***Biological & visual data mining for***

***Predicting Cardiovascular Disease***

***by machine learning algorithms***

Table of Contents

[Abstract 3](#_Toc132554596)

[Problem Introduction 4](#_Toc132554597)

[Goals & Objectives to achieve 4](#_Toc132554598)

[Literature Review 4](#_Toc132554599)

[From Data mining’s point of view 5](#_Toc132554600)

[Machine learning Algorithms 5](#_Toc132554601)

[Deep Learning 6](#_Toc132554602)

[KNN and Random Forest 6](#_Toc132554603)

[Naive Bayes models 7](#_Toc132554604)

[Proposed System 7](#_Toc132554605)

[System Architecture 7](#_Toc132554606)

[Dataset 8](#_Toc132554607)

[Methods 8](#_Toc132554608)

[Support Vector Machine (SVM) 8](#_Toc132554609)

[Random Forest 8](#_Toc132554610)

[Decision Tree Algorithm 9](#_Toc132554611)

[Logistic Regression Algorithm 10](#_Toc132554612)

[Ada-Boost 10](#_Toc132554613)

[Analysis of Results 10](#_Toc132554614)

[Future dimensions & Conclusion 11](#_Toc132554615)

# 

# 

# Abstract

The use of machine learning has spread throughout the globe. The state of health is not remarkable. Learning machines might potentially diagnose cardiac, movement, and other conditions. In the not-too-distant future, the data may aid medical professionals modify patient diagnosis and treatment. Learning algorithms on computers will be able to predict human cardiac conditions. Comparisons will be made between the effectiveness using random forests, random forests logistical regression, decision tree models, and support vector machines (SVM). The purpose of this study is to show a model for classification that can learn data using an extensive number of instances and verify that it is correct using both strong as well as weak classification techniques. This article made a comparison of several models and proposed a classifier that is more accurate and predictive.

# Problem Intro

According to the WHO, 12 million humans die annually by coronary disease. Heart disease is one among the main contributors to worldwide morbidity and mortality. The early identification of cardiovascular diseases is one of the most crucial topics in the segment on data analysis. In recent years, cardiovascular disease has swiftly spread around the globe. Numerous studies are being conducted in the hope to accurately predict the risk level and identify the most significant cardiovascular disorder risks. Even today, cardiovascular disease is known as a "silent killer" due to the fact that it murders individuals without apparent symptoms. When cardiovascular disease is diagnosed early in patients at high risk, it is simpler to implement lifestyle modifications that reduce the risk of consequences. Using machine learning, the enormous quantity of data produced in the healthcare sector can be successfully analyzed and predicted. Using a machine-learning algorithm, this study aims to classify patient data and predict an individual's future risk for cardiovascular disease. Consequently, methods for machine learning can be advantageous. Even though coronary disease can occur in a variety of contexts, there are a standard set of fundamental risk factors that determine whether or not a person is at risk for developing coronary disease. We can conclude that this method is ideally adapted for predicting cardiovascular disease because it collects data from multiple sources, arranges it under the appropriate headings, and then conducts an analysis to extract the necessary details.

# Goals & Objectives to achieve

The primary objective of the study is to present the cardiovascular disease predictor model, which forecasts the commencement of heart disease. The most effective classification system for determining whether a patient may have heart disease is an additional objective of this investigation. This effort receives backing at multiple levels of evaluation by compared research and analysis utilizing classification techniques Random Forest, the Decision Tree, and Naive Bayes. Although these are prevalent machine learning techniques, cardiovascular disease diagnosis is a significant endeavor that requires the highest degree of precision. The end result is the evaluation of three algorithms using a number of evaluation methodologies and levels. Researchers and healthcare providers will have improved access to this.

# Literature Review

Along with advancements in machine learning as well as medical science, countless investigations and trials have resulted in significant publications in recent years.

