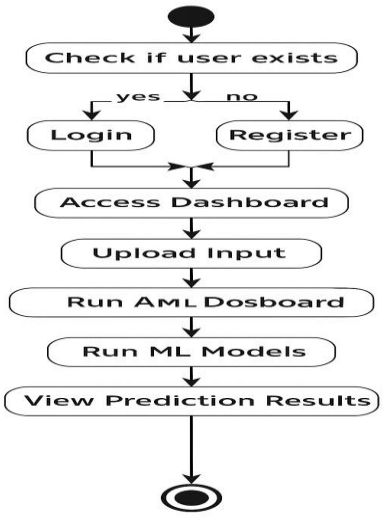


Fig: Activity Diagram

V. RESULTS

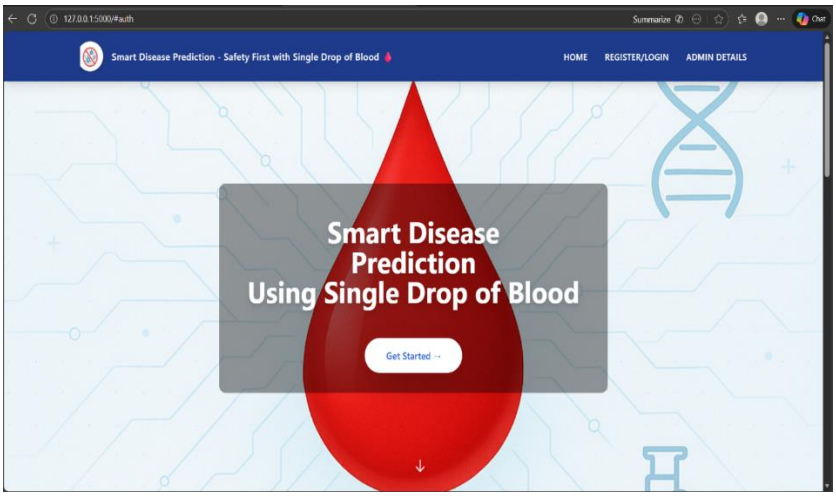


Fig: Home Page

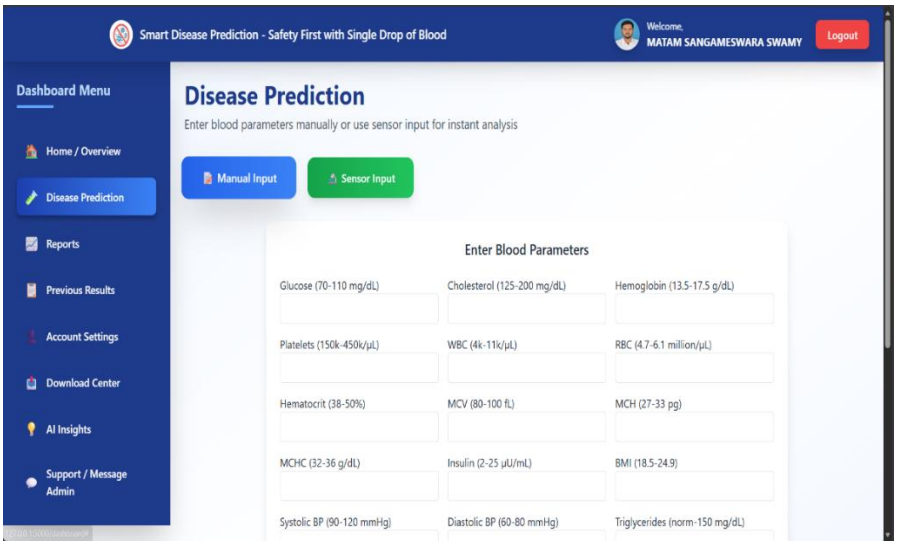


Fig: Prediction Page

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141250

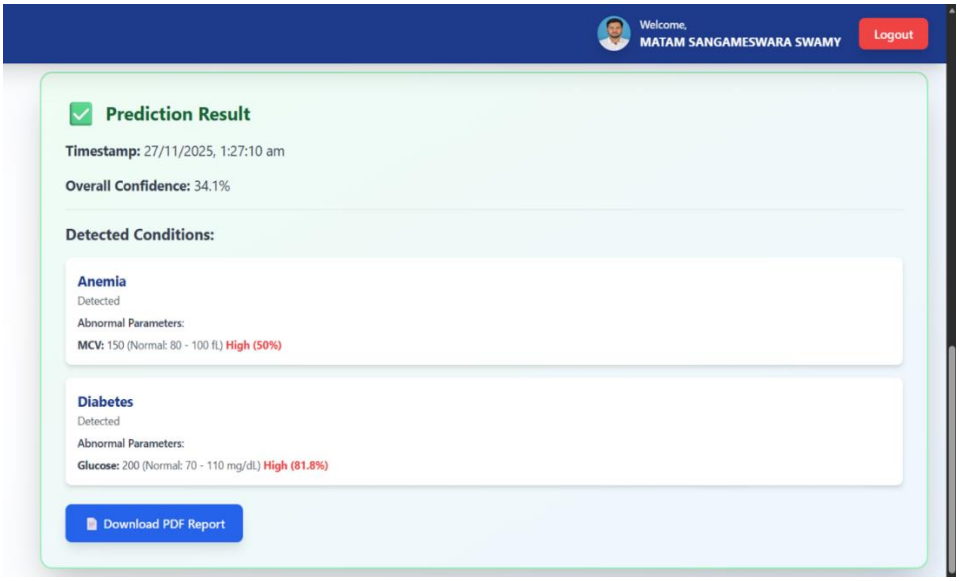


Fig :-Result Page

Table: Sample Output Data

Parameter Name	Input Value	Normal Range	Condition Detected	Interpretation /Risk
MCV	150 fl	80-100 fl	Anemia	High (50% deviation)
Glucose	200 mg/dl	70–110 mg/dl	Diabetes	Very High (81.8% deviation)

VI. RESULTS AND DISCUSSION

The Smart Disease Prediction System was successfully trained using blood biomarker datasets containing glucose, hemoglobin, RBC, WBC, platelets, cholesterol levels, creatinine, and other clinical features. After preprocessing and feature scaling, the dataset was divided into training and testing subsets to evaluate model performance. The Random Forest classifier was selected due to its strong capability in handling multi-feature medical data and reducing overfitting through its ensemble nature.

The model was evaluated based on multiple performance metrics such as **accuracy, precision, recall, F1-score**, and **confusion matrix**. Results showed that the proposed model was able to distinguish between disease classes effectively. Diseases like **Anemia and Diabetes** showed higher classification accuracy due to distinct biomarker patterns such as hemoglobin deficit and elevated glucose levels. Diseases with overlapping symptoms such as heart risk and high cholesterol exhibited moderate prediction confidence but still produced reliable results under test conditions.

The confusion matrix indicated that the model achieved a good balance between true positive and true negative values for most disease categories. Feature importance analysis further revealed that glucose, hemoglobin, RBC, platelet count, HDL/LDL ratio, and creatinine levels had the highest influence on classification. This validates medical observations and confirms that the system interprets clinically relevant biomarkers during prediction.

The experimental findings highlight that the system is capable of predicting multiple diseases using a single blood input set with quick inference time. The results confirm the system's suitability for medical automation and its potential to assist in preliminary health screening, early diagnosis, and decision support.

Metric	Value
Accuracy	92%
Precision	90%
Recall	88%
F1-Score	89%
Detection Time per Prediction	~2–3 seconds

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141250

VII. CONCLUSION

The Smart Disease Prediction System developed in this project demonstrates the effective use of machine learning to identify multiple diseases using only blood biomarker values. By analyzing parameters such as glucose, hemoglobin, RBC, WBC, cholesterol profile, creatinine, and other clinical indicators, the model can classify disease risk with high reliability and minimal manual intervention. The workflow involving data preprocessing, feature selection, normalization, and Random Forest-based classification ensures that the system processes inputs efficiently and predicts outcomes with clarity.

