International Journal of Electrical, Electronics and Computer Engineering **8**(1): 40-45(2019)

ISSN No. (Online): 2277-2626

A Literature review of Diagnosis of Heart Disease using Data Mining Techniques

Leena Sarvaiya, Himanshu Yadav and Chetan Agrawal
Department of Computer Science Engineering,
RITS, Bhopal (Madhya Pradesh), INDIA

(Corresponding author: Leena Sarvaiya)

(Received 21 February 2019 Accepted 23 May, 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In health care industry, it produces huge amount of information which is not so easy to handle and getting essential information from them because there is no availability of its much amount of doctors. So extraction of essential information is vital task. Heart disease is one the disease which occurs to the most of the people nowadays and cure of this disease also available in the market. Data mining is one of the extensively used areas of research in this field. In this, a comprehensive review of literature work and data mining technique is discussing. The data mining techniques namely decision tree, SVM, Naïve Bayes and neural network etc.

Keyword: Data Mining, Heart Disease, SVM, Neural network.

I. INTRODUCTION

Heart disease is nothing but the class of diseases that involve the heart or blood vessels (arteries and veins). Today most countries face high and growing rates of heart disease and it has become a leading cause of debilitation and death worldwide in men and women over age sixty-five and today in many countries heart disease is viewed as a "second epidemic," replacing infectious diseases as the leading cause of death [1]. Most countries face high and increasing rates of heart disease or Cardiovascular Disease. Even though, modern medicine is generating huge amount of data every day, little has been done to use this available data to solve the challenges that face a successful interpretation of heart disease examination results. Data mining techniques applied on educational field increases day by day. Data mining techniques when applied to the educational environment named as Educational Data Mining. A research community of such emerging field involves student learning experiences by predicting student's performance. As compared [22] to all data mining techniques, classification is most important one. The techniques used for improving students performance and finding the finest formation of set of courses for existing environment is foremost plan behind this research.

Data mining is a knowledge discovery technique to analyze data and encapsulate it into useful information. The current research intends to predict the probability of getting heart disease given patient data set. Predictions and descriptions are principal goals of data mining, in practice. Prediction in data mining involves attributes or variables in the data set to find unknown or future state values of other attributes. Description emphasize on discovering patterns that explains the data to be interpreted by humans. The purpose of predictions in data mining is to help discover trends in patient data in order to improve their health [1]. Due to change in life styles in developing countries, like South Africa, Cardio Vascular Disease (CVD) has become a leading cause of deaths. CVD is projected to be a single largest killer worldwide accounting for all deaths. An endeavor to exploit knowledge, experience and clinical screening of patients to diagnose or recognize heart attacks is regarded as a treasured opportunity. In the health sectors data mining plays an important role to predict diseases. The predictive end of the research is a data mining model and figure 1 shows the stages involved in data mining.

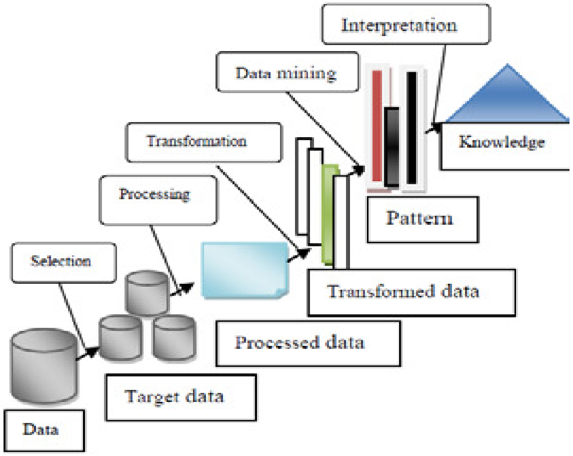

Fig. 1. Steps involved in data mining.

