International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4– No.6, December 2012 – www.ijais.org

Data Mining in Clinical Data Sets: A Review

Shomona Gracia Jacob
Faculty of Information and Communication Engineering
Anna University,
Guindy, Chennai, India.

R Geetha Ramani, PhD.
Associate Professor
Department of Information Science and Technology, Anna University (CEG Campus)
Guindy, Chennai, India.

ABSTRACT

Data mining is one of the extensively researched areas in computer science and information technology owing to the wide influence exhibited by this computational technique on diverse fields that include finance, clinical research, multimedia, education and the like. Adequate survey and literature has been devoted to Clinical data mining, an active interdisciplinary area of research that is considered the consequent of applying artificial intelligence and data mining concepts to the field of medicine and health care. The aim of this research work is to provide a review on the foundation principles of mining clinical datasets, and present the findings and results of past researches on utilizing data mining techniques to mine health care data and patient records. The scope of this article is to present a brief report on preceding investigations made in the sphere of mining clinical data, the techniques applied and the conclusions recounted. Albeit extensive research has led to remarkable advancement in the field of clinical data mining and has paved the way for incredible enhancements in medical practice, the most recent research findings that can further unveil the potential of data mining in the realm of health care and medicine are clearly presented in this review.

General Terms

Data Mining, Classification

Keywords

Clinical data mining, Outlier Detection, Feature Selection, Clustering, Classification

1. INTRODUCTION

Data mining refers [1] to a collection of techniques that provide the necessary actions to retrieve and gather knowledge from an exhaustive collection of data and facts. Data is available in enormous magnitude, but the knowledge that can be inferred from the data is still negligible [2]. Data mining concepts [3] [4] are focused on discovering knowledge, predicting trends and eradicating superfluous data. Discovering knowledge [5] in medical systems and health care scenarios is a herculean yet critical task. Knowledge discovery [2] [3] describes the process of automatically searching large volumes of data for patterns that can be considered additional knowledge about the data [4]. The knowledge obtained through the process may become additional data that can be used for further manipulation and discovery [3].Application of data mining concepts to the medical arena has undeniably made remarkable strides in the sphere of medical research and clinical practice saving time, money and life [5-9]. Clinical data mining is the application of data mining techniques using clinical data [7]. Clinical Data-Mining (CDM) involves the conceptualization, extraction, analysis, and interpretation of available clinical data for practical knowledge-building, clinical decision-making and practitioner reflection [9]. The main objective of clinical data mining is to haul new and previously unknown clinical solutions and patterns to aid the clinicians in diagnosis, prognosis and therapy[8][9][10]. Moreover application of software solutions to store patient records in an electronic form is expected to make mining knowledge from clinical data less stressful [11].

There is a growing need in the health care scenario to store and organize sizeable clinical data, analyze the data, assist the health care professionals in decision making, and develop data mining methodologies to mine hidden patterns and discover new knowledge from clinical data[4][11]. The aforementioned factors are the target issues of clinical data mining. The basic steps involved in clinical data mining include data sampling, data analysis, data modernization, data modeling and data ranking[6][7][10]. The focus of this research is to explore and present an overview of the fundamental models and frameworks of mining clinical data, investigate existing results of mining patient records of varied nature, and brief about the challenges encountered in mining patient records.

2. SURVEYS ON CDM

There have been a great number of surveys and studies in the area of data mining, and each of the phases in data mining viz, Clustering, Feature selection, Outlier Detection and Classification play a major role in unearthing significant clinical patterns from patient records and inferring previously unknown knowledge [12][13][14][15]. The following sections present a brief survey on previous and recent reviews on mining clinical data.

2.1 Mining Clinical Data

Roddick et.al, [11] presented the experiences of the authors in applying exploratory data mining techniques to medical health and clinical data. This enabled the authors to elicit a number of general issues and provided pointers to possible areas of future research in data mining and knowledge discovery from a broad perspective. Hanauer [12] reported the challenges and solutions in mining electronic data for research and patient care. The Michigan Health system statistics were utilized for their research. However the author was concerned and focused on the hurdles involved in text mining alone. The challenges that the author inferred included affirmation of accurate diagnosis and natural language processing of electronic health records. The author had provided a solution called EMERSE (Electronic Medical Record Search Engine) that provided keyword searches for basic users and advanced features for power users. The interface was user-friendly, secure and compliant with privacy regulations and practical for implementation. He concluded stating that EMERSE was intuitive and powerful. However the system needed more training and the searching procedures continued to raise complexity. Iavindrasana et.al [7], used the nine data mining steps proposed by Fayyad in 1996 [8] as the main themes of the review. MEDLINE [16] was used as the primary source and 84 papers were retained by the authors for analysis. Their results identified three main objectives of data mining that were stated as follows: understanding of the clinical data, providing assistance to healthcare professionals, and formulating a data analysis methodology to explore clinical data. Classification was stated to be the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines). A myriad of quantitative performance measures were proposed with a predominance of accuracy, sensitivity, specificity, and ROC curves. Further work was reported by Lalayants et.al [17] who described a practice-based, mixed-method research methodology stating Clinical Data-Mining (CDM) to be a strategy for engaging international practitioners for describing, evaluating and ruminating upon endogenous forms of practice with the ultimate goal of improving practice and contributing to knowledge[9]. Such knowledge contributions were considered to be localized, but through conceptual reflection with empirical replication they could be generalized. Their article detailed CDM methodology, discussed its strengths and limitations, and illustrated international applications in Australia, Hong Kong, Israel, New Zealand, and the United States. A more elaborate account on mining clinical data was done by Epstein and Irwin as stated in the following lines. Epstein and Irwin [9] provided a clear definition of clinical data mining (CDM) as a practice-based, retrospective research strategy whereby practitioner-researchers alone or with the assistance of a research consultant, systematically extracted, codified, investigated and interpreted available qualitative and quantitative data from their own and other agency records in order to reflect on the practice, program and policy implications of their findings. Methodologically, they identified three types of clinical studies namely quantitative data, narrative data and qualitative data. Their work included a number of CDM research descriptions spanning an assortment of clinical populations covering paediatric diabetes, adolescent mental health, domestic violence, liver transplantation, geriatrics, and palliative care across the lifespan [18][19][20]. According to their survey, CDM was considered advantageous in terms of cost savings in using existing data within practices compared to creating data for prospective analysis. However they also brought about the existing flaws in CDM viz, missing data and discrepancies amongst workers on work description and record maintenance that could significantly affect and reduce the reliability and validity of practitioner-generated information.

The above reviews have focused on the broad domain of mining clinical data, the challenges encountered and the results inferred from clinical data analysis. The ensuing sections deal with specific techniques to mine and disinter patterns, associations and clusters of medical records. Moreover methods to manage and analyze voluminous patient records have also been reviewed in this work.

2.2 Data Mining Models in CDM

Clinical data mining analysis crafts effective and worthwhile knowledge that is indispensable for precise and accurate decision making [21]. Various types of mining models have been used in the past to represent interesting facts and latent patterns and trends in clinical datasets with copious applications in medical practice [22] [23]. In this subsection some of the data mining models applied to healthcare are briefly reviewed.

2.2.1 Feature Relevance Models

Clinical data are generally voluminous in nature and need special attention by virtue of data storage and analysis. Feature relevance analysis[24][25] is a phase in data mining that enables researchers to filter out certain predictors of ailments from further exploration under the pretext of being less contributory to the detection of an ailment[26]. For instance, a patient's health record may contain the concerned Patient ID, Address, and Occupation along with the evidenced clinical findings and laboratory investigation results among other details. The former factors are highly inessential in diagnosing the patient's state of health and time spent on analysis of such details is a huge squander. Such attributes need to be filtered out from further analysis and this would certainly save time and lessen computational complexity.

