International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Survey on Prediction of Heart Disease Using Data Mining Techniques

Dr. B. Umadevi[1], M. Snehapriya[2]

[1]Assistant Professor & Head, P.G. & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India

[2]Research Scholar, P.G. & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, TamilNadu, India

Abstract: ***The major disease which makes sudden demise for the people is the heart diseases in the medical field. It is imperative to predict the disease at a premature phase. The computer aided systems help the doctor as a tool for predicting and diagnosing heart disease. The medical field is dealing with huge amount of data regularly. Handling that large data by traditional way may affect the results. Advanced data mining techniques are used to find out facts in the database and for medical research, particularly in heart disease prediction. The massive amounts of data generated for prediction of heart disease which is too difficult and baggy to be processed and analyzed by conventional methods. Data mining provides the methodology and technology to transform these data into useful information for decision making. Use of data mining algorithms will result in quick prediction of disease with high accuracy.***

Keywords: Data mining, Heart disease prediction, Data mining Techniques.

1. Introduction

Data mining is a novel field for exploring the hidden information patterns from huge raw data sets. In a medical organization like hospitals and medical centers, generates a large amount of data which contains a wealth of hidden information, but these data are not used properly. Hence, that unused data can be converted into useful information by using different data mining techniques [1]. In the modern world, cardiovascular diseases are the highest flying diseases and in every year more than 12 million deaths occur worldwide due to heart problems. Cardiovascular Diseases also cause maximum casualties in India and its diagnosis is a very complicated practice. Health Informatics is a rapidly growing field that is concerned with evolving Computer Science and Information Technology to medical and health data. Medical Data Mining is a domain of challenge which involves a lot of misdiagnosis and uncertainty. A general framework proposed for medical data mining is shown in Figure.1 [2].

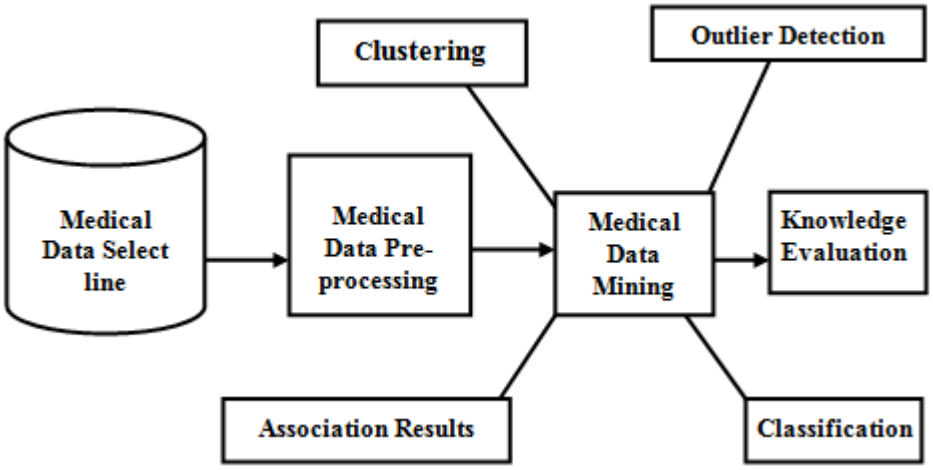

Figure 1: Framework for Medical Data Mining

2. Heart disease

The heart is an important organ of all living individuals, which plays an essential role of blood pumping to the rest of the organs through the blood vessels of the circulatory system. If the circulation of blood in the body is improper the organs like brain suffer and if the heart stops working altogether and death occurs. Life is completely dependent on the proper working of the heart. The term Heart disease refers to disease of heart & the blood vessel system within it. Different types of heart related cardiovascular diseases along with description are given in Table1. Various risk factors along with its symptoms that contribute to heart attack are presented in Table 2.