A. Applications of data mining to healthcare data

Data mining scholars have long studied the application of tools and equipment in improving the process of data analysis in large and complex datasets. Adopting data mining techniques in the medicine field is of high importance in diagnosing, predicting and deeply understanding of healthcare data. These applications include treatment centers analysis aimed at improving treatment policies and prevention of any mistake in hospitals, early diagnosis of diseases, prevention of diseases and hospital death reduction. Heart specialist's record and store large amounts of patients' data. This provides a great opportunity for extracting a valuable knowledge from such datasets. Researchers are adopting statistical approaches as well as data mining techniques to help treatment and healthcare specialists diagnose and determine heart disease risk factors in patients. Statistical analyses have identified a number of risk factors for heart diseases including age, blood pressure, smoking, total cholesterol, The number of diabetic patient is growing day by day and various causes has been discover to diagnosis the diabetes but early prediction of the cause is very indispensable to get rid of from this disease. Data Mining is a technique which plays an imperative role by unhidden the important data related to diabetes. These data can be utilized for rapid and improved clinical decision making for protective and [23] suggestive medicine. This paper proposes a hybrid approach SVM classifier using KNN and GA technique

which can effectively discover the effectual data to diagnose the diabetes disease. The simulation and experimental analysis of the propose approach is done using MATLAB toolbox and measuring parameter specificity, accuracy and sensitivity. The simulation results of proposed approach give improved value than the existing approach**.** diabetes, and hypertension, heart disease background in family, obesity and lack of physical activity. The awareness of heart disease risk factors assists treatment and healthcare specialists to identify patients who are subject to high risk factors.

II. FACTOR AFFECTING THE HEART

The circumstances or habits that make a person more likely to develop a disease are Risk factors. They can also boost the probability of an existing disease will get worse [2].

A. Controllable Risk Factors

Smoking: The chemicals in tobacco smoke promote the development of blood clots and increase the cause heart attacks by building-up of plaque in artery walls.

Weight: If body pound increases, the risk of heart disease also rises. This is especially factual for people who carry extra body fat around the waist. To reduce the risk of heart disease numerous dietary factors that can be used.

Cholestrol: Excessive cholesterol in the blood building up in the walls of the arteries can cause a process called atherosclerosis, a form of heart disease.

Diabetes: Diabetes can cause heart disease by growing the risk of high blood pressure and high cholesterol in the blood. It promotes injury to the artery walls and formation of blood clots.

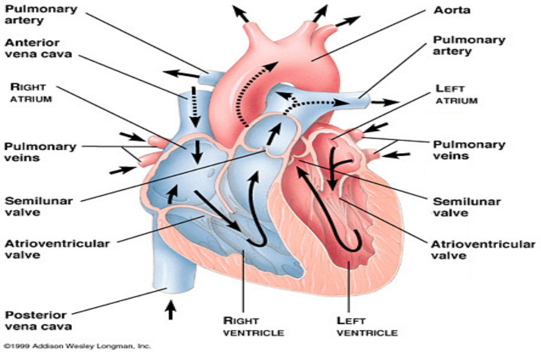

Fig. 2. Structure of heart.

Blood pressure: Blood pressure is the force of the blood against the inner walls of the blood vessels , generated when the heart pumps blood. When a person has hypertension, the arteries are under increased pressure and the heart has to pump harder, which may lead to injury of the artery walls, atherosclerosis, and coronary heart disease.

B. Uncontrollable Risk Factors

Age: Heart Related disease usually occurs in women after menopause and in men above the age of 40, and most people who die of heart attacks are above the age of 65.

Sex: Men have got more risk of heart attack than women, and men generally suffer from heart attacks at earlier ages.

Family history: For the person who is having a close relative who had heart attack may be at risk of heart disease.