2.2.2 Clustering Models

Clustering is derived from mathematics, statistics, and numerical analysis [27] [28]. In this technique the dataset is partitioned into two or more factions (clusters) of similar records [29]. The clustering algorithms aim at grouping records keeping in mind the ultimate objective of maximizing a similarity metric between the members of the cluster [30]. In most cases, closeness is the similarity metric and the aim is to maximize the cumulative closeness between data records in a cluster [29] [30]. The researchers then explore the properties of the members of the generated clusters.

2.2.3 Outlier Detection Models

Outlier detection models signify novelty, anomaly, noise, variation or could be categorized under mining exceptions [31]. As quoted in [32], definition derived from Barnette & Lewis (1994) stated that an outlying observation, or outlier, is one that appeared to deviate markedly from other members of the sample in which it occurs. Indicated in study, outlier [33] normally being considered as noise, and recently under data mining approaches outliers were considered as significant details to drill out important information. One of the steps towards obtaining a coherent analysis is detection of outlying observations [34].Anomaly detection [35] provides a set of techniques that are capable of identifying rare events in large datasets. In the medical scenario, there is ample possibility for human error and negligence in making clinical data entry, interpreting clinical findings and inadequate domain knowledge leading to misconception of results [36] [37] [38]. Detection of extreme observations could eradicate incorrect data while at the same time presence of Outliers could lead to novel insights in clinical knowledge discovery [39] [41]. Hence Outliers pose a challenge in the domain of CDM and need to be handled appropriately.

2.2.4 Classification Models

A classification algorithm [38] assigns a patient's data record with specific attributes and attribute- values to a predefined class. The classification techniques in healthcare are generally applied for diagnostic purposes [40]. A classification model is built using a set of relevant attribute-values (records) derived from clinical facts and findings that lead us to generate different categories representing different nature of records[42] [43]. On comparison of a new patient's record with those of patients in different classes, one can determine to which class the new patient belongs, for instance a benign class (Non-Cancerous) or a malignant one (Cancerous).

2.2.5 Association Models

Association rule(X) Y is defined over a set of transactions T where X and Y are sets of items [44]. In a Clinical setting, the set T can be patients' clinical records and items may be symptoms, measurements, observations, or diagnosis corresponding to the patients' clinical records[44][45]. Given S as a set of items, support(S) is defined as the number of transactions in T that contain all members of the set S [47]. The confidence of a rule (X) Y is defined as support(X(Y)/support(X)), and the support of this rule is support(X(Y)). The discovered association rules show hidden patterns in the mined dataset [47]. For example, the rule: ({People who are alcoholic})/ {People needing dialysis} with a high confidence signifies that the number of people requiring dialysis is high among people who are alcoholic.

The following sections are devoted to a review on past work on clinical data investigated through data mining techniques and models.

2.3 Data Mining Techniques in CDM

This survey is aimed at providing a review on past researches in the domain of mining clinical datasets comprising of patient records and clinical findings. Data mining techniques commonly applied in the medical domain aim at classification of disease nature or prediction of the course of an ailment. Clustering and Association rule mining have been utilized in cases where similar patient records and related symptoms needed investigation [38] [39] [40] [41]. Early work on CDM was reported by Prather et.al, [48] who stated that clinical databases tend to accumulate large quantities of information about patients and their medical conditions. Relationships and patterns within the data could provide new medical knowledge. Since few methodologies existed to discover this hidden knowledge, the authors aimed to study the techniques of data mining to detect relationships in a large clinical database. Specifically, data accumulated on 3,902 obstetrical patients were evaluated for factors potentially contributing to preterm birth using exploratory factor analysis. Three factors were identified by the investigators for further exploration. Their work described the process involved in mining a clinical database including data warehousing, data query and cleaning, and data analysis. Bontempi et.al [24] made a detailed study on the impact of Feature Selection methods in Bio-informatics and stated the following as the rewards of using feature selection to mine clinical data: Facilitation of data visualization and data comprehension, reduction in the measurement and storage requirements, reduction in training and utilization times, and the ability to defy the curse of dimensionality to improve prediction performance. Venkataraman et.al [26], proposed an alternative to Univariate statistics to detect population differences in functional connectivity. The authors proposed a feature selection method based on a procedure that could search across subsets of the data to isolate a set of robust and predictive functional connections. A metric called Gini Importance was introduced that could summarize multivariate patterns of interaction, that could not be captured by Univariate techniques. The authors also compared the proposed metric with Univariate statistical tests to evaluate functional connectivity changes induced by Schizophrenia. The empirical results indicated that Univariate features showed dramatic variation across subsets of the data and had little classification power. Their evaluation proved that results based on Gini Importance were considerably more stable and permitted accurate prediction and diagnosis of a test subject.

Way et.al [49] conducted a simulation study to evaluate the performance of various combinations of classifiers and feature selection techniques and their dependence on the class distribution, dimensionality, and the training sample size. Three feature selection techniques, the stepwise feature selection (SFS), sequential floating forward search (SFFS), and principal component analysis (PCA), and two commonly used classifiers, Fisher's linear discriminant analysis (LDA) and support vector machine (SVM), were investigated. Samples were drawn from multidimensional feature spaces of multivariate Gaussian distributions with equal or unequal covariance matrices and unequal means, estimated from a clinical data set. Classifier performance was quantified by the area under the receiver operating characteristic (ROC) curve Az. The mean Az values obtained by resubstitution and hold-out methods were evaluated for training sample sizes ranging from 15 to 100 per class. The number of simulated features available for selection was chosen to be 50, 100, and 200.Their results indicated that the relative performance of the different combinations of classifier and feature selection methods depended on the feature space distributions, the dimensionality, and the available training sample sizes. The LDA and SVM with radial kernel performed similarly for most of the conditions evaluated in their study.PCA was comparable to or better than SFS and SFFS for LDA at small samples sizes, but inferior to SVM with polynomial kernel. For the class distributions simulated from clinical data, PCA did not show advantages over the other two feature selection methods. Under this condition, the SVM with radial kernel performed better than the LDA when few training samples were available, while LDA performed better when a large number of training samples were available. The authors concluded by stating that none of the investigated feature selection-classifier combinations provided consistently superior performance under the studied conditions for different sample sizes and feature space distributions. The performance of the SVM with radial kernel was better than, or comparable to, that of the SVM with polynomial kernel under most conditions studied.

Chih Lee et.al, [50] investigated the Linear Discriminant Analysis, Sequential Probability Ratio Test(SPRT) and a modified SPRT (MSPRT) empirically using clinical datasets from Parkinson's disease, colon cancer, and breast cancer. The authors assumed the same normality assumption as LDA and proposed variants of the two SPRT algorithms based on the order in which the components of an instance were sampled. Leave-one-out cross-validation was used to assess and compare the performance of the methods. Their results indicated that two variants, SPRT-ordered and MSPRT-ordered, were superior to LDA in terms of prediction accuracy. Moreover, on average SPRT-ordered and MSPRT-ordered examined fewer components than LDA before arriving at a decision. The reported results suggested that SPRT-ordered and MSPRT-ordered were the preferred algorithms over LDA.

Jacob and Ramani, [37] addressed the task of grading the performance of feature selection and classification algorithms on clinical datasets of varied nature. In 2011, the authors investigated the performance of sixteen data mining classification algorithms viz. Random Tree, Quinlan's decision tree algorithm (C4.5), K-Nearest Neighbor algorithm etc., on the 'Wisconsin Breast tissue dataset' (derived from the UCI Machine Learning Repository) that comprised of 11 attributes and 106 patient records. The analysis indicated the level of training accuracy and other performance measures of the algorithms in detecting the presence of breast cancer and the associated breast tissue conditions that raised the risk of developing cancer in future. Moreover the importance of feature selection in improving the performance of classification algorithms was recorded. The classification algorithm Random Tree produced 100 percent accuracy for classification of all the training data under multiple classes.

