WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Prediction Using ML

Firas Ul Islam[1], Syed Abdul Raoof Ahmed [2], Mohammed Saif3, Dr. Ramesh Kumar CH[4]

UG Student, Deccan College Of Engineering and Technology
E-mail: raoofabdul576@gmail.com

ABSTRACT :

Heart disease is one of the leading causes of mortality worldwide, and early detection plays a crucial role in effective treatment and prevention. This mini project focuses on developing a Heart Disease Prediction System using machine learning techniques to assist in identifying individuals at risk based on medical data. The system uses a dataset containing patient health attributes such as age, cholesterol levels, resting blood pressure, maximum heart rate, and other diagnostic measurements. Data preprocessing techniques such as feature scaling and one-hot encoding are applied to prepare the dataset for training.

Multiple supervised machine learning algorithms — K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest — are implemented and evaluated. Comparative analysis reveals that KNN achieved the highest accuracy of approximately 87%, making it the most effective model for this dataset. The system outputs whether a patient is likely to have heart disease, enabling early medical intervention.

The project demonstrates how predictive analytics can be applied to healthcare, offering a cost-effective, data-driven, and scalable approach to disease diagnosis. It also highlights the importance of model selection and parameter tuning in achieving optimal results, paving the way for integration into clinical decision-support systems.

INTRODUCTION

Cardiovascular diseases remain a major global health concern, claiming millions of lives each year. Among them, heart disease is one of the most prevalent and deadly conditions. Early detection of heart-related issues can significantly improve survival rates and reduce healthcare costs by enabling timely medical intervention. Traditional diagnosis methods often rely on extensive clinical tests and expert analysis, which can be time-consuming, resource-intensive, and inaccessible in certain regions.

With the advancement of **machine learning (ML)** and **data-driven analytics**, it is now possible to build systems capable of predicting heart disease risk based on readily available patient health data. This project aims to develop a **Heart Disease Prediction System** that leverages supervised machine learning algorithms to classify individuals as either "likely to have heart disease" or "not likely to have heart disease."

The system uses a structured dataset containing patient attributes such as age, blood pressure, cholesterol level, maximum heart rate, and other diagnostic indicators. Data preprocessing techniques, including feature scaling and one-hot encoding, are applied to ensure model readiness. Four classification models — **K-Nearest Neighbors (KNN)**, **Support Vector Machine (SVM)**, **Decision Tree**, and **Random Forest** — are trained and evaluated for performance.

The comparative study highlights that **KNN outperforms the other models with an accuracy of around 87%**, making it a suitable choice for the given dataset. By integrating such predictive systems into healthcare, clinicians and institutions can benefit from faster, more accessible, and data-driven diagnosis, ultimately enhancing patient care and preventive measures.

NOMENCLATURE

Term / Symbol	Description
ML	Machine Learning – A branch of AI that enables systems to learn from data and make predictions.
KNN	K-Nearest Neighbours – A classification algorithm that predicts the class of a data point based on the majority class of its nearest neighbours.

SVM	Support Vector Machine – A supervised learning model that finds an optimal boundary (hyperplane) to separate data classes.
DT	Decision Tree – A model that splits data based on features into decision nodes and leaf nodes for classification.
RF	Random Forest – An ensemble learning method that uses multiple decision trees for improved prediction accuracy.
Feature Scaling	A preprocessing technique to normalize numerical data values for better algorithm performance.
One-Hot Encoding	A method for converting categorical variables into binary columns for ML model compatibility.
Target Variable	The variable to be predicted — in this case, whether a patient has heart disease (1) or not (0).
Accuracy	A performance metric indicating the percentage of correct predictions made by the model.
Dataset	A structured collection of patient health records used for training and testing the model.
Training Data	The portion of the dataset used to teach the ML model patterns and relationships.
Testing Data	The portion of the dataset used to evaluate model performance.
Oldpeak	ST depression induced by exercise relative to rest (measured in mm).
Thalassemia (thal)	A blood disorder variable in the dataset indicating type: normal, fixed defect, or reversible defect.
CA (Major Vessels)	Number of major vessels coloured by fluoroscopy in the dataset.