## From Data mining’s point of view

Using data mining techniques including Neural Network, Nave Bayes, and Decision Tree, a prototype Smart Predicting Cardiovascular Disease System was created. When it comes to accomplishing the specified mining objectives, these results indicate that each method has its own distinct advantages. In the words of Sellappan Palaniappan et al. ((Palaniappan & Awang, 2008)), IHDPS, or can answer sophisticated "what if" queries in a manner that traditional decision support systems cannot. The results demonstrated the rudimentary effectiveness of each methodology in comprehending the stated mining objectives. Unlike conventional support for decisions systems, IHDPS could respond to questions. It made it simpler to implement crucial information, such as the connections and trends between medical factors associated with cardiovascular disease. The IHDPS remains user-friendly, reliable,scalable, and internet-expandable.

## Machine learning Algorithms

In their study entitled "Predictions of Cardiovascular Illness with Machine Learning Algorithms," Santhana Krishnan J, et al. used both the decision tree and the Naive Bayes algorithm for predicting cardiovascular disease. The decision tree algorithm constructs a tree and produces True/False decisions based on a set of conditions. Depending on the dependent variables, the outcomes of techniques like SVM and KNN depend on either vertical or horizontal division conditions. With the aid of a decision tree, the significance of the dataset's characteristics can be determined. A decision tree is a tree with a root node, leaves, and branches contingent on a decision taken in each tree. Additionally, they utilized the Cleveland collection of data. The dataset is divided into 70% training and 30% assessment using an assortment of strategies. The accuracy of this algorithm is 91%. The second utilizes the classification method of Naive Bayes. It is suitable for a dataset on cardiac disease, which is similarly complex, dependent, and nonlinear, because it can manage intricate, dependent, and nonlinear data. The method in question has an 87% rate of accuracy(Krishnan.J & S, 2019),

The research presented in Sonam Nikhar et al.'s article titled "Predictions of Heart Disorders Through Machine Learning Algorithms" provides an in-depth examination of both naive bayes algorithms as well as decision tree classifier, which are commonly used in the estimation of cardiovascular disease. During a portion of the research, it was hypothesized that the identical data would benefit to the implementation of this anticipated data mining technique. As a consequence, the Decision Tree method was found to be more accurate than the Bayesian classifier. Detection of Cardiovascular Disease Using Data Mining Techniques (Nikhar & Karandikar, Prediction of Heart Disease Using Machine Learning Algorithms, 2016).

Aditi Gavhane et al. prepare and evaluate the data using a neural network that generates a multi-facet perceptron. In this approach, an input, an output, and one or more concealed layers are present. The hidden layer connects each input node to the output layer. The weights of this connection are determined that random. The additional input, bias, is weighted according to demand. A form of feed forwarding or return link is conceivable between the nodes (Gavhane, Pandya, Kokkula, & Devadkar, 2018).

According to "Machine Learning Methods for Heart Disease Prediction" by Lakshmi Rao et al., there are additional factors that influence cardiovascular disease. Consequently, identifying coronary artery disease can be difficult. Using various neural network and data mining techniques, the intricate nature of cardiovascular disease among people is assessed. (Lakshmanarao, Swathi, & Sri Sai Sundareswar, 2019)

## Deep Learning

Abhay Kishore et al. suggested "Heart Attack Prediction Using Deep Learning". Recurrent neural networks and deep learning algorithms predict heart-related infections. This deep learning-data mining model is the most accurate and error-free. This publication is a good reference for another heart attack prediction model. (Kishore, Kumar, Singh, Punia, & Hambir, 2018).

Senthil Kumar Mohan et al. presented "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" to improve heart disease accuracy. KNN, LR, SVM, and NN calculations provide crossover irregular woodland direct model cardiovascular disease accuracy of 88.7%. (MOHAN, THIRUMALAI, & SRIVASTAVA, 2019).

## KNN and Random Forest

Aditi Gavhane et al. identified heart failure to prevent deaths. Machine learning dominates this paper. This forecast saves lives. KNN and randomly generated forest algorithms predict heart attacks here (Gavhane, Prediction of Heart Disease Using Machine Learning, 2018).

