A SURVEY ON HEART DISEASE AND BREAST CANCER PREDICTION BY USING DATA MINING TECHNIQUES

K.Yuvaraj[1] and D.Manjula[2]

ABSTRACT

Data mining is the process to analyze and extract the huge data. Data mining plays a outstanding task for predicting diseases in health concerns. The data are too large and complex to predict heart disease for processing. Multiple tests are needed to predict heart disease from the patient. By using data mining technique the tests can be reduced. A quick and regimented detection technique to reduce number of deaths from heart diseases is needed. Cardiovascular disease is the major source of deaths pervasive and the prediction of Heart Disease is significant at an ill-timed phase.

Keywords : Data Mining, data mining techniques, Heart Disease, Breast Cancer

I. INTRODUCTION

The main objective of our paper is to survey on the different techniques of data mining which is used in prediction of heart disease by using different data mining tools [1]. Heart is essential part of our body, incase if it is not working properly it would affect all other parts of the body [2]. The heart disease can be increased by so many factors and cause death. [3]

The diseases like Hepatitis, lung cancer, liver disorder, breast cancer etc are predicted by using data mining techniques to give accurate result.

In this survey the Heart disease, and Breast cancer disease predictions are analyzed by using data mining techniques[4].

II. EXISTING ALGORITHMS II

FACTORS OF HEART DISEASE

High blood pressure :

High blood pressure is a stress to your heart which also affects heart or kidney for human being. Because of high blood pressure it may cause stroke or heart attack.

Smoking :

Lung disease was caused by smoking. It damages small air sacs which was found in lungs. The main effect of causing lung cancer is due to cigarette smoking.

Family history of heart disease :

Heart disease may also caused by family histories. If a person in a family may have cardiovascular disease it may also affect their families.

Cholesterol :

High blood cholesterol is one of the major risk factors for heart disease. A risk factor is a condition that increases your chance of getting a disease. In

[1] Associate Professor, Department of Computer Science, Karpagam University

[2] Associate Professor, Department of Computer Science, Karpagam University

fact, the **higher** your **blood cholesterol** level, the greater your risk for developing heart disease or having a heart attack.

Lack of physical exercise : - The high blood pressure can be control by physical action. By doing this physical action it help us to control our body weight and make more pretty and strong in your heart ,so the stress can be lowered. So it can develop the healthy heart which makes good in blood pressure of your body.

Figure 1 : Causes of Heart Diseas

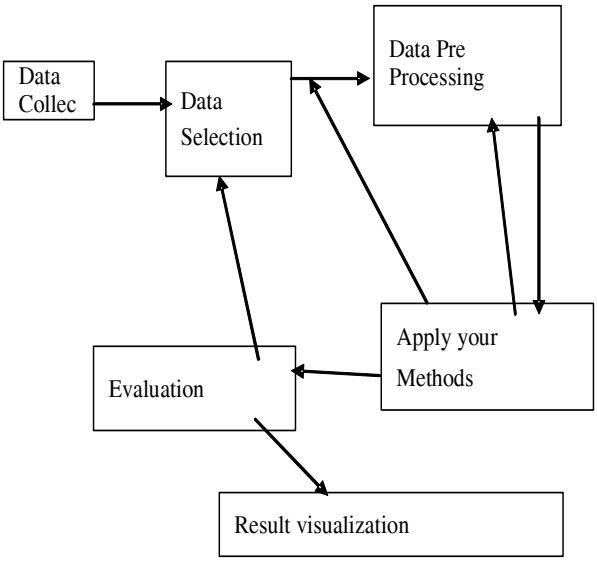

Figure 2 : Different data mining processes

Table 1 :

Heart disease dataset using different data mining

Author	Year	Technique	Accuracy
Chaitrali et al,	2012	Naive Bayes	89.74%
DT		89.62%	
NN		100%	
Indira S. Fal Dessai	2013	PNN	92.6%
DT		74.2%	
NB		64%	
BNN		70.4%	
Jesmin et al	2013	Naive Bayes	91.08%
M. anbarasi et al	1999	Naive Bayes	86.5%
Decision Tree		89.2%	
Classification via clustering Naive Bayes		89.3%	
Matjaz et al	1999	exercise ECG(NN) exercise ECG(NN)	54%
myocardial scintigraphy(NN)		65%	
Naïve bayes		58%	
T. John et al.	2012	Naïve bayes technique	96.18%