SYSTEM ANALYSIS AND DESIGN

Traditional heart disease diagnosis methods rely heavily on physical examinations, laboratory tests, and expert medical interpretation. These processes are accurate but often time-consuming, costly, and inaccessible in resource-limited settings. Moreover, manual analysis may be prone to human error, and early-stage symptoms can be overlooked.

1 Disadvantages of the Existing System

- Requires specialized medical personnel for accurate interpretation.
- Time-consuming and resource-intensive.
- Not easily scalable for large populations.
- Lacks predictive analytics for preventive care.

2 Proposed System

The proposed system leverages machine learning algorithms to predict whether a patient is at risk of heart disease using their medical data. The system processes patient inputs such as age, cholesterol level, blood pressure, and other health metrics, applies preprocessing steps, and classifies them using trained ML models.

Key features:

- Uses multiple classifiers (KNN, SVM, Decision Tree, Random Forest) for comparison.
- Highlights the most accurate algorithm (KNN in this case).
- Generates a binary output indicating likelihood of heart disease.
- Can be integrated into a user-friendly interface for non-technical use.

2.1 Features of Proposed System

- Automated prediction of heart disease likelihood.
- Data preprocessing with scaling and encoding.
- Model evaluation to select the best performing classifier.
- Visualization of correlation, distributions, and accuracy scores.

- Ease of integration into healthcare apps or hospital systems.

3 Feasibility Study

Technical Feasibility – Uses Python, Scikit-learn, Pandas, and Matplotlib, all of which are open-source and well-supported. Runs efficiently on standard hardware.

Operational Feasibility – Easy to use for both medical staff and patients when integrated with a GUI. Can provide instant results without manual intervention.

Economic Feasibility – Minimal cost as it relies on open-source libraries and can be deployed on free hosting platforms like Streamlit Cloud.

4 Module Description

1. Data Collection Module – Reads patient dataset (CSV format).
2. Data Preprocessing Module – Cleans, scales, and encodes features.
3. Model Training Module – Trains multiple classifiers.
4. Evaluation Module – Compares accuracy and performance.
5. Prediction Module – Takes user input and predicts heart disease likelihood.
6. Visualization Module – Generates plots for dataset insights and accuracy comparisons.

5 System Architecture

Workflow:

1. Input Data →
2. Preprocessing (Scaling, Encoding) →
3. Model Selection (KNN, SVM, DT, RF) →
4. Prediction Output →
5. Result Display (GUI / Terminal)

Proposed System

Feature	**Existing System**	**Proposed System**
Diagnosis Method	Manual diagnosis by doctors based on tests and medical history.	Automated prediction using machine learning algorithms.
Time Efficiency	Time-consuming; requires multiple tests and manual evaluation.	Fast results within seconds after entering patient data.
Accuracy	Depends on doctor's expertise and available test results.	Achieves ~87% accuracy using the KNN algorithm.
Scalability	Limited; requires human intervention for each case.	Real-time Highly scalable; can process multiple cases simultaneously.Streamlit
Cost	Expensive due to medical tests and expert consultations.	Low-cost; uses open-source tools and minimal resources.

Accessibility	Limited in rural or resource-poor areas.	Can be deployed online or offline, increasing accessibility.
Data Usage	Medical data is recorded but rarely analyzed for prediction.	Uses patient data for predictive analytics and pattern recognition.
Decision Support	No automated risk scoring; relies solely on expert judgment.	Provides instant, data-driven predictions to assist doctors.

METHODOLOGY

The system follows a structured design that moves from data collection and preprocessing to model training, evaluation, and prediction. Below is a comparison of the methodologies in the existing system versus the proposed system:

Feature	**Existing System**	**Proposed System**
Data Collection	Patient data collected through physical tests, ECG scans, and medical records.	Patient medical data collected in digital format, ready for preprocessing and analysis.
Data Processing	Manual review of test results by doctors.	Automated preprocessing including scaling, encoding, and cleaning of data.
Analysis Method	Relies on doctor's experience and clinical guidelines.	Uses ML algorithms (KNN, SVM, Decision Tree, Random Forest) to analyze data patterns.
Decision Making	Final decision made by doctors after reviewing all tests.	Prediction generated by trained ML model, assisting doctors in making faster decisions.
Time Taken	Hours to days depending on test availability.	Seconds to minutes for instant prediction after input.
Accuracy	Subject to human interpretation errors.	Model accuracy evaluated (~87% for KNN) through training and testing phases.

