iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.77600

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 III Mar 2026- Available at www.ijraset.com

An AI-Based Smart Health Monitoring System for Multi-Disease Risk Prediction Using Machine Learning

Md. Sharukh, Amaan Ahmad, Mohd. Fuzail Khan
Department of Computer Science & Engineering, Integral University, Lucknow, India

Abstract: ***Chronic illnesses like diabetes and heart disease continue to place a heavy burden on global health, making early detection more important than ever. This study introduces an artificial intelligence framework designed to predict the risk of both conditions while keeping user data private. Using Random Forest ensemble learning, the system analyzes two well-known datasets: the PIMA Indians Diabetes dataset and the Cleveland Heart Disease dataset. To ensure reliable results, the researchers applied stratified train-test splitting and five-fold cross-validation.The diabetes model reached an accuracy of nearly 76% with a strong ROC-AUC score of 0.813, while the heart disease model performed even better, achieving over 81% accuracy and a ROC-AUC of 0.947. Importantly, the framework doesn't just provide predictions—it also highlights which features matter most, aligning with established medical risk factors. This makes the system more transparent and clinically meaningful.Compared to traditional methods like Logistic Regression and Decision Trees, the ensemble approach proved more robust. By processing inputs locally, the framework ensures privacy while offering dependable insights. Overall, the result of this research shows how AI can support preventive healthcare, empowering individuals and clinicians to act early and reduce long-term complications.***
Keywords: Machine Learning, Random Forest, Diabetes Prediction, Heart Disease Prediction, Preventive Healthcare

I. INTRODUCTION

Non-communicable diseases like diabetes and heart disease are among the biggest threats to human health today, responsible for millions of deaths worldwide. Cardiovascular disease remains the leading killer, while diabetes continues to rise due to lifestyle factors such as poor diet, lack of exercise, obesity, and aging populations. Because of this, early detection and preventive care are critical to reducing illness and saving lives.

Traditional diagnosis often depends on doctors manually assessing risk factors, which can be effective but also time-consuming and influenced by personal judgment. Artificial Intelligence (AI) and Machine Learning (ML) are changing this landscape by offering faster, data-driven insights. These models can uncover complex patterns in medical data that humans might miss, providing more accurate risk predictions.

Among the many algorithms available, Random Forest stands out for healthcare applications. It works well with moderately sized datasets, reduces errors through its ensemble approach, and offers transparency by showing which features matter most—something vital in medicine. While most studies focus on predicting one disease at a time, few attempt to combine multiple conditions into a single system. This research fills the gap by creating a unified, privacy-preserving framework that predicts both diabetes and heart disease risks, ensuring reliable, interpretable, and secure healthcare support.

A. CONTRIBUTIONS OF THIS WORK

The key contributions of this research are:

1) Development of a unified dual-disease risk prediction framework integrating diabetes and heart disease assessment.
2) Implementation of Random Forest ensemble learning with stratified sampling and five-fold cross-validation.
3) Integration of feature importance ranking to enhance interpretability and clinical transparency.
4) Comparative evaluation against Logistic Regression and Decision Tree classifiers.
5) Design of a privacy-preserving architecture that processes user data locally without external transmission.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

II. LITERATURE REVIEW

Machine learning has become a cornerstone of modern healthcare, offering powerful tools for diagnosis and prediction. Logistic Regression remains popular because it is easy to interpret and provides probabilistic outputs, but its reliance on linear boundaries limits its ability to capture the complex, nonlinear nature of medical data. Decision Trees, on the other hand, can model nonlinear relationships but often suffer from overfitting, especially when datasets are small. Support Vector Machines and Neural Networks have also shown promise, yet deep learning models typically demand large datasets and significant computational resources, which are not always available in clinical contexts like the PIMA and Cleveland datasets.

Random Forest, introduced by Breiman, offers a balanced solution. By combining multiple decision trees through bootstrap aggregation and random feature selection, it reduces variance and improves generalization. In healthcare datasets, Random Forest consistently delivers stable and robust performance compared to single classifiers. Recent research highlights the importance of explainability and validation in medical AI, making techniques like feature importance analysis and cross-validation essential for clinical trust. Despite these advances, few studies attempt to unify predictions for multiple diseases within one framework. This study addresses that gap by presenting a consolidated, privacy-preserving system capable of predicting both diabetes and heart disease risks with reliability and transparency.

