International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 10 3003 – 3008

Review on Heart Disease Prediction System using Data Mining Techniques

Beant Kaur
Research Scholar,
Dept of Computer Engineering,
Punjabi University, Patiala, India

Williamjeet Singh
Asst. Professor,
Dept of Computer Engineering,
Punjabi University, Patiala, India

Abstract: Data mining is the computer based process of analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict future trends, allowing business to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally taken much time consuming to resolve. The huge amounts of data generated for prediction of heart disease are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. By using data mining techniques it takes less time for the prediction of the disease with more accuracy. In this paper we survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. Result from using neural networks is nearly 100% in one paper [10] and in [6]. So that the prediction by using data mining algorithm given efficient results. Applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.

Keywords: *- Heart disease, Data mining, Data mining techniques*

*****

1. Introduction

The main objective of our paper is to learn the different techniques of data mining used in prediction of heart disease by using different data mining tools. Life is dependent on efficient working of heart because heart is essential part of our body. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. Heart disease is a disease that affects on the operation of heart. There are number of factors which increases risk of Heart disease. Nowadays, in the world Heart disease is the major cause of deaths. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. In 2008, 17.3 million people died due to Heart Disease. Over 80% of deaths in world are because of Heart disease. WHO estimated by 2030, almost 23.6 million people will die due to Heart disease as written in [10]. Prediction by using data mining techniques gives us accurate result of disease. IHDPS (intelligent heart disease prediction system) can discover and extract hidden knowledge associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus help healthcare analysts and practitioners to make intelligent clinical decisions which traditional decision support systems cannot. In this paper analysis of various data mining techniques given in tables which were used and helpful for medical analysts or practitioners for accurate heart disease diagnosis.

1.1 The risk factor for heart disease

Family history of heart disease: - most people know that the heat disease can run in families. That if anybody has a family history of heart disease, he/she may be at greater risk for heart attack, stroke and other heard diseases.

Smoking: - smoking is major cause of heart attack, stroke and other peripheral arterial disease. Nearly 40% of all people who die from smoking tobacco do so due of heart and blood vessel diseases. A smoker's risk of heart attack reduces rapidly after only one year of not smoking.

Cholesterol: - abnormal levels of lipids (fats) in the blood are risk factor of heart diseases. Cholesterol is a soft, waxy substance found among the lipids in the bloodstream and in all the body's cells. High level of triglyceride (most common type of fat in body) combined with high levels of LDL (low density lipoprotein) cholesterol speed up atherosclerosis increasing the risk of heart diseases.

High blood pressure: - High blood pressure also known as HBP or hypertension is a widely misunderstood medical condition. High blood pressure increase the risk of the walls of our blood vessels walls becoming overstretched and injured. Also increase the risk of having heart attack or stroke and of developing heart failure, kidney failure and peripheral vascular disease.

Obesity:-the term obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Being obese puts anybody at a higher risk for health problem such as heart disease, stroke, high blood pressure, diabetes and more.

Lack of physical exercise: -lack of exercise is a risk factor for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD, because it also increases the risk for diabetes and high blood pressure.

2. Literature Survey

Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today's time. This paper aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction. Table 1 shows different data mining techniques used in the diagnosis of Heart disease over different Heart disease datasets. In some papers this is given that they use only one technique for diagnosis of heart disease as given in Shadab et al [12**],** Carlos et al **[5]** etc. but in case of other research work more than one data mining techniques are used

for the diagnosis of heart disease as given in Ms. Ishtake et al.[3] , MA.JABBAR, et al[2], Shantakumar et al[7] etc.

