World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Cardiac ailment recognition using ML techniques in E-healthcare

Calabe P S *, Prabha R and Veena Potdar

Department of CS & E, Dr. Ambedkar Institute of Technology, (Affiliated to VTU, Belagavi). Bengaluru, Karnataka, India.

World Journal of Advanced Research and Reviews, 2023, 17(01), 302–307

Publication history: Received on 22 November 2022; revised on 07 January 2023; accepted on 09 January 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.17.1.0010

Abstract

Heart ailments can take numerous forms, and they are frequently referred to as cardio vascular illnesses. These can range from heart rhythm problems to birth anomalies to blood vessel disorders. It has been the main cause of death worldwide for several decades. To recognize the illness early and properly manage, it is critical to discover a precise and trustworthy approach for automating the process. Processing massive amounts of data in the field of medical sciences necessitates the application of data science. Here we employ a range of machine learning approaches to examine enormous data sets and aid in the accurate prediction of cardiac diseases. This paper explores the supervised learning models of Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, in order to provide a comparison investigation for the most effective method. When compared to other algorithms, K-Nearest Neighbor provides the best accuracy at 86.89%.

Keywords: Heart Disease Prediction; Support Vector Machine; Naïve Bayes; K-Nearest Neighbor; Decision Tree

1. Introduction

Cardiovascular diseases have been the main cause of death worldwide over the previous 10 years. According to the World Health Organization, more than 17.9 million people die each year as a result of cardiovascular diseases, with coronary artery disease and cerebral stroke accounting for 80% of these fatalities [1]. Personal and occupational behaviors, as well as inherited predisposition, all have a role in the development of heart disease. A range of risk factors, including smoking, excessive alcohol and caffeine use, stress, and physical inactivity, as well as physiological variables such as obesity, hypertension, high blood cholesterol, and pre-existing heart disease, are commonly associated with cardiac illness. An efficient, precise, and early medical diagnosis of heart disease is necessary in order to take preventative measures to avoid the challenges that these illnesses bring.

The provision of high-quality services, as well as effective and accurate prediction, is now the most crucial matter confronting the area of medical sciences. Automation, such as data mining and machine learning, can be used to automate the latter issue. The practice of obtaining valuable information from a large collection of raw data is known as data mining. It comprises analyzing patterns in big data sets using multiple tools. Data collection, warehousing, and computer processing must all be done effectively. Machine learning (ML) in data mining successfully handles large, well-formatted datasets. Machine learning has the potential to be utilized in the medical business to diagnose, detect, and forecast a wide range of disorders. A number of machine learning methods are investigated to see which is the most accurate, including Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Decision Tree ensemble approach. In this case, the dataset on cardiac disease from the UCI repository is used. This paper discusses and compares the current categorization systems. The paper also analyses the potential for growth and the scope of future research.

* Corresponding author: Calabe P S

2. Related Work

- Ramalingam et al [1]. Predicted several diseases using decision trees and naïve bayes data mining methods. They were primarily concerned in predicting diabetes, breast cancer, and heart disease. The findings were obtained using the confusion measures.
- Paul CJ et al [2] developed a machine learning model that compares five techniques. In terms of accuracy, Rapid Miner outperformed Matlab and Weka. This article studied the classification accuracy of the approaches Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM. The decision tree algorithm gave the best output.
- Hoke TR et al [3]. Investigates many ML algorithms that can be used to classify heart illness. Through study, the accuracy of the classification methods Decision Tree, KNN, and K-Means was compared. The study discovered that Decision Trees had the highest accuracy, and it was decided that it might be made more successful by merging multiple approaches and fine-tuning its parameters.
- K Gomathi et al [4]. Predicted several diseases using decision trees and naïve bayes data mining methods. They were primarily concerned in predicting diabetes, breast cancer, and heart disease. The findings were obtained using the confusion measures.
- Bouali H et al [5]. Used many classification approaches to conduct a survey to predict heart disease. Several classification algorithms, including Naive Bayes, KNN (K- Nearest Neighbor), decision trees, neural networks, and others, were used to examine the accuracy of the classifiers for a variety of features.