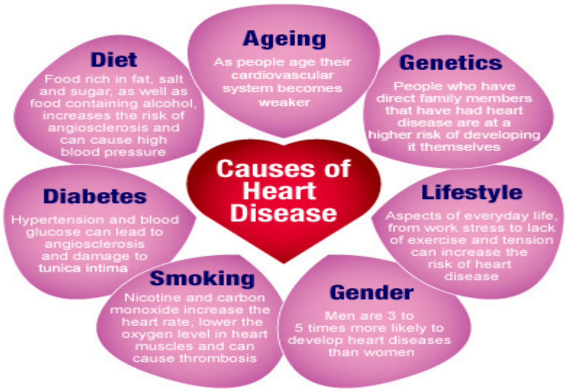
Diet
Food rich in fat, salt and sugar, as well as food containing alcohol, increases the risk of angiosclerosis and can cause high blood pressure
Ageing
As people age their cardiovascular system becomes weaker
Genetics
People who have direct family members that have had heart disease are at a higher risk of developing it themselves
Causes of Heart Disease
Diabetes
Hypertension and blood glucose can lead to angiosclerosis and damage to tunica intima
Lifestyle
Aspects of everyday life, from work stress to lack of exercise and tension can increase the risk of heart disease
Smoking
Nicotine and carbon monoxide increase the heart rate, lower the oxygen level in heart muscles and can cause thrombosis
Gender
Men are 3 to 5 times more likely to develop heart diseases than women

Fig. 3. Causes of Heart Attack.

III. HEART DISEASE TYPES

Although coronary heart disease (CAD) is the most common type of heart disease, there are several other types. Some of the other more common types of heart disease include abnormal heart rhythms (arrhythmias), heart failure, heart valve disease, heart muscle disease, and congenital heart disease.

(i) Coronary artery disease (CAD) is the most common type of heart disease. In CAD, the arteries carrying blood to the heart muscle (the coronary arteries) become lined with plaque, which contains materials such as cholesterol and fat. This plaque buildup (called atherosclerosis) causes the arteries to narrow, allowing less oxygen to reach the heart muscle than it needs to work properly. When the heart muscle does not receive enough oxygen, chest pain (angina) or heart attack can occur.

(ii) Heart valve disease occurs when one or more of the four valves in the heart are not working properly. Heart valves help to ensure that the blood being pumped through the heart keeps flowing forward. Disease of the heart valves (e.g., stenosis, mitral valve prolapse) makes it difficult for the heart to work efficiently.

(iii) Heart muscle disease (cardiomyopathy) causes the heart to become enlarged or the walls of the heart to become thick. This causes the heart to be less able to pump blood throughout the body and often results in heart failure.

(iv) Arrhythmia is an irregular or abnormal heartbeat. This can be a slow heart beat (bradycardia), a fast heartbeat (tachycardia), or an irregular heartbeat. Some of the most common arrhythmias include atrial fibrillation (when the atria or upper heart chambers contract irregularly), premature ventricular contractions (extra beats that originate from the lower heart chambers, or ventricles), and bradyarr hythmias (slow heart rhythm caused by disease of the heart's conduction system).

(v) Heart failure (congestive heart failure, or CHF) occurs when the heart is not able to pump sufficient oxygen-rich blood to meet the needs of the rest of the body. This may be due to lack of force of the heart to pump or as a result of the heart not being able to fill with enough blood.

(vi) Congenital heart disease is a type of birth defect that causes problems with the heart at birth and occurs in about one out of every 100 live births. Some of the most common types of congenital heart disease include:

(a) Atrial septal defects (ASD) and ventricular septal defects (VSD), which occur when the walls that separate the right and left chambers of the hearts are not completely closed

(b) Patent ductus arteriosus (PDA), which occurs when the ductus arteriosis doesn't close properly after birth.

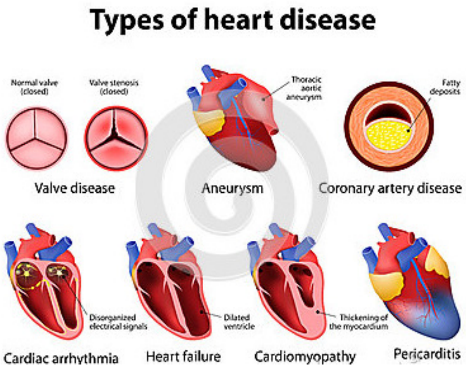

Fig. 4. Types of Heart Disease.