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

Author	***Year***	***Technique Used***	***attributes***
Carlos et al	2001	association rules	25
Dr. K. Usha Rani	2011	Classification	13
		Neural Networks	
Jesmin Nahar , et al	2013	Apriori	14
		Predictive Apriori	
		Tertius	
Latha et al.	2008	genetic algorithm	14
		CANFIS	
Majabbar et al	2011	*Clustering*	14
		Association rule mining,	
		Sequence number,	
Ms. Ishtake et al.	2013	Decision Tree	15
		Neural Network	
		Naive Bayes	
Nan-Chen et al	2012	(EVAR)	
		Machine learning	
		Markov blanket	
Oleg et al.	2012	artificial neural network	
		genetic polymorphisms	
Shadab et al	2012	Naive bayes	15
Shantakumar et al	2009	MAFIA	13
		Clustering	
		K-Means	

2.1. Data Mining

Data Mining is main concerned with the analysis of data and Data Mining tools and techniques are used for finding patterns from the data set. The main objective of Data Mining is to find patterns automatically with minimal user input and efforts. Data Mining is a powerful tool capable of handling decision making and for forecasting future trends of market. Data Mining tools and techniques can be successfully applied in various fields in various forms. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Mining tools and techniques, various fields of business get benefit by easily **evaluate** various trends and pattern of market and to produce quick and effective market trend analysis. Data mining is very useful tool for the diagnosis of diseases.

2.2. Techniques used in data mining

A. Association: - Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in heart disease prediction as it tell us the relationship of different attributes used for analysis and sort out the patient with all the risk factor which are required for prediction of disease.

B. Classification: -Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

C. Clustering: -Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. For example In prediction of heart disease by using clustering we get cluster or we can say that list of patients which have same risk factor. Means this makes the separate list of patients with high blood sugar and related risk factor n so on.

D. Prediction: - The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

2.3. Comparative statement

The following table presents the comparative statement of various data mining trends from past to the future taken from Venkatadr et al[32].

Table 2: Table presents the comparative statement of various data mining trends from past to the future**.**

Data mining Trends	*Algorithms/ Techniques Employed*	*Data formats*	*Computing Resources*
Past	Statistical, Machine	Numerical data and	Evolution of 4G PL and

	Learning Techniques	structured data stored in traditional databases	various related techniques
Present	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats includes structured, semi structured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi instance objects etc.	Multi-agent technologies and Cloud Computing

2.4 Data Mining used in Various Applications some of them are given below

Business intelligence: Business intelligence is a set of theories, methodologies, architectures, and technologies that transform raw data into meaningful and **useful** information for business purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new opportunities. BI, in simple words, makes interpreting voluminous data friendly. Making use of new opportunities and implementing an effective strategy can provide a competitive market advantage and long-term stability. BI technologies provide historical, current and predictive views of business operations.

Sports: -Sports are ideal for application of data mining tools and techniques. In the sports world the **vast** amounts of statistics are collected for each player, team, game, and season. Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning

Analyze Students Performance: - The classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's management system, to predict the performance at the end of the semester. This paper investigates the accuracy of Decision tree techniques for **predicting** student performance.

Telecommunication Industry: - Telecommunication services have grown from local and long distance voice communication services to fax, pager, cellular phones and e-mails. Now the **telecommunication** services have integrated with the computer, internet, and network and with other communication technologies. Due to **the** advancements in telecommunication technologies and to work these technologies effectively, Data Mining techniques integrated with these technologies to produce effective results. Data Mining helps to identify telecommunications patterns, fraud activities and also helps to better use of resources and improve the quality of services.

Retail Industry: -Data Mining plays an important role in the retail industry also. Retail industry involves large amount of data that includes transportation, sales and consumptions of goods and services. This data grows rapidly due to increase in purchase and sales in business. These days, E-commerce is growing fast with the growth of companies and also improving the online experience. Electronic commerce describes the buying and selling of products, services, and information via computer networks including the Internet.

Table 3: Table shows 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%	
	J48		94.85%	
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	.NET	96.5%,	

	networks	platform	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

2.5 Open source tools for data mining

WEKA Tool: -WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state of- the-art facility for developing machine learning techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a 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 with this package. WEKA is open source software issued. The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file.

TANAGRA: -Tanagra is free data mining software for academic and research purposes. It proposes several data mining method from exploratory data analysis, statistical learning, machine learning and database area. Tanagra is an open source project as every researcher can access to the source code and add his own algorithms, as far as he agrees and conforms to the software distribution license. The main purpose of Tanagra project is to give researchers and students an easy to use data mining software, conforming to the present norms of the software development in this domain and allowing to analyze either real or synthetic data.

