MEDICAL DATA MINING: A REVIEW

Amrita Kundu[1], Pallavi[2]

[1,2] *Department of Computer Science, Banasthali University, (India)*

ABSTRACT

Data Mining refers to the mining of useful and interesting patterns from large data sets. Since its advent in the early 1980s data mining has made remarkable progress because of its use in industry, medical science, scientific applications, web etc. Medical data now a day is available in abundance but without proper mining they cannot be used. 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 systems. In this paper light has been thrown on various data mining approaches for efficient management of medical data that can serve the mankind which is the prime motto behind any research work.

Keywords: Data Mining, Decision trees, KDD, Medical data, Naïve Bayes.

I INTRODUCTION

The goal of data mining is to learn from data [4]. The strategy used may vary as per the requirement. Data mining is an interdisciplinary field and is gaining popularity because of exploring Database technology, Information Science, Machine learning and Neural networks along with the Statistical techniques. Though data mining algorithms are not applied on the medical data by common people but the knowledge obtained can be very useful for them if shared with in an understandable form. Some of the applications of data mining on medical data include classification of several medical images like X-ray images or MRI images can serve for better diagnosis of any abnormality in the body, clustering the patient records any chronic disease to obtain the knowledge about the spread of the disease, analyzing data in healthcare or survival chances of a patient suffering from diseases like cancer.

II LITERATURE REVIEW

2.1 In year 2004, Mary K. Obenshain [1] applied data mining in three healthcare arena namely hospital infection control, ranking hospitals and identifying high-risk patients.

1. Hospital infection control- Infectious diseases breakout at different times in different geographic areas and there has been a rapid increase in the number of drug-resistant infections too. A surveillance system has been discussed that uses association rule on culture and patient care data obtained from laboratory information systems to generate useful patterns monthly that can be used by an expert taking care of infection control [1].

2. Ranking hospitals- Mining on the reports from the healthcare providers have been discussed to rank hospitals and healthcare plans. According to the reports provided the hospitals should be ranked according to their standard by the organizations [1].

3. Identifying high-risk patients- A robust data mining and model-building solution has been suggested for identifying the patients tending towards high-risk conditions so that the quality of healthcare of patients can be improved [1].

Critic: The role of association rule is to segregate the uninteresting patterns from the interesting ones using a support-confidence framework. For ranking hospitals it is necessary that the hospitals provide data correctly else patient mortality rate cannot be predicted accurately. Identifying high-risk patients may help decrease their number in future.

2.2 In year 2005, authors Md. Rafiqul Islam et.al. [2]made an attempt to implement data mining in image archiving systems using an algorithm based on an inductive decision tree to learn the attributes of lung cancer. Tomograms of 250 patients were considered for the experiment from which patients with pulmonary nodules of size up to 5 cm were chosen . Before application of the algorithm the images were preprocessed and feature subset was selected for enhancing the accuracy of experiment and reducing unnecessary data.

Critic: There is abundance of image data but only applying proper data mining techniques can help extract the attributes necessary to study any disease with accuracy.

2.3 In year *2006,* authors Abdelghani Bellaachia et.al have given their study [3]. Breast cancer is a common form of cancer now a day. The chances of survivability of cancer patients vary from case to case and the stage to stage. The authors used the Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithm to predict that and compared the three.

TABLE-1 Classification techniques used with their accuracy

CLASSIFICATION TECHNIQUE	ACCURACY (%)
Naïve Bayes	84.5
Neural Net	86.5
C4.5	86.7

Critic: A decision tree is similar to flow chart having a tree structure where tests are performed at each level except the last one which gives the result. It is a powerful means for classifying a data set. Predicting the survival time of a cancer patient by this means can actually help the doctor to plan further treatment procedures and the patient to take care of himself so that the time can be extended as much as possible.

