# Heart Disease Prediction System

Pratik Dave, Saloni Choudhary, Aryaman Jha, Jenil Kumbhani, Govind Wakure
*Dept. of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai, India*

**Abstract: A prevalent cardiovascular condition that can alter an individual's life, heart disease is very expensive for both individuals and healthcare systems. Effective prevention strategies and early diagnosis are necessary as it progresses to be a major global societal well-being concern. Improving patient outcomes and cutting costs associated with healthcare requires early detection and proactive management. Based on a variety of patient characteristics and medical data, machine-learning characteristics, classifiers, and algorithms have demonstrated promise in the past several years in the prediction of cardiac disease risk. This abstract describes the application of machine learning methods to the construction of a prediction system for cardiac disease. This abstract analyzes patient data, such as electronic health records (EHR), medical histories,vital signs, and diagnostic test results, using sophisticated algorithms. The principal aim of the system is to precisely detect individuals who are susceptible to heart failure through the application of particle swarm optimization, forward selection, and backward elimination techniques. This will facilitate early intervention and tailored care, with 0 signifying the non-existence of heart disease and 1 signifying its presence.**

**Keywords: Machine Learning Algorithms, EHR, Particle Swarm Optimization, Forward Selection, Backward Elimination.**

## I. INTRODUCTION

Heart disease is a major and common medical illness that people around the world in millionssuffer from. It happens when the pumping of the blood to the heart is majorly impacted, whichreduces the amount of oxygen and nutrients that reach the body's tissues. Heart failure can haveserious side effects, such as a lower quality of life, more hospital admissions, and even death. The creation of precise and timely cardiac disease prediction systems is crucial. The project's objective is to present a Heart Disease Prediction System that uses machine learning classifierslike Logistic Regression, Gradient Boost, Random Forest, Decision Trees, Naïve Bayes, and Support Vector Machines and data analytics to help with early detection and risk assessment of heart disease in people. The dataset used over here is the Cleveland Heart Disease dataset taken from the UCI Machine Learning repository. The dataset consists of 303 individuals' datathat has 14 columns in it comprising of no missing values and its significance lies in performingthe classification task where 0 denotes the absence whereas 1 denotes the presence of the heartdisease.

## II. LITERATURE SURVEY

Several studies in a variety of ways have been done on the Heart Disease Prediction System. In [1] authors present an effective heart disease prediction using machine learning techniques. Authors utilized heart disease and machine learning classification techniques like logistic regression and others for classifying heart-related disease using different models and a realistic data. In [2], the focus is on using different performance metrics like accuracy, specificity, sensitivity, etc on these classifiers used in the developed model helps to generate promising results.In [3] authors present a novel approach for heart disease prediction using strength scores with significant predictors. Authors Armin Yazdani, Kasturi Dewi Varathan, Yin Kia Chiam, Asad Waqar Malik, and Wan Azman Wan Ahmad emphasized on the features that appeared a greaternumber of times are highlighted as the most impactful features of the heart disease prediction system. In [4] authors present an effective heart disease prediction using hybrid machine learning techniques for classification of the heart disease based on the random forest algorithm used inmodel. In [5] authors created a model for the prediction of heart disease using machine learning algorithms where prediction algorithms were shortlisted on the basis of the evaluation that makes heart disease prediction quick and accurate. The keywords/parameters considered for the system are the prediction algorithms and the cardiovascular disease. In [6] the literature work machine learning-based heart disease prediction: a study for home personalized care focuses on the usage of the random forest classifier on the c level and heart disease dataset that returns a confusion matrix representing the presence of the disease in the patient.