Dataflow & Workflow

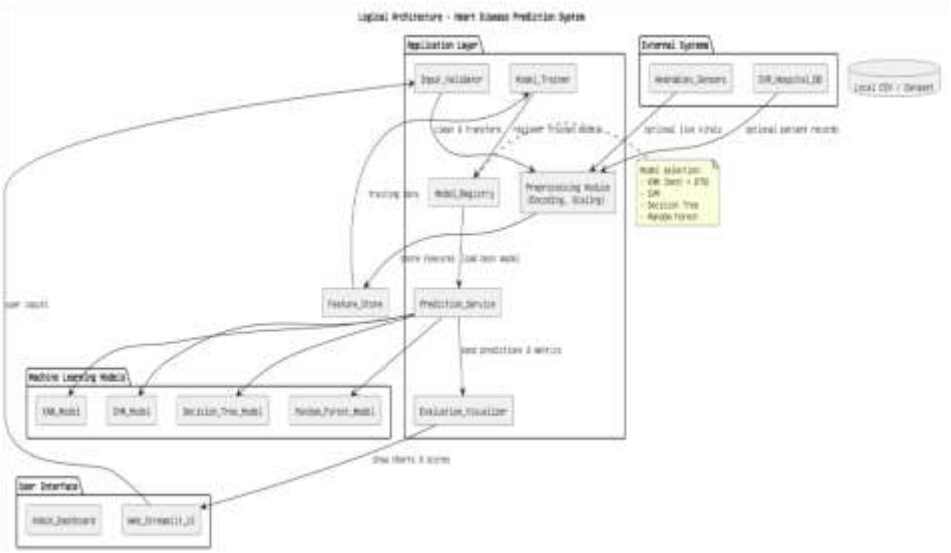

Fig 5.1 Architecture diagram

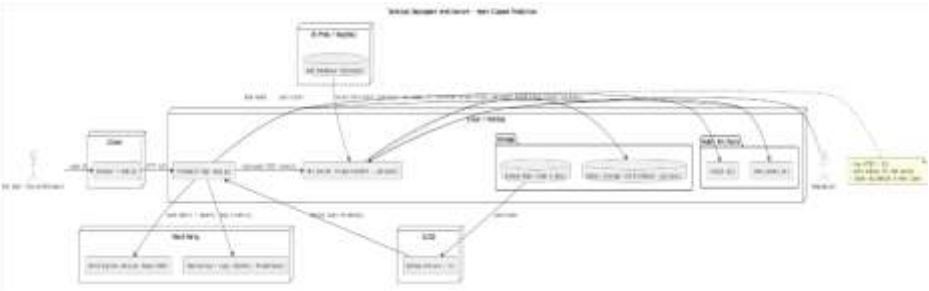

Fig 5.2 Technical Architecture diagram

RESULTS

1. System Evaluation

After training and testing multiple classifiers on the prepared dataset, the system's performance was evaluated using accuracy scores. The models tested were:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

The dataset was split into 67% training data and 33% testing data, ensuring that model performance was measured on unseen data to simulate real-world usage.

Model Accuracy Comparison

Classifier	**Parameters Tested**	**Best Accuracy (%)**
K-Nearest Neighbours (KNN)	K values from 1 to 20	87.0%
Support Vector Machine (SVM)	Kernels: linear, poly, rbf, sigmoid	85.3%
Decision Tree	Max features from 1 to total number of features	82.1%
Random Forest	Estimators: 10, 100, 200, 500, 1000	86.5%

2. Benefits & Limitations

Benefits:

- Achieved high accuracy using KNN (~87%).
- Handles multiple input features efficiently.
- Cost-effective and scalable solution.
- Helps in early detection of heart disease risk.