Fig. 1. A decision tree

2.4 In year 2007, Ghim-Eng Yap et. al.[4] had given their study:

a) Whenever we deal with data sets errors are evident. These errors may have considerable effect on the output of results when not diagnosed and this effect may be very serious when it's about the medical diagnosis of a disease. In this paper the authors made an attempt that deal with this issue using a knowledge discovery approach by Bayesian network learning [4]. Also a novel procedure for handling the error is implemented to deal with the uncertainty caused by them. This ultimately leads to reliable distinction between cancer affected (here ovarian cancer) and normal patients. The use of Bayesian network has been shown in three steps:

1. Presenting serum expression profile to a learned Bayesian network in case of an unseen sample. .
2. Based on it the network is allowed to update the posterior probabilities of all the nodes it has.
3. Concluding if a patient is suffering from cancer or not based on the largest posterior probability.

Before this erroneous markers are discovered and randomness in the behavior of studied protein is eliminated using a threshold value which is considered 0.1.

Critic: The biological data if noisy can lead to major problem in diagnosis result while managing this noise can give equally good result in cancer detection even with the use of only ten proteins. Such type of tool discussed can prove to be very efficient in screening the patients.

b) Several diseases have a hairline difference between them and are considered as similar. However distinguishing them is necessary to provide the most appropriate treatment to the patient. In this paper an attempt has been made by the authors to meet this challenge using an open source data mining toolbox. Here the two diseases considered are thrombolic brain stroke and embolic brain stroke as given by Petra Kralj et.al. [5]. The experiment included collecting the CT of total 290 patients belonging to the above mentioned categories and using decision tree induction to obtain the contrast sets. Other than this the physical examination of the patients, their ECG, laboratory and other diagnosis measures were also considered [5]. Then data mining was applied to mine the contrast sets to classify the patients belonging to the two categories. This can be done using the Apriori algorithm. A learning rule is finally applied.

Critic: Subgroup discovery approach involves the mining of the interesting groups with the help of a target variable. In this the predictive and descriptive induction are combined. Contrast set mining is used to differentiate between the contrasting groups in a data set. It can be implemented using special algorithm STUCCO or decision tree induction and learning rule[5]. The authors performed contrast set mining through subgroup discovery to differentiate between the patients belonging to the two different classes.

2.5 In year 2008, Dimitris Bertsimas [6] helped with their comments:

a) The most common problem faced by the patients is not the disease they are suffering from but the expenditure of treatment. The common people are the main victims of this problem. So the authors made an attempt to estimate the medical cost and using the past cost pattern to predict the future cost. The data set used was the medical and pharmaceutical claims data for 838,242 individuals. The diagnostic, procedure and drug related information are present in the claims [6]. Cost bucketing is performed on the samples and the Baseline method is used along with classification trees and clustering to group the members with similar cost characteristics [6].

Critic: Care must be taken of the errors for accurate prediction of medical costs. Here the measures of errors used are the hit ratio, the penalty error and the absolute prediction error (APE) [6]. This method can be used by the insurance companies too for pricing the health insurances. Also the comparison with past expenses can be beneficial in predicting the rise in future too.

b) The authors in this paper tried to use sequential data mining to prevent atherosclerosis, a disease characterized by thickening of the arterial walls by deposition of white blood cells. The risk factors of the disease have been taken into account. Case study: STULONG has been used which is specially done to detect the risk factors and other related conditions of atherosclerosis. The windowing approach is used to decompose the data into several disjoint windows. OAR algorithm and AR mining procedure are applied for mining association rules as given according to Jiˇr´ı Kl´ema et.al. [7]. Episode rule mining is used to mine the sequential patterns.

Critic: Sequential data mining deals with discovering patterns in data that are statically relevant. Sequential data has significant potential in medical science but they are not present in a form suitable for the direct application of general data mining algorithms.

Fig. 2. Atherosclerosis

2.6 In year 2009, the author Ruban D. Canlas Jr [8] explains the use of data mining techniques in medical research and public health. Through data mining fraudulent insurance claim can be detected, better health-policy making health centers can be analyzed. In Health Sector Use of data mining and KDD was introduced in this area by Wilson et al [8]. The importance of data mining in medicine and public health can be summarized as-

1. Data Overload: There is a bulk of computerized health record. The medical breakthroughs have slowed down and complexity of the present-day medical information is high, so data mining is best-suited to discover the knowledge. [8]

2. Evidence-based medicine and prevention of hospital errors: In this we apply data mining on their existing data; it can discover useful and life-saving knowledge. By mining hospital record we reduce the rate of error that have made. [8]

3. Policy-making in public health: Open source java based data mining tool Weka and J48 [8] is used to find out the similarities between the community health centers. Data mining and decision support methods, can lead to better performance in decision making. [8]

4. More value for money and cost savings: Through data mining we extract more knowledge from existing data at minimal cost. Data mining is applied to discover knowledge about fraud in credit cards and insurance claims. [8]

5. Early detection and/or prevention of diseases: Classification algorithm is used for early detection of heart disease and heart related disease is the major public concern in the world. Data mining tool is to aid in monitoring trends in the clinical trials for cancer vaccines. [8]

6. Early detection and management of pandemic diseases and public health policy formulation: Health experts have decided to apply data mining in early detection and management of pandemics diseases. In this we have applied simulation and spatial data mining to find interesting characteristics of disease out-break. In 2005 introduced WSARE [8], an algorithm which is used to detect outbreak in early stages. [8]

7. Non-invasive diagnosis and decision support: Thangvel et al. (2006) decided to use K-means clustering algorithm to analyze the cervical cancer patients. The predictive result concluded from K-means clustering algorithm is better than existing medical opinion. Data mining is used to enhance computer-aid diagnosis and endoscopic ultra sonographic elastography and this create new non-invasive cancer detection [8].

8. Adverse drug events (ADEs): Data mining used in US food and drug administration to discover knowledge about drug side effects in their database. The algorithm which is used is called MGPS (Multi-item Gamma Poisson Shrinker) and it was able to detect 67% of ADEs five years before than they were detecting through traditional approach. [8]

Critic: Application of data mining techniques in medical field is a big challenge due to the changing behavior from patient to patient. But still data mining algorithms are used quite successfully to extract better information in case of fraud detection in health policies and as well as in early detection of disease such as heart and pandemics from existing data which is stored in the medical database.

Critic: Swarm Intelligence (SI) is the collective behavior of decentralized, self-organized systems natural or artificial introduced by Gerardo Beni and Jing Wang in the year 1989 and is used in artificial intelligence.

Fig. 3. Swarm Intelligence in fishes.

Comparison table-

TABLE-2 Algorithms used in the paper and their significance

Algorithm Used	Significance
ACO	Works well on nominal and categorical attributes
PSO	Low in Spatial Complexity
Combined ACO/PSO	To find one classification rule at a time

2.7 In Year 2011, Veenu Mangat, [9] the authors have discussed the use of data mining in medical records for enhancement of strategic decisions. Weka version 3.6.0 is used as software for the data mining analysis. Data set were converted into Weka data format after that different associative algorithm was executed on data set and their results were comparatively analyzed. There are four algorithms in which comparative analysis is done. After comparative analysis it was found that predictive Apriori is best fitted algorithm in strategic decisions with highest and lowest accuracy values being 0.99498 and 0.9733 Yilmaz GOKSEN et.al. [10].

Apriori: It generates association rules by using frequent item sets and it generate longer candidate item sets from shorter ones. It reduces minimum support until it find required number of association rules. [10]

Filter Associator: It passess the data through filter before it reaches to the associator. User can configure both base associator and filter. [10]

Predicitive Apriori: To find out best ‘n’ association rule it merges confidence and support into single measure of predictive accuracy. [10]

Tertius: Rules are find according to a confirmation measure. Rules were seeked with multiple conditions like apriori difference is that conditions are applied in OR operations together instead of using and opertion [10].

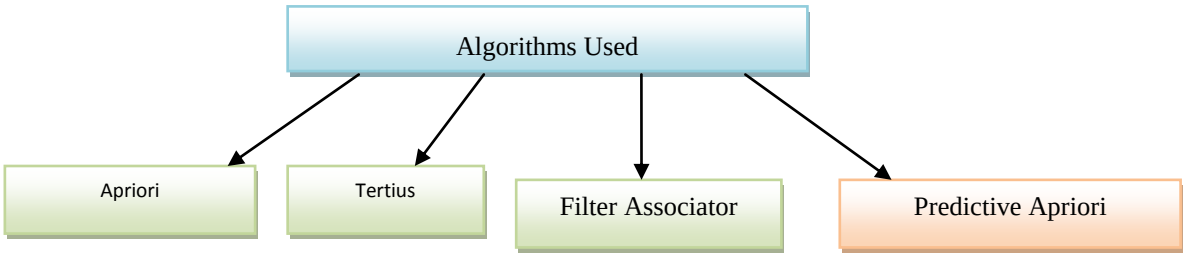

Fig. 4. Algorithms discussed in the paper.

Critic: We have applied association algorithm to extract association rule, but it is noticed sometimes data sets itself might not be suited for association tasks in data mining. No data mining tools and model will provide 100% accuracy by itself in pure and robust automated system. We need experts and managers for checking the final result.

2.8 In year *2012*, Shweta Kharya [11],has discussed some classifiers*:*

a) In the paper author have explored the applicability of decision trees to find a group with high-susceptibility of suffering from breast cancer. One of the most widely used and practical methods for classification is decision tree learning. To generate best decision tree be have used different parameter like confidence factor, pruning. For anomaly detection, classification and automatically categorize medical images on real mammograms two data mining techniques, association rule mining and neural networks is used. There are certain classifier which is explained in the paper and used for diagnosis and prognosis of cancer disease. [11]

1. Association rule based classifier: In this we find out association rule for the medical data. One basis of association rule classification system is constructed and it will categorize the mammograms as normal, malign or benign. [11]

2. Neural network based classifier system: In neural network we classify the medical data set and train the neural network with breast cancer data base by using feed forward network and back propagation learning algorithm with momentum and learning rate. [11]

3. Naïve Bayes Classifier: Analysis of the prediction of survivability rate of breast cancer is done by using naïve Bayes, back- propagated neural network, and decision tree [11].

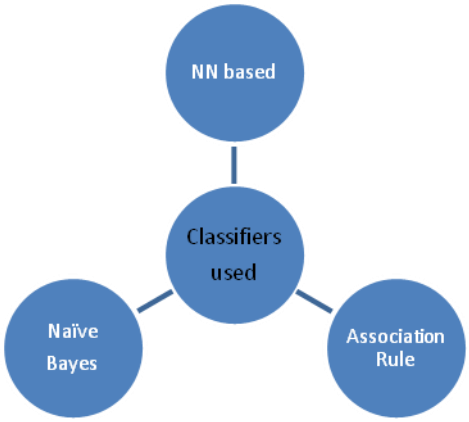

Fig. 5. The classifiers discussed in the paper.

Critic: In diagnosis and prognosis of cancer disease we have used different data mining classifiers and soft computing approaches. Decision tree is found to be best predicator with greater amount of accuracy as compared to other techniques.

b) In this paper the authors have discussed the issue that availability of huge amounts of medical data leads to need for powerful data analysis tools to extract useful knowledge. Five data mining techniques classification by decision tree induction, Bayesian classification, Neural networks, Support Vector Machines (SVM) and classification based

on associations have been analyzed and it was found that decision tree and SVM are most effective for heart disease diagnosis as given by Aqueel Ahmed et.al. [12]. Various classification techniques for data mining are examined and reported that data mining technique shows the 92.1 % - 91.0 % accuracy for the heart diseases. Here usage of various parameters (age, fasting blood sugar, sex etc.) increases the accuracy for the heart disease patient using data mining techniques.

Critic: Healthcare system has huge data available but effective analysis of those using proper tools is the job of data mining techniques. Finding out the hidden relationships and patterns in the data can help predicting a disease very accurately.

2.9 In year 2013, Taranath NL et.al.[13] focused on novel framework of medical decision support system for missing data. In this author have used the machine learning approach, automatic learning is used in certain tasks such as medical imaging. In this there are two tasks to be performed: identification and extract informative sentence on disease , fine grained classification according to semantic relation of the sentence on the basis of diseases and treatment. [13]

Citric: In order to offer the required accuracy for this domain both the approaches of integrating machine learning and ontological reasoning is found best approaches. The framework of medical decision support system is emerging application in medical sector.

b) The paper contain the idea of developing a prediction model that can predict heart disease cases based on measurements taken from transthoracic echocardiography examination. KDD has been used. Prediction system built with aid of data mining techniques like decision trees, naïve Bayes and neural network [14]. The models used along with the accuracy measures are:

TABLE-3 Models used in the paper and their accuracy

Model Used	**Accuracy(%)**
J48 unpruned with all attributes	94.29
J48 pruned with all attributes	95.41
Naive Bayes with all attributes	91.96
Naive Bayes with selected attributes	92.42

Citric: The goal of paper is to develop prediction model, but researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy. On the basis of this model that can predict specific heart diseases.

2.10 In year 2014, Matthew Herland et.al. given their review on big data in the field of health informatics[15]:

a) Since the amount of data in health informatics is very vast, the authors in this paper discussed the concept of big data analysis for improving the quality of healthcare. Problems regarding analyzing such huge data in a reliable

manner are also discussed. Big data tools and approaches at various levels are dealt with live molecular, tissue, patient and population [15].

TABLE-4 Summary of studies in paper

Data level	Question level(s) answered
Molecular	Clinical
Tissue	Human scale biology, Clinical
Patient	Clinical
Population	Epidemic-scale, Clinical

Critic: In future work needs to be done on handling vast amounts of gene probe and select the kind of subset that can provide the best correlation. Difficulty is faced while dealing with MRI data that are high resolution data and numerous samples are required to perform the research work of acceptance level with those.

b) Apriori algorithm is a general algorithm in data mining .In the paper , year 2014 theauthors Gitanjali J et.al.[16], the authors used this algorithm to find out the frequency diseases using the data set of months, diseases and the instances of their occurrences. Experimental results are shown in the form of bar graphs to show frequency of occurrence of different diseases. The steps used are:

1. Scanning data set D to generate list of candidates [16].

2. Compare candidate support count with min support count [16].

3. Scan D for count of each candidate and compare candidate support count with min support count remaining in list [16].

Critic: The frequency of occurrence of diseases varies with geographic area, habits of people, season etc.

III CONCLUSION

Medical data available is huge and hence analyzing or classifying them with efficiency is a challenge. Also high resolution data like MRI are to be mined in large amount for predicting critical diseases of brain with accuracy. Several classifiers have been implemented till date including artificial neural networks but proper learning technique must be implemented to get the desired results. In the presence of all this challenges also data mining algorithms have shown eye catching results in different domains of medical science but more work is required to be done in managing different variety of diseases, major or minor because mostly minor disease affect people which may turn major if not treated correctly on time.

REFERENCES

[1] Mary K. Obenshain, “Application of Data Mining Techniques to Healthcare Data,” Statistics for Hospital Epidemiology. North Carolina, vol. 25 No. 8, pp. 690-695, August 2004.

[2] Md. Rafiqul Islam, Morshed U. Chowdhury and Safwan Mahmood Khan*, Medical Image Classification Using an Efficient Data Mining Technique. Dhaka, Bangladesh, vol.12, Complexity International, 2005, pp.01-09.

[3] Abdelghani Bellaachia and Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques," Washington DC 20052, The George Washington University, 2006, pp. 01-04.

[4] Ghim-Eng Yap, Ah-Hwee Tan and Hwee-Hwa Pang, "Learning Casual Models for Noisy Biological Data Mining: An Application to Ovarian Cancer Detection," Singapore 639798. Association for the Advancement of Artificial Intelligence 2007, pp.354-359.

[5] Petra Kralj1, Nada Lavra·c1;2, Dragan Gamberger3, Antonija Krsta·ci, "Contrast Set Mining for Distinguishing Between Similar Diseases," Slovenia and Croatia, Technology Project "Knowledge Technologies",2007.

