ARTICLE

SIGNIFICANCE OF DATA MINING TECHNIQUES IN DISEASE DIAGNOSIS AND BIOMEDICAL RESEARCH – A SURVEY

Thara Lakshmipathy[1] and Gunasundari Ranganathan[2]

[1]Department of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, INDIA

[2] Department of Information Technology, Karpagam University, Coimbatore, Tamil Nadu, INDIA

ABSTRACT

In the Biomedical sector, the current applications of data mining are the most successful and they have become the subfield of medical research. The basis for this growth is that nowadays most of the health related data are stored in small datasets scattered through various Hospitals, Clinics and Research centers. The healthcare sector is presently facing both the economic predictability and the methodological prospect of a data based approach for quality administration. Hence, abundant data mining techniques have been proposed to process medical data stream. Applying data mining techniques to the centralized database will offer to doctors investigative and foretelling tools from the surface of the data. Modern clinical practices also undertake transformation not only in diagnosis and treatment methods, but also in understanding of health and illness concepts. Data Mining can generate information that can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices. Though data mining methods and tools have been applied in diverse areas for more than 40 years already, their applications in healthcare are comparatively young. Therefore, this paper presents a survey on the significance of Data Mining techniques in Disease Diagnosis and Biomedical Research.

KEY WORDS
data mining techniques; diagnosis; health care sector; medical diagnosis; classification; decision tree.

Published: 2 Nov 2016

***Corresponding Author**
Email:
[1]kltharavijay@gmail.com,
[2]gunasoundar04@gmail.com
Tel.: +91- 9976907003

INTRODUCTION

Infection Practically in every country, the cost of healthcare is increasing more rapidly than the readiness and the capability to pay for it. At the same time, more and more data is being obtained around healthcare processes in the form of Electronic Health Records (EHR), health insurance claims, medical imaging databases, disease registries, spontaneous reporting sites, and clinical trials. As a result, data mining has become important to the healthcare world. On the one hand, EHR offers the data that gets data miners excited. However, on the other hand, it is accompanied by challenges such as 1) the unavailability of large sources of data to academic researchers, and 2) limited access to data-mining experts [6]. But, still the quality of health care service at a cheap cost continues to be a difficult issue in developing countries. Although several countries have taken firm steps in providing aid services, the reach of those services to the poor, still remains as a difficulty. In different dimensions, un-imaginable services like separation of dual births, discovery of latest medicines for deadly diseases are happening. However, a few cases of death occur due to poor clinical diagnosing system and inadequate treatment. Most of the time, the clinical selections are created by the doctor's experience and skill. Hospitals do have info systems, call Support Systems, Image and Scan process systems. However, all hospitals do not have these facilities or their applications are restricted.

Call support systems with diagnosing support for naive doctors and for knowledgeable doctors would be a tenet for higher clinical cognitive process. The primary initiative is that the prospective costs and effects of diagnosis can be measured and used to enhance the task. Furthermore, the transitional enquiry steps such as asking for more systematic studies, or more examination done by human professionals, can also be computed in terms of cost and effect. Based on these cost and effect, the system can learn to recommend the optimal action. Providing precious services at cheap prices may be a major constraint encountered by the aid organizations (hospitals, medical centers). Valuable quality service denotes the correct diagnosing of patients and providing economical treatment. Acceptable computer-based info and/or call support systems will aid in achieving clinical tests at a reduced value. Due to the accessibility of integrated info through huge patient repositories, there is a swing within the insight of clinicians, patients and payers from qualitative health care sectors.

IMPORTANCE OF DISEASE DISCOVERY & ROLE OF DATA MINING IN CLINICAL DIAGNOSIS

President Obama has recently expressed a strong conviction that "Science offers great potential for improving health". The President has announced a research initiative that aims to accelerate progress toward a new era of precision medicine. The time is accurate for this idealistic initiative, and the National Institutes of Health (NIH) and other partners will work to attain this vision [7]. The concept of precision medicine speaks of prevention and treatment strategies that take individual variability into account. Even though it is not new, the impact of practicality is still deprived. If one considers blood typing, for instance, has been used to guide blood transfusions for more than a century. But the vision of applying this concept broadly has been intensely improved by the latest development of large-scale biologic databases (such as the human genome sequence), powerful methods for typifying patients (such as proteomics, metabolomics, genomics, diverse cellular assays, and even mobile health technology), and computational tools for analyzing large sets of data. Francis S.Collins et al. stated that, "What is necessary now is a broad research program to encourage resourceful methodologies to precision medicine, test them rigorously, and eventually use them to build the evidence base needed to guide clinical practice" [7].

