

Research Article

Volume 19 Issue 1 - December 2023
DOI: 10.19080/JOCCT.2023.19.556004

J Cardiol & Cardiovasc Ther

Copyright © All rights are reserved by Mohan Raja Pulicharla

A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare

Sreenivasulu Ramisetty^{1*}, Thirupurasundari Chandrasekaran², Vamsi Krishna Eruvaram³ and Dr. Mohan Raja Pulicharla⁴

¹Data Architect, USA

²Sr Project Manager

³Data Engineer, USA

⁴Data Engineer Staff, USA

Submission: November 01, 2023; **Published:** December 20, 2023

***Corresponding author:** Mohan Raja Pulicharla, MCA, 2004, Madras University, India

Abstract

The heart, as the second most vital organ after the brain, is integral to maintaining bodily equilibrium, and disruptions to its function have profound health consequences. Heart disease, a leading global cause of mortality, often arises from cumulative daily physiological changes, emphasizing the importance of timely illness prediction. In healthcare, the fusion of data mining and machine learning, explored in this study using Support Vector Machine, Decision Tree, and Random Forest algorithms, addresses the challenges of diagnosing prevalent conditions like heart disease, particularly crucial in the field of cardiology.

Our proposed machine learning-based approach for diagnosing cardiac disease employs a range of classification algorithms and advanced feature selection techniques, demonstrating superior accuracy in detecting heart diseases from extensive datasets of unprocessed medical images. This technological advancement holds the potential to significantly enhance patient care in various healthcare settings, showcasing the promising impact of artificial intelligence tools on improving the quality of life for billions worldwide.

Keywords: Machine learning; Heart disease; Algorithms; Cardiovascular disease; Regression analysis

Abbreviations: ML: Machine Learning; LCS: Learning Classifier Systems; ILP: Logic Programming; ANN: Artificial Neural; SVMs: Network Support Vector; Machines; CVDs: Cardiovascular Diseases; TP: True Positive; TN: True Negative; FN: False Negative; FP: False Positive; LDL: Low Density; Lipoprotein; ECG: Electrocardiogram; CXR: Chest X-Ray

Introduction

After the brain, the heart is regarded as the second-most significant organ. Every heart disruption causes the entire body to become upset. Heart disease is one of the top five killer diseases in the world. Disorders, including heart disease, are a result of the changes that occur to us daily. Consequently, it is crucial to predict a sickness at the appropriate time. Data mining is a fundamental and fundamental process for defining and discovering relevant data and uncovering hidden patterns in massive databases. By predicting and diagnosing various diseases, data mining, and machine learning techniques are used in the medical sciences to address genuine health-related challenges. This study compares the performance of three machine learning algorithms-support

vector machine, decision tree, and random forest-for the prediction of heart disease.

Machine Learning-Based Approach for Diagnosing Cardiac Disease

The study emphasizes the critical need for swift and accurate heart disease identification, proposing a machine learning approach with Support Vector Machines, Logistic Regression, Artificial Neural Networks, K-Nearest Neighbors, Naive Bayes, and Decision Trees for classification (Figure 1). Efficiency is enhanced through feature selection algorithms and a conditional mutual information method, ensuring commendable accuracy,

particularly with Support Vector Machines. This makes it a promising tool for rapid implementation in medical settings, crucial for early identification and interrupting cardiac disease progression. The analysis of diverse datasets identifies key features for heart disease prediction, utilizing seven machine

learning methods. A hybrid dataset is created and analyzed with Python's Scikit-learn module using a univariate feature selection technique, offering a comprehensive approach to discern crucial factors in predicting and preventing heart disease.



Figure 1: Various algorithms of artificial intelligence and Machine Learning applied in cardiovascular medicine.

Provides the Maximum Accuracy, Several Parameters Relating to Various Algorithms:

Datasets are split into training and testing using holdout and cross-validation techniques, adjusting algorithm parameters for maximum accuracy. Evaluation metrics, including a classification report and confusion matrix, gauge performance. Majority voting, combining logistic regression, SVM, and naive Bayes, achieves 88.89% accuracy on the first dataset, while the hybrid dataset lags individual ones. Project outcomes are compared with prior methodologies. Machine learning, an algorithmic system falling under AI, learns without explicit programming, relying on statistics and data for outcome prediction. Linked to data mining

and Bayesian modeling, it operates by taking data as input and generating answers through algorithms, seen in applications like personalized recommendations, fraud detection, and predictive maintenance [1].