## III. METHODOLOGY

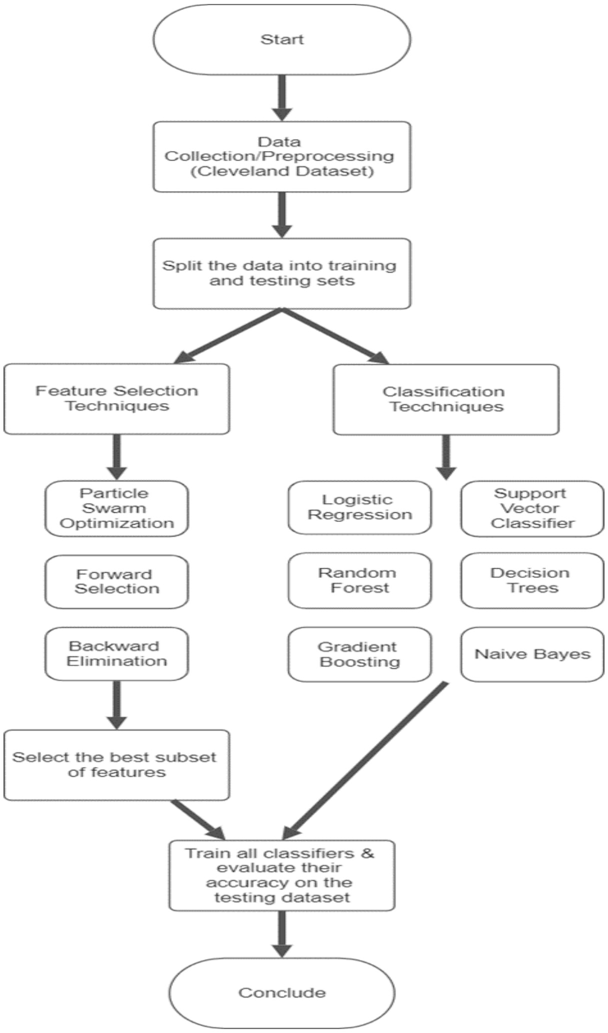

Start
Data Collection/Preprocessing (Cleveland Dataset)
Split the data into training and testing sets
Feature Selection Techniques
Classification Tecchniques
Particle Swarm Optimization
Forward Selection
Backward Elimination
Logistic Regression
Support Vector Classifier
Random Forest
Decision Trees
Gradient Boosting
Naive Bayes
Select the best subset of features
Train all classifiers & evaluate their accuracy on the testing dataset
Conclude

Fig 1. Process Flowchart

When discussing the process flow of a heart disease prediction system, the first step is to choosethe most comprehensive and dependable dataset. Our examination of the Z-Alizadeh Sani, Statlog, and Cleveland Heart Disease datasets gave us a clearer picture of how the Cleveland Heart Disease dataset might affect the suggested approach. After the Cleveland dataset has been chosen, it is divided into training and testing sets using a percentage ratio of either 80:20or 70:30, with the former portion serving as the training set and the latter as the testing set. Three feature selection methods—Particle Swarm Optimization, Forward Selection, and Backward Elimination—are used to the training dataset. Here, six distinct classification methods are employed: decision trees, logistic regression, random forests, support vector machines, gradient boosting, and naive bayes algorithms. This system was developed using a methodology that included training a model on a specific classifier using a given feature selection technique. To train the model, the logistic regression classifier is chosen. A feature selection technique is now employed to choose the top features from the 13 features included in the Cleveland Dataset for the Heart Disease.