IV. RELATED WORK

Author/ Researchers	Description
Lamia Abed Noor Muhammed [4]	presented and discussed the experiment that was executed with naïve bayes technique in order to built predictive model as an artificial diagnose for heart disease based on data set which contains set of parameters that were measured for individuals previously. Then compare the results with other techniques according to using the same data that were given from UCI repository data
Ranganatha S., Pooja Raj H.R., Anusha C., Vinay S.K. [5]	stored medical information of patients who come for hospitalization for heart disease and algorithms are run on that information and result will be provided in the form of user understandable words and graph. When very large data sets are present, data mining algorithms (here considering only ID3 and Naïve Bayesian algorithms) are used. ID3 outputs the result in the form of decision tree which can be easily understood. Naïve Bayesian predicts the chances of heart disease based on conditions given
G. Vaishali, V. Kalaivani [6]	developed a centralized patient monitoring system using big data. In the proposed system, large set of medical records are taken as input. From this medical dataset, it is aimed to extract the needed information from the record of heart patients using map reduce technique. Heart disease is a major health problem and it is the leading causes of death throughout the world. Early detection of heart disease has become an important issue in the medical research fields. For heart disease detection, some features are analyzed such as RR interval, QRS interval and QT interval. The classification process states whether the patient is normal or abnormal and in the detection step using map reduce technique to detect the disease and reduce the dataset. Thus, the proposed system helps to classify a large and complex medical dataset and detect the heart disease.

Author/ Researchers	**Description**
Ankita Dewan, Meghna Sharma[7]	Developed a prototype which can find out and extract unknown knowledge (patterns and relations) related with heart disease from a past heart disease database record. It can resolve complicated queries for detecting heart disease and hence assist medical practitioners to make smart clinical decisions which traditional decision support systems were not able to. By providing proficient treatments, it can help to decrease costs of treatment.
B. Venkatalakshmi, M.V Shivsankar [8]	Design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree and Naïve Bayes algorithms. In this proposed work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naïve Bayes outperforms when compared to Decision tree.
S. U. Amin, K. Agarwal, and R. Beg [9]	Implemented a hybrid system that uses global optimization benefit of genetic algorithm for initialization of neural network weights. The prediction of the heart disease is based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, smoking, alcohol intake and obesity.
A. K. Sen, S. B. Patel, and D. P. Shukla[10]	Proposed a layered neuro-fuzzy approach to predict occurrences of coronary heart disease simulated in MATLAB tool. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work efficiency in performing analysis for coronary heart disease occurrences
M. Jabbar, P. Chandra, and B. Deekshatulu [11]	Proposed a new approach for association rule mining based on sequence number and clustering transactional data set for heart disease predictions. The implementation of the proposed approach was implemented in C programming language and reduced main memory requirement by considering a small cluster at a time in order to be considered scalable and efficient.
P. Chandra, M., Jabbar, and B. Deekshatulu [12]	Created class association rules using feature subset selection to predict a model for heart disease. Association rule determines relations amongst attributes values and classification predicts the class in the patient dataset. Feature selection measures such as genetic search determines attributes which contribute towards the prediction of heart diseases.
B.Venkatalakshmi, M.V Shivsankar [23]	This project intends to design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree and Naïve Bayes algorithms. In this proposed work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naïve Bayes outperforms when compared to Decision tree

IV. DATA MINING TECHNIQUES

Researchers have employed different data mining techniques to help specialists and physicians diagnose heart disease [13]. Some techniques are more common such as Naïve Bayes, decision tree and K-nearest neighbor. However, there are other classification-based data mining techniques such as kernel density, neural network, bagging algorithm, sequential minimal optimization, direct Kernel self-organizing map and support vector machine. The next section briefly explains those techniques which were used in this study.

