Survey on Data Mining Algorithms in Disease Prediction

V.Kirubha[1], S.Manju Priya[2]
[1] Research Scholar, Department of Computer Science, Karpagam University,Coimbatore, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India

Abstract*—Data mining is the process of extracting hidden interesting patterns from massive database. Medical domain contains heterogeneous data in the form of text, numbers and images that can be mined properly to provide variety of useful information for the physicians. The patterns obtained from the medical data can be useful for the physicians to detect diseases, predict the survivability of the patients after disease, severity of diseases etc. The main aim of this paper is to analyse the application of data mining in medical domain and some of the techniques used in disease prediction.*

Keywords — *Datamining, medical data, disease prediction.*

I. INTRODUCTION

Data Mining is the discovery of knowledge in databases. Techniques of data mining help to process the data and turn them into useful information. Prediction results from data mining are useful in various fields like Business Intelligence, Bioinformatics, Healthcare Management, Finance etc. Medical field has wide amount as well as variety of data for processing and there exist many challenging tasks. This field requires accurate and timely mannered diagnosis which can save many patients life. Data mining techniques plays a vital role in healthcare analysis.Early detection and accurate results are achievable by physicians using data mining algorithms. Different algorithms will be used for varied disease diagnosis. Based on the data used the accuracy and performance also vary.

II. RELATED WORK

Parvathi I and Siddharth Rautaray [1] presented a hybrid methodology by combining classification tree and association rule mining techniques. They have discussed about the application of data mining in various fields and the data mining techniques used for analysis. Advantages and disadvantages of data mining in medical domain and the algorithms used for medical diagnosis have been explained.

In the paper proposed by Dhanya P Varghese and Tintu P B [2], the data mining classification techniques used on medical system and also the various papers presented on medical data mining using classification techniques are discussed. They have also explained the importance of data mining in healthcare domain.

Vahid Rafe and Roghayeh Hashemi Farhoud [3] examined the importance and the data mining techniques in medicine. They have also discussed about applications of data mining in medicine.

III. MEDICAL DATA MINING

Data mining has been used to uncover patterns from the large amount of stored information and then used to build predictive models. Medical field contains large amount of data that are needed to be processed. Data mining in medical field improves the quality of patient care and the prediction of healthcare patterns. Data mining tools helps us to discover unknown patterns, group the related items and decision making of healthcare oriented problems. Medical care is necessity; it gives patient and hope for a fruitful life. The collected data when published is used forsocial causes without harming the dignity of the patients. Early detection of disease can increase the survivability of patients[1]. Data mining techniques such as classification and prediction, clustering, association rule mining and various mining methods can be useful to apply on medical data. The collected data in the form of images can also be used to mine healthcare data. Nowadays, many image mining techniques improved the disease prediction and health care decision making task as easiest. This paper presents the application of data mining algorithms in field of medicine.

IV. HEART DISEASE PREDICTION

Heart disease is an umbrella term for any type of disorder that affects the heart. Heart disease means the same as cardiac disease but not cardiovascular disease. Cardiovascular disease refers to disorders of the blood vessels and heart, while heart disease refers to just the heart.According to WHO (World Health Organization) and the CDC, heart disease is the leading cause of death in the UK, USA, Canada and Australia[4].

The association's 2015 Heart disease and Stroke update compiled by the American Heart Association, the Centres for Disease Control and Prevention, the National Institutes of Health and other government sources presents that Cardiovascular disease is the leading global cause of death, accounting for 17.3 million deaths per year, a number that is expected to grow to more than 23.6 million by 2030. From 2001 to 2011, the death rate from heart disease has fallen about 39 % – but the burden and

risk factors remain alarmingly high. The risk factors of heart disease are smoking, overweight, cholesterol, high blood pressure, diabetes, unhealthy diet and physical activity[5].