III. Problems And Challenges

The difficult task in medical field and their profession is applying data mining[2].The hypothesis which are results are adjusted to fit the hypothesis in medical research is begins with data mining. By using this we can starts with dataset this differs from standard data not including apparent hypothesis. [11]Patterns

and trends in dataset are mainly concerned with traditional data mining, but in medical data mining they are not conformed. According to the doctor intuition the clinical decision are often made. The quality of service provided to patients is affected due to unwanted bias, errors and undue medical cost. Data mining have the capacity to generate a knowledge-rich environment. It can help to improve the significant quality of clinical decision. [5] In the survey of [3] the three supervised machine learning algorithms are used.

The above algorithms has been used to analyse the heart disease. In this algorithms classification accuracy is used. To predict the heart disease with minimum number of attributes the above work is used.

In the survey of [3] the heart disease is predicted by using association rule data mining technique. The author introduced an algorithm that uses search constraint to decrease the number of rules. In future this work should be extended by using fuzzy learning models to find the accuracy of time to decrease the number of rules[9]. In the survey of [4] the author proposed a new concept that uses weighted association rule for classification. In future this work can be extended by using association rule hiding technique in data mining. In the survey[5] to predict heart disease the minimal subset of attributes has been proposed. The datas are collected from different health organizations to compare the accuracy with all the techniques in datamining. In the survey of [6] attributes of a diabetic patient has been predicted. Weka tool was able to classify 74% of the input correctly.

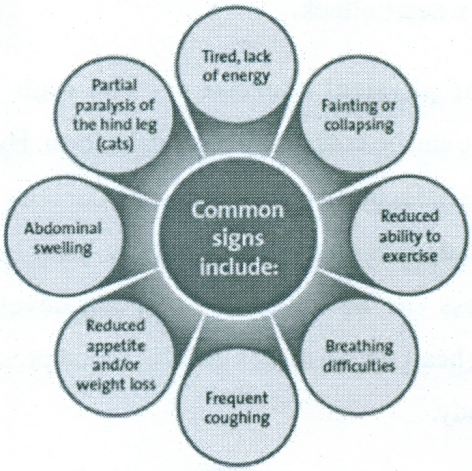

Figure 3 : Symptoms of Heart disease

IV. Techniques In Data Mining :

Classification: It is a function which targets the items on a group to achieve classes. The main aim of classification is to find exact point of the objective class in each folder in data. For example, a **classification** model is used to identify the loan applicants as low, medium or high risks. This classification is used on machine learning this classification can be based.

Clustering: - Cluster is a collection of objects that belongs to the identical class. Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups[1]. **Cluster analysis** is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. The objects which are assigned

can be defined in classes and where the object can be clustered in predefined classes.

The below example In prophecy of heart disease by using clustering we get crowd together otherwise the similar threat of patients can be prohibited[3]. The patients with high blood sugar and linked threat and so on can be identified.

Prediction:- classification model to predict the treatment outcome for a new patient, and it is used to be predict. The prophecy as it name that identity of one thing and which is based purely on the description of another, related thing.. Here we take an example that predict analysis can be needed in sale for the purpose of predict profit for the upcoming days. In this independent variable is an sale and dependent variable is a profit. By using a profit data and ancient sale the profit prophecy can be drawn by a built in regression bow.