SVM outperformed Naivy Bayes, KNN, and decision tree in Kumar et al.'s data mining. These methods were trained on the 14-feature, 303-sample UCI machine learning dataset (Kumar, Koushik, & Deepak, 2018).

Sonam Nikhar and colleagues wrote "Heart Disease Prediction through Machine Learning Algorithms." This paper describes our decision tree and neural Bayes classifiers, which predict cardiovascular disease. According to research that applied predictive data mining approaches utilizing the same data, Decision Tree beats the Bayesian classification scheme. (Nikhar & Karandikar, Prediction of Heart Disease Using Machine Learning Algorithms, 2016)

et al., Nagaraj M. Lutimath predicted cardiovascular illness using SVM and Naive Bayes classification. Presentation estimates are Mean Outright Mistake, Amount of Squared Blunder, and Root Mean Squared Blunder. SVM is more precise than Credulous Bayes (Lutimath, C, & Pol, 2019).

## Naive Bayes models

M. Satish used the "pure classifier association rule" with naive Bayes models and decision trees to identify heart disease. He used an archived cardiovascular disease dataset for this strategy. (Al-Bayaty & Joshi, 2016)

According to a literature analysis and survey for this work, the studies indicated above have explored cardiovascular disease machine learning ideas. Using many machine learning methods to assess the predictor with maximum accuracy may provide positive results.

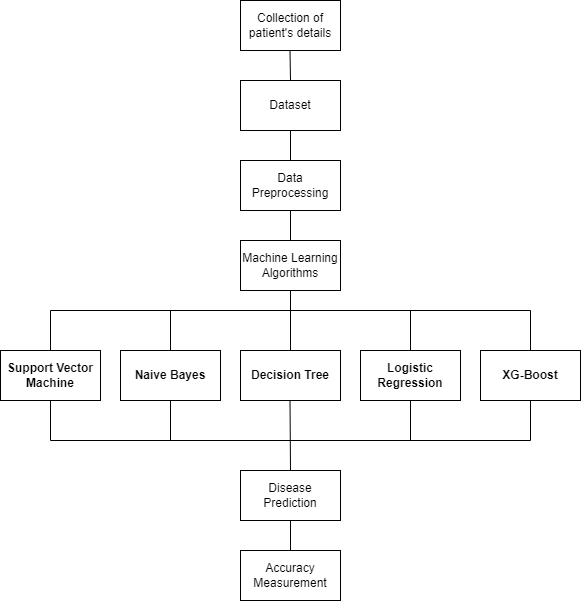
# Proposed System

Data capture and feature selection are the initial steps in system functioning. Then the data is formatted. Assessment & training data are segregated. Along with to these methods, training data trains models. Matching system efficiency to test data determines accuracy. These modules implement the system.

* Data compilation
* Attribute selection
* Data Preprocessing
* Data Balancing
* Disease Prediction

# System Architecture

The system's architecture provides a high-level comprehension of how the system operates. Following is a description of how the system operates: A collection is the process of acquiring data collections containing patient-specific information. Using an attribute selection technique, the components that are crucial for predicting heart disease are chosen. After locating all available data sources, they are selected, classified, and given the appropriate structure. In order to accurately predict cardiac disease, an array of methods for classification would be applied to preprocessed datasets. The accuracy Metrix is to assess the efficiency in different implemented algos.