Table 1: Types of Cardiovascular Diseases

Heart-related cardiovascular disease	*Description*
Acute coronary syndromes	Blood-supply to the heart muscle is swiftly obstructed
Angina	Chest pain due to a lack of blood to the heart muscle
Arrhythmia	Atypical heart rhythm
Cardiomyopathy	Heart muscle disease
Congenital heart disease	Heart disfigurements that are present at birth
Coronary heart disease	Arteries supplying blood to heart muscle becomes obstructed
Rheumatic heart disease	Rheumatic fever

Table 2: Risk Factors and Symptoms of Heart Attack

Risk factors	Symptoms of Heart Attack
• Age • Angina • Blood cholesterol levels • Diabetes • Diet • Genes • Hypertension • Obesity • Physical Inactivity • Smoking • Work	• Chest Discomfort • Coughing • Nausea • Vomiting • Crushing chest pain • Dizziness • Dyspnoea (shortness of breath) • Restlessness

In many cases, diagnosis is generally based on patient's current test reports & the doctor's experience. Thus the diagnosis is a complex task that requires high skill & much experience. Datasets of heart disease patients can be collected from various Universities like UCI, Cleveland, etc. for our intelligence system. The attributes like age, sex, chest pain, resting blood pressure, cholesterol mg/dl, blood sugar, maximum heart rate, etc. Data pre-processing is done to extract relevant data and then those data should be converted into the format necessary for the prediction of risk. Cleaning and filtering of datasets is done sometimes to remove duplicate records, normalize the values, accounting for missing data and removing irrelevant data items. Table 3. Describes about the data set for heart disease prediction.

Table 3: Data set description for Heart disease Prediction

Technique	**Data Set**	**Attribute**	**Description**
Classification	Heart	Age, sex, chest pain type Restbloodpressure, serumcholestoral, Slope Fastingbloodsugar, Old peak RestingbloodSugar, Restelectrocardingraphic,	Obesity, smoking Maximum heart attack achieved The slope of peak exercise, Diagnosis of Heart disease.

3. Healthcare Data Mining

Data mining holds immense potential for the healthcare industry to set up health systems to systematically use data and analytics for determining inefficiencies and the finest practices that improve care and reduce costs. Most hospitals today use sort of hospital information system to manage huge and voluminous amounts of patient's data. There is a wealth of hidden knowledge in these data that is largely untapped, which using data mining [3-5] can be turned into beneficial information that can allow healthcare practitioners to take intelligent clinical decisions. Medical science is another field where large amount of data is generated using different clinical reports and other patient symptoms. Data mining can also be used heavily for the same purpose in medical datasets also. These explored hidden patterns in medical datasets can be used for clinical diagnosis. However, medical datasets are widely dispersed, heterogeneous, and huge in nature. These datasets need to be organized and integrated with the hospital management systems. This disease attacks a person so instantly that it hardly gets any time to get treated with. So diagnosing patients correctly on a timely basis is the most challenging task for the medical. A wrong diagnosis by the hospital leads to earn a bad name and loosing reputation. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. Some of the prediction based data mining techniques are as follows:

3.1 Bayesian Classifiers

Using Bayesian classifiers, the system can discover the concealed knowledge associated with diseases from historical records of the patients having heart disease. Bayesian classifiers predict the class membership probabilities, in a way that the probability of a given sample belongs to a particular class statistically. Bayesian classifier is based on Bayes' theorem. We can use Bayes theorem to determine the probability that a proposed diagnosis is correct, given the observation. A simple probabilistic, the naive Bayes classifier is used for classification based on which is based on Bayes' theorem. According to naïve Bayesian classifier the occurrence or an occurrence of a particular feature of a class is considered as independent in the presence or absence of any other feature. When the dimension of the inputs is high and more efficient result is expected, the chief Naïve Bayes Classifier technique [6-8] is applicable. The Naïve Bayes model identifies the physical characteristics and features of patients suffering from heart disease. For each input, it gives the possibility of attribute of the expectable state. Naïve Bayes is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value of a given class is independent of the values of other attributes. The advantage of using naïve bayes is that one can work with the Naïve Bayes model without using any Bayesian methods.

