A Survey on Medical Text Mining

Revathi M Nair
Department of Computer Science
College of engineering Poonjar

Sindhu L.
Department of Computer Science
College of engineering Poonjar

ABSTRACT
Medical diagnosis is considered as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system will be very useful for the medical field. Due to recent technology advances, large masses of medical data are available. These large data contain valuable information for diagnosing diseases. Text mining techniques are using to extract useful patterns from these mass data. It provides a user- oriented approach to the novel and hidden patterns in the data. This paper intends to provide the survey of various medical text mining techniques used in medical field. The purpose of this survey is to obtain a most suitable text mining technique for the medical data.

General Terms
Text Mining, Medical Text Mining methods

Keywords
Information Extraction, Summarization, Clustering, Classification, Topic Tracking, Information Visualization, Concept Linkage, Association Rule Mining, Question Answering.

1. INTRODUCTION
The automation of disease diagnosis system is growing, since all medical data are easily obtainable. In previous, an enhancement has been witnessed in the accuracy and sensitivity of diagnostic tests, from observing external symptoms and using sophisticated lab tests and complex imaging methods. That permits detailed internal examinations. This improved accuracy has resulted in an exponential increase in the patient data available to the physician. The process of getting evidence to detect a probable cause of patient's key symptoms from all other possible causes of the symptom are known as establishing a medical diagnosis.

Data Mining Techniques applied in many application domains like e-business, Marketing, Health care and Retail have led to its application in other industries and sectors. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical field. These patterns can be utilized for clinical diagnosis. But the available raw medical data are widely in the form of distributed in nature and large. These data need to be compiled in an organized pattern. Medical diagnosis is regarded as an important yet complex task that needs to be done accurately and efficiently. The Healthcare environment is still information rich but knowledge poor.

Text mining technology provides a user oriented approach to the novel and hidden patterns in the data. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are used in order to extract patterns. Many techniques available in text mining such as classification, clustering, association rule, Summarization etc[2].

The paper is organized as follows: section 2 brief introduction of Text Mining, Section 3, followed by major text mining methods. Section 4 brief description of related works and section 5 covers conclusion

2. TEXT MINING
Text mining is similar to data mining except that data mining can only handle structured data sets but text mining can handle unstructured or semi-structured data set like emails, HTML files etc. So text mining is better for handle large volume of data[1].

3. TEXT MINING METHODS
Major text mining methods includes Information Extraction, Topic Tracking, Summarization, Categorization or Classification, Clustering, Concept Linkage, Information visualization and Association Rule Mining. Classification includes K-Nearest Neighbor, Decision Tree(DT), Support Vector Machine(SVM), Neural Network(NN) and Bayesian Method Clustering includes Hierarchical clustering, Partition Method and Density Based Clustering. The rest of this section briefly describes these models.

3.1 Information Extraction
Information extraction is a starting point for computers to examine unstructured text. Information extraction technique uses a process called pattern matching this process is looking for predefined structures in the text from this it identifies key phrases and relationships within a text[3]. IE can be used to extract meaningful information from large volume of text. One difficulty of IE is that it cannot represent the extracted information directly into a structured format so it requires a post processing.

3.2 Topic Tracking
Topic tracking mechanism works by using the topic searched by the user, or previously viewed by the user and then predicts the topic interest to the user based on the previously searched topic[4]. Topic tracking have some limitations that is if a user search for a topic s/he will receive numerous results but very few are really on text mining.

3.3 Summarization
Text summarization is the process of creating the short version of a source text by preserving the main points without losing its meaning. Summarization helps the users to easily read and understood the huge documents. Two classifications of summarization are extractive and abstractive. Extractive summarization selects main sentences and paragraphs etc and then concatenating them into abstract form. The main sentences are selected based on the statistical and linguistic features of the sentence. Abstractive summarization attempts to understand the main concepts of text and then express the main concepts in natural language. It uses linguistic methods to examine the text to find the concepts One of the challenge faced by this technique is difficulty to teach software to analyze semantics and to interpret meaning[5].

3.4 Classification/Categorization

Text classification is the task of classifying the document and assigning the document into predefined classes based on its content[6]. In the medical field classification can be used to divide patients into high risk or low risk patient on the basis of their disease pattern. Binary and multilevel are the two methods of classification. Binary includes only two classes but multilevel include more than two classes. Dataset are partitioned as training and testing. Using training dataset we learned the classifier and correctness is checked by test dataset. In healthcare classification is most widely using for text mining. Following are the various classification algorithms:

3.4.1 K-Nearest Neighbor

K-Nearest Neighbor (K-NN) classifier is one of the simplest non parametric method of classification. It discovers an unknown object from known neighbor object(nearest neighbor). That is object is classified by a majority vote of its neighbors, object assigned to the class most common among its k nearest neighbors[7].

3.4.2 Decision Tree (DT)

DT is a tree-like graph in which every leaf represents class labels and branches represent conjunctions of features that lead to those class labels. The top most node labels in the tree is called root node. Building a decision for any problem doesn't need any type of domain knowledge. Decision Tree can be used in operations research analysis for calculating conditional probabilities [8]. Using Decision Tree, visually and explicitly represent decisions and decision making. Decision makers can choose best alternative from root to leaf that indicates unique class separation based on maximum information gain.

3.4.3 Support Vector Machine (SVM)

The support vector machine classifier constructs hyper plane or multiple hyper planes in a high or infinite dimensional space. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space. Various kernel function such as polynomial, Gaussian, sigmoid *etc.*, are used for this purpose. The goal of svm is to find a decision boundary between two classes that is maximally far from any point in the training data.

3.4.4 Neural Network (NN)

It is an algorithm for classification; they are crude electronic networks of neurons based on the neural structure of the brain. They process one record at a time, and learn by comparing their classification of the record with the known actual classification of the record. The errors occurred in classification of the first record is fed back into the networks algorithm, it repeats for many iterations[8]. Neural Network is used for classification and pattern recognition.

3.4.5 Bayesian Method

The classification based on bayes theory is known as Bayesian classification. It is based on Bayesian networks. As a classifier it learns from training data from the conditional probability of each attribute given the class label. Bayes rule is used to compute the probability of the classes since given the particular instance of the attributes, class prediction is done by identifying the class with the highest posterior probability[9]. The major problem with Naïve Bayes Classifier is that it assumes that all attributes are independent with each other while in medical field attributes such as patient symptoms and disease are related with each other. Rather than this problem, Naïve Bayesian classifier has great performance in terms of accuracy and efficiency so if attributes are independent with each other then we can use it in medical field.

3.5 Clustering

Clustering is an unsupervised data mining (machine learning) technique used for grouping the data elements without advance knowledge of the group definitions[10]. The objective of clustering is to find the intrinsic grouping in a set of unlabelled data. Transform the set of features into subsets so that features in the same subset are similar in some sense.

3.5.1 Hierarchical Clustering

Hierarchical clustering organizes data in a tree structured clusters. The result is a binary tree or a dentrogram. The root node represent the whole data and each leaf node represent the data objects. The intermediate node describe the extent that the objects are proximal to each other. Hierarchical clustering classified into two agglutinative and divisive. Agglutinative cluster starts with each of the units in a separate cluster and recursively merges two or more clusters and ends up with a single cluster of all units and then form new clusters by dividing those single cluster until ends up in clusters containing individual units[11].

3.5.2 Partition Method

Partition algorithms construct partitions of a database of N objects into a set of k clusters. The partitioning clustering algorithm usually adopts the Iterative Optimization. Iterative optimization means different relocation schemes like k-medoid, k-means that iteratively reassign objects between the *k* clusters to improve the quality. K means is very simple and easy to implement. It is a simple iterative method to recover the user specified number of cluster k which is symbolized by the centroids. It presents no limitations on attribute types the choice of methods is dictated by the position of a predominant fraction of points inside a cluster[10]. The grouping of person on the basis of high blood pressure and cholesterol level into high risk and low risk of having heart disease using K-means clustering.

3.5.3 Density based Clustering

The density based cluster is discovering the clusters of arbitrary shapes and the noise in a spatial database. It uses two parameters Epsilon and Minimum Points of each cluster and at least one point from the respective cluster. The number of neighbours is greater than or equal to minimum points, a cluster is formed. It vends all the neighbour points within distance epsilon of the starting point. The cluster is fully expanded then the algorithm proceeds to iterate through the remaining invested points in the dataset. One disadvantage is cannot cluster data sets well with large differences in densities. It maintains the set of objects in three different categories such as classified, unclassified and noise. Every classified object has an associated cluster –id. A noise object may also have an associated dummy cluster –id. Unclassified object does not have any cluster –id[10].

3.6 Concept Linkage

Concept linkage is a method of identifying related documents which share similar concepts. Major goal of concept linkage is to promote browsing for information not for searching information[12]. Concept linkage is very useful in medical

field for identifying links between symptoms and diseases, and diseases and treatments.

3.7 Question Answering

Many websites that are allowing question answering technology, allow end users to "ask" question to the computer and get exact or related answer [13]. Question and Answering technique utilizes multiple text mining methods. First step question answering technique is the passage retrieval (PR) method. It allows passages with the highest probability of containing the answer to be retrieved, instead of simply recovering the passages sharing a subset of words with the question. Second step is Answer Extraction-aims to establish the best answer for a given question. It is based on a supervised machine-learning approach. It consists of two modules, one for attribute extraction and the other for answer selection.

3.8 Information Visualization

Information visualization is the visualization of data. It is the mapping of large textual data in a visual hierarchy. It provide browsing capabilities addition to simple searching. It includes three steps: (1) Data preparation: i.e. collecting original data of visualization and form original data space. (2) Data analysis and extraction: i.e. analyze and identifying visualization data needed from original data and create visualization data space. (3) Visualization mapping: i.e. employ certain mapping algorithm to map visualization data space to text cluster result.

3.9 Association Rule Mining

Association Rule Mining is a technique used to identify relationships among a large set of variables in a data set. It is also known as market basket analysis due to its capability of discovering the association among purchased item or unknown patterns of sales of customers in a transaction database[14]. For example if a customer is buying a laptop then the chance of buying antivirus software is high. This information helps the storekeeper to further enhance their sales . Association also has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms.

4. RELATED WORKS

Hu et al[15]. used different classification method such as decision tree, SVM and ensemble approach for analyzing microarray data. This research work performed comparative analysis of above mentioned classification method using 10-fold cross validation approach on the data set obtained from Kent Ridge Bio Medical Dataset repository. The experiment results indicate that among all classification method ensemble achieved good accuracy. Bestsimas et al[16]. used classification tree approach to predict the cost of healthcare by using the dataset of 3 years collected from the insurance companies to perform the experiment. The first two year data was used to train the classifier and last one year data was used for comparing the predicted results of classifier.

Jen *et al[17]*, 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 analyse 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. Shouman *et al[18]*, used K-NN classifier for analyzing the patients suffering from heart disease. The data was collected from UCI and experiment was performed using without voting or with voting K-NN classifier and it is found that K-NN achieve better accuracy without voting in diagnosis of heart diseases as compare to with voting K-NN. Chien *et al.[19]*, 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.

Liu *et al[20]*., proposed an Optimization (PSO) was also used for specifying fuzzy strength constraint and neighbourhood size.

Soliman *et al[21]*., used SVM classification approach for classification of various diseases and SVM together with k-means clustering was applied on microarray data for identifying the diseases. E.Avci[22] proposed a system using genetic SVM classifier for analyzing the heart valve disease. This system extracts the important feature and classifies the signal obtained from the ultrasound of heart valve.

Er *et al[23]*., construct a model using Artificial Neural Network (ANN) for analyzing chest diseases and a comparative analysis of chest diseases was performed using multilayer, generalized regression, probabilistic neural networks. An ensemble neural network methodology is proposed by Das *et al.[24]*, for diagnosis of heart disease in order to develop effective decision support system.

Curiac *et al[25]*., analyze the psychiatric patient data using BBN in making significant decision regarding patient health suffering from psychiatric disease and performed experiment on real data obtain from Lugoj Municipal Hospital.

Tapia et al[26]. analyzed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm

Another research work explores the application of Data Mining techniques in healthcare. Balasubramanian *et al[27]*., analyze the impact of ground water on human health using clustering technique. They discovered the causes of risk related with the fluoride content in water with the help of k-means clustering. Using this, author identified the valuable information in order to make decision regarding human health. Escudero *et al[28]*., used k-means clustering to classify the Alzheimer's disease (AD) data feature into pathologic and non-pathologic groups. This research work used the concept of Bioprofile and K-means clustering for early detection of AD.

Chipman et al[29]., proposed the hybrid hierarchical clustering approach for analyzing microarray data.The research work combines both top-down and bottom-up hierarchical clustering concepts in order to effectively utilize the strength of this clustering approach. Belciug et al[30] use the hierarchical clustering approach for grouping the patients according to their length of stay in the hospital that enhance the capability of hospital resource management.

Clebi et al[31], The research work extracts the useful and interesting patterns from biomedical images using density based clustering. This research discovers the area of homogeneous colour in biomedical images. This method separates the unhealthy skin or wound from healthy skin and discovers the sub regions of varied colour or spotted part inside the unhealthy skin which is again useful for classification and association task.

Ji *et al[32]*., used association in order to discover infrequent casual relationships in Electronic health databases . Healthcare organization widely used Association approach for

discovering relationships between various diseases and drugs. It is also used for detecting fraud and abuse in health insurance. Association is also used with classification techniques to enhance the analysis capability of Data Mining. Soni *et al[33].*, used an integrated approach of association and classification for analyzing health care data. This integrated approach is useful for discovering rules in the database and then using these rules an efficient classifier is constructed. This study performed experiment on the data of heart patients and also generate rules using weighted associative classifier.

Table 1. Comparison of different Text Mining Methods

Method	Related work	Working Mode	Advantage	Limitations
Information Extraction	No works done.	i) Based on the pattern matching process. ii) mapping of natural language texts into predefined, structured representation.	i) Easy to implement. ii) Computationally efficient. iii) Extract meaningful information from large volume of text.	i) Cannot represent the extracted information directly into a structured format. ii) Requires a post processing
Topic Tracking	No works done.	i) Based on topic searched or previously viewed by the user. ii) predicts the topic interest to the user.	i)Ranking of document ii)Does not consider index inside a document (ie,all weights are binary)	ii)Search for a topic receive numerous results. ii) Very few are really on text mining.
Method	Related Work	Working Mode	Advantages	Limitations
Summarization	No works done.	i) Select main sentences and paragraphs. ii) Concatenate them into abstract form. or iii) Understand main concepts of text. iv. Express main concepts in natural language.	i)Time saving ii) Easy to understand meaning of long text.	i)Difficulty to teach software to analyze semantics and to interpret meaning.
Categorization or classification	i) Hu et al[15] ii)Bestsimas et al[16].	i) Classify the doccument. ii) Assign document into predefined classes.	i) Training is done in faster manner. ii) Better Accuracy as compare to other methods.	i) It is restricted to one output attribute .
Clustering	i) Tapia et al[26].	i) Transform the features into subset. ii) Group similar subsets.	i) Easy to implement. ii) Documents can appear in multiple subtopics	i) Clustering user need to specify the number of cluster in advance.
Question Answering	No Work Done.	i) End user to allow ask questions. ii) Find passage with highest probability to contain answer. iii) Recovering the passages sharing a subset of words with the question. iv) Answer extraction.	i)Provide best answer for a question.	i) Specify the queries as Boolean expressions.
Information Visualization	No Work Done.	i)Data preparation. ii)Data analysis and extraction. iii) Visualization mapping	i) Users to see, explore, and understand large amounts of information at once. ii)Provides browsing capabilities.	i)Difficult to implement. ii)Require large memory.

Association Rule Mining	i) Ji *etal[32]*. ii) Soni *et al[33]*.	i) Identify relationships among a large set of variables in a data set.	i) Can relate many topics.	i) Obtaining non interesting rules. ii) Huge number of discovered rules. iii) Low algorithm performance.

Table 2. Comparison of different Classification Methods

Method	Related work	Working Mode	Advantages	Limitations
K-Nearest Neighbor	i) Jen *et al[17]*, ii) Shouman *et al[18]*, iii) Liu *et al[19].*,	i) Discovers the unidentified data point using the previously known data points(nearest neighbor).	i)It is easy to implement. ii) Training is done in faster manner.	i) It requires large storage space. ii) Sensitive to noise. iii). Testing is slow.
Decision Tree(DT)	i) Liu *et al[20]*.	i) Based on topic searched or previously viewed by the user. ii) predicts the topic interest to the user.	i)There are no requirements of domain knowledge in the construction of decision tree. ii). It minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. iii). It can easily process the data with high dimension.	i)It is restricted to one output attribute. ii) It generates categorical output. iii) It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset.
Support Vector Machine(SVM)	i) Soliman *et al[21]*. ii)E.Avci [22]	i) Select main sentences and paragraphs. ii) Concatenate them into abstract form. or iii) Understand main concepts of text. iv. Express main concepts in natural language.	i)Better Accuracy as compare to other classifier. Ii) Easily handle complex nonlinear data points. iii) Over fitting problem is not as much as other methods.	i)Computationally expensive. ii) The main problem is the selection of right kernel function. For every dataset different kernel function shows different results.
Neural Network	i) Er *et al[23]*. ii) Das *et al.[24]*,	i)Uses gradient descent method. ii) Based on biological nervous system. iii) Having multiple interrelated processing elements known as neurons.	i)Easily identify complex relationships between dependent and independent variables. ii) Able to handle noisy data.	i)Local minima. ii) Over-fitting. iii) The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks.
Method	Related Works	Working Mode	Advantages	Limitations
Bayesian Method	i) Curiac *et al[25]*.	i)Based on bayes theory. ii) concentrates on prior, posterior and discrete probability distributions of data items.	i)It makes computations process easier. ii) Have better speed and accuracy for huge datasets.	i)It does not give accurate results in some cases where there exists dependency among variables.

Table 3. Comparison of different clustering Methods

Method	Related work	Working Mode	Advantage	Limitations
K-means Clustering	i)Balasubramanian *et al[27]*. *ii)*Escudero *et al[28].*,	i)Based on iterative optimization. ii) Recover the user specified number of cluster k which is symbolized by the centroids.	i)Simple clustering approach. ii) Efficient. iii) Less complex method	i) Requires number of cluster in advance. ii) Problem with handling categorical attributes. iii) Not discover the cluster with non-convex shape. 4. Result varies in the presence of outlier.
Hierarchical Clustering	i) Chipman et al[29]. ii) Belciug et al[30].	i) Organizes data in a tree structured clusters. ii) The root node represent the whole data and each leaf node represent the data objects. iii) Intermediate node describe the extent that the objects are proximal to each other.	i)Easy to implement. ii)Having good visualization capability. iii) There is no need to specify the number of clusters in advance.	i)Have cubic time complexity in many cases so it is slower. ii. Decision regarding selection of merge or split point. Once a decision is made it cannot be undone.
Density Based Clustering	i)Clebi et al[31],	i) Discovering the clusters of arbitrary shapes. ii) Uses two parameters Epsilon and Minimum Points of each cluster. iii) The number of neighbours is greater than or equal to minimum points, a cluster is formed.	i)No need to specify number of cluster in advance. ii) Easily handle cluster with arbitrary shape. iii) Worked well in the presence of noise.	i)Not handle the data points with varying densities. ii) Results depend on the distance measure.

4. CONCLUSION

In this paper, different text mining techniques in medical field were studied and their advantages and drawbacks have been discussed. The different methods of text mining have been used to extract the useful patterns and thus the knowledge from this variety databases. Selection of data and methods for text mining is an important task in medical diagnosis and needs the knowledge of the domain. The main focus of this survey of text mining techniques is to how the text mining techniques applied in medical field. Each technique is suitable for some medical applications. An efficient mining method should be selected that suits desired task. The performance of clustering method is suitable for medical diagnosis since documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results. The grouping of related information from a patient's health history, physical examination, and laboratory results as part of the process of making a diagnosis.

5. REFERENCES

[1] Shah Neha K "Introduction of Text Mining and An Analysis of Text Mining Techniques" Paripex- Indian Journal of Research , volume :2,Issue:2,February 2013

[2] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies In Web Intelligence, VOL. 1, NO. 1, AUGUST 2009

[3] N. Kanya and S. Geetha ,"Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Comm. Technology in Electrical Sciences, IEEE(2007), Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111- 1118.

[4] Sungjick Lee and Han-joon Kim, "News Keyword Extraction for Topic Tracking", 4th International conference on Networked Computing and Advanced Information Management, IEEE (2008), Korea. 554-559.

[5] Vishal Gupta,Guruprit Lehal,"A survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August 2010.

[6] Jinshu, Su., Zhang, Bofeng., and Xin, Xu (2006). Advances in Machine Learning Based Text Categorization, Journal of Software, vol.17, No.9, pp1848-1859.

[7] Eui-Hong (Sam), Han George Karypis, Vipin Kumar, "Text Categorization Using Weight Adjusted *k*-Nearest Neighbor Classification", Army HPC Research Center University of Minnesota.