Jacob et.al, [39] further explored the performance of classification algorithms on the Breast Cancer dataset through Data mining algorithms. Their research aimed at recognizing the significance of feature selection in classifying Breast Cancer data under two classes namely Benign (non-cancerous) and Malignant (cancerous). They examined the performance of six feature relevance algorithms on the Wisconsin Breast Cancer (WBC) dataset comprising of 699 patient records and the Wisconsin Diagnostic Breast Cancer (WDBC) dataset comprising of 569 patient details obtained from the UCI Machine Learning Repository. A comparison of twenty classification algorithms based on their Misclassification Rate was portrayed. Their results indicated that the Random Tree Classification algorithm produced 100 percent training accuracy in classifying the input datasets. The Lymphography dataset comprising of 18 predictor attributes and 148 patient records was also explored by the authors to highlight the performance of sixteen classification algorithms and to estimate the accuracy of the classification algorithms in performing multi-class categorization of medical data [51]. Furthermore their research placed emphasis on the performance of four feature selection algorithms and their impact on the classification accuracy. The authors found that the Random Tree algorithm and the Quinlan’s C4.5 algorithm produced 100 percent classification accuracy with all the predictor features and also with the feature subset selected by the Fisher Filtering feature selection algorithm. Moreover ReliefF feature selection algorithm gave improved results for Radial Basis Function algorithm improving the classification accuracy by 1.35%. It is also stated here that the C4.5 algorithm offered more efficient classification since the decision tree size generated comprised of fewer nodes than the Random Tree algorithm.

Jacob et.al, [38] analyzed the performance of twenty classification algorithms on the (Oncovirus) cancerous Hepatitis C virus dataset from the UCI Machine Learning Repository [52]. The authors performed binary classification on the dataset, comprising of 155 instances and 19 predictor features. The performance of the classification algorithms revealed that Random Tree Classification and Quinlan's C4.5 algorithm classified the patient records as infected and healthy with 100% accuracy. A similar work was carried out on the Splice Junction DNA sequence data which involved Multi-class categorization of DNA sequence records into Exon, Intron and Neither category [53]. This study involved the execution of nine classification algorithms on the Splice junctions of 3190 DNA sequences taken from the Keel data repository, each having 60 nucleotides, to detect the boundaries between introns and exons that will further aid in the process of analyzing genetic markers and understanding the mechanism of protein synthesis. The Quinlan’s C4.5 algorithm and the Random Tree classification algorithm revealed 99.97% classification accuracy on the DNA data.

In a another major attempt to explore the classification algorithms, Jacob et.al [41] executed classification algorithms on diverse clinical data comprising of patient records related to the following ailments viz, Mammography masses, Heart Disease, Dermatology infection, Orthopaedic ailment and Thyroid diseases. The authors made a careful selection of clinical data from varied domains in order to identify the performance of the data mining algorithms on different types of clinical datasets. The research work proposed the design of a data mining framework that generated an efficient classifier, trained on all the clinical datasets stated, by learning patterns and rules framed by executing classification algorithms. The results validated that the Quinlan’s C4.5 classification algorithm and the Random Tree algorithm yielded 100 percent classification accuracy on SPECTF Heart, Orthopaedic (Vertebral Column) ailments, Thyroid and Dermatology infection datasets while Binary Logistic Regression and Cost Sensitive Classification with Least Misclassification Cost also produced 100 percent classification accuracy on the SPECTF Heart Dataset while Multinomial Logistic Regression classified the Dermatology dataset with 100 percent accuracy. However on the Mammography training dataset, the classification accuracy produced by Random Tree and Quinlan’s C4.5 algorithms was only 91.36%. The Decision tree generated by the Quinlan’s C4.5 algorithm is smaller than the decision tree given by the Random Tree classification technique.

Churilov et al. [54] described a clustering method using an optimization approach to extract risk grouping rules for prostate cancer patients. The data record fields were the patient’s age, tumor stage, Gleason score, and PSA level. The clustering algorithm generated 10 clusters that were later grouped to low, intermediate and high risk categories.

Jacob et.al [55] analyzed the Cardiotocography dataset from the UCI Irvine Machine Learning Repository comprising of 2126 Fetal Heart Rate (FHR) and morphology pattern (MP) records and 21 predictor attributes providing life saving information on the state of the fetus in the womb. They classified the fetal records into three target classes and their analysis state 100 percent classifier accuracy for Random Tree and Quinlan’s C4.5 algorithm in the case of the FHR dataset. They also aimed at detecting the outlier records and revealed the improved classification rate of the classification algorithms post outlier detection. Possible clusters in the Cardiotocography dataset that unearthed the significance of each attribute in deciding the cluster generation were determined using Variable Clustering, a Top Down approach.

Some classical examples for inward procedures of Outlier detection can be found in [32] [56] [57]. The classification to inward and outward procedures also pertains to multivariate outlier detection methods. However, existing outlier detection methodology focuses mostly on detection of anomalous data entries in the datasets. This pilots a pre-conceived notion that all outliers (anomalies) are events that occur with a low prior probability. In the medical scenario, this may lead to neglecting or overlooking patients suffering from a rare disease or exhibiting a rare combination of symptoms [58] [59]. To account for this conditional aspect of outlier detection in medicine Hampel [60] introduced, developed and continued developing a new conditional anomaly detection framework that sought to detect unusual values for one or a subset of variables given the values of the remaining variables. Hampel [60] [61] introduced the concept of the *breakdown point*, as a measure for the robustness of an estimator against outliers. The breakdown point is defined as the smallest percentage of outliers that could cause an estimator to take arbitrary large values.

Chauhan et.al, [30] made a detailed study of Hierarchical, Partitioning-based, Density-based and Grid-based clustering algorithms. Their conclusions suggested the use of K-Means and Hierarchical Agglomerative Clustering for mining clinical databases. Wilson et al. [36] discussed potential uses of data mining techniques in pharmaco-vigilance to detect adverse drug reactions.

Ramesh Kumar [44] had proposed an algorithm named nVApriori to mine interesting rules from HIV infected Patient's Treatment records. This is a n-cross validation based Apriori (nVApriori) algorithm to mine domain irrelevant rules. Acquired Immuno Deficiency Syndrome (AIDS) is a critical disease in the medical domain. The author proposed a new dataset for AIDS/HIV infected patients' case history. The data were collected from Midwest clinics, London. The nVApriori algorithm was applied to the proposed dataset. A number of interesting rules were mined thus providing novel insights to domain experts. Their results stated that the proposed algorithm performed better than traditional Apriori algorithm.

Exarchos et.al [62] devised a new automated methodology based on Association rules for detection of Ischemic beats in long duration Electrocardiographic recordings (ECG). The proposed approach comprised of three stages viz, Pre-processing, Discretization and Classification. Noise removal and extraction of required features was done during the Pre-processing phase. The continuous valued features were transformed to categorical during the second phase. An association rule extraction algorithm was utilized and a rule-based classification model was formulated. According to the proposed methodology, electrocardiogram (ECG) features extracted from the ST segment and the T-wave, as well as the patient's age, was used as inputs. The output was the classification of the beat as ischemic or non-ischemic. Various algorithms were tested both for discretization and for classification using association rules. To evaluate the methodology, a cardiac beat dataset was constructed using several recordings of the European Society of Cardiology ST-T database. The obtained sensitivity (Se) and specificity (Sp) was 87% and 93%, respectively. The proposed methodology combined high accuracy with the ability to provide interpretation for the decisions made, since it was based on a set of association rules. Their obtained sensitivity (Se) and specificity (Sp) was recorded as 87% and 93%, respectively.