In precision medicine, the proposed initiative has two main components: a near-term emphasis on cancers and a longer-term goal to generate knowledge pertinent to the whole array of health and diseases. Due to advances in elementary research, both components are now within the scope, including molecular biology, genomics, and bioinformatics. Moreover, the initiative taps into congregating trends of improved connectivity, through social media and mobile devices, and Americans' budding desire to be dynamic partners in medical research [7]. The components are:

(i) The cancer-focused component of this initiative will be intended to address some of the hindrances that have already been encountered in "precision oncology" [17]. Precision medicine's more customized, molecular approach to cancer will enrich and modify, but not replace, the successful staples of oncology - prevention, diagnostics, some screening methods and effective treatments while providing a robust framework for accelerating the adoption of precision medicine in other spheres. The most obvious of those domains are inherited genetic disorders and communicable diseases, but there is assurance for many other diseases and conservational reactions.
(ii) The initiative's second component involves pursuing research progresses that will enable better assessment of disease risk, understanding 8of disease mechanisms, and prediction of optimal therapy for many more diseases, with the goal of expanding the benefits of precision medicine into countless aspects of health and health care [17].

Data mining may be a crucial step in discovery from giant data sets. A wide type of areas such as client relationship management, engineering, medicine, crime analysis, knowledgeable prediction, web mining, mobile computing rely on the utilization of data processing. In recent years, data processing has found its important hold in each field including health care. Mining method is over the info analysis which has classification, clustering, association rule mining and prediction. Varied fields related to medical services like prediction of effectiveness of surgical procedures, medical tests, medication, and also the discovery of relationships among clinical and diagnosing information are yet to use data processing methodologies [7].

NEED OF DATA MINING TO DIAGNOSE VARIOUS DISEASES IN MEDICAL HEALTHCARE SYSTEM

It is attainable to predict the effectiveness of medical treatments by strengthening the information mining applications. The real-life data processing applications are desirable since they supply information miners with varied set of issues, time and tests.

Lately, the practicality of data mining techniques has been realized in Healthcare domain. This realization is in the wake of explosion of complex medical data. As on today, lot of research is found in the field of biomedicine. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. The data mining techniques especially are more useful in predicting heart diseases, lung cancer, and breast cancer and so on [15].

Healthcare entities are reluctant to release their internal data to academic researchers and in most cases there is a limited interaction between industry practitioners and academic researchers working on related problems. According to the report [6] from the National Center for Health Statistics of USA, adoption of EHR in the USA as the most prominent medical information system is shown in Figure 1. It elucidates a linearly raising amount of non-sparse, but continual data reflecting patients' clinical continuity together with the treatment which took place and medication being used.

Due to the fact that the healthcare sector is very diverse and its entities as well as actors have different objectives and fields of activities, they employ different methods and tools in their operations [31, 19, 27].

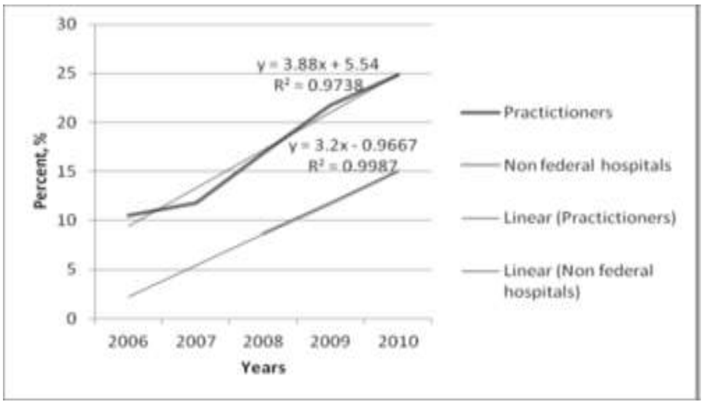

Fig. 1: EHR Adoption in USA

Any proposed automated medical diagnosis system, which is constructed using data mining methodology, will help the doctors to diagnose the correct disease in less time [33]. Table:1 highlights a few of such system's leading objectives along with their associated authors working in the field of predicting medical disease(s) using data mining technique. Common data mining techniques which are used in almost all the sectors are listed as: Naive Bayes, Decision Tree, Artificial neural network (ANN), K- nearest neighborhood (KNN), Support vector machine (SVM) etc. In order to improve the success of healthcare organization and health of the patients, the knowledge gained by data mining can be exercised for booming research problems in the associated arena [33].