Machine Learning vs Traditional Programming

Traditional Programming: This problem is meant to be solved via machine learning. The computer creates a rule after learning how the input and output data are related. Every time there is fresh data, the programmers do not need to design new rules. The algorithms change because of fresh information and experiences, increasing their efficacy over time (Figure 2).

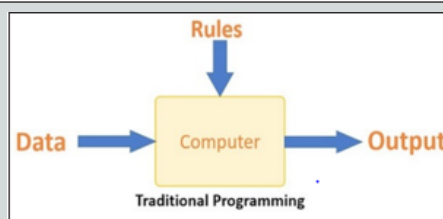


Figure 2: Traditional programming.

Machine Learning Approach

Machine learning mimics human learning through experience, succeeding in familiar scenarios with easier predictions. Like humans, machines train by observing examples for precise predictions but struggle with new instances. Central to machine

learning is learning and inference, primarily achieved by identifying patterns in data. A crucial skill for data scientists is selecting data to create a feature vector, simplifying reality with sophisticated algorithms. This feature vector transforms the learning step into a condensed model that describes the data (Figure 3-6).

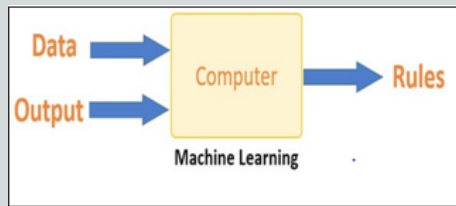


Figure 3: Machine Learning.

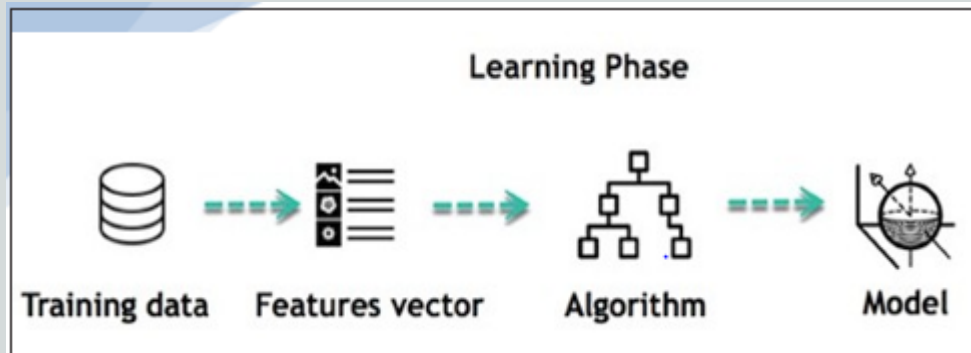


Figure 4: Learning phase.

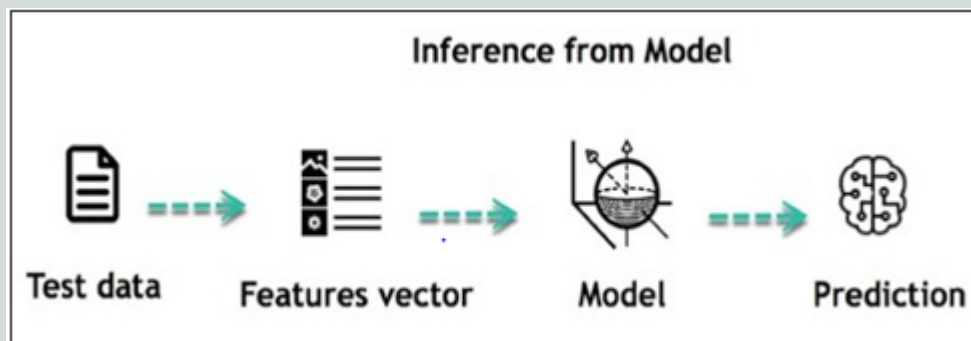


Figure 5: Inference model.