## IV. SYSTEM ARCHITECTURE

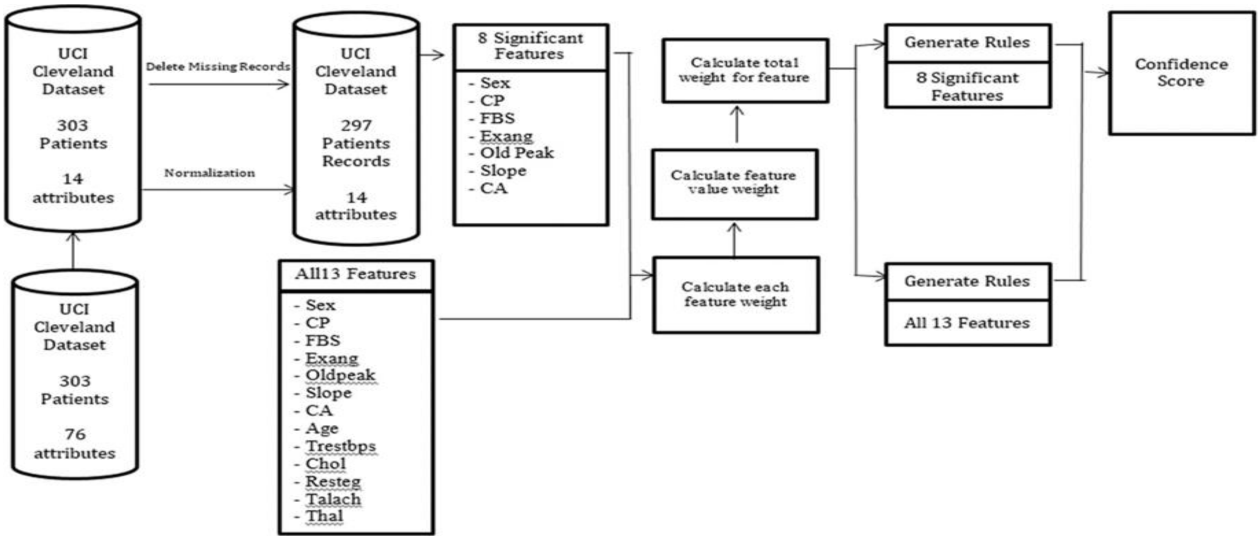
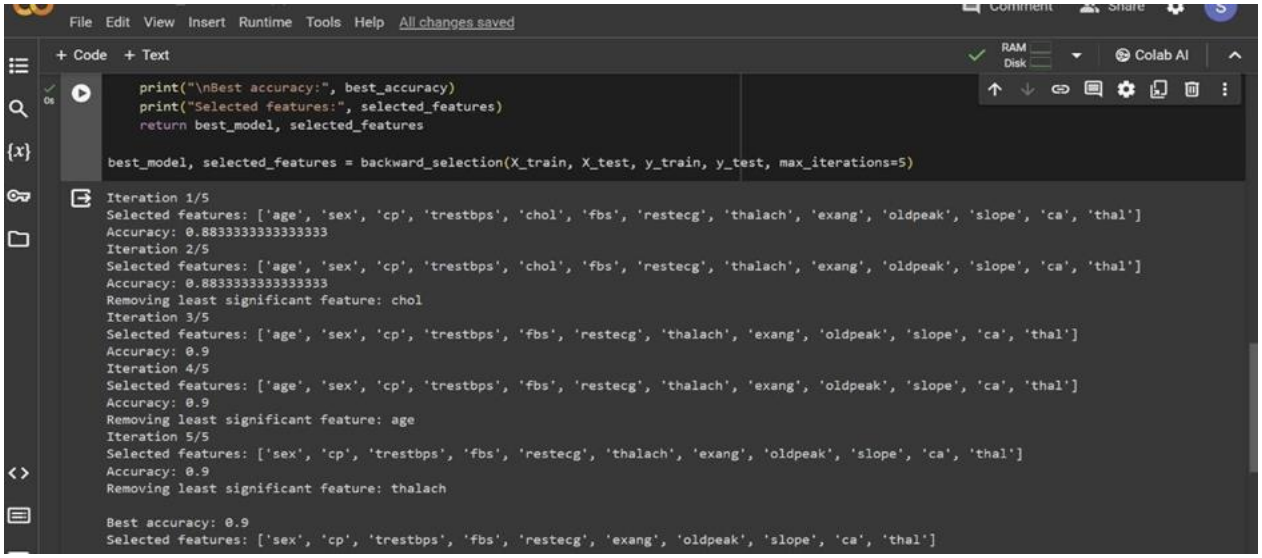


Fig 2. Data Pre-processing Methodology of the Cleveland Dataset

Of all the datasets that are accessible, the Cleveland dataset is the one with highest precision. This dataset, which was utilised from the UCI Machine Learning Repository, includes 303 patients' health-related information and displays 13 significant characteristics about them. Any model that deals with an individual's heart health is developed using these elements as its foundation. When it comes to data pre-processing, the first steps are to normalize the dataset and remove any missing records. After 6 items are removed during normalization, we are left with 297 patient records and 14 attributes (13 essential features and 1 target variable).