Table 1: Details of Related Work Done for Various Diseases

Disease	Objective	Tools & Algorithms used	References
Breast cancer	An Efficient Prediction of Breast Cancer Data using Data Mining Techniques.	Tool: Weka Algorithm:SVM classifier.	[10]
Heart Disease	Diagnosis of Heart Disease using Data mining Algorithm	Tools: Weka & MATLAB Algorithms: Decision Stump, Random Fores and LMT Tree	[35]
	Predictive data mining for medical diagnosis: An overview of Heart disease prediction.	Tool: TANAGRA Algorithms: Naive Bayes, Decision Tree and ANN.	[36]
	Proposed a system for Heart disease prediction using data mining techniques.	Traditional Classification algorithms like j48, Naive Bayes, Neural Networks and GNP	[37]
Heart Disease	Applying K-nearest neighbor in diagnosing Heart disease patients	Decision tree, Classification via clustering and Naive Bayes	[39]
	To design a predictive model for heart disease detection to enhance their liability of Heart disease diagnosis.	Tool: Weka Algorithms: Decision Tree, Neural Network and Bayesian Classifier	[40]
Kidney Stone	Statistical and data mining aspects on kidney stones: A systematic review and meta analysis.	Classification techniques: Decision tree, ANN and Naive Bayes	[38]
Lung Cancer & Skin Cancer	Early prevention and detection of skin cancer and lung cancer risk using data mining.	Apriori Tid, Decision Tree, K-means and Medoid based clustering	[41],[42]
Liver Disorder	Performance evaluation of different data mining classification algorithm and predictive analysis.	Classification of medical data using Bayesian Ying Yang (BYY) Algorithm	[43]
Lung Cancer	Diagnosis of lung cancer prediction system using data mining classification techniques	Rule based Classification algorithm like IF-THEN Rule, Decision tree, Bayesian classifiers and Neural networks.	[44]
Kidney dialysis	Performance comparison of three data mining techniques for predicting kidney dialysis survivability	Tool: TANAGRA Algorithms: ANN, Decision Tree(C5), Logical Regression	[45]
Acute Abdominal Pain	An improved Medical Diagnosing of Acute Abdominal Pain with Decision Tree	Tools: C++ & MATLAB Algorithms: CART & OC1	[46]
Existence of Helicobacter pylori Microbe	Helicobacter pylori microbe and detecting with data mining Algorithms	Tool: Cross Validation model Algorithms: RBF Network, Naïve Bayes, PART, Decision Tree, Logistic Regression.	[47]
Liver Based Diseases	Classification of Liver Based Diseases using Random Tree	Tool: Weka Algorithm: Random Decision Tree	[34]
Gastric Cancer	Clinical Data Analysis Reveals Three Sub types of Gastric Cancer	Tool: MATLAB Algorithm: Hierarchical Clustering.	[32]
	A new algorithm to extract hidden rules of Gastric cancer data based on Ontology	Tool: MATLAB Algorithm: Mixture of Apriori Algorithm and Ontology	[48]

APPLICATION OF DATA MINING TECHNIQUES IN BIOMEDICINE

Well-known data mining techniques include the Artificial Neural Network (ANN), decision tree, Bayesian classifiers, Support Vector Machine (SVM) and many others [16]. The Text Mining Algorithms are designed assuming that a document is a collection of words with rules (model called, bag of words) and applied to perform Text Summarization, Document Retrieval, Text categorization, Document Clustering, Identifying key phrases and Entity and information extraction [26]. S. L. Ting et al. [25] introduced some basic data mining techniques, namely unsupervised learning and supervising learning, and reviewed the application of data mining in biomedicine. The researchers tried to combine both unsupervised and supervised methods for the analysis as advanced Data Mining Techniques, for instance, Hierarchical clustering, C-means clustering, self-organizing maps (SOM), Support Vector Machines (SVM) and multidimensional scaling techniques. Springer 2005 discusses the new concepts, technologies and practices of biomedical knowledge management, data mining and text mining that are beginning to bring useful “knowledge” to Biomedical professionals and Researchers [20].

DATA MINING IN BIO MEDICAL RESEARCHES

The researchers in the medical field determine and predict the diseases besides proffering effective take care of patients with the help of information mining techniques. Several different studies have investigated completely different technologies for the assessment of medical diagnosis systems.

Due to many factors, the diagnosis for sickness or symptoms identified may face a multi-layered drawback and results in false assumptions and erratic report. So it seems affordable to undertake utilizing the information and skill of many specialists collected in databases towards aiding the diagnosing method. The information mining techniques are used by a good type of works within the literature to diagnose numerous diseases including: polygenic disorder, Hepatitis, Cancer, Heart diseases and so on. Information related to the sickness, prevailing within the style of electronic clinical records, treatment info, sequence expressions, pictures and more were utilized altogether for diagnosis. In the recent past, the info mining techniques were used by many authors to gift diagnosing approaches to various kinds of heart diseases.