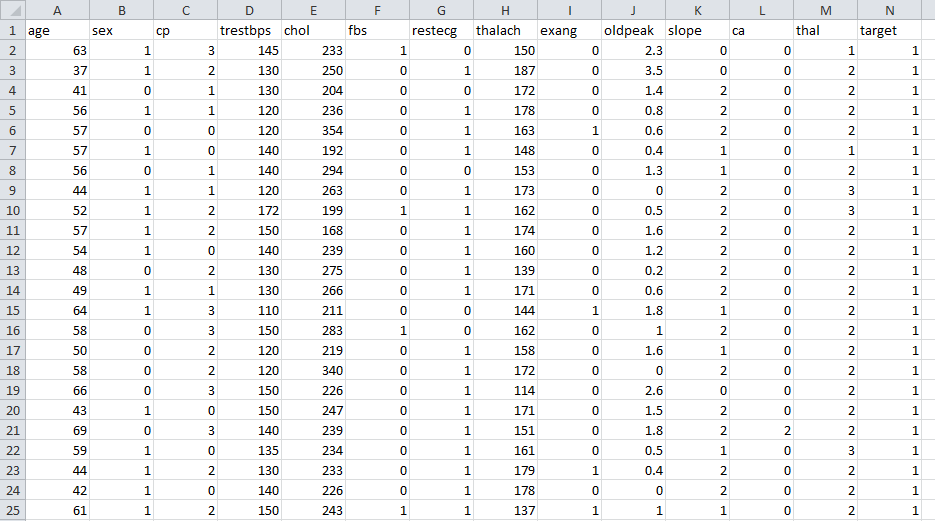
Figure 1 architecture of the system

Figure 2 shows dataset attributes

# Dataset

Fourteen of the seventy-six classes present in the dataset are considered for the most accurate classification. (Fedesoriano, 2022).

# Methods

## Support Vector Machine (SVM)

SVM is a popular supervised classification and regression algorithm. Machine Learning classification problems often use it. SVM generates a decision limit or optimum line to classify n-dimensional spaces quickly categorize data points. Hyper-plane is the best choice limit. SVM selects extreme coordinates for a hyperplane.

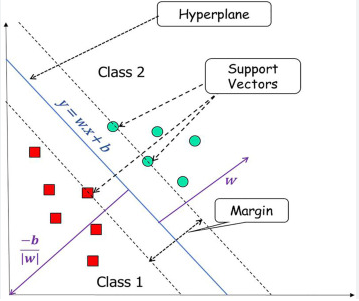


Figure 3 Support Vector Machine working

The technology is called Support Vector Machine because support vectors describe these unusual circumstances. SVM regression and classification methods are effective and adaptable supervised machine-learning methods. Thus, categorization issues utilize all of them. (Kanade, 2022)

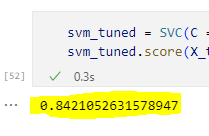


Figure 2 Support Vector Machine accuracy score

## Random Forest

Random Forest modifications for machine-learning classifiers include the decisions tree. Trees use a randomly selected vector and combine tree predictors. Every tree gets split uniformly. Random forest modeling uses the best prediction subgroup, randomly picked from node elements, instead of factors.

The most likely scenario for learning with Random Forests features a temporal overhead O (M (dnlogn)), where M is the overall number of growing trees, n is the total amount of incidents & d is the data set dimension. (G, 2022).

It can do classifications, regression, and regression. It also accurately assesses the feature's value. Feature selection, picture classification, and systems for recommendations use Random Forests. Ensemble analysis combines classification approaches to handle difficult issues and increase model performance. A random forest uses forecasts from each tree and guesses the result using a majority of predictions. The additional trees in a forest boost accuracy and prevent overfitting.

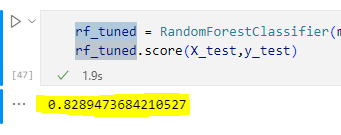


Figure 3 Random Forest accuracy score

## Decision Tree Algorithm

Despite being favored for regression, decision trees may address classification and regression issues. Nodes indicate data properties and routes describe the classification's decision-making process in this classification, and leaf nodes representing the classification's outcome. Decision and Leaf Nodes make up a decision tree. Leaf nodes represent the consequences of choices, whereas decision nodes generate decisions and include multiple branches. Evaluation is based on dataset features arrive at the determination. It refers to the visual representation of obtaining all potential solutions to a problem using predetermined parameters. The is referred to as a tree of decision-making because, similar to shaped like a tree, it starts at the base and radiates out. Trees are built using the CART technique. The decision tree asks a question and separates into sub-trees depending on the answer (Yes/No). The objective of this strategy appears to be the development of a model that predicts the probable outcome of a specified quantity. To achieve this, a choice tree was utilized, where a leaf node corresponds to the class identifier and the internal nodes represent characteristics. Remember that building a machine-learning model requires selecting the most effective approach for a given dataset and problem. (Greenwell, 2020).

