A REVIEW ON DISEASE PREDICTION USING DATA MINING TECHNIQUES

P.B. Tintu, S. Manju Priya*

Abstract

Data Mining extracts are useful and interesting patterns which are from large datasets. Health industry contains a diverse amount of data which shall not be used without the appropriate application of Data Mining techniques. Using Data Mining in medical areas can help in analyzing critical issues and studying the risk factors of several diseases. The mined data provide knowledge to the physicians for detection of disease. The main focus of this review article is to analyze the usage and importance of Data Mining.

Keywords: Data Mining, KDD process, Medical Area, Predicting Diseases.

I. INTRODUCTION OF DATA MINING

The data to be mined analyzes, discovers and summarizes patterns from enormous dataset and converts it into an overall goal to extract information.Data Mining is the analysis of "knowledge discovery in database". During the current epoch, Data Mining is becoming widespreadin the healthcare field because it helps in timely analyzes and accurate diagnosis of diseases which can save the lives of many patients. It also helps the health care providers for making efficient research for making efficient healthcare policies and health profits for individuals. Different algorithms are used for various diseases diagnoses. Gleaned from the information used the accuracy along with execution will vary.

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

II. KDD PROCESS

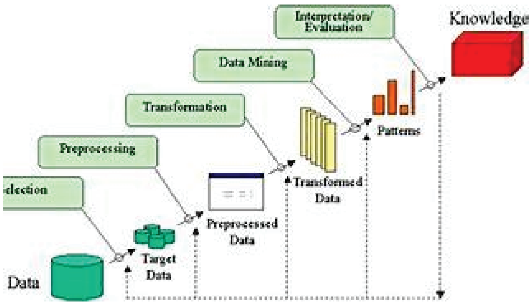

Fig 1 Steps of KDD process[1]

A huge amount of information is hidden among a vast collection of information.

Selection

Data required for analysis is selected according to condition for data selection.

Preprocessing

In this stage,the information which is not required is removed. This stage is also known as the data cleaning stage.

Transformation

In this stage, data is transformed based on the necessity of research.

Data Mining

In the knowledgediscovery process,Data Mining is used where this stage extracts useful and meaningful information.

Interpretation and evaluation

Patterns which are identified by the system are interpreted as knowledge into this stage.

III. RELATED WORK

P.Ahmad, Saqib Qamar and S.Q.A.Rizvi [2] presented classification, clustering, association, etc… through Data Mining approach in the health area. In future work, they have discussed DataMining techniques.

In the paper proposed by Parathion I and Siddhartha Autarky, [3] they had specified about hybrid technology through classification algorithm and association mining method. Various applications and techniques of Data Mining are discussed. How Data Mining is used in the medical area and its advantages and disadvantages are discussed.

P. Nayak and Sayeesh [4] discussed usage of Data Mining in medical services. Advantages and Disadvantages of Data Mining approach and analytical commencement regarding medicaldataset in health care are also discussed.

IV. DATA MINING IN MEDICAL AREA

Medical Data Mining is a vital area of research which is used to expose patterns from information which was used to develop predictive replicas. All health care organizationsstore data in electronic order. The data contain all details about patients and health care providers. By using traditional methods, it is difficult to extract meaningful information. Data mining is used in situationswhere a large collection of healthcare data is used. Data Mining tools help us to discover information from unknown patterns. Techniques like classification, prediction, clustering and association rules are some techniques which are used on medical information. This review paper presents DataMining techniques in the field of medical area.

V. DATA MINING FOR CANCER PREDICTION

Cancer will affect any part of the body which denotes a large group of diseases, about twenty five percent cause for death is cancer. Second common cancer, fifth cause of death is breast cancer worldwide. For women between the ages 33 to 55, it is found as a malignant tumor. Survival of disease infected can be improved on early detection. Cancer replicates into abnormal cells which grow very fast, affecting human parts and it spreads to different parts; this process is known to be metastasizing. Different methods to monitor and identify breast cancer are

(a) Luminescence techniques and

(b) Mammograms,

Genetic data can be used for analyzing breast cancer. Many researchers have been attracted towards statistics fields and computational intelligence. Tobacco, obesity, unhealthy diet, lack of exercise, alcohol, air pollution, smoke from plastic and fuels are the major factors of cancer.