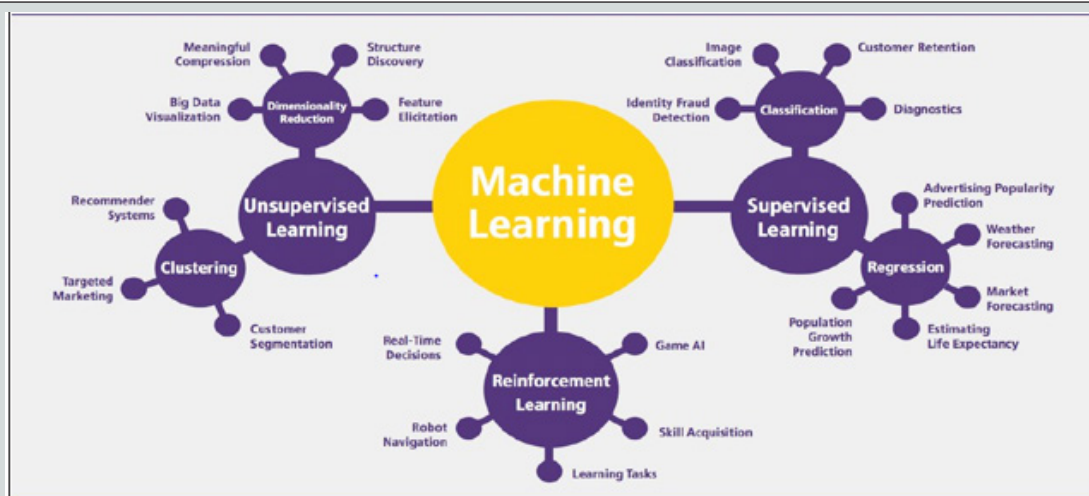


Figure 6: Machine Learning Algorithms

Key Points of Machine Learning Programs (Table 1)

Table 1:
The Life of Machine Learning Programs is Straightforward and Can Be Summarized in The Following Points:
1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

Machine learning Algorithms

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms.

Supervised Learning: An algorithm learns the link between given inputs and a particular output using training data and feedback from humans. For instance, a practitioner can forecast sales using input data such as marketing expenses and weather predictions.

There are two categories of supervised learning:

- Classification task

- Regression task

Classification: To determine a customer's gender for a commercial, information is extracted from the database, including height, weight, occupation, salary, and purchase history. The classifier's goal is to assign a probability label (male or female) based on these features. Once the model learns to distinguish between genders, it can be used for predictions with new data. For example, if the classifier predicts a 70% probability of being male and 30% female, the algorithm confidently assigns the customer as male. Classifiers can have multiple classes for predicting items, like glass, table, shoes, each representing a different class [2,3].

Types of Algorithms: (Table 2)

Algorithm Name	Description	Type
Linear regression	Finds a way to correlate each feature to the output to help predict future values.	Regression
Logistic regression	Extension of linear regression that's used for classification tasks. The output variable is binary (e.g., only black or white) rather than continuous (e.g., an infinite list of potential colors)	Classification
Decision tree	Highly interpretable classification or regression model that splits data feature values into branches at decision nodes until a final decision output is made	Regression Classification
Naive Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event	Regression Classification
Support vector machine	Support Vector Machine, or SVM, is typically used for the classification task. SVM algorithm finds a hyperplane that optimally divides the classes. It is best used with a non-linear solver.	Regression (non-very common Classification)
Random forest	The algorithm, based on a decision tree, significantly enhances accuracy by employing a random forest. This approach generates numerous simple decision trees and utilizes the 'majority vote' method to determine the final label for classification tasks. In regression tasks, the final prediction is the average of all the trees' predictions.	Regression Classification
AdaBoost	Classification or regression technique that uses models to come up with a decision but weighs them based on their accuracy in predicting the outcome	Regression Classification
Gradient-boosting trees	Gradient-boosting trees is a state-of-the-art classification/ regression technique. It is focusing on the error committed by the previous trees and tries to correct it.	Regression Classification

Unsupervised Learning: In unsupervised learning, an algorithm explores input data without being given an explicit output

variable (e.g., explores customer demographic data to identify patterns) (Table 3).

Algorithm	Description	Type
K-means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommender system	Help to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension

Machine Learning (ML) Algorithm and Practical Application

Numerous machine learning algorithms exist, chosen based on specific goals. In the following flower prediction example, ten algorithms predict flower types based on petal dimensions. The dataset is depicted in the top left image, divided into red, light

blue, and dark blue groups. Classifications include the upper left of the second image in the red group, the middle exhibiting ambiguity and light blue, and the bottom in the dark category. Subsequent images illustrate various algorithms attempting to categorize the data (Figure 7 & 8).

