Volume 7, Issue 3, March 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Health Data Using Data Mining Techniques

Dr. A. R. Pon Periasamy
Associate Professor of Computer Science
Nehru Memorial College
Puthanampatti, Trichy (DT)
Tamilnadu, India

S. Mohan
Assistant Professor of Computer Science
Dr.S.Ramadoss Arts & Science College
Periyavadavadi, Vridhachalam
Tamilnadu, India

Abstract— Data Mining refers to the mining of useful and interesting patterns from large data sets. Medical data now a day is available in abundance but without proper mining they cannot be used. The goal of data mining is to turn data that are facts, numbers, or text which can be processed by a computer into information and knowledge. Using data mining techniques on medical data several critical issues can be understood better and dealt with starting from studying risk factors of several diseases to identification of the diseases occurring frequently or taking care of hospital information For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The huge amounts of data generated by healthcare transactions 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. The main aim of this survey is, analysis of the uniqueness of medical data mining. Therefore, this paper aims to understand about data mining and its importance in medical systems.

Key Words: KDD Process, Classifiers, Artificial Neural Networks, Decision Tree, K- Nearest Neighbor.

I. INTRODUCTION

Data mining is the process of analyzing and summarizing data from different perspectives and converting it into useful information. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods.

It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals Medical databases are very bulky that need computerized programs to find latent trends that will help in medical diagnosis and treatment.

In the wake of data mining techniques, especially medical data mining techniques, the health care domain has made significant progress in using the technologies in prevention and diagnosis of disease. The American Medical Informatics Association defined health Informatics as "all aspects of understanding and promoting the effective organization, analysis, management, and use of information in health care".

The data mining processes include formulating a hypothesis, collecting data, performing preprocessing, estimating the model, and interpreting the model and draw the conclusions.

1.1 Stages of KDD process

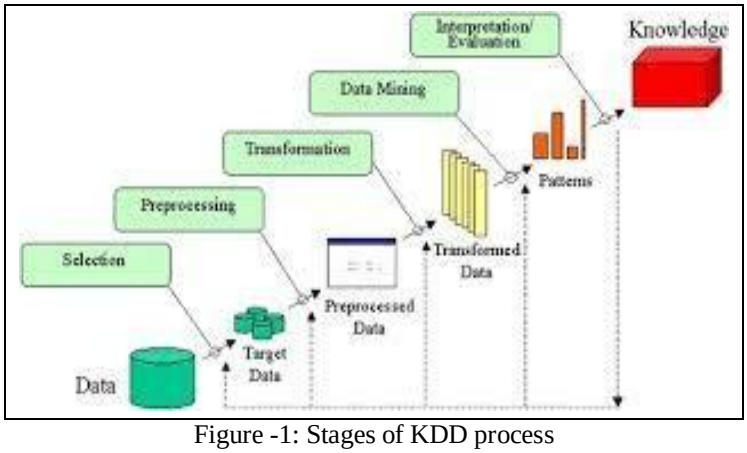

Figure -1: Stages of KDD process

A number of relationships are hidden among such a large collection of data for example a relationship between patient data and their number of days of stay

Selection

The data is selected according to some criteria in this stage. For example, a bicycle owns by all those people, we can determine subsets of data in this way.

Preprocessing

This stage removes that information which is not necessary. For example while doing pregnancy test it is not necessary to note the sex of a patient. It is also known as data cleansing stage.

Transformation

This stage transformed only those data which are useful in a particular research .For example only data related to a particular demography is useful in market research.

Data mining

Data mining is a stage knowledge discovery process. This stage is useful for extracting the meaningful patterns from data.

Interpretation and evaluation

The meaningful patterns which the system identified are interpreted into knowledge in this stage. This knowledge may be then useful for making useful.

II. SIGNIFICATION OF DATA MINING IN HEALTH CARE

Generally all the healthcare organizations across the world stored the healthcare data in electronic format. Healthcare data mainly contains all the information regarding patients as well as the parties involved in healthcare industries. The storage of such type of data is increased at a very rapidly rate. Due to continuous increasing the size of electronic healthcare data a type of complexity exist in it. In other words, we can say that healthcare data becomes very complex. By using the traditional methods it becomes very difficult in order to extract the meaningful information from it. But due to advancement in field of statistics, mathematics and very other disciplines it is now possible to extract the meaningful patterns from it. Data mining is beneficial in such a situation where large collections of healthcare data are available.

Recently researchers uses data mining tools in distributed medical environment in order to provide better medical services to a large proportion of population at a very low cost, better customer relationship management, better management of healthcare resources, etc. It provides meaningful information in the field of healthcare which may be then useful for management to take decisions such as estimation of medical staff, decision regarding health insurance policy, selection of treatments, disease prediction *etc.*, [1-4].

III. DATA MINING METHODOLOGIES

Well-known data mining techniques include the Artificial Neural Network (ANN), Decision tree, Bayesian classifiers, and K nearest neighbor classification techniques. In this section, we introduce these four widely used data mining techniques.

3.1 Artificial Neural Network Human Performance

Artificial neural networks models have been studied for many years in the hope of achieving human like performance in several fields. In Neural Networks, basic elements are neurons or nodes. These neurons are interconnected and within the network they worked together in parallel in order to produce the output functions. From existing observations they are capable to produce new observations even in those situations where some neurons or nodes within the network fails or go down due to their capability of working in parallel. An activation number is associated to each neuron and a weight is assigned to each edge within a neural network. In order to perform the tasks of classification and pattern recognition neural network is mainly used [5].

ANN is based on the biological neural networks in the human brain and described as a connectionist model

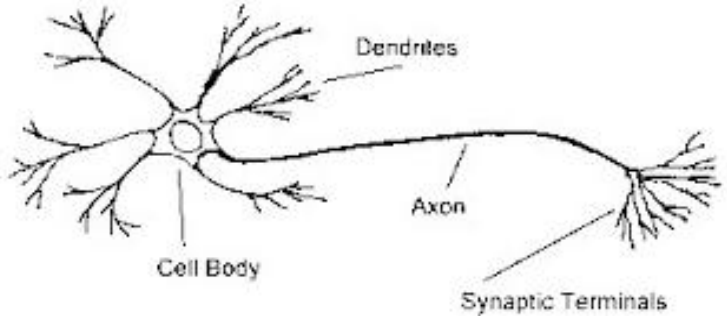

Figure -2: A Sketch of a Neuron in the Human Brain

It is based on the neuron, a cell that processes information in the human brain [6]. The neuron cell body contains thenucleus, and has two types of branches, the axon and the dendrites. The axon transmits signals or impulses to other neurons while the dendrites receive incoming signals or impulses from other neurons. Every neuron is connected and communicates through the short trains of pulses [6]. The nodes are the artificial neuron and the directed edges represented the connection between output neurons and the input neurons. In training phase, the internal weights of the neural network are adjusted according to the transactions used in the learning process. For each training transaction the neural network receives in addition the expected output. This allows modification of weight.

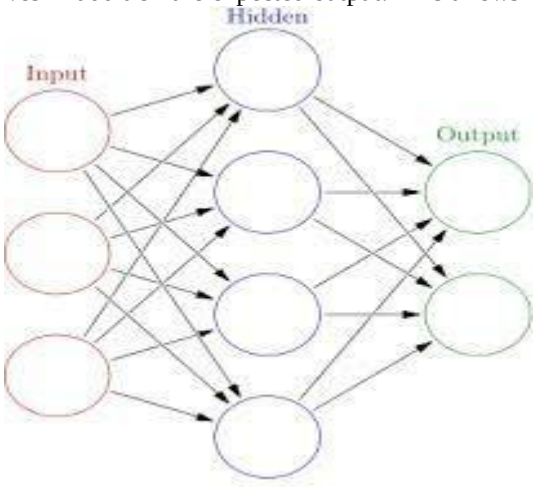

Figure -3: Artifical Neural Networks

3.2 Decision Tree Stages in DM

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making The goal is to create a model that predicts the value of a target variable based on several input variables. Decision trees used in data mining are of two main types:

Classification tree analysis is when the predicted outcome is the class to which the data belongs.

Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital

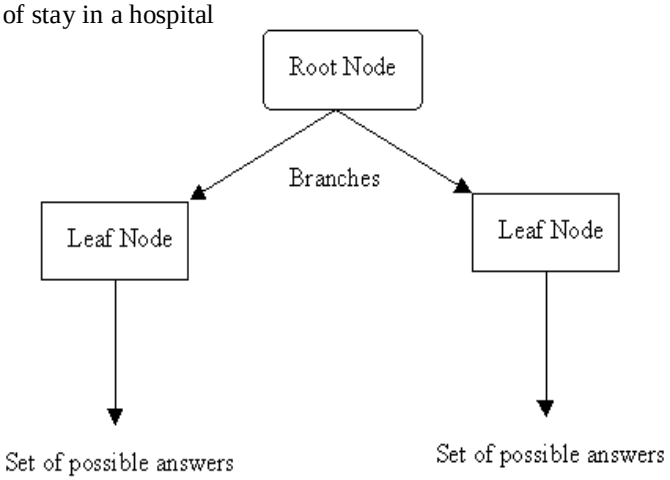

Figure -4: Decision Tree

3.3 Bayesian Classifiers Theorem of statistics

For probabilistic learning method Bayesian classification is used. Bayes theorem of statistics plays a very important role in it. While in medical domain attributes such as patient symptoms and their health state are correlated with each other but Naïve Bayes Classifier assumes that all attributes are independent with each other. This is the major disadvantage with Naïve Bayes Classifier. If attributes are independent with each other then Naïve Bayesian classifier has shown great performance in terms of accuracy. There are two types of probabilities

Posterior Probability [P(H/X)]

Prior Probability [P(H)]

where X is data tuple and H is some hypothesis. According to Bayes' Theorem,

$P(H/X)= P(X/H)P(H) / P(X)$

A. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a

closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

B. Probabilistic model

Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$ representing some *n* features (independent variables), it assigns to this instance probabilities

$$p(C_k|x_1, \ldots, x_n)$$

for each of *K* possible outcomes or *classes*.

The problem with the above formulation is that if the number of features *n* is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

In plain English, using Bayesian probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features F_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C_k, x_1, \ldots, x_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned} p(C_k, x_1, \ldots, x_n) &= p(C_k)\ p(x_1, \ldots, x_n|C_k) \\ &= p(C_k)\ p(x_1|C_k)\ p(x_2, \ldots, x_n|C_k, x_1) \\ &= p(C_k)\ p(x_1|C_k)\ p(x_2|C_k, x_1)\ p(x_3, \ldots, x_n|C_k, x_1, x_2) \\ &= p(C_k)\ p(x_1|C_k)\ p(x_2|C_k, x_1)\ \ldots p(x_n|C_k, x_1, x_2, x_3, \ldots, x_{n-1}) \end{aligned}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$, given the category C. This means that

$$p(x_i|C_k, x_j) = p(x_i|C_k),\ p(x_i|C_k, x_j, x_q) = p(x_i|C_k),\ p(x_i|C_k, x_j, x_q, x_l) = p(x_i|C_k),$$

and so on, for $i \neq j, q, l$. Thus, the joint model can be expressed as

$$\begin{aligned} p(C_k|x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\propto p(C_k)\ p(x_1|C_k)\ p(x_2|C_k)\ p(x_3|C_k)\ \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k)\,. \end{aligned}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(C_k|x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

where the evidence $Z = p(\mathbf{x})$ is a scaling factor dependent only on $x_1, \ldots, x_n$, that is, a constant if the values of the feature variables are known.

Constructing a classifier from the probability model

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the *maximum a posteriori* or *MAP* decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some *k* as follows:

$$\hat{y} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

3.4 *k*-Nearest Neighbors K-NN Algorithm

In pattern recognition, the ***k*-Nearest Neighbors algorithm** (or ***k*-NN** for short) is a non – parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature spaces. The output depends on whether *k*- NN is used for classification or regression:

In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing *etc*. There are various advantages of KNN classifiers.[7] These are: ease, efficacy, intuitiveness and competitive classification performance in many domains. If the training data is large then it is effective and it is robust to noisy training data. A main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. If there is a big sample then its response time on a sequential computer will also large.[8]

IV. CLASSIFICATION TECHNIQUES ILLUTION IN HEALTHCARE

Gunasundari et al., used ANN for discovering the lung diseases. This research work analyze the chest Computed Tomography (CT) and extract significant lung tissue feature to reduce the data size from the Chest CT and then extracted textual attributes were given to neural network as input to discover the various diseases regarding lung [9].

Soni et al., proposed the associative classification approach for better analyzing the healthcare data. The proposed approach was the combined approach that integrated the association rules as well as classification rules. This integrated approach was useful for discovering rules in the database and then using these rules to construct an efficient classifier. In this research, experiments on the data of heart patients were performed in order to find out that accuracy of associative classifiers was better than accuracy of traditional classifiers.

Apart from this, the research also generated the rules using weighted associative classifier [10].

Breast cancer is one of the fatal and dangerous diseases in women. Potter et al, had performed experiment on the breast cancer data set using WEKA tool and then analyzed the performance of different classifier using 10-fold cross validation method [11]

Classification techniques were used for predicting the treatment cost of healthcare services which was increased with rapid growth every year and was becoming a main concern for everyone [12].

Khan et al., used decision tree for predicting the survivability of breast cancer patient [13].

Chang et al., used an integrated decision tree model for characterize the skin diseases in adults and children. The main focus of this research was to analyze the results of five experiments on the six major skin diseases. The main aim of this research was to construct the best predictive model in dermatology by using the decision tree and combine this decision tree with the neural network classification methods.

On the basis of experimental result, it has been found that neural network has 92.62% accuracy in prediction of skin diseases [14].

Das et al., proposed a intelligent medical decision support system based on SAS software for the diagnosis of heat diseases. In order to construct the proposed system, neural networks method was mainly used. In this research, experiments were performed on the data taken from Cleveland heart disease database. On this basis of experiments, it has been found that neural networks have

89.01% accuracy [15]

Cheng et al. [16] applied classification algorithm to diagnose cardio vascular diseases. For classification effectiveness they focused on two feature extraction techniques namely automatic feature selection and expert judgment.

Xue et al. [17] proposed and applied Bayesian Network algorithm for diagnosis of an ailment known as Coronary Heart Disease (CHD).

Abraham et al. [18] proposed discrimination techniques to improve the accuracy of classification of medical data using Naive Bayesian classifier algorithm

Doron Shalvi and Nicholas DeClaris, [19] discussed medical data mining through unsupervised neural networks besides a method for data visualization. They also emphasized the need for preprocessing prior to medical data mining.

In the year 2000 Krzysztof J. Cior [20], bioengineering professor, identified the need for data mining methods to mine medical multimedia content.

Tsumoto [21] identified problems in medical data mining. The problems include missing values, data storage with respect to temporal data and multi-valued data, different medical coding systems being used in Hospital Information Systems (HIS).

Shim and Xu [22] proposed a classification methodbased on Bayesian Ying Yang (BYY) which is a three layered model. They applied this model to classify liver disease through automatic discovery of medical trends.

Hai Wang, and Shouhong Wang [23] studied on the role of medical experts in medical data mining. Medical experts can give expert advice that can beused as input in medical data mining.

CHAO and WONG [24] proposed a decision tree learning methodology which could interpret attributes in medical data classification for higher accuracy when compared with Incremental Tree Induction (ITI) algorithm.

Tu et al. [25] proposed an intelligent medical decision support system which provides diagnosis of heart diseases through decision tree algorithm C4.5 and bagging algorithm Naïve Bayes

Khan et al.used decision tree for predicting the survivability of breast cancer patient [26] and Chien etal.proposed a universal hybrid decision tree classifier for

classifying the activity of patient having chronic disease. They further improved the existing decision tree model to classify different activities of patients in more accurate manner [27].

In the similar domain, Moon et al. exemplify the patterns of smoking in adults using decision tree for better understanding the health condition, distress, demographic and alcohol [28].

Chang et al.,also used an integrated decision tree model for characterize the skin diseases in adults and children [29].

Moon et al., used decision tree algorithm in order to characterize the smoking behaviors among smokers by assessing their psychological distress, psychological health status, consumption of alcohol, and demographic variables. The classification analysis was conducted on the basis of decision tree algorithm in order to find the relationship between the average numbers of cigarette consumption per day [30].

Jena et al., used K-NN and Linear Discriminate Analysis (LDA) for classification of chronic disease in order to generate early warning system. This research work used K-NN to analyze the relationship between cardiovascular disease and hypertension and the risk factors of various chronic diseases in order to construct an early warning system to reduce the complication occurrence of these diseases [31].

R. Bhubaneswar and K. Kalaiselvi [32], presents the significance of Naïve Bayes algorithm in healthcare sector. Using this technique, a Decision Support System is presented for diagnosing patients with Heart Disease.

Manjusha K. K, K.Sankaranarayanan and Seena P [33], proposes a system to determine the presence of different dermatological diseases in Kottayam and Alappuzha. The system is built using Naïve Bayes classifier which is based on Bayesian theorem. Author collected data from various healthcare areas and used Naïve Bayes Algorithm as they produces higher predictive accuracies. The percentage of eight skin diseases is predicted effectively using the implementation in Java platform. On the basis of imported inputs, the prediction window gives the results indicating the chance of occurrence of diseases.

V. CONCLUSIONS

In this review we identified and evaluated the most commonly used DM algorithms resulting as well-performing on medical databases, ased on recent studies. For any algorithm its accuracy and performance is of greater importance. But due to presence of some factors any algorithm can greatly lost the above mentioned property of accuracy and performance. Classification is also belongs to such an algorithm. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm but also degrades its performance. Hence, before applying classification algorithm it must be necessary to remove all those attributes from datasets who later on acts as noisy attributes. This paper has provided the summary of data mining techniques used for medical data mining It also throws light into the importance of locally frequent patterns and the mining techniques used for the purpose.

REFERENCES

[1] C. McGregor, C. Christina and J. Andrew, "A process mining driven framework for clinical guideline improvement in critical care", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). http://ceur-ws. org, vol. 765, **(2012)**.

[2] P. R. Harper, "A review and comparison of classification algorithms for medical decision making", Health Policy, vol. 71, **(2005)**, pp. 315-331.

[3] V. S. Stel, S. M. Pluijm, D. J. Deeg, J. H. Smit, L. M. Bouter and P. Lips, "A classification tree for predicting recurrent falling in community-dwelling older persons", J. Am. Geriatr. Soc., vol. 51, **(2003)**, pp. 1356-1364.

[4] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines", Int. J. Med. Inform., vol. 77, **(2008)**, pp. 81-97.

[5] M. H. Dunham, "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc., **(2003)**.

[6] A. K. Jain, et al.Artificial neural network : a tutorial[Online].

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

[8] Alpaydin, E. (1997), Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, p. 115–132.

[9] S. Gunasundari and S. Baskar, "Application of Artificial Neural Network in identification of Lung Diseases", Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on. IEEE, (2009)

[10] S. Soni and O. P. Vyas, "Using Associative Classifiers for Predictive Analysis in Health Care Data Mining",

[11] R. Potter, "Comparison of classification algorithms applied to breast cancer diagnosis and prognosis", advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, (2007) July, pp. 40-49.

[12] G. Beller, "The rising cost of health care in the United States: is it making the United States globally noncompetitive?", J. Nucl. Cardiol., vol. 15, no. 4, (2008), pp. 481-482.