## V. RESULTS



```
    print("\nBest accuracy:", best_accuracy)
    print("Selected features:", selected_features)
    return best_model, selected_features

best_model, selected_features = backward_selection(X_train, X_test, y_train, y_test, max_iterations=5)
```

```
Iteration 1/5
Selected features: ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
Accuracy: 0.8833333333333333
Iteration 2/5
Selected features: ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
Accuracy: 0.8833333333333333
Removing least significant feature: chol
Iteration 3/5
Selected features: ['age', 'sex', 'cp', 'trestbps', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
Accuracy: 0.9
Iteration 4/5
Selected features: ['age', 'sex', 'cp', 'trestbps', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
Accuracy: 0.9
Removing least significant feature: age
Iteration 5/5
Selected features: ['sex', 'cp', 'trestbps', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
Accuracy: 0.9
Removing least significant feature: thalach

Best accuracy: 0.9
Selected features: ['sex', 'cp', 'trestbps', 'fbs', 'restecg', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
```

Fig 3. Accuracy score of all the classifiers with respect to backward selection technique

Here, the best results are generated from the backward elimination technique when performedon a logistic regression classifier model. Gradient Boost classifier gives same results when performed on the Cleveland Dataset as well as even after performing feature selection technique using backward elimination. A drop in the accuracy of the system has been observed when the classifiers like Support Vector Machines and Naïve Bayes are used for the Heart Disease Prediction System. In the realm of the disease prediction system for heart health, achieving high accuracy rates is paramount for reliable diagnoses and effective patient care. Notably, utilizing forward selection with logistic regression yielded an impressive accuracy peak of 83.9%, showcasing its efficacy in feature selection for predictive modelling. On the other hand, the utilization of Particle Swarm Optimization (PSO) coupled with Random Forestas an objective function led to the identification of the Support Vector Classifier as the top performer, achieving a commendable accuracy of 81.97%. These findings underscore the significance of employing diverse methodologies, from traditional regression techniques to advanced optimization algorithms, in optimizing feature selection and classifier performance.

## VI. CONCLUSION

A heart disease prediction system's conclusion would normally include an overview of the mainconclusions, implications of the system's performance, and an analysis of its possible effects. The application of this heart disease prediction system may benefit the healthcare industry by lessening the number of hospitalizations caused by heart failure, enhancing patient satisfaction, and maybe lowering medical expenses. The comparison of various classifiers and feature selection techniques on the Cleveland dataset for heart disease prediction yields insightful conclusions. Firstly, the significant accuracy improvement from 78% to 90% with backward elimination in logistic regression underscores its effectiveness in refining feature subsets for optimal predictive performance.

## VII. FUTURE SCOPE

In the realm of developing a system for heart disease prediction using machine learning, future advancements hold promising prospects. By harnessing personalized medicine approaches, algorithms can tailor risk evaluations and treatment recommendations based on individual genetic, clinical, and lifestyle factors. Integration with Electronic Health Records (EHRs) facilitates seamless identification of at-risk patients, enabling healthcare professionals to intervene promptly through automated alerts and real-time warnings. Furthermore, telehealth initiatives stand to benefit greatly from predictive systems, offering remote monitoring solutions that not only reduce readmissions at the healthcare institutions but also empower thelevel of lifestyle system for individuals with early-stage cardiac disease. Together, these developments pave the way for more precise, proactive, and patient-centric approaches to cardiovascular healthcare.

## REFERENCE

[1] Chicco, Davide, and Giuseppe Jurman. *"Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone."* BMC medical informatics and decision making 20.1 (2020): 16.

[2] Chaudhary, K., Ji, L., & Bhagwan, R. (2020). *“Deep learning predictive analytics for heartfailure diagnosis.”* In 2020 IEEE International Conference on Systems, Man, and Cybernetics(SMC) (pp. 4705-4711).

[3] Ahmed, M., & Mahmood, A. N. (2017). *“Heart disease prediction system using data mining techniques.”* Journal of King Saud University-Computer and Information Sciences.

[4] Sudharshan, S., Neelanarayanan, P., Ramachandran, P., Vaseeharan, B., & Rao, P. S. (2021). *“Early Prediction of Heart Failure Using Machine Learning Algorithms.”* In Advances in Machine Learning and Data Science (pp. 67-77). Springer.

[5] Ponikowski, P., Voors, A. A., Anker, S. D., Bueno, H., Cleland, J. G., Coats, A. J., ... & Ruschitzka, F. (2016). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). European journal of heart failure, 18(8),891-975.