International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 | Aug-2016 www.irjet.net p-ISSN: 2395-0072

Heart Disease Prediction System using Data Mining Techniques: A study

Syed Immamul Ansarullah[1], Pradeep Kumar Sharma[2], Abdul Wahid[3], Mudasir M Kirmani[4]

[1]Research Scholar, School of CS & IT, MANUU, Hyderabad, India, syedansr@gmail.com
[2] Assoc. Professor, School of CS & IT, MANUU, Hyderabad, India, drpkumar1402@gmail.com
[3] Professor, School of CS & IT,MANUU, Hyderabad, India, hod.csit@manuu.ac.in,
[4]Asstt. Prof., SKUAST-K, J&K, India, mmkirmani@gmail.com

Abstract - *Data Mining is the process of non-trivial extraction of implicit, previously unknown and potentially useful information from data. A pattern is interesting if it is valid for a given test data with some degree of certainty, novel, potentially useful and easily understood by humans. The huge amount of data generated for prediction of heart disease is too complex and voluminous to be processed and analyzed by traditional methods. Advanced Data Mining tools overcome this problem by discovering hidden patterns and useful information from complex and voluminous data. Researchers reviewed literature on prediction of heart disease using data mining techniques and reported that Neural Network technique overcome all other techniques with higher levels of accuracy[5][6]. Applying Data Mining techniques on healthcare data can help in predicting the likelihood of patients getting heart disease. This paper highlights the important role played by data mining tools in analyzing huge volumes of healthcare related data in prediction and diagnosis of disease.*

Keywords: Heart disease, Data mining, Data mining techniques, Neural Networks, Decision trees,

1. INTRODUCTION

Data mining is the process of finding previously unknown patterns and hidden information from healthcare datasets. Data mining combines statistical analysis, machine learning algorithms and database technology to extract hidden patterns and relationships from large databases [13]. Nowadays, data mining is becoming popular in healthcare domain as there is need of efficient analytical methodology for detecting unknown and valuable information in healthcare domain. Cardiovascular disease (also called heart disease) is a class of diseases that involve the heart or blood vessels (arteries, capillaries, and veins). Cardiovascular disease is the leading cause of deaths worldwide, since 1970s the cardiovascular mortality rates have declined in many high-income countries. At the same time, cardiovascular deaths and disease have increased at a fast rate in low and middle income group countries [14]. Although cardiovascular disease usually affects older adults, the antecedents of cardiovascular disease, notably atherosclerosis; begins at early stages of life making primary prevention efforts necessary from childhood. Therefore, increased emphasis on preventing atherosclerosis by modifying risk factors, evidence suggests a number of risk factors for heart disease such as age, gender, high blood pressure, high serum cholesterol levels, smoking, excessive alcohol consumption, sugar consumption, family history, obesity, lack of physical activity, psychosocial factors, diabetes mellitus, air pollution and using tobacco. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure – a condition that causes around half of the deaths from stroke and heart disease. Heart disease is the major cause of casualties in the different countries including India. Heart disease kills one person in every 34 seconds in the United States [15]. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on a doctor's experience and knowledge. This leads to unwanted results and excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial. This research work is an attempt to present the detailed study about the different data mining techniques which can be deployed in these automated systems.

2. METHODOLOGY

This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate heart disease diagnosis. The main methodology used for this paper was survey of journals and publications in the fields of medicine, computer science

and engineering with specific reference to applications of computer science in healthcare sector.

3. LITERATURE REVIEW

This paper aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction. Different data mining techniques have been used in the diagnosis of CVD over different Heart disease datasets. Some papers use only one technique for diagnosis of heart disease and other researchers use more than one data mining techniques for the diagnosis of heart disease.

- **Jyoti Sonia, et.al.** [6] in year 2011 presented three classifiers Decision Tree, Naïve Bayes and Classification via clustering to diagnose the presence of heart disease in patients. Classification via clustering: Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. Experiments were conducted with WEKA 3.6.0 tool. Data set of 909 records with 13 different attributes. All attributes were made categorical and inconsistencies were resolved for simplicity. To enhance the prediction of classifiers, genetic search was incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time.