3. Data Source

The Cleveland heart data sample from the machine learning repository at UCI was used in the experiments. There are 303 occurrences and 14 attributes in the collection. There are eight category categories and six number qualities.

Table 1 Description of the Dataset

Attribute	Description	Range
Age	Age of person in years	29-79
Sex	Gender of person (1-M 0-F)	0,1
Cp	Chest pain type	1,2,3,4
Trestbps	Resting blood pressure in mm Hg	94-200
Chol	Serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar in mg/dl	0,1
Restecg	Resting Electrocardiographic results	0,1,2
Thalach	Maximum heart rate achieved	71-202
Exang	Exercise Induced Angina	0,1
OldPeak	ST depression induced by exerciserelative to rest	1-3
Slope	Slope of the Peak Exercise ST segment	1,2,3
Ca	Number of major vessels colored byfluoroscopy	0-3
Thal	3 – Normal, 6 – Fixed Defect, 7 –Reversible Defect	3,6,7
Result	Class Attribute	0,1

The above given table provides information on the dataset. This dataset contains patients ranging in age from 29 to 79. Male and female patients are identified using gender values of 1 and 0, respectively. There are four types of chest pain that might indicate heart disease. Type 1 angina is caused by narrowed coronary arteries, which restrict blood flow to the heart muscles. Type 1 angina causes chest pain due to mental or emotional stress. Chest pain that isn't caused by angina might have a variety of causes and isn't necessarily connected to true heart disease. The fourth

category, asymptomatic, might not be an indicator of heart disease. The next attribute is resting blood pressure measurement it is denoted by the symbol trestbps. The amount of cholesterol is chol. Fbs stands for fasting blood sugar level; a value of 1 is given if it is less than 120 mg/dl and a value of 0 if it is more. The term "restecg" stands for "resting electrocardiographic result," "thalach" for "maximum heart rate," "exang" for "exercise-induced angina," "oldpeak" for "exercise-induced ST depression," "slope" for "peak exercise ST segment," "ca" for "number of major vessels coloured by fluoroscopy," "thal" for "exercise test duration in minutes," and "num" for "class attribute." Patients with heart disease have a value of 1 for the class attribute, while normal people have a value of 0.

4. Methodology

4.1. Classification Algorithms:

Classification is a supervised learning strategy used to predict outcomes using past data. This study proposes a way for identifying heart disease using classification algorithms. Unique classifiers are trained by using training dataset, which is divided into a training set and a test set in an 80:20 ratio. The efficacy of the classifiers is assessed using the test dataset. The following section explains how each classifier works.

4.1.1. Decision tree

A decision tree classification method may be utilized with both categorical and numerical data. Decision trees are used to build structures that look like trees. A tree-shaped graph's data is straightforward to construct and analyze. This algorithm separates the data into two or more related sets based on the main indicators. The information gain or entropy of each characteristic is used to segregate the data, with predictors having the largest information gain or the lowest entropy. The produced findings are easy to read and comprehend. This method is gives good accuracy since it examines the dataset in a tree-like graph. However, the data may be over classified, and only one characteristic is checked for decision-making at a time.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad \text{——— (1)}$$

$$IG(Y, X) = E(Y) - E(Y|X) \quad \text{——— (2)}$$

4.1.2. Naïve Bayes

A supervised algorithm is the Naive Bayes classifier. It is a straightforward classification method based on the Bayes theorem. It presumes that each value is independent. To determine the likelihood, mathematicians apply the Bayes theorem. The predictors don't have any connections to one another or a correlation with one another. To increase the likelihood, each quality separately contributes. Naive Bayes classifiers are used in many difficult real-world scenarios.

$$P(X/Y) = P(Y/X) \times P(X)/P(Y) \quad \text{——— (3)}$$

Note:

P(X/Y) is the posterior probability,
P(X) is the class prior probability,
P(Y) is the predictor priorprobability,
P(Y/X) is the likelihood, probability of predictor.

4.1.3. Support Vector Machine

A Support Vector Machine (SVM) model is just a hyper-plane in multidimensional space that represents several classes. SVM will construct the hyper-plane through an incremental approach in order to reduce error. SVM aims to classify datasets in order to identify a Maximum Marginal Hyper-plane (MMH).

