Data Mining Techniques for Medical Data: A Review

Subhash Chandra Pandey
Computer Science & Engineering Department
Birla Institute of Technology, Ranchi -Allahabad Campus
Naini, Allahabad (UP), India
subh63@yahoo.co.in

***Abstract*— Data mining is an important area of research and is pragmatically used in different domains like finance, clinical research, education, healthcare etc. Further, the scope of data mining have thoroughly been reviewed and surveyed by many researchers pertaining to the domain of healthcare which is an active interdisciplinary area of research. In fact, the task of knowledge extraction from the medical data is a challenging endeavor and it is a complex task. The main motive of this review paper is to give a review of data mining in the purview of healthcare. Moreover, intertwining and interrelation of previous researches have been presented in a novel manner. Furthermore, merits and demerits of frequently used data mining techniques in the domain of health care and medical data have been compared. The use of different data mining tasks in health care is also discussed. An analytical approach regarding the uniqueness of medical data in health care is also presented.**

Keywords*— *Medical data;Data mining tasks;Data mining applications on medical data style; Data mining techniques; Uniqueness of medical data

I. INTRODUCTION

Medical data means databases that stores healthcare information, like patient's records. With the development of Information Technology, lots of such medical data are stored in electronic forms. These databases contain large volume of data. Medical data is available from different sources for example; X-ray, computed tomography scans (CT), magnetic resonance images (MRI), ultrasound, etc. Thus, the increase in the volume of data and the databases required to store the digitized data has increased exponentially [1]. Further, raw medical data is usually huge and dissimilar in nature and it may be collected from different sources like, images, interviews with the patient, laboratory data, and the physician's observations and evaluations [2]. Medical data are of the various types. It can be in the form of images, datasets, signals, wavelengths etc. In present scenario, due to researches and development in the field of information gathering tools, we can witness huge amount of information or data available in electronic format. It is obvious that to store such a large amount of data or information the sizes of databases also increase substantially [3].

Medical data are available in hundreds of public and private databases, which has only been possible by novel database technologies and the Internet [4]. It has been estimated that healthcare industry may generate terabytes of data every year [5]. Actually, the job of extracting useful information for quality healthcare is tricky and important and

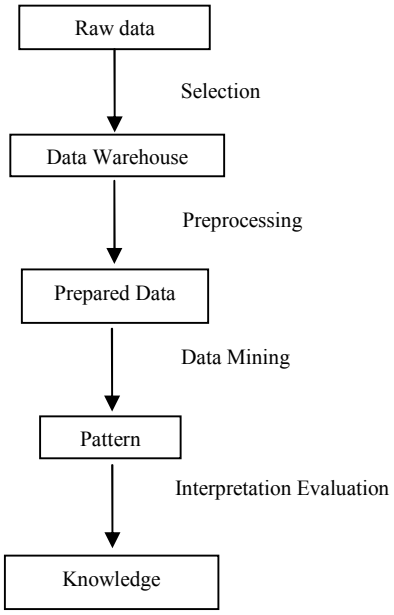

Fig. 1: Role of data mining in knowledge discovery process

nowadays we have lots of data available in our databases for this purpose. However, the knowledge that is extracted from it is nearly negligible. Thus, effective organization, analysis and interpretation of data are of the paramount importance so that tangible extraction of knowledge could become possible. In fact, different computational techniques are required to manage these large databases of medical data to discover useful patterns and hidden knowledge from them [4]. Often in data mining process we analyze enormous and large observational datasets and subsequently extract the useful hidden patterns for the purpose of data classification. Today, data mining has also started its tryst with healthcare and medical data. It is because of the fact that there is dire need of efficient techniques for detecting unknown and valuable hidden information from medical data [6] so that complex interrelation among the patients, their medical conditions, and treatments can be analyzed in a lucid manner [7]. The use of data mining in healthcare and medical field is pervasive and it has many applications like, detection of fraud in health insurance, providing better medical solutions to patients at a lower cost, detection and causes of diseases, and identification of efficient medical treatments methods. Indeed, data mining is a core process of a broader prospect known as the

978-1-5090-4620-1/16/$31.00 ©2016 IEEE

knowledge discovery. The inter-relation between the data mining and knowledge discovery is shown in the Figure 1.

II. Data Mining Tasks and its Use in Healthcare

There are different data mining models varying from one application domain to another. However, it can be broadly categorized in two groups. Namely: Predictive Model and Descriptive Model. Some important data mining tasks pertaining to medical and healthcare domain are enumerated below.

- Summarization
- Association
- Classification
- Clustering
- Trend analysis
- Regression.

(i) Summarization: In summarization, the set of data is abstracted that results into a smaller set of data which gives us a general overall review of the data. Thus, summarization is the abstraction or generalization of the data. Summarization can be done till many levels of abstraction and it can be

viewed from different perspectives. For example, rather than looking at the details of the call, it can be summarized into duration of the call, number of call, and cost incurred during the call. In the same way, calls can also be summarized on the basis of national calls or international calls. These combinations of different levels of abstraction tell us about the various patterns and regularities present in the data [8].

(ii) Association: Association is looking for togetherness or connection of objects in large databases. Such kind of connection is known as association rule. An association reveals relationships existing among objects. Its main purpose is to find interesting correlations existing among the objects, i.e., existence of a set of objects in some other object [9]. Association rules are usually used in marketing, commodity management, advertising, etc. From these association rules associations and patterns are extracted that exist among various attributes. Indeed, association based data mining aims to find associations between attributes and then generate rules from those data sets [10]. For example, an association rule that "call waiting" is associated with "call display", says if a customer is subscribed to the "call waiting" service, that customer is very likely to subscribe to "call display" service as well.

(iii) Classification: Classification divides data sets into target classes. Classification techniques predict the target classes for each of the data instance present. For example, using classification techniques a patient can be classified into "high risk" or "low risk" on the basis of their disease patterns. In this approach the classes are known and thus it is a kind of supervised learning. There are two methods of classification task. These are: binary and multilevel. In classification task the dataset is divided into training and testing data sets. Further, the classifier is trained with the help of training data set and subsequently the correctness of the classifier is tested on test dataset. The classification task of data mining is generally used in healthcare industries [6]. The classification task is often used to predict the treatment cost of different disease [11].

