Heart Disease Prediction using Learning Techniques

Bhagwan Thorat[1], Shantanu Pattanshetti[2], Sahil Varde[3], Deepika Sidral[4], Saburi Nikam[5]
Bhagwan.thorat@vit.edu[1] , shantanu.pattanshetti24@vit.edu[2] , sahil.varde24@vit.edu[3]
,deepika.sidral24@vit.edu[4] , nikam.saburi24@vit.edu[5]
Artificial Intelligence and Data Science
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Abstract — Heart disease prediction is a vital area of research in healthcare, aiming to reduce the mortality rate through early diagnosis and intervention. This project focuses on developing an advanced system that leverages Artificial Intelligence and Machine Learning (AIML) techniques to predict the risk of heart disease with high accuracy. The system provides healthcare professionals and patients with a reliable tool for assessing cardiovascular health, enabling proactive management and timely treatment. The proposed model is developed using a combination of machine learning algorithms and data analytics. It performs three key stages in heart disease prediction: patient data preprocessing, feature extraction, and risk classification, ensuring a fast, real-time, and efficient predictive solution.

Keywords — Heart disease prediction, Artificial Intelligence, Machine Learning, Risk classification, Healthcare analytics

I. Introduction

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized the field of healthcare by enabling the development of predictive systems capable of identifying critical health conditions early. Among these, heart disease prediction has emerged as a significant area of research due to its potential to reduce mortality rates through timely diagnosis and intervention. Machine learning techniques, such as Decision Trees, Support Vector Machines (SVM), and Neural Networks, provide powerful tools to process large datasets and uncover patterns indicative of cardiovascular risks.

This research article explores the development and application of a machine learning-based heart disease prediction system. The primary objective is to leverage AI and ML techniques to design a model that accurately and efficiently predicts the likelihood of heart disease based on patient data, enabling early detection and preventive measures. This system aims to support healthcare professionals and patients alike by providing actionable insights that improve clinical decision- making and outcomes.

The methodology detailed in this study includes data preprocessing, feature selection, and model training using state-of-the-art machine learning algorithms. The integration of relevant patient information, such as demographic data, medical history, and diagnostic parameters, ensures a comprehensive approach to risk prediction. Furthermore, the evaluation of the system's performance is carried out using standard benchmarks and real-world datasets, demonstrating its accuracy and reliability.

II. Literature Review

Intisar Ahmed's paper proposes a heart disease prediction system that explores how machine learning algorithms can be utilized to predict heart disease. The study focuses on optimizing models to improve accuracy, minimize misdiagnosis, and detect early abnormalities in patients, ultimately benefiting both patients and the healthcare system. [1]

Similarly, the research conducted by S. Baral, Suneeta Satpathy, Dakshya Prasad Pati, and P. Mishra aims to compare and analyze the performance of different machine learning algorithms—such as Support Vector Machines, Artificial Neural Networks, Logistic Regression, Random Forest, and Decision Trees—in predicting cardiovascular diseases (CVDs). The study provides insights to support the development of a clinical decision-making tool for the early detection and prevention of heart disease, thereby reducing morbidity and mortality rates associated with these conditions. [2]

The study Heart Disease Prediction Using Machine Learning Method presents the application of machine learning techniques for predicting heart disease based on clinical information. The study compares different models, wherein Support Vector Machine (SVM) had the best accuracy rate of 92.1%. The study highlights the requirement of preprocessing data, such as feature selection and transformation, for improving model performance. The findings suggest the potential of machine learning in enhancing early detection and prevention of heart disease[3].

The research Prediction of Cardiovascular Diseases Using Neural Networks and Machine Learning compares traditional machine learning algorithms such as Decision Trees, Random Forest, and XGBoost with neural networks for the prediction of cardiovascular disease. Based on the study, ensemble learning methods, particularly Random Forest, provide the maximum predictive accuracy. The paper highlights the significance of AI in the field of medicine, demonstrating the capability of machine learning algorithms in predicting cardiovascular disease risk accurately based on basic clinical parameters[4].