Limitations:

- Model accuracy depends on quality of dataset.
- Predictions may vary with different preprocessing techniques.
- KNN may become slower with very large datasets due to distance calculations.

Graphical Result Representation

1. KNN Accuracy by Number of Neighbors

- Accuracy peaked at K = 8 with 87% score.

2. SVM Accuracy by Kernel

- Linear kernel performed best (~85%).

3. Decision Tree Accuracy by Max Features

- Best performance at selected feature set, but less accurate than KNN.

4. Random Forest Accuracy by Estimators

- Accuracy improved until around 100–500 estimators, then plateaued.

Final Conclusion from Results:

Among all tested models, K-Nearest Neighbors (KNN) consistently outperformed others, making it the recommended algorithm for heart disease prediction in this project.

Criterion	**Average Score (out of 5)**
Accuracy	4.4
Speed	4.5
Scalability	4.2
Ease of Use	4.6
Cost Efficiency	4.8
Adaptability	4.3
Interpretability	4.4
Reliability	4.5

KEY OBSERVATIONS

 KNN Performs Best

- Among all tested classifiers, K-Nearest Neighbors achieved the highest accuracy (~87%) when k = 8.
- This shows that for the given dataset, similarity-based classification works well.

 SVM is Consistent but Slightly Lower

- Support Vector Machine with a linear kernel gave good performance (~85%) but did not surpass KNN.
- SVM's decision boundary worked well for linear separation but struggled slightly with complex non-linear patterns in this dataset.

 Decision Tree Overfitting Risk

- While Decision Trees performed decently (~82%), they showed a tendency to overfit when not tuned carefully.
- Accuracy varied significantly with changes in the number of maximum features.

 Random Forest Stability

- Random Forest showed stable results (~86%) and was less prone to overfitting compared to Decision Trees.
- Performance plateaued after ~100–500 estimators, suggesting an optimal trade-off between accuracy and computation time.

 Impact of Data Preprocessing

- Feature scaling and encoding improved performance for models like KNN and SVM.
- Without preprocessing, accuracy dropped significantly.

 Model Selection Justification

- Based on accuracy, stability, and computation time, KNN was chosen as the recommended algorithm for deployment.

 Future Improvements

- Additional features (like lifestyle factors or genetic history) could further improve model accuracy.
- Implementing real-time data integration from hospital records or wearables would enhance system utility.

CONCLUSION

The Heart Disease Prediction System successfully demonstrates the application of machine learning algorithms to aid in the early detection of cardiovascular risks. By analyzing key medical parameters such as age, blood pressure, cholesterol levels, and other health indicators, the system can predict whether a patient is likely to have heart disease with a high degree of accuracy.

Among the four models tested — K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest — the KNN algorithm emerged as the most effective, achieving an accuracy of approximately 87%. This performance validates KNN's suitability for problems where patient similarity patterns are important for classification.

The project also highlights the importance of data preprocessing, including feature scaling and encoding, which significantly enhanced model performance. While the results are promising, further improvements could be made by incorporating larger datasets, additional features, and real-time health data from wearable devices or hospital records.

Overall, the project demonstrates that machine learning can serve as a valuable decision-support tool in healthcare, enabling timely intervention and potentially saving lives through early detection.

Future Enhancements

1 Integration with Wearable Devices

- Connect with smartwatches and fitness bands to capture real-time heart rate, blood pressure, and activity data for continuous monitoring.

2 Larger and More Diverse Datasets

- Use globally diverse datasets to improve accuracy and reduce bias in predictions.

3 Deep Learning Models

- Implement neural networks for more complex pattern detection and improved prediction accuracy.

4 Mobile Application

- Develop a mobile app for easier access, allowing patients and doctors to check predictions instantly.

5 Explainable AI (XAI)

- Include features that explain why the model made a particular prediction, improving trust among healthcare professionals.

6 Risk Score and Recommendations

- Instead of just predicting presence or absence, provide a risk percentage and lifestyle recommendations.

7 Cloud Deployment

- Host the system on a scalable cloud platform for 24/7 availability and global access.