Historically, the well-mined info is painted as a model of the linguistics structure of the dataset. It would be attainable to use the model within the prediction and classification of the latest information. R. D. Wilson et al. [3] started to classify and collect medical publications where knowledge discovery and DM techniques were applied or researched from 1966 till 2002. According to their study, “…some authors refer to DM as the process of acquiring information, whereas others refer to DM as utilization of statistical techniques within the knowledge discovery process”.

APPLICATIONS OF DATA MINING ALGORITHMS - A LITERATURE REVIEW

The availability of enormous amounts of medical data leads to the need for dominant data analysis tools to extract useful knowledge. There is a lot of data available within the healthcare systems. However, there is a task of finding effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domains. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis of large data sets. Disease diagnosis is one of the applications where data mining tools are producing successful results [2].

A Computer-Aided Detection of prostatic adenocarcinoma in MRI [9]

Prostate cancer is one in every of the most important causes of cancer death for men within the western world. Magnetic Resonance Imaging (MRI) is being more and more used as a modality to find prostatic adenocarcinoma. Therefore, computer-aided detection of prostatic adenocarcinoma in MRI pictures has become a vigorous space of analysis. In this paper the authors tend to investigate a totally automatic computer-aided detection system that consists of 2 stages. In the 1st stage, they found initial candidates victimization multi-atlas-based prostate segmentation, voxel feature extraction, classification and native maxima detection. In the second stage segments, the candidate regions and victimization classification were acquired with cancer likelihoods for every candidate [9]. Options represent pharmacokinetic behavior, symmetry and look among others. The system was evaluated on an outsized consecutive cohort of 347 patients with MR-guided diagnostic test. This set contained 165 cancer patients. Performance analysis relies on lesion-based free-response receiver operative graphical record and patient-based receiver operative characteristic analysis.

Mining Time dynamical information Streams [8]

Most applied mathematics and machine-learning algorithms assume that the information may be a random sample drawn from a stationary distribution. Sadly, most of the massive databases offered for mining nowadays violate this assumption. They were gathered over months or years, and also the underlying processes generating them modified throughout this point, generally radically. Though numerous amount of algorithms are projected for learning time-changing ideas, they typically do not scale well to terribly giant databases. In this paper, the authors tend to associate economical algorithmic program for mining call trees from continuously-changing information streams, which supports the ultra-fast VFDT call tree learner. This algorithmic program known as CVFDT, stays current by creating the foremost of recent information using another sub tree. CVFDT learns a model that is comparable in accuracy to the one that may be learned by reapplying VFDT to a moving window of examples. Each time a brand new example arrives with O(1) complexity per example and hostile O(w), where w is the size of the window. Experiments on a collection of huge time-changing information streams demonstrate the utility of this approach.

Mining time-changing information streams is of great interest. The elemental issues are the way to effectively determine the numerous changes and organize new coaching information to regulate the out-of-date model. Geoff Hulten et al. proposed a vigorous learning system to handle these problems. Whenever the suspected changes are indicated, it exploits a light-weight uncertainty sampling algorithmic program to settle on the foremost informative instances to label. With these tagged instances, it more tests the reality of the suspected changes. If the changes so cause important performance deterioration of the present model, it evolves the recent model. Thus, this technique is sensitive to important changes and strong to clanging changes, and might quickly adapt to concept-drift. Experimental results from each artificial and real-world information ensure the benefits of the system.

Application of Data Mining Algorithms in Heart Disease Prediction

Indians are undoubtedly victims to peculiar kinds of cardiovascular diseases that result in worse outcomes like anaemia cardiovascular disease - a condition characterized by reduced blood provided to the guts. Machine Intelligence is employed for Medical data processing wherever giant assortment of Medical information is well-mined for attention-grabbing pattern. 2% of total world deaths are because of Cardio tube sickness (CVD), which is anticipated to be the leading cause for deaths in developing countries because of modification in life style, work culture and food habits. Hence, additional careful and economical strategies of diseases diagnosis and periodic examinations are of high importance.

Numerous studies have been done that have focus on diagnosis of heart diseases. They have applied different data mining techniques for diagnosis and achieved different probabilities for different methods [14]. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [4]. The resultant dataset contains records of patients having heart diseases. Three constraints were introduced to decrease the number of patterns [22], the result of which diagnoses the presence or absence of heart disease.