- **K. Srinivas et al**. [ ] presented application of Data Mining Technique in Healthcare and Prediction of Heart Attacks. The potential use of classification based data mining techniques such as rule based Decision Tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consisted of 3000 instances with 14 different attributes. The instances in the dataset were representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers was evaluated and their results were analyzed. The results of comparison were based on 10 tenfold cross-validations. The comparison made among these classification algorithms out of which the Naive Bayes algorithm showed better performance.

- **Nidhi Bhatla et al.** [1] observations revealed that the Neural Networks with 15 attributes did perform better in comparison with other data mining techniques [1]. The research study concluded that Decision Tree technique showed better performance with the help of genetic algorithms using featured subset selection. This research work also proposed a prototype of Intelligent Heart Disease Prediction system using data mining techniques namely Decision Tree, Naïve Bayes and Neural Network. A total of 909 records were obtained from the Cleveland Heart Disease database. The results reported in the research work justified the better performance of Decision Tree techniques with 99.6% accuracy using 15 attributes. However, Decision tree technique in combination with genetic algorithm the performance reported was 99.2% using 06 attributes.

- **Chaitrali S. Dangare and Sulabha S. Apte** [3] showed that Artificial Neural Network outperforms other data mining techniques such as Decision Tree and Naïve Bayes. In this research work, Heart disease prediction system was developed using 15 attributes [3]. The research work included two extra attributes obesity and smoking for efficient diagnosis of heart disease in developing effective heart disease prediction system.

- **Abhishek Taneja** [10] research work was aimed to design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography. The models were built on the preprocessed Transthoracic Echocardiography dataset with three different supervised machine learning algorithms J48 Classifier, Naïve Bayes and Multilayer Perception using WEKA 3.6.4 machine learning software. The performance of the models was evaluated using the standard metrics of accuracy, precision, recall and F-measure. The most effective model to predict patients with heart disease appeared to be a J48 classifier implemented on selected attributes with a classification accuracy of 95.56%. From a total of 15 attributes that were available, 8 attributes that were highly relevant in predicting heart disease from Transthoracic Echocardiography dataset were selected in the research work.

- **R. Chitra et.al.** [12] Researchers in year 2013 presented Hybrid Intelligent Techniques for the prediction of heart disease. Some Heart Disease classification system was reviewed in this study and

concluded with justification importance of data mining in heart disease diagnosis and classification. Neural Network with offline training is good for disease prediction in early stage and the good performance of the system can be obtained by preprocessed and normalized dataset. The classification accuracy can be improved by reduction in features.

- **Vikas Chaurasia, et.al**. [9] In their research work used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a larger dataset. Observation showed that performance of CART algorithm was better when compared with other two classification methods.
- **V. Manikandan et al.** [2] proposed that association rule mining is used to extract the item set relations. The data classification was based on MAFIA algorithms which resulted in better accuracy. The data was evaluated using entropy based cross validation and partition techniques and the results were compared. MAFIA (Maximal Frequent Itemset Algorithm) used a dataset with 19 attributes and the goal of the research work was to have highly accurate recall metrics with higher levels of precision.
- In year 2014, **Williamjeet Singh and Beant Kaur** [5] published a research paper in IJRITCC "Review on Heart Disease using Data Mining Techniques". The author mentioned the work of large number of researchers and compared various data mining techniques based on performance & accuracy.
- **Aditya Methaila et.al.** [11] In their research work focused on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.
- The researchers **Hlaudi Daniel Masethe and Mosima Anna Masethe** [8] proposed a model for prediction of heart disease using J48, Bayes Net, and Naïve Bayes, Simple CART and REPTREE Algorithms using patient data set from Medical Practitioners. Evaluation of the confusion matrix showed that J48, REPTREE and SIMPLE CART show a prediction model of 89 cases with a risk factor positive for heart attacks. The techniques strongly suggested that data mining algorithms are able to predict a class for diagnoses.
- **B.Venkatalakshmi and M.V Shivsankar** in year 2014 [7] performed an analysis on heart disease diagnosis using data mining techniques Naïve bayes and Decision Tree techniques. Different sessions of experiments were conducted with the same datasets in WEKA 3.6.0 tool. Data set of 294 records with 13 attributes was used and the results revealed that the Naïve Bayes outperformed the Decision tree techniques.