4.1.4. K-Nearest Neighbor

The K-Nearest Neighbor method is a supervised classification methodology. Objects are classified depending on their nearest neighbor. It is a strategy to learning based on instances. The Euclidean distance is used to determine how far a property is from its neighbors. It employs a collection of named points for the purpose of designating another point. The data is classified based on how closely they resemble one another. The K-NN approach is simple to use without the need for a model or any assumptions. This method is adaptable and may be used for search, regression, and classification. Despite being the most straightforward technique, K-accuracy NNs are influenced by noise and irrelevant information.

5. Flow Chart of the Model

The flowchart depicts the process of using the dataset while developing a prediction model. This flow chart is essential for understanding this study report.

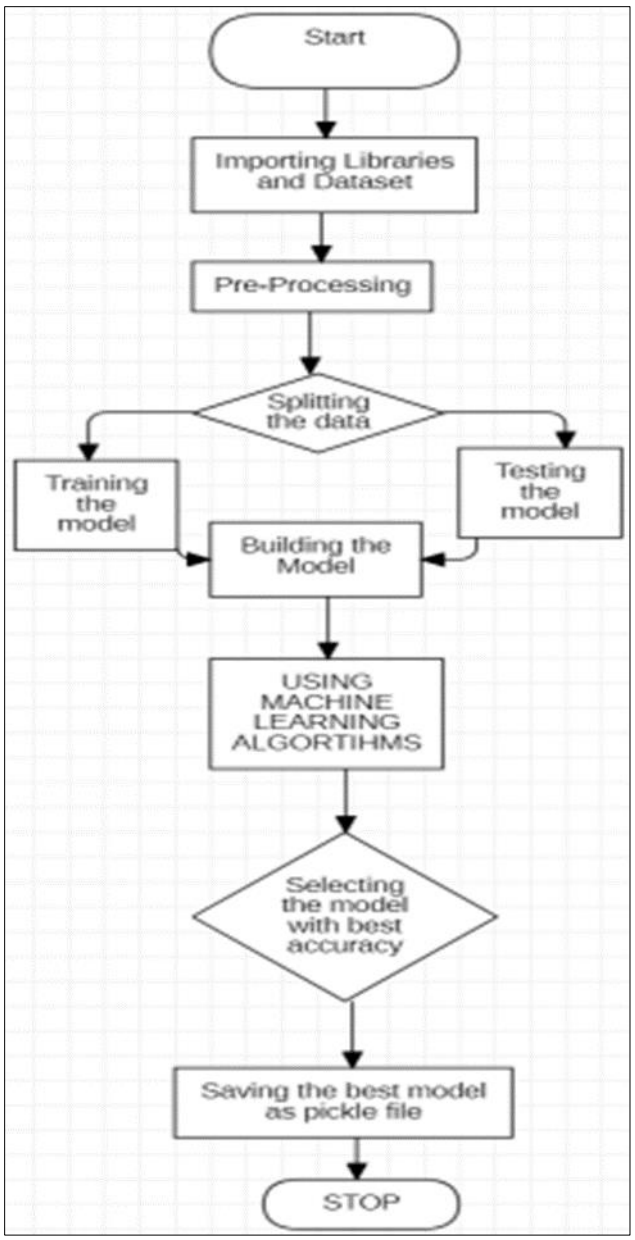

Figure 1 Flow chart of the Model

6. Result and Discussion

The purpose of this study is to evaluate the effectiveness of several classification algorithms in order to identify the best reliable algorithm for predicting whether or not a patient would have cardiac disease. On the UCI dataset, this study used Naive Bayes, Support Vector Machine, K-Nearest Neighbor, and Decision Tree approaches.

Python was used to train the models, separate the dataset into training and test data, and measure model correctness. The table below details how accurate they were and graph given below shows how the algorithms performed in comparison.