1. The attributes have to appear on only one side of the rule.
2. Separate the attributes into groups, i.e. uninteresting groups.
3. In a rule, there should be a limited number of attributes.

Sellappan Palaniappan et al. [21] proposed “An Intelligent Heart Disease Prediction System (IHDPS)” using data mining techniques, Naive Bayes, Neural Network, and Decision Trees. Each method has its own strength to get appropriate results. To build this system, hidden patterns and relationship between them is used. It is web-based, user friendly and expandable.

An overview on “Data Mining Techniques to Find out Heart Diseases” suggested that the medical diagnosis for Heart Diseases is an extremely important but complicated task that should be performed accurately and efficiently. It proposed to find out the heart diseases through data mining, Support Vector Machine (SVM), Genetic Algorithm, rough set theory, association rules and Neural Networks. It concluded that, out of the above techniques, Decision tree and SVM are the most effective ones for the heart disease. So it is observed that the data mining could help in the identification or the prediction of high or low risk heart diseases [2].

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. [18]. The dataset contains records with 13 attributes in each record. The supervised network, i.e. Neural Network with back propagation algorithm is used for training and testing of data.

A “Study of Heart Disease Prediction using Data Mining” [13] points out that the doctors and experts available are not in proportion with the population. Also, symptoms are often neglected. Heart disease diagnosis is a complex task which requires much experience and knowledge. Heart disease is the single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry, the data mining is mainly used for predicting the diseases from the datasets. So far, the Data Mining techniques, namely Decision Trees, Naive Bayes, Neural Networks, Associative classification, Genetic Algorithm have been applied on Heart disease databases.

The healthcare industry collects huge amounts of health connected data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advance data mining techniques can help remedy this situation [24]. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely Decision Trees, Naive Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex "what if"; queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar, it can predict the likelihood of patients getting heart diseases. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the .NET platform [24].

Application of Data Mining Algorithms in Breast Cancer prediction

Breast cancer is one of the leading cancers affecting women when compared to all other cancers [29]. It is the second most common cause of cancer death in women. Breast cancer risk in India revealed that 1 in 28 women develop breast cancer during her lifetime. This is higher in urban areas, being 1 in 22 in a lifetime compared to rural areas where this risk is relatively much lower being 1 in 60 women developing breast cancer in their lifetime. In India the average age of the high risk group is 43-46 years unlike in the west where women aged 53-57 years are more prone to breast cancer [29].

Breast cancer has become the most hazardous type of cancer among women in the world. Early detection of breast cancer is essential in reducing life losses [10]. The accuracy of classification techniques is evaluated based on the selected classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for medical applications. The performance of SVM shows a higher level compared with other classifiers. Hence SVM shows the concrete results of Breast Cancer in the patient records. Therefore, SVM classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance [10].

Abdelghani Bellaachia et al.[1] presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Public-Use Data. The pre-processed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. The problem investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, the result of the research found out that C4.5 algorithm has a much better performance than the other two techniques [1]. This paper has outlined various issues and identified the algorithms and techniques for the problem of breast cancer survivability prediction in SEER database. The authors approach takes into consideration, besides the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD). The experimental results show that their approach clearly shows that the preliminary results are promising for the application of the data mining methods to the survivability prediction problem in medical databases. The analysis does not include records with missing data; future work will include the missing data in the EOD field from the old EOD fields prior to 1988. This might increase the performance as the size of the data set will increase considerably. Finally, the research tried the survival time prediction of certain cancer data such as respiratory cancer where the survivability is seriously low [1].

G. Ravi Kumar et al. [10] presented a comparison among the different Data mining classifiers on the database of Wisconsin Breast Cancer (WBC), by using classification accuracy. This paper aims to establish an accurate classification model for Breast cancer prediction, in order to make full use of the invaluable information in clinical data especially that is usually ignored by most of the existing methods when they aim for high prediction accuracies. Experiments were done on WBC data. The dataset was divided into training set with 499 and test set with 200 patients. In this experiment, they compared six classification techniques in Weka software and comparison results showed that Support Vector Machine (SVM) has a higher prediction accuracy than those methods. Different methods for breast cancer detection were explored and their accuracies are compared. With these results, it was concluded that the SVM are more suitable in handling the classification problem of breast cancer prediction, and recommended the use of these approaches in similar classification problems [10].

The aim of this paper is also to investigate the performance of different classification techniques. They have analyzed the breast Cancer data available from the Wisconsin dataset from UCI machine learning with the aim of developing accurate prediction models for breast cancer using data mining techniques. In this research, a total of 683 rows and 10 columns had been used to test, by using classification accuracy. The experiment result proves that, they compared three classification techniques in Weka software and comparison results shows that Sequential Minimal Optimization (SMO) has a higher prediction accuracy, i.e. 96.2% than IBK and BF Tree methods.