Due to its tree-like structure, the theory underlying the decision tree is readily interpretable; as a result, decision trees typically mimic the human decision-making process.

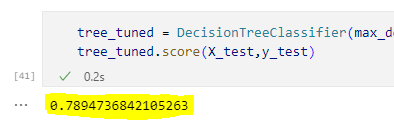


Figure 4 Decision Tree accuracy score

## Logistic Regression Algorithm

Logical regression is a popular Supervised Learning technique. It may evaluate the category wise variables using mentioned components.

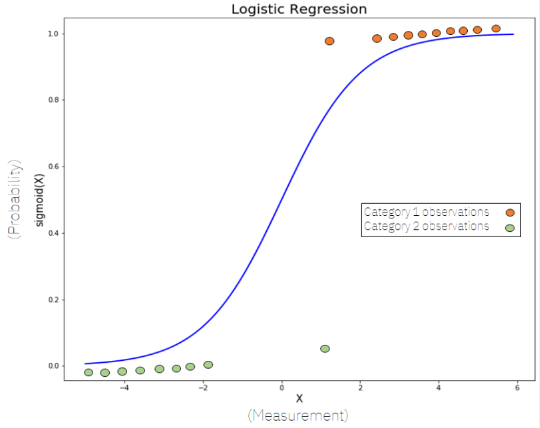


Figure 5 Logistic Regression

Logistic regression predicts categorical data using dependent variables. The output should include discrete or categorical values. It estimates probability between zero and one, not exact numbers. Yes, No, 0, 1, etc. The methods of logistic regression & linear regression are similar except for the application process. Linear regression handles regression concerns, whereas logistic regression addresses classification issues. Instead of a regression range, logistic regression creates a "S"-shaped logistic function. The arc of a LR function indicates the probabilities of numerous factors, such as the identified that a cell is cancerous, that the mice will be fat depending on its weight, etc. Logistic regression is main in machine learning because it classifies discrete & continuous information (IBM).

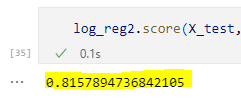


Figure 6 accuracy of Logistic Regression Algorithm

## Ada-Boost

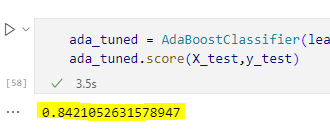
Adaboost was the first superior binary classification method. "Adaptive boosting," additionally referred to as "Adaboost," is a popular boosting approach that combines numerous "sensitive classifiers" into one "strong and resilient classifier." Despite its simplicity, it has won many information competitions recently (Polzer, 2022). 

Figure 7 AdaBoost accuracy

# Analysis of Results

We use machine learning to analyze and train a supported vector machine model, which is more precise than previous approaches. The table below shows which SVM or AdaBoost is are better with 84.2% accuracy. Only a confusion matrix for every method may assess accuracy.

Table 1 Accuracies of all models

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| Support Vector Machine | 84.2% |
| Ada-Boost | 84.2% |
| Random Forest classifier | 82.8% |
| Logistic Regression classifier | 81.5% |
| Decision Tree classifier | 78.9% |

