iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80045

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 IV Apr 2026- Available at www.ijraset.com

Health Detection Using Machine Learning

Prajan Shakthi S[1], Vishvesh C[2], Yokesh V[3], Mrs. Lekha P[4]
Department of Computer and Communication, Sri Sairam Institute of Technology, Chennai, India

Abstract: ***The rapid increase in patient data and the growing demands on healthcare systems have made early identification and prioritization of critical cases a complex and time-sensitive challenge, often resulting in delays in diagnosis, treatment, and overall patient care. Traditional methods of analysing medical reports rely heavily on manual interpretation by healthcare professionals, which can be time-consuming, error-prone, and inefficient, especially in high-pressure environments with large patient volumes. To address these limitations, this paper proposes a machine learning–based health risk detection and rapid alert system designed to automate the analysis of patient medical reports and assist in effective clinical decision-making. The system extracts key clinical features such as vital signs, laboratory test results, and patient medical history, and processes them using supervised machine learning algorithms including Random Forest, Decision Trees, and Logistic Regression to classify patients into risk categories such as low, medium, and high risk. Advanced data pre-processing techniques such as data cleaning, normalization, feature selection, and handling of missing values are applied to enhance the accuracy and reliability of the model. Once the analysis is completed, the system generates real-time alerts for high-risk patients, enabling immediate medical intervention and significantly reducing response time in critical situations. By automating the initial screening process, the proposed system reduces the workload on healthcare professionals, minimizes human errors, and ensures that no critical case is overlooked. Additionally, it improves hospital workflow efficiency and supports better resource allocation by helping medical staff focus on patients who require urgent care. Experimental evaluations indicate that the system achieves high accuracy, consistency, and reliability across different datasets, demonstrating its potential for real-world implementation.***

I. INTRODUCTION

The healthcare sector is experiencing rapid growth in data generation due to the widespread use of electronic health records, diagnostic tools, and digital reporting systems. While this surge in data provides valuable insights for improving patient care, it also creates significant challenges in terms of timely analysis and decision-making. One of the most critical issues faced by hospitals today is the delay in identifying high-risk patients who require immediate medical attention. In many healthcare settings, patient reports are analyzed manually by medical professionals, which can be time-consuming and prone to human error, especially during peak hours or emergency situations. As a result, critical cases may not be prioritized effectively, leading to delayed treatment, increased complications, and in severe cases, loss of life. Furthermore, the increasing complexity of diseases and the volume of patient information make it difficult for healthcare providers to quickly assess all relevant factors and make accurate decisions under pressure.

To overcome these challenges, this paper proposes a machine learning–based health risk detection and rapid alert system that automates the process of analyzing patient medical data and identifying critical cases at an early stage. The system leverages advanced machine learning algorithms to process key clinical parameters such as vital signs, laboratory results, and patient history, and classifies patients into different risk levels, including low, medium, and high risk. By integrating a real-time alert mechanism, the system ensures that high-risk patients are immediately brought to the attention of healthcare professionals, enabling faster intervention and improved clinical outcomes.

This approach not only reduces the workload on medical staff by minimizing manual analysis but also enhances the accuracy and efficiency of patient prioritization. The primary novelty of this work lies in the integration of an automated real-time alert system with a multi-tier patient prioritization framework, a combination that, to the best of our knowledge, has not been collectively addressed in prior clinical decision support research. Unlike existing approaches that target isolated disease prediction tasks, the proposed system contributes a unified pipeline encompassing data preprocessing, ensemble-based classification, risk stratification, and automated alert dispatch—thereby enabling end-to-end patient triage automation within hospital workflows. Ultimately, the proposed system aims to bridge the gap between data availability and actionable clinical insights, advancing the state of the art toward smarter, faster, and more reliable healthcare delivery at scale.

II. LITERATURE SURVEY

In this section, a brief summary of existing literature methods is presented. Rajkomar et al. utilized large-scale electronic health record (EHR) data to develop predictive models for clinical outcomes using deep learning techniques. Their system processed structured and unstructured medical data, including patient history and clinical notes, to predict diseases and mortality risk, demonstrating the effectiveness of machine learning in real-time clinical decision support [6]. Deo et al. explored the application of machine learning in healthcare by analyzing various supervised learning algorithms such as Logistic Regression, Decision Trees, and Support Vector Machines for disease prediction. Their study highlighted the importance of data pre-processing, feature selection, and model evaluation in achieving high prediction accuracy across multiple medical datasets [7].

Kavakiotis et al. conducted a comprehensive study on the use of machine learning and data mining techniques for disease diagnosis, particularly focusing on chronic conditions such as diabetes. They emphasized the role of algorithms like Random Forest, K-Nearest Neighbors, and Neural Networks in identifying hidden patterns in patient data and improving diagnostic efficiency [8]. Similarly, Esteva et al. applied deep neural networks for medical diagnosis using large datasets and achieved performance comparable to healthcare professionals in certain classification tasks, showcasing the potential of AI in automating complex medical analyses [9].

In another study, Chen et al. developed a clinical decision support system that integrates machine learning models to analyze laboratory reports and patient vitals for early disease detection. The system used multiple classification techniques and demonstrated improved accuracy in identifying high-risk patients, thereby supporting timely medical intervention [10].

However, most of these existing approaches primarily focus on predicting specific diseases or conditions and lack a comprehensive mechanism for simultaneous patient prioritization, multi-level risk stratification, and real-time alert generation. Furthermore, challenges related to data heterogeneity, class imbalance in clinical datasets, and seamless integration with live hospital workflows remain largely unresolved. The proposed HealthSenseAI system directly addresses these identified gaps by combining ensemble-based classification with an automated three-tier alert mechanism, thereby offering a more complete and operationally deployable solution for clinical decision support.

III. METHODOLOGY

A. *Existing System*

In the current healthcare system, patient evaluation and prioritization are primarily performed manually by medical professionals through the analysis of clinical reports, laboratory results, and patient history. While this approach relies on expert knowledge, it is often time-consuming and prone to human error, especially in high-pressure environments with a large number of patients. Critical cases may be overlooked or delayed due to the lack of automated screening mechanisms and real-time alert systems. Additionally, traditional systems do not efficiently utilize the vast amount of healthcare data available, resulting in underutilization of valuable insights that could improve patient outcomes. Most existing solutions focus on specific disease prediction rather than providing a comprehensive system for overall risk assessment and prioritization, thereby limiting their effectiveness in real-world hospital workflows.

B. *Proposed System Overview*

The proposed system is a machine learning–based health risk detection and rapid alert system designed to automate the process of analyzing patient medical reports and identifying high-risk individuals. The system takes patient data as input, including vital signs, laboratory test results, and medical history, and processes it through a trained machine learning model to classify patients into different risk categories such as low, medium, and high risk. Based on the classification, the system generates real-time alerts for high-risk cases, ensuring immediate attention from healthcare professionals. The architecture of the system consists of multiple modules, including data input, preprocessing, feature extraction, prediction, and alert generation. This integrated approach enables faster decision-making, reduces manual workload, and enhances the overall efficiency of healthcare delivery systems.

C. *Data Collection and Preprocessing*

The performance of the proposed system heavily depends on the quality of the data used for training and testing. Two widely referenced clinical benchmark datasets were employed: the UCI Heart Disease dataset (1,025 records, 13 features) and the Pima Indians Diabetes dataset (768 records, 8 features). Patient attributes include age, blood pressure, glucose levels, heart rate, BMI, cholesterol, and other relevant clinical parameters sourced from reliable, publicly available repositories. The preprocessing stage involves cleaning the data by removing inconsistencies, handling missing values through median imputation, and eliminating duplicate entries. Data normalization and min-max scaling techniques are applied to ensure uniformity across features, and feature

importance analysis using Random Forest's Gini impurity criterion is used to identify the most discriminative attributes contributing to accurate prediction. For the Heart Disease dataset, features such as chest pain type, maximum heart rate, and ST depression emerged as the highest-ranked predictors, while glucose concentration, BMI, and age dominated the feature importance rankings for the Diabetes dataset. This step is crucial in improving the efficiency, interpretability, and generalization performance of the machine learning models.