Vararuk et.al, [63] made use of data mining techniques to extract and investigate patterns in HIV/AIDS patient data. These patterns aimed to provide better management of the disease and targeting of resources. A total of 250,000 anonymised records from HIV/AIDS patients in Thailand were imported into a database. IBM's Intelligent Miner was used for clustering and association rule discovery. Clustering highlighted groups of patients with common characteristics and also errors in data. Association rules identified associations that were not expected in the data and were different from traditional reporting mechanisms utilised by medical practitioners. It also allowed the identification of symptoms that co-existed or proved to be precursors of other symptoms. The significance of the study was stated as follows: Identification of symptoms that were precursors of other symptoms could allow the targeting of the former so that the later symptoms could be avoided. The research showed provisioning a pragmatic and targeted approach to the management of resources available for HIV/AIDS treatment. The authors work suggested implementation of a quality monitoring system to target available resources.

Ordonez et al. [64] proposed a new algorithm to mine association rules in medical data with additional constraints on the extracted rules and applied the method to predict heart disease. A decision tree-based classification approach was applied to mass spectral data to help diagnosis of ovarian cancer suspects. While association rule classifiers have been applied to diagnose breast cancer using digital mammograms, Land et al. [65] made use of Neural Network based classification approach for the same purpose. Association rules mining was also applied over data of human sleep time. Duch et al. [66] compared various data mining methods supporting diagnosis of Melanoma skin cancer.

The following section provides a brief review on the design and use of data mining frameworks to mine Clinical data.

3. CLINICAL DATA MINING SYSTEMS

Clinical data mining [7] refers to the collection of algorithms, techniques and methods to discover previously unknown, new patterns from clinical data that could aid clinicians, heath-care practitioners, medical researchers, and scientists in disease diagnosis and prognosis, genetic marker detection and drug therapy [67]. We deem it essential to brief about the methodology involved in the design of a framework to mine clinical data whose domain covers basic medical case sheets to full-length genome sequence and amino acid substitutions.

The basis for any data mining framework involves a preliminary learning phase during which the problem is modeled followed by the test phase that validates the constructed model [40]. The learning process can be accomplished either in a Supervised or Unsupervised manner [1] [43]. Supervised Learning [10] requires the training data to be accompanied by class labels and the test data is classified based on the training set, whereas in unsupervised learning, the class label is unknown and the aim is to establish the existence of clusters or classes in the data[55][58]. Models required to mine data are classified into Predictive and Descriptive models. Clustering and Association Rule Models are descriptive while Classification and Regression models are stated to be predictive.

3.1 Clinical Data Mining Frameworks

The general approach to mine clinical data comprises of the following phases namely Data collection, Data Pre-processing, Feature Selection, Classification and Evaluation [37] [38] [39] [40]. Inclusion of Outlier detection prior to Classification could reduce computational complexity and remove sparse and unrelated patient data [55]. We also attempt to summarize the Clustering techniques to group similar medical records into classes. Moreover the dependencies among symptoms and diseases can also be identified through Association Rule Mining.

Abe et.al [68] introduced the concept of categorized and integrated data mining. The authors reviewed the rapid progress in medical science, medical diagnosis and treatment and perceived the need for an integrated and cooperative research among medical researchers, biology, engineering, cultural science, and sociology. Hence they proposed a framework called Cyber Integrated Medical Infrastructure (CIMI), a framework of integrated management of medical data on computer networks consisting of a database, a knowledge base, and an inference and learning component, connected to each other in the network. The framework had the capacity to deal with diverse types of data which required integrated analysis of diverse data. In their study, for medical science, they analyzed the features and relationships among various types of data and revealed the possibility of categorized and integrated data mining. Anand et.al, [69] presented a framework to incorporate parallelism in mining data. This led to the enhancement of algorithms being developed within this framework to be parallel and was hence expected to be efficient for large data sets, a definite need for medical data mining. The parallelism within the framework permitted distribution and heterogeneity. The framework could be easily updated and new discovery methods could be readily incorporated within the framework. The framework provided a spontaneous view of handling missing data during the discovery process using the concept of Ignorance borrowed from Evidence Theory. The framework incorporated the possibility of representing data and knowledge, and methods for data manipulation and knowledge discovery. They suggested an extension of the conventional definition of mass functions in Evidence Theory for use in Data Mining, as a means to represent evidence of the existence of rules in the database. The discovery process within EDM consisted of a series of operations on the mass functions. Each operation was carried out by an EDM operator. Also included was a classification for the EDM operators based on the discovery functions performed by them and a discussion of induction, domain and combination operator classes was carried out. The application of EDM to two separate Data Mining tasks was also addressed, highlighting the advantages of using a general framework for Data Mining and, in particular, using one that was based on Evidence Theory. Lin and Haug, [70] proposed an approach to data preparation that utilized information from the data, metadata and sources of medical knowledge. Heuristic rules and policies were defined for the three types of supporting information. Compared with an entirely manual process for data preparation, their proposed approach could potentially reduce manual work by achieving a degree of automation in the rule creation and execution. A pilot experiment demonstrated that data sets created through the approach lead to better model learning results than a fully manual process. This study was conducted using data extracted from the enterprise data warehouse (EDW) of Intermountain Health Care (IHC) in Salt Lake City [71]. The data was captured during routine clinical care documented in the HELP [4] hospital information system [72]. IHC had established a working process that duplicated data from the HELP system to a data mart in the EDW called the “HELP” data repository. The authors developed a system to detect patients who were admitted to the hospital with pneumonia. The data set included data of patients who were discharged from the hospital with pneumonia as primary diagnosis as well as a group of control patients. Both groups were sampled from patients admitted to LDS Hospital from the year 2000 to 2004. The manual approach was to acquire variables relevant to pneumonia according to domain knowledge and the medical literature. Keyword searches were used on the code description field to find a list of candidate data codes. The candidate codes were inspected and the most suitable codes were chosen. The earliest observed value for each code was selected as the summary value for the chosen period. Each time no value was found for an instance of a variable, the variable was discretized and a state called ‘missing’ was added to it. By using the aforementioned process, a data set in flattened-table format was created from the original data. In their experimental approach, two types of heuristic rules were used to select variables. One was to pre-screen data elements based on their statistical characteristics and their gross categorization in the data dictionary. This allowed selection of data subsets that were relevant to the clinical model being developed. The second was to select data elements that were able to differentiate the specific clinical problem according to comparative statistics calculated from the test and control groups across all candidate variables. The candidate variable list was then manually inspected to remove obviously irrelevant variables. The numbers of patients in the case and control groups were 1521 and 1376 respectively. The two 95% confidence intervals of the difference of ROC were found to be above zero, indicating that the difference was statistically significant (a=0.05). The results revealed the fact that the two tested model learning algorithms performed better with the data set prepared by the framework.

Hwang et.al [73], suggested a data-mining framework that utilized the concept of clinical pathways to facilitate automatic and systematic construction of an adaptable and extensible detection model. The proposed approaches were evaluated objectively by a real-world data set gathered from the National Health Insurance (NHI) program in Taiwan. The empirical experiments show that the detection model was efficient and capable of identifying some fraudulent and abusive cases that were not detected by a manually constructed detection model. In the said research, the authors proposed a process-mining framework that utilized the concept of clinical pathways to facilitate the automatic and systematic construction of an adaptable and extensible detection model. They assumed a data-centric point of view and consider healthcare fraud and abuse detection as a data analysis process. The focus of their approach was to apply process-mining techniques to collected clinical-instance data to construct a model that differentiated fraudulent behaviors from normal activities. The proposed automatic approach eradicated the need to manually analyze and encode behavior patterns, as well as the guesswork in selecting statistics measures. The proposed framework was evaluated via real-world data to demonstrate its efficiency and accuracy. The authors had outlined a framework that facilitated the automatic and systematic construction of systems that detected healthcare fraud and abuse. The authors also investigated the mining of frequent patterns from clinical instances and the selection of features that depicted higher discrimination power. The proposed approaches were evaluated objectively using a real-world data set gathered from the NHI program in Taiwan. The empirical experiments showed that the detection model was efficient and capable of identifying certain fraudulent and abusive cases that were not detected by a manually constructed detection model.