(iv)Clustering: There is subtle difference between classification and clustering. Classification is a supervised learning whereas clustering is an unsupervised learning method. Classification has the information of the class leveled but in clustering the information regarding the class leveled is not known. In clustering similar data are placed in the same cluster and dissimilar data are placed in some other cluster [12]. Clustering needs very less or no information for partitioning the data. The drawback of clustering is that first we have to identify the clusters and then assign a new instance to the clusters [13].

(v) Trend analysis: We can observe a lot of time dependent data in literature. In different walks of life such that: sales of a company, credit card transactions of a customer, and stock prices are all time series data. Such data can be viewed as objects with a 'time' attribute. It is interesting to find patterns and regularities in the data along the dimension of time. Trend analysis discovers these interesting patterns [9].

(vi) Regression: Regression is learning a function which can map a data item to a real – valued prediction variable [14]. Indeed, regression establishes a relationship between unknown and independent estimated variable and known dependent variable. Regression is a widely used technique for prediction

A. Data Mining for Healthcare

In healthcare industries dependence on data is increasing day by day [15]. In medical science, diagnosis of any disease and treatment of patients is the most important task. In recent days, doctor's hand written notes have been converted to electronic records with an aim of reducing cost incurred during treatment and improves efficiency of the treatment [16].

Data mining applications in healthcare can be further divided into following categories:

a. Diagnosis and prediction of diseases – When it comes to healthcare industries, diagnosis and prognosis of diseases is very important [17], it is one of the most important purpose of using data mining for healthcare. Use of data mining for healthcare has helped doctor's to improve the health services provided by them [15]. One cannot waste time and money by choosing some incorrect treatment for a patient, which can also harm patient's health [18].

b. Ranking of various hospitals – Data mining techniques are used to study all the details of various hospitals in order to rank them [19]. Organizations rank various hospitals on the basis of their capability to handle patients with serious illness, i.e., hospitals with a higher rank are more suitable for handling high–risk patients, as it is their highest priority whereas this is not the case in lower ranked hospitals because they do not even consider the risk factor.

c. Better treatment techniques – With the help of data mining techniques, both the doctor and patient can choose the best

treatment option by comparing among all the treatment techniques. They can select the best treatment techniques both in terms of effectiveness and cost. Through data mining they can also find out the side effects of various treatments and thus decreases risk to patients [6].

d. Effective treatments– By comparing factors like causes, symptoms, side effects, and cost of treatments data mining is used to analyze the effectiveness of treatments. For example, one can compare the results of treatments of different patients which were suffering from the same disease but were treated with different drugs. In this way, we can find which treatment is effective in terms of the patient's health and cost [20].

e. Better quality services provided to patients– With the advancement in technology, we already have voluminous data stored in digitized form. Data mining when applied on this huge medical data can help us in extracting many of the interesting unknown patterns. With the help of these patterns we can improve the quality of services and care provided to patients. Data mining also helps in knowing patients needs and more of their requirements so that they can be better treated [6]. Milley has also stated that data mining can help in analyzing specific patient's needs in order to enhance services provided by healthcare organizations [21].

f. Infection control in hospitals– Hospital infections affects millions of patients every year and the number of infections which are drug resistant is really high [22]. Inspection for infection is done through data mining to identify some irregular patterns in the data of infection control [15]. For infection control, these patterns are further studied by a knowledgeable person. Such a surveillance system that uses data mining techniques for discovering unknown patterns in infection control data was implemented at the University of Alabama [23].

g. Identifying high risk patients–American Health ways helps hospitals with diabetes disease management services to improve the quality and reduce the cost of diabetic patients. To differentiate between high–risk and low–risk patients, American Health ways used predictive modeling technique. Using predictive modeling technique, high–risk patients who needed more concern regarding their health were identified by the healthcare providers [24].

h. Reduction in insurance fraud and abuse–Healthcare insurer constructs a model to identify unusual patterns of claims by patients, physicians, hospitals, etc [25]. In 1998, Texas Medicaid Fraud and Abuse Detection System saved million dollars by detecting fraud and abuse through data mining techniques [26].

i. Proper hospital resources management – Management of hospital resources is an important task in healthcare industries. Data mining constructs a model for managing hospital resources. Group Health Cooperative uses data mining and provides services to hospitals at a lower cost [27]. Blue Cross manages diseases efficiently by reducing the cost and improving the outputs with the help of data mining [28].

j. Medical device industry – Without medical devices, healthcare industry could not exist. Mobile communications and inexpensive wireless bio-sensors are the most important

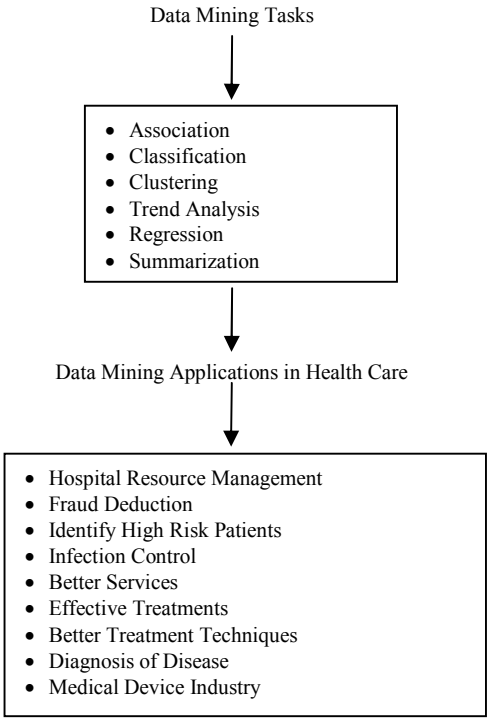

Fig. 2: Data mining tasks and applications in healthcare

aspect of mobile healthcare applications which provides a safe method for studying important signs of patients [29]. Ultimately, the success of data mining in healthcare totally depends on the availability of clean and organized healthcare data. Thus, the healthcare industries must look into this factor as well, i.e., how to capture and store data so that it could be properly mined subsequently [30]. The applications of data mining techniques in healthcare along with various data mining tasks are diagrammatically shown in the Figure 2.

III. KNOWLEDGE MANAGEMENT AND DATA MINING IN HEALTHCARE

Medical healthcare has been recently gaining increasing attention and popularity. Due to advances in technologies like molecular, biomedical techniques, medical imaging, and medical records of patients, large amount of medical data is generated every day. From clinical practices to individual research, these medical data is being stored in hundreds of private as well as public databases after the digitization of medical information like patient records, lab reports etc. Today, the rate of data accumulation is much faster than the rate of data extraction. Thus, this data needs to be well organized and stored in order to be useful. New information technology techniques are required to handle these large data repositories of medical data and to extract useful patterns from it. Basically, knowledge management and data mining have been adopted in various medical domains in recent years.

