Review on Prediction of Heart Disease Using Data Mining

Ayushi Vaishya [1], Anjali Mehta [2], Kavita Joshi [3]

[1,2,3]*Department of CSE, Devbhoomi Instotute of Technology, Dehradun, Uttarakhand, India*

([1]ayushimalaiwa99@gmail.com, [2] anshuanjali343@gmail.com, [3]kjkavita86@gmail.com*)*

***Abstract*—**We are living in an era of advanced technologies along with a highly expandable database which is being increased day-to-day. Living in an informative age, we are in a dearth of physical well-being. The unhealthy lifestyle is making us prone to several heart- related diseases. Most of the cases, this disease is acknowledged at the last stage. Data Mining is a process of identifying the hidden patterns from the available data set and converting them to information. There are various data mining algorithms such as Regression, Classification, and Association which are used to predict the likelihood of suffering from heart diseases. The main aim of this paper is to encapsulate previous research papers based on prediction of heart disease using various data mining tools and techniques. We have analyzed the results of these research papers to make the prediction more efficient at early stage.

Keywords*—Heart disease, Data Mining, Data Mining techniques, Data Mining tools*

I. Introduction

Heart Disease is one of the major causes of death in today's world. It has been estimated that about 17.9 million people globally died due to heart disease in 2016. Out of 31% of global death, 85% are due to heart disease. In 2015 out of total 17 million under age death of 70 (premature death), 37% are due to cardio vascular disease. According to WHO, by 2030, 2.63 million people of the world will be died due to heart diseases [12]. Due to increasing the number of people suffering from heart diseases world-wide it has been become very necessary to easily predict the disease so that it can be prevented at the very early stages.

Data mining is one of the best processes of analyzing the hidden data patterns inorder to get useful information which is then collected and assembled in data warehouses, for data mining algorithms, efficient analysis, business decision making inorder to cut cost and increase gain.

Different types of data mining techniques such as clustering, Classification Algorithms like Decision Tree, Naïve Bayes, Artificial Neural Network etc, Association Rule Mining are used for the prediction of various kind of disease such as heart disease, lung cancer, breast cancer etc. HPDS is the user- friendly system that helps the doctor to predict heart disease status based on the data of the patients. There is numerous numbers of data mining tools and algorithms, which helps to predict the disease at the early stage. The main objective of our paper is to provide brief introduction about the algorithms and tools which can be used to predict the heart diseases. The information collected can be used by medical analyst to diagnosis the disease.

A. Heart Disease

The heart is the muscular organ of the body, which pumps blood through the blood vessel of the body. Blood helps to provide the body with oxygen and nutrients and it also helps to remove the metabolic wastes from the body. But due to various reasons our heart starts to get weaker and weaker and it stop functioning well. Due to insufficient circulation of blood in our body various organ like brain suffers and if it stop completely to function, end occurs within few seconds. Heart disease is one of the crucial reasons for patients in countries like United States, India.

There are numerous numbers of risk factors which result in the failure of the heart [13]:

a) *Age:* The older you get the higher chance of risk.

b) *Gender:* Than female male are at higher risk level until the age of 75.

c) *Obesity:* According to study published in the journal JAMA Cardiology in April 2018, concluded the adults between ages 40 to 59 who are obese are likely to have a risk of heart disease.

d) *High Blood Pressure:* It leads to the narrowing and blocking of blood vessels which increases the risk of heart failure.

e) *Smoking:* Smoking leads to damage the lining of our arteries, which leads to build of atheroma (fatty material) and can cause a heart attack, stroke and angina. The carbon monoxide (CO) present in tobacco smoke reduces the amount of oxygen from the body.

f) *Family History of Heart Disease:* Genes can increase the risk of cardiovascular disease, and they can be responsible for increasing high blood pressure and cholesterol in the body.

g) *Hyper Tension:* High Blood Pressure result to hyper tension and vice versa.

h) *Physical Inactivity:* Less active and less fit person have almost 30-50% chance of high blood pressure. It is a risk factor for cardiovascular disease itself.

i) *Poor Diet:* High-salt consumption will increase blood pressure and hence increase high chance of heart attack and failure.

j) *High Cholesterol:* Diet with low cholesterol, simple sugars, regular exercise, medications etc will help to lower the level of cholesterol and reduce the chance of heart disease.

k) *Diabetes:* With healthy diet, medications, regular exercise, walking etc will reduce the chance of diabetes and reduce the risk of heart failure.

l) *Poor Hygiene:* Due to poor hygiene various kinds of bacteria can easily inject to body and thus result in the chance of heart failure.

II. Data mining tools

A. Weka

WEKA (Waikato Environment for Knowledge Analysis) is open source software written in Java and developed at University of Waikato, New Zealand. It is a collection of machine learning algorithms to perform data mining tasks. It contains tools for data classification, preparation, clustering, regression, visualization and association rules mining.

B. RapidMiner

It is a platform for data science software developed by RapidMiner. It provides integrated environment for machine learning, data preparation, deep learning, predictive analysis and text mining. It is developed on an open core model.

III. Data Mining techniques

A. Decision Tree

Decision Tree classifier divides the provided dataset into smaller subsets. It generalizes and classifies a given data set. The value of target variable is predicted on the basis of input variables. Leaf nodes represent the value of target variable.

Interior nodes represent the input data. The edges represent the possible values of input variables.

B. Naïve Bayes Classifier

Naïve – Bayes algorithm is a classifier in machine learning and data mining. It is probability based conditional theorem. The variables in the training data set are independent of each other.

C. K-Nearest Neighbour (KNN) Algorithm

KNN algorithm is a classifier algorithm which is non-parametric. No assumption is made on the underlying data distribution. Most of the training data is used in testing phase. There is no need of prior knowledge about the distribution.

D. Artificial Neural Network (ANN)

ANN is a machine learning algorithm that works as a human brain does. This network contains artificial neurons which are interconnected nodes. It consists of three layers. The first layer is the input layer which consists of input data sets. The second layer is hidden layer and third layer is output layer.

E. Clustering

Clustering technique is the grouping of similar objects or datasets in same class and aggregating the data for desired information analysis. In this, the clusters in the data are discovered such that if two objects belong to same cluster then their degree of association is highest and if they are in different clusters then the degree of association is lowest.

F. Regression

The relationship among variables is identified. The properties of dependent variable changes if any of the independent variable is varied.

G. Support Vector Machine (SVM)

SVM technique is used for classification and regression problems. SVM analyzes the input dataset and separates the data into various classes. This classification helps in analyzing any new data.

H. Apriori Algorithm

In the algorithm, various association rules of if-then format are produced on the basis of given dataset. According to the rules, if frequency of occurrence of a dataset is high then the frequency of occurrence of its sub –datasets is also high.

I. Random Forest

Random forest is a collection of decision trees and they are merged together in order to get more stable and accurate result. In decision tree we have a collection of dataset and on the base of certain conditions we split the node in order to reduce entropy and to increase information gain hence, accuracy. In random decision tree, hyper parameter is used either to increase the speed of model or to increase predictive power of the model. It removes the problem of over fitting in machine learning as it consists of large number of trees.

IV. Literature REview

As the age passes, different types of researches have been done on the prediction of heart diseases by using different types of machine learning algorithms and data mining techniques.

M. Anbarasi [1] et al. proposed enhanced prediction of heart disease with feature subset selection using genetic algorithm. The main aim was to predict the likelihood of a patient suffering from heart disease with reduced number of attributes so that the patient has to take less number of tests. The dataset consists of 909 records. 13 attributes were provided for the experiment. Genetic search is used to increase the accuracy of prediction. The dataset for genetic search is provided with initial zero attributes which were to be examined on an initial population with randomly generated rules.

The initial population is evolved until the rules are satisfied. With the help of genetic search the attributes were reduced to six. Naive Bayes algorithm, classification by clustering and Decision Tree algorithm were used as classifiers having input dataset of six attributes. The accuracy of Naïve Bayes, Decision Tree and Classification via clustering was 96.5%, 99.2% and 88.3% respectively.

T.John Peter and K.Somasundaram [2] (2012), presented the use of data mining and pattern recognition techniques inorder to predict risk models in the clinical domain of cardiovascular medicine. There were the limitations in the conventional medical scoring system which was handled by classification models which can internally detect possible interactions between predicator variables as well as nonlinear complex relationship between independent and dependent variables.

The dataset consists of huge number of data which consumes high classification time hence the researcher used the attribute selection methods inorder to reduce the data size. The researcher used different classification algorithms like Naïve Bayes, Decision Tree, K-Nearest Neighbour and Neural Network inorder to find the best accuracy on the reduced data set. Naïve Bayes gave the best accuracy 83.70% for the prediction by using CFS attribute selection method.

Chaitrali S. Dangare [3] et al. proposed a heart disease prediction system using data mining techniques in which two more attributes, i.e. obesity and smoking, other than the 13 attributes, were used to predict the likelihood of patient suffering from heart disease. The data was collected from the Cleveland heart disease database and Statlog heart disease database.

The data mining techniques used were Decision Trees, Naïve Bayes and Artificial Neural Networks and the results from these techniques were analyzed. Weka 3.6.6 tool was used for data mining. Total numbers of 573 records were collected for the experiment, which was divided into two datasets. The training dataset consists of 303 records and the testing dataset consists of 200 records. The accuracy of Neural networks, Decision Tree and Naïve Bayes came out to be 100%, 99.62% and 90.74% respectively. The result shows that prediction of heart disease with Neural Networks technique has the highest accuracy.

T.Revathi and S.Jeevitha [4] proposed a comparative study on heart disease prediction system using data mining techniques. 14 parameters were used to perform the experiment for prediction of heart disease. The data mining approaches used were Back-propagation network, Naïve Bayes algorithm and Decision Tree algorithm. The record was collected from Cleveland database.

Out of the 76 attributes present in the database, 14 were chosen for the analysis. The accuracy from back-propagation network, Naïve Bayes algorithm and Decision Tree algorithm came out to be 100%, 90.74% and 99.62%.

Walid Moudani [5] proposed a research on dynamic feature selections for heart disease classification. The main aim was to predict Coronary Heart Disease (CHD) which is the major cause of heart attacks by using data mining and Random Forest technique. Dynamic programming is used to generate the subsets of reduced features dynamically by using rough subsets technique. Random Forest Decision tree classifier is used to check the risky state of heart disease. 512 adults' data was collected in this system. This system has contributed in providing the CHD risk.

K.Gomathi and Dr.Shanmugapriyaa [6] (2016), presented an analysis on heart disease on the male patient by using data mining techniques like J48 decision tree, ANN (Artificial Neural Network) and Naïve Bayes in order to analyze the dataset which is based on the attributes of diseases of heart. All the available 8 fields from the database are presented in the preprocessed data set which consists of 210 records.

The data mining tool used by them was WEKA (Waikato Environment for knowledge Analysis) which is written in Java. The goal is to achieve high accuracy, beside high precision and recall metrics. The performance of Naïve Bayes was more accurate with 79.9043% in 0.01 Seconds than ANN (Artificial Neural Network) with 76.555% in 1.55 Seconds and J48 with 77.0335% in 0.01 Seconds.

Jagdeep Singh, Amit Kamra and Harbhag Singh [7] (2016), developed a framework for early detection of heart disease by using techniques of associative classification. Different data mining techniques such as Naïve Bayes, ZeroR, J48, k-nearest neighbor and J48 along with association algorithm such as FP-Growth and Aprior are used for prediction of heart disease.

The main goal of this research was to introduce a method that can produce CARs (Classification Association Rules) efficiently and to measure which method can give the highest percentage for prediction of early heart diseases. Dataset from University of California Irvine (UCI) machine learning repository is used to test on different techniques of data mining. The prediction accuracy of 99.19% is obtained by using classification associative rules (CARs) of hybrid technique.

Theresa Princy and J.Thomas [8] (2016), presented the survey on different classification techniques for predicting the risk level based on age, gender, pulse rate etc for each individual. The data mining techniques such as Naïve Bayes, KNN, Artificial Neural Network etc along with classifiers are used and found that using more number of attributes result gives high accuracy of risk level.

By using above techniques the patient record is predicted and classified continuously, if any kind of changes occurs, the patient and doctor will be informed immediately and at the early stage the doctor can immediately diagnosis the heart disease. By using KNN and ID3 algorithm the prediction of heart disease was done and for different number of attributes accuracy level is provided.

Sushmita Manikandan [9] (2017), presented prototype implementation of a heart disease prediction system inorder to save time and efforts of doctor by automating predicting the risk by using binary classifier web based graphical user interface and Naïve Bayes algorithm is used to build binary classifier by using Anaconda v2.7 packages and Rapid Miner is used for cleansing the dataset which is collected from UCI's Machine Learning Repository. The final datasets consist of 13 predictor variables and only one response variable named num. If num=0, the prediction is 'low risk' else 'high risk'. Naïve Bayes gave the best accuracy of 81.25%.

AH Chen [10] et al. developed a system which can predict the heart disease and can assist to medical professionals to predict heart disease by using clinical data of patients. The researchers have used data from UCI machine learning repository. It contained about 303 instances out of which 139 belonged to heart disease. For each instance 14 clinical features have been recorded. It involves 3 approaches:

Firstly, they will select 13 important clinical feature of patient such as age, sex, pulse rate, cholesterol, chest pain, blood pressure, fasting blood sugar, old peak, number of vessels colored, exercise induced angina, resting ecg and thal.

Secondly, they will classify heart disease based on clinical feature by using an Artificial Neural Network (ANN).Used C as a tool inorder to implement disease prediction and classification. The prediction of accuracy is near to 80%.

Lastly, they develop a user-friendly HDPS (Heart Disease Predict System) developed from C# and C environment. HPDS convert code into C and integrated result into C# interfaces. HPDS system consisted of various features, such as ROC curve display section which displays the ROC curve of the HPDS, input clinical data section which uses 13 pieces as input of clinical data and prediction performance display section which displays the performance of HPDS including sensitivity, specificity, accuracy and prediction result.

Used Dataset from UCI, they collected total of 303 instances, 164 belonged to healthy and remaining 139 belonged to heart disease.

Uma K. and Dr. M. Hanumanthappa [11] proposed a research on feature selection for heart disease prediction with data mining techniques. The data was collected from University of California, Irvin (UCI) and Machine Learning Repository which consists of 689 records with 18 attributes.

The ReplaceMissingValues filter of Weka tool is used for data pre-processing. This research was divided into two sets of experiments. In first experiment, the pre-processed data was provided as input to five classification techniques: Support Vector Machine(SVM), Bagging, Naïve- Bayes algorithm, Regression technique and J48 Decision Tree which results in the following accuracy : 99.7%, 91.7%, 92.7%, 99.7% and 92% respectively.

This classified the presence and absence data of heart disease. As a result, prediction through SVM and regression is more efficient. In the second experiment, the prediction is done using feature selection method. The five classification techniques were provided with many attribute selection methods, namely, CfsSubsetEval, Information Gain, Gain Ratio, Correlation and Wrapper methods. The CfsSubsetEval results to the lowest accuracy.