Mc. Gregor et.al, [74] presented a framework for process mining in critical care. The CRoss Industry Standard Process for Data Mining (CRISP-DM) [75] was developed in 1996, with the goal of being industry, tool and application-neutral. Constant references to the methodology by analysts have established it as the de facto standard for data mining and Knowledge Discovery in Databases (KDD) [76]. The model breaks the life cycle of a data mining project into six phases namely business understanding, data understanding, data preparation, modeling, evaluation and deployment. The proposed framework utilized the CRISP-DM model[77], extended to incorporate temporal and multidimensional aspects (CRISP-TDMn), in conjunction with the Patient Journey Modeling Architecture (PaJMa), to provide a structured approach to knowledge discovery of new condition onset pathophysiologies in physiological data streams[74]. The approach was based on temporal abstraction and mining of physiological data streams to develop process flow mappings that could be used to update patient journeys; instantiated in critical care within clinical practice guidelines. The author's demonstrated the framework within the neonatal intensive care setting, where current clinical research in relation to pathophysiology within physiological streams of patients diagnosed with late onset neonatal sepsis was being carried out. A portray of the instantiation of the framework for late onset neonatal sepsis was given, using CRISP-TDMn for the process mining model and PaJMa for the knowledge representation. Their research presented a generalized framework to support process mining in critical care that enabled knowledge discovery of new condition onset pathophysiologies using temporal data mining of physiological data streams and constructed process flow mappings that could be used to update patient journeys as instantiated within clinical practice guidelines. The research was demonstrated within the context of neonatal intensive care and specifically as it relates to LONS and the potential of reduced heart rate variability and apnea to be pathophysiologies for the said clinical condition. They had demonstrated the ability for this framework to be used by clinical researchers as a platform to perform knowledge discovery, accrue the new knowledge, and decipher the knowledge to updated clinical guidelines.

Kazemzadeh et.al, [78] focused on encoding, sharing, and using the results of data mining analyses for clinical decision making at the point of care. With the aforesaid objective in mind, a knowledge management framework was proposed that addressed the issues of data and knowledge interoperability by adopting healthcare and data mining modeling standards, HL7 [79] and PMML [80] respectively. A prototype tool was developed as part of their research that provided an environment for clinical guideline authoring and execution capable of applying and interpreting data mining results. Moreover, three real world case studies were presented. The authors also described a novel framework for dissemination and application of the data and mined knowledge among the heterogeneous healthcare information systems. For data interoperability, HL7 Clinical Document Architecture (CDA) schema [81] was used to define the required structure for encoding patients' health related data. The healthcare researchers extracted knowledge by mining existing healthcare data in an off-line operation and stored in proprietary databases. They used the PMML specification to encode the produced mined knowledge to achieve knowledge interoperability between sources of knowledge and their users. This was reported to be the first methodology to make this type of knowledge portable and available at the application sites. Further on, decision modules could access patient data from CDA documents and supply them into the data mining models from the PMML documents. The results of this operation were also provided as CDA documents to allow interoperability of the results. Moreover, the authors utilized the mined knowledge in the proposed extension to the GLIF3 [82] clinical guideline modeling language that provided recommendations and warnings to the healthcare personnel based on the results of knowledge application.

Rapid and extensive research has attracted science and engineering professionals to work in a cohesive manner to the advancement in the domain of clinical data mining. The following section depicts a brief outlook on the rising spheres of CDM.

4. APPLICATIONS OF CLINICAL DATA MINING

Several reviews and surveys have been reported in the past that have portrayed the impact of data mining techniques in refining health care applications[5][6][8][9][68]. A concise view of the recent work in the area of clinical data mining and their contribution to the advancement of clinical practice, data management and research is presented in this section.

4.1 Data Mining in Clinical Data Management

Data pre-processing techniques have been widely used in management of medical data and patient records [14] [15] [70]. The large volume of data available needs to be formatted and collected in a manner that will permit secure and simple retrieval when needed, faster and efficient mining of credential information and economic utilization of storage space and computation time. Electronic health records [12] [14] were the first attempt to securely manage the patient records and are currently used in practice in several medical institutions and health care centres around the world. Distributed network [83] of medical records was another innovation that spurred a renaissance in the medical field that allowed clinicians to share patient information for the purpose of obtaining an expert opinion or sharing the storage space available in another network and even for providing backup facility. Mining of information from such distributed records is currently an intense area of research.

4.2 Knowledge-Based Systems/Clinical Decision Support Systems

Several studies [38][39][40][41]have been reported on the results of mining medical data by application of data mining techniques that include feature selection, outlier detection and classification/ prediction. Each of the algorithms is evaluated and the technique that produces the best classification accuracy is chosen. The rules generated by the classification algorithm and the medical data records on which the data mining techniques were executed constitute the Knowledge Base which is the core component of any data mining framework[53][55]. Following this any medical record relative to the particular ailment under study can be input to the classifier and the precision in classification can be verified from the clinical decision of the system. Hence such classifier systems offer support to the medical practitioners in predicting the course of a disease based on the existing symptoms, proposing drugs, identifying the need for hospitalization and predicting possible time for recuperation. Other data mining applications related to clinical practice include associating the various side-effects of treatment, collating common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub-populations that respond differently from the mainstream population to certain drugs, and determining proactive steps that can reduce the risk of affliction.

Data mining techniques thus provide better assessment of patient needs, better information about clinical fidelity, superior idea about patient outcomes and association between interventions and outcomes[4][36][42][54]. This has stimulated application of computational techniques [84] [85] in mining of medical data streams, customer relationship management and fraud detection related to non-compliance with security/ethical issues of clinical data.

4.3 DNA Sequence Analysis for Genetic Marker Detection

Data mining techniques have proven to produce improvement in the analysis, classification of the affection status of more individuals and by locating more single nucleotide polymorphisms related to the disease [86]. Molecular genetic markers represent one of the most influential tools for the analysis of genomes and enable the association of inborn traits with underlying genomic diversity [87]. Molecular marker technology has developed rapidly over the last decade and two forms of sequence based markers, Simple Sequence Repeats (SSRs), also known as microsatellites, and Single Nucleotide Polymorphisms (SNPs) now preponderate applications in modern genetic analysis [88]. The diminishing price of DNA sequencing has led to the availability of large sequence data sets derived from whole genome sequencing and large scale Expressed Sequence Tag (EST) discovery have enabled the mining of SSRs and SNPs [89]. These can later be applied to diversity analysis, genetic trait mapping, association studies, and marker assisted selection. These markers are economical, require minimal labour to produce and can frequently be associated with annotated genes.