In the 20th century, management along with psychology and cognitive sciences led to the evolution of knowledge management [31]. The term 'knowledge management' came into existence in 80s and the academic discipline was developed in 1995 [32]. Indeed, knowledge management is the managerial approach to collect, manage, use, analyze, share, and discover the knowledge in order to maximize the performance [33]. There is no definition for what constitutes knowledge, but it is something abstract and inferential and is needed to support hypothesis generation and decision making. Recently researchers have done studies which showed that knowledge management has good effects on organizational and operational performance [34, 35]. A knowledge management model proposed in [36] gave substantial information regarding the healthcare industries and it said that the knowledge management processes lead to better organization learning and decision making which in turn leads to better organization performance. Knowledge management methodologies and techniques have been used to support storing, retrieving, sharing and management of data to make it explicit to biomedical knowledge. It is used in both scientific and business domains recently. There are many goals and challenges for knowledge management in companies. This is due to the following reasons; knowledge management could increase their performance, evaluate risks, help in developing partnerships, organize the management as well as enhance their economic value [37]. There are some criticisms also for knowledge management given by T.D. Wilson, [38]. However, knowledge management could succumb these criticisms mainly because of the fact that companies and organization really need knowledge management.

Methods and techniques in knowledge management can be categorized into three sections: people and technology, requirements elicitation, and measurement of value. Today frameworks take humans as well as technical perspectives into account. When we talk about human perspectives: it is about motivation and adoption. The employees are motivated either by giving financial or non–financial incentives in order to use knowledge management, not only for the sake of technology but also because it would affect the company. In [39], it is suggested that apart from giving incentives there should be a win-win system, both for the employee as well as the company and not a win-lose reward system. Other issue related to knowledge motivation was knowledge adoption; since people were not ready to use knowledge management. In [40], a model is proposed which discussed about issues of knowledge adoption. Indeed, data mining is a core step of a broader prospect known as knowledge discovery and it is used in different domain e.g.; to discover different biological, drug and patient care knowledge. It is also used for statistical analysis of the patterns. Perhaps, data mining is frequently used technique in medicine [27]. The basic objective of data mining is to analyze a set of raw data or data and to identify and extract novel and useful patterns [41]. Various data mining techniques such as neural networks, decision trees, fuzzy sets, support vector machines, bayesian networks and genetic algorithms are used to discover knowledge and patterns that are not known to the system and the users [42, 33]. In biomedical data mining, patient data should not be 'individually identifiable', i.e., no record should give sufficient data about the patient so that no one can identify the patient [2].

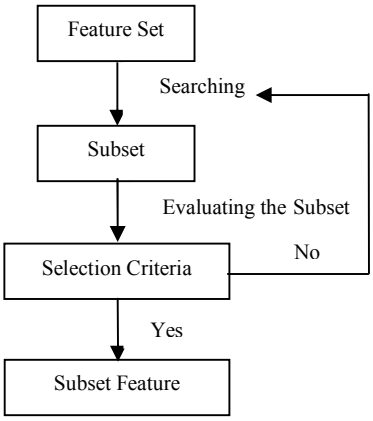

Fig. 3: Process of feature selection.

IV. Data Mining Techniques for Healthcare

Data mining uses various techniques for mining medical data. In fact, data mining techniques are used for feature selection. Feature Selection can be described as the process of selecting a minimum subset of features which are actually essential for classification. The feature set may be redundant and it may decrease the efficiency. Feature selection is a problem in the field of medical diagnosis [43]. The feature subset generation is also known as data reduction that is a step in data preprocessing [44]. Further, feature selection minimizes the number of essential features required for maximizing the accuracy of the model. It helps in reducing the space required by the feature set.

It also removes the redundant noise that might be present in the feature set and thus it increases the efficiency of the data mining algorithm [45]. The objective of feature selection is to produce cost effective and efficient model [46].

Fig. 3 shows complete process of the feature selection. It mainly consists of four stages: subset formation, evaluation of the subset, a selection criterion which is used as stopping criteria, and the final subset feature [44]. In the first step the feature set is searched after eliminating some inconsistencies like null values etc and redundancies that are present. Then the process of subset generation starts after searching the feature set. Subsequently, attribute evaluator evaluates the subset generated [47]. The phase of subset generation and evaluation continues until the selection/stopping criteria are fulfilled. Only after that the final subset feature set is selected.

A. Neural Networks

Neural networks were developed in the early days of the 20th century [48]. Neural networks are used in medicines as one of the most popular data modeling algorithm. Before the invention of decision trees and Support Vector Machine, neural networks were the best classification algorithm [49].The main objective of using neural networks is for pattern recognition and performing the tasks of classification [50]. The neural network system is modeled like a human brain. The human brain consists of millions of interconnected neurons. In a similar way, the neural network is an interconnection of artificial neurons and each connection has associated weight. By adjusting the weights, due to its adaptive nature it helps in minimizing the error [3].These neurons work together in parallel to produce the output function. In the learning phase the network will learn by adjusting the weights to predict the correct class label of the input. Neural Networks have added advantage because they can predict nonlinear relationship unlike simple modeling methods [51]. Neural networks play an important role in analysis of medical data. Applications of neural networks in this field consists tissue classification, disease prediction and drug development. Prediction of heart diseases can be done with the help of a neural network [52]. There are a few architectures of neural networks which are enumerated below:

i. Multi Layer Neural Network (MLNN): This type of neural networks use hidden layers with the help of which it solves the classification problem for non linear sets [53]. These hidden layers are usually interpreted as hyper-planes. This kind of neural networks is used for classifying different categories of data.

ii. Polynomial Neural Network (PNN): Polynomial neural networks have neurons like units as multilayer perceptrons which produce multivariate polynomial mappings.