Paper title	Author	PublishedDate	Technology Used	Limitations
Heart disease prediction system using machine learning	Intesar Ahmad	December2022	logisticRegressionalgorithm	Mayfaceschallengesduetoimbalanceddatasets
Evaluation and Comparison of ML Algorithms for Predicting Cardiovascular Diseases (CVDs)	S.Baral,DakshyaPrasadPati,P.Mishra	March 2024	Decision tree Algorithm, and Logistic regression algorithm	Modelsmayoverfittotriningdata
Developm	ChintanM.B	Febru	Kaggle Cardiovas	Redundancy may reduce

ent of ML Model for Predicting Cardiovascular Diseases	hatt, Parth Patel, Parth Patel, Pier Luigi Mazzeo	ary 2023	cular disease datasets decision tree algorithm and logistic regression algorithm	models Accuracy

The research article Advancements in Heart Disease Prediction: A Machine Learning Approach for Early Detection and Risk Assessment compares and contrasts different machine learning classifiers like Logistic Regression, Decision Tree, K-Nearest Neighbors, Neural Networks, and Support Vector Machine (SVM). Among them, SVM yields the highest accuracy of 91.51%. The study emphasizes the use of machine learning techniques in cardiovascular medicine in proving their potential to improve risk assessment, early detection, and personalized treatment plans[5].

The article Novel Deep Learning Architecture for Predicting Heart Disease Using CNN presents a deep learning-based model based on a 1D Convolutional Neural Network (CNN) that classifies patients as non-healthy or healthy.

The model employs various different regularization methods to prevent overfitting and achieves a very good training accuracy of 97% along with test accuracy of 96%. The research compares its performance against traditional machine learning methods and concludes that CNN-based models provide more accurate predictions in heart disease diagnosis[6].

Heart Disease Prediction Using Machine Learning	R. Kaur, V. Chauhan	September 2022	K-mean clustering	Lack of transparency making it harder for medical professionals to trust
Ensemble Framework for Cardiovascular Disease Prediction	Achyut Tiwari, Aryan Chugh, Aman Sharma	June 16, 2023	ExtraTrees Classifier, Random Forest, and XGBoost	Possible overfitting because of model complexity; needs large datasets for successful training
An Improved Heart Disease Prediction Using Stacked Ensemble Method	Md. Maidul Islam, Tanzina Nasrin Tania, Sharmin Akter, Kazi Hassan Shakib	April 12, 2023	Stacked ensemble approach integrating nine algorithms, including Random Forest, Multi-Layer Perceptron, and XGBoost	High computational intensity required; interpretation of the model can be hard
Novel Deep Learning Architecture for Heart Disease Prediction using Convolutional Neural Network	Shadab Hussain, Santosh Kumar Nanda	May 22, 2021	1D Convolutional Neural Network (CNN)	Risk of overfitting narrow generalizability to varied populations
Machine Learning and Ensemble Approach Onto Predicting Heart Disease	Aaditya Surya	November 16, 2021	Ensemble techniques combining classifiers like Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Gaussian Naive Bayes, Random Forest, and Multi-Layer Perceptron	Ensemble methods may be computationally expensive and difficult to interpret; possible issues with model interpretability
Prediction of Cardiovascular Diseases Using Neural Networks and Machine Learning	Jegadeesan Ramalingam	January 30,2023	Neural networks and Random Forest for classification	Needs good-quality feature selection; dataset bias can affect accuracy
Heart Disease Prediction Using Machine Learning Method	Mafia Rasheed Muhammad Adnan Khan	Oct-22	Data mining techniques and ML algorithms for prediction	Processing of huge datasets is heavy on resources; accuracy is data quality-dependent