Application of Data Mining Algorithms in Gastric & lung Cancer prediction

The problem of “Discovering Association of Diseases in the Upper Gastrointestinal Tract using Text Mining Techniques” measures the negative health impacts on acidic environment of the stomach. Historically, it was widely believed that the highly acidic environment of the stomach would keep the stomach immune from infection. Having too little or no gastric acid is known as hypochlorhydria or achlorhydria respectively and are conditions which can have negative health impacts [26]. Gastric cancer is the fourth most common cancer and second leading cause of cancer-related death worldwide [32]. Gastric cancer, or stomach cancer, refers to tumors that develop in the lower part of the esophagus, in the stomach, or in the uppermost part of the small intestine. It is the 4th most common cancer and second leading cause of cancer-related death worldwide [28].

The clinical data of over 1,500 gastric cancer patients was analyzed. It is interesting to find that gender is a major factor for gastric cancer subtype characterization. Actually several types of cancer, including stomach, liver, and those of colon, are far more common in men than in women. Some scientists have hypothesized that variances in lifestyle such as diet and smoking, may account for the role of gender factor. On the other hand, growing evidence also submits that the variances are rooted in basic biological differences between men and women. For example, recent research indicates that “Estrogen” protects against gastric cancer [23].

Cancer research is commonly clinical and/or biological in nature, and data driven statistical research has become a command counterpart. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. Cancer is the most vital reason for death of both men and women. The early detection of cancer can be supportive in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is growing. Lung cancer is a disease that is commonly misdiagnosed. Early diagnosis of Lung Cancer saves many lives, failing which may lead to other severe problems causing sudden incurable end. Its cure proportion and prediction depends mainly on the early detection and diagnosis of the disease [30].

Decision Tree results are easier to read and interpret. The drill through feature to access detailed patients’ profiles is only available in Decision Trees. The decision tree shown in [Fig. 2] was built from the very small training set [Table 2]. In this table each row corresponds to a patient record. The data set contains 3 predictor attributes, namely age, gender, intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient has a definite disease or not [30].

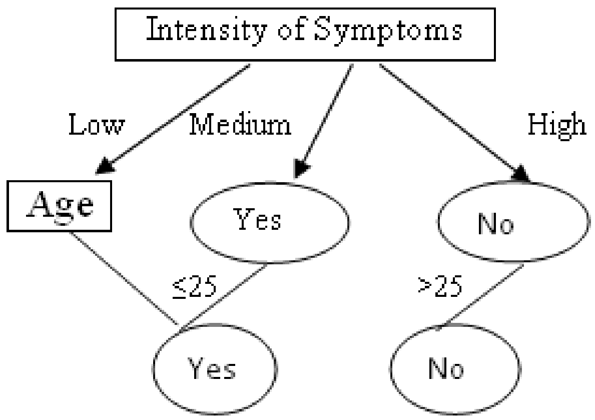

Fig 2: Decision tree

Table 2: Data set used to build decision tree of Fig: 2

Age	Gender	Symptoms	Disease
25	Male	Medium	Yes
32	Male	High	Yes
30	Female	Low	No
21	Male	Low	No
34	Male	Medium	No
18	Female	Low	No
55	Male	Medium	No

In DM, IF-THEN prediction rules are very popular, which signify the discovered knowledge at a high level of abstraction. In the health care system, it can be applied as: (Indications) (Earlier--- history) → (Cause-of-disease).

Example: If_then_rule induced in the diagnosis of level of alcohol in blood.

IF Sex = MALE AND Unit = 8.9 AND Meal = FULL THEN Diagnosis = Blood_alcohol_content_HIGH.

CONCLUSION

Application of Data mining Algorithms in healthcare industry plays a major role in prediction and diagnosis of various diseases. This paper plotted the need and application of Data mining in several disease discoveries and outlined the associated research problems which were proposed in the Prediction and Detection of diseases using Data Mining Algorithms. Thus the survey presents the Significance of Data Mining Techniques that had been employed for Bio Medical Research. Therefore, it was initially agreed to define the scope of this paper as Data Mining applications in the healthcare providers' institutions. The paper screens the commonly accepted belief that data mining is widely used in medicine by comparing academic advances with practical achievements in the field. This will positively help when incorporated into a knowledgeable decision making system.