The summary of reviewed literature along with the number of attributes used for the prediction of CardioVascular Disease (CVD) is given in table1.

Table 1: Table shows different data mining techniques used in the diagnosis of Heart disease.

Author/Researcher	Data Mining Technique used	Year	Number of Attributes Selected
Jyoti Sonia, et.al.	Naïve Bayes, Decision Tree, KNN	2011	13
K.Srinivas et.al.	Naïve Bayes, knn and D.L.	2011	14
Nidhi Bhatla et.al.	Naïve Bayes, Decision Tree, Neural Network	2012	15 and 13
Chaitrali S.Dangare & Sulabha S.Apte	Naïve Bayes, Decision Tree, Neural Network	2012	13 and 15
Abhishek Taneja	Naïve Bayes,J48 unpruned tree, Neural Network	2013	15 and 8
Author/Researcher	Data Mining Technique used	Year	Number of Attributes Selected
R. Chitra et. al.	Hybrid Intelligent Techniques	2013	15
Vikas Chaurasia, et.al.	CART,ID3,Decision Table	2013	Not Mentioned
V. Manikandan et al.	K-Mean based on MAFIA, K-Mean based on MAFIA with ID3, K-Mean based on MAFIA with ID3 and C4.5	2013	19
Beant Kaur & Williamjeet	Papers Reviewed	2014	Nil

Aditya Methaila et. al.	Decision Tree, Naive Bayes, Neural Network ,Genetic Algorithm	2014	15 and 6
Hlaudi Daniel Masethe, Mosima Anna Masethe	J48,REPTREE,Naïve Bayes, Bayesnet, Simple CART	2014	15
B.Venkatalakshmi and M.V Shivsankar	Decision Tree and Naïve Bayes	2014	13

4. DATA MINING TECHNIQUES USED :-

Naïve Bayes: - Bayesian classifiers [13] can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of other attributes this assumption is called class conditional independence.
The Bayes theorem is as follows: Let X={x_1, x_2…….. x_n} be a set of 'n' attributes. In Bayesian, X is considered as evidence and H is some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Decision tree: - Decision Trees (DTs) [14] are a non-parametric supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. Decision trees are commonly used in operations research, mainly in decision analysis. Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48.

ID3:- ID3 stands for Iterative Dichotomiser 3 [16]. ID3 adopt a greedy (i.e., nonbacktracking) approach in which decision trees are constructed in a top-down recursive divide and conquer manner .The resulting tree is used to classify the future samples.

J48:- J48 decision tree is the implementation of ID3 algorithm [17] developed by WEKA project team. J48 is a simple C4.5 decision tree for classification. With this technique, a tree is constructed to model the classification process. Once the tree is build, it is applied to each tuple in the database and the result in the classification for that tuple.

Neural Network: - An artificial neural network (ANN) [13], often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal x_i with weights w_{ji} of respective connections from input layer. The output Y_j is function of $Y_j = f(\Sigma\, w_{ji}\, x_i)$ Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

5. PERFORMANCE OF DIFFERENT DATA MINING TECHNIQUES:-

All the research papers referred in this research work have used 13 input attributes and the same is given in table 2 for prediction of CVD.