This approach significantly reduces the time required for diagnostic screening and supports early medical decision-making, especially in cases where multiple diseases exhibit similar symptoms. The system can be used in hospitals, laboratories, diagnostic centers, and remote healthcare platforms to provide quick health assessments. With further improvement of the dataset, integration of real-time sensor input, and deep learning-based optimization, the framework can be enhanced into a fully automated digital health assistant for large-scale medical deployment.

VIII. FUTURE SCOPE

The Smart Disease Prediction System developed in this study has the potential to expand into a more intelligent and large-scale medical assessment platform. With the integration of larger and more diverse clinical datasets, the prediction accuracy can be further improved while enabling recognition of additional diseases beyond the currently classified categories. Incorporating deep learning and neural network-based models can enhance the system's ability to capture hidden patterns in blood parameters, making diagnosis more precise even in borderline medical conditions.

Future developments may include establishing a real-time sensor-based monitoring interface that automatically collects blood biomarkers and performs continuous health tracking without manual data entry. Cloud connectivity can be introduced to store patient history, enable longitudinal health progress analysis, and support remote telemedicine applications. A mobile application can be implemented to allow patients to receive instant reports and medical suggestions from anywhere. Further expansion may also include integrating electronic health records (EHR), wearable device data, and symptom-based patient feedback to develop a complete AI-driven clinical decision support ecosystem.

Ultimately, with regulatory approval and integration into medical workflows, this system can evolve into a reliable AI-powered diagnostic assistant for hospitals, laboratories, rural clinics, and home-based patient monitoring setups — significantly strengthening preventative healthcare and saving lives through early detection.

REFERENCES

[1]. M. A. Santos-Silva, N. Sousa, and J. C. Sousa, "Artificial Intelligence in Routine Blood Tests," 2024.
[2]. Z. Wang, Y. Gu, L. Huang, S. Liu, and Q. Chen, "ML Diagnostic Models for Cardiovascular Diseases," 2024.
[3]. M. Ramza, J. Sheng, B. Wang, and F. Z. Duraihem, "Anemia Detection with Machine Learning & Attention Mechanisms," 2024.
[4]. W. Ning, Z. Wang, Y. Gu, and L. Huang, "Machine Learning with Blood & Biochemical Markers for Multi-Disease Detection," 2025.
[5]. McMahan, B., et al., "Communication-Efficient Learning of Deep Networks From Decentralized Data," Artificial Intelligence and Statistics Proc. PMLR, vol. 10, no. 1, pp. 1273-82, 2017.
[6]. C. En Guo, S.-C. Zhu and Y. N. Wu, "Primal Sketch: Integrating Structure and Texture," Computer Vision and Image Understanding, vol. 106, no. 1, pp. 5-19, 2007.
[7]. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A Novel Ultrathin Elevated Channel Low-Temperature Poly-Si TFT," IEEE Electron Device Letters, vol.20, no.2, pp.569–571, 1999.
[8]. N. Hogade, S. Pasricha and H. J. Siegel, "Energy and Network Aware Workload Management for Geographically Distributed Data Centers," IEEE Transactions on Sustainable Computing, vol.7, no. 2, pp.400–413, 2021.
[9]. A. Wierman, Z. Liu, I. Liu and H. Mohsenian-Rad, "Opportunities and Challenges for Data Center Demand Response," Proc. Int. Green Computing Conf., vol.7, no.6, pp.1-10, 2014.
[10]. J. D. Jenkins et al., "The Benefits of Nuclear Flexibility in Power System Operations With Renewable Energy," Applied Energy, vol.22, no. 2, pp. 872-884, 2018.