MATLAB: - MATLAB is a high 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 us to explore multiple approaches and reach a solution faster than with spreadsheets of traditional programming languages, such as C/C++ of JAVA.

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. This is packed with features for data analytics.

.NET Framework: -.net framework is a software framework developed by Microsoft 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 and seamless and secure communication.

RapidMiner: -RapidMiner 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. Thousand of applications of RapidMiner in more than 40 countries give their users a competitive edge.

Table 4: Table shows heart disease dataset using different data mining techniques

Author	Year	Technique	Accuracy
Chaitrali et al,	2012	Naive Bayes	90.74%
		DT	99.62%
		NN	100%
Indira S. Fal Dessai	2013	PNN	94.6%
		DT	84.2%
		NB	84%
		BNN	80.4%
Jesmin et al	2013	Naive Bayes	92.08%
		SMO	96.04%
		IBK	95.05%
		AdaBoostM1	96.04%
		J48	96.04%
		PART	96.04%
M. anbarasi et al	1999	Naive Bayes	96.5%
		Decision Tree	99.2%
		Classification via clustering Naive Bayes	88.3%
Matjaz et al	1999	exercise ECG(NN) exercise ECG(NN)	74%
		myocardial scintigraphy(NN)	85%
N. Aditya Sundar et al.,	2012	WAC	84%
		Naïve bayes	78%
T. John et al.	2012	Naïve bayes	85.18%
		Multilayer	78.88%

		J48	85.18%
		KNN	85.55%
Tanawut et al	2008	BNN	74.5%
		DK-SOM	80.4%

2.6 Methodology Used in Data Mining

Data Mining is core part of Knowledge Discovery Database (KDD). Many people treat Data Mining as a synonym for KDD since it's a key part of KDD process. Knowledge discovery as a process is depicted in Figure 1 and consists of an iterative sequence of the following steps:

- ata Cleaning - To remove noise or irrelevant data.
- Data Integration - Where multiple data sources may be combined.
- Data Selection - Where data relevant to the analysis task are retrieved from the database.
- Data Transformation - Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining - An essential process where intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation - To identify the truly interesting patterns representing knowledge based on some interestingness measures.
- Knowledge Presentation - knowledge representation techniques are used to present the mined knowledge to the user.

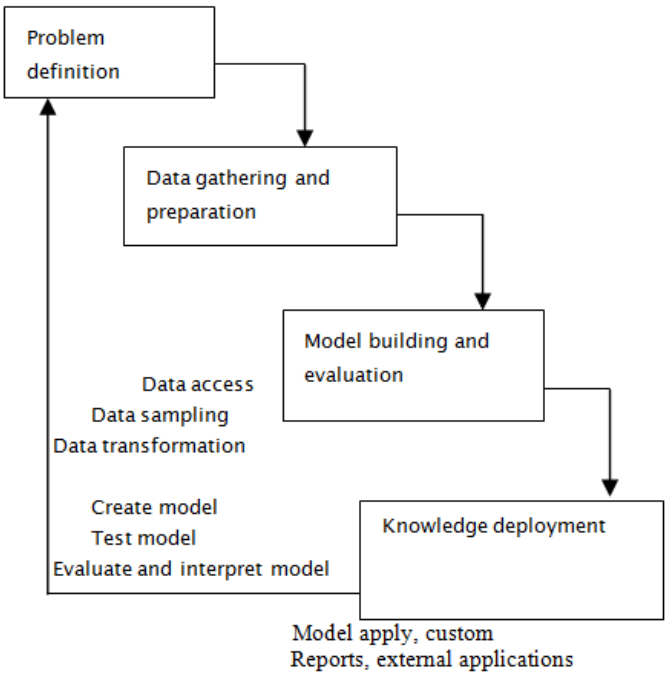

Figure 1:- Data mining

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repository.

3. Data mining techniques used in diagnosis of other diseases

Table5: Table shows research papers in which data mining techniques used for diagnosis of different diseases.

Author	Year	Disease	Technology
Humar et al.	2008	diabetes heart diseases	Classification, Backpropagation,Fuzzy neural network.
Marcel Et al	2007	Carcinoid heart disease	Bayesian classification
Mohammad et al	2012	breast cancer heart disease	C4.5, C5.0
M.Akhil et al.	2012	Pima Indian Diabetes Breast Cancer Heart Disease Data	Associative Classification and Genetic Algorithm

Data mining techniques used in various fields as given above like in student performance, in business, fraud detection, banking, marketing, law, insurance etc. in given papers data mining techniques used for prediction of more than one disease like in paper Mohammad et al [1] C4.5 and C5.0 data mining algorithms are used for prediction of heart disease and also for breast cancer. In paper Humar et al. [22] Classification, Backpropagation, Fuzzy neural network techniques used for diseases diabetes and heart diseases. And in case of M.Akhil et al [29] Associative Classification and Genetic Algorithm are used for three different diseases as Pima Indian Diabetes, Breast Cancer and Heart Disease Data.

4. Conclusion

The objective of our work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that different technologies are used in all the papers with taking different number of attributes. So, different technologies used shown the different accuracy to each other. In some papers it is shown that neural networks given the accuracy of 100% in prediction of heart disease. On the other hand, this is also given that Decision Tree

has also performed well with 99.62% accuracy by using 15 attributes [6]. So, different technologies used shown the different accuracy depends upon number of attributes taken and tool used for implementation. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of heart disease. Although applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success, the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention.

References

[1] 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.

[2] 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.

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

[4] Dr. K. Usha Rani,analysis of heart diseases dataset using neural network approach,International Journal of Data Mining & Knowledge Management Process, 2011.

[5] Carlos Ordonez, Edward Omiecinski, Mining Constrained Association Rules to Predict Heart Disease, IEEE. Published in International Conference on Data Mining (ICDM), p. 433-440, 2001.

[6] Nidhi Bhatla Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT), 2012.

[7] 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.

[8] Abhishek taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[9] 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,2010.

[10] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology, 2012.

[11] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra,performance analysis of classification data mining techniques over heart diseases data base, international journal of engineering science and advanced technology, 2012.

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

[13] Latha Parthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Medical Sciences, 2008.

[14] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Elsevier, 2013.

[15] Nada Lavrac, Selected techniques for data mining in medicine, Elsevier, 1999.

[16] Tanawut Tantimongcolwat, Thanakorn Naenna, Identification of ischemic heart disease via machine learning analysis on Magnetocardiograms, Elsevier,2008.

[17] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, Effective diagnosis of heart disease through neural networks ensembles, Elsevier, 2009.

[18] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur Diagnosis of valvular heart disease through neural networks ensembles, Elsevier, 2009.

[19] Oleg Yu. Atkov, Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters, Elsevier, 2012.

[20] Marcel A.J. van Gerven, Predicting carcinoid heart disease with the noisy-threshold classifier, Elsevier, 2007.

[21] Matjaz' Kukar, Analysing and improving the diagnosis of ischaemic heart disease with machine learning, Elsevier, 1999.

[22] Humar Kahramanli, Novruz Allahverdi, Design of a hybrid system for the diabetes and heart diseases, Elsevier, 2008.

[23] Jesmin Nahar, Tasadduq Imam, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, Elsevier, 2013.

[24] Nan-Chen Hsieh & Lun-Ping Hung & Chun-Che Shih, Intelligent Postoperative Morbidity Prediction of Heart Disease Using Artificial Intelligence Techniques, J Med Syst, 2012.

[25] Adebayo Peter Idowu, Data Mining Techniques for Predicting Immunize-able Diseases: Nigeria as a Case Study, International Journal of Applied Information Systems, 2013.

[26] Monali Dey, Siddharth Swarup Rautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies(2014).

[27] T. John Peter, K. Somasundaram, study and development of novel feature selection framwork for Heart disease prediction, International Journal of Scientific and Research Publications, 2012.

[28] Indira S. Fal Dessai, Intelligent Heart Disease Prediction System Using Probabilistic Neural Network, International Journal on Advanced Computer Theory and Engineering, 2013.

[29] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu, Heart Disease Prediction System using Associative Classification and Genetic Algorithm, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012.

[30] Chaitrali S. Dangare,Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications, 2012.

[31] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

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