Table 1. Literature Review

III. System Architecture

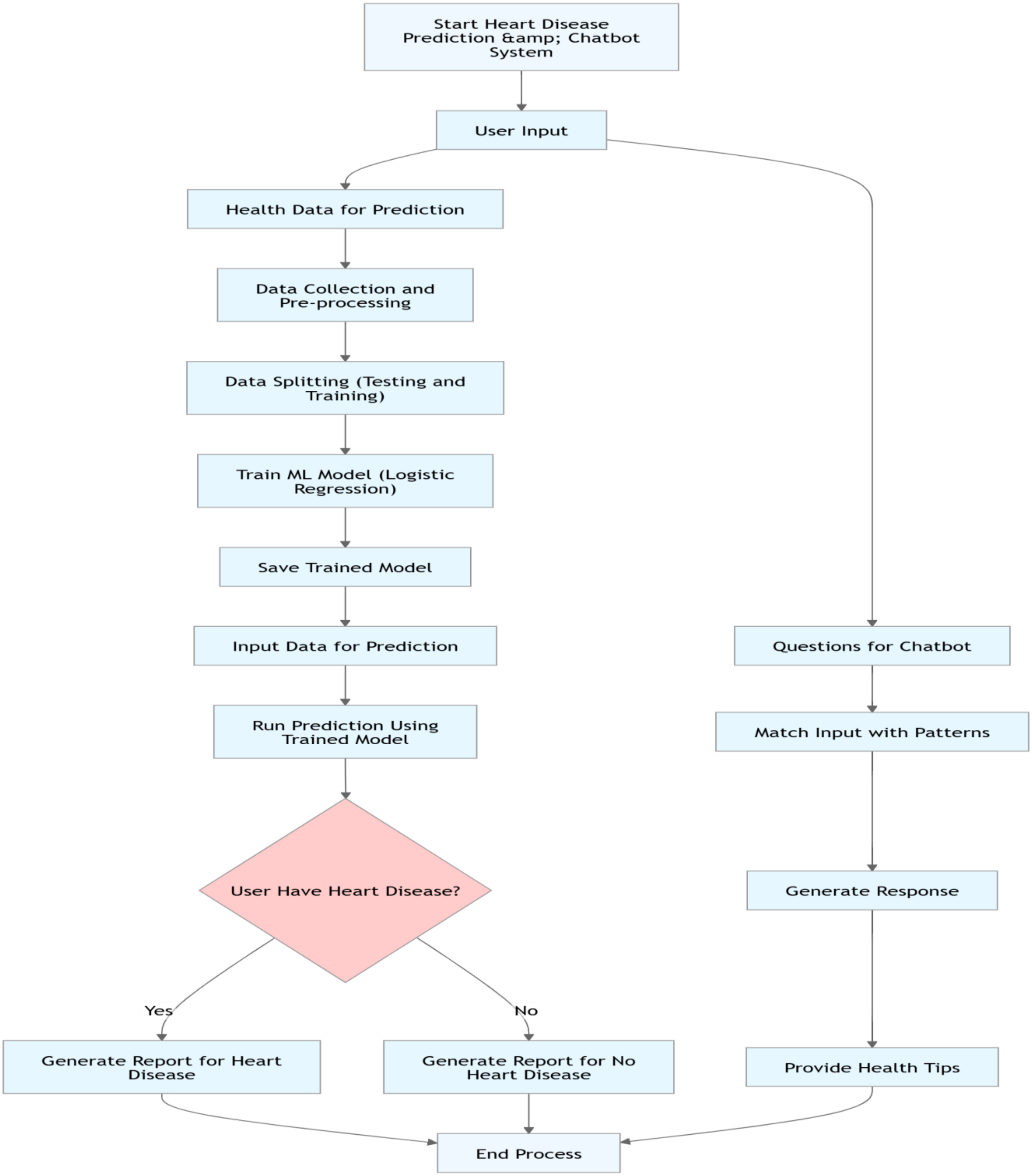

Start Heart Disease Prediction &amp; Chatbot System
User Input
Health Data for Prediction
Data Collection and Pre-processing
Data Splitting (Testing and Training)
Train ML Model (Logistic Regression)
Save Trained Model
Input Data for Prediction
Run Prediction Using Trained Model
User Have Heart Disease?
Yes
No
Generate Report for Heart Disease
Generate Report for No Heart Disease
Questions for Chatbot
Match Input with Patterns
Generate Response
Provide Health Tips
End Process