FUTURE DIRECTION

In future, many diseases can be diagnosed in terms of various parameters through the patient's symptoms gathered from the clinical data centers. An intelligent disease diagnosis system using the novel hybridized classification approach may be developed. In this hybridization approach, clustering may be done before classification and in the iteration of classification, data pruning may be done. By doing so, better classification accuracy can be obtained.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Purohit Abdelghani Bellaachia, Erhan Guven.[2006] Predicting Breast Cancer Survivability Using Data Mining Techniques, Department of Computer Science -The George Washington University Washington DC 20052, 58(13).

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

[3] A Wilson, L Thabane, A Holbrook. [2003] Application of DM techniques in Pharmacovigilance, British Journal of Clinical Pharmacology 57(2): 127-134

[4] Carlos Ordonez. [2004] Improving Heart Disease Prediction Using Constrained Association Rules, Seminar Presentation at University of Tokyo.

[5] D Castro. [2009] Explaining International IT Application Leadership: Health IT, The Information Technology at Innovation Foundation.

[6] Division of Health Care Statistics, "NCHS Health E-Stat Report", National Center for Health Statistics of US 2011.

[7] Francis S. Collins MD. [2015] PhD and Harold Varmus MD, A New Initiative on Precision Medicine", The New England Journal of Medicine, N Engl J Med 2015; 372:793-795,

[8] Geoff Hulten, Laurie Spencer, Pedro Domingos. [2013] Mining Time dynamical information Streams Computer Science & Engineering,

[9] G Litjens, O Debats, J Barentsz, N Karssemeijer, H Huisman. [2014] A Computer-Aided Detection of prostatic Adenocarcinoma in MRI, IEEE Transaction on Medical Imaging, 33(5)

[10] G Ravi Kumar, GA Ramachandra, K Nagamani.[2013] An Efficient Prediction of Breast Cancer Data using Data Mining Techniques, International Journal of Innovations in Engineering and Technology (IJIET), 2 (4).

[11] Healthcare Information and Management Systems Society, "Electronic Health Records- A Global Perspective - White paper, HIMSS Enterprise Systems Steering Committee and the Global Enterprise Task Force, 2010.

[12] HR Wulff, SA Pedersen, R Rosenberg.[1990] Philosophy of Medicine an Introduction, Blackwell Scientific Publications, Oxford,.

[13] K. A. Stroetmann, J. Artmann, and V. N. Stroetmann, "European Countries on their Journey Towards National eHealth Infrastructures", Final European Progress Report, European Commission, DG Information Society and Media, ICT for Health Unit, 2011.

[14] K.Sudhakar, Dr. M. Manimekalai. [2014] Study of Heart Disease Prediction using Data Mining International Journal of Advanced Research in Computer Science and Software Engineering, 4(1), ISSN: 2277 128X, January

[15] Mohammed Abdul Khaleel, Sateesh Kumar, Pradham GN Dash. [2013] A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases, International Journal of Advanced Research in Computer Science and Software Engineering, 3(8)

[16] Muhamad Hariz Muhamad Adnan,Wahidah Husain, Nur'Aini Abdul Rashid, "Data Mining for Medical Systems: A Review", Proc. of the International Conference on Advances in Computer and Info. Technology - ACIT 2012.

[17] National Research Council, Toward Precision Medicine: building a knowledge network for biomedical research and a new taxonomy of disease" Washington, DC: National Academies Press, 2011.

[18] Niti Guru, Anil Dahiya, Navin Rajpal.[2007] Decision Support System for Heart Disease Diagnosis Using Neural Network, Delhi Business Review, 8(1)

[19] P Treigys, V Saltenis, G Dzemyda, V Barzdziukas, A Paunksnis,[2008] Automated optic nerve disc parameterization", Informatica 19(3):403-420.

[20] S Fuller, Carol Friedman, William Hersh. [2005] Medical Informatics: Knowledge Management and Data Mining in Biomedicine, Edited by Hsinchun.Chen & Sherrilynne, Springer

[21] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

[22] Shantakumar B.Patil, Y.S.Kumaraswamy. [2009] Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network" - ISSN 1450-216X 31 (4) :642-656.

[23] Sheh A et al. [2011] Prevent Gastric cancer by modulating leukocyte recruitment and oncogenic pathways in

Helicobacter pylori -infected INS-GAS male mice, 17 (lowercase beta) – estradiol and Tamoxifen, Cancer prevention research.

[24] S Palaniappan, R Awang. [2008] Intelligent heart disease Prediction system using data mining techniques, Computer Systems and Applications, IEEE/ACS International Conference, 31.03.2008 - 04.04.

