Advanced Techniques in Biology & Medicine
ISSN: 2379-1764

OPEN ACCESS Freely available online

Research Article

Effective Heart Disease Detection Using Machine Learning Techniques

Kashish Agarwal*, **Ayush Singh**, **Hrithik Maheshwari**

Department of Information Technology, Galgotias College of Engineering and Technology, Uttar Pradesh, India

ABSTRACT

According to World Health Organization (WHO) Heart diseases are the major cause of death all across the globe, estimating about 17.9 million lives (32% of overall deaths) each year. This group of disorder includes coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. The most of the daily activity behavioural risk factors of heart disease and stroke includes unbalanced diet, physical inactivity, inertness, consumption of tobacco and alcohol. These risk factors may show up among people as raised blood pressure, raised blood sugar level, raised blood lipids and obesity. These intermediate risks factors can be measured in primary care facilities and helps in indicating increased risk of heart complications such as heart attack, stroke, heart failure.

As a traditional method, detection of disease is done by a doctor based on the laboratory test reports. This process involves consultation with multiple doctors by the patient in order to decrease the human error coefficient which not only costs a lot of money but also takes huge time. As a solution to solve this problem, various machine learning based techniques are used to provide non-invasive solutions. In this paper, we propose to use such machine techniques which can be used to check whether a patient has some kind of heart disease or not. We evaluate our approach on several benchmark datasets and show that it outperforms existing state-of-the-art and makes significant contribution.

Keywords: Machine learning; Heart diesase detection; Naive bayes; Decision tree; KNN; Support vector machine

INTRODUCTION

One of the leading causes of death worldwide in humans is heart disease which occurs due to inability of heart to push required amount of fresh oxidized blood to the overall body. Hence, it is very important to check the death rate by identifying the disease correctly at a very initial stage. Major symptoms that a heart patient faces are weakness of the body, shortness of breath and swollen feet [1]. Although there have been existing methods that identifies the same but their complexity is one of the biggest reasons which affect the standard of life as well as mortality rate in some of the developed states in the world. Hence some enhanced techniques are required that can help patients in living their daily life activities. The conventional/ invasive methods for detecting coronary artery diseases are regulated through examination of patient's medical records but it is not an effective method to diagnose the same [2].

As an advanced approach, we can use data mining technologies to uncover an understanding from the datasets. This discovered knowledge can then be used by the healthcare personnels to improve the quality of service. The main goal of this paper is to assist the task of doctors by providing a tool that can detect a disease at an early stage and will help in providing an effective treatment as per the need and avoid severe consequences. Machine learning in this plays a very important role by detecting the hidden discontinuous patterns and thereby analyzing the given data [3]. This paper presents performance analysis of various machine learning techniques such as Naive bayes, decision tree, logistic regression and random forest for predicting heart disease at an early stage. The main objective of this paper on effective heart disease detection using machine learning techniques is to develop a system which will be simple and easy to use, as here one must provide the patient's medical details and based on the features extracted, the algorithm will then detect the heart disease at an early stage [4].

Correspondence to: Kashish Agarwal, Department of Information Technology, Galgotias College of Engineering and Technology, Uttar Pradesh, India, Tel: 8057024076; E-mail: kashish.agarwal0101@gmail.com

Received: 01-Mar-2023, Manuscript No. ATBM-23-23998; **Editor assigned:** 06-Mar-2023, PreQC No. ATBM-23-23998 (PQ); **Reviewed:** 20-Mar-2023, QC No. ATBM-23-23998; **Revised:** 22-Jun-2023, Manuscript No. ATBM-23-23998 (R); **Published:** 29-Jun-2023, DOI: 10.35248/2379-1764.23.11.406

Citation: Agarwal K, Singh A, Maheshwari H (2023) Effective Heart Disease Detection Using Machine Learning Techniques. Adv Tech Biol Med. 11:406.

MATERIALS AND METHODS

Algorithms

Naive bayes algorithm: It is a type of supervised learning algorithm, which is established on the basis Bayes theorem and is highly used for solving different classification problems. Basically, it is described on the basis of two words *i.e.* Naive and Bayes [5]:

Naive: It is called Naïve because it presumes that the occurrence of one trait is unrelated to the occurrence of other features, it is known as naive. Such as if the vegetable is identified on the bases of color, shape and taste, then red, spherical, and sour vegetable is recognized as a tomato. Hence each feature individually contributes to identify that an attribute value on a given class is independent of the values of other attributes [6].

Bayes: As a result of its reliance on the Bayes' theorem principle, it is known as Bayes. It depends on the conditional probability.

$P(A|B)=P(B|A)\ P(A)/P(B)$

$P(A|B)$ is posterior probability, $P(B|A)$ is likelihood probability, $P(A)$ is prior Probability, $P(B)$ is marginal probability.

One of the biggest advantage of the naive Bayes classifier is that it needs a minuscule amount of training data in order to estimate various parameters (like means and variances of the variables) which are required for the classification. Since independent variables are taken, only the each class variances are determined rather than the entire covariance matrix [7].

Decision tree algorithm: It is a type of supervised learning technique that is used well for both classification as well as regression problems, However, we mostly prefer it for evaluating classification problems. Most commonly used two nodes are the decision node and the leaf node. It is basically a tree-structured classifier, in which the features of a dataset are represented as internal nodes, the decision rules are represented as branches and the result as the leaf node. In order to make any decision and have multiple branches, we use decision node whereas Leaf nodes are the output of those decisions and do not contain any further branches [8]. We can use information gain techniques for attribute selection method.

Information gain=Entropy(S)-[(Weighted Avg) × Entropy (each feature)]

Entropy(s)=-P(yes) log2 P(yes)-P(no) log2 P(no)

Where

S=Total number of samples P(yes)=probability of yes P(no)=probability of no

K nearest neighbor algorithm: KNN algorithm (also called case based reasoning, k nearest neighbor, example based reasoning, instance based learning, memory based reasoning or lazy learning is an algorithm that is used to classify new cases based on similarity measures of previously stored cases. Select the number K of the neighbors. Identify the Euclidean distance between K neighbours. Choose the K closest neighbours based on the Euclidean distance estimate. Among these k neighbours, count the total the amount of data points in each category. Put the additional data points to the category where the neighbour count is at its highest. Our model is ready [9].

Euclidean distance between the points= $((x_2-x_1)^2+(y_2-y_1)^2)^{1/2}$

The similarity between two instances with n attribute values can be calculated in a variety of ways. Every measure has the following three requirements. Let d (X, Y) be the distance between two points X,Y then d(X,Y)>=0 and d(X,Y)=0

if X=Y

d(X,Y)=d(Y,X) d(X,Z)<=d(X,Y)+d(Y,Z)

Property 3 is called as "Triangle in equality", according to which the shortest distance between any two points is a straight line. Most common distance measures used is Euclidean distance. For continuous variables Z score standardization and min max normalization are used [10].

Support Vector Machine (SVM) algorithm: It is an important concept of statistics and computer science for a set of correlated supervised learning methods that examine data and acknowledge pattern. SVM have shown good performance in a number of application areas. It constructs a hyperplane or set of hyperplanes in an infinite-dimensional space, which is used well for classification, regression, or other tasks. SVM's are very much useful in data classification. SVM are used to classify data on the basis of optimal hyper plane which separates the dimensional data into its two classes with a maximum interclass margin. They make use of so-called kernel functions so that the data can be casted into a higher dimensional space where the data is separable. [11].

In the case of support vector machines, any data point can be seen as a p-dimensional vector (*i.e.* a list of p numbers), and our task is to know whether we can separate such points with the use of (p-1)- dimensional hyper plane. This is called a linear classifier. It plots all the training vectors in high dimensional space and labels them by their class. It minimizes the error rate as it is based on the principle of risk minimization (Figure 1).

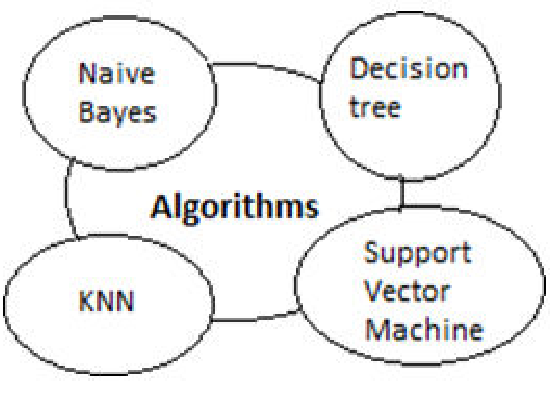

Figure 1: Types of algorithms.

RESULTS AND DISCUSSION

The Cleveland UCI Heart Disease Dataset is used for the training and validation of the model. This dataset consists of approximately 303 patient's record which consists of a total of 13 features columns and 1 target column. Out of these total columns 5 were numerical data, 8 were categorical data, and 1 column was of Boolean data. They are well shown in Table 1 along with their feature type.

Table 1: The cleveland UCI heart disease dataset.