[6] Dimitris Bertsimas, Margrét V. Bjarnadóttir, Michael A. Kane, J. Christian Kryder, Rudra Pandey, Santosh Vempala and Grant Wang, "Algorithmic Prediction of Health-Care Costs," Operations Research, vol.56, No.6, pp. 1832-1892, November-December 2008.

[7] Ji˘r´ı Kl´ema, Lenka Nov´akov´a, Filip Karel, Olga ˇ St˘ep´ankov´a, and Filip ˇ Zelezn´y, "Sequential Data Mining: A Comparative Case Study in Development of Atherosclerosis Risk Factor," IEEE Transactions on System, Man and Cybernatics. vol. 38 ,No.1, pp. 01-12, January 2008.

[8] Ruban D. Canlas Jr., "Data Mining in Healthcare:Current Applications and Issues"MSIT, August 2009

[9] Veenu Mangat, "Swarm Intelligence based Technique for Rule Mining in the Medical Domain"International Journal of Computer Applications ,vol. 4, no.1, July 2010

[10] Yilmaz GOKSEN, Mete EMINAGAOGLU and Onur DOGAN, "Data Mining in Medical Records for the Enhancement of Strategic Decisions:A Case study"Scientific Bulletin – Economic Sciences, Vol. 10 / Issue 1,2011

[11] Shweta Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease",International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[12] Aqueel Ahmed, Shaikh Abdul Hannan., "Data Mining Techniques to Find Out Heart Diseases,"International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.

[13] Taranath NL, Dr. Shantakumar B Patil, Dr. Premajyothi Patil, Dr. C.K. Subbaraya, "Medical Decision Support System for the Missing Data using Data Mining-A Survey"International Conference on Contemporary Computing and Informatics,2014 IEEE.

[14] Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques" Oriental journal of Computer science & technology**,** December 2013, Vol. 6, No.4: pp. 457-466.

[15] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald*,"A Review of Data Mining Using Big Data in Health Informatics ," Herland et al. Journal of Big Data 2014, 1:2, USA.

[16] Gitanjali J, C.Ranichandra,M.Pounambal, "APRIORI Algorithm Based Medical Data Mining for Frequent Disease Identification," International Journal of Information Technology (IIJIT), vol.2, issue 4, April 2014.

Year	**Technique used**	**Objective of the paper**
2004	Decision trees, Neural Network,	Applications of data mining techniques to health care data.

	Naïve Bayes classification, Logistic Regression	
2005	Decision trees, Neural Network	Classification of medical image.
2006	Weka toolkit, C4.5 decision tree	Predicting Breast cancer survivability.
2007	a) Clustering b) KDD using Data mining tools	a) Analysis of clinical courses of chronic Hepatitis. b) Managing noisy data while ovarian cancer detection.
2008	a) Bucketing, clustering, classification trees b) Trend analysis, windowing	a) Predicting health-care costs. b) Development of Atherosclerosis risk factor.
2009	KDD using Data mining tools	Studying current applications and issues related to healthcare.
2010	Swarm Intelligence algorithms	Rule mining in medical domain.
2011	Estimation, prediction, classification, clustering, association	Mining medical records for the enhancement of strategic decisions.
2012	a) Classification, Neural Network, Association rule mining, C4.5 decision tree, Naïve Bayes. b) Association rule, clustering	a) Diagnosis and Prognosis of cancer disease. b) Finding out heart diseases.
2013	a) Knowledge-based and learning-based system b) Neural network, decision tree, Naïve Bayes	a) Medical decision support system for the medical data. b) Designing heart disease prediction system.
2014	a) Big Data tools b) Apriori algorithm	a) Data mining using big data in health informatics. b) Frequent disease identification.

Annexure

Biographical Notes

Ms. Amrita Kundu is presently pursuing M.Tech. first year in Computer Science Engineering Department from Banasthali Vidyapith, India.

Ms. Pallavi is presently pursuing M.Tech. first year in Information Technology Engineering Department from Banasthali Vidyapith, India.