Comparative Analysis of Data Mining Classification Techniques for Cardiovascular Disease Prediction

Alpa Makvana
Department of Computer Engineering,
V.V.P. Engineering College, Rajkot
Gujarat, India

Devangi Kotak
Department of Computer Engineering,
V.V.P. Engineering College, Rajkot
Gujarat, India

Abstract:- Cardiovascular diseases are the main cause of death around the word. Every year, more people die because of these diseases than any other disease. Data mining techniques are widely used for the analysis of diseases, including cardiovascular conditions from person's health data. For such analysis so many techniques are available in data mining. In data mining some of the classification techniques are used to predict the heart diseases. This can be taken as evidence that the proposed method can be used assertively as decision making support to diagnose a patient with cardiovascular disease.

Keywords:- Neural networks, data mining, cardiovascular disease, support vector machine

1. INTRODUCTION:

Heart disease is one of the prevalent disease that can lead to reduce the lifespan of human begins nowadays. Each year 17.5 million people are dying due to heart disease [1]. Life is dependent on component functioning of heart, because heart is necessary part of our body. Heart disease is a disease that affects on the function of heart [2]. An estimate of a person's risk for coronary heart disease is important for many aspects of health promotion and clinical medicine. A risk prediction model may be obtained through multivariate regression analysis of a longitudinal study [3]. Due to digital technologies are rapidly growing, healthcare centers store huge amount of data in their database that is very complex and challenging to analysis. Data mining techniques play a vital role in analysis of different data in medical centers. Common attributes used for heart disease are Age, Sex, Fasting Blood Pressure, Chest Pain Type, Resting ECG (test that measures the electrical activity of the heart), Threst Blood Pressure(high blood pressure), Serum Cholestrol, Thalach (maximum heart rate achieved), ST depression (finding on an electrocardiogram), Fasting blood sugar, smoke, Hypertension, Food habits, weight, height, and obesity[4]. Table I summarizes the most common types of the heart disease are follows.

Table I Different types of heart disease [5]

Arrhythmia	The heart beat is improper whether it may irregular, too slow or too fast.
Cardiac arrest	An unexpected loss of heart function, consciousness and breathing occur suddenly.
Congestive heart failure	The heart does not pump blood as well as it should, it is the condition of chronic.
Congenital heart disease	The heart's abnormality which develops before birth.
Coronary artery disease	The heart's major blood vessels can damage or any disease occurs in the blood vessels.
High blood pressure	It has a condition that the force of the blood against the artery walls is too high.
Peripheral artery disease	The narrowed blood vessels which reduce flow of blood in the limbs, is the circulatory condition.
Stroke	Interruption of blood supply occur damage to the brain.

Figure 1 depicts the parts of human heart such as Left atrium, Right atrium, Right ventricle, Left ventricle, Aorta, pulmonary vein, Pulmonary valve, Pulmonary artery, Tricuspid valve, Aortic valve, Mitral valve, Superior vena cava and Interior vena cava.
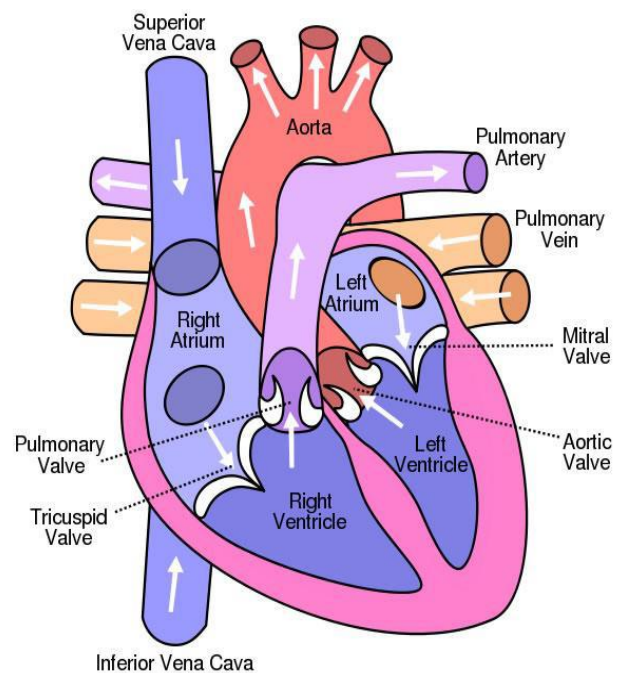

Figure 1 Human Heart [6]

2. LITERATURE REVIEW

There are numerous works has been done related to disease prediction system using different data mining techniques.

Mohammad Shafenoor Amin et al, [7] This research aims to identify significant features and data mining techniques to predict heart disease. Proposed method uses 2 classifiers Naïve bayes and Logistic Regression. Number of features used for heart disease prediction is 9. Accuracy for vote is 87.41%. This research can be extended by conducting the same experiment an a large scale real world dataset
.
Sarath Babu et al, [8] The main focus of this paper is on using algorithms in data mining and sequence of several attributes for effective heart disease prediction. Proposed heart disease prediction using genetic algorithm, k- means algorithm, MAFIA algorithm, Decision algorithm. After applying genetic algorithm, decision tree has tremendous efficiency.

Meenal saini et al, [9] Survey on some recent techniques used to predict heart disease risk. 9 classifiers used for hybrid system. Develop a hybrid approach, Hybrid classifier with weighted voting(HCWV). HCWV gives accuracy is 82.54%.

Purushottam et al, [10] The main study to help the non-specialized doctors to make correct decision about heart disease level. KEEL tool is used for implementation. Classification Decision rule generated for classification. They gives the accuracy is 86.7%.

Rishabh wadhawan et al, [11] In this paper used framework implementation in visual studio c# to develop a system prototype which help determine and extract knowledge from dataset. Result for precision is 0.78, Recall is 0.67 and accuracy is 74%. To design a framework using heart disease patient prescription and utilizing web mining and data warehouse technology which can be extremely useful for effective and precise prediction system.

Bandarage shehsni et al, [12] In this paper comparative study of classification techniques for heart disease prediction. Proposed method uses Algorithm is Naïve bayes, neural network, decision tree. Accuracy of naïve bayes is 86.5%. Accuracy of neural network is 89%. Accuracy of decision tree is 85.5%. To analyze the reason of heart disease and investigating a trainable combining method such as Bayesian combiner.

D. Karthick et al, [13] The system development for predicting the chances of occurances of cardiovascular disease. Software R studio is used. Naïve bayes and random forest algorithm use in this proposed system. The output is a binary classification which will give 1 or 0; 1 being the chances to occur and 0 being the chances not to occur. The model uses the patients data having age within fifty years; so it is helpful for predicting the heart disease and a patient can cure enough at the young age.

Marjia et al, [14] developed heart disease prediction using k star, j48, SMO, and bayes net and multilayer perception using weka software. Based on performance from different factor SMO and Bayes net achieve optimum performance than KStar, Multilayer perception and J48 techniques using k-fold cross validation. The accuracy performances achived by those algorithm are still not satisfactory. Therefore, the accuracy's performance is improved more to give better decision to diagnosis disease. Few research work in the domain of interest are summarized in Table II.

Table II Comparative analysis of Literature Review

Purpose	Technique Used	Accuracy
To select significant features and classifier techniques for heart disease prediction	Naïve Bayes and Logistic Regression as a hybrid technique names vote	Very good accuracy by 87.41%
Early detection of heart disease and its diagnosis correctly on time and providing treatment with affordable cost.	Genetic algoritm K means clustering Mafia algorithm Decision tree	Better result by using decision tree after k means
Accurate and timely diagnosis of heart disease	Combination of 9 classifier are used	HCWV gets 82.54% accuracy
Make efficient use of medical data and generate a heart disease prediction system	Decision rule making algorithm	Compared with other classifiers, EHDPS gives better accuracy by 86.7%.
Generate a system for identifying coronary illness of patient	K means clustering Apriori algorithm	Accuracy is 74%
To find a best classifier which give better accuracy of disease prediction to the patient	Naïve bayes Neural network Decision tree	Accuracy decrease by increasing no. of attributes.
Detect heart disease risk at early stage for young people	Svm Naïve bayes	Binary classification result 1 high risk 0 low risk

3. PROPOSED WORK

Inspired by the growing rate of patient's death owing to heart disease each year, there is an increasing availability of patient's data which can help experts to extract important information by data mining techniques. This important information can help human experts to cure the heart disease. Moreover, it can facilitate the design of a model that can help the hospital management to encourage and give advice to the experts related to the diagnosis and proper treatment given to the patients having heart disease. This paper describes some standard

Classifier for disease prediction. In the proposed system early diagnosis of the heart disease is carried using the data mining techniques. The proposed framework is shown in figure 2.
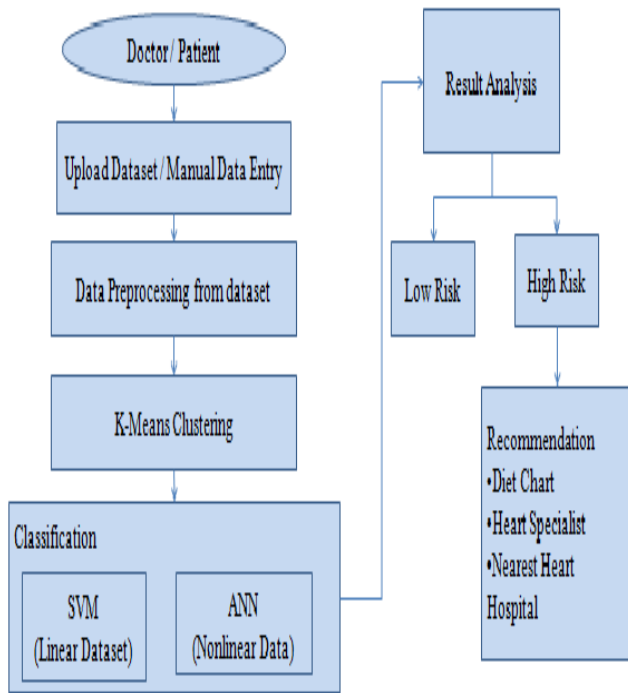

Figure 2 Blocked diagram of Proposed work

4. DATASET:

The heart disease dataset from [15] has been utilized for training and testing purpose. It consists of 76 attributes; However, only 14 of them have been used because for remaining attributes, the values were missing. We obtain accurate result with reduced number of features. Additionally processing is done by replacing the missing values of the attribute (column) by the column's arithmetic mean, in case of nominal data it is replaced with the mode.

Table III list the chosen attributes of the heart disease dataset. The performance of all the classifiers is accessed and their outcomes are then analyzed on the basis of accuracy. Some researchers, however, have used the Cleveland, Hungarian and long-bench-via Switzerland dataset consisting of 14 attributes, which along with the values and their possible data types are described in table III.

Table III Data Set Description

Attribute Name	Discription	Values
age	Age	In year
sex	Sex	1=Male, 2=Female
chest	Chest pain type	1,2,3,4
Resting_blood_pressure	Resting blood pressure	1=high, 0=normal
Serum_cholestoral	Serum cholestoral in mg/dl	Serum Chol in mg/dl
Fasting_blood_sugar	Fasting blood sugar>120 mg/dl	72 to 99 mg/dl
Resting_electrocardio Graphic_results	Resting electrocardiographic Results(0,1,2)	60 to 100 bpm

Maximum_heart Rate_achieved	Maximum heart rate achieved	150 to 200 bpm
Exercise_induced_ Angina	Exercise induced Angina	Up to 225 mm/hg
Oldpeak	Oldpeak=ST Depression induced by exercise relative to Rest	ST depression Induced by Exercise relative to rest
Slope	The slope of the peak Exercise ST segment	1=unslop, 2=flat, 3=downslop
Number_of_major_ Vessels	Number of major vessels(0-3) colored by flourosopt	0-3 colored by flourosopy
Thal	3=normal; 6=fixed defect; 7=reversible defect	3=normal, 6=fixed, 7=reverse
Class	class	0=no risk, 1= heart disease

5. RESEARCH METHODOLOGY

This segment summarized various methodologies in data mining, which are used in the diagnosis of heart disease.

A. Support Vector Machine

Support vector machine algorithm make good judgement for data points that are outside the preparing set. There are two classes of information in SVM. The data points are separated such that they could draw a horizontal line on the figure. The line is made in a way that it isolates every one of the focuses on one side of alternate class. When such circumstances happens. Then the data are linearly separable. The line used to isolate the dataset is known as a separating hyperplane. The points nearest to the isolating hyperplane are called as support vectors. Kernels are utilized to extend SVMs to a bigger number of datasets. Mapping of one feature space to another is finished by kernel. Kernel method, maps the information(in some cases likewise called as nonlinear information) from a little dimensional space to an extensive dimensional space. In a bigger measurement, it decides straight issue that is non liner in smaller dimensional space. The radial bias function(RBF) is a prominent kernel that measures the separation among two vectors.

B. Artificial Neural Network

Artificial neural network(ANN) is a mathematical structure in view of biological neural networks. Artifical neural network is depends on perception of a human brain. Human brain is extremely web of neurons. Analogically artificial neural network ia arrangement of three simple units specifically input, hidden and output unit. The parameters that are passed as input to the following structure a first layer. In medical finding patients hazard factors are treated as input to the neural network.

6. CONCLUSION AND FUTURE WORK

By using different types of data mining techniques to predict the occurrence of heart disease have summarized. Determine the prediction performance of each algorithm and apply the proposed system for the area it needed. Use more relevant feature selection methods to improve the accurate

performance of algorithms. There are several treatment methods for patient, if they once diagnosed with the particular form of heart disease. Data mining can be of very knowledge from such suitable dataset.

The various heart disease prediction techniques are analyzed in this paper. The data mining techniques used to predict heart disease are discussed here. Heart disease is a mortal disease by its nature. This disease makes several problems such as heart attacks and death in long run. By using SVM and ANN, the aims is to get the better accuracy. The future work is to provide the recommendation of prevention for heart disease risk. Recommendation system gives proper diet chart, heart specialist doctors list, heart hospitals details.

7. REFERNCES:

[1] Animesh Hazara,Arkomita Mukherjee,Amit Gupta,Asmita Mukherjee, "Heart Disease Diagnosis and prediction using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, July 2017, pp.2137-2159

[2] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications, February 2016

[3] Guizhou Hu, Martin M. Root, "Building Prediction Models for Coronary Heart Disease by Synthesizing Multiple Longitudinal Research Findings", European Science of Cardiology, 10 May 2005

[4] T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Heart Disease Prediction Model using SVM- Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications, vol. 68, 16 April 2013

[5] http://www.medicalnewstoday.com/articles/257484.Php

[6] NImai Chand Das Adhikari, Arpana Alka, and rajat Garg, "HPPS: Heart Problem Prediction System using Machine Learning"

[7] Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan, "Identification of Significant Features and data mining techniques in predicting heart disease" 2018 ELSEVIER, telematics and Informatics

[8] Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, "Heart Disease Diagnosis Using Data Mining Technique" 2017 IEEE International Conference on Electronics, Communication and Aerospace Technology(ICECA)

[9] Meenal Saini, Niyati Baliyan, Vineeta Bassi, "Prediction of Heart Disease Severity with Hybrid Data Mining" International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)-IEEE2017

[10] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System" 2016 ELSEVIER Procedia Computer Science

[11] Rishabh Wadhawan, "Prediction of Coronory Heart Disease Using Apriori Algorithm With Data Mining Classification"

[12] Bandarage Shehani Sankheta Rathnayake,Gamage Upeksha Ganegoda, "Heart disease prediction with Data Mining and Neural Network Techniques" 2017-IEEE 4th International Conference For Convergence in Technology

[13] D.Karthick, B.Priyadarshini, "predicting the chances of occurrence of Cardio Vascular Disease(CVD) in people using Classification Techniques within fifty years of age" Proceedings of the Second International Conference on Inventive Systems and Control(ICISC 2018)-IEEE

[14] Marjia Sultana, Afrin Haider, "Heart Disease Prediction Using WEKA tool and 10-Fold cross-validation", The Institute of Electrical and Electronics Engineers, March 2017.

[15] UCI repository accessed on 20 March 2017, available online at http://archive.ics.uci.edu/ml/datasets/Heart+Disease.