3.2 Decision Tree

Decision Tree is a popular classifier which is simple and easy to implement. There is no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees. The presentation of the Decision Tree technique [9][10] in the treatment of heart disease has been investigated by the researchers with significant success. The decision tree is a tree like structure, which consists of internal nodes, branches and leaf nodes, in which each

branch denotes an attribute value, each internal node denoted a test on an attribute which is used for and a leaf node represents the predicted classes or class distributions. The classification starts from the root node, then traverses the tree based on the predictive attribute value. The methodology involves data partitioning, data classification, decision tree category selection, and the request of reduction of fault trimming to create trimmed decision trees.

3.3 Neural Networks

In practical applications, neural networks are well known to generate highly accurate results. By using a feed forward neural network model [11] variable learning rate and back propagation learning algorithm with momentum, the neural network is trained with the Heart Diseases database. The design of the model is as follows: It starts with the input of clinical data and progresses to develop ANN algorithm. After training model, it can produce the prediction [12] results. The computational steps of a neural network algorithm begin with the classification of clinical data into two equal parts randomly. One is used for testing and the other is used for training.

3.4 Support Vector Machines

SVM is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory . It is a method for classification of both linear and non-Linear data. The training data is converted into non-dimensional data using non-linear transformation method. Then the algorithm search for the best hyper-plane to separate the transformed data into two different classes. SVM performs classification tasks by maximizing the margin of the hyper-plane separating both classes while minimizing the classification errors. **T**he SVM algorithm predicts the occurrence of heart disease by plotting the disease, predicting attributes in multidimensional hyperplane and classifies the classes optimally by creating the margin between two data clusters. This algorithm attains high accuracy by the usage of nonlinear [13] functions called kernels.

4. Open source tools available for data mining

4.1 Rapid Miner

Rapid Miner is unquestionably the world-leading open-source system for data mining. It is available as a stand alone application for data analysis and as a data mining engine for the integration into own products. It offers an integrated environment useful in machine learning, text mining, data mining, business analytics and predictive analytics. The tool supports various steps useful in data mining including result optimization, visualization and validation.

4.2 Weka

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be appropriate straight to a dataset or called from your own Java code. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. It shows you various relationships between the data sets, clusters, predictive modeling, visualization etc.

4.3 Orange

Orange is an Open source data visualization and analysis for noise and experts in Data mining through visual programming or Python scripting. Orange incorporates various components useful in data preprocessing, feature filtering and scoring, model evaluation, exploration and modeling techniques.

4.4 Tanagra

Tanagra is a free data mining software for academic and research purposes. It proposes several data mining methods for data analysis, statistical learning, machine learning and databases area. The main purpose of Tanagra project is to give researchers and students easy to use data mining software to analyse either real or synthetic data.

4.5 Matlab

A proprietary programming language developed by the MathWorks, Matlab allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python. When doing data mining, a large part of the work is to manipulate data. Indeed, the part of coding the algorithm can be quite short since Matlab has a lot of toolboxes for data mining.

4.6 KNIME

KNIME developed as a proprietary product, is a comprehensive open-source data integration, processing, analysis, and exploration platform. Using modular data pipelining concept, it integrates various components for machine learning and data mining.

5. Related Works

Carlos Ordonez [14] did a study on prediction of heart disease with the help of Association rules. They used a simple mapping algorithm. This algorithm constantly treats attributes as numerical or categorical. This is used to convert medical records to a transaction format. An improved algorithm is used to mine the constrained association rules. A mapping table is prepared and attribute values are mapped to items. The decision tree is used for mining data because they automatically Split numerical values [14]. The split point chosen by the Decision tree is of little use only. Clustering is used to get a global understanding of data.

Usha Rani [15] have proposed a system for predicting heart disease with the help of artificial neural network, which is a combination of feed forward and back propagation algorithm. The experiment is carried out by considering single and multilayered neural network models. Parallelism is implemented to speed up the learning process at each neuron in all hidden and output layers.

T. Revathi and S. Jeevitha [16] analyzed the data mining algorithms on prediction of heart disease. The clinical data related to heart disease is used for analysis. The results of Neural Network, Naïve Bayes, and Decision Tree algorithms are compared, Neural Network achieved good accuracy.

