A COMPARATIVE STUDY OF
DATA MINING APPLICATIONS IN DIAGNOSING DISEASES

Mrs.V.Priyavadana[1] ,Ms.A.Sivashankari[2], Mr.R.Senthil Kumar[3]

[1] Research Scholar, Department of Computer Science, D.K.M College for Women, Tamil Nadu, India
[2] Head of the Department, Department of Computer Science, D.K.M College for Women, Tamil Nadu, India
[3] Head Of the Department, Department of Computer Science, RTG College of Arts & Science, Tamil Nadu, India

---***---

Abstract - *Data Mining has emerged as a technique of extracting and discovering new knowledge in data implicit in a large data warehouse so as to enable better decisions and strategy formulation. It is an effective and rapid process of data searching, retrieval and semantic analysis from different perspectives. Data mining helps in presenting and summarizing information extracted from data warehouse in a meaningful manner that can be used to increase business performance, enhance revenue and reduce costs. It is an ultimate process of finding patterns or correlations among relational databases. Data mining has revealed novel biomedical and healthcare acquaintances for clinical decision making that has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Disease diagnosis is one of the applications where data mining tools are establishing the successful results. This paper summarizes some techniques on medical diagnosis and prognosis. It has also been focused on current research being carried out using the data mining techniques to enhance the disease(s) forecasting process.*

Key Words: *Data Mining Techniques, Data Mining Tools, Data Mining Applications.*

1. INTRODUCTION

Data Mining and Knowledge Discovery is demand of today's business as it is presently being successfully applied in diverse areas of banking, healthcare, industry, transportation, finance, retail, surveillance, science and engineering, spatial data and telecommunication etc. With the huge advancement of technology in areas of information retrieval, digitization, electronic access, data archiving, distributed network and online analytical processing, data mining provides the best solution to all. It gives the way for sophisticated data search using statistical algorithms to discover correlations and patterns in large preexisting data warehouse (Farlex, 2010)[1] where techniques are applied not only on flat files and relational databases but also on non-numerical data. Data mining techniques are used not only for Knowledge Discovery in Databases (KDD) but also for Customer Relationship Management (CRM), Business Intelligence (BI) and E-commerce. It is used for a variety of purposes both in private and public organizations for supporting better management of customer relationships by leveraging large data warehouse effectively in real time. The process of data mining becomes inevitable for discovering deeper information and categorization of data embedded in multiple source contents (Folorunso and Ogunde, 2010)[2].The future of data mining and knowledge discovery in databases (KDD) has been growing in leaps and bounds (Han and Kamber, 2001)[3]. Data mining, therefore consists of vital functional elements that transform data stored in multidimensional data warehouse, facilitates data access to analysts using application tools and techniques and meaningful presentation to managers for quick decisions and strategy achievement. According to the Gartner Group, "data mining is the process of discovering meaningful new correlation patterns and trends by shifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" (Larose, 2010)[4].

Past Trends- data mining was done on structured and numerical data stored in traditional databases by techniques of statistics and machine learning. It was primarily focused on serving business needs through techniques of fourth generation programming languages.

Current Trends- currently data mining explores not only simple homogeneous data but also heterogeneous data in multiple formats of structured, semi structured and unstructured data. It uses techniques of pattern recognition, artificial intelligence, machine learning and statistics to serve web, business and healthcare sectors through high end storage devices, high speed networks and parallel distributed computing.

Future Trends-Future focus of data mining is on complex data objects like graph; multi represented objects, noise in time series, high speed data streams etc. through techniques of soft and parallel computing like neural networks, fuzzy logic and genetic programming etc. It is going to serve more complex scientific and research fields, social networking, medical diagnosis, web and business using cloud computing and multi agent technologies.

2. DATA MINING IN MEDICAL SECTOR

Data mining is used in healthcare industries to analyze immense data of medical research, patient, staff and doctor's records, biotech and medicines and compounds in pharmaceutical industry thus enabling discovery of relationship between diseases, research of new drugs, treatments effectiveness, genetic network analysis and market activities in drug delivery services.

Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories.

i) Treatment value

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

ii) Healthcare organization

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

iii) Patron association management

Patron association management is a core approach to managing interactions between commercial organizations-typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

iv) Scam and Exploitation

Detect scam and exploitation establishes norms and then identifies unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

v) Medical gadget Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless bio-sensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients. Ubiquitous Data Stream Mining (UDM) techniques such as light weight, one-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge and available memory.

vi) Pharmaceutical Production

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in the Pharmacy data is vital to a firm's competitive position and organizational decision-making.

vii) Hospital Administration

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized. Three layers of hospital management:

- ✓ Services for hospital management
- ✓ Services for medical staff
- ✓ Services for patients

3. ANALYSIS OF MEDICAL PROBLEMS USING DATA MINING

Data mining tools are used to predict the successful results from the data recorded on healthcare problems. Different mining tools are used to predict the accuracy level in different healthcare problems. The following medical problems has analyzed and evaluated:

- Heart disease
- Cancer
- HIV/AIDS
- Tuberculosis
- Diabetes Mellitus
- Kidney dialysis
- Dengue
- IVF
- Hepatitis C

3.1 DATA MINING APPLICATIONS IN HEALTHCARE

The diseases are the most critical problems in human. To analyze the effectiveness of the data mining applications for diagnosing the disease, the traditional methods of mathematical / statistical applications are also given and compared. Listed eleven problems are taken for comparison with this work.

Table – 1: Comparison between diseases with different data mining techniques.

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naïve	Probability	60
2	Cancer	WEKA	Classification	Rules. Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naïve Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

Graph chart formed by using this table with the values of health care problems, Data Mining tools and Accuracy Level is as illustrated in the following Figure. In this chart, the prediction accuracy level of different data mining applications has been compared.

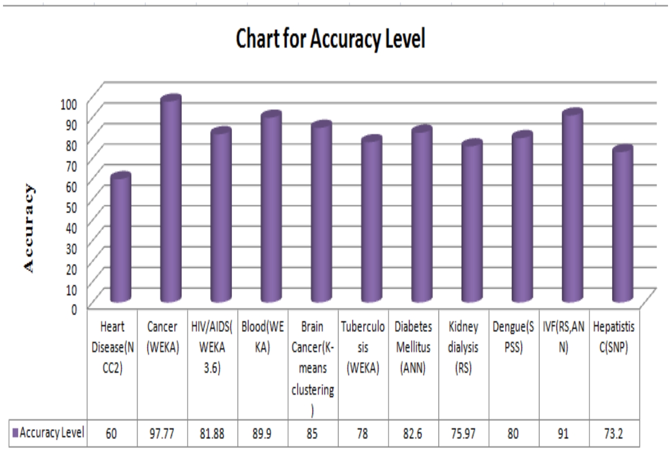


Chart -1: Accuracy Level of Various Diseases

4. COMPARATIVE STUDY OF IVF TREATMENTAND TUBERCULOSIS

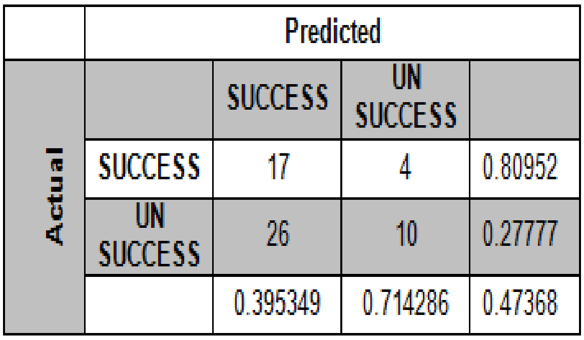
4.1 IVF SUCCESS RATE PREDICTION

The section deals with the comparative study of a data mining application for predicting the success rate of IVF treatment. The process of data mining applications, its advantages and results obtained are compared. The detailed study of selected works gives a broad idea about the application of data mining techniques.

4.1.1 Rough set theory for medical informatics data analysis

The research work aims to analyze the medical data by applying Rough Set Theory of data mining approach. The data reduction process has been done using rough set theory reduction algorithm. Rough set is mainly used to reduce the attributes without compromising its knowledge of the original. To analyze the fertilization data, ROSETTA tool kit reduction algorithm is used in this work to produce the optimal reduct set without affecting the original knowledge. The treatment success rate is predicted and tabulated as depicted in the following table,

Table -2: Actual and Predicted output using Rough set Theory

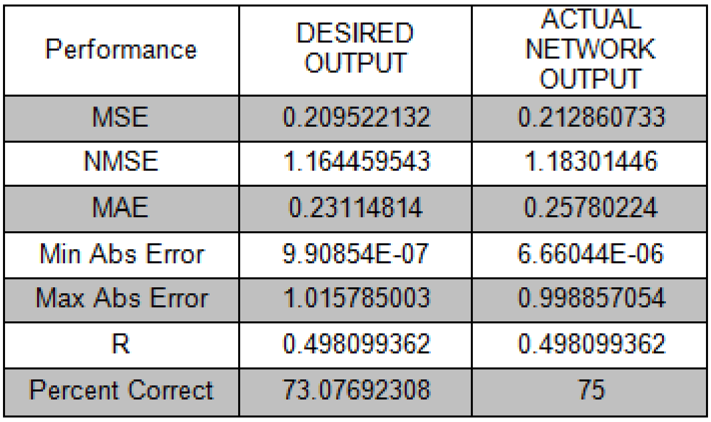


		Predicted		
Actual		SUCCESS	UN SUCCESS	
	SUCCESS	17	4	0.80952
	UN SUCCESS	26	10	0.27777
		0.395349	0.714286	0.47368

4.1.2 Artificial neural network for IVF treatment

This research work is mainly aimed to predict and classify the IVF treatment results using Artificial Neural Network (ANN). The artificial neural network is constructed with multi-layer perception and back-propagation training algorithm, and constructed network is trained, tested and validated using patients' sample IVF data. This work finally compares the success rate between desired output which is field recorded data and actual output which is predicted output of neural network. In the Table 3, the comparison between desired and actual output of the neural network is illustrated,

Table -3: Comparison between desired and actual output of ANN



Performance	DESIRED OUTPUT	ACTUAL NETWORK OUTPUT
MSE	0.209522132	0.212860733
NMSE	1.164459543	1.18301446
MAE	0.23114814	0.25780224
Min Abs Error	9.90854E-07	6.66044E-06
Max Abs Error	1.015785003	0.998857054
R	0.498099362	0.498099362
Percent Correct	73.07692308	75

4.2 TUBERCULOSIS PREDICTION

Bakar and Febriyani (2007) [5] applied Rough Neural Networks for classifying the tuberculosis types. Data set has 233 records, which has 14 attributes, firstly reduced as a result of preprocessing of data. The decisive data set is having 8 attributes which are gender, age, weight, fevers, night sweats, (cough>3 weeks), blood phlegm and sputum test. 70% (131 instances) of the data set is used for training and 30% (56 instances) is used for testing. Discretization is applied on the numeric and continuous attributes using Rough Set application. After then, neural network is applied for training the data. Sanchez and et al. (2009)[6] implemented data mining technique to classify TB related handicaps to determine patients' sickness. This study was classifying tuberculosis diagnostic categories based on are used as raw data set. Those 56 attributes are reduced into 5 attributes which are antecedents, bacteriology results, age category, pulmonary tuberculosis and extra pulmonary tuberculosis. Exhaustive CHAID is selected for generating decision trees for classes.

4.2.1 Classification

Classification, which can be described as analyzing of data in order to obtain models that are used to characterize data classes, is the most usual task in data mining. This task focuses on predicting the value of the decision class for an input object within set of classes which are predefined. There are many different classification approaches in the literature. Each of them are developed and proposed by various researchers. The most known techniques are decision tree based classification, neural network based classification, statistical classification, rough set based and genetic algorithm classifiers. We can divide data classification task into two phases. The first phase is called as the learning step, and the second phase is called as testing step. In the learning step, a model which defines predetermined set of classes will be constructed. This operation is made by analyzing a set of training data. For this data, each set of elements are assumed to belong a specific, predefined class. In the testing step, which is the second leg of classification, the constructed model is tested by using different set of data. In this phase, the accuracy of the classification is estimated by using one of the several proposed approaches. If the estimation of accuracy shows an adequate result, then the generated model can be used for classification of new input sets whose class labels are unknown. Before applying the generated classification model, some data preprocessing techniques can be executed in order to obtain a better accuracy and efficiency for the classification model.

a) **Data Cleaning**

The removal of noisy data and filling of missing data is considered in this step. There are many different approaches designed by researches in the literature for data cleaning issue.

b) **Feature Selection**

In the initial data set, there may be some attributes which are not related or not important for the learning step of model. The removal of inappropriate and unnecessary attributes from the data set is applied on this step of classification. After this process, the reduced data set is used to generate the classification model. For feature selection task, numerous applications are implemented by various researchers.

c) Data Discretization

The data set which will be used by classification algorithms may have some attributes that cannot be handled by the algorithm itself without applying some transformations. Such as, numerical scaled values are needed to be converted into nominal or discrete values in order to make some of the algorithms work correctly. This conversion step is considered in the data Discretization part of classification.

4.2.2 PREPARING TUBERCULOSIS DATA SET

Data set that we used in this study contains information about 667 patients who were examined at a private clinic. Each of those records consists of 30 different variables.

We categorized all of these parameters into 3 groups: clinical findings, medical laboratory findings and radiological findings.

If we take a closer look about the meaning of these parameters we can briefly explain them as follows: In clinical findings, the gender parameter indicates whether the patient is male or female. Age group parameter indicates the age group that patient belongs to. All ages are grouped into 7 classes. These classes are “18-24”,“25-32”, “33-40”, “41-45”, “46-51”, “52-57” and “58+”. Weight parameter indicates the weight of the patient in kilograms. Smoke addiction parameter defines whether the patient is a smoker or not. Rates are grouped into 4 classes. “0” means the patient is not a smoker. “1” means the patient smokes less than 5 cigars per day. “2” means the patient smokes between 6 to 10 cigars per day. And “3” means the patient smokes more than 11 cigars per day. Alcohol addiction parameter indicates if the patient is addicted to any kind of alcohol or not. BCG vaccine parameter shows the whether the patient has BCG vaccine or not. Malaise, arthralgia, exhaustion, unwillingness for work, loss of appetite, loss in weight, sweating at nights, chest pain, back pain and coughing are binary valued parameters. They indicate if these parameters are positive for the patient or not. Hemoptysis means coughing up blood from the respiratory tract. This parameter identifies whether the patient has hemoptysis or not. Fever is classified into 3 categories: “0” means normal fever value which is nearly 36.5 degrees Celsius,“1” means fever value is in high ranges and “2” means subfebrile fever value which does not exceed 38.5 degrees Celsius.

In medical laboratory findings we categorized some of blood and skin tests’ parameters. PPD parameter identifies whether the patient has the result of the PPD test positive (labeled as“1”) or negative (labeled as “0”). Erythrocyte is the red blood cells. They are responsible for delivering oxygen to the body tissue. We grouped this parameter into 3 categories. “0” means erythrocyte level is in normal range (about 4.5 to 5 million permicroliter). “1” means low (which is less than 4.5 million permicroliter) and “2” means patient has high erythrocyte level (more than 5 million per microliter). We also grouped the Haematocrit parameter into 3 categories. “0” means the patient has normal haematocrit percentage (about 40% to 45%). “1” means low (less than 40%) and “2” means patient has high haematocrit percentage (more than 45%). Haemoglobin is the iron-containing oxygen-transport metalloproteinase in the red blood cells. In our parameter values, “0” means the patient has normal haemoglobin values (about 14 to 16 g/dl). “1” means low (less than 14 g/dl) and “2” means the haemoglobin value is considered as high (more than 16 g/dl). Leucocytes are white blood cells. They are responsible for defending the body against infections, diseases etc. In our parameter values, “0” means the patient has normal leucocyte values (about 5000 to 10000 in 1mm3 blood). “1” means leucocyte values are low (less than 5000 in 1mm3 blood) and “2” means the leucocyte value is considered as high (10000 in 1mm3 blood). Number of leucocyte type parameter shows the density of leucocytes. If there is a normal density (labeled as “0”) or a lymphocytic density (labeled as “1”) or macrophage density (labeled as “2”). Sedimentation parameter is a measure of the settling of red blood cells in a tube of blood during one hour.

Table -4: Full List of Variables

Clinical Findings	Medical laboratory findings	Radiological findings
Gender	Loss of appetite	Erythrocyte
Age group	Loss in weight	Haematocrit
Weight	Sweating at nights	Haemoglobin
Smoke Addiction	Chest pain	Leucocyte
Alcohol Addiction	Back pain	Number of leucocyte types
BCG vaccine	Coughing	Active specific lung lesion
Malaise	Hemoptysis	Calcify tissue

Arthralgia	Fever	Cavity
Exhaustion	Sedimentation	Pneumonic infiltration
Unwillingness for work	PPD	Pleural effusion

The full list of those variables is based on World Health Organization's standard of Direct Observation of Therapy is given in table 4,

In radiological findings, active specific lung lesion parameter indicates whether there is a radiological proof of a tuberculosis lung lesion on the patient or not. Calcific tissue shows that whether the patient has had tuberculosis disease before. If this parameter is positive, it indicates that the patient has had tuberculosis disease at least once. Cavity parameter states if there are opening-like lesions on the patient's lung or not. Positive value means those kinds of lesions exist on lung. Pneumonic infiltration parameter is positive if a pneumonia-like lesion is seen on the chest x-ray of patient. Pleural effusion means the accumulation of excessive pleural fluid in pleura. This parameter is positive if such thing is seen on the chest x-ray of patient. Before generating ANFIS model, attribute ranking function is applied using information gain ranking filter in WEKA platform. By applying this function, we chose the most important parameters that will affect the fuzzy model mostly. The variables which are ranked less than 10% were eliminated. According to this reduction on the data set, BCG vaccine, Arthralgia, chest pain, smoking addiction, gender, malaise, coughing, back pain, alcohol addiction and pleural effusion variables were ignored.

4.2.3 METHODS

In this research, ANFIS, Multilayer Perceptron and PART methods are used for classification. The following subsections contain brief descriptions about these methods.

i) Adaptive Neuro Fuzzy Inference System (ANFIS)

ANFIS is a neural-fuzzy system which contains both neural networks and fuzzy systems. A fuzzy-logic system can be described as a non-linear mapping from the input space to the output space. This mapping is done by converting the inputs from numerical domain to fuzzy domain. To convert the inputs, firstly, fuzzy sets and fuzzifiers are used. After that process, fuzzy rules and fuzzy inference engine is applied to fuzzy domain. The obtained result is then transformed back to arithmetical domain by using defuzzifiers. Gaussian functions are used for fuzzy sets and linear functions are used for rule outputs on ANFIS method. The standard deviation, mean of the membership functions and the coefficients of the output linear functions are used as network parameters of the system. The summation of outputs is calculated at the last node of the system. The last node is the rightmost node of a network. In Sugeno fuzzy model, fuzzy if-then rules are used.

i) Multilayer Perceptron

An artificial neural network is a simulation system based on mathematical models. Those systems are called as "neural networks" because their working principles are inspired from biological neural networks. Artificial neural networks are basically non-linear statistical data modeling tools. They usually have many inputs (each having different weights) and one output. A neural network has multiple layers. Those layers are mostly input layer, hidden layer and output layer. At input layer, the network gets its values from a vector of variables. At the hidden layer, each input is multiplied by their weight and the results are summed to produce a combined value. And then, this value is fed into a function which will generate the output of the network. Multilayer perceptron is an artificial neural network which has a feed forward structure. Feed forward means that the values only move through the network layers, no resultant values are fed back to any previous inner network layer. A multilayer perceptron network must have an input and an output layer. But the number of hidden layers may change due to the network architecture.

ii) Partial Decision Trees

A partial decision tree is indifferent from conventional decision trees which are having branches to other sub-trees. To generate this kind of tree, a recursive algorithm is required to divide the instances into smaller subsets. The rules for partial decision trees are generated different than standard approach. Rule generation process is done by building a pruned decision tree for the current set of instance and the leaf which has the largest coverage is promoted as a rule. The partial tree generation contains iteration for every subset. On each step, the selected subset is expanded. This process is repeated until there is no subset left unexpanded.

iii) Findings

In this section, the results of this research will be stated. As stated in the previous sections, different models on the data were applied which are ANFIS, Multilayer Perceptron, and PART (a Partial Decision Trees algorithm implementation). Each model generated close results to each other but in an overall point of view, ANFIS has the best scores when compared to other methods. The table 5 summarizes the test data benchmarking results for these methods.

If we consider the overall scores, ANFIS generated best results for testing data as it is clearly shown on Table

2. It is also clear that results of both Multilayer Perceptron and PART methods are very close to ANFIS. When comparing sensitivity and correctness, ANFIS has obviously better scores. ANFIS, Multilayer Perceptron and PART algorithms generated RMSE values of 0.18, 0.19 and 0.20 respectively. Since a lower RMSE value shows more reliability in testing of data, it can be stated that ANFIS scored best RMSE result among other methods.

Table -5: Testing Result of Models

Method Name	Sensitivity	Specificity	Precision	Correctness
ANFIS	0.95	0.97	0.89	0.97
Multilayer Perceptron	0.89	0.97	0.90	0.89
PART	0.85	0.96	0.87	0.85

5. CONCLUSION

This paper aimed to summarize the various categories of medical issues and the comparison between different data mining application in the healthcare sector for extracting useful information. The prediction of diseases using Data Mining applications is a challenging task but it drastically reduces the human effort and increases the diagnostic accuracy. Developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise. Exploring knowledge from the medical data is such a risk task as the data found are noisy, irrelevant and massive too.. It is observed from this study that a combination of more than one data mining techniques than a single technique for diagnosing or predicting diseases in healthcare sector could yield more promising results. The comparison study shows the interesting results that data mining techniques in all the health care applications give a more encouraging level of accuracy high.

REFERENCES

[1] The Free Dictionary by Farlex, 2010

[2] Folorunso O and Ogunde A O, Data Mining as a technique for knowledge management in business process redesign, The Electronic Journal of Knowledge Management, 2(1) (2004) 33-44, Available online at:www.ejkm.com

[3] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001

[4] Larose D T, Discovering knowledge in data: an introduction to data mining, (John Wiley; New York), 2005.

[5] Bakar AA, Febriyani F. Rough Neural Network Model for Tuberculosis Patient Categorization. In: Proceedings of the International Conference on Electrical Engineering and Informatics; 2007; Indonesia. p. 765-768.

[6] Sánchez MA, Uremovich S, Acrogliano P. Mining Tuberculosis Data. In: Berka P, Rauch J, Zighed DA, editors.Data Mining and Medical Knowledge Management: Cases and Applications. New York: Medical Information Science Reference; 2009

[7] Atiq Islam, Syed M. S. Reza, and Khan M. Iftekharuddin, Senior Member, IEEE , "Multifractal Texture Estimation for Detection and Segmentation of Brain Tumors" IEEE Transactions On Biomedical Engineering, VOL. 60, NO. 11, NOVEMBER 2013.

[8] T.Rajesh, R. Suja Mani Malar "Rough Set Theory and Feed Forward Neural Network Based Brain Tumor Detection in Magnetic Resonance Images" Proceedings of the International Conference on Advanced Nanomaterials & Emerging Engineering Technologies" (ICANMEET-2013).

[9] Abdul Kalam Abdul salam,"Assessment on brain Tumour Detection using Rough set Theory" International Journal of Advance Research in Computer Science and Management Studies,Volume 3, Issue 1, January 2015.

[10] ShusakuTsumoto and Shoji Hirano, ―Temporal Data Mining in Hospital Information Systems‖.

[11] Arun K Punjari, ―Data Mining Techniques‖, Universities (India) Press Private Limited, 2006.

[12] Margaret H.Dunham, ―Data Mining Introductory and Advanced Topics‖, Pearson Education (Singapore) Pte.Ltd.,India. 2005.

[13] PrasannaDesikan, Kuo-Wei Hsu, JaideepSrivastava, ―Data Mining For Healthcare Management‖, 2011SIAM International Conference on Data Mining, April, 2011.

[14] N. AdityaSundar, P. PushpaLatha and M. Rama Chandra, ―Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base‖, International Journal of Engineering Science & Advanced Technology, (2012).

[15] K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010).