Dr. N. Subhash Chandra,Dr. G. Narsimha, V. Krishnaiahand [5] used the techniques of Data Mining in analysis of cancer in Lungs. Symptoms and risk factors of lung cancer are explained. Decision Tree, Bayesian Network, Rule set classifiers, Neural Network were used for analysis. A. Govardhan and Jothi Prabha [6] presented classification techniques on attributes of breastcancer. Decision Tree algorithm using J48, Naïve Bayes classification and dataset of chronic disease was analyzed. They found that Naïve Bayes classification gave good and better results. Deepika R and Durairaj M [7] used Microarray technique to give a review about Acute Myeloid Leukaemia (AML), Myelodysplastic Syndrome (MDS) and prediction. About ten papers were reviewed by authors for diagnosing disease. Accuracy of the algorithm was compared by using the WEKA tool. It was found that the decision table and 1BK gave accurate results. Vikas Chaurasi, Saurabh Pal [8] detected breast cancer using Simple Logistic, RepTree and RBF Network. They used the WEKA tool for analysis. 74.47% accuracy was achieved by a simple logistic algorithm.

VI. DATA MINING FOR HEART DISEASE PREDICTION

An umbrella term for heart disorder is heart disease which affects the human heart. A disorder which affects the blood vessels and heart is known as cardiovascular disease. According to the CDC and World Health Organization, heart disease leads to the cause of death in Australia, Canada, UK, and USA.

The factors which are main reasons for heart diseases are overweight, diabetes in patients, smoking habits, blood pressure level, junk foods, cholesteroland nonphysical activities [9]. Jeevitha.S and T. Revathi [10] analyzed heart disease prediction using algorithms in Data Mining. Clinical data is used for analysis. Naïve Bayes, Neural Network algorithms were used for comparison to achieve perfect accuracy. Manimekalai.K, [11] for predicting heart disease, used various Data Mining algorithms. In comparison with C 5.0, KNN, Neural Network and Naïve Bayesian, genetic algorithms with SVM classifiersgave better prediction accuracy. Tushar Mahajan, Devendra Ratnaparkhi and Vishal Jadhav [12] used Naïve Bayes andproposed heart disease prediction system. They compared the results with Decision Tree algorithms and Neural Network. According to the experiment, Naïve Bayes algorithm gave good prediction.

VII. DATA MINING FOR CANCER PREDICTION LIVER DISEASE PREDICTION

Any disturbance in the liver leads to disease of liver function. Damage will be caused to the human body if the liver, which is responsible for various functions in the human body is injured or diseased. Another name for liver disease is hepatic disease. In countries like England and Wales disease in the liver is the fifth after cancer, heart disease, respiratory and stroke disease . Hepatitis C is estimated in every five people out of every six who are unaware of their infectious disease . Risk factors Autoimmunity problem, Alcoholism, toxins, viruses, hereditary conditions are some of the risk factors .C. Jothi Venkateshwaran and A. S. Aneeshkumar[13] presented a classification method for approaching liver disorder. Fuzzy based classification gave better results for diagnosis in liver disorder. Jemina Priyadarshini and Sindhuja [14], described in their paper about classification techniques which are used for analyzing liver disease. Algorithms such as SVM, Decision Tree, and Back propagation, C 4.5, Classification, Naïve Bayes and Regression Tree were compared to mark the merits and demerits of these algorithms. When compared C 4.5 gave good performance when compared with other algorithms. Sugumaran, Shankar Sowmein, Karthikeyan C. P. and .Vijayaram T. R [15] predicted liver disease using the C.4.5 decision tree, which provides accurate and good results in predicting diseases in liver. Mr. S. Dhayanand and Dr. S. Vijayarani [16] used Naïve Bayes and SVM methods for predicting liver disease. Dataset in Indian liver, when cross checked the accuracy and efficiency of algorithms. it was concluded that than Naïve Bayes which takes minimum time for execution, SVM gives better accuracy?

