Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Prediction of Heart Disease Using Machine Learning

Sangya Ware[1*] and Shanu k Rakesh[2]

[1]M. tech scholar CSE Department, Chouksey Engineering College Lalkhadan Bilaspur(CG)
[2]Assistant professor, CSE Department, Choukesy Engineering College Lalkhadan Bilaspur(CG*)*

ABSTRACT

Heart disease is most common nowadays and yet very serious problem. According to recent survey by WHO 17.5 million people die of the heart disease every year. Machine learning provides a best way for prediction of heart attack. We are living in an “information age”. The health care industry generates a huge amount of data publicly. The aim of this paper is to develop a simple, light weight approach for detection of heart disease by machine learning implementation using supervised learning method. Heart disease prediction would be made using the training model constructed from publicly available data set. The main reason for death in the world over the last decade is due to heart disease. Most of the Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. However using machine learning technique can reduce the number of test that are required. There have to be a quick and efficient detection technique in order to reduce heart disease. Heart disease is a deadly disease that large population of people around the world suffers from. When considering death rates and large number of people who suffers from heart disease, it is revealed how important early diagnosis of heart disease. Traditional way of diagnosis is not sufficient for such an illness. Developing a medical diagnosis system based on machine learning for prediction of heart disease provides more accurate diagnosis than traditional way. This paper includes heart disease prediction system by using SVM algorithm and random forest classifier. 13 clinical features were used as input for the SVM and then the SVM was trained to predict absence or presence of heart disease. “Accuracy” is used as a performance metric for the analysis and comparison of classification models. By using SVM, the overall accuracy is 86.88%. And by using Random forest accuracy is 81.11%. After comparing both the classifier the performance of SVM is slightly higher than random forest.

KEYWORDS: SVM (Support Vector Machine), Heart Diseases Prediction System, dataset, machine learning, Random forest.

***Corresponding Author**

Sangya Ware

Department of computer science and Engineering,

Chouksey Engineering College,

Bilaspur (C.G.) India

E-mail: sangyaware@gmail.com

INTRODUCTION

Heart disease has created a lot of serious concerns among researches; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. Early techniques have not been so much efficient in diagnosis[1].There are various medical instruments available in the market for predicting heart disease but there are two are very much expensive and secondly, they are not efficient enough to be able to calculate the chance of heart diseases. According to latest survey conducted by WHO, the medical professionals are able to correctly predict only 67% of heart diseases[2]. So, there is a need to find better and efficient approach to diagnose heart diseases at early stage. With advancement of computer science in different research areas including medical sciences, this has been made possible. As application areas of computer science varies from meteorology to ocean engineering and medical sciences. In last decade, artificial intelligence has gain momentum because of the improved technologies and machine learning algorithms. Machine learning implementations are applicable to vast research areas including depression predictions, image and speech recognition, medical sciences, genomics and natural language processing etc. A machine-learning system is trained rather than explicitly programmed. It is presented with many examples relevant to a task, and it finds statistical structure in these examples that eventually allows the system to come up with rules for automating the task[3]. Machine learning could be a better choice for achieving high accuracy for detection of heart diseases. This survey paper is dedicated for wide scope survey in the field of machine learning technique in prediction of heart disease. Later part of this survey paper will discuss about various machine learning algorithms and their relative comparison based on various performance metrics like F1 Score, specificity, accuracy etc.

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. These systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

The heart is very important part of human body. Which pumps blood into the entire body? If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

PROPOSED SYSTEM

The dataset comprises of 14 attributes in total, out of which 13 are predictor variables and one feature is a binary responsible variable. Therefore, the dataset represents binary classification problem.SVM was used for the classification process.

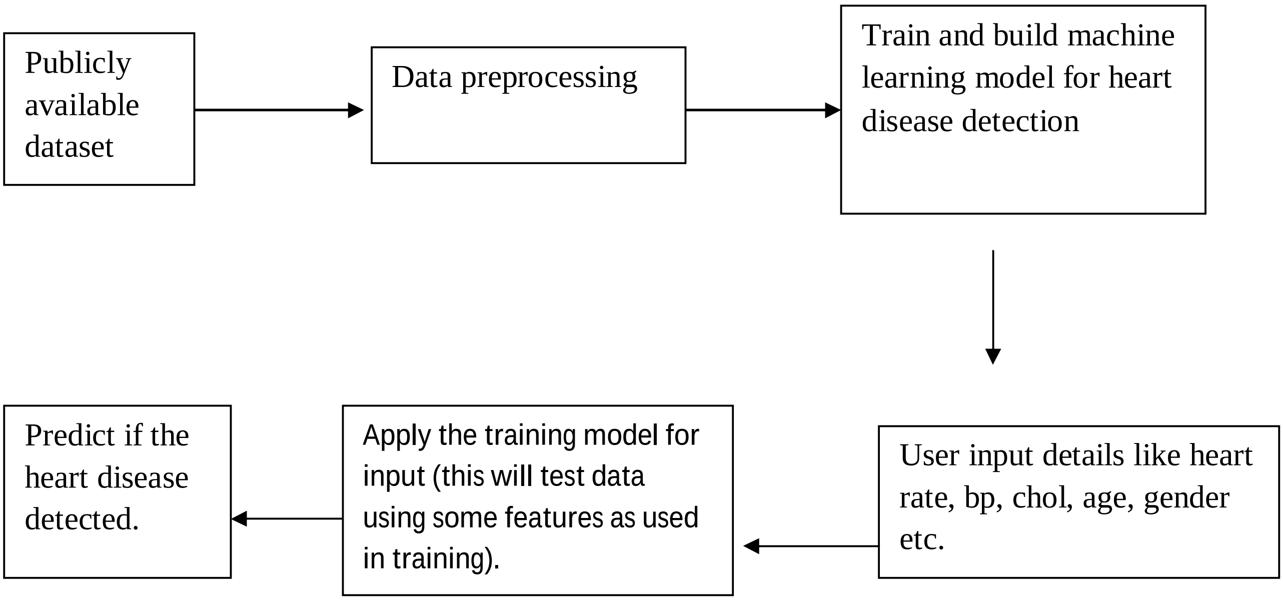
Publicly available dataset
Data preprocessing
Train and build machine learning model for heart disease detection
User input details like heart rate, bp, chol, age, gender etc.
Apply the training model for input (this will test data using some features as used in training).
Predict if the heart disease detected.

Figure 1: Block Diagram of Proposed System

The Figure 1 describes the block diagram of proposed system. We have built a model by training on publicly available dataset for heart disease detection. We have optimized the analysis on standard matrices (f1 score, accuracy etc). The optimized model will then be tested on the user data to predict if the user has a heart disease or not.

PROJECT PLAN AND IMPLEMENTATION

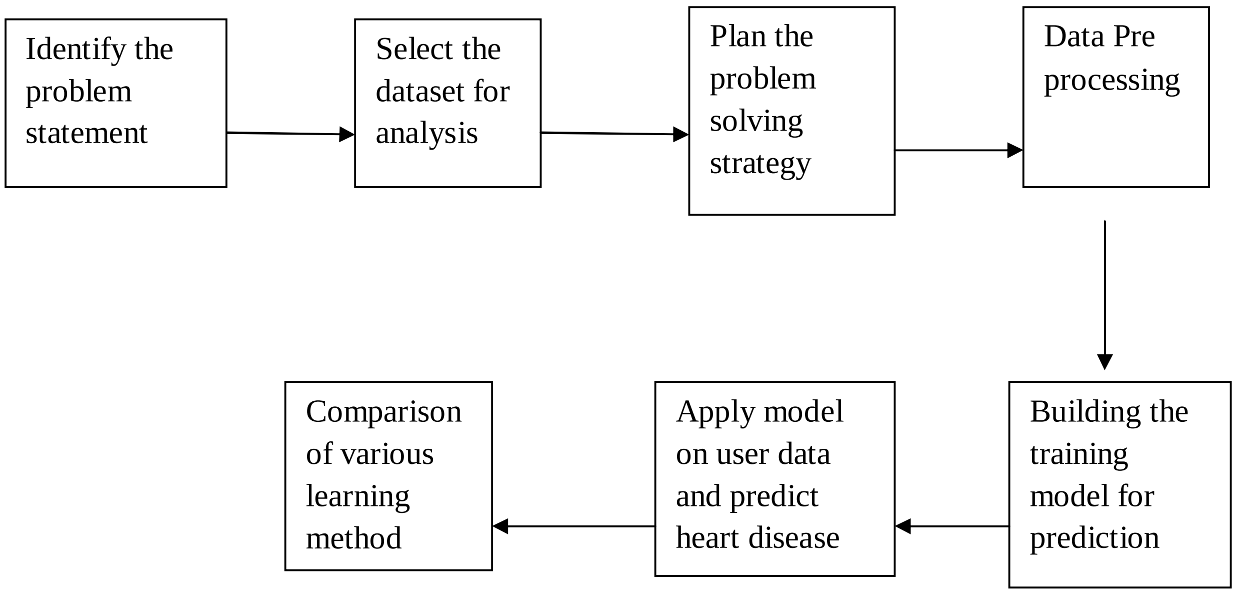
Identify the problem statement
Select the dataset for analysis
Plan the problem solving strategy
Data Pre processing
Building the training model for prediction
Apply model on user data and predict heart disease
Comparison of various learning method

Figure 2: Project Plan

Figure 2 shows the project plan where first step is to identify the problem statement, where the problem is identified weather the person has heart disease or not. The second step is to select the dataset for analysis. There are many datasets available in UCI machine learning repository such as Hungarian dataset, Switzerland dataset and Cleveland dataset. The third step is plan the problem solving strategy, for this we are using two classifiers one is SVM i.e. support vector machine and Random forest for comparison. The next step is Data preprocessing. Data preprocessing is a process of removing noisy and missing data from the data set. The next step is building a training model for prediction in this we will build our model by publicly available dataset such as sex, age etc. The next step is Apply the model on user data and predict heart disease. The last step is comparison of various learning methods based on classification results. In the last step it will compare both the classifiers SVM and Random forest for better choice.

METHODOLOGY

Problem Statement

Study the dataset (Cleveland dataset) and to predict whether a person has heart disease or not. If a person has a heart disease it is represented by 1 and if a person has no heart disease it is represented by 0.

Dataset

The heart disease dataset is a very well studied dataset by researchers in machine learning and is freely available at the UCI machine learning dataset repository https://archive.ics.uci.edu/ml/datasets/Heart+Disease.Though there are 4 datasets in this, we have used the Cleveland dataset. The dataset has 76 attributes and 303 records. However, only 14 attributes are used for this study & testing as shown in Table 1.

Table 1: Selected Heart Disease Attribute

Attributes	**Description**
Age	Age in years
Sex	0 = female 1 = male
Cp	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Resting blood pressure (in mm Hg)
Chol	Serum cholesterol in mg/dl
Fbs	Fasting blood sugar>120 mg/dl:1-true 0=False
Exang Continuous Maximum heart rate achieved	Exercise induced angina:1 = Yes 0 = No
Thalach	Maximum heart rate achieved
Old peak ST	Depression induced by exercise relative to rest
Slope	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	3 = normal 6 = fixed defect 7= reversible defect
Restecg	resting electrocardiographic results Diagnosis
Class	resting electrocardiographic results Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 1. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1) from absence (value 0).

Problem Solving Strategy

Data mining classification techniques for good decision making in the field of health care addressed are namely support vector machine, decision trees, Artificial Neural Networks and Naive Bayes. We are using two classifiers one is Support Vector Machine (SVM) and other one is Random forest for comparison.

SVM

The SVM is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory. SVM is method for classification of both linear and non-linear data. It uses a non-linear mapping to transform the original training data into a higher dimension. Within this new dimension it searches for linear optimal separating hyper plane. With an appropriate non linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. The SVM find this hyper plane using support vectors and margins. SVM performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors[4].

Random Forest

A random forest is a data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. This type of algorithm helps to enhance the ways that technologies analyze complex data. Random forest algorithm can use both for classification and the regression kind of problems. Random Forest is a supervised learning algorithm. It creates a forest to evaluate results. Random Forest builds multiple decision trees by picking ‘K’ number of data points point from the dataset and merges them together to get a more accurate and stable prediction..For each ‘K’ data points decision tree we have many predictions and then we take the average of all the predictions. Random forest is an Ensemble learning Algorithm. Ensemble learning is the process by which multiple models combine together to predict one result. Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Data Preprocessing

Data preprocessing is a process of removing noisy and missing data from the data set.

Data Pre-processing is one of the most important data mining task which includes preparation and transformation of data into suitable form to mining procedure. Data pre-processing aim is to reduce the data size, find the relation between data, Normalize data, remove outliers and extract features for data. It includes all techniques like data cleaning, integration, transformation and reduction[5].

Data set sample size. The main aim of analysis is to create multiple classifiers for predicting heart disease and evaluate and compare their performance. The dataset uses the “diagnosis of heart disease” as the class label and we are using 13 features.

We have used fairly large dataset with total sample size of 303 (total rows).

Out of 303, following is the sample distribution for each class-

164 samples belong to class 0 (Value 0: < 50% diameter narrowing)

139 samples belong to class 1 (Value 1: > 50% diameter narrowing).

Normalization

This method works by a adjusting the data values into a specific range such as between 0-1 or -1+1. This method is useful for mining techniques like classification, clustering and artificial neural networks[4].

- Min Max Normalization

Min Max Normalization - Min Max Normalization performs a linear alteration on the original data. The values are normalized within the given range. The computation given by:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where x is a current cell value and *min* and *max* are the minimum and maximum values in *x* given its range.

COMPARISON

It is the last step where the comparison is done between both the classifiers. We will comparesupport vector machine (SVM) and Random Forest for the better results.

RESULT AND DISCUSSION

The analysis of the two algorithms considered for the prediction of heart disease based on the accuracy. Based on the analysis, it clearly shows that the Support Vector Machine has higher accuracy compared to the Random Forest. We have performed binary classification, i.e. for 2 class labels (0 and 1).

We constructed classification models using-Support Vector machine (SVM) using RBF kernel and Random forest classifier. We divided our training and testing data using 60% for training and 40% for testing. “Accuracy” is used as a performance metric for the analysis and comparison of classification models.

Table 2: Comparative analysis of data mining technique

Algorithm	*Accuracy*
SVM	86.88%
Random Forest	81.11%

Using SVM, we got an overall accuracy of 86.88%.And by using Random forest accuracy is 81.11%.We can see that the performance of SVM is slightly higher than random forest. We got very high accuracy values using both the classifiers.

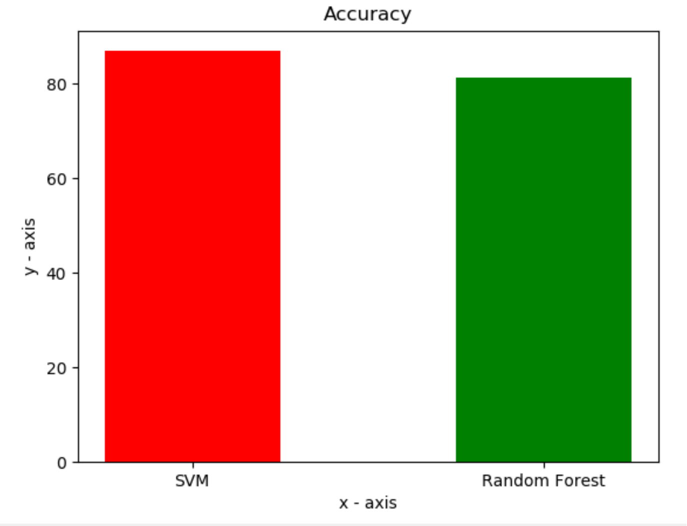

Figure 3: Accuracy of algorithm

Figure 3 shows the accuracy of various algorithms. SVM has the highest accuracy with 86.88% and random forest has the accuracy of 81.11%.

CONCLUSION

There are different data mining techniques that can be used for the identification and prediction of Heart disease among patient. In this paper there are two different algorithms are compared, which are used to predict heart disease. Form the comparison study; it is observed that the Support Vector Machine model turned out to be best classifier for Heart disease prediction. In future, the plan is to explore and merge more datasets so that will get more effective dataset which includes large diversity of population. Also plan to implement various other classification methods to effectively support prediction.

REFERENCE

1. M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset Selection," Int. Conf. Intel. Syst. Des. Appl. ISDA, 2012; 628–634.
2. V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," 2016; 38(3): 124–128.
3. Chollet, Francois. Deep learning with python. Manning Publications Co., 2017.
   S. Shylaja; R. Muralidharan "A novel method to predict heart disease using SVM Algorithm. ISSN: 2278-4853 June 2018; 7(6).
4. Saud A. Alasadi et al "Review of Data Pre-processing techniques in data mining", Journal of Engineering and applied science, ISSN: 1816949 X. 2017; 12(16): 4102-4107.