Type	Name
Numerical data	Age
	Resting blood pressure
	Cholesterol
	Max heart rate achieved
	Depression
Categorical data	gender
	Chest pain type
	Fasting blood sugar
	Rest ECG
	Slope
	Number of major vessel
	Thalassemia
	Condition
Boolean data	Exercise induced angina

The methodology involves the following steps:

Data pre-processing: This involves the process of carrying out the few operations on the data so that the dataset becomes accurate and error free.

Data cleaning: It involves the process of removing the fields which does not have any value by substituting it with the mean value of the column.

Feature scaling: It involves the process of feature scaling for the proper functioning of objective function since the range of values of raw data varies widely.

Factorization: The process of assigning a specified meaning to the values so that the algorithm doesn't confuse between them.

Modeling/Training: It involves the process of training various classification models on the training set and see which one yields the highest accuracy. For this we need to compare the accuracy of K-NN (K-Nearest Neighbors), SVM (Support Vector Machine), Naive bayes classifier, Decision trees. Finally, the Confusion matrix will hence determine true positives and true negatives.

Predicting heart disease is difficult and crucial in the medical field. However, we can control the mortality rate at a larger level if the disease is detected at an early stage by adopting necessary preventative measures as soon as possible. As a traditional method, detection of disease is done by a doctor based on the laboratory test reports. This process involves consultation with multiple doctors by the patient in order to decrease the human error coefficient which not only costs a lot of money but also takes huge time. As an advanced approach, we have used data mining technologies to uncover an understanding from the datasets. In order to consider the existing relationships between variables, application of data mining plays an important role as it helps in analysing the medical data. From our proposed approach we have shown how mining helps in retrieving the required correlation even from the attributes that are not directly related to the class that we are trying to predict.

Challenges

Medical diagnosis is related to the life of a being hence it is considered as a one of the most significant yet intricate task that needs to be carried out precisely and error-free. The automation of the same would be highly beneficial. In many cases, the clinical decisions/diagnosis is based on doctor's intuition and experience rather than on the knowledge rich data available in the database. This practice sometimes even lead to unwanted biases, errors and most importantly excessive medical costs which highly affects the patient in terms of the quality of service provided. Data mining have the potential to improve the quality of clinical decisions by generating a knowledge-rich environment provided that the model is well trained using error free training dataset.

CONCLUSION

In this paper, we have covered such machine techniques which can be used to check whether a patient has some kind of heart disease or not. We have evaluated our approach on several benchmark datasets and show that it outperforms existing state-of-the-art and makes significant contribution. The purpose of this work was to compare different machine learning algorithms such as KNN, Naïve Bayes, Decision Tree, and Support Vector Machine with different performance measures. Here all data was pre-processed and then used for test prediction. Some algorithms work better in some situations while others in some other cases. The suggested methods are compared to supervised algorithms based on the underlying approximate sets and classification accuracy measurements are used to evaluate the performance and accuracy of the proposed approaches.

This paper can be considered as the initial step in acquiring knowledge in the diagnosis of heart disease with the automatic learning and it can be extended for future research as well. There are several limitations to this study as well. This may be firstly due to restricted author's knowledge base, secondly, due to the tools that are used in the study and thirdly due to the limited time limit constraint for the study.

REFERENCES

1. Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. Natur Rev Cardiol. 2011;8(1):30-41.
2. Durairaj M, Ramasamy N. A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate. Int J Control Theory Appl. 2016;9(27):255-260.
3. Allen LA, Stevenson LW, Grady KL, Goldstein NE, Matlock DD, Arnold RM, et al. Decision making in advanced heart failure: A scientific statement from the American heart association. Circulation. 2012;125(15):1928-1952.
4. Vanisree K, Singaraju J. Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int J Comp Appl. 2011;19(6):6-12.
5. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. IEEE. 2008;108-115.
6. Colin C, Yiming Ying, Learning with support vector machines. Morgan Claypool, 2011.
7. Barakat N, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Transactions Infon Technol Biomed. 2010;14(4):1114-11120.
8. Mertik M, Kokol P, Zalar B. Gaining features in medicine using various data-mining techniques. IEEE 3rd Int Confer Comp Cybernet. 2005;13:21-24.
9. Suganya G, Dhivya D. Extracting diagnostic rules from support vector machine. J Comp Appl. 2011;4(4):2011.
10. Barakat NH, Bradley AP. Rule extraction from support vector machines: A sequential covering approach. IEEE Transact Knowledge Data Eng. 2007;19(6):729-741.
11. Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R. SVM ranking with backward search for feature selection in type II diabetes databases. IEEE Int Confer Syst Man Cyberneti. 2008;12:2628-2633.