B. Decision Tree

A decision tree is one which has terminal and non-terminal nodes. Each non-terminal node represents a test or condition on a data item. Decision trees classify the instances by sorting them down from the non-terminal to the terminal nodes [54]. The output that which branch will be selected completely depends on the outcome of the test. For example, we have a decision tree for medical readmission. With the help of this tree we can decide whether a patient needs readmission or not [3]. Decision trees basically create a visual representation of various pros and cons and potential values of each option [55]. Decision trees are commonly used for calculating conditional probabilities in operations research analysis [56]. Best alternatives can be chosen with the help of decision trees and based on maximum information gain the traversal from root to leaf node indicates unique class separation [57]. In some other applications of data mining, like in marketing, the accuracy of a prediction could be all that they need. It may not be important to know about the working of the model. For example, when a marketing professional wants to launch a marketing campaign, he would require the overall descriptions of customer segments. For these types of applications, the decision tree algorithm is very suitable [58].

C. Fuzzy Sets

Fuzzy sets and fuzzy logic are the best methodology used in data mining that is generally used for representing and processing uncertainty. It is one of the best methods to deal with imperfect and noisy data [51]. This fuzzy set theory was introduced by Zadeh [59], which helps us in handling vague data. Fuzzy sets and fuzzy logic are needed to implement the proposed expert system. With the help of fuzzy logic we can calculate the probability of any particular case to fall in any cluster and after that based on the value, decisions can be made [60].

D. Support Vector Machine (SVM)

The concept of SVM was proposed first time in. [61-62]. It provides the most accurate results in comparison to all the other algorithms. It is a classification technique and it works on the basis of statistical learning theory [62-63]. For various kernels, SVM has been used as a universal approximator [64]. The subset of the learning data is called support vector and with the help of this the support vector machines is defined. Absence of local minima is one of the main features of SVM. The SVM model is a representation of the training data and with the help of support vectors one can extract the condensed data set [65]. SVM finds an optimal separating hyper-plane which maximizes the margin between the examples of two different classes. SVM was developed for problems related to binary classification but then it can easily be extended to problems related to multiclass problems. This is one of the most important reasons for SVM to gain popularity [66-67]. In a binary classification task, such as predicting ICU mortality, the hyper-plane is the division between two outputs. To be useful for tasks it can create single as well as multiple hyper-planes. There are two methods for implementing SVM's. The first method involves mathematical programming and the second method employs kernel functions. The main task of using hyper-planes is that it will maximize the separation between data points [3]. In noisy data, error is minimized by maximizing the margin between the examples of two different classes and the hyper-plane is defined as the center line of the separating space. There are two types of SVMs. The first one is Linear SVMs which separates the data points with the help of a linear decision boundary. It performs well on the datasets that can easily be separated into two parts. But sometimes complex datasets are difficult to classify with the help of a linear kernel for which the second kind of SVMs is used i.e., Non–linear SVMs which separates the datasets with the help of non linear decision boundary. It is the most powerful algorithm as it can obtain maximal generalization when predicting the classification of data [45]. The SVM shows accuracy in binary classification problems like valve classification/heart beat etc [68-70].

E. Bayesian Networks

Bayesian network is a specific type of network which represents knowledge about uncertain domain. It belongs to the domain of probabilistic graphical models (GMs). In Bayesian network nodes represent the variables and various edges represent probabilistic dependencies among those

variables [71-73]. Bayesian network specifies two types of information for each variable [74].

F. Rough Set

The concept of rough sets theory is similar to the concept of fuzzy sets theory. The only difference is that in this theory the uncertainty is described as a boundary region of a set. Every subset that is defined through upper and lower approximations is called a rough set. This definition also needs mathematical concepts since it is defined by topological operations known as approximations. They are usually combined with other methods such as classification, clustering [51].

G. Genetic Algorithm

The genetic algorithm is a search and optimization techniques which is based on genetics and selection. Genetic algorithms are basically used in neural sets which act as a guide for the learning process of data mining algorithms rather than for finding patterns. They are also used in the form of association rules or some other formalism in data mining to formulate hypothesis about variables and dependencies among them. The basic idea of genetic algorithm is that we can obtain a much better solution by combining the good parts of other solutions which is said in schemata theory, in a way like nature does by combining the DNAs of living creatures [75]. In a genetic algorithm there is a population that is composed of many individuals which evolve under specific selection rules to a state where fitness is maximized [76]. Initially a population of rules is created at random, each rule representing a solution to the problem. Then pairs of rules are selected as parents which are usually the strongest rules and these pairs of rules are then combined to produce offspring [77]. A genetic algorithm basically consists of three operators, namely, selection, crossover and mutation. In selection, on the basis of fitness a suitable string is selected for breeding a new generation, then crossover combines these suitable good strings to produce better offspring, mutation then alters a string locally so that the genetic diversity is maintained from one generation of a population to another. In every generation the population is evaluated for the termination of the algorithm, if the termination criteria are not satisfied it again is operated by the three operators and then again it is evaluated.

V. Machine Learning Methods in Healthcare

There is plethora of research in machine learning domain and it is mostly application driven. Machine learning researches are widely used in healthcare domain. Machine learning methods are able to identify areas in which an increase in research would lead to advances. In conditions where algorithmic solutions are not present and there is lack of formal codes or there is poor definition of knowledge about the application domain, machine learning methods come into existence. Machine learning includes many methods, but we can broadly classify them as symbolic and sub-symbolic based on the nature of manipulation while learning [78]. When we talk about symbolic learning method, knowledge required and the level of inference performed are different, like in decision trees [79]. On the other hand genetic algorithms [80] and artificial neural networks [81] are examples of sub-symbolic methods of classification.

When we talk about machine learning methods in healthcare domain, these techniques and tools can help in diagnosis and prognosis of diseases, prediction of disease progression, or extraction of medical knowledge. Symbolic classification like inductive learning is used to add learning and knowledge management to expert systems [82]. Machine learning tools help us in handling few characteristic features of medical domain like missing values, random noise or only few patient records available [83]. Sub-symbolic learning methods like neural networks help in improving the decision making because they are able to handle these datasets [84]. A major application in medical diagnosis is to interpret the medical image which provides significant assistance [85]. Indeed, as the healthcare domains is becoming more and more reliant on computer systems, machine learning methods can substantially help the physician's in many cases and enable diagnosis in real time.

Apart from making medical decisions, machine learning improves the efficiency and quality of medical decision making systems [86]. Issues like how well a medical expert can understand and use the results obtained from a system depend considerably on machine learning methods used. Many researchers worked on medical expert systems for ECG diagnosis by implementing machine learning techniques to improve the knowledge of the medical expert system.