Fig. 1. Architecture of System

IV.Methodology

Logistic Regression is the supervised algorithm to be applied when the machine wants to be fed with problems where the output variables are supposed to have just two possible values- for instance 0 and 1, or Yes/No, Disease or No Disease, etc. A statistical model for estimating the conditional probability of class membership in response to some vector of predictor inputs

We choose logistic regression for this project because:

Binary Classification Problem:

The primary goal is to classify whether a patient has heart disease (1) or does not have heart disease (0) based on multiple health parameters. Logistic regression is specifically designed for such binary classification tasks.

Interpretable Model:

Unlike complex models like neural networks, logistic regression provides clear feature importance, showing how different health factors contribute to heart disease risk.

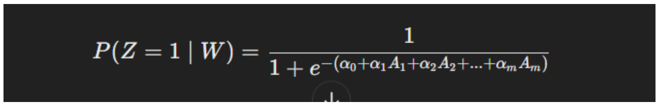
Probabilistic Output:

Instead of just generating a class label (Yes/No), logistic regression returns a probability score. For example, 0.85 means an 85% chance of having a condition. This helps doctors and patients make informed decisions.

Logistic regression is based on the **logistic (sigmoid) function**, which converts any input into a probability value between 0 and 1.

Equation of Logistic Regression

The logistic regression model predicts the probability of an outcome **Z = 1** (condition present) given input features **A_1, A_2, ..., A_m**:



$$P(Z = 1 \mid W) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + \ldots + \alpha_m A_m)}}$$

Where:

- **P(Z=1|W)** → Probability of having the condition.
- **α_0** → Intercept (bias term).

  α_1, α_2, ..., α_m → Coefficients (weights) associated with each feature.

associated with each feature.

- **A_1, A_2, ..., A_m** → Input features (e.g., age, cholesterol level, blood pressure, smoking status, etc.).
- e → Euler's number (≈2.718).

The logistic function (sigmoid function) transforms the linear equation into a probability output:

σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}

σ(z)=1+e−z1

Where:

- *y=α0+α1A1+α2A2+...+αmAmy = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + ... + \alpha_m A_my*=α0 +α1 A1 +α2 A2 +...+αm Am
- *y*= **α_0 + $\alpha_1 A_1$ + $\alpha_2 A_2$ + ... + $\alpha_m A_m$**
- *y=α0+α1A1+α2A2+...+αmAmy = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + ... + \alpha_m A_my*=α0 +α1 A1 +α2 A2 +...+αm Am
- The output of this function is always between **0 and 1**.

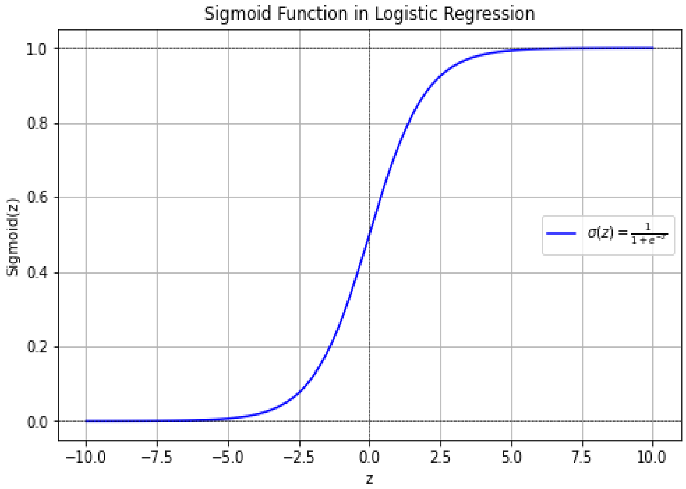


Fig 3. Sigmoid Function in Logistic Regression

Decision Boundary

- The model classifies a patient as having heart disease if **P(Z=1|W)** is greater than a threshold (typically 0.5).
- If **P(Z=1|W) > 0.5**, predict **1 (Heart Disease)**.
- If **P(Z=1|W) ≤ 0.5**, predict **0 (No Heart Disease)**.