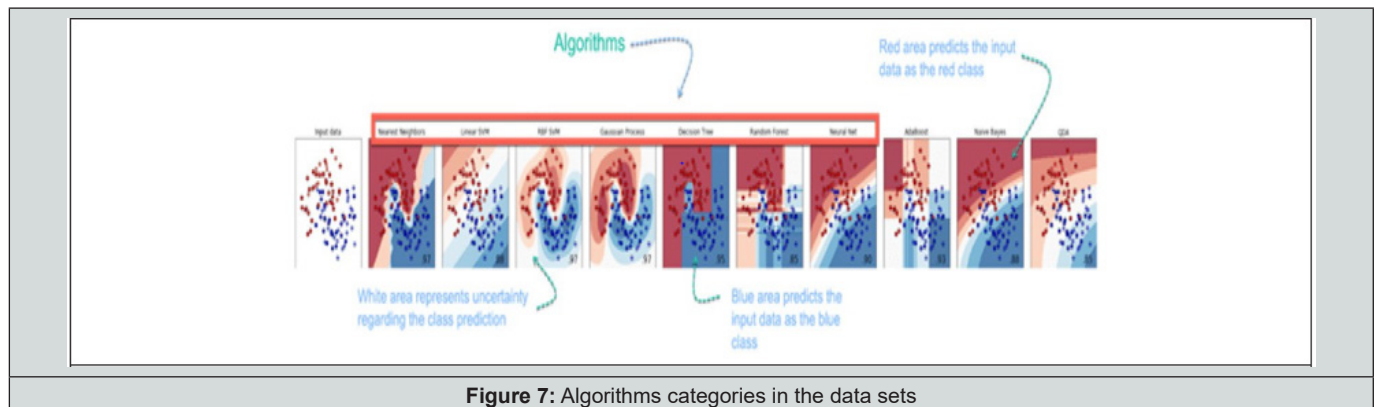


Figure 7: Algorithms categories in the data sets

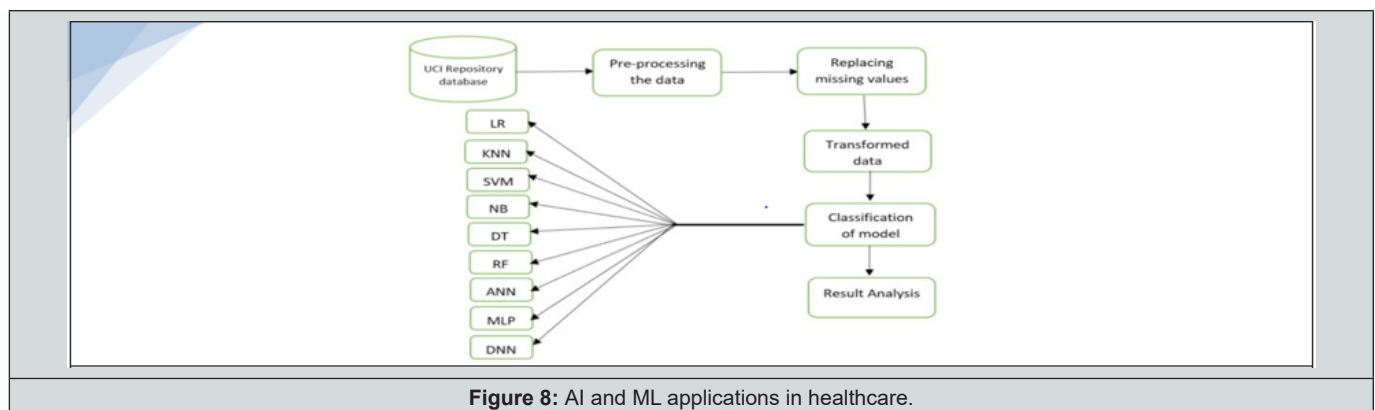


Figure 8: AI and ML applications in healthcare.

Objectives of the study

- To study the Machine Learning application in heart disease identification.
- To analyze the factors impacting on heart disease through
- using the Machine Learning application under Artificial Intelligence systems.
- To study the factors that affect heart disease risk the most across three separate datasets.

- To combine two separate datasets to create a hybrid dataset and to use three datasets.
- One hybrid dataset to apply several machine learning algorithms for the prediction of heart disease.
- To suggest efficient and effective measures to diagnose the heart disease identification through Machine Learning

Methodology

This project utilizes three datasets from Kaggle and the UCI machine learning repository, employing Python for data analysis through the Anaconda distribution. Various supervised machine learning algorithms, such as support vector machines, decision trees, k-nearest neighbors, naive Bayes, random forest, and logistic regression, are employed, including an ensemble technique, the majority voting classifier. The study explores different parameters and methods, such as adjusting C and gamma values for support vector machines and tuning k values for k-nearest neighbors. Feature selection is conducted using the univariate method, and a hybrid dataset is created from two distinct datasets, standardized, and evaluated. The research compares outcomes with literature,

focusing on achieving the best accuracy in diagnosing heart disease [4,5].