[8] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf, **(2003)**.

[9] Ranjit Abraham, Jay B. Simha, S. Sitharama Iyengar, "Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining" International Journal of Computational Intelligence Research.ISSN 0974-1259 Vol.4, No.X (2008).

[10] Periklis Andritsos, "Data Clustering Techniques", University of Toronto, March 11, 2002

[11] Prof. M.S. Prasad Babu, K. Swapna, Tilakachuri Balakrishna, Prof. N.B.Venkateswarulu, "An Implementation of Hierarchical Clustering on Indian Liver Patient Dataset", IJETCAS 14-495; © 2014,

[12] Rowena Chau, Ah Chung Tsoi, Markus Hagenbuchner, Vincent C.S. Lee, "A ConceptLink Graph for Text Structure Mining", , Wellington, New Zealand, January 2009.

[13] Yllias Chali, Shafiq R. Joty, Sadid A. Hasan, "Complex Question Answering: Unsupervised Learning Approaches and Experiments", Journal of Artificial Intelligence Research 35 (2009).

[14] Ji Hoon Kang, Dong Hoon Yang, Young Bae Park, and Seoung Bum Kim, "A Text Mining Approach to Find Patterns Associated with Diseases and Herbal Materials in Oriental Medicine", International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.

[15] H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, "A Comparative Study of Classification Methods For Microarray Data Analysis", Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, **(2006)**, pp. 33-37.

[16] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala and G. Wang, "Algorithmic prediction of health-care costs", Oper. Res., vol. 56, no. 6, **(2008)**, pp. 1382-1392.

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

[18] M. Shouman, T. Turner and R. Stocker, "Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients", International Conference on Knowledge Discovery (ICKD-2012), **(2012)**.

[19] D. Y. Liu, H. L. Chen, B. Yang, X. E. Lv, N. L. Li and J. Liu, "Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease", Journal of Medical System, Springer, **(2012)**.

[20] 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-September 1.

[21] T. H. A. Soliman, A. A. Sewissy and H. A. Latif, "A Gene Selection Approach for Classifying Diseases Based on Microarray Datasets", 2nd International Conference on Computer Technology and Development (lCCTD 2010), **(2010)**.

[22] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier", Expert Systems with Applications, Elsevier, vol. 36, **(2009)**, pp. 10618-10626.

[23] O. Er, N. Yumusakc and F. Temurtas, "Chest diseases diagnosis using artificial neural networks", Expert Systems with Applications, vol. 37, **(2010)**, pp. 7648-7655.

[24] 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.

[25] D. I. Curiac, G. Vasile, O. Banias, C. Volosencu and A. Albu, "Bayesian Network Model for Diagnosis of Psychiatric Diseases", Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, **(2009)** June 22-25.

[26] J. J. Tapia, E. Morett and E. E. Vallejo, "A Clustering Genetic Algorithm for Genomic Data Mining", Foundations of Computational Intelligence, vol. 4 Studies in Computational Intelligence, vol. 204, **(2009)**, pp. 249-275.

[27] T. Balasubramanian and R. Umarani, "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, **(2012)** March 21-23.

[28] J. Escudero, J. P. Zajicek and E. Ifeachor, "Early Detection and Characterization of Alzheimer's Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, **(2011)** August 30-September 3.

[29] H. Chipman and R. Tibshirani, "Hybrid hierarchical clustering with applications to microarray data", Biostatistics, vol. 7, no. 2, **(2009)**, pp. 286-301.

[30] T. S. Chen, T. H. Tsai, Y. T. Chen, C. C. Lin, R. C. Chen, S. Y. Li and H. Y. Chen, "A Combined K-Means and Hierarchical Clustering Method for improving the Clustering Efficiency of Microarray", Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems, **(2005)**.

[31] M. E. Celebi, Y. A. Aslandogan and R. P. Bergstresser, "Mining Biomedical Images with Density-based Clustering", Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), **(2005)**.

[32] J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, "Mining Infrequent Causal Associations in Electronic Health Databases", 11th IEEE International Conference on Data Mining Workshops, **(2011)**.

[33] S. Soni and O. P. Vyas, "Using Associative Classifiers for Predictive Analysis in Health Care Data Mining", International Journal of Computer Applications (0975 – 8887), vol. 4, no. 5, **(2010)** July.