VIII. DATA MINING FOR KIDNEY DISEASE PREDICTION

A condition where Kidney cannot filter blood means that our kidney is damaged ,where by waste spreads up in the human body. Kidney damage in many people occurs at a slow pace for years; this is often due to high blood pressure or diabete. The chronic kidney disease is growing every day as a problem [17]. Chronic kidney disease affects nearly 10 percent of the worldwide population. Without affordable treatment many people die each year. Chronic kidney disease was ranked 27th by Global Burden of Disease study, during 2010.various factors which causes disease in kidney are obesity , cigarette smoking, cholesterol, autoimmune disease which cause bladder obstruction, cirrhosis ,atherosclerosis, and liver failure and even bladder cancer. [18]. Mr. S. Dhayanand and Dr. S. Vijayarani, [19] used SVM for predicting kidney disease. Classification accuracy was compared and execution time of these algorithms was compared. Out of comparison SVM took minimum time for execution and ANN provided accurate classification

Narendra Ku. Kamila and Lambodar Jena [20] used UCI repository dataset of chronic kidney disease. Decision trees, Naïve Bayes were used for comparing accuracy classification. For predicting chronic kidney diseases, a multilayerperceptron algorithm was used which gave better classification accuracy and prediction. Hajar Mousannif, Basma Boukenze, and AbdelkrimHaqiq [21] focused on big data evolution in the healthcare system. SVM, Bayesian and Decision Tree used a dataset from UCI Repository to predict chronic Kidney patients with kidney failure. Better accuracy and minimum execution time gave results with better execution time and good accuracy. PushpaPatil .M [22] using data mining classifiers surveyed various research papers in predicting chronic kidney disease. Classifiers such as Bayes , Rule based , Decision Tree, Back Propagation, were also presented. Multilayer Preceptor, K-Nearest Neighbor Classifiers gives results with higher accuracy.

IX. DATA MINING FOR DIABETES PREDICTION

When the pancreas does not produce insulin, chronic diabetes is formed. A hormone which regulates sugar in blood is Insulin. A raised blood sugar and Hyperglycemia, is caused by unmanageable diabetes and it leads to significant issues in our human body especially blood vessels and nerves. There is no specific age group for diabetes. Nearly 1.5 million people were dead in 2012 due to diabetes. There are many risk factors for diabetes which include environmental factors , weight, family history, dietary factors and inactivity and abnormal level of cholesterol[23]. Amit Kumar Dewangan and Pragati Agarwal [24] used data mining techniques in diagnosing diabetes Mellitus. Methods used are classification are Support Vector Machine [SVM],k fold cross validation, class wise K- Nearest Neighbor [C KNN], They also presented that, on diabetic dataset SVM gives better accuracy. Prof. Suvarna Pawar andMsNilamChandgude [25] presented in their paper that different classification algorithms were used for diagnosing diabetes. Support Vector Machine, CART algorithms, Decision Tree, , ID3, Naïve Bayes, C 4.5 were compared in this paper, it was found that for better accuracy CART is used than other algorithms . Nagarajan .N and Thirumal P. C.[26] used various data mining applications to predict mellitus diabetes. Dataset used for analysis is mellitus The Pima Indians diabetes. After data pre-processing Naïve Bayes Classifier, algorithms such as KNN gives better exactness than K Nearest Neighbour provided with bottom most accuracy. Dr. Dhenakaran and K. Rajalakshmi [27] used for predicting issues in the health management system. Approaches such as Neural Network , SVM , Decision Tree and Bayesian Classifiers were also presented. The SVM algorithm performed better than various other algorithms in predicting diabetics.

Table1 expresses the algorithm used for comparing diseases.

Algorithm disease	Naives Bayes	Neural Network	Fuzzy	SVM	Decision Tree
Cancer	✓	✓			
Hheart		✓		✓	
Liver			✓		
Kidney		✓	✓	✓	✓
Diabetics	✓	✓	✓	✓	✓

X. CONCLUSION

This review paper aims to analyze medical data by using Data Mining for predicting diseases. From the analysis, it is noted that answers may differ based on techniques and tools used. Good results are provided by Data Mining in diagnosing diseases.

REFERENCES

[1] Fayyad, Piatetsky-shapiro,Smyth,"from Data mining to knowledge Discovery: An overview",pp.1-34,CA1996.

[2] P. Ahamad,SaqibQamar,S.Q.A Rizvi "Techniques of Data Mining in Healthcare:A Review" ,vol.120(15) pp. 38-50,June 2015

[3] Parvathi I, Siddharth Rautaray, —Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain ‖ , International Journal of Computer Science and Information Technologies, Vol. 5 (1), 838-846, ISSN: 0975- 9646, 2014.

[4] KamnaSolanki ,ParulBerwal,Sudhir —Analysis of Application of Data Mining Techniques in Health care, International Journal of Computer Applications, Vol. 148 (2), August 2016.

[5] V.Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, Vol.4 (1), ISSN: 0975-9646, 2013.]