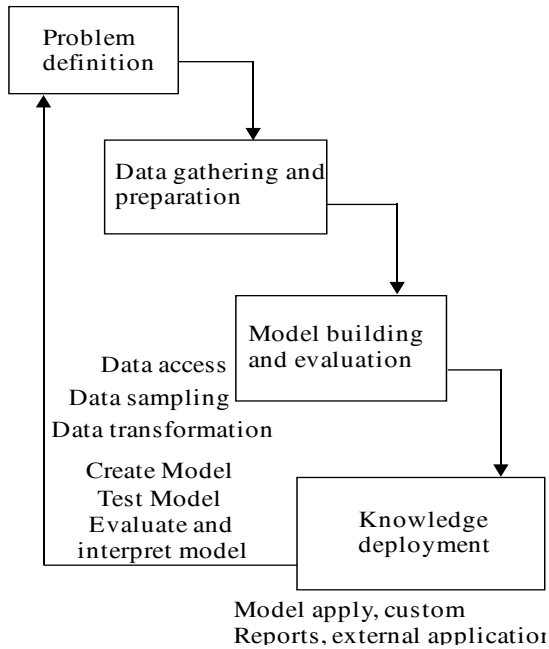

Figure 4 : Process of data mining

Some of the Open source tools for data mining

WEKA Tool : -

It is system developed by the university of Waikato in new Zealand. It implements datamining algorithm using JAVA. It is a stste of the art facility to develop machine education techniques. These are applied to dataset. WEKA includes visualization tools. It implements algorithms for data processing, classification, clustering etc. With this package new machine learning schemas can be developed. It is also known as open source software. The data file ARFF file format is used. It consists of special tags.

TANAGRA : -

The free data mining software for scholastic and research purposes can be done by using tanagra tool. It can be used for data mining methods for accurate data analysis and some learning purpose. The main motto of this software is to use the data very easy for mining software and present norms of the software development in this area can be confirmed for analyzing the factual data or fake data.

MATLAB : -

It is an interactive environment for programming, visualization, numerical computation. By using this we can analyse data and to improve the algorithms and to create models and applications. The math function enable us to discover various approaches and it reaches the solution faster.

Table 2 : Different data mining tools used on heart disease predictions with accuracy.

Author	***Technique used***	***Data mining tool***	***Accuracy***	***Objective***
Abhishek et al (2013)	J48	Weka 3.6.4	95.56%	HDP System Using DM Techniques
Naive Bayes			92.42%	
94.85%				
J48				
Chaitrali et al (2012)	Neural Network	Weka 3.6.6	100%	Prediction of HD
Monali Et al	C4.5	WEKA		Study and Analysis of Data mining Algorithms for Healthcare Decision Support System
Multilayer Perceptron				
Naïve Bayes				
Nidhi et al (2012)	Naive Bayes	Weka 3.6.6	90.74%, 99.62%, 100%	Analysis of HDP using Different DM Techniques
Decision Trees		TANAGRA	52.33%, 52%, 45.67%	
Weka 3.6.0			86.53%, 89%, 85.53%	
Neural networks		.NET platform	96.5%, 99.2%, 88.3%	
Resul et al (2009)	Neural networks	SAS base software 9.1.3	97.4%	Diagnosis of valvular HD
Rashe-Dur Et al (2013)	Neural Network	WEKA	79.19%	Comparison of Various Classification Techniques
Fuzzy Logic		TANAGRA	83.85%	
Decision Tree		MATLAB		
Resul et al (2009)	Neural networks	SAS base software 9.1.3	89.01%	diagnosis of HD

V. Breast Cancer Prediction

According to the survey of united state is affected by breast cancer[16].To predict the breast cancer the mammography is a conventional method and it also showed that how it was interpret with mammogram by considering radiologists. Elmore states that very few (3%) of cancers are recognized and it is indicated 90% radiologists case[15].

To diagnosis the breast cancer the excellent pointer desire cytology for producing accurate accuracy. Hence the correct average identification rate is near to 90% [17].

In this survey our major motto is to connected research which are matching and decide among patients with breast cancer(called Bad group) and patients without breast cancer (called gentle group).

Here it can be predicted by 3 ways

1) cancer reaction can be calculated.

2) Reappearance of cancer can be calculated.

3) Cancer growth can be survived.

BY author point of view it states that the traditional analytical factor for breast cancer is the American Joint Commission on Cancer (AJCC). It is production system based on the TNM system (T, tumor; N, node; M, metastasis) [7] and frequency of breast cancer can be measured by survival of the person is still alive from the date of tightening where cancer has recurred at a specific time.

VI. ISSUES AND CHALLENGES

In [8] the author states the problems, algorithms, and techniques to predict breast cancer. The records are not included in this analysis. In future work this work may enhance by including missing data.