Table 2: Description of 13 input attributes

Serial No.	Attributes	Description
1	Age	Age in years
2	Sex	Male or female
3	Cp	Chest pain type
4	Thestbps	Resting blood pressure
5	Chol	Serum cholesterol
6	Restecg	Resting electrographic results
7	Fbs	Fasting blood sugar
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise
11	Slope	relative to rest
12	Ca	Slope of the peak exercise ST
13	Thal	Segment

In order to improve the accuracy of results in predicting heart disease two more parameters were used and the same are listed in table 3.

Table 3: Description of 2 incorporated input attributes

Serial No.	Attributes	Description	Value
1	Obes	Obesity	1=yes, 0=No
2	Smoke	Smoking	1=Past,2=Current,3=Never

The performance of all the data mining algorithms that were used in Prediction of Heart Disease is shown in below table 4.

The analysis shows that Neural Network with 15 attributes has shown the highest accuracy i.e. 100% so far. On the other hand, Decision Tree has also performed well with 99.62% accuracy by using 15 attributes. Moreover, in combination with Genetic Algorithm and 6 attributes, Decision Tree has shown 99.2% efficiency.

6. OPEN SOURCE TOOLS USED FOR DATA MINING

- ***WEKA Tool***: - WEKA is a data mining tool developed by the University of Waikato, New Zealand that implements data mining algorithms using the JAVA language. WEKA is one of the state-of-art software used for developing machine learning techniques and their application in real world data mining problems. The data mining algorithms are applied directly to the dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; it also includes visualization tools. The new machine learning schemes can also be developed using this package. WEKA is open source software were the data file are in ARFF file format, which consists of special tags to indicate different things in the data file.
- ***TANAGRA:*** - TANAGRA is a data mining suite developed using graphical user interface algorithms. The main purpose of Tanagra project is to give researchers and students easy-to-use data mining software, and allowing analyzing either real or generated data sets. Tanagra is powerful system that contains clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms
- **.NET *Framework***: - .NET framework is a software framework developed by Microsoft Corporations that runs primarily on Microsoft windows and provides languages interoperability across several programming languages. For developers the .NET Framework provides a comprehensive and consistent application that has visually stunning user experiences with seamless and secure communication.
- ***RapidMiner:*** -RapidMiner is one of the leading open source systems for data mining. RapidMiner is available as a stand-alone application for data analysis and as a data mining engine for the integration into one product. Thousand of applications of RapidMiner in more than 40 countries give their users a competitive edge.
- ***Orange***: - orange is an open data visualization and analysis for novice and experts. Data mining used through visual programming of python scripting, Components for machine learning. Add-ons used for bioinformatics and text mining.
- ***MATLAB***: - MATLAB is a high level language and interactive environment for numerical computation, visualization and programming. Using MATLAB we can analyze data, develop algorithms and create models and applications. The language, tool and built-in math functions enable a researcher to explore multiple approaches and reach a solution faster in comparison to spreadsheets of traditional programming languages.

7. CONCLUSION

The objective of this research work is to provide an insight of different data mining techniques that can be employed in automated heart disease prediction systems. Heart disease is one of the leading causes of deaths worldwide and the early prediction of heart disease is very important. The computer aided heart disease prediction system facilitates the physician as a tool for heart disease diagnosis. Some of the Heart Disease classification systems were reviewed in this study and based on different research studies it was concluded that data mining plays a major role in heart disease classification. Neural Network with offline training is a good tool for disease prediction at early stage. The good performance of the system can be obtained by preprocessed and normalized dataset. The classification accuracy can be improved by reduction in features and by using various techniques. After analyzing different results reported in reviewed research studies

Neural Networks based techniques with 15 attributes has performed better in comparison with Decision Tree based techniques which showed better performance of 99.62% using 15 attributes. The different Data mining tools can help an expert in effective decision making and better diagnosis which will result in proper mitigation of disease in healthcare sector.

8. REFERENCES:

[1] Nidhi Bhatla, Kiran Jyoti, " An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering and Technology Vol.1 issue 8 2012.