VI. Uniqueness of Data Mining in Healthcare

In this section, we will render the unique features of medical data mining to make the expert system dealing with healthcare more constraint free specifically while mining the large heterogeneous medical data because medical data itself is very rewarding and difficult to mine in comparison to other datasets. The medical datasets are huge and contain large amount of medical information. At the same time, medical data also possess distinct legal, ethical, and social constraints [2]. Precisely, there are four main points that should be discussed regarding the uniqueness of medical data.

i. Medical data is heterogeneous in nature: As we already know raw medical data is voluminous and heterogeneous. It may be collected from various sources like images, physician's observations, interviews with patients, laboratory data. All these help in diagnosis and prognosis of diseases and

TABLE 1. ADVANTAGES AND DISADVANTAGES OF DIFFERENT TECHNIQUES USED IN HEALTHCARE

S. No.	Name of the Technique	Advantages	Disadvantages
1.	Neural Networks	1. It is able to handle noisy data properly for training. 2. It is capable of producing complex relationships between input and output. It can analyze and organize data based on its own features without any external help. 3. Various neural networks can be used for clustering and prototype creation.	1. It does not work well with hundreds or thousands of input features and even it does not work well for complex problems. 2. Local minima. 3. Over fitting. 4. It is difficult to understand the model built by neural network and requires high processing time.
2.	Decision Trees	1. It can handle all types of variables, variables with missing values as well and it is easy to interpret. 2. For constructing decision trees one does not need to know about the domain. Even it can handle numerical and categorical data. 3. It can process high dimension data easily and it minimizes ambiguity of complex decisions and assigns exact values to the outputs.	1. For numeric dataset, it generates complex decision trees. 2. It is an unstable classifier, i.e., performance of a classifier depends on the dataset. 3. It is restricted to one output attribute and generates categorical data. 4. Performance of decision trees is not affected by co-linearity and linear-separability problems.
3.	Fuzzy sets	1. Unsupervised approach. 2. Converges approach.	1. Larger computational time. 2. Sensitivity to speed, local minima. 3.Sensitivity to noise, and one expects zero or low noise level.
4.	Support Vector Machines	1. Provides better accuracy in comparison to other classifiers and it is effective in high dimensional spaces. 2. It is effective in cases where the number of dimensions is greater than the number of samples. 3. It easily handles complex non linear data points and over fitting is not a problem like in other cases. 4. It is memory efficient because it uses a subset of training sets in support vectors. 5. It is versatile because different kernel functions can be specified for the decision functions.	1. It gives poor performances when the number of features is much greater than the number of samples. 2. It is computationally expensive and even the training process takes more than in comparison to other methods. 3. Selection of right kernel function is a problem because for every dataset different kernel function shows different results. 4. SVM was developed to solve the problems of binary class. Thus, it solves problem of multi class by breaking it into pair of two classes. 5. It does not provide probability estimates directly. These are calculated using an expansive five – fold cross validation.
5.	Bayesian Networks	1. It is fast and accurate for huge datasets as well. 2. It makes computations easier.	1. In some cases, where there is dependency among variables, it does not gives accurate results.
6.	Rough Sets	1. It does not need any additional knowledge about data like probability in statistics. 2. Identifies relationships that would not be easily found using statistical methods. 3. From data it produces sets of decision rules.	1. Some new discretization methods are required for quantitative attributes. Even more research is needed in this field. 2. Studies of new approach to missing data are also needed.
7.	Genetic Algorithms	1. Here the fitness function is a flexible expression of modeling criteria.	1. Finding fitness function is critical.

thus are very important in nature and cannot be ignored. One of the areas in the heterogeneity of medical data is the volume and complexity of medical data. It is worth to mention here that the heterogeneity is in the sense that we have data in numeric as well as images form. Further, the huge medical data requires lots of storage space and needs new tools to analyze the data. In fact, un-stored and un-organized data are considered less pragmatic in healthcare domain.

The second area in the heterogeneity of medical data is the importance of physician's observations. It may be in the form of images, signals and is usually written in English and is difficult to standardize and mine.

Even experts from the same field find it difficult to understand because of reasons like different grammatical constructs used for describing relations between medical entities or different names used for same disease. It is said that a part of the solution for the processing of physician's interpretation may be held by the computer translation [87-89].

The third area in the heterogeneity of medical data is the specificity analysis and sensitivity – almost all diagnoses and finding of effective treatments in medicine have some associated errors and it is not easy to measure which specificity analysis and sensitivity should used.

For understanding the concept of specificity and sensitivity we should first understand what a test is. A test is basically one of those values that are used to characterize the condition of a patient. Sensitivity measures how many times you find what you are looking for. Specificity measures how many times what you find is what you are looking for.

The fourth area in the heterogeneity of medical data is due to the poor mathematical characterization of medical data. Moreover, another unique feature of mining medical data is that the underlying structures of medical data are poorly characterized and less emphasized mathematically in comparison to other fields of science. Medicine has no formal structure into which information can be organized by a data miner. Perhaps, the main reason of heterogeneity in medical data is difference in the logic of medicine from the logic of physical sciences [90-93].

ii. Legal, ethical and social issues: Medical data is basically patient's data. So any misuse of medical data would lead to patient's abuse. Thus, there is a large ethical and legal traditions designed to prevent misuse of medical data.

A point of discussion under legal, ethical and social issues is 'ownership of data'. Theoretically, ownership is entitlement to sell an item of property [94]. The question of data ownership within the purview of medical data is quite complicated because human data cannot be actually sold. Human medical data is available in thousands of terabytes for data mining and it is very often a heterogeneous databases. In addition, it is scattered without any format throughout the medical care establishment. That's why; it is hard to decide the actual ownership of medical data.

Second point of discussion is fear of lawsuits. This is another unique feature in the mining of medical data and it restricts the health care providers and physicians. Medical care in some places is expensive than other places. Because of this, physicians and other producers of medical data are reluctant to handle their medical data to mining experts for the mining purposes which in turn cause untoward events.

Security and privacy is the third unique feature concerned with human data. In different countries, guidelines are set by government agencies for concealment of patient identification. This renders the patients to be frank with their physicians. Moreover, patients are assured that their personal data would not be made public. Another issue is data security or rather data handling and data transfer. Since the data is transferred electronically it is insecure. It has been noted in US federal documents [95-97], that there are two research needs for re-identification of de-identified medical data which are important in nature. Some important cases of this domain are: Firstly, accidental duplicate records of the same patient should be prevented. Secondly, there might be a need to refer to re-identify the records to verify patient data or to obtain some additional information.