Anaconda Distribution Package Configuration

Anaconda, a free and open-source distribution for R and Python, simplifies large-scale data processing and package management. This study leverages Anaconda's graphical user interface, Anaconda Navigator, for program launch and package, environment, and channel management. The navigator integrates various tools, such as RStudio, Spyder, Orange, Jupiter, and Jupiter Notebook. Specifically, Jupiter Notebook is used for running essential data analysis codes in an interactive, web-based computational environment.

Algorithm for Disease Prediction

Algorithm: Detection of heart disease using classifiers

Input: Heart disease dataset with several attributes

Output: Accuracy score/Confusion matrix/Classification report of predicted values (Figure 9).

Process:

```

Step 1: Import libraries of sklearn, pandas, numpy

Step 2: Import the classifier functions

Step 3: Import train_test_split function

        Or Import cross_val_score function

Step 4: i) Import accuracy_score function

        ii) Import confusion_matrix function

        iii) Import classification_report function

Step 5: Load the CSV file containing data using read_csv() function

Step 6: Separate the input and target attributes

Step 7: i) For holdout method, separate the train and test data using train_test_split() function

        ii) Model the classifier using model.fit() function

        iii) Predict the test data using model.predict() function

        Or apply cross validation using cross_val_score() function

Step 8: i) Find accuracy using accuracy_score() function

        ii) Find confusion matrix using confusion_matrix() function

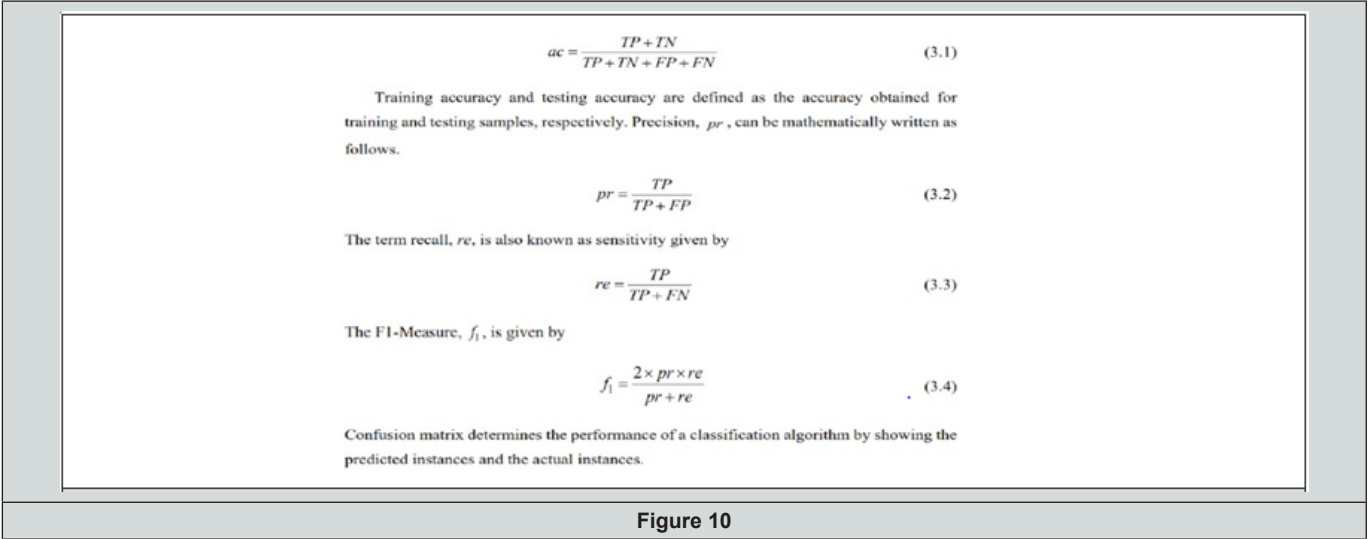
        iii) Find classification report using classification_report() function
    
```

Figure 9

Performance Metrics

This study evaluates heart disease identification using various metrics, such as training accuracy, testing accuracy, precision, recall, and F1-measure, employing terms like True Positive (TP), True Negative (TN), False Negative (FN), and False Positive

(FP). TP indicates correctly identified heart disease cases, TN represents correctly identified healthy cases, FN refers to missed heart disease cases, and FP signifies incorrectly identified cases. Accuracy, denoted by *ac*, measures the percentage of correctly identified vectors among all normal and abnormal samples (Figure 10).



Overview of Conceptual and Theoretical Aspects of Machine Learning Application in Heath Industry

Introduction

Machine learning in healthcare can enhance diagnostic tools for medical image analysis. ML algorithms applied to medical imaging, such as X-rays or MRI scans, utilize pattern recognition to identify specific conditions. These applications analyze vast datasets, aiding healthcare professionals in making more

informed judgments [6,7].