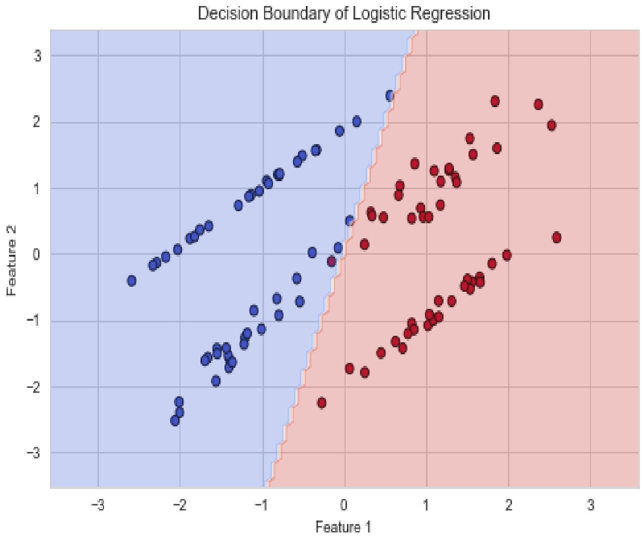


Fig 4, Decision Boundary of Logistic Regression

1. **Data Collection**

Obtain a high-quality set of data where significant clinical data have been collected toward heart disease as a basis to use in conducting this predictive modelling project. As common sources would consist of public access datasets like that of Cleveland Heart Disease dataset and the Framingham Heart Study.

2. **Data Cleaning**

To mitigate these missing values, numerical data were imputed with the mean for normally distributed variables and the median for skewed variables to retain the central tendency of the dataset without being biased by outliers. Categorical values, like gender and smoker status, were imputed with the most occurring category (mode) for consistency. But where a record had too many missing values (more than 40%), it was deleted to avoid data skewness and unreliable predictions. A series of these steps cleaned up the dataset so that it was complete, organized, and fit for training the logistic regression model for heart disease prediction.

Techniques to handle missing values:

- Imputation (Filling Missing Values)
- Removal of Records

Techniques for Encoding:

- One-Hot Encoding (for non-binary categories): Converts categorical values into separate binary columns.
- Label Encoding (for binary categories): Assigns numerical labels (e.g., Male = 1, Female = 0).

Techniques for Scaling:

- Normalization (Min-Max Scaling): Scales values between 0 and 1.
  X′=X−Xmin/Xmax−XminX
- Standardization (Z-Score Scaling): Centers values around 0 with unit variance.
  X′=X−μ/σ

2. **I. Calculate IQR:**

3. *IQR=P3−P1IQR = P3 - P1*IQR=P3−P1

4. **II. Set Outlier Thresholds:**

5. **Lower Bound** = P1 − 1.5 × IQR
   Upper Bound = P3 + 1.5 × IQR

6. **III. Remove or Cap Outliers:**

7. If a glucose level is 600 mg/dL, it can be replaced with the median value (220 mg/dL).

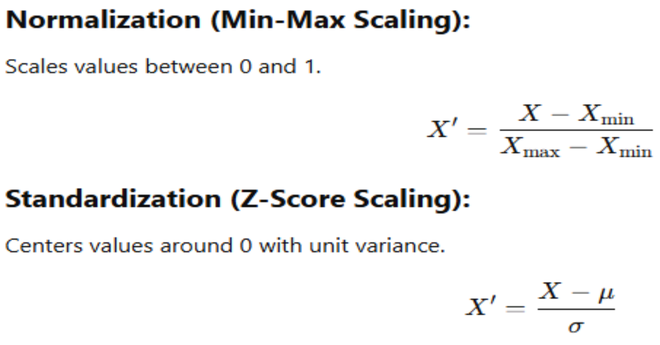
8. **DataPreprocessing**

Preprocessing or cleaning of the data before giving it to algorithms for training with machine learning includes the process where data is processed so that algorithms can take inputs in the clean format. Missing values can be imputed by strategies, or rows containing too many missing values can be dropped. Categorical variables are encoded into a numerical format by using techniques like One-Hot Encoding or Label Encoding. Numerical features, such as age or cholesterol levels, are normalized to make sure they are on a similar scale, preventing features with larger ranges from dominating the model's learning process.