T. Revathi and S. Jeevitha [6] analysed 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 [7] proposed a heart disease prediction system using Naïve Bayes and compared the results with Neural Network and Decision Tree algorithms. According to that method, Naïve Bayes algorithm provides good prediction.

K. Manimekalai [8] 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, J 4.8, decision tree and Fuzzy mechanism algorithms.

Jyoti Rohilla and Preeti Gulia [9] analysed some of the data mining algorithms to predict heart disease. They have used a heart disease dataset from UCI machine learning repository and analysed using WEKA tool, shown that decision tree algorithms performed well in predicting heart disease.

V. KIDNEY DISEASE PREDICTION

Kidney disease means that the kidneys are damaged and can't filter blood like they should. This damage can cause wastes to build up in the body. For most people, kidney damage occurs slowly over many years, often due to diabetes or high blood pressure. This is called chronic kidney disease. When someone has a sudden change in kidney function—because of illness, or injury, or have taken certain medications—this is called acute kidney injury. This can occur in a person with normal kidneys or in someone who already has kidney problems. Kidney disease is a growing problem[10]. 10% of the population worldwide is affected by chronic kidney disease (CKD), and millions die each year because they do not have access to affordable treatment.

According the 2010 Global Burden of Disease study, chronic kidney disease was ranked 27th in the list of causes of total number of deaths worldwide in 1990, but rose to 18th in 2010. Risk factors of kidney disease are cigarette smoking, obesity, high cholesterol, diabetes (types 1 and 2), autoimmune disease, obstructive kidney disease, including bladder obstruction caused by benign prostatic hyperplasia (BPH), atherosclerosis, cirrhosis and liver failure, narrowing of the artery that supplies your kidney, kidney cancer, bladder cancer, kidney stones, kidney infection[11].

Dr. S. Vijayarani, Mr S. Dhayanand [12] focused on predicting kidney diseases and analysed the prediction of kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network. They have compared the classification accuracy and execution time of these algorithms. And results shown that, ANN achieves accurate classification performance and SVM has taken minimum execution time.

Lamboder Jena, Narendra Ku. Kamila [13] analysed a chronic kidney disease dataset from UCI machine learning repository. They have used algorithms such as Naïve Bayes, Multilayer Perceptron, SVM, J48, Conjunctive rule and Decision tree for comparing the classification accuracy. And presented that multilayer perceptron algorithm gives better classification accuracy and prediction performance to predict chronic kidney diseases.

Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq [14] 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.

Pushpa M. Patil [15] surveyed on various research papers on prediction of chronic kidney disease using data mining classifiers. Basic concepts of Decision Tree, Bayes classification, Rule based classification, Back Propagation algorithm, Support Vector Machine (SVM), K-Nearest Neighbour classifier are presented. Classifiers with higher accuracy are Multilayer Perceptron, Random forest, Naïve Bayes, SVM, K-Nearest Neighbour and Radial Basis Function.

VI. LIVER DISEASE PREDICTION

Liver disease is any disturbance of liver function that causes illness. The liver is responsible for many critical functions within the body and should it become diseased or injured, the loss of those functions can cause significant damage to the body. Liver disease is also referred to as hepatic disease. Liver disease is the fifth 'big killer' in England & Wales, after heart, cancer, stroke and respiratory disease[16].

The World Health Organisation (WHO) reports that approximately 3% of the world's population are infected with hepatitis C. 170 million people are chronically infected and 3-4 million are newly infected each year. The World Health Organisation (WHO) reports that approximately 3% of the world's population are infected with hepatitis C. 170 million people are chronically infected and 3-4 million are newly infected each year. An estimated five out of every six people with chronic hepatitis C are unaware of their infection. According to the Health Protection Agency, around 191,000 people aged 15-59 have hepatitis C (2003) and 142,000 in this age group have chronic disease [17]. Risk factors of liver disease are

alcoholism, autoimmune diseases, and exposure to toxins, hereditary conditions and viruses [18].

P. Sindhuja and R. Jemina Priyadarshini, [19] in their paper, described classification techniques for analysing liver disorder. And advantages and disadvantages of algorithms such as C 4.5, Naïve Bayes, Decision Tree, Support Vector Machine, Back propagation and Classification and Regression Tree are compared. They have presented that C 4.5 gives better performance than other algorithms.

A. S. Aneeshkumar, C. Jothi Venkateshwaran, [20] in their paper, presented an approach for liver disorder classification using data mining techniques. They have usedFuzzy based classification and achieved better accuracy for diagnosis of liver disorder.

According to the study on diagnosis of Hepatitis using decision tree algorithm by V. Shankar Sowmein, V. Sugumaran, C. P. Karthikeyan, T. R. Vijayaram[21] the C 4.5 decision tree algorithm provides an efficient result in liver disease prediction.

Dr. S. Vijayarani, Mr S. Dhayanand [22] presented a method for liver disease diagnosis. They have used SVM and Naïve Bayes approach and the Indian Liver Dataset is used for their study. Compared the performance of both the algorithm and found that SVM gives higher accuracy and Naïve Bayes has taken minimum execution time.

VII.DIABETES PREDICTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2012 diabetes was the direct cause of 1.5 million deaths and high blood glucose was the cause of another2.2 million deaths. Risk factors of diabetes include family history, environmental factors, dietary factors, weight, inactivity, high blood pressure, abnormal cholesterollevels [23].

Pragati Agarwal, Amit kumar Dewangan [24] concentrated in the diagnosis of diabetes Mellitus using data mining techniques. They have analysed k-fold cross validation, classification method, class wise K- Nearest Neighbour[CKNN], Support Vector Machine [SVM], LDA Support Vector Machine and Feed Forward Neural Network, Artificial Neural Network, Statistical Normalization and Back propagation methods for diabetic diagnosis. And presented that, SVM gives better accuracy on diabetic dataset.

Ms Nilam Chandgude and Prof. Suvarna pawar, [25] in their paper presented the different classification algorithm used on the diagnosis of diabetes. They used neural network, Decision Tree, Naïve Bayes, Support Vector Machine, ID3, C 4.5, CART algorithms and compared the performance of these algorithm. And found that CART provides better accuracy than other algorithm.

Thirumal P. C. and Nagarajan .N [26] presented various data mining techniques to predict diabetes mellitus. The Pima Indians diabetes dataset is used for analysis. After pre-processing the data, algorithms such as Naïve Bayes Classifier, C4.5 algorithm, SVM, KNN are applied. C4.5 algorithm provided higher accuracy and KNN provided lower accuracy.

K. Rajalakshmi and Dr. S. S. Dhenakaran [27] analysed data mining prediction techniques in healthcare management systems. Data mining techniques such as Decision Tree, Bayesian Classifier, Neural Network and SVM are presented. Various data mining techniques are compared on different disease prediction. SVM algorithm performed well on predicting diabetic.

VIII. CANCER PREDICTION

Cancer is a generic term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumours and neoplasms. One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organs, the latter process is referred to as metastasizing. Metastases are the major cause of death from cancer. Cancers figure among the leading causes of morbidity and mortality worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths in 2012. The most common causes of cancer death are cancers of: lung (1.59 million deaths), liver (745 000 deaths), stomach (723 000 deaths), colorectal (694 000 deaths), breast (521 000 deaths), oesophageal cancer (400 000 deaths). Tobacco use, being overweight or obese, unhealthy diet with low fruit and vegetable intake, lack of physical activity, alcohol use, urban air pollution, indoor smoke from household use of solid fuels are the major risk factors of cancer[28].

Vikas Chaurasia and Saurabh Pal [29] concentrated on the detection of breast cancer through a diagnosis system based on RepTree, RBF Network and Simple Logistic. They have used the data provided by the University Medical Centre, Institute of Oncology, and WEKA tool for experiments. Simple Logistic algorithm achieved 74.47% accuracy for diagnosis of breast cancer.

V. Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra [30] examined the data mining classification techniques on diagnosis of Lung cancer. They have presented the symptoms and risk factors of lung cancer and explained the data mining concepts. The algorithms such as Rule set classifiers, Decision Tree, Neural Network, and Bayesian Network are used for analysis. From the comparison results, found that Naïve Bayes is better than other algorithms used.

Durairaj M and Deepika R [31] presented review about the prediction of Myeloid Dysplastic Syndrome (MDS) and Acute Myeloid Leukaemia (AML) pathogenesis. The Microarray technique in data mining is also explained. They have reviewed ten papers on disease diagnosis. Compared the accuracy of algorithms using WEKA tool and found that, 1BK and Decision table provides accurate results.

Jothi Prabha A and A. Govardhan [32] presented the application of classification techniques on various attributes of Breast Cancer. Attributes of breast cancer disease by observing the intensity levels of each attribute are reviewed. Naïve Bayes classification, J48 decision tree algorithm and chronic disease dataset is used for analysis. Naïve Bayes classification provided better results.

The below table 1 shows the comparison of algorithms used for disease prediction.

TABLE I: COMPARISON OF ALGORITHMS

Disease / Algorithm	Heart	Kidney	Liver	Diabetes	Cancer
Decision Tree	√	√	√	√	√
Naïve Bayes	√				√
Neural Networks	√	√			
Fuzzy			√		
SVM	√	√	√	√	
Multilayer Perceptron		√			√
Simple Logistic				√	

IX. CONCLUSION

This paper aimed to analyse the application of data mining in medical domain and some of the algorithms used to predict diseases. It is observed that results may vary for different disease diagnosis based on the tools and techniques used. Data mining provides good results in disease diagnosis when appropriate tools and techniques applied. Hence data mining is the promising field for healthcare predictions.

REFERENCES

[1] 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.

[2] Dhanya P Varghese, Tintu P B, "*A Survey on Health Data using Data Mining Techniques*", International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Oct-2015.

[3] Vahid Rafe, Roghayeh Hashemi Farhoud, "*A Survey on Data Mining Approaches in Medicine*", International Research Journal of Applied and Basic Sciences, Vol 4 (1), ISSN 2251-838X, 2013.

[4] www.medicalnewstoday.com/articles/237191.php

[5] www.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm-480086.pdf

[6] 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.

[7] 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.

[8] K.Manimekalai, "P*rediction 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.

[9] 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, July 2015.

[10] www.niddk.nih.gov/health-information/health-communication-programs/nkdep/learn/causes-kidney-disease/pages/disease-basics.asp

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

[12] 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.

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

[14] Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq, *"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.

[15] 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.

[16] www.medicinenet.com

[17] www.britishlivertrust.org.uk/about-us/media-centre/facts-about-liver-disease/

[18] www.healthcommunities.com/liver-disease/causes.html

[19] D.Sindhuja, R. Jemina Priyadarsini, *"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.

[20] A.S.Aneeshkumar, Dr. C.Jothi Venkateswaran, *"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.

[21] V.Shankar sowmien, 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.

[22] Dr. S. Vijayarani, Mr.S.Dhayanand, *"Liver Disease Prediction using SVM and Naïve Bayes Algorithms*", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, ISSN: 2278 – 7798, April 2015.

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

[24] Pragati Agrawal, Amit kumar Dewangan, *"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. Nilam chandgude, 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 Datamining 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 2015.

[28] www.who.int/mediacenter/factsheets/fs297

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

[30] 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.

[31] Durairaj M, Deepika R, *"Prediction of Acute Myeloid Leukemia Cancer Using Data Mining- A Survey*", Volume I, Issue 2, ISSN: 2394 – 6598, February 2015.

[32] 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.