III. DATASET DESCRIPTION

This study makes use of two well-established benchmark datasets that are widely recognized in medical prediction research: the PIMA Indians Diabetes dataset and the Cleveland subset of the UCI Heart Disease dataset. Both are publicly available resources that have been extensively applied in machine learning studies focused on healthcare. Their popularity stems from the fact that they provide structured, clinically relevant data, making them ideal for testing predictive models. By relying on these datasets, the research ensures comparability with prior work while also grounding its framework in data that has been validated and trusted across numerous studies. This choice strengthens the credibility of the proposed system and allows for meaningful evaluation against existing approaches.

A. PIMA INDIANS DIABETES DATASET

The PIMA Indians Diabetes dataset contains 768 instances collected from female patients of Pima Indian heritage. Each record consists of eight clinical attributes used to predict the presence or absence of diabetes. The target variable is binary:

- 1 → Diabetic
- 0 → Non-diabetic

The predictor variables include:

- Pregnancies
- Plasma Glucose Concentration
- Diastolic Blood Pressure
- Skin Thickness
- Insulin Level
- Body Mass Index (BMI)
- Diabetes Pedigree Function
- Age

Glucose concentration and BMI are clinically recognized as strong indicators of diabetes risk. The dataset exhibits moderate class imbalance, which was addressed using stratified sampling during train-test splitting.

No categorical encoding was required as all features were numeric. The dataset was divided into 80% training data and 20% testing data while preserving class distribution.

It is important to note that the PIMA Indians Diabetes dataset contains records exclusively from female patients of Pima Indian heritage. Therefore, the trained diabetes prediction model is inherently influenced by gender-specific physiological characteristics. While the dataset is widely used as a benchmark in machine learning research, caution must be exercised when generalizing the findings to broader populations that include male patients or different ethnic groups.

B. CLEVELAND HEART DISEASE DATASET

The Cleveland Heart Disease dataset originally contains 303 instances with 14 attributes. After removing missing values represented by "?", 297 instances were retained for analysis. The original target variable ranges from 0 to 4, representing different severity levels of heart disease.

For this research, the target variable was converted into binary classification:

- 0 → No Heart Disease
- 1 → Presence of Heart Disease (values > 0)

The dataset includes the following clinical attributes:

- Age
- Sex
- Chest Pain Type (cp)
- Resting Blood Pressure (trestbps)
- Serum Cholesterol (chol)
- Fasting Blood Sugar (fbs)
- Resting Electrocardiographic Results (restecg)
- Maximum Heart Rate Achieved (thalach)
- Exercise Induced Angina (exang)
- ST Depression (oldpeak)
- Slope of Peak Exercise ST Segment (slope)
- Number of Major Vessels (ca)
- Thalassemia (thal)

Rows containing missing values were removed to ensure numerical consistency and avoid imputation bias. Stratified sampling preserved class distribution in training and testing sets.

C. PREPROCESSING STRATEGY

The preprocessing pipeline included:

- Data loading using Pandas
- Missing value removal (heart dataset)
- Binary conversion of target variable
- Stratified 80–20 train-test split
- Five-fold cross-validation

Feature scaling was not applied because Random Forest is a tree-based algorithm and does not require normalization.

IV. METHODOLOGY

The predictive system is built around Random Forest, chosen as the primary classification algorithm because of its strong performance and reliability in medical datasets. To provide a meaningful benchmark, Logistic Regression and Decision Tree models were also implemented as baseline comparisons. This setup allows the study to highlight the advantages of ensemble learning over simpler, single-model approaches, demonstrating how Random Forest achieves greater stability and interpretability while maintaining clinical relevance.

A. RANDOM FOREST CLASSIFIER

Random Forest is a powerful ensemble learning method that improves prediction accuracy by combining many decision trees. It works through a process called bootstrap aggregation, or bagging. In this approach, each tree is trained on a random subset of the data, selected with replacement, which means some samples may appear multiple times while others may be left out. At every split within a tree, only a random subset of features is considered, which introduces diversity among the trees and reduces the chance of overfitting.

Once all the trees are built, the model makes predictions by aggregating their outputs. For classification tasks, this is done through majority voting—each tree casts a "vote" for a class, and the class with the most votes becomes the final prediction. This ensemble mechanism reduces variance compared to a single decision tree and provides more stable, generalizable results.

In healthcare applications, this balance of accuracy, robustness, and interpretability makes Random Forest especially valuable, as it can highlight which features contribute most to predictions while maintaining strong performance on moderate-sized datasets.

B. GINI IMPURITY CRITERION

The splitting criterion used in Random Forest is based on Gini impurity:

$$Gini = 1 - \sum(pi)^2$$

where $p_i$represents the probability of class $_i$ at a node. Lower Gini values indicate higher node purity and better class separation.

C. LOGISTIC REGRESSION (BASELINE)

Logistic Regression models the probability of class membership using the sigmoid function:

Where

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta 0 + \beta 1x1 + \cdots + \beta nxn)}}$$

$$z = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn$$

interpretable, Logistic Regression may struggle with nonlinear relationships present in clinical datasets.

D. DECISION TREE (BASELINE)

Decision Tree recursively partitions the dataset using threshold-based splits. While capable of modeling nonlinear relationships, it may suffer from overfitting compared to ensemble methods.

V. EXPERIMENTAL SETUP

All experiments were conducted using Python and Scikit-learn libraries. The datasets were split into 80% training and 20% testing subsets using stratified sampling. Five-fold cross-validation was implemented to evaluate model stability.
The following evaluation metrics were used:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC Score
- Confusion Matrix

VI. RESULTS AND ANALYSIS

A. Diabetes Prediction Results

The Random Forest classifier achieved:

- Accuracy: **75.97%**
- ROC-AUC: **0.813**
- Cross-validation mean**:** 0.769 ± 0.026

Confusion Matrix: $\begin{bmatrix} 86 & 14 \\ 23 & 31 \end{bmatrix}$

Table 1. Performance Metrics for Diabetes Prediction

Metric	Value
Accuracy	75.97%

Precision	0.69
Recall	0.57
F1-score	0.63
ROC-AUC	0.813
Cross Validation Mean	0.769
Cross Validation Std	±0.026

As shown in Table 1, the model demonstrates reliable classification capability with stable cross-validation variance.

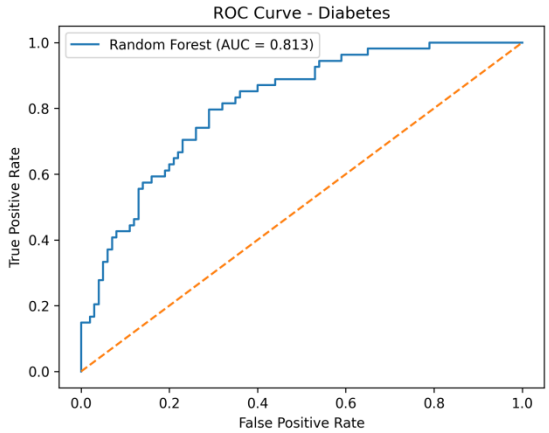


Figure 1. ROC Curve for Diabetes Prediction (AUC = 0.813)

The ROC curve for diabetes prediction (Figure 1) illustrates strong discriminatory power. Feature importance ranking (Figure 2) identifies Glucose, BMI, and Age as dominant predictors.

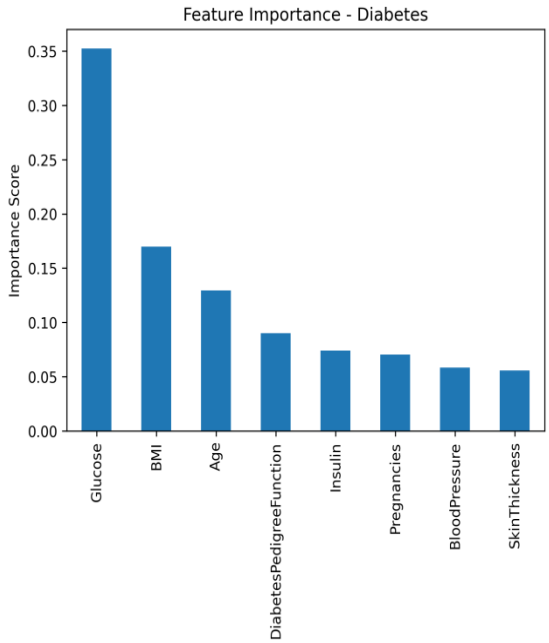


Figure 2. Feature Importance Ranking for Diabetes Model

B. Heart Disease Prediction Results

The Random Forest classifier achieved:

- Accuracy: **81.67%**
- ROC-AUC: **0.947**
- Cross-validation mean: **0.824 ± 0.031**

Confusion Matrix: $\begin{bmatrix} 23 & 3 \\ 8 & 20 \end{bmatrix}$

Table 2. Performance Metrics for Heart Disease Prediction

Metric	Value
Accuracy	81.67%
Precision	0.87
Recall	0.71
F1-score	0.78
ROC-AUC	0.947
Cross Validation Mean	0.824
Cross Validation Std	±0.031

The ROC curve for heart disease prediction (Figure 3) indicates excellent class separability.

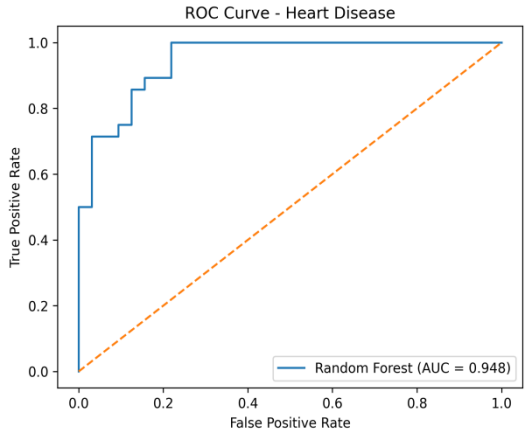


Figure 3. ROC Curve for Heart Disease Prediction (AUC = 0.947)

Feature importance ranking (Figure 4) highlights Thalassemia, Chest Pain Type, and Number of Major Vessels as dominant predictors.

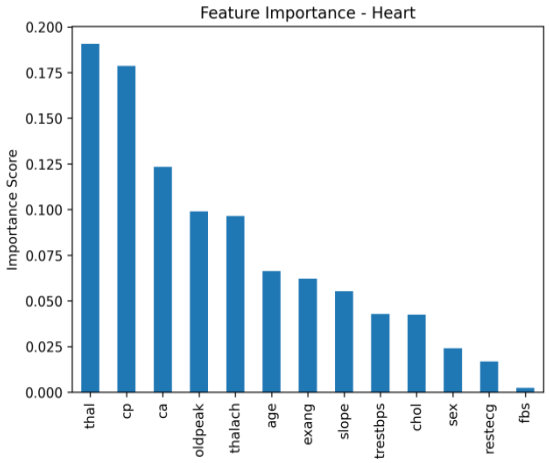


Figure 4. Feature Importance Ranking for Heart Disease Model

The confusion matrix visualization is presented in Figure 5.

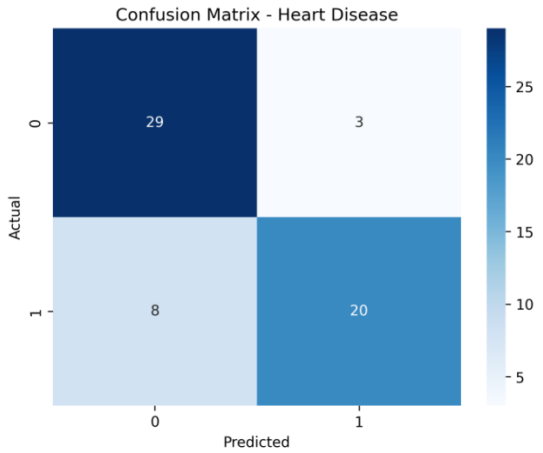


Figure 5. Confusion Matrix Visualization for Heart Disease Prediction

C. Comparative Analysis

To evaluate the robustness of the Random Forest classifier, comparative experiments were conducted using Logistic Regression and Decision Tree classifiers for both datasets.

Table 3. Comparative Analysis of Classification Algorithms

Model	Diabetes Accuracy	Heart Accuracy
Random Forest	75.97%	81.67%
Logistic Regression	71.42%	83.33%
Decision Tree	72.72%	70%

Although Logistic Regression achieved slightly higher accuracy (83.33%) for heart disease prediction, Random Forest demonstrated stronger cross-validation stability and better interpretability through feature importance analysis. Decision Tree showed comparatively lower stability due to potential overfitting.

The ensemble approach provided balanced performance across both diseases, validating its suitability for multi-disease predictive frameworks.

VII. DISCUSSION

The experimental results demonstrate that the proposed multi-disease prediction framework provides reliable and interpretable performance across both datasets.

The heart disease prediction model achieved a higher ROC-AUC score (0.947) compared to the diabetes model (0.813). This difference can be attributed to stronger feature separability in the Cleveland dataset. Clinical indicators such as chest pain type, thalassemia, and number of major vessels create clearer classification boundaries compared to metabolic indicators in diabetes prediction.

For diabetes prediction, the model showed strong performance in identifying non-diabetic cases while maintaining moderate recall for diabetic cases. False negatives in medical diagnosis are clinically significant, indicating opportunities for further optimization through threshold tuning or cost-sensitive learning approaches.

Feature importance rankings aligned closely with established clinical knowledge:

- Diabetes: Glucose, BMI, Age
- Heart Disease: Thalassemia, Chest Pain Type, Major Vessels

This alignment enhances interpretability and increases trust in the predictive framework. Cross-validation standard deviation values (±0.026 for diabetes and ±0.031 for heart disease) confirm stable generalization performance.

Compared to prior studies focusing on single-disease classification, this research integrates dual-disease prediction within a unified framework while maintaining privacy-preserving local deployment. This integrated approach enhances practical applicability in preventive healthcare systems.

VIII. LIMITATIONS AND FUTURE WORK

While the framework shows strong promise, it is important to recognize a few limitations. First, both the PIMA and Cleveland datasets are relatively moderate in size and may not fully capture the diversity of global populations. Expanding to larger, more varied datasets would help improve generalization and make the system more universally applicable. Second, the heart disease prediction task was simplified into a binary classification problem—presence or absence of disease. In reality, heart disease severity exists on a spectrum, and future work could extend this approach to multi-class prediction for more nuanced clinical insights. Third, although Random Forest provides interpretability through feature importance, more advanced explainable AI methods such as SHAP (SHapley Additive Explanations) could offer deeper transparency by quantifying the contribution of each feature to individual predictions. Addressing these limitations in future research would strengthen the framework's clinical relevance, scalability, and trustworthiness.

Since the diabetes dataset is restricted to female participants, the model's ability to generalize across genders may be limited.

Future work may include:

- Hyper-parameter optimization using Grid Search or Bayesian Optimization
- Integration of deep learning hybrid ensemble models
- Real-time wearable sensor data integration
- Deployment validation in clinical environments
- To make the framework more widely applicable, future research could draw on datasets that better reflect diversity—balancing gender and including participants from multiple ethnic backgrounds.

IX. CONCLUSION

This research introduced a privacy-preserving AI framework for predicting risks of both diabetes and heart disease using Random Forest ensemble learning. The system achieved strong results: the diabetes model reached 75.97% accuracy with a ROC-AUC of 0.813, while the heart disease model performed even better with 81.67% accuracy and a ROC-AUC of 0.947. Cross-validation confirmed the stability of these outcomes, and feature importance analysis provided interpretability by highlighting clinically relevant risk factors.

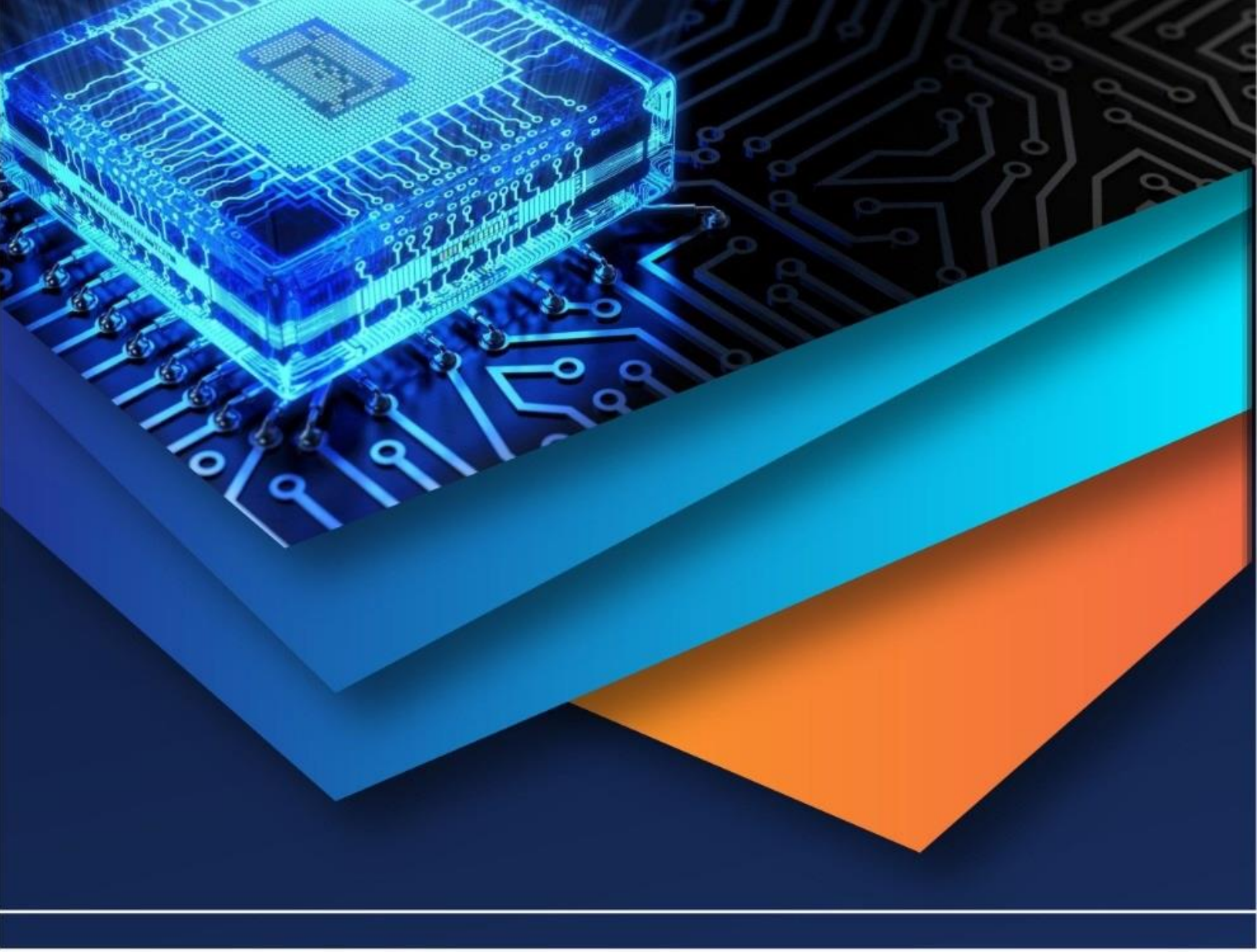
By integrating dual-disease prediction into a single system and ensuring that all data is processed locally, the framework balances predictive reliability with privacy protection. This combination makes the approach practical for real-world preventive healthcare, where both accuracy and trust are essential. Ultimately, the study demonstrates how AI-driven analytics can support early detection, empower clinicians, and strengthen decision-making in managing chronic diseases.

X. ACKNOWLEDGEMENTS

The authors extend their heartfelt thanks to the Department of Computer Science & Engineering at Integral University, Lucknow, for their invaluable guidance and technical support throughout the course of this research. Their mentorship and resources played a crucial role in shaping the study, ensuring both academic rigor and practical relevance. This acknowledgment reflects the collaborative effort behind the work and the importance of institutional support in advancing meaningful research outcomes.

REFERENCES

[1] Breiman, L. "Random Forests." *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[2] UCI Machine Learning Repository. "Heart Disease Dataset." University of California, Irvine. Available: https://archive.ics.uci.edu
[3] Kaggle. "PIMA Indians Diabetes Database." Available: https://www.kaggle.com
[4] World Health Organization. *Global Report on Diabetes*. WHO Press, Geneva, 2016.
[5] Rajkomar, A., Dean, J., Kohane, I. "Machine Learning in Medicine." *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
[6] Deo, R. "Machine Learning in Medicine." *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.
[7] Topol, E. "High-performance medicine: the convergence of human and artificial intelligence." *Nature Medicine*, vol. 25, pp. 44–56, 2019.
[8] Beam, A. L., Kohane, I. S. "Big Data and Machine Learning in Health Care." *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.
[9] Lundberg, S. M., Lee, S.-I. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[10] Johnson, A. E. W., et al. "Reproducibility in Machine Learning for Health." *NPJ Digital Medicine*, vol. 2, Article 77, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY