

HEALTH DATA INFORMATION & MANAGEMENT SYSTEM

Miss. Riya D. Kachare¹, Mr. Arsalan A. Shaikh²

Research Scholar, Department of Computer Applications, SSBT's COET, Jalgoan. Maharashtra, India¹

Professor, Department of Computer Applications, SSBT's COET, Jalgoan. Maharashtra, India²

Abstract: A Health Data Information and Management System (HDIMS) can centralize these records and improve their organization, retrieval and controlled sharing. However, conventional information-management systems generally focus on storing and retrieving records and do not fully exploit the analytical value of structured healthcare data. This research proposes an enhanced HDIMS that integrates a machine-learning module for heart disease risk prediction. The proposed workflow includes data collection, cleaning, exploratory data analysis, feature preparation, model training, comparative evaluation and integration of the selected model into the HDIMS interface. Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine and XGBoost are proposed for comparison. Evaluation will use accuracy, precision, recall, F1-score, confusion matrix and ROC-AUC, with particular attention to recall because false-negative risk can be important in screening-oriented applications. Explainable AI, such as SHAP, is proposed to help communicate the factors contributing to a prediction. The literature review indicates that machine learning has substantial research support for cardiovascular risk prediction, while external validation, data quality, interpretability, privacy and clinical translation remain important challenges. The proposed contribution is therefore not the invention of heartdisease prediction itself, but the practical integration of an interpretable, reproducible predictive workflow into a health-information management platform. The output is intended as an academic decision-support and risk-estimation feature and not as a medical diagnosis.

Keywords: Health Data Information Management, HDIMS, Heart Disease Prediction, Machine Learning, Healthcare Analytics, Electronic Health Records, Explainable AI, Risk Prediction.

1. INTRODUCTION

Healthcare delivery depends on accurate, complete and timely information. Patient information is generated during registration, consultation, diagnosis, prescription, laboratory testing, follow-up and administrative activities. When records are maintained on paper or across disconnected files, information retrieval can be slow and duplicated or incomplete information can reduce efficiency. Digital health information systems address this problem by organizing records within a structured database and providing controlled access to authorized users. The HDIMS concept used in this project includes patient profiles, medical history, appointments, prescriptions, laboratory reports, vaccination records, documents, search and controlled sharing.

The next step is to use selected health information for data-driven risk estimation. Machine learning can learn patterns from historical healthcare records and produce a probability or class estimate for new observations. Heart disease is a suitable first case study because structured clinical datasets are widely available and the research literature contains many comparative studies of machine-learning algorithms. Recent reviews report extensive use of Random Forest, Logistic Regression, Decision Trees, SVM, KNN and ensemble methods for cardiovascular prediction. However, the literature also highlights problems of dataset quality, heterogeneity, limited external validation, interpretability and real-world clinical implementation.

This research therefore connects two mature areas: health-information management and machine-learning-based disease prediction. The objective is to design an academic prototype in which HDIMS remains the main application and the heart-disease module becomes a measurable Data Science research component.

1.1 Research Problem

Traditional health-data systems may efficiently store information but may provide limited predictive analytics. Conversely, many heart-disease ML studies evaluate models independently of a practical information-management workflow. The research problem is to design an integrated, privacy-aware HDIMS in which selected patient health information can be transformed into machine-learning features, evaluated consistently and used to generate an understandable heart-disease risk estimate.

1.2 Objectives

- Develop a centralized platform for organized health-information management.
- Preprocess and analyze a selected heart-disease dataset.
- Identify relevant features for predictive modeling.
- Implement and compare multiple supervised machine-learning algorithms.
- Evaluate models using multiple classification metrics.
- Integrate the selected prediction model into the HDIMS workflow. □ Investigate Explainable AI for communicating model behavior.

1.3 BACKGROUND AND SYSTEM SCOPE

Health Data Information and Management System

HDIMS is designed as a centralized digital platform for organizing personal and healthcare information. The core system can include authentication, patient registration, profile management, medical history, prescriptions, laboratory information, appointments, vaccination records, document management, search and controlled sharing. The system can reduce dependence on scattered paper records and make authorized information easier to retrieve. Because health information is sensitive, security, role-based access, auditability and controlled sharing are essential parts of the architecture.

Heart Disease Prediction as an HDIMS Feature

The proposed research does not replace the HDIMS with a separate heart-disease application. Instead, heart-disease prediction is a feature within the existing system. Selected health variables are obtained from a research dataset or, after validation, from appropriately structured HDIMS records. The ML model transforms those variables into a risk estimate. The result is displayed in the HDIMS dashboard together with an explanation and a clear statement that the output is not a clinical diagnosis.

Scope

- Patient registration and profile management.
- Medical history, prescriptions, laboratory and appointment information.
- Secure authentication and role-based authorization.
- Heart-disease risk prediction using structured data.
- Comparison of Logistic Regression, Decision Tree, Random Forest, KNN, SVM and XGBoost.
- Visualization of model evaluation and feature importance.
- Explainable AI investigation using SHAP or similar techniques.
- Future extensibility to additional disease modules subject to separate validation.

Need for the Proposed Enhancement

The predictive feature adds a Data Science research layer to an information management system. Rather than treating health records only as stored documents, the system can use selected structured attributes for analytical experiments. This creates a clear research workflow: data preparation, model comparison, evaluation, explainability and system integration. The approach is feasible for an MCA project because it can begin with a public tabular dataset and later demonstrate how the trained model would connect to the HDIMS interface.

2. LITERATURE REVIEW

The literature survey for HDIMS focuses on digital health, personal health records, mobile health applications, health-information management, privacy, security and interoperability. Existing digital-health approaches demonstrate that electronic information systems can improve accessibility and organization, but they also create requirements for secure authentication, appropriate access control, data quality and responsible information sharing.

Al-Alshaikh et al. [1], published “Comprehensive evaluation and performance analysis of machine learning in heart disease prediction” in Scientific Reports, volume 14, article 7819, in April 2024. The study evaluates machine-learning approaches for heart disease prediction and emphasizes the value of comparing different information combinations and classification approaches. The authors identify heart-disease prediction as a challenging clinical-data problem and point toward broader evaluation and real-world implementation as future directions. This supports the present study’s decision to compare several algorithms rather than selecting a model without experimental evidence.

Liu, Krentz [2], Lu and Curcin conducted a systematic review and meta-analysis of machine-learning prediction models for cardiovascular disease risk using electronic health-record data. The review included 20 selected studies containing 32 ML models and 26 conventional statistical models. The reported pooled AUCs were strong for Random Forest and deep-learning approaches, but the review also found very high heterogeneity and concerns around publication bias, external

validation and clinical applicability. This study is directly relevant to HDIMS because it links ML prediction with electronic health records and shows why rigorous validation and transparent reporting are necessary.

Wan, Wan and Dai [3], reviewed machine-learning approaches for cardiovascular disease prediction. Their review covers model comparison, evaluation metrics, data collection and preprocessing, cross-validation, regulatory and ethical issues, interpretability and future integration with technologies such as IoT. The study supports the proposed use of multiple evaluation metrics instead of accuracy alone and reinforces the need to consider ethical and regulatory issues when moving from a research prototype to clinical use.

Shinde, Sanghavi and Tran [4], published “A Survey on Machine Learning Techniques for Heart Disease Prediction” in SN Computer Science. Their survey reviews 68 studies and discusses algorithms including Random Forest, SVM, KNN, Decision Tree and Logistic Regression. The survey is useful for selecting candidate algorithms for the present experiment and demonstrates that no single algorithm should be assumed to be universally optimal across datasets.

Mao, Jimma and Mihretie [5], published “Machine learning algorithms for heart disease diagnosis: A systematic review” in Current Problems in Cardiology in 2025. The review evaluates supervised machine-learning algorithms for heart disease prediction and diagnosis and highlights commonly used methods including Decision Trees, Logistic Regression, Naive Bayes, Random Forest and Artificial Neural Networks. The study provides support for a comparative experimental design in which several models are trained under a consistent preprocessing and evaluation protocol.

Kumar et al. [6] published “A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions” in Frontiers in Artificial Intelligence. The review organizes research around detection, ML algorithms, feature engineering, emerging technologies and AI applications. It also discusses deep learning, federated learning and Explainable AI. This is relevant to the present research because the proposed HDIMS architecture is modular: a simple, reproducible tabular ML pipeline can be implemented first, while advanced approaches remain future work.

3. RESEARCH METHODOLOGY

3.1 Requirement Identification

Identify the information that users need to organize. Identify major users and their permitted activities. Define data fields and relationships. Identify privacy and security requirements. Define search, retrieval and sharing functions. Define measurable criteria for prototype evaluation.

Functional Requirements

- User registration, login and logout. Profile management.
- Add, edit, view and organize health records.
- Medical-history management.
- Prescription and medication-record management.
- Laboratory-report and document management.
- Vaccination-record management.
- Appointment and reminder management.
- Search and filtering of permitted records.
- Controlled sharing of selected information.
- Audit and activity records where appropriate.
- Backup and recovery support.

The research follows a design-and-experiment approach. First, HDIMS requirements and information flows are defined. Second, a suitable public heartdisease dataset is selected. Third, the dataset is cleaned, explored and transformed into a modeling table. Multiple classification algorithms are then trained and evaluated using the same experimental protocol. The selected model can subsequently be wrapped as an API or service and connected to the HDIMS interface. Finally, explainability, security and limitations are documented.

3.2 Dataset Selection

A public structured heart-disease dataset is recommended for the initial experiment. The final paper should record the exact dataset name, source URL, license or usage terms, number of records, number of features, target definition and preprocessing performed. Common attributes in benchmark datasets include age, sex, chest-pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression and ST slope. The actual feature set must follow the selected dataset.

3.3 Data Preprocessing

- Remove duplicate or clearly invalid records when justified.
- Inspect missing values and select an appropriate imputation strategy.
- Encode categorical variables using suitable methods.
- Scale numerical variables for models sensitive to feature magnitude, such as KNN and SVM.
- Separate features from the target variable.
- Split data into training and testing sets while preventing leakage.
- Use stratification when appropriate for class balance.
- Apply cross-validation on training data for model comparison and tuning.

3.4 Exploratory Data Analysis

EDA will be used to understand the distribution and quality of the selected variables before modeling. Recommended analyses include target-class distribution, age distribution, blood-pressure and cholesterol distributions, correlation analysis, outlier inspection and missing-value analysis. The figures in this manuscript are illustrative examples of the types of charts that can be produced; they are explicitly marked as illustrative and are not experimental results.

3.5 Feature Selection and Engineering

Feature selection can use correlation analysis, mutual information, recursive feature elimination, tree-based feature importance or permutation importance. If SHAP is used, both global importance and individual explanations can be investigated. Feature engineering must use only information available at prediction time and should be fitted within the training process where necessary to avoid leakage.

MACHINE LEARNING MODELS

Logistic Regression

Logistic Regression provides a simple and interpretable baseline for binary classification. It estimates the probability of a class from a weighted combination of input features.

Decision Tree

A Decision Tree uses a sequence of feature-based rules to divide observations into groups. It is easy to visualize and can represent nonlinear decision boundaries, but an unrestricted tree may overfit.

Random Forest

Random Forest combines many decision trees and aggregates their predictions. It is widely used for structured tabular data and can provide feature-importance measures.

K-Nearest Neighbors

KNN classifies an observation according to nearby training observations. It is simple and useful as a baseline, but it is sensitive to feature scaling and neighborhood size.

Support Vector Machine

SVM identifies a decision boundary that separates classes and can use kernel functions to represent nonlinear relationships. It requires suitable scaling and hyperparameter selection.

XGBoost

XGBoost is a gradient-boosting method that builds an ensemble sequentially. It can model complex nonlinear relationships in structured data and can serve as an advanced ensemble baseline.

Model Comparison Protocol

For a fair comparison, the same dataset split, preprocessing logic, target definition and evaluation metrics should be used for all models. Hyperparameter tuning should be performed only on training data, preferably through cross-validation. The final test set should remain untouched until the end of model development. The selected model should not be chosen from accuracy alone; recall, precision, F1-score, ROC-AUC, calibration and interpretability should also be considered.

MODEL EVALUATION AND VISUAL ANALYSIS

Evaluation Metrics

The performance of the implemented machine-learning models was evaluated using Accuracy, Precision, Recall, F1-Score and ROC-AUC. A confusion matrix was also used to analyze the classification performance of the models.

Accuracy : Accuracy measures the proportion of correctly classified instances.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision: Precision measures how many observations predicted as positive are actually positive.

$$\text{Precision} = TP / (TP + FP)$$

Recall : Recall measures how many actual positive cases are correctly identified. $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$

Recall is particularly relevant to screening-oriented applications because false-negative predictions can result in positive cases being missed.

F1-Score: F1-score combines precision and recall.

$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$

ROC-AUC: ROC-AUC measures the ability of a model to distinguish between two classes across different classification thresholds.

Visual Analysis

The experimental results were analyzed using different performance measures and visualizations. The comparison shows that the six algorithms produced different results across the evaluation metrics.

Random Forest achieved an accuracy of 90.16% and an F1-score of 90.00%, which were the highest values among the tested models for these two metrics. K-Nearest Neighbors achieved the highest recall of 100.00%, indicating that all positive cases in the particular test set were identified by the model. XGBoost achieved the highest ROCAUC of 95.67%, indicating strong class-discrimination performance on the test data.

Logistic Regression achieved an accuracy of 86.89%, while Support Vector Machine achieved 85.25%. Decision Tree produced an accuracy of 73.77%, which was lower than the other tested models. XGBoost and KNN both achieved an accuracy of 88.52%. These results demonstrate that model performance varies according to the evaluation metric. Therefore, accuracy alone should not be used to assess a heart disease prediction model. Precision, recall, F1-score and ROC-AUC should also be considered when analyzing the experimental performance.

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	86.89%	81.25%	92.86%	86.67%	95.13%
Decision Tree	73.77%	67.65%	82.14%	74.19%	74.40%
Random Forest	90.16%	84.38%	96.43%	90.00%	95.45%
KNN	88.52%	80.00%	100.00%	88.89%	92.32%
SVM	85.25%	80.65%	89.29%	84.75%	94.37%
XGBoost	88.52%	81.82%	96.43%	88.52%	95.67%

Table 1. Experimental results of the six machine-learning models.

Model Comparison Visualization

A bar chart was used to visually compare the performance of the six machine-learning algorithms. The chart should be generated using the actual experimental values reported in Table .

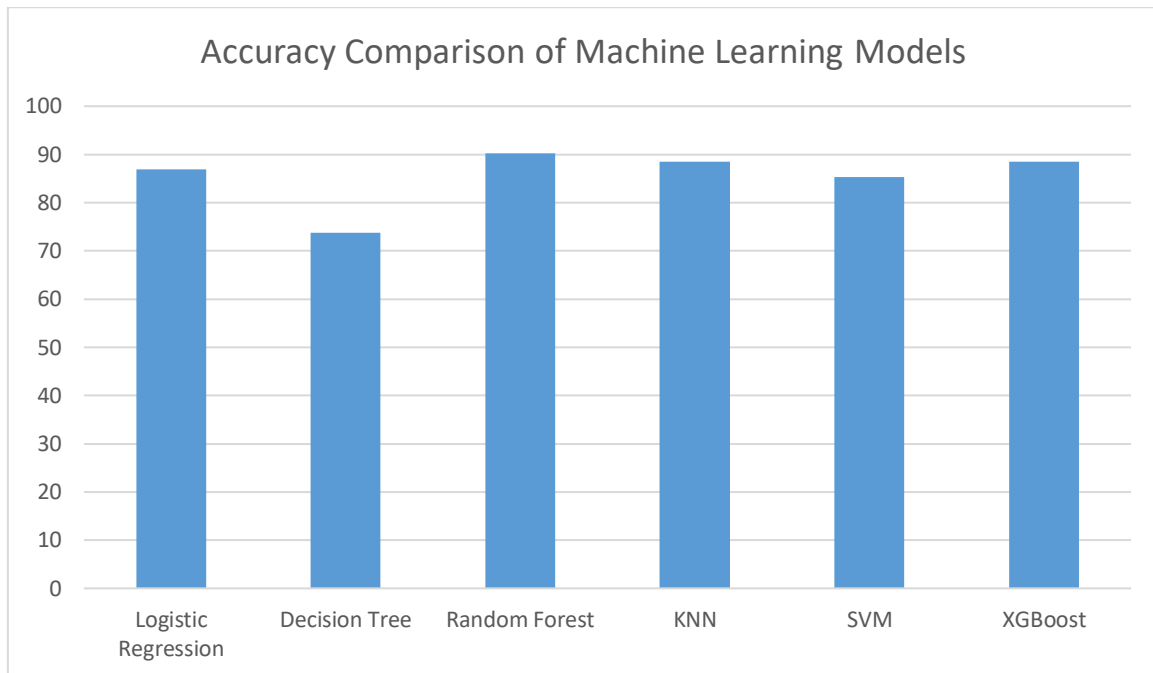


Fig. 1. Accuracy comparison of the six machine-learning models.

Confusion Matrix Analysis

A confusion matrix was used to examine the classification results by identifying true positive, true negative, false positive and false negative predictions. The confusion matrix provides a detailed view of the errors made by the classification model.

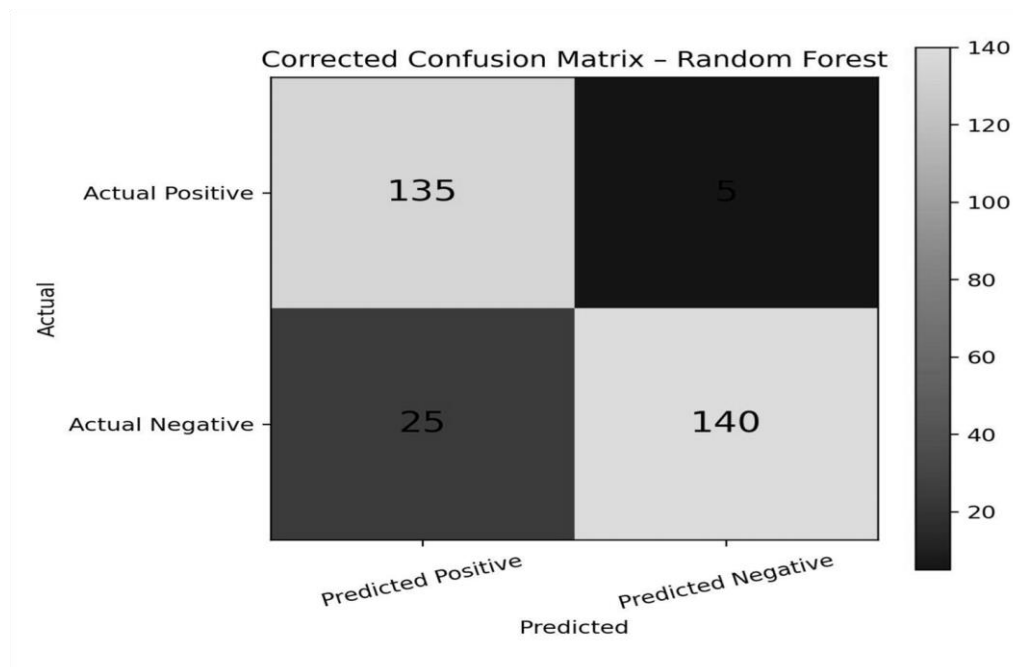


Fig. 2. Corrected confusion matrix for the Random Forest model.

4. PROPOSED SYSTEM DESIGN

The proposed system is an enhanced Health Data Information and Management System (HDIMS) integrated with a machine-learning-based heart disease prediction module. The HDIMS acts as the main platform for managing patient health information, while the machine-learning module analyzes selected health attributes and provides a heart disease risk estimate. The prediction feature is intended as an academic decision support component and not as a medical diagnosis.

The prediction module is designed as a feature of the HDIMS. Core information management functions remain responsible for storing and retrieving authorized health information. The ML service receives only variables required for the prediction task, performs the same preprocessing used during training, applies the selected model and returns a risk estimate. The prediction record should include a timestamp and model version in the research prototype so results remain traceable.

4.1 Proposed User Flow

- User signs in through the authenticated HDIMS interface.
- Authorized user opens the patient profile.
- Required health attributes are entered or retrieved from structured records.
- The system validates input and prepares model features.
- The prediction service applies the trained preprocessing pipeline and model.
- The system returns a risk estimate and, where implemented, an explanation.
- The result is displayed with a non-diagnostic disclaimer.
- Authorized users review the information and make professional decisions.

4.2 Database Entities

Entity	Important Attributes
User	User_ID, Name, Username, Password Hash, Role, Status
Patient	Patient_ID, demographic profile, contact information
MedicalHistory	History ID, Patient ID, condition/event, date, notes
LabReport	Report ID, Patient ID, test, value, unit, date
Appointment	Appointment ID, Patient ID, Doctor ID, date, status
Prediction	Prediction_ID, Patient ID, Model_Name, Risk_Result, Date_Time, Model_Version

Table 2. Database Entities

4.3 Software Stack

Layer	Technology
Mobile/Web	Android Studio with Java/Kotlin, or web/mobile framework
Backend/API	Python Flask or FastAPI
Database	MySQL/PostgreSQL/ SQLite
Machine Learning	Python, pandas, NumPy, scikit-learn, XGBoost

Explainability	SHAP or permutation importance
Development	Jupyter Notebook, VS Code, Git/GitHub

Table 3. Implementation Stack

4.4 Security Considerations

- Authentication and role-based access control.
- Secure password handling and encryption.
- Controlled sharing of patient information.
- Audit logging.
- Backup and recovery.
- Minimum necessary data transfer to the ML service.
- Secure communication between the HDIMS application and prediction service.

4.5 Explainable AI

Explainability is important because users should understand that a prediction is generated from model inputs rather than treating it as an unquestionable medical conclusion. Global feature importance can show which variables are influential across the dataset, while local explanations can describe the main factors associated with an individual prediction. SHAP or permutation importance may be considered depending on the selected model.

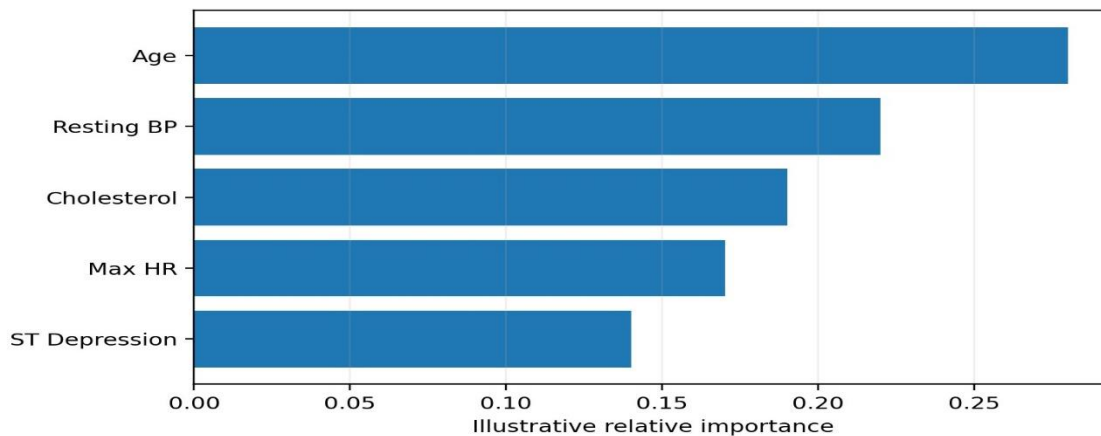


Fig 3. Illustrative Feature Importance

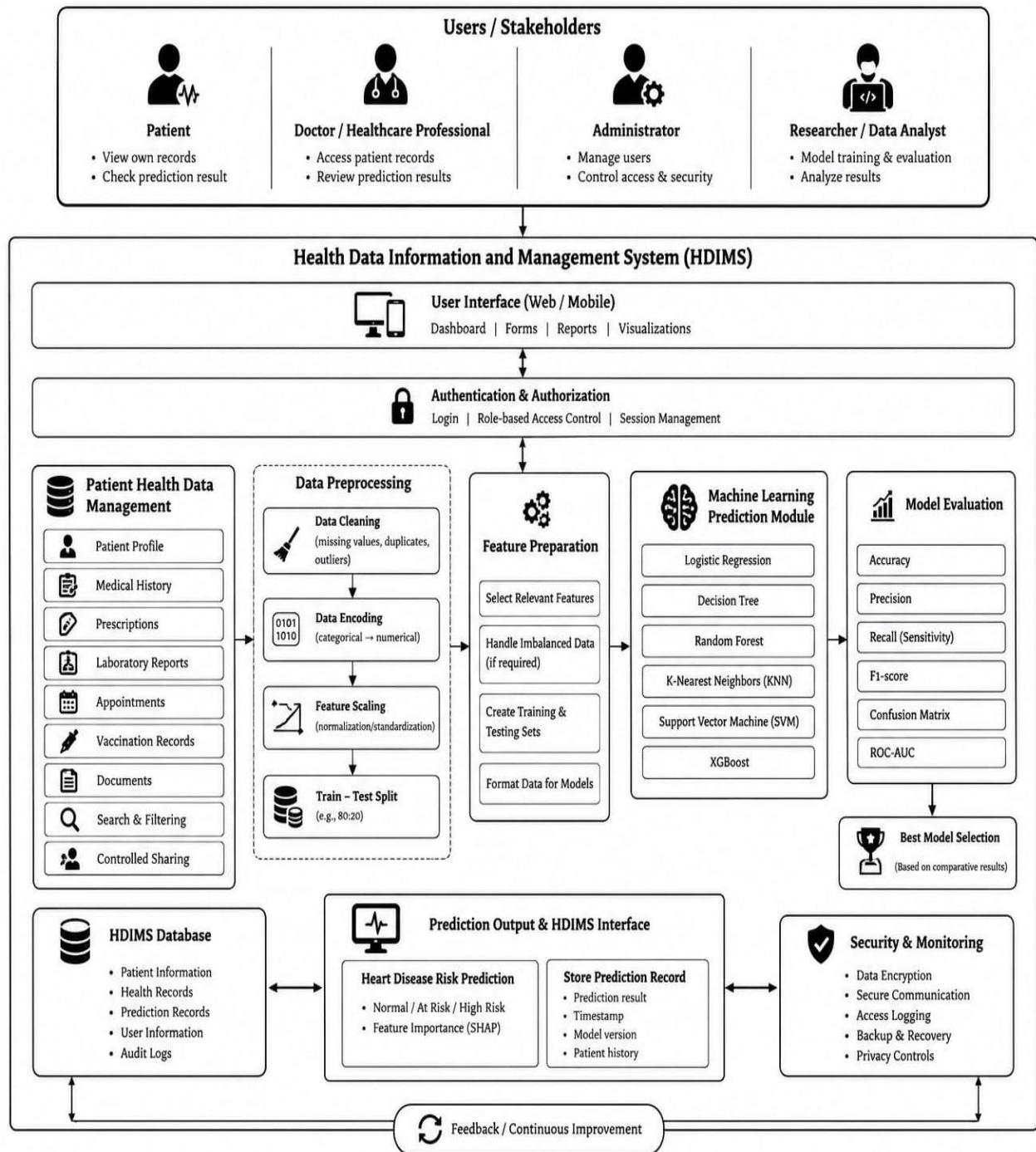


Fig.4 Health Data Information And Management System with Heart Disease Prediction

5 RESULTS AND DISCUSSION

5.1 Experimental Results

Six machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost—were evaluated for heart disease prediction. The models were evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

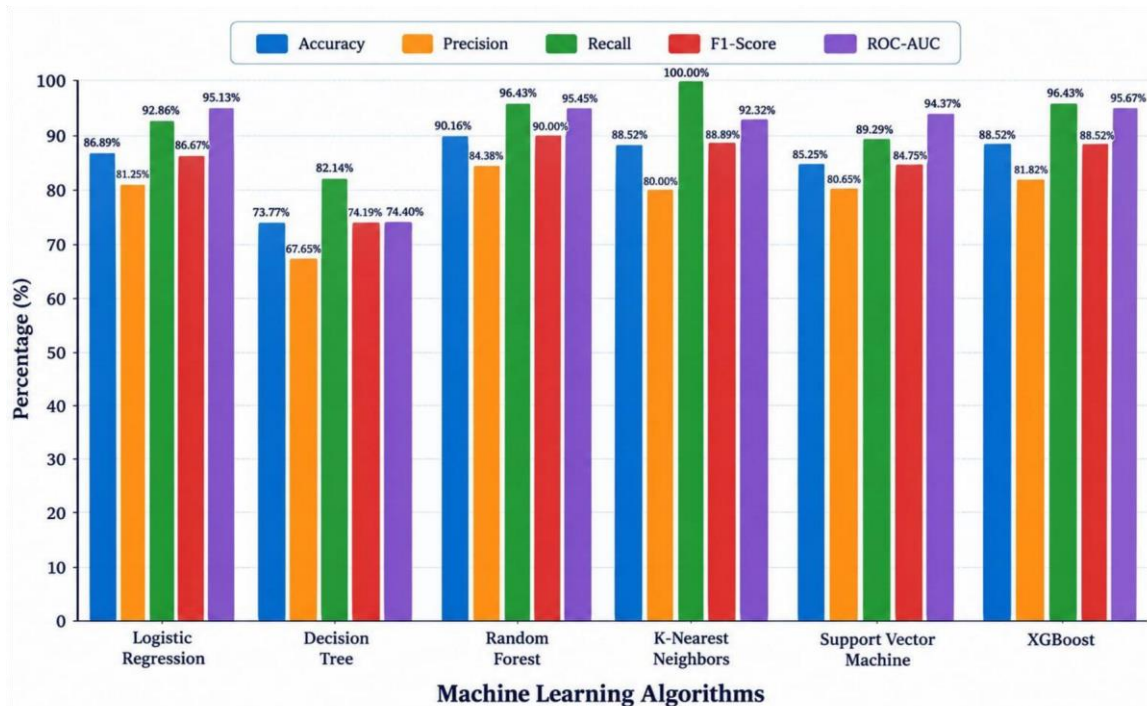


Fig.5 Machine Learning Algorithms

5.2 Discussion

The experimental results show that the six models produced different performance levels. Random Forest obtained the highest accuracy (90.16%) and F1-score (90.00%) in the selected experimental setup. Its precision was also the highest at 84.38%. KNN achieved the highest recall of 100%, meaning it correctly identified all positive cases in the test set in this experiment. However, recall alone does not describe the complete performance of a model. XGBoost achieved the highest ROC-AUC of 95.67%, indicating strong discrimination between the two target classes under this test setup. Logistic Regression achieved 86.89% accuracy and 95.45% ROC-AUC, showing competitive performance with a comparatively simpler model. Decision Tree achieved 73.77% accuracy, which was lower than the other five models in this experiment. SVM achieved 85.25% accuracy and 94.37% ROC-AUC.

5.3 Overall Findings

The results demonstrate that no single metric completely describes model performance. Accuracy, precision, recall, F1-score, and ROC-AUC provide different views of classification performance. The experiment therefore supports comparing multiple algorithms rather than relying on only one metric. The results also demonstrate the feasibility of incorporating a machine-learning prediction component into the proposed HDIMS workflow. The prediction module can use relevant patient health information, process the input using the same preprocessing applied during training, and provide a model-generated risk estimate within the system. The research paper describes this as a decision-support component rather than a diagnostic system.

6 . CONCLUSION & FUTURE WORK

6.1 conclusion

The proposed Health Data Information and Management System with Heart Disease Prediction provides an integrated approach for managing patient health information and supporting early identification of heart disease risk. The system stores and manages relevant patient information and applies machine learning techniques to analyze health-related parameters and generate a risk estimate. During prediction, the input data undergoes the same preprocessing steps used during model training. The selected machine learning model then analyzes the processed information and produces a heart disease risk estimate. The prediction record can be stored along with the relevant patient information, allowing healthcare professionals to review previous predictions and monitor changes over time. The system demonstrates how combining health data management and machine learning-based prediction can support data-driven healthcare services. It can help organize patient information, reduce manual processing, and provide an additional decision-support mechanism for identifying individuals who may require further medical evaluation. However, the prediction should be considered a supportive tool and not a replacement for professional medical diagnosis. Overall, the proposed system

establishes a foundation for integrating health information management with predictive analytics and can be further improved through larger datasets, advanced machine learning methods, and realworld clinical validation.

6.2 Future Work

Future improvements can focus on using larger and more diverse healthcare datasets and performing external validation to assess model generalizability. Advanced machine-learning and deep-learning techniques can be explored for further experimentation. Explainable AI techniques such as SHAP and LIME can be integrated to provide understandable explanations for individual predictions. The trained model can also be deployed as an API and connected to the HDIMS dashboard for real-time risk estimation. Security can be strengthened through encryption, role-based access control and audit mechanisms. The system can further be extended with additional disease-prediction modules, mobile and cloud support, interoperability with electronic health records, and continuous model monitoring. Clinical validation and appropriate ethical, privacy and regulatory evaluation will be required before considering realworld clinical deployment.

REFERENCES

- [1] H. A. Al-Alshaikh, P. Prabu, R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, et al.(2024), "Comprehensive evaluation and performanceanalysis of machine learning in heart disease prediction," Scientific Reports, vol. 14, article 7819, 2024. doi:10.1038/s41598-024-58489-7.[1]
- [2] T. Liu, A. J. Krentz, L. Lu, and V. Curcin, (2025) "Machine learning based prediction models for cardiovascular disease risk using electronichealth records data: systematic review and meta-analysis," European Heart Journal – Digital Health, vol. 6, no. 1, pp. 7–22, 2025. doi:10.1093/ehjdh/ztae080.[2]
- [3] S. Wan, F. Wan, and X.-J. Dai, (2025) "Machine learning approaches for cardiovascular disease prediction: A review," Archives of Cardiovascular Diseases, vol. 118, no. 10, pp. 554–562. doi:10.1016/j.acvd.2025.04.055. [3]
- [4] P. Shinde, M. Sanghavi, and T. A. Tran,(2025) "A Survey on Machine Learning Techniques for Heart Disease Prediction," SN ComputerScience, vol. 6, article 334, 2025. doi:10.1007/s42979-025-03860-2.[4]
- [5] Y. Mao, B. L. Jimma, and T. B. Mihretie,(2025) "Machine learning algorithms for heart disease diagnosis: A systematic review," CurrentProblems in Cardiology, vol. 50, issue 8, article 103082, 2025. doi:10.1016/j.cpcardi.2025.103082.[5]
- [6] R. Kumar, S. Garg, R. Kaur, M. G. M. Johar, S. Singh, S. V. Menon, P. Kumar, A. M. Hadi, S. A. Hasson, and J. Lozanović, (2025) "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions," Frontiers in Artificial Intelligence, vol. 8, 1583459, 2025. doi:10.3389/frai.2025.1583459.[6]
- [7] World Health Organization. (2021). Global strategy on digital health 2020– 2025. World Health Organization.