Devendra Ratnaparkhi, Tushar Mahajan and Vishal Jadhav [17] proposed a heart disease prediction system using Naïve Bayes and compared the results with Neural Network and Decision Tree algorithms. According to that method, the Naïve Bayes algorithm provides good prediction.

K. Manimekalai [18] enlightened various data mining techniques to predict heart disease. From the experimental results, SVM classifier with genetic algorithm provides better prediction accuracy while compared with Naïve Bayesian, C 5.0, Neural Network, KNN, J4.8, decision tree and Fuzzy mechanism algorithms.

Jyoti Rohilla and Preeti Gulia [19] analyzed some of the data mining algorithms to predict heart disease. They have used a heart disease dataset from the UCI machine learning repository and analyzed using WEKA tool, shown that decision tree algorithms performed well in predicting heart disease.

Shadab Adam Pattekari and Asma Parveen [20] developed a Decision Support in Heart Disease Prediction System using Naive Bayesian Classification technique. The system discovers the hidden knowledge from a past heart disease database. This is the most effective model to predict patients with heart disease. This model could respond to complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

Nilakshi P. Waghulde, Nilima P. Patil [21] did an experiment with Heart Disease dataset by taking
into consideration of Multilayer Neural Network along with Back propagation Learning Algorithm for training the network. To optimize the initialization of neural network Weights genetic algorithm is used. This work shows the result of the Genetic Neural Network for prediction of heart disease by improving the accuracy as 98% using optimize neural network architecture, it predicts whether the patient is suffering from heart disease or not.

Upasana Juneja et al. [22] performed a work on Heart Disease Analysis under the intelligent system that can be adapted by a doctor and parameter based fuzzification that will perform the analysis based on some parameters. The proposed work is analysis on the patient symptom information based on which a pre-level decision is taken to identify the chances of a heart disease. Specifically, the whole application software finds the frequent illnesses with medication. This research provides important facts like correlations between medical issues related to disease finally.

Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq [23] focused on, the evolution of big data in healthcare system. In their paper applied, Support Vector Machine (SVM), Decision Tree (C4.5) and Bayesian Network machine learning algorithms. Chronic Kidney disease dataset from UCI Machine Learning Repository is used to predict patients with chronic kidney failure disease and patients who are not suffering from chronic kidney disease. C4.5 classifier provided results with minimum execution time and better accuracy.

6. Conclusion

Heart disease is the leading cause of death for both men and women. Know the warning signs and symptoms of a heart attack so that you can act fast if you or someone you know might be having a heart attack. The chances of survival are greater when emergency treatment begins quickly. This paper mainly focuses on the study of various approaches of heart attack disease prediction research papers are analyzed and studied. The prediction accuracy of existing systems can be improved, so In future, new algorithms and techniques are to be developed which overcome the drawbacks of the existing system.

References

[1] B. Umadevi, D.Sundar, Dr.P.Alli, “A Study on Stock Market Analysis for Stock Selection – Naïve Investors’ Perspective using Data Mining Technique”, International Journal of Computer Applications (0975 – 8887), Vol 34– No.3,2011.
[2] R. Alizadehsani, J. Habibi, B. Bahadorian, et al., “Diagnosis of coronary artery stenosis using data mining”,J MED Signals Sens, vol. 2, pp. 153-9,2012.
[3] Chaurasia V, “Early prediction of heart diseases using data mining techniques”, Caribbean Journal of Science and Technology, 1:208–17, 2013.
[4] Alzahani SM, Althopity A, Alghamdi A, et al., “An overview of data mining techniques applied for heart disease diagnosis and prediction”, Lecture Notes on Information Theory, 2(4):310–5,2014.
[5] Swathi P, Yogish HK, Sreeraj RS, “Predictive data mining procedures for the prediction of coronary artery disease”, International Journal of Emerging Technology and Advanced Engineering, 5(2):339–42,2015.
[6] Liu X, Lu R, Ma J, Chen L ,“ Privacy-preserving patient centric clinical decision support system on naïve bayesian classification”, IEEE Journal of Biomedical and Health Informatics, 20(2):655–88,2016.
[7] Patil RR,“A heart disease prediction system using naïve byes' and Jelinek-Mercer smoothing”, International Journal of Advanced Research in Computer Science and Communication Engineering, 3(5):6787–9,2014.
[8] Pattekari SA, Parveen A, “Prediction system for heart disease using naive bayes”, International Journal of Advanced Computational and Mathematical Sciences, 3(3):290–4, 2012.
[9] Komal G, Vekariya V, “Novel approach for heart disease prediction using a decision tree algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, 3(11):11544–1, 2015.