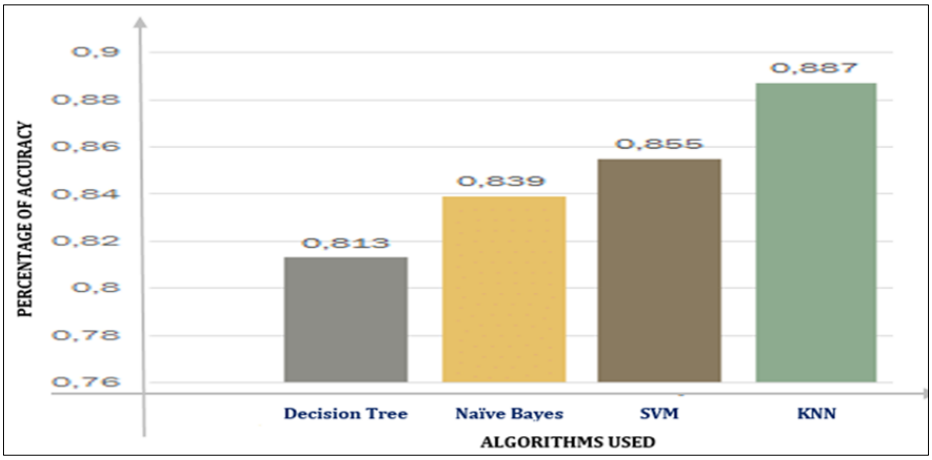

Figure 2 Accuracy Percentage of Algorithms used

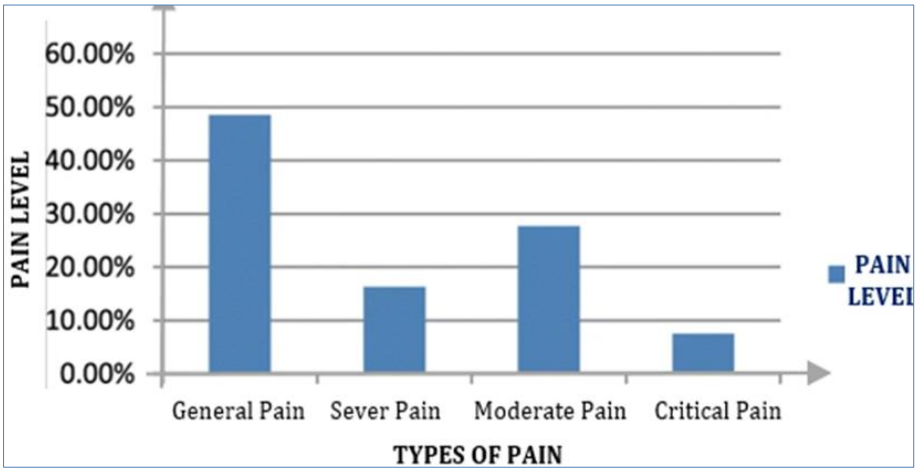

Figure 3 Cardiac Pain Level Percentage

Table 2 Accuracy Obtained by different Algorithm

Algorithm	Accuracy
Naïve Bayes	83.05%
Support Vector Machine	85.77%
K-Nearest Neighbor	86.89%
Decision Tree	81.05%

The above graph details the different pain level observed in the patient such as general pain, severe pain, moderate pain and critical pain.

7. Conclusion

The overarching goal is to identify several data mining methods that may be effectively used to forecast cardiac disease. The aim of this study is to develop efficient and effective prediction methods using fewer features and tests. The model employed pre-processed data that had been previously altered. The most effective algorithms are Support Vector Machine with 85.77% and K Nearest Neighbor with 86.89%. However, Decision Tree performed with 81.05% accuracy, which was less accurate. This research may be expanded by including more data mining methods, such as time series, clustering and association rules, and other ensemble approaches. Given the limitations of this study, more complicated and combined models must be used in order to improve the accuracy of heart disease early prediction.

Compliance with ethical standards

Acknowledgments

The authors would like to appreciate all subjects who participated in the present study.

Disclosure of conflict of interest

The authors have no conflicts of interest to declare.

References

[1] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[2] Paul CJ, Kim D, Miranda ML, Hull AP, Galeano MA. Changes in blood lead levels associated with use of chloramines in water treatment systems. Environ Health Perspect. 2007;115:221–225. [PMC free article] [PubMed] [Google Scholar]

[3] Hoke TR, Seckeler MD. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.

[4] K.Gomathi Heart Disease Prediction Using Data Mining Classification

[5] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.

[6] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine- learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[7] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p.204–207