Next is the theory of expected benefits. Patient's data in public databases cannot be mined without justifying that this will creates some obvious benefits to the society otherwise one cannot perform data analysis legally and ethically. US federal guidelines specify a number of administrative policies for patient privacy that would not be required for non– medical data mining [98].

iii. Statistical philosophy: Data mining methods, especially statistics may be different for medical data. Primarily, medicine is a patient – care entity and secondly it is also used as a research resource. Generally, justification of patient's benefits is given before collection or rejection of medical data. Therefore, to reduce such complexities; statistical philosophy in medicine is incorporated.. When classical statistical tests are designed, rules are set up in advance on the basis of the idea that the experiment would be repeated. So, we cannot change rules in the middle of the experiment otherwise it would lead to meaningless formulas and distributions. Thus, classical statistical tests in medicine may lead to ambiguous results. If one's mind changes during the investigation, then interpretation of the data will be polluted even if the observed values are not changed. Suppose we are taking a neural network to examine the dataset then different training strategies will produce different outputs [94]. Here, it is defeating to conceal a subset of cases from the training data set. The second point is that data mining is a superset of statistics. Data mining and statistics share a great deal together since both aims at discovering underlying structures in data. The difference is that data mining must deal with heterogeneous data fields. Further, because of large volume and heterogeneous nature of medical databases, it is not plausible that any data mining tool can succeed with raw and unorganized data [99]. When we talk about medical data

mining and knowledge discovery, it is important to follow a set of rules from problem specification to application of the results [100]. Knowledge discovery is a non–trivial process of determining valid, new, useful, and understandable patterns from large sets of data [101].

iv. Status of medicine: Medicine is a need, a must for a patient. It is not a luxury or pleasure for any human being. The outcome of healthcare is life or death which applies to all humans. Medicine has a special status in daily life and is a popular subject of common interest of humanity. Medical care is sometimes risky and when it fails the desire for revenge is intense. Medical information of a patient is private and the public is fearful about its disclosure. We enjoy the benefits of medical research, but very few of us are ready to contribute our personal details for research purpose. Moreover, when medical data are published, it is expected that the researchers will maintain the confidentiality regarding the identity of an individual patient, and the results will be used for benefit of the society [102]. As a matter of fact, researchers must follow that scientific advancement are for overall development i.e., it could be used for betterment of good as well as bad [103].

VII. Results and Discussion

We elucidate different aspects and techniques of machine learning regarding the medical data and healthcare. There are many techniques used in healthcare. However, this paper mainly focused on these techniques namely; neural networks, decision trees, fuzzy sets, support vector machines, bayesian networks, rough sets, and genetic algorithms. Each technique has its own advantages and associated disadvantages e.g.,NN is able to handle noisy data properly for training but it does not work well with hundreds or thousands of input features and even it does not work well for complex problems.Further, decision tree can handle all types of variables but its use is restricted to one output attribute. SVM provides better accuracy in comparison to other classifiers and it is effective in high dimensional spaces. However, it gives poor performances when the number of features is much greater than the number of samples. Bayesian Networks is fast and accurate for huge datasets as well but in some cases result obtained from it is wrong. Therefore, it is hard to say that which method is best. Indeed, in different scenario different technique renders best while the same technique performs worst in other application. A comparatice study is shown in table 1.

VIII. Conclusions

In this paper, we have discussed that data mining can be beneficial in medical domain. Due to rapid increase in the volume of medical data, data mining techniques have high utility in this field. Various tasks and applications related to data mining are analyzed within the purview of healthcare organizations. This paper explores different data mining techniques, their advantages and drawbacks. Perhaps, there is no single data mining technique which can give consistent results for all types of healthcare data. Indeed, the performance of techniques varies from one dataset to other dataset. For effective utilization of these techniques in healthcare domain, there is a need to enhance and secure health data sharing among various parties. This paper also addresses uniqueness of data mining with respect to medical data. Further, the constraints and difficulties related to privacy sensitivity and large volume of medical data play vital role in selection of the particular data mining technique. Moreover, ethical and legal aspects of medical data are also important aspects. Medical data can have a special status based on its applicability to all people.

Acknowledgment (HEADING 5)

The author is pleased to acknowledge the sincere effort and help extended by his student Ms. Akanksha Verma. Ms. Akanksha has completed his M.Tech (Computer Engg.) from Birla Institute of Technology (Allahabad Campus). She completed her M.Tech thesis on medical data classification under the supervision of the author.

References

[1] S. Mitra, S.K.Pal & Mitra , P., Data mining in soft computing framework: A survey, IEEE transactions on neural networks, 13(1), 3-14,2002.

[2] Krzysztof J. Cios, G.William Moore, Uniqueness of medical data mining, Artificial Intelligence in Medicine 26, 1–24, 2002.

[3] Parvez Ahmad, Saqib Qamar, Syed QasimAfser Rizvi, Techniques of Data Mining in Healthcare : A Review, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2015.

[4] Hsinchun Chen, Sherrilynne, S. Fuller, Carol Friedman and William Hersh, Knowledge Management, Data Mining and text mining inmedical informatics.

[5] V. krishnaiah, G. Narsimha, & N. Subhash Chandra, A study on clinical prediction using Data Mining techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239 248, March 2013.

[6] Divya Tomar and Sonali Agarwal , A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2013.

[7] Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya, A survey of data mining techniques on medical data for finding temporally frequent diseases", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013.

[8] S.D.Gheware, A.S.Kejkar, S.M.Tondare, Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2014.

[9] Yongjian Fu , Data Mining : Tasks, Techniques and Applications http://academic.csuohio.edu/fuy/Pub/pot97.pdf

[10] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2002.

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

[12] Pang-Ning Tan, Michael Steinbach ,Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005.

[13] Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009, International Conference on, vol. no., pp.1,6, 22-24 July 2009.

[14] Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2002.

[15] A. S. Elmaghraby, et al. Data Mining from multimedia patient records. 6, 2006.

[16] Nada Lavrac, BlažZupan, "Data Mining in Medicine" in Data Mining and Knowledge Discovery Handbook, 2005.

[17] Soni J, Ansari U, Sharma D, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.

[18] Naren Ramakrishnan, David Hanauer, Benjamin J. Keller, Mining Electronic Health Records, IEEE Computer 43(10): 77-81, 2010.

[19] O. Mary K, Mat, "Applications of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, August 2004.