[10] Shouman M, Turner T, Stocker R, "Using decision tree for diagnosing heart disease patients", Proceedings of the 9th Australasian Data Mining Conference (AusDM'11); Ballarat, Australiap. 23–30, 2011.
[11] Rani KU, "Analysis of heart diseases dataset using neural network approach", IJDKP, 1(5):1–8,2013.
[12] B. Umadevi D.Sundar, Dr.P.Alli, "An Optimized Approach to Predict the Stock Market Behavior and Investment Decision Making using Benchmark Algorithms for Naive Investors", Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on (IEEE Xplore Digital Library),pg1 -5.
[13] B. Umadevi, ,D.Sundar, Dr.P.Alli An Effective Time Series Analysis for Stock Trend Prediction Using ARIMA Model for Nifty Midcap-50, International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.3, No.1,2013.
[14] Carlos Ordonez, Edward Omincenski and Levien de Braal, "Mining Constraint Association Rules to Predict Heart Disease",IEEE International Conference on Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, pp: 433-440,2001.
[15] Usha. K Dr, Analysis of Heart Disease Dataset using neural network approach",IJDKP,Vol 1(5),2011.
[16] T. Revathi, S. Jeevitha, "Comparative Study on Heart Disease Prediction System Using Data Mining Techniques", Volume 4 Issue 7, ISSN (Online): 2319-7064,2015.
[17] Devendra Ratnaparkhi, Tushar Mahajan, Vishal Jadhav, "Heart Disease Prediction System Using Data Mining Technique",International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 08, e-ISSN: 2395 -0056, p-ISSN: 2395-0072,2015.
[18] K.Manimekalai, "Prediction of Heart Diseases using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 2, ISSN(Online):2320 - 9801, ISSN (Print):2320-9798, 2016.
[19] Jyoti Rohilla, Preeti Gulia, "Analysis of Data Mining Techniques for Diagnosing Heart Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, ISSN: 2277 128X, 2015.
[20] Shadab Adam Pattekari and Asma Parveen, "Prediction System For Heart Disease Using NaiveBayes", International Journal of Advanced Computational and Mathematical Sciences, ISSN 2230 - 9624, Vol 3, Issue 3, pp 290-294, 2012.
[21] Nilakshi P. Waghulde, Nilima P. Patil, "Genetic Neural Approach for Heart Disease Prediction", International Journal of Advanced Computer Research (ISSN (print): 2249-7277, Vol 4 Number-3 Issue-Sept 2014.
[22] Upasana Juneja et. al., "Multi Parametric Approach Using Fuzzification on Heart Disease Analysis", IJESRT, Juneja et al., 3(5) ISSN: 2277-9655, Page No.492-497,2014.
[23] Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq, "Performance of Data MiningTechniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease", International Journal of Database Management System, Vol.8, No.3, 2016.

Authors Profile

Dr. B. Umadevi has received her Doctoral degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, India. Currently working as Assistant Professor & Head - P.G and Research Department of Computer Science, Raja Doraisingam Government Arts College, Sivagangai-Tamilnadu, India. She has over 22 years of Teaching Experience and published her research papers in various International, National Journals and Conferences. Her research interests Include Data Mining, Soft Computing and Evolutionary Computing. She got the Best Paper Award For her publication in the IEEE International Conference on Computational Intelligence and Computing Research held on 27th Dec 2013 at VICKRAM College of Engineering and Technology.

M. Snehapriya is an M.Phil Research Scholar in PG & Research Department Of Computer Science, Raja Doraisingam Government Arts College, Sivaganga, Tamilnadu, India. Her research Interest include in Data mining.