Applications for machine learning (ML)

Machine Learning (ML) applications are pervasive and play a vital role in diverse practical domains, notably healthcare and patient data security. This study explores the utilization of ML to analyze medical records and predict diseases, addressing gaps in effective ML methods and applications within the healthcare industry.

Highlights of Machine Learning: (Table 4)

Table 4:
Python is an interpreted language, allowing runtime processing without the need for pre-compilation, like PHP and PERL.
Python is Interactive - To write programs, sit at a Python prompt and communicate with the interpreter directly.
Python's support for object -oriented programming, which encapsulates code within objects, makes it an object-oriented language.
Python is a fantastic language for beginning programmers and facilitates the creation of a wide range of programs, including simple text processing, web browsers, and games.

Machine Learning

Machine learning, a subset of artificial intelligence, creates predictive systems by learning from experiences and building models on datasets to uncover hidden patterns. In healthcare, machine learning applications optimize trial samples, increase

data points, and play a pivotal role in early epidemic detection. The study emphasizes the transformative impact of machine learning on healthcare operations, allowing professionals to focus on patient care and addressing global healthcare challenges (Figure 11).

Machine Learning Algorithms:

<p>Supervised Learning Algorithms:</p> <ul style="list-style-type: none"> • Linear Regression • Logistic Regression • Decision Trees • Random Forest • Support Vector Machines (SVM) • K-Nearest Neighbors (KNN) • Naive Bayes <p>Unsupervised Learning Algorithms:</p> <ul style="list-style-type: none"> • K-Means • Hierarchical Clustering • Principal Component Analysis (PCA) • Independent Component Analysis (ICA) <p>Semi-Supervised Learning Algorithms:</p> <ul style="list-style-type: none"> • Self-Training • Multi-View Learning <p>Reinforcement Learning Algorithms:</p> <ul style="list-style-type: none"> • Q-Learning • Deep Q Network (DQN) <p>Neural Network Architectures:</p> <ul style="list-style-type: none"> • Feedforward Neural Networks: • Convolutional Neural Networks (CNN) • Recurrent Neural Networks (RNN) • Long Short-Term Memory (LSTM) • Generative Adversarial Networks (GAN) <p>Ensemble Learning Algorithms:</p> <ul style="list-style-type: none"> • AdaBoost • Gradient Boosting Machines (GBM) • XGBoost <p>Clustering Algorithms:</p> <ul style="list-style-type: none"> • DBSCAN (Density-Based Spatial Clustering of Applications with Noise) • Mean Shift • Affinity Propagation 	<p>Dimensionality Reduction Techniques:</p> <ul style="list-style-type: none"> • t-Distributed Stochastic Neighbor Embedding (t-SNE) • Autoencoders <p>Time Series Forecasting:</p> <ul style="list-style-type: none"> • ARIMA (AutoRegressive Integrated Moving Average) • Prophet <p>Anomaly Detection:</p> <ul style="list-style-type: none"> • Isolation Forest • One-Class SVM <p>Natural Language Processing (NLP) Algorithms:</p> <ul style="list-style-type: none"> • Word2Vec • BERT (Bidirectional Encoder Representations from Transformers) • Named Entity Recognition (NER) <p>Recommender Systems:</p> <ul style="list-style-type: none"> • Collaborative Filtering • Content-Based Filtering • Matrix Factorization <p>Transfer Learning:</p> <ul style="list-style-type: none"> • Fine-tuning • Domain Adaptation <p>Hyperparameter Tuning Algorithms:</p> <ul style="list-style-type: none"> • Grid Search • Random Search <p>Evolutionary Algorithms:</p> <ul style="list-style-type: none"> • Genetic Algorithms • Particle Swarm Optimization (PSO) <p>Ensemble Learning Techniques:</p> <ul style="list-style-type: none"> • Stacking • Bagging (Bootstrap Aggregating)
--	---

Results and Discussion

This section presents the results of data analysis for predicting heart diseases, considering variables such as age, chest pain type, blood pressure, blood glucose level, ECG, heart rate, exercise angina, and four types of chest pain. The heart disease dataset undergoes effective preprocessing, eliminating unrelated records and handling missing values. The K-means algorithm is then applied to compose the preprocessed dataset, discussing four types of heart diseases: asymptomatic pain, atypical angina pain, non-anginal pain, and non-anginal pain. Histogram analysis

shows a higher risk of heart disease in the age range of 50 to 55, where the development of coronary fatty streaks begins (Figure 12) [8-11].

Figure shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease (Figure 13).