[20] Hian Chye K, Gerald T, Data mining applications in healthcare, Journal of healthcare information management: JHIM.19 (2): 64-72, (2005).

[21] A. Milley, "Healthcare and data mining", Health Management Technology, vol. 21, no. 8, pp. 44-47, 2000.

[22] Gaynes R, Richards C, Edwards J, et al. Feeding back surveillance data to prevent hospital-acquired infections. Emerg Infect Dis 2001;7:2 95-298, 2001.

[23] Brosette SE, Spragre AP, Jones WT, Moser SA. A data mining system for infection control surveillance. Methods Inf Med,39: 303-310, 2000.

[24] M. Ridinger, "American Healthways uses SAS to improve patient care", DM Review, vol. 12, no.139, 2002.

[25] M. Durairaj, V.Ranjani, Data mining applications in healthcare sector: A Study, International Journal Of Scientific & Technology Research Volume 2, Issue 10, ISSN 2277-8616, October 2013.

[26] Anonymous. Texas Medicaid Fraud and Abuse Detection System recovers $2.2 million, wins national award. Health Management Technology, vol. 20, no. 10, 1999.

[27] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, 2005.

[28] B. K. Schuerenberg, "An information excavation", Health Data Management, vol. 11, no. 6, pp. 80-82, 2003..

[29] P. D. Haghighi et. al., Mobile Data Mining for Intelligent Healthcare Support, IEEE xplore, 2009.

[30] Neelamadhab Padhy, Pragnyaban Mishra and Rasmita Panigrahi, The survey of data mining applications and feature scope, Asian Journal Of Computer Science And Information Technology 2: 4, 68– 77, 2012.

[31] Wiig, K, Knowledge Management: An Emerging Discipline Rooted in a Long History Knowledge Management (pp. 352): Butterworth-Heinemann, 1999.

[32] Stankosky M, Creating the Discipline of Knowledge Management: Butterworth-Heinemann, 2005.

[33] Chen H and Chau M. "Web Mining: Machine Learning for Web Applications," Annual Review of Information Science and Technology, 38, 289-329, 2004.

[34] Chen C-J & Huang J-W. Strategic human resource practices and innovation performance - The mediating role of knowledge management capacity. Journal of Business Research 62: 104-114, 2009.

[35] Fugate BS, Stank TP & Mentzer JT. Linking improved knowledge management to operational and organizational performance. Journal of Operations Management in Press, Corrected Proof, 2008.

[36] Orzano A.J, McInerney CR, Scharf D, Tallia AF & Crabtree BF. A knowledge management model: Implications for enhancing quality in health care. Journal of the American Society for Information Science & Technology 59: 489-505, 2008.

[37] Christo El Morr and Julien Subercaze, Knowledge Management in Health care, DOI: 10.4018/978-1-61520-670-4.ch023, pp 490-510.

[38] Wilson T. D, The nonsense of knowledge management, Information Research, 8(1), 2002.

[39] Zand D.E. The Leadership Triad: Knowledge, Trust and Power, New York: Oxford University Press, 1997.

[40] Sussman S. W & Siegal W. S, Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption, Information Systems Research, 14(1), 47-65, 2003.

[41] Fayyad U. M, Piatetsky-Shapiro G and Smyth P, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, 17(3), 37-54, 1996.

[42] Dunham M. H, Data Mining: Introductory and Advanced Topics, New Jersey, USA: Prentice Hall, 2002.

[43] Jihoon Yang and Vasant Honavar, Feature subset selection using Genetic Algorithm, IEEE Intelligent Systems, 1998.

[44] Hany M. Harb, Abeer S. Desuky, Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization, International Journal of Computer Applications (0975 – 8887) Volume 104 – No.5, October 2014.

[45] G. Ravi Kumar, Dr. G.A.Ramachandra, K.Nagamani, An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.

[46] V.Sangeetha, J.Preethi, M.Sreeshakthy, Survey on Medical Data Cluster analysis using Feature Selection and Neural Networks, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014.

[47] Megha Aggarwal, Amrita, Performance Analysis Of Different Feature Selection Methods In Intrusion Detection, International Journal Of Scientific & Technology Research Volume 2, Issue 6, June 2013.

[48] Anderson J. A and Davis J., An introduction to neural networks, MIT, Cambride, 1995.

[49] Obenshain M. K, Application of data mining techniques to healthcare data Infect. Control Hosp. Epidemiol, 25(8):690–695, 2004.

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

[51] A. Shameem Fathima, D. Manimegalai and Nisar Hundewale, A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, ISSN (Online): 1694-0814, November 2011.

[52] K. Usha Rani, Analysis Of Heart Diseases Dataset Using Neural Network Approach, International Journal Of Data Mining &Knowledge Management Process (Ijdkp) Vol.1, No.5, September 2011.

[53] Haykin. S, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.

[54] Emina Alickovic and Abdulhamit Subasi, Data Mining Techniques for Medical Data Classification, The International Arab Conference on Information Technology (ACIT) ,2011.

[55] S. Anto, Dr.S.Chandramathi, Supervised Machine Learning Approaches for Medical Data Set Classification - A Review, IJCST Vol. 2, Issue 4, Oct - Dec 2011.

[56] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, 2003.
http://ir.iit.edu/~nazli/cs422/CS422-Slides/DMClassification.

[57] Apte & S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsapteweissue_with_cover.pdf, 1997.

[58] V.Gayathri, M.Chanda Mona, S.Banu Chitra, A survey of data mining techniques on medical diagnosis and research. V.Gayathri, M.Chanda Mona, S.Banu Chitra, International Journal of Data Engineering (IJDE) Singaporean Journal of Scientific Research(SJSR) Vol.6.No.6 2014.

[59] L.A. Zadeh, "Some reflection on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent system", Soft computing,vol. 2, 1998.

[60] Kalyani Mali & Samayita Bhattacharya., Soft computing on Medical - Data (SCOM) for a Countrywide Medical System using Data Mining and Cloud Computing Features, Global Journal of Computer Science

and Technology Cloud and Distributed, Volume 13 Issue 3 Version 1.0 Year 2013.

[61] V. Vapnik, "Statistical Learning Theory", Wiley, ISBN: 978-0-471-03003-4, 1998.

[62] V. Vapnik, "The support vector method of function estimation", AT & T Labs – Research, John Wiley and Sons, New York, USA,1998.

[63] N. Cristianini and J.Shawe-taylor, An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 1995.