TABLE I. COMPARISION TABLE OF VARIOUS ALGORITHMS IN LITERATURE REVIEW

Author	Purpose	Technique Used	Accuracy
M. Anbarasi et. al.	Enhanced prediction of heart disease with feature subset selection using genetic algorithm	1)Naïve Bayes 2)Decision Tree 3)Classification via clustering	1)96.5% 2)99.2% 3)88.3%
T.John Peter and K.Somasundaram	Heart Disease by using Classification Data Mining Techniques.	1)Naïve Bayes 2)Decision Tree 3)K-NN 4)Neural Network	1)83.70% 2)76.66% 3)75.18% 4)78.148%
Chaitrali S. Dangare et. al.	Improved study of prediction heart disease system using data mining classification techniques	1)Decision Trees 2)Naïve Bayes 3)Neural Networks	1)99.62% 2)90.74% 3)100%
T. Revathi et. al.	Comparative study on heart disease prediction system using data mining techniques	1)Back-propagation network 2)Decision Tree 3)Naïve Bayes	1)100% 2)99.62% 3)90.74%
Walid Moudani	Dynamic feature selections for heart disease classification	Dynamic Programming, Random forest technique	
K.Gomathi and Dr.Shanmugapriyaa	Prediction Of Heart Disease using Data Mining	1)Naïve Bayes, 2)ANN and 3)J48	79.9043% 76.555% 77.0335%
Jagdeep Singh, Amit Kamra and Harbhag Singh	Prediction Of Heart Disease Using Associative Classification	Naïve Bayes, ZeroR, J48, k -nearest neighbor and J48 along with association algorithm such as FP-Growth and Aprior	99.19% using K-Nearest Neighbour with Aprior Associative Algorithm
Theresa Princy and J.Thomas	Human Heart Disease Prediction System using Data Mining Techniques	Naïve Bayes, KNN, Neural Network etc with classifiers	more number of attributes result in high accuracy of risk level
Sushmita Manikandan	Heart Attack Prediction System	Naïve Bayes	81.25%

AH Chen et al.	HPDS: Heart Disease Prediction System	Developed a system which can predict the heart disease and can assist to medical professionals to predict heart disease by using clinical data of patients.	The accuracy of classification is near to 80%with 85% sensitivity And 70% specificity.
Uma K. et. al.	Feature selection for heart disease prediction with data mining techniques	1) SVM 2) Bagging 3)Naïve Bayes 4)Regression 5)J48 Decision tree	1)99.7% 2)91.7% 3)92.7% 4)99.7% 5)92%

V. Conclusion

Our main goal is to predict heart diseases at the very early stage by using different data mining techniques, algorithms and tools so that the doctor and the patient can be made aware about the problem. Different kinds of techniques have been used by number of researchers' inorder to find best technique which can easily predict the heart disease at low cost and in less time. The survey is conducted from 2010-2017 which gives us the idea about the available techniques and methods. It is concluded that the Artificial Neural Network and the back propagation network gives the highest efficiency, i.e. 100% accuracy to predict the heart disease. The future scope in this field can be the reduction of attributes that predicts the occurrence of heart disease at very early stage which can subsequently reduce the number of cost, time and test. This will be beneficial for the low-income and middle-income countries.

References

[1] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376.

[2] Manikandan, S. (2017, August). Heart attack prediction system. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 817-820). IEEE.

[3] Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

[4] Revathi, T., & Jeevitha, S. (2013). Comparative Study on Heart Disease Prediction System Using Data Mining Techniques. International Journal of Science and Research (IJSR) ISSN (Online), 2319-7064.

[5] Moudani, W. (2013). Dynamic features selection for heart disease classification. World Academy of Science, Engineering and Technology, 7.

[6] Gomathi, K., & Priyaa, D. D. S. (2016). Multi Disease Prediction using Data Mining Techniques. International Journal of System and Software Engineering, 12-14.

[7] Singh, J., Kamra, A., & Singh, H. (2016, October). Prediction of heart diseases using associative classification. In 2016 5th International Conference on Wireless Networks and Embedded Systems (WECON) (pp. 1-7). IEEE.