[6] Jothi Prabha A, A.Govardhan, "Application of Classification Techniques on Various Attributes of Breast Cancer", Vol. 4, Issue 6, and ISSN (Online): 220-9801, ISSN (Print): 2320- 9798, June 2016.

[7] Durairaj M, Deepika R, "Prediction of Acute Myeloid Leukemia Cancer Using Data Mining- A Survey ‖ , Volume I, Issue 2, ISSN: 2394 – 6598, February 2015. [8] Jothi Prabha A, A.Govardhan, "Application of Classification Techniques on Various Attributes of Breast Cancer", Vol. 4, Issue 6, and ISSN (Online): 2320-9801, ISSN (Print): 2320- 9798, June 2016.

[8] Vikas Chaurasia, Saurabh Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", Vol. 3, Issue. 1, ISSN 2320–088X, January 2014.

[9] www.heart.org/idc/groups/ahamahpublic/@wcm/@sop/@smd/documents/downloadable/ucm480086.pdf

[10] T. Revathi, S. Jeevitha, — Comparative Study on Heart Disease Prediction System Using Data Mining Techniques ‖ , Volume 4 Issue 7, ISSN (Online): 2319-7064, July 2015.

[11] K.Manimekalai, —Prediction of Heart Diseases using Data Mining Techniques ‖ , International Journal of Innovative Research in Computer andCommunication Engineering, Vol. 4, Issue 2, ISSN(Online):2320-9801, ISSN (Print):2320- 9798, February 2016.

[12] 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, Nov-2015.

[13] A.S.Aneeshkumar, Dr. C.JothiVenkateswaran, "A novel approach for Liver disorder Classification using Data Mining Techniques", Engineering and Scientific International Journal (ESIJ), Volume 2, Issue 1, ISSN 2394- 7179 (Print), ISSN 2394-7187 (Online), January - March 2015.

[14] D.Sindhuja, R. JeminaPriyadarsini, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", International Journal of Computer Science and Mobile Computing,

Vol.5, Issue.5, ISSN 2320– 088X, May 2016.

[15] V.Shankarsowmien, V.Sugumaran, C.P.Karthikeyan, T.R.Vijayaram, "Diagnosis of Hepatitis using Decision tree algorithm", International Journal of Engineering and Technology (IJET), Vol 8 No 3, e-ISSN : 0975-4024, p-ISSN : 2319-8613, Jun-Jul 2016.

[16] Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction using SVM and ANN algorithms", International Journal of Computing and Business Research (IJCBR), Volume 6, Issue 2, ISSN (Online):2229-6166, March 2015.

[17].www.niddk.nih.gov/health-information/healthcommunication-programs/nkdep/learn/causes-kidneydisease/pages/disease-basics.asp

[18] www.kidney.org/kidneydisease/global-facts-about-kidneydisease/

[19] Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction using SVM and ANN algorithms", International Journal of Computing and Business Research (IJCBR), Volume 6, Issue 2, ISSN (Online):2229-6166, March 2015.

[20] Lambodar Jena, Narendra Ku. Kamila, "Distributed Data Mining Classification Algorithms for Prediction of ChronicKidney-Disease", International Journal of Emerging Research in Management &Technology, Volume-4, Issue-11, and ISSN: 2278-9359, November 2015.

[21] Basma Boukenze, Hajar Mousannif and AbdelkrimHaqiq, "Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease", International Journal of Database Management Systems (IJDMS), Vol.8, No.3, June 2016.

[22] Pushpa M. Patil, "Review on Prediction of Chronic Kidney Disease using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 5, ISSN 2320–088X, Issue. 5, May 2016.

[23] www.who.int/medicentre/factsheets

[24] Pragati Agrawal, Amit kumarDewangan, "A Brief Survey on the Techniques used for the Diagnosis of Diabetes-Mellitus"International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 03, e-ISSN: 2395 - 0056, p-ISSN: 2395-0072, June 2015.

[25] Ms. Nilamchandgude, Prof. Suvarna pawar, "A survey on diagnosis of diabetes using various classification algorithm", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 12, ISSN: 2321-8169, 6706 – 6710, December 2015.

[26] Thirumal P. C, Nagarajan N, ―Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus- A Case Study", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 1, ISSN 1819-6608, January 2015.

[27] K. Rajalakshmi, Dr. S. S. Dhenakaran, "Analysis of Data mining Prediction Techniques in Healthcare Management System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, ISSN: 2277 128X, April 20.