Figure shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease (Figure 14).



Figure 11: Smart features of machine learning for healthcare domain.



Figure 12: Histogram of variation of age for each target class.

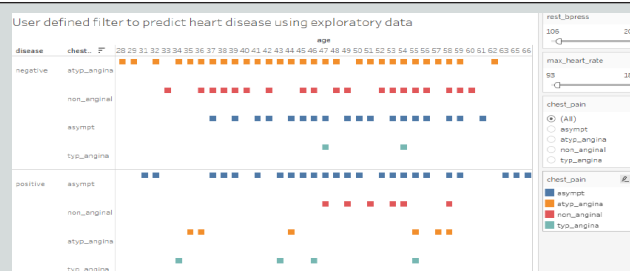
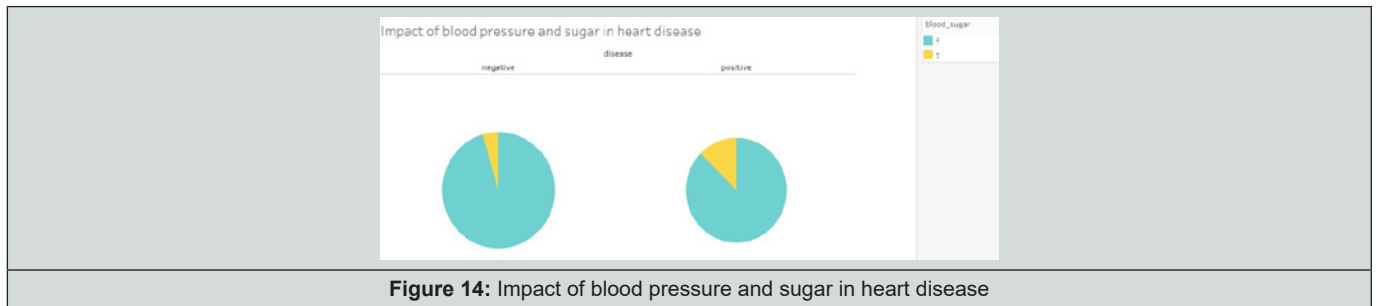


Figure 13: shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease.



K-means Clustering: The K-means clustering algorithm is chosen for its efficiency, simplicity, capacity to generate even-sized clusters, and scalability in handling the dataset, ensuring

accurate outputs with a minimum sum of squares. The dataset comprises 209 observations with 7 variables (Figure 15).

$$\sum_{k=1}^K \sum_{i \in S_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$$

where S_k is the set of observations in the k th cluster and \bar{x}_{kj} is the j th variable of the cluster center for the k th cluster.

Figure 15

Chest Pain Type: Asymptomatic

The plot of Age vs. Max Heart Rate broken down by Disease.

Colour shows details about disease. The screen shot of the clustering are described below (Table 5 & 6), (Figure 16 & 17) & (Tables 7-12).

Table 5: The plot of Age vs. Max Heart Rate broken down by Disease factors.	
Summary of Diagnostics	
No. of Clusters	2
No. of Points	102
Between-group Sum of Squares	20.285
Within-group Sum of Squares	9.5649
Total Sum of Squares	29.85

Table 6: Chest Pain Type: Asymptomatic.				
No. of Clusters	Items	Ages (in Sum)	Sum Of Maximum Heart Rate	Disease
Cluster1	75	49.853	124.03	Positive
Cluster2	27	48.556	136.59	Negative

Table 7: Chest Pain Type: Atypical Angina: Factors.	
Summary of Diagnostics	
No. of Clusters	2
No. of Points	65
Between-group Sum of Squares	5.5109
Within-group Sum of Squares	8.3246

Table 8: Chest Pain Type: Atypical Angina.

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster1	59	45.492	147.47	Positive
Cluster2	6	47.5	139.5	Negative

Table 9: Chest Pain Type: Non-Angina: Factors.

Chest Pain Type: Non-Angina				
Summary of Diagnostics				
No. of Clusters				3
No. of Points				36
Between-group Sum of Squares				8.89
Within-group Sum of Squares				2.251
Total Sum of Squares				11.141

Table 10: Chest Pain Type: Non-Angina.

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster 1	15	39.533	162.8	Negative
Cluster 2	14	54.571	133.43	Negative
Cluster 3	7	52.857	140.29	Positive

Table 11: Chest Pain Type: Typical Anginal Pain: Factors.

No. of Clusters				3
No. of Points				6
Between-group Sum of Squares				2.3779
Within-group Sum of Squares				0.52542
Total Sum of Squares				2.9033

Table 12: Chest Pain Type: Typical Anginal Pain.