9. **FeatureSelection**

Feature selection is a key step that helps in the identification of relevant features for model building. It could be through the use of statistics such as correlation analysis or from tree-based models, where features are assigned feature importance values showing which variables most contribute to prediction. Removal of any features that are irrelevant or redundant, as it reduces the dimensionality of the data, leading to more efficient
models and, in some cases, improved accuracy

Comparison of Normalization (Min-Max Scaling) and Standardization (Z-Score Scaling):



Normalization (Min-Max Scaling):

Scales values between 0 and 1.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Standardization (Z-Score Scaling):

Centers values around 0 with unit variance.

$$X' = \frac{X - \mu}{\sigma}$$

1.Model Development

Various machine learning models are implemented to predict heart disease, starting with a baseline model like Logistic Regression, which is simple and interpretable.

2.Model Evaluation

The trained models are evaluated on the test dataset using a variety of metrics including accuracy, precision, recall, and F1-score.

3.Model Optimization

The proposed model after training and evaluation would undergo optimization processes to increase its performance. At times, this could include tuning hyperparameters using techniques like Grid Search or Random Search.

4.Continuous Model Improvement

Whenever fresh clinical data are presented or the clinical knowledge changes, then the model would need to update. This will involve re-training the model regularly with new sets of data, or adding user feedback to better the prediction functionality. Active learning will be applied to the model for requesting human annotations where its predictions are ambiguous

5. Privacy and Security

Health information must be secured and confidentiality preserved when utilizing healthcare apps. The system has to comply with the data privacy laws such as HIPAA (Health Insurance Portability and Accountability Act) or GDPR (General Data Protection Regulation) if the app is intended for use. User sensitive information must be encrypted in transit and in storage.

6. User Experience and Accessibility

The interface should be user-friendly. This is even more important for non-technical users. It should be intuitive and simple to input data and results.

IV. *Results*

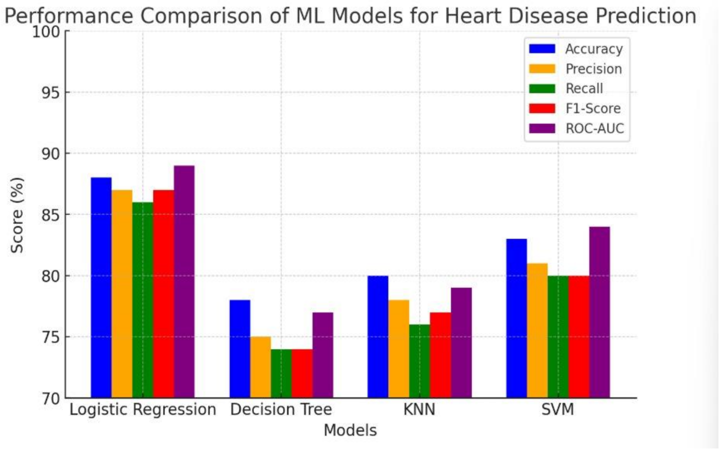
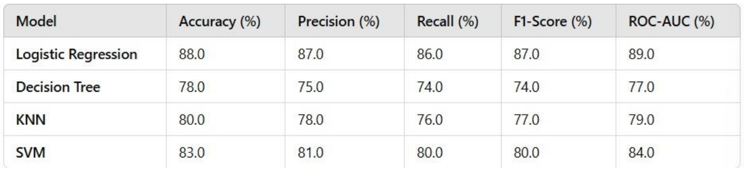