[25] SL Ting , CC Shum , SK Kwok , AHC Tsang, WB Lee. [2009] Data Mining in Biomedicine: Current Applications and Further Directions for Research, J Software Engineering & Applns, 2: 150-159,.

[26] SS. Saraf, GR Udupi, Santosh D. Hajare. [2011] Data Mining in Biomedical, Imaging, Signaling and Systems, Chapter 14, Discovering Association of Diseases in the Upper Gastrointestinal Tract Using Text Mining Techniques, Edited by Taylor and Francis Group.

[27] S. Wasan, V. Bhatnagar, and H. Kaur. [2006] The impact of data mining techniques on medical diagnostics", Data Science Journal 5, , 119–126.

[28] Thun MJ, DeLancey JO, Center MM, Jemal A, Ward EM.[2010] The global burden of cancer: Priorities for Prevention, Carciogenesis, 31: 100-110, 2010.

[29] Vikas Chaurasia, Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engg 2(1)

[30] V Krishnaiah, G Narsimha, N Subhash Chandra. [2013] Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, 4 (1):39 – 45.

[31] V. Speckauskiene and A. Lukosevicius, [2009] Methodology of adaptation of data mining methods for medical decision support: Case Study", Electronics and Electrical Engineering 2(90): 25–28.

[32] Xinxin Wang, Zhana Duren, Chao Zhang, Lin Chen, Yong Wang, [2012] Clinical Data Analysis Reveals Three Subtypes of Gastric Cancer, IEEE 6th International Conference on Systems Biology

[33] Shubpreet Kaur , RK Bawa.[2015] Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System, International Journal of Energy, Information and Communications 6(4):.17-34,

[34] Parminder Kaur and Aditya Khamparia, [2015]Classification of Liver Based Diseases using Random Tree", in International Journal of Advances and Engineering & Technology, 8(3).

[35] Asha Rajkumar, B Sophia Reena, [2010] Diagnosis of Heart Disease Using Data mining Algorithm" , Global Journal of Computer Science and Technology,10(10):38 - 43.

[36] Jyoti Soni ,Ujma Ansari, Dipesh Sharma, Sunita Soni. [2011] Predictive Data Mining for Medical Diagnosis, An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), 17(8)

[37] M Akhil Jabbar, Bulusu Lakshmana Deekshatulu, Priti Chandra, [2012]Heart Disease Prediction System using Associative Classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies.

[38] DD Kaladhar, KA Rayavarapu, Varahalarao Vadlapudi. [2012] Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis Journal of Biometrics and Biostatistics, 1(12):1-5.

[39] Mai Shouman, Tim Turner, Rob Stocker.[2012] Using Data Mining Techniques In Heart Disease Diagnosis And Treatment" ,Proceedings in Japan Egypt Conference on Electronics, Communications and Computers, IEEE,2 :174-177.

[40] Abhishek Taneja. [2013] Heart Disease Prediction System Using Data Mining Techniques", Oriental journal of Computer science & technology, 6(4): 457-466, December 2013.

[41] Kawsar Ahmed, Abdullah Al Emran, Tasnuba Jesmin, et al. [2013]Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, 14: 595-598.

[42] Kawsar Ahmed, Tasnuba Jesmin,Md.Zamilur Rahman. [2013] Early Prevention and Detection of Skin Cancer Risk using Data Mining, International Journal of Computer , 62(4):1-6,

[43] SF Shazmeen, MMA Baig, MR Pawar.[2013] Performance Evaluation of Different Data Mining Classification lgorithm and Predictive Analysis, Journal of Computer Engineering, 10(6): 01-06.

[44] V Krishnaiah. [2013] Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies (IJCSIT), 4(1): 39 –45

[45] K R Lakshmi, Y Nagesh, M. VeeraKrishna. [2014] Performance comparison of three data mining Techniques for predicting kidneydisease survivability", International Journal of Advances in Engineering & Technology, 7(1): 242-254.

[46] Dariusz JANKOWSKI and Konrad JACKOWSKI.[2012] An improved Medical Diagnosing of Acute Abdominal Pain with Decision Tree", Journal of Medical Informatics & Technologies, 20, ISSN 1642-6037.

[47] Amir Hossein Rasekh, Zeinab Liaghat, Alireza Tabebordbar. [2013] Helicobacter pylori microbe and detecting with data mining Algorithms, Open Journal of Gastroenterology, 93-98.

[48] Seyed Abbas Mahmoodi, Kamal Mirzaie, Seyed Mostafa Mahmoudi, [2016] A new algorithm to extract hidden rules of Gastric cancer data based on Ontology, SpringerPlus, 5:312,