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster 1	2	40	177.5	Positive
Cluster 2	2	49	145.5	Positive
Cluster 3	2	50.5	143.5	Negative

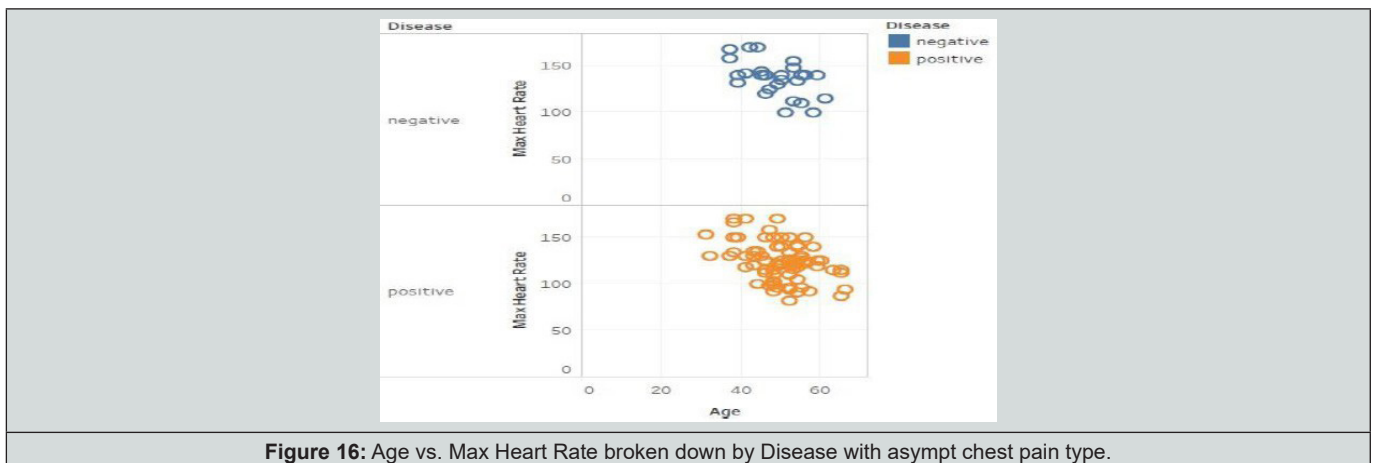

Figure 16: Age vs. Max Heart Rate broken down by Disease with asympt chest pain type.



Figure 17: Age vs. Max Heart Rate broken down by Disease with atypical angina chest pain type.

References

1. M Gudadhe, K Wankhade, S Dongre (2010) Decision support system for Decision support system for heart disease based on support vector machine and artificial neural network," 2021 Int. Conf. Compute. Common. Technol. ICCCT-2010 pp. 741-745.
2. K Thenmozhi, P Deepika (2014) Heart Disease Prediction Using Classification with Different Decision Tree Techniques. Int J Eng Res Gen 2(6): 6-11.
3. PPR Patil, PSA Kinariwala (2017) Automated Diagnosis of Heart Disease using Data Mining Techniques. Int J Adv Res Ideas Innov 3(2): 560-567.
4. SK Mohan, C Thirumalai, G Srivastava (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 7.
5. SP Bingulac (1994) On the Compatibility of Adaptive Controllers. Proc Fourth Ann Allerton Conf Circuits and Systems Theory Pp: 8-16.
6. S Nikhar, AM Karandikar (2016) Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Advanced Engineering, Management and Science (IJAEMS) 2(6): 617-627.
7. IS Jacobs, CP Bean (1963) Fine particles, thin films and exchange anisotropy. In Magnetism III, GT Rado, H Suhl Eds. New York: Academic Pp: 271-350.
8. Aditi G, Gouthami K, Isha P, Kailas D (2018) Prediction of Heart Disease Using Machine Learning. Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).
9. Abhay K, Ajay K, Karan S, Maninder P, Yogita H (2018) Heart Attack Prediction Using Deep Learning. International Research Journal of Engineering and Technology (IRJET) 5(4): 4420-4423.
10. A Lakshmana Rao, Y Swathi, PSS Sundareswar (2019) Machine Learning Techniques for Heart Disease Prediction. International Journal of Scientific & Technology Research 8(11): 374-377.



This work is licensed under Creative Commons Attribution 4.0 License
DOI: [10.19080/JOCCT.2025.19.556004](https://doi.org/10.19080/JOCCT.2025.19.556004)

Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission

<https://juniperpublishers.com/online-submission.php>