A. Decision tree

There are different types of decision trees. They only differ in the mathematical model they use to select the class of attribute during rule extraction. Gain ratio decision tree is the most common, successful type [14]. It is a relationship between entropy (information gain) and classified information. In entropy technique, the attribute which minimizes entropy and maximizes information gain is selected as the tree root. To select tree root, it is first necessary to calculate the information gain of each attribute. Then, the attribute maximizing information gain should be selected. Information gain, or entropy[15].

$$E = -\sum_{i=1}^{k} p_i log_2^{p_i}$$

Where *k* is the number of response variable classes, *pi* is the ratio of the number of the ith class events to total number of samples (occurrence probability of *i*).

B. Bayesian network

Bayesian network is a statistical technique predicting the membership class of the studied sample using the probability theory. Bayesian network practices classification process in accordance with Bayes' theorem. It assumes that the influence of the value of a theorem on a class is independent from the influence of other attributes. This assumption is called "class conditional independence". This assumption was made to simplify engaged calculation and this is why it was named "Naïve", i.e., simple. This technique calculates the prior probability of the response variable and the conditional probability of other variables. The prior and conditional probabilities of the initial training are calculated. Then, for every test dataset sample, the probability of the occurrence (presence) of each case of response variable is calculated. Afterwards, the response variable with the highest occurrence probability is selected. The probability of test sample for the response variable value is[16].

$$P(v = c_i = P(c_i) = \sum_{j=1}^{n} P(a_j = v_j | class = c_i)$$

Where V, ci, aj and vj are test sample, response variable value, data attribute and the test sample value, respectively.

C. K-nearest neighbor

This classification technique is called a memory-based technique, since the training samples should be stored in memory during runtime [17-21]. If a is the first sample denoted by (a1, a2,…, an), and b is the second sample denoted by (b1, b2,…, bn), the distance between them is calculated by relation.

$$\sqrt{(a_1 - b_1)^2(a_2 - b_2)^2(a_n - b_n)^2}$$

D. Support vector machine

Given availability of support vectors, Support Vector Machine (SVM) is the boundary determining the best data classification and separation. In SVM, only those data lying inside support vectors are used as the base data for machine and building a model. This means that this algorithm is not sensitive to other data. It aims to find the best data boundary with the farthest possible distance from all classes (their support vectors). SVM transfers data to a new space with respect to their predetermined classes so that data can be classified and separated linearly (using hyperplanes). Then, it searches for support lines (or support planes in multi-dimensional space) and tries to determine the equation of a straight line that maximizes the distance between each two classes. Each support vector is characterized with an equation describing the boundary line of each class.

E. Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and has no predefined classes. They are used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Moreover, they are used for data compression, outlier detection, understand human concept formation. Some of the applications are Image processing, spatial data analysis and pattern recognition. Classification via Clustering is not performing well when compared to other two algorithms. All these algorithms are implemented with the help of WEKA tool for the diagnosis of heart diseases.

Data set of 294 records with 13 attributes. These algorithms have been used for analyzing the heart disease dataset. The Classification Accuracy should be compared for this algorithm. After the comparison attributes are to be reduced for further purpose.

F. Neural Networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [19]. A Multi-layer Perceptron Neural Networks (MLPNN) is used. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of Yj = f (Σ wji xi) Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

V. CONCLUSION

The major no. of people is suffering from the heart disease problem so the diagnosis of this disease becomes vital task. Data mining techniques provides huge amount of dataset which helps in extracting the essential information heart patient. In this paper, we present the review of literature of the heart disease diagnosis using data mining. In this we also discuss various data mining techniques and analyze that the neural network and decision tree technique gives approximately same value of accuracy and some are less efficient in diagnosis of heart disease. So in future work design such technique which use the best features of two or more data mining technique which can efficiently determine and diagnose heart disease.

REFERENCES

[1]. A. Aziz, N. Ismail, and F. Ahmad, (2013). "Mining Students' Academic Performance", *Journal of Theoretical & Applied Information Technology,* vol. **53**, no. 3, pp, 485-496.

[2]. S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita (2016). "Prediction of Heart Disease using Data Mining Techniques", *Indian Journal of Science and Technology,* Vol. **9**(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016.