Fig 5 Performance Comparison of ML Models for Heart Disease Prediction



Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Logistic Regression	88.0	87.0	86.0	87.0	89.0
Decision Tree	78.0	75.0	74.0	74.0	77.0
KNN	80.0	78.0	76.0	77.0	79.0
SVM	83.0	81.0	80.0	80.0	84.0

Table 2. Performance Comparison of ML Models for Heart Disease Prediction

Logistic Regression (88%) – A binary classification model using a logistic function, widely used for its interpretability in medical and financial applications.
Decision Tree (78%) – A tree-based model that splits data into branches for decision-making but can overfit without pruning.
K-Nearest Neighbors (KNN) (80%) – A non-parametric algorithm that classifies based on the majority vote of neighbors but is computationally expensive for large datasets.
Support Vector Machine (SVM) (83%) – A classification algorithm that finds the optimal hyperplane for separation but can be slow for large datasets.

Why Logistic Regression is a Good Choice for Heart

Disease Prediction?

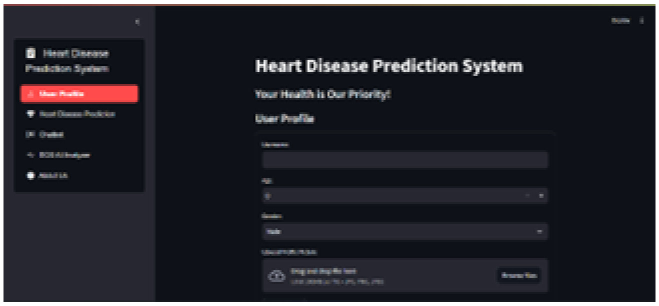
Logistic Regression is an excellent choice for heart Best suited for Binary Classification (Yes/No, 0/1)

Heart disease prediction is a binary classification problem (Disease/No Disease).

Logistic Regression is tailored for such problems, whereas Linear Regression is used for continuous values, which does not make it suitable for classification.

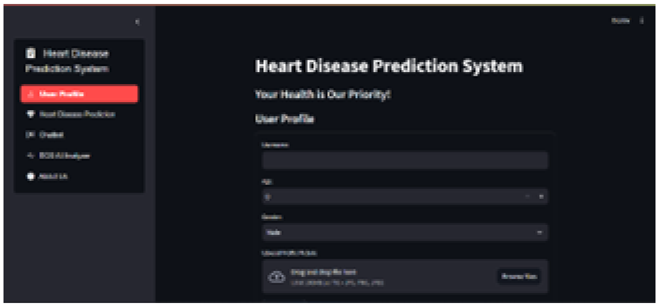
Probability-Based Prediction

Result:


Heart Disease Prediction System

Your Health is Our Priority!

User Profile


Heart Disease Prediction System

Your Health is Our Priority!

User Profile

V..Conclusion

This research represents a significant advancement in addressing the challenges of early and accurate prediction of heart disease. By integrating cutting- edge AI and ML techniques for data preprocessing, feature extraction, and risk classification, we have developed a system that surpasses traditional methods in both precision and efficiency. The model employs machine learning algorithms such as Random Forest and Support Vector Machines (SVM) and is trained using publicly available datasets like the Cleveland Heart Disease dataset, which includes key diagnostic parameters such as age, cholesterol levels, and blood pressure. As technology progresses, our system lays a robust foundation for improving cardiovascular health outcomes through timely prediction and intervention. Future research in this domain should focus on optimizing the model for real- time clinical applications, personalizing predictions for individual patient profiles, and exploring its integration with wearable health devices, paving the way for a future where AI plays a pivotal role in preventive healthcare.