In recent years there has been an upsurge in the rate of acquisition of biomedical data. Advances in molecular genetics technologies, such as DNA microarrays [86] [87] [88] [89] [90] allow laymen for the first time to obtain a comprehensive view of the cell. Machine learning and statistical techniques applied to gene expression data have been used to focus on the questions of distinguishing tumour morphology, predicting post-treatment outcome, and finding molecular markers for diseases. On record today, the microarray-based classification of different morphologies, lineages and cell histologies can be performed successfully in many instances. The performance in predicting treatment outcome or drug response has been quite limited although some results are quite promising. Most results of microarray analysis still require further experimental validation and follow up study. In a few cases the results of microarray analysis have found their way into more serious consideration in clinical use. Recent advances in data mining applications include Gene Selection that is a process of attribute selection, which finds the genes most strongly related to a particular class[91][92][93][94], Clustering aims at finding new biological classes or refining existing ones [95][96][97][98][99][100] and Classification aims at classifying diseases or predicting outcomes based on gene expression patterns, and includes identifying the best treatment for a given genetic signature [38-43][50][53][55][65][100-105]

The influence exerted by data mining techniques can span wider avenues only when the current obstacles in mining medical data are handled in an appropriate manner. The following section narrates the existing challenges in mining clinical data and patient records.

5. CHALLENGES IN CLINICAL DATA MINING

Clinical data mining is certainly limited by the ease of access to medical findings, since required facts for data mining often exist in different settings, forms and systems, viz, administration, clinics, laboratories and other. This calls for a strategy to gather and integrate data before data mining can be done [106] [107]. While several authors and researchers have suggested the need for a data warehouse prior to mining clinical data, the expenses involved challenge their utility. However, Intermountain Health Care have successfully implemented a warehouse from five different sources— a clinical data repository, acute care case-mix system, laboratory information system, ambulatory case-mix system, and health plans database and imparted better evidence-based clinical solutions[84]. Research by Oakley [83] suggested a distributed network topology instead of a data warehouse for more efficient data mining. Another imposing hurdle in medical data collection include missing, distorted, conflicting, and non-homogenous data, such as bits of information recorded in different formats in diverse data sources. Precisely, the lack of a standardized clinical vocabulary is a serious hindrance to data mining. Cios and Moore [108] have posed a dispute that data problems in healthcare are the result of the dimensionality, intricacy and assorted nature of medical data and their low mathematical characterization and non conformance to a certain protocol. Moreover ethical, legal and social issues encountered in CDM also have to be appropriately handled. The issue of obtaining patterns of diverse nature on exhaustive mining of data needs to be deliberated upon. Extensive research may reveal many interesting patterns and relationships not necessarily valuable. The successful application of data mining requires expertise in data mining methodology and tools not ignoring realistic knowledge of medical practice. Data mining applications in healthcare can have tremendous potential and efficacy. However, the success of healthcare data mining hinges on the availability of clean healthcare data [4-6] [17] [23]. In this respect, it is critical that the healthcare industry consider diverse ways and means of capturing, storing, processing and mining data [42] [44] [45]. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. Further, as healthcare data are not limited to patient records, it is necessary to explore the use of text and image mining approaches to expand the scope and nature of clinical data mining[109][110].

6. CONCLUSION

Clinical data mining is a practice –based research strategy by which practitioners and researchers retrieve, analyze and interpret available qualitative and quantitative information from available medical records. CDM is a highly motivated area of research due to the extensive influence exerted by this multi-domain research area that brings together interests of medical practitioners, computer science researchers and health care professionals. Mining of clinical facts is highly essential due to the availability of exhaustive and enormous volume of medical records. This paper presented a review of clinical data mining concepts and the data mining techniques applied in clinical practice. Designs of Data mining framework for clinical data mining systems have been reviewed to provide researchers an initiative to formulate new techniques for clinical record analysis and exploration, besides reforming the flaws in the existing systems. Also stated are the principles behind the existing applications of clinical data mining, the challenges existing in CDM and future directions for research. This research study is expected to be a significant contribution to researchers and practitioners in the data mining and clinical industry.

7. ACKNOWLEDGEMENT

This research work is a part of the All India Council for Technical Education(AICTE), India funded Research Promotion Scheme project titled “Efficient Classifier for clinical life data (Parkinson, Breast Cancer and P53 mutants) through feature relevance analysis and classification" with Reference No:8023/RID/RPS-56/2010-11, No:200-62/FIN/04/05/1624.

8. REFERENCES

[1] Ian H. Witten; Eibe Frank; Mark A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques" (3 Ed.). Elsevier. ISBN 978-0-12-374856-0

[2] Cabena, Peter, Pablo Hadjnian, Rolf Stadler, Jaap Verhees and Alessandro Zanasi (1997). "Discovering Data Mining: From Concept to Implementation" Prentice Hall, ISBN 0-13-743980-6.

[3] Xingquan Zhu, Ian Davidson (2007). "Knowledge Discovery and Data Mining: Challenges and Realities." Hershey, New York. p. 18. ISBN 978-1-59904-252-7.

[4] Debahuti Mishra , Asit Kumar Das, Mausumi and Sashikala Mishra, "Predictive Data Mining: Promising Future and Applications", Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010

[5] Dave Smith, SAS, Marlow, UK, "Data Mining in the Clinical Research Environment", PhUSE 2007.

[6] Prasanna Desikan, Hsu, Srivastava, "Data mining for health care management", 2011 SIAM International Conference on Data mining.

[7] Iavindrasana J et.al, Clinical data mining: a review. Med Inform. 2009:121-33. Review.

[8] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. Retrieved 2008-12-17.

[9] Epstein, Irwin. (2010). Clinical data-mining: Integrating practice and research. London. Oxford University Press

[10] Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005). Introduction to Data Mining. ISBN 0-321-32136-7

[11] John F.Roddick, Peter Fule, Warwick J.Graco, "Exploratory Medical Knowledge Discovery: Experiences and Issues", 2004.

[12] David Hanauer, MD, MS Mining clinical electronic data for research and patient care: Challenges and solutions, Clinical Assistant Professor University of Michigan, USA, 2007 September

[13] R. Agrawal et al., Fast discovery of association rules, in Advances in knowledge discovery and data mining pp. 307–328, MIT Press, 1996.

[14] Bennett CC and TW Doub. (2010) "Data mining and electronic health records: Selecting optimal clinical treatments in practice". Proceedings of the 6th International Conference on Data Mining. pp. 313-318.

[15] M.F. Ochs et al. (eds.), "Clinical Research Systems and Integration with Medical Systems", Biomedical Informatics for Cancer Research,DOI 10.1007/978-1-4419-5714-6_2, © Springer Science Business Media, LLC 2010

[16] Medline Resources http://www.nlm.nih.gov/bsd/pmresources.html

[17] Lalayants et.al, "Clinical data-mining: Learning from practice in international settings", International Social Work March 27, 2012, doi: 0020872811435370

[18] Jerome Beker, Anthony J Grasso Dsw, Irwin Epstein, Boysville Of Michigan, Information Systems in Child, Youth, and Family Agencies, Published October 11th 1993 by CRC Press

[19] Irwin Epstein, Susan Blumenfield, Clinical Data-Mining in Practice-Based Research, May 7th 2002 by Routledge

[20] Irwin Epstein, Ken Peake, Daniel Medeiros, Clinical and Research Uses of an Adolescent Mental Health Intake Questionnaire, August 14th 2005 by Routledge

[21] Gregory Piatetsky-Shapiro, Pablo Tamayo, "Microarray Data Mining: Facing the Challenges" SIGKDD Explorations. Volume 5, Issue 2.

[22] Weiss and Indurkhya. Predictive Data Mining. Morgan Kaufmann

[23] Riccardo Bellazzi, Blaz Zupanb, Predictive data mining in clinical medicine: Current issues and guidelines"", international journal of medical informatics 7 7 (2 0 0 8) 81–97.

[24] G. Bontempi. "Structural feature selection for wrapper methods". In Proceedings of ESANN 2005, European Symposium on Artificial Neural Networks, 2005.