[13] M. U. Khan, J. P. Choi, H. Shin and M. Kim, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", 30th AnnualInternational IEEE EMBS Conference Vancouver, British Columbia, Canada, (2008) August 20-24.

[14] L. Chang and C. H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis", Expert Systems with Applications, Elsevier, vol. 36, (2009), pp. 4035-4041.

[15] R. Das, I. Turkoglub and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, vol. 36, (2009), pp. 7675-7680.

[16] Tsang-Hsiang Cheng, Chih-Ping Wei, Vincent S. Tseng (n.d). Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches . IEEE. p1-6.

[17]Weimin Xue, Yanan Sun, Yuchang Lu (n.d). Research and Application of Data Mining in TraditionalChinese Medical Clinic Diagnosis. IEEE.p1-4.

[18] Ranjit Abraham, Jay B.Simha, Iyengar (n.d). A comparative analysis of discretization methods for MedicalDatamining with Naïve Bayesian classifier.

[19] Doron Shalvi and Nicholas DeClaris., (n.d). An Unsupervised Neural Network Approach to Medical DataMining Techniques. IEEE. 0 (0), p1-6.

[20] Krzysztof J. Cior , Medical Data Mining and Knowledge Discovery. (n.d). From the guest Editor. IEEE. p1-2

[21] Shusaku Tsumoto (n.d). Problems with Mining Medical Data. IEEE. p1-2.

[22] Jeong-yon shim, lei xu (n.d). Medical data mining model for oriental medicine via byy binary independent factor analysis. Ieee. P1-4.

[23] Hai Wang, Shouhong Wang 1. (n.d). Medical Knowledge Acquisition through Data Mining. IEEE. 0 (0),p1-4.

[24] Sam chao, Fai wong, "An incremental decision tree learning methodology regarding attributes in medical data mining". Proceedings of the eighth international conference on machine learning and cybernetics, baoding,12-15 july 2009.

[25] My Chau Tu AND Dongil Shin (2009). A Comparative Study of Medical Data Classification MethodsBased on Decision Tree and Bagging Algorithms. IEEE.P1-5

[26] M.U.Khan,J.P.Choi,H.Shin and M.Kim, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, (2008)August 20-24

[27] C.Chien and G.J.Pottie, "A Universal Hybrid Decision Tree Classifier Design for Human Activity Classification", 34th Annual International Conference of the IEEE EMBS San Diego, California USA,(2012)August 28-September1

[28] S.S.Moon,S.Y.Kang,W.JitpitaklertandS.B.Kim,"Decision tree models for characterizing smoking patterns of older adults", Expert Systems with Applications,Elsevier, vol. 39,(2012), pp. 445-451

[29] C.L.Chang and C.H.Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis", Expert Systems with ApplicationsElsevier, vol. 36,(2009)

[30] S. S. Moon, S. Y. Kang, W. Jitpitaklert and S. B. Kim, "Decision tree models for characterizing smoking patterns of older adults", Expert Systems withApplications, Elsevier, vol. 39, (2012), pp. 445-451.

[31] H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S.Chen, "Application of classification techniques ondevelopment an early-warning systemfor chronic illnesses", Expert Systems with Applications, vol. 39, (2012), pp. 8852-8858.

[32] R. Bhuvaneswari and K. Kalaiselvi, "Naive Bayesian Classification Approach in Healthcare Applications", International Journal of Computer Science and Telecommunications", Volume 3, Issue 1, January 2012.

[33] Manjusha K. K, K.Sankaranarayanan and Seena P,"Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification", International Journal of Advanced Research in Computer Science and Software Engineering 4(1), January -2014, pp. 864-868.

[34] Elma Kolçe (Çela) ,NekiFrasheri:A Literature Review of Data Mining Techniques Used in Healthcare Databases,ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857- 7288

[35] M.Kumari, S. Godara: Comparative Study of Data Mining Classification Methods inCardiovascular Disease Prediction, IJCST ISSN : 2229- 4333 Vol. 2, Issue 2, June 2011.