In [9] the author states that the duplicate neural network has been trained in many hospitals by using the variables of patient. The author describes that the numeral networks are more precious tools. The data collected from more current era and to find new factors included in a neural association model. In[10] the author describes the multiple prophecy models by using large datasets. it also contains cross validation methods.

VII Conclusion

This survey paper is developed using four data mining classification modeling techniques.. This ancient heart disease database can be extracted the hidden knowledge. Here lot of techniques and data mining machine learning tool are used for skillful and successful heart disease diagnosis in past years. By using different technology the different number of attributes can be taken from more papers. By using this technologies different exactness can be showed to each other. Hence applying data mining methods are used for providing health care profession on heart disease can be predict and some of them is success ,for heart disease the data mining method can be used to identify for treatment and the out come of patient has given low attention. In further work it can be used for the various types of disease prediction and the work can be stretched and better for the automation of the disease. By using this attributes we can find another disease also with huge amount.

References

1. Ho, T. J. : *"Data Mining and Data Warehousing",* Prentice Hall, 2005.

2. Siri Krishan Wasan1, Vasudha Bhatnagar2 and Harleen Kaur, (2006) *"The impact of Data Mining Techniques onMedical Diagnostics",* Data Science Journal, Volume 5, 119-126.

3. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni *"Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction"* IJCSE Vol. 3 No. 6 June 2011.

4. M. Anbarasi, e. Anupriya, n.ch.s.n.iyengar, *"Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm",* International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370- 5376 [6].

5. G. Parthiban, A. Rajesh, S.K.Srivatsa *"Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method"*

6. Choi J.P., Han T.H. and Park R.W., *"A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis",* J Korean Soc Med Inform, 2009, pp. 49-57

7. Venkatadri. M, Dr. Lokanatha C. Reddy a review on data mining from past to the future. International Journal of Computer Applications, 2011.

8. Bellaachia Abdelghani and Erhan Guven, *"Predicting Breast Cancer Survivability using Data Mining Techniques,"* Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Confere nce on Data Mining," 2006.

9. Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , *"Artificial Neural Networks Applied to Survival Prediction in Breast Cancer",* Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

10. Delen Dursun , Walker Glenn and Kadam Amit , *"Predicting breast cancer survivability: a comparison of three data mining methods,"* Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005.

11. Ruben D. Canlas Jr., *"DATA MINING IN HEALTHCARE : CURRENT APPLICATIONS AND ISSUES",* August 2009.

12. Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, A Prototype of Cancer/ Heart Disease Prediction Model Using Data Mining, International Journal of Applied Engineering Research, 2012.

13. Ma.jabbar, Dr.prirti Chandra, B.L.Deekshatulu, cluster based association rule mining for heart attack prediction, Journal of Theoretical and Applied Information Technology,2011.

14. Ms. Ishtake S.H ,Prof. Sanap S.A., *"Intelligent Heart Disease Prediction System Using Data Mining Techniques",* International J. of Healthcare & Biomedical Research,2013.

15. Shantakumar B.Patil, Dr.Y.S. Kumaraswamy, Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction, (IJCSNS) International Journal of Computer Science and Network 228 Security ,2009

16. Wingo PA, Tong T, Bolden S, *"Cancer statistics",* 1995, CA Cancer J Clin 45 (1995), no. 1, 8-30.

17. Fentiman IS, *"Detection and treatment of breast cancer",* London: Martin Duntiz (1998).

18. B.Pushpalatha, C.Willson joseph, *" Credit card fraud detection based on transaction by using datamining techniques",* IJIRCCE, 5(2), 2017.

Author's Biography

K.Yuvaraj completed his M.Phil in computer science from Bharathidasan University in 2014. He is working as assistant professor in department of computer science, Karpagam University, Coimbatore. His experience is 1 years 3 months. He has presented in international conference. His research area is networking.

D. Manjula completed her M.Phil in computer science from Bharathiyar University in 2015. she is working as assistant professor in department of computer science, Karpagam University, Coimbatore. Her experience is 6 Months. She has presented in international conference. Her research area is data mining.