[25] Jiang et.al, Feature Mining Paradigms for Scientific Data, Copyright © by SIAM

[26] Archana Venkataraman, Marek Kubicki, Carl-Fredrik Westin, Polina Golland, "Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies", 978-1-4244-7028-0/10/$26.00 ©2010 IEEE

[27] M. Sacha. (2008) "Clustering of a periodical medical Knowledge -Constrained K-means Clustering with Background data." in Proceedings of the Eighteenth http://www.mareksacha.com/blog/clustering-of-an-International Conference on Machine Learning, 2001, a periodical-medical-data. pp. 577 - 584.

[28] G. Y. Hang, D. Zhang, J. Ren, and C. Hu, "A Machine Learning Repository: Hierarchical Clustering Algorithm Based on K-Means http://archive.ics.uci.edu/ml/support/Zoo with Constraints," in Fourth International Conference on Innovative Computing, Information and Control, Kaohsiung, Taiwan, 2009, pp. 1479-1482

[29] Lin W. and C. Le "Model-based cluster analysis of microarray gene expression data". Genome Biology, 3(2): research0009.1-0009.8, (2002).

[30] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010

[31] V.Elango, R.Subramanian,V.Vasudevan, "A Five Step Procedure for Outlier Analysis in Data Mining" , European Journal of Scientific Research, ISSN 1450-216X Vol.75 No.3 (2012), pp. 327-339.

[32] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons. 3rd edition.

[33] Zalizah Awang Long , Abdul Razak Hamdan and Azuraliza Abu Bakar , "Framework on Outlier Sequential patterns for Outbreak Detection", 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

[34] Balakrishnan, N.; Childs, A. (2001), "Outlier", in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4

[35] Chandola V, Banerjee A, Kumar V."Anomaly Detection - A Survey". 41(3):2009. ACM Computing Surveys.

[36] Valko M, et al. "Conditional anomaly detection methods for patient-management alert systems". ICML Workshop on Machine Learning in Health Care Applications. 2008

[37] Mrs.Shomona Gracia Jacob and Dr. R.Geetha Ramani,"Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data", International Journal of Computer Applications (IJCA), 32(7): 46-53, October 2011a DOI: 10.5120/3920-5521. Published by Foundation of Computer Science, New York, USA.

[38] Shomona Gracia Jacob, Dr.R. Geetha Ramani, Nancy .P (2012), "Efficient Classifier for Classification of Hepatitis C Virus Clinical Data through Data Mining Algorithms and Techniques", Proceedings of the International Conference on Computer Applications, Pondicherry, India, January 27-31, 2012,Techno Forum Group, India. ISBN: 978-81- 920575-8-3: DOI: 10.73445/ISBN_0768, ACM#.dber.imera.10.73445.

[39] Shomona Gracia Jacob, Dr.R.Geetha Ramani, P.Nancy (2011 b), "Feature Selection and Classification in Breast Cancer Datasets through Data Mining Algorithms", Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC'2011), Kanyakumari, India,, IEEE Catalog Number: CFP1120J-PRT, ISBN: 978-1-61284-766-5. Pp. 661-667

[40] Shomona G J, R.Geetha Ramani, "Evolving Efficient Classification Rules from Cardiotocography data through Data mining methods and Techniques", Vol.78, Issue.3 PP. 668-680.

[41] Shomona G.J. R.Geetha Ramani, "Mining of Classification Patterns in Clinical data through data mining methods and techniques", Proceedings of the International Conference on Systemics, Cybernetics and Informatics, Held at Chennai, India, during August 3-5, 2012, pp. 997-1003

[42] Serhat Özekes And A.Yilmaz Çamurcu, "Classification And Prediction In A Data Mining Application", Journal Of Marmara For Pure And Applied Sciences, 18 (2002) 159-174Marmara University, Printed In Turkey

[43] Khan J. et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks". Nature Medicine, Volume 7, Number 6, June 2001.

[44] K.Ramesh Kumar, "Extracting Association Rules from HIV Infected Patient's treatment dataset", Trends in Bioinformatics 4(1):35-46, 2011, ISSN 1994-7941 / DOI:10.3923/tb.2011.35.46

[45] Roberto J Bayardo, Rakesh Agarwal, "Mining the Most Interesting Rules", Appears in Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 145-154, 1999.

[46] Maragatham G et al.,"A Recent Review on Association Rule Mining", Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166 Vol. 2 No. 6 Dec 2011-Jan 2012 832-835.

[47] J. Han and M. Kamber, Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers, 2000

[48] Prather et.al. "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", 1091-8280/97/$5.00 0 (1997) AMIA, Inc.

[49] Ted W. Way, Berkman Sahiner, Lubomir M. Hadjiiski, and Heang-Ping Chan, "Effect of finite sample size on feature selection and classification: A simulation study", Med Phys. 2010 February; 37(2): 907–920.

[50] Chih Lee, Brittany Nkounkou, and Chun-Hsi Huang, "Comparison of LDA and SPRT on Clinical Dataset Classifications", Biomed Inform Insights. 2011 April 19; 4: 1–7.doi: 10.4137/BII.S6935

[51] Shomona Gracia Jacob, Dr.R.Geetha Ramani, Nancy.P, „Discovery of Knowledge Patterns in Lymphographic Clinical Data through data mining methods and Techniques", International Conference on Artificial Intelligence, Soft Computing and Application (AIAA, 2012), Held at Chennai, India, July 13th -15th 2012, Advances in Computing and Information Technology, AISC 178,pp.129-140. Springer Proceedings

[52] University of California, Irvine, Machine Learning Repository, www.ics.uci.edu/~mlearn/.

[53] Shomona Gracia Jacob, Dr.R.Geetha Ramani, Nancy.P, „Classification of Splice Junction DNA sequence data through Data mining techniques", ICFCCT, 2012, held at Beijing, China, May 19-20, 2012. Pp.143-148, ISBN:978-988-15121-4-7.

[54] Leonid churilov , Adyl Bagirov , Daniel Schwartz , Kate Smith , Michael Dally „ "Data mining with combined use of optimization Techniques and Self-Organizing Maps for Improving Risk Grouping Rules: Application to Prostate Cancer Patients", Journal of Management Information Systems Issue: Volume 21, Number 4 / Spring 2005, Pages: 85 - 100

[55] Shomona Gracia Jacob, Dr.R.Geetha Ramani, "Evolving Efficient Classification Rules from Cardiotocography data through Data mining Techniques", European Journal of Scientific Research,June, 2012 Vol.78, Issue 3, pp.468-480. (SNIP:0.010:SJR:0.071)

[56] Bernard Rosner, "Percentage Points for a Generalized ESD Many-Outlier procedure", Technometrics, Vol.25, No.2, May, 1983.

[57] Hawkins, D.M., (1980) "The Identification of Outliers", Chapman and Hall, London.

[58] Rich Caruana, Alexandru NiculescuMizil, "Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria", KDD'04, August 22–25, 2004, Seattle, Washington, USA.Copyright 2004 ACM 1581138881/04/0008

[59] D.V. Chandra Shekar and V.Sesha Srinivas, "Clinical Data Mining – An Approach for Identification of Refractive Errors", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I, IMECS 2008, 19-21 March, 2008, Hong Kong

[60] Hampel.F., "A General Quanlitative definition of Robustness, Ann. Math. Statist, 42, 1887-1896., 1971.

[61] Hampel.F, "The influence curve and its role in robust estimation", JASA 69, 383-393.