[64] Hammer, B. and Gersmann, K., "A Note On theUniversal Approximation Capability of SupportVector Machines", Neural Process Lett 17, pp. 1061 - 1085, 2003.

[65] Vapnik, V.N., "The Nature of Statistical LearningTheory", Springer, New York, 2005.

[66] N. Chistianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines, and other kernel-based learning methods", Cambridge University Press, 2000.

[67] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, 2000.

[68] Argyro Kampouraki, Christophoros Nikou, George Manis, "Robustness of Support Vector Machine-based Classification of Heart Rate Signals", Proceedings of the 28th IEEE, EMBS Annual International Conference, New York, USA, Aug30-sep 3,2006, 1995.

[69] Samjin Choi, "Detection of valvular heart disorders using wavelet packet decomposition and support vector machine, Elsevier", Expert Systems with Applications, 35, pp 1679-1687, 2008.

[70] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos, Antonis Stasis, "Support vector machine based identification of heart valve diseases using heart sounds", Elsevier, Computer Methods and Programs in Biomedicine ,95, pp. 47-61, 2009.

[71] Friedman N., Geiger, D.Goldszmidt M, "Bayesian network classifiers. Machine Learning 29: pp. 131-163, 1997.

[72] Friedman N., Koller D, "Being Bayesian About Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks", Machine Learning 50(1): pp. 95-125, 2003.

[73] Finn V. Jensen, An Introduction to Bayesian Networks, Springer, New York, 1996.

[74] Sebe N., Ira Cohen, Ashutosh Garg and Thomas Huang S. "Machine Learning in Computer Vision", Springer, Netherlands, pp. 130-133, 2005.

[75] Ankita Agarwal, "Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.

[76] Jihoon Yang and Vasant Honavar. Feature subset selection using Genetic Algorithm. IEEE Intelligent Systems, 1998.

[77] Sang Jun Lee, Keng Siau, A review of data mining techniques, Industrial Management & Data Systems, 101/1,MCB University Press [ISSN 0263-5577], 2001..

[78] George D., Magoulas and Andriana Prentza. Machine Learning in Medical Applications, Proceeding machine learning and its applications:Advance lectures, pp. 300-307, 2001.

[79] Quinlan J.R, "Induction of decision trees", Machine Learning, 1, 1, 81-106, 1986.

[80] Goldberg D, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[81] Rumelhart D.E and Mc Clelland, J.L. (eds.), Parallel Distributed Processing, Vol. 1: Foundations. MIT Press, Cambridge, MA: MIT Press,1986.

[82] Bourlas Ph, Sgouros N, Papakonstantinou G and Tsanakas P, "Towards a knowledge acquisition and management system for ECG diagnosis", In Proceedings of 13th International Congress Medical Informatics Europe-MIE96, Copenhagen, 1996.

[83] Zupan B., Halter J.A and Bohanec M., "Qualitative model approach to computer assisted reasoning in physiology", In Proceedings of Intelligent Data Analysis in Medicine and Pharmacology-IDAMAP98, Brighton, UK, 1998.

[84] Akay Y.M, Akay M, Welkowitz W and Kostis J.B, "Noninvasive detection of coronary artery disease using wavelet-based fuzzy neural networks", IEEE Engineering in Medicine and Biology, 761-764, 1994.

[85] Coppini G, Poli R and Valli G, "Recovery of the 3-D shape of the left ventricle from echo cardio graphic images", IEEE Transactions on Medical Imaging, 14, 301-317, 1995.

[86] Moustakis V and Charissis G, "Machine learning and medical decision making". In Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence- ACAI99, Chania, Greece, 1-19, 1996.

[87] Manning CD, Schuetze H, Foundations of statistical natural language processing. Cambridge (MA): MIT Press, 2000.

[88] Ceusters W, Medical natural language understanding as a supporting technology for data mining in healthcare. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, p. 32–60 [chapter 3], 2000.

[89] Friedman C, Hripcsak GW, Evaluating natural language processors in the clinical domain. Meth Inform Med,37:334–44, 1998.

[90] Brewka G, Dix J, Konolige K. Non monotonic reasoning: an overview. CSLI Lecture Notes No. 73, ISBN 1-881526-83-6, pp. 179, 1997.

[91] Moore GW, Hutchins GM. Effort and demand logic in medical decision making. Meta medicine,1:277–304, 1980.

[92] Moore GW, Hutchins GM, Miller RE. Token swap test of significance for serial medical databases. Am J Med,80:182–90, 1986.

[93] Zadeh LA. Fuzzy sets and information granularity. In: Gupta MM, et al., editors. Advances in fuzzy set theory and applications. Dordrecht: North-Holland, pp. 3–18, 1979.

[94] Moore GW, Berman JJ. Anatomic pathology data mining. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, p. 61–108 [chapter 4], 2000.

[95] US Department of Health and Human Services. 45 CFR (Code of Federal Regulations). Parts 160–164. Standards for Privacy of Individually Identifiable Health Information. Final Rule. Fed Regist 28 , 65(250):82461–610, 2000. (http://aspe.hhs.gov/admnsimp/).

[96] US Code of Federal Regulations, 45 CFR Subtitle A, 10-1-95 ed. Part 46. 101 (b) (4). US Department of Health and Human Services (CommonRule),56:28003,1991.

http:// ohrp.osophs.dhhs.gov/humansubjects/guidance/45cfr46.htm).

[97] US National Cancer Institute's Confidentiality Brochure, 2000. (http://www-cdp.ims.nci.nih.gov/policy.html).

[98] Saul J M, Legal policy and security issues in the handling of medical data. In: Cios KJ, editor, Medical data mining and knowledge discovery. Heidelberg: Springer, p. 17–31 [chapter 2], 2000.

[99] Pawlak Z, Rough classification, Int J Man-Mach Stud, 20, pp.469–83, 1984.

[100] Cios KJ, Kurgan LA. Trends in data mining and knowledge discovery. In: Pal NR, Jain LC, Teodoresku N, editors. Knowledge discovery in advanced information systems. Berlin: Springer, 2002.

[101] Fayyad U M, Piatesky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. Boston: AAAI Press/MIT Press, 1996.

[102] Saul J M, Legal policy and security issues in the handling of medical data. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, pp. 17–31 [chapter 2], 2000.

[103] Changeux J-P, Connes A, Conversations on mind, matter, and mathematics [De Bevoise MB, Trans.]. Princeton (NJ): Princeton University Press, 1995.