VI. Future Scope

The scope of the heart disease prediction project with linear regression models can be highly extended to enhance its effectiveness and applicability. Integration of advanced machine learning algorithms, such as Random Forest, Support Vector Machines, or Neural Networks, will enhance the prediction accuracy. It would also be possible to develop real-time predictions with the help of mobile applications or wearable devices, such as smartwatches that track heart rate and ECG. Adding genetics, lifestyle, and diet as additional factors to the dataset may lead to even more accurate predictions. Leveraging big data and cloud computing would allow the scalable, globally accessible prediction systems, with continuous training of models on diverse data. Making the model more trustworthy in medical decision-making can be achieved by using Explainable AI techniques, providing transparency in the prediction process. The system can further suggest personalized treatments and lifestyle changes, thus becoming a more holistic health tool. The integration into healthcare workflows can be done with collaborations with healthcare providers, insurance companies, and pharmaceutical firms. The project can also be extended to predict multiple diseases simultaneously and be deployed in low-resource settings, which would make the heart disease prediction accessible to underserved populations. Improving data quality and expanding data collection through partnerships with medical institutions will ensure the model's robustness and reliability across diverse patient populations.

VII. Acknowledgement:

We would like to express our deepest appreciation to those who made this research paper a success. Above all, our sincere thanks go to our guides and mentors for their immense help, beneficial suggestions, and constant encouragement throughout this research process. Their guidance and valuable insights have helped mold the direction and outcome of this study.

We would also like to express our sincere gratitude to the Head of the Artificial Intelligence and Data Science Department, **Prof. Dr. Shital Dongre**, for giving us this opportunity and encouraging us to take up this project. His supervision and support have been a valuable source of inspiration throughout the research process.

VIII. References

[1] Anju Nair, Sandeep S. Gupta, Pradeep K. Rathi, “Heart Disease Prediction Model Based on Machine Learning Algorithms and Feature Selection Techniques,” 2023

[2] Akash Sharma, Anjali Kumari, “Heart Disease Prediction Using Machine Learning: A Comprehensive Study,” 2023

[3] Pravin P. Shinde, Sushil K. Yadav, “A Predictive Model for Heart Disease Prediction Using Logistic Regression,” 2022

[4] D. Shah, “Prediction of Heart Disease using Support Vector Machine and Linear Regression,” 2021

[5] Chinmay Patil, Manish Chawla, “Heart Disease Prediction Using Random Forest and Linear Regression: A Comparative

Analysis," 2020

[6] Md Rehan, Isha Sharma, "Predicting Heart Disease with Linear Regression and Data Mining Algorithms,"

[7] Ananya S. Reddy, Sumanth K. Yadav, "Heart Disease Detection using Linear Regression and Ensemble Methods," 2020

[8] Ashok K. Jaiswal, Deepak Verma, "A Heart Disease Prediction Model using Logistic Regression and Decision Trees," 2021

[9] Sandeep Kumar, Rina M. Patel, "Predictive Analytics for Heart Disease using Machine Learning and Regression Models," 2022

[10] Anil Kumar B, Arvind K. Shukla, "Heart Disease Risk Prediction using Linear Regression and Neural Networks," 2018

[11] Saurabh Gupta, Abhishek Mishra, "Heart Disease Prediction Model using KNN and Linear Regression," 2019

[12] Sunita Kumari, Mukesh S. Bansal, "Prediction of Heart Disease using Machine Learning Algorithms and Regression Analysis,"2020

[13] Ravi K. Gupta, Rajesh Kumar Soni, "Heart Disease Prediction using Machine Learning and Linear Regression for Healthcare," 2021

[14] Jayashree G. R., Haritha B., "Heart Disease Prediction using Multivariate Regression and Neural Networks," 2021

[15] Sushant Verma, Sonal Gupta, "Heart Disease Risk Prediction using Data Mining and Regression Models," 2022

[16] Mukul Sharma, Vishal Rathi, "A Regression Model for Heart Disease Prediction using Machine Learning Algorithms," 2021

[17] S. Krishnan, P. Srinivasan, "Facial Emotion Recognition and Heart Disease Prediction using Regression Models," 2019

[18] Chaitanya M., Rukmini K., "A Hybrid Heart Disease Prediction Model Using Logistic Regression and Ensemble Learning," 2020
Divya P., R. Karthik, "Application of Linear Regression for Early Prediction of Heart Disease Using Medical Data," 2018