[62] ExarchosT.P. Papaloukas, C. ; Fotiadis, D.I. ; Michalis,L.K. , "An association rule mining-based methodology for automated detection of ischemic ECG beats", IEEE Transactions on Biomedical Engineering. Volume: 53 , Issue: 8 Page(s): 1531 - 1540

[63] A. Vararuk, I. Petrounias, V. Kodogiannis, (2007) "Data mining techniques for HIV/AIDS data management in Thailand", Journal of Enterprise Information Management, Vol. 21 Iss: 1, pp.52 – 70

[64] Ordonez.C, et.al, "Mining constrained association rules to predict Heart disease", IEEE International Conference on Data Mining, pp.433-440.

[65] W. H. L. Jr, T. Masters, J. Y. Lo, D. W. McKee, and F. R. Anderson. "New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data". In 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications, pages 47–52, 2001.

[66] Rules for Melanoma skin cancer diagnosis, URL = http://www.phys.uni.torun.pl/publications/kmk/

[67] Nassif.et.al, "Information Extraction for Clinical Data Mining: A Mammography Case Study" Appears in Proceedings of the 2009 IEEE International Conference on Data Mining Workshops.

[68] Akinori Abe, Norihiro Hagita, Michiko Furutani, Yoshiyuki Furutani and Rumiko Matsuoka, "Categorized and Integrated Data Mining of Medical Data, Communications And Discoveries From Multidisciplinary Data", Studies in Computational Intelligence, 2008, Volume 123/2008, 315-330, DOI: 10.1007/978-3-540-78733-4_19.

[69] Sarabjot S. Anand, David A. Bell, John G. Hughes, "EDM: A general framework for Data Mining based on Evidence Theory, Data & Knowledge Engineering" Volume 18, Issue 3, April 1996, Pages 189–223

[70] Jau-Huei Lin, M.D. and Peter J. Haug, M.D. "Data Preparation Framework for Pre-processing Clinical Data in Data Mining", AMIA Annu Symp Proc. 2006; 2006: 489–493.

[71] Primary Children's Medical Center. Salt Lake City, UT, USA. (Personal Communication).

[72] Clayton PD, Narus SP, Huff SM, Pryor TA, Haug PJ, Larkin T, Matney S, Evans RS, Rocha BH, Bowes 3rd WA, Hoston FT, Gundersen ML. "Building a comprehensive clinical information system from components. The approach at Intermountain Health Care". Methods INF Med 2003; 42:1-7.

[73] S. Y. Hwang, C. P. Wei, and W. S. Yang, "Process mining: Discovery of temporal patterns from process instances," Computers in Industry, vol. 53 no. 3 pp. 345-364, 2004.

[74] Carolyn McGregor, Christina Catley and Andrew James, "A Process Mining Driven Framework for Clinical Guideline Improvement in Critical Care" 13th Conference on Artificial Intelligence in Medicine, Bled, Slovenia, July,2011.

[75] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. CRISP-DM 1.0,http://www.crisp-dm.org/download.htm

[76] Shearer, C. "The CRISP-DM Model: The New Blueprint for Data Mining". J. Data Warehousing. 5, 13-22 (2000)

[77] Catley, C., Smith, K., McGregor, C., Tracy, M. "Extending CRISP-DM to Incorporate Temporal Data Mining of Multidimensional Data Streams: a Neonatal Intensive Care Unit Case study." In: Computer-Based Medical Systems, pp. 1-5 (2009)

[78] Reza Sherafat Kazemzadeh, Kamran Sartipi and Priya Jayaratna, "A Framework for Data and Mined Knowledge Interoperability in Clinical Decision Support Systems". Clinical Data and Knowledge Interoperability, 2008.

[79] Health Level-7. URL = http://www.hl7.org. [Online; accessed 1-August-2008].

[80] Data Management Group (DMG). Predictive Model Markup Language (PMML) version 3.0 specification.URL = http://www.dmg.org/pmml-v3-0.html.

[81] Health Level 7. The Clinical Document Architecture (CDA) standard specification. URL = http://www.hl7.org. [Online; accessed 1-August-2008].

[82] Guideline Interchange Format (GLIF) 3.5 - technical specification. URL = http://smi-web.stanford. edu/projects/intermed-web/guidelines/GLIF TECH SPEC May 4 2004.pdf. [Online; accessed 1-August-2008], May 2004.

[83] Oakley, S. (1999). "Data mining, distributed networks and the laboratory". Health Management Technology, 20(5), 26-31

[84] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", Journal of Healthcare Information Management — Vol. 19, No. 2

[85] Ming Hua a, JianPei b,n, "Clusteringin applications with multiple data sources—A mutual subspace clustering approach", Neurocomputing 92 (2012) 133–144

[86] Schena, M. et al, "Quantitative monitoring of gene expression patterns with a cDNA microarray". Science 270:467-470 (1995).

[87] DeRisi J, et al. “Use of a cDNA microarray to analyze gene expression patterns in human cancer”. Nat Genet 1996 Dec; 14(4):457-60.

[88] Hegde P. et al. “A concise guide to cDNA microarray analysis”. Biotechniques. 2000 Sep; 29(3):548-50, 552-4, 556.

[89] Marchal K et al “Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays”. JOURNAL OF BIOLOGICAL SYSTEMS 10 (4): 409-430 DEC (2002).

[90] Chipping Forecast 1999, 2002, The Chipping Forecast. Special Supplement. Nature Genet. 21, Jan. 1999.

[91] Baldi P and AD Long. “A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes”. Bioinformatics, 17: 509- 519, (2001).

[92] Tusher VG et al. “Significance analysis of microarrays applied to the ionizing radiation response”. PNAS, 98:5116-5121, (2001).

[93] Dudoit S et al. “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”. Statistica Sinica, 12:111-139, (2002).

[94] Ideker, T. et al. “Testing for differentially-expressed genes by maximum likelihood analysis of microarray data”. Journal of Computational Biology, 7, 805-817 (2000).

[95] Eisen M. et al. “Cluster analysis and display of genome-wide expression patterns”. PNAS, 95:14863-14868 (1998).

[96] Tamayo P. et al. “Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation”. PNAS, 96:2907-2912, (1999).

[97] Hastie T. et al. “Supervised harvesting of expression trees”. Genome Biology, 2(1):research0003.1-0003.12, (2001).

[98] Li H and F. Hong. “Cluster-Rasch models for microarray gene expression data”. Genome Biology, 2(8)}:research0031.1-0031.13, (2001).

[99] Lin W. and C. Le “Model-based cluster analysis of microarray gene expression data”. Genome Biology, 3(2):research0009.1-0009.8, (2002).

[100] Golub T. et al. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”. Science, 286:531-537, 1999

[101] Alizadeh L. et al. “Identification of clinically distinct types of diffuse large B-cell lymphoma based on gene expression patterns”. Nature 403: 503-511 (2000).

[102] Bittner M. et al. “Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling”. Nature 406: 536-540 (2000)

[103] Ramaswamy S. et al. “Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures”, PNAS 98: 15149-15154.

[104] Tibshirani R, et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression" PNAS 2002 99:6567- 6572 (May 14).

[105] Ramaswamy S. et al. “Evidence for a Molecular Signature of Metastasis in Primary Solid Tumors”. Nature Genetics, vol. 33,January 2003, pp. 49-54.

[106] Milley, A. (2000). “Healthcare and data mining” Health Management Technology, 21(8), 44-47.

[107] Kolar, H.R. (2001). “Caring for healthcare”. Health Management Technology, 22(4), 46-47.

[108] Cios, K.J. & Moore, G.W. (2002). “Uniqueness of medical data mining”. Artificial Intelligence in Medicine, 26(1), 1-24.

[109] Ceusters, W. (2001). “Medical natural language understanding as a supporting technology for data mining in healthcare”. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica- Verlag Heidelberg, New York, 41-69.

[110] Megalooikonomou, V. & Herskovits, E.H. (2001). “Mining structure function associations in a brain image database”. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica-Verlag Heidelberg, New York, 153-180