# Future dimensions & Conclusion

Heart disease is the primary reason of mortality on a global scale; therefore, the use of cutting-edge technologies like machine learning to foresee cardiovascular illness early on has a substantial societal impact. Due to the earlier detection of cardiac disease, a significant advancement in the discipline of medical care may entail high-risk patients deciding on adjustments in lifestyle to minimize complications. Numerous individuals are diagnosed annually with cardiac disease. It requires both earlier detection and action requirement. Patients and the healthcare industry will benefit from appropriate management assistance in this discipline. The SVM, a Random Forest classifier, a Decision Tree algorithm with Logical Regression, and the Adaboost are the five different machine learning techniques examined in this study. All of these techniques were applied to the information at hand. This data set, consisting of 76 attributes, contains the anticipated features that contribute to heart disease in humans, & fourteen additional classes chosen to aid in determining the system. All of the other features are considered, the creator is left with a system that is significantly less efficient. To increase efficiency, attribute selection is employed. In this scenario, m number of qualities must be analyzed to evaluate the model’s performance with maximum level of precision. Countless data classes are excluded because their co-relations are easily identical. If we take unnecessary attribute in the data, the precision decreases dramatically. Before developing a prediction model, the precision of all five machine-learning methods is evaluated. SVM and Adaboost have a success rate of 84.2% if comparing all five methods.

# Bibliography

Al-Bayaty, B. F., & Joshi, S. (2016). Comparative Analysis between Naïve Bayes Algorithm and Decision Tree to Solve WSD Using Empirical Approach. *Lecture Notes on Software Engineering*, 1-5.

Fedesoriano. (2022). *Heart Failure Prediction Dataset.* Retrieved Jan 20, 2023, from kaggle: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

G, M. (2022, April 5). *Disease Prediction using Machine Learning.* Retrieved Jan 20, 2023, from IJRASET Journal for Research in Applied Science and Engineering Technology: https://www.ijraset.com/research-paper/disease-prediction-using-ml

Gavhane, A. (2018). Prediction of Heart Disease Using Machine Learning. *Conference: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*.

Gavhane, A., Pandya, I., Kokkula, G., & Devadkar, P. K. (2018). Prediction of Heart Disease Using Machine Learning. *Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018)*, 1-4.

Greenwell, B. B. (2020, Fev 1). Hands-on machine learning with R. *Chapter 9 Decision Trees*. Retrieved from Chapter 9 Decision Trees.

IBM. (n.d.). *What is logistic regression?* Retrieved Jan 20, 2023, from IBM: https://www.ibm.com/topics/logistic-regression

Kanade, V. (2022, Sep 29). *What is a support vector machine? working, types, and examples*. Retrieved Jan 20, 2023, from All You Need to Know About Support Vector Machines: https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/

Kishore, A., Kumar, A., Singh, K., Punia, M., & Hambir, Y. (2018). Heart Attack Prediction Using Deep Learning. *International Research Journal of Engineering and Technology (IRJET)*, 1-4.

Krishnan.J, M., & S, D. (2019). Prediction of Heart Disease Using Machine Learning Algorithms. *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 1-5.

Kumar, M. N., Koushik, K. V., & Deepak, K. (2018). Prediction of Heart Diseases using Data Mining and Machine Learning Algorithms and Techniques.

Lakshmanarao, A., Swathi, Y., & Sri Sai Sundareswar, P. (2019). Machine Learning Techniques For Heart Disease Prediction. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 1-4.

Lutimath, N. M., C, C., & Pol, B. S. (2019). Prediction of Heart Disease using Machine Learning. *International Journal of Recent Technology and Engineering (IJRTE)*, 1-4.

MOHAN, S., THIRUMALAI, C., & SRIVASTAVA, A. G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *SPECIAL SECTION ON SMART CACHING, COMMUNICATIONS, COMPUTING AND CYBERSECURITY FOR INFORMATION-CENTRIC INTERNET OF THINGS*, 1-13.

Nikhar, S., & Karandikar, A. (2016). Prediction of Heart Disease Using Machine Learning Algorithms. *International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication*, 1-5.

Nikhar, S., & Karandikar, A. (2016). Prediction of Heart Disease Using Machine Learning Algorithms. *International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication*, 1-5.

Palaniappan, S., & Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. *IJCSNS International Journal of Computer Science and Network Security*, 349.

Polzer, D. (2022, Aug 10). *AdaBoost, step-by-step.* Retrieved Jan 20, 2023, from Medium: https://towardsdatascience.com/adaboost-in-7-simple-steps-a89dc41ec4