[3]. J. Banupriya, S. Kiruthika (2016). "Heart Disease Using Data Mining Algorithm on Neural Networks and Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering,* **6**(8), August-2016, pp. 40-42.

[4]. Lamia Abed Noor Muhammed (2012). "Using Data Mining technique to diagnosis heart disease", In proceeding of IEEE, 2012.

[5]. Ranganatha S., Pooja Raj H.R., Anusha C. and Vinay S.K. "Medical Data Mining And Analysis For Heart Disease Dataset Using Classification Techniques", In proceeding of IEEE, 2013.

[6]. G. Vaishali and V. Kalaivani (2016). "Big Data Analysis for Heart Disease Detection System Using Map Reduce Technique", In proceeding of IEEE, 2016.

[7]. Ankita Dewan and Meghna Sharma (2015). "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", In proceeding of IEEE 2015.

[8]. B. Venkatalakshmi and M.V. Shivsankar (2014). "Heart Disease Diagnosis Using Predictive Data mining", *International Conference on Innovations in Engineering and Technology (ICIET'14) On* 21st & 22nd March, Volume **3**, Special Issue 3. In proceeding of IJIRSET.

[9]. S.U. Amin, K. Agarwal, and R. Beg, (2013). "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in *Proceedings of 2013 IEEE Conference on Information and Communication Technologies* (ICT 2013), 2013, no. Ict, pp. 1227–1231.

[10]. A.K. Sen, S.B. Patel and D.P. Shukla, (2013). "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," *International Journal of Engineering and Computer Science,* vol. **2**, no. 9, pp. 1663–1671, 2013.

[11]. M. Jabbar, P. Chandra, and B. Deekshatulu, (2011). "Cluster Based Association Rule Mining For Heart Attack Prediction,". *Journal of Theoretical & Applied Information Technology,* vol. **32**, no. 2, pp. 196–201, 2011.

[12]. P. Chandra, M.. Jabbar, and B. Deekshatulu, (2012). "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628–634.

[13]. Helma C., Gottmann E., Kramer S. (2000) Knowledge discovery and data mining in toxicology. *Statistical Methods in Medical Research,* **9**: 329-358.

[14]. Quinlan, J.R. (1986) Decision trees and multi-valued attributes. In: Hayes, Michie D (eds.) Machine intelligence. Oxford University Press.

[15]. Han J. and Kamber M. (2006). Data Mining Concepts and Techniques: Morgan Kaufmann Publishers.

[16]. Bramer M (2007) Principles of data mining: Springer.

[17]. Alpaydin E (1997). Voting over multiple condensed nearest neighbors. *Artificial Intelligence Review***, 11**: 115-132.

[18]. National Center for Chronic Disease Prevention and Health Promotion (2013). Know the facts about heart disease. Vol. 5-1, pp,1-2.

[19]. Rajkumar, M. and Reena, G.S. (2010). Diagonsis of Heart Disease using Data mining Algorithm. *Global Journal of Computer Science and Technology,* **10**: 38-43.

[20]. Shouman, M. (2014). Prototype Development of a Novel Heart Disease Risk Evaluation Tool Using Data Mining Analysis. Zagazig University, Egypt. Vol. **119**, pp, 38-48.

[21]. Tu, M.C., Shin, D., Shin, D. (2009). Effective Diagnosis of Heart Disease through Bagging Approach. *Paper presented at the 2nd International Conference on Biomedical Engineering and Informatics.*

[22]. Ankita Katare and Shubha Dubey, (2017). A Study of various Techniques for Predicting student Performance under Educational Data Mining, *International Journal of Electrical, Electronics and Computer Engineering* **6**(1): 24-28(2017).

[23]. Asma Aziz Khan and Vipin, Verma, (2017). Prediction of Diabetes Disease Using Entropy and Gain based Data Mining Approach, *International Journal of Electrical, Electronics and Computer Engineering,* **6**(1): 164-172.