[2] V. Manikandan and S. Latha, "Predicting the Analysis of Heart Disease Symptoms Using Medical Data Mining Methods "International Journal of Advanced Computer Theory and Engineering", Vol. 2, Issue. 2, 2013.

[3] Chaitrali S. Dangare, Sulabha S. Apte, ―Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques; International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[4] Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012.

[5] Beant Kaur and Williamjeet Singh.," Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC ,October 2014.

[6] Jyoti Soni et.al. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction; International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[7] B.Venkatalakshmi, M.V Shivsankar, Heart Disease Diagnosis Using Predictive Data mining; International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014.

[8] Hlaudi Daniel Masethe, Mosima Anna Masethe- prediction of Heart Disease using Classification Algorithms; Proceedings of the World Congress on Engineering and Computer Science 2014.

[9] Vikas Chaurasia, et al, Early Prediction of Heart Diseases Using Data Mining Techniques; Caribbean Journal of Science and Technology ISSN 0799-3757, Vol.1,208-217, 2013.

[10] Abhishek Taneja, Heart Disease Prediction System Using Data Mining Techniques; Oriental Journal of computer science & Technology ISSN: 0974-6471 December2013.

[11] Aditya Methaila, Early Heart Disease Prediction Using Data Mining Techniques; CCSEIT, DMDB, ICBB, MoWiN, AIAP pp. 53–59, 2014.

[12] R. Chitra, Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques; Ictact Journal On Soft Computing, July 2013,volume: 03, Issue: 04

[13] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco.

[14] Shanthi Mendis; Pekka Puska; Bo Norrving; World Health Organization (2011).

Global Atlas on Cardiovascular Disease Prevention and Control (PDF). World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-3.

[15] World Health Orginazation ;Cardiovascular Diseases(CVDs) Fact Sheet Reviewed June 2016

[16] Badr HSSINA et.al. A comparative study of decision tree ID3 and C4.5. (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.

[17] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

Table 4: Performance of Different Data Mining Techniques with Accuracy and **number of attributes incorporated**

Researcher/s	Data Mining Techniques	Accuracy with		Precision	Recall	Timing to build Model	Correctly Classified Instances	Incorrectly Classified Instances
		13 Attributes	15 Attributes					
Nidhi Bhatla et.al.	Decision Tree	96.66%	99.62%					
	Neural Network	99.25%	100%					
	Naïve Bayes	94.44%	90.74%					
V. Manikandan et al	k-mean based on MAFIA		74%	0.78	0.67			
	k-mean based on MAFIA with ID3		85%	0.81	0.85			
	k-mean based on MAFIA with ID3 and C4.5		92%	0.82	0.92			
Chaitrali S.Dangare & Sulabha S.Apte	Decision Tree		99.62%					
	Neural Network		100%					
	Naïve Bayes		90.74%					
Jyoti Sonia, et.al.	Naïve Bayes		96.50%					
	Decision Tree		99.20%					
	Classification via Clustering		88.30%					
B.Venkatalakshmi and M.V Shivsankar	Naïve Bayes		85.03%					
	Decision Tree		84.01%					
Hlaudi Daniel Masethe, Mosima Anna Masethe	J48		99.07%			0	107	1
	REPTREE		99.07%			0	107	1
	Naïve Bayes		97.22			0	105	3
	Simple CART		99.07			0.1	107	1
K.Srinivas et.al.	Naïve Bayes		52.33%			609 ms		
	Decision Tree		52%			719 ms		
	K-NN		45.67%			1000 ms		
Aditya Methaila et. al.	Naïve Bayes		86.53%					
	Decision Tree		89%					
	Artificial Neural Network		85.53%					
Vikas Chaurasia, et.al.	CART		83.49%			0.23	253	50
	ID3		72.93%			0.02	221	75
	Decision Table		82.50%			0.03	250	53
Abhishek Taneja	J48	95.56%	95.41%	0.943				
	Naïve Bayes	92.42%	91.96%	0.919				
	Neural Network	94.85%	93.83%	0.938				