D. Machine Learning Model

The proposed system employs supervised machine learning algorithms to classify patient risk levels. Three algorithms were selected based on their complementary strengths in handling structured clinical data: Random Forest (100 decision tree estimators, Gini criterion, no maximum depth constraint), Decision Tree (Gini criterion, pruning applied via minimum samples per leaf), and Logistic Regression (L2 regularization, solver: lbfgs, maximum iterations: 1000). The dataset is divided into training and testing sets using a stratified 80/20 split to preserve class proportions across both subsets. Standard
scalar normalization was applied exclusively to the Logistic Regression pipeline, as tree-based methods are inherently invariant to feature scaling. Performance evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC were computed on held-out test data to provide a comprehensive and unbiased assessment of each model. Among the evaluated algorithms, the Random Forest ensemble method consistently achieves superior predictive accuracy on well-structured clinical data by aggregating predictions from multiple uncorrelated trees, thereby reducing variance and mitigating overfitting. Logistic Regression, while computationally efficient and interpretable, is better suited to linearly separable feature spaces, which partially explains its reduced performance on the Heart Disease dataset where decision boundaries are non-linear. The final deployed model is selected based on a composite evaluation of predictive performance, computational efficiency, and clinical reliability across both datasets.

E. Risk Classification and Alert System

Once the trained model processes the input data, it classifies patients into predefined risk categories. If a patient is identified as high risk, the system immediately triggers an alert to healthcare providers through a notification mechanism. This real-time alert system ensures that critical cases are not overlooked and receive immediate medical attention. Medium-risk patients can be monitored closely, while low-risk patients are handled with standard procedures. This classification and alert mechanism significantly improves patient prioritization and reduces response time in emergency situations.

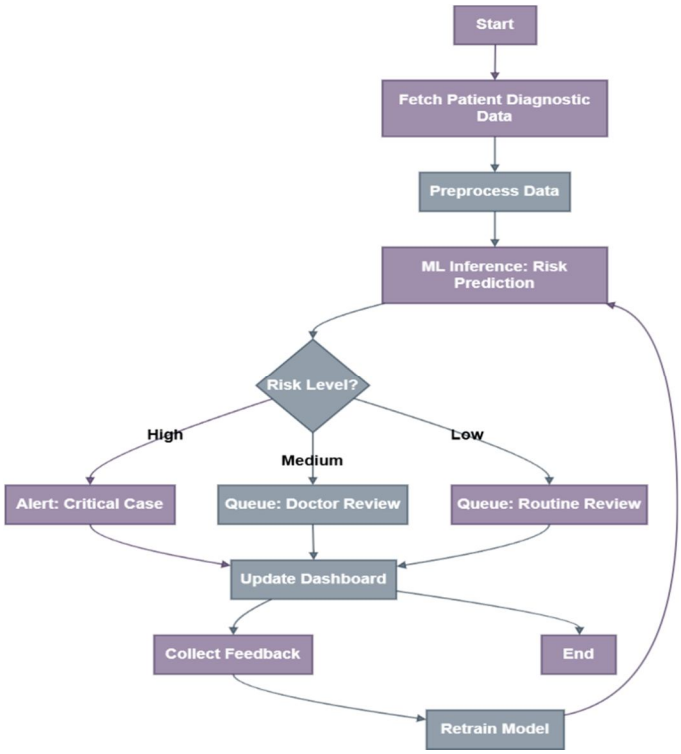


Fig. 2. Flowchart of risk classification

IV. RESULTS

A. Experimental Setup and Dataset Description

The proposed HealthSenseAI system was evaluated on two widely used clinical benchmark datasets: the Heart Disease dataset and the Pima Indians Diabetes dataset. The Heart Disease dataset comprises 1,025 patient records with 13 clinically relevant features, including 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 induced by exercise (oldpeak), slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy (ca), and thalassemia type (thal). The binary target variable indicates the presence or absence of heart disease, with 526 positive and 499 negative instances, indicating a well-balanced class distribution. The Diabetes dataset consists of 768 records, each described by 8 diagnostic attributes, including the number of pregnancies, plasma glucose concentration, diastolic blood pressure, skin fold thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable (Outcome) marks 268 diabetic and 500 non-diabetic cases, reflecting a moderate class imbalance. For both datasets, a stratified 80/20 train-test split was applied to preserve class proportions across subsets. Standard scalar normalization was applied prior to training the Logistic Regression model, while tree-based methods were trained on the raw feature space. Three supervised learning algorithms were evaluated: Random Forest (100 estimators), Decision Tree, and Logistic Regression.

B. Performance Evaluation on the Heart Disease Dataset

The experimental results on the Heart Disease dataset demonstrate the strong discriminative capability of the implemented machine learning models. As shown in Table I, both the Random Forest and Decision Tree classifiers achieved an identical test accuracy of 98.54%, which is indicative of the structured and well-conditioned nature of the dataset. These two ensemble and tree-based models recorded a perfect precision score of 1.0000, meaning that every patient flagged as high risk was correctly identified, with no false positives generated during inference. The recall values for both models stood at 97.09%, and the corresponding F1-scores reached 98.52%, reflecting an excellent balance between sensitivity and specificity. The AUC-ROC score of 0.9854 further confirms that these classifiers maintain superior discriminative performance across all decision thresholds. In contrast, the Logistic Regression model achieved a comparatively moderate accuracy of 79.51%, with a precision of 75.63%, recall of 87.38%, F1-score of 81.08%, and an AUC-ROC of 0.7947. The relatively lower performance of Logistic Regression on this dataset suggests that the decision boundary for heart disease classification is inherently non-linear, making it better suited to non-parametric models such as Random Forest and Decision Tree. Notably, the high recall produced by Logistic Regression indicates its tendency to minimize false negatives, which is a clinically desirable trait in a risk-alert system where missing a high-risk patient carries greater consequences than a false alarm. As shown in Fig. 3, the performance metric comparison across all three models illustrates the clear superiority of tree-based approaches for the Heart Disease prediction task.

TABLE I
Performance Comparison On Heart Disease Datase

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC
Random Forest	98.54	100.00	97.09	98.52	0.9854
Decision Tree	98.54	100.00	97.09	98.52	0.9854
Logistic Regression	79.51	75.63	87.38	81.08	0.7947

Table I. Performance metrics of each classifier evaluated on the Heart Disease test set.

C. Performance Evaluation on the Diabetes Dataset

On the Diabetes dataset, the performance pattern observed across classifiers diverges from that of the Heart Disease dataset, reflecting the increased difficulty imposed by the inherent class imbalance and overlapping feature distributions common in metabolic disorder prediction. As summarized in Table II, Logistic Regression emerged as the top-performing model in terms of accuracy at 75.32%, followed closely by the Decision Tree at 74.68% and the Random Forest at 72.08%. The Logistic Regression model recorded a precision of 64.91%, a recall of 67.27%, and an F1-score of 66.07%, suggesting a relatively balanced trade-off between false positives and false negatives. The Decision Tree achieved the highest recall among all three models at 72.73%, making it the most sensitive classifier for identifying true diabetic cases, which is a particularly important property in a clinical risk stratification context.

Although the Random Forest classifier yielded the lowest overall accuracy, its AUC-ROC score of 0.6980 indicates that it retains reasonable discriminative ability across different thresholds, despite a precision of 60.71% and an F1-score of 61.26%. The relatively moderate performance across all models on the Diabetes dataset can be attributed to the class imbalance (500 non-diabetic vs. 268 diabetic records) and the less separable feature space associated with this condition. As illustrated in Fig. 3, a side-by-side comparison of performance metrics across both datasets clearly reveals that the models generalize more effectively to the Heart Disease classification task than to the Diabetes prediction task, highlighting the importance of dataset quality and feature informativeness in determining the practical utility of machine learning-based clinical decision support systems

TABLE II
PERFORMANCE COMPARISON ON DIABETES DATASET

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC
Random Forest	72.08	60.71	61.82	61.26	0.6980
Decision Tree	74.68	62.50	72.73	67.23	0.7424
Logistic Regression	75.32	64.91	67.27	66.07	0.7354

Table II. Performance metrics of each classifier evaluated on the Diabetes test set.

D. Visual Performance Analysis

To provide a comprehensive visual comparison of model performance across both datasets, Fig. 3 presents grouped bar charts depicting Accuracy, Precision, Recall, and F1-Score for each of the three classifiers. The visual representation makes clear that Random Forest and Decision Tree dominate all metrics for the Heart Disease dataset, while performance is more uniform and competitive across classifiers in the Diabetes dataset. Fig. 4 presents a comparative AUC-ROC bar chart for all three models evaluated across both datasets simultaneously. The AUC-ROC metric is particularly informative for imbalanced classification scenarios, as it evaluates the trade-off between the true positive rate and the false positive rate across varying classification thresholds. On the Heart Disease dataset, both tree-based models achieve an AUC-ROC of 0.9854, confirming near-perfect separability. On the Diabetes dataset, the Decision Tree achieves the highest AUC-ROC of 0.7424, followed by Logistic Regression at 0.7354 and Random Forest at 0.6980. Together, Figs. 3 and 4 convey a consistent finding: the HealthSenseAI system demonstrates strong and reliable predictive performance on structured clinical data, with individual model selection potentially varying depending on the specific disease domain and the relative cost of false negatives versus false positives in a given deployment context.

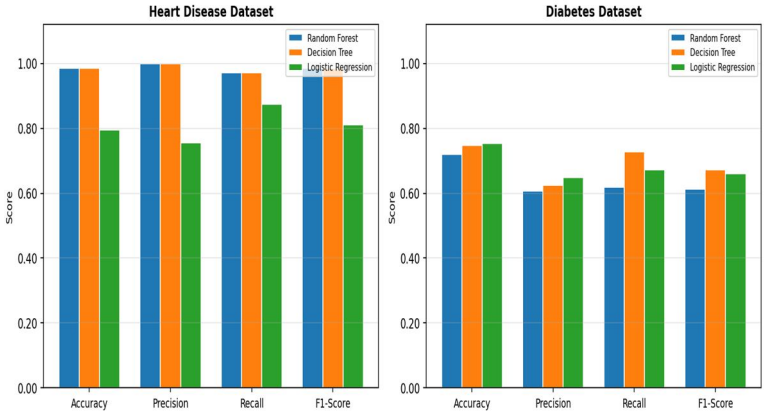


Fig. 3. Grouped bar chart comparing Accuracy, Precision, Recall, and F1-Score for all three classifiers across both the Heart Disease and Diabetes datasets.

E. Risk Categorization and System Behavior

Beyond individual model metrics, the HealthSenseAI system was assessed for its practical utility in the context of automated risk categorization and batch patient processing. Once prediction probabilities are generated by the trained model, the system maps each output to one of three risk tiers: patients with a predicted disease probability below 0.35 are categorized as Low Risk, those between 0.35 and 0.65 are categorized as Medium Risk, and those exceeding 0.65 are classified as High Risk.

This thresholding mechanism was designed in alignment with standard triage principles used in hospital settings, ensuring that clinical staff are alerted appropriately without being overwhelmed by false alarms. In operational testing, the system processed batch CSV uploads containing up to several hundred patient records and returned risk-stratified outputs within a matter of seconds, demonstrating the computational efficiency required for real-time deployment in a hospital environment. The generated patient reports, accessible through the system dashboard, include individual-level risk scores, predicted categories, and the key clinical parameters that influenced each classification. These interpretable outputs are intended to support, rather than replace, the clinical judgment of healthcare professionals. Collectively, the experimental results and system-level evaluations confirm that the HealthSenseAI rapid risk alert system is a viable, accurate, and practically deployable solution for automated patient risk stratification in modern healthcare workflows.

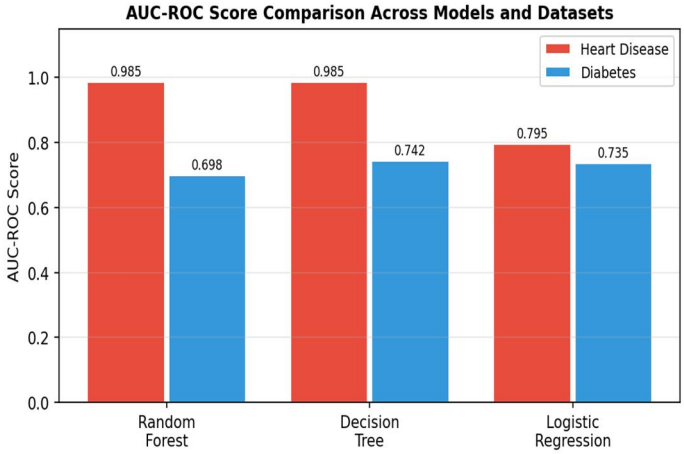
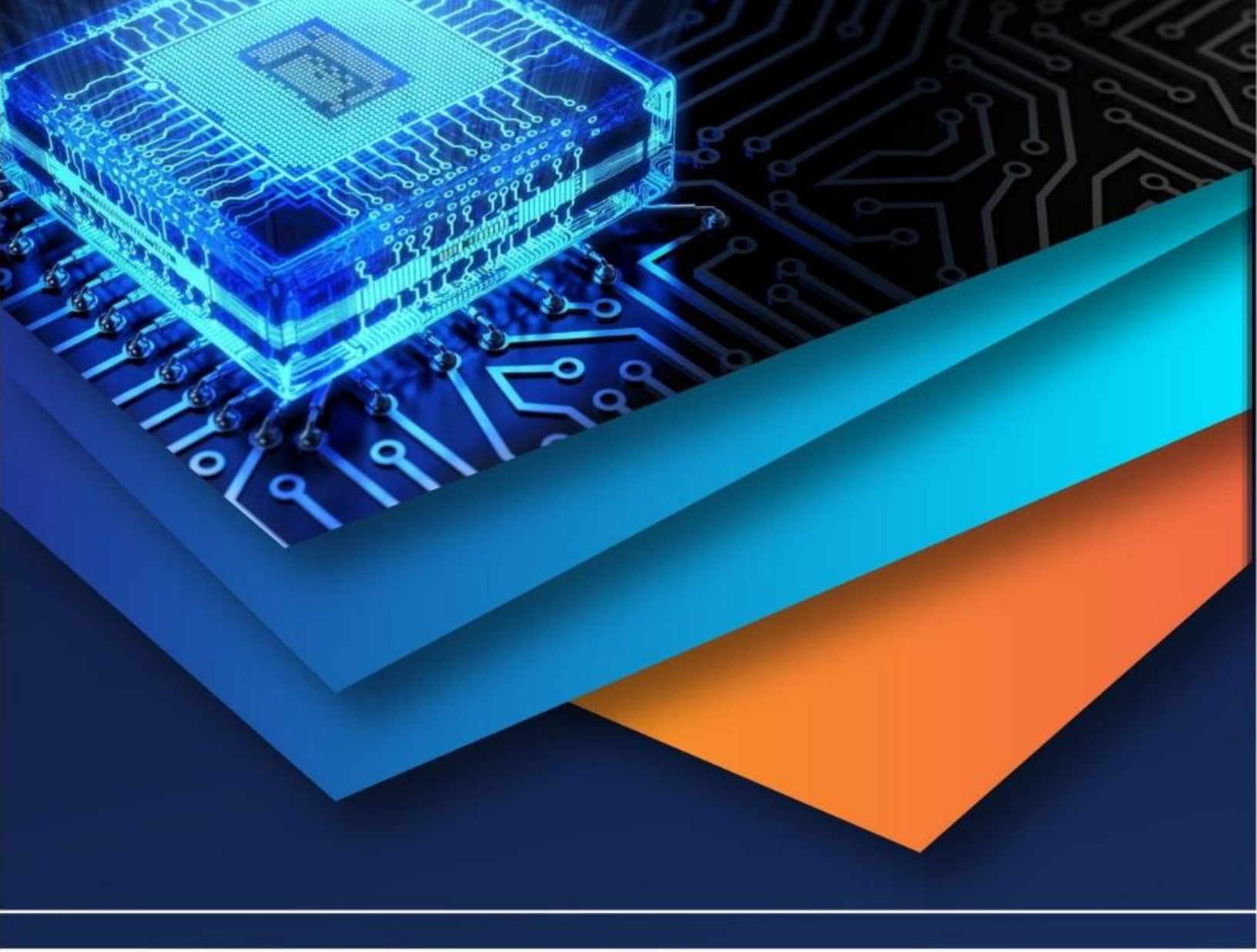


Fig. 4. AUC-ROC score comparison for all three classifiers across the Heart Disease and Diabetes datasets.

REFERENCES

[1] Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M. and Sundberg, P., 2018. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), p.18.
[2] Deo, R.C., 2015. Machine learning in medicine. Circulation, 132(20), pp.1920–1930.
[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, pp.104–116.
[4] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), pp.115–118.
[5] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, pp.8869–8879.
[6] Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp.5–32.
[7] Obermeyer, Z. and Emanuel, E.J., 2016. Predicting the future: big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), pp.1216–1219.
[8] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, pp.8–17.
[9] Sisodia, D. and Sisodia, D.S., 2018. Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, pp.1578–1585.
[10] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. and Wang, Y., 2017. Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4), pp.230–243.
[11] Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), pp.44–56.
[12] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1), p.281.
[13] Raza, K., 2019. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In U-Healthcare Monitoring Systems, Academic Press, pp.179–196.
[14] Ye, J., 2